KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

# Multimodal Fusion

Konstantinos Kontras

Supervisors:
Prof. dr. ir. M De Vos
Prof. dr. ir. J. Suykens
Dr. C. Chatzichristos

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

January 2025

# Multimodal Fusion

**Konstantinos KONTRAS**

Examination committee:
Prof. dr. ir. J De Schutter, chair
Prof. dr. ir. M De Vos, supervisor
Prof. dr. ir. J. Suykens, supervisor
Dr. C. Chatzichristos, supervisor
Prof. dr. ir. M. Blaschko
Prof. dr. ir. A. Bertrand
Prof. dr. ir. T. Tuytelaars
Prof. dr. Paul P. Liang
(MIT Media Lab and Department of Electrical Engineering and Computer Science)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

January 2025

© 2025 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Konstantinos Kontras, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

# Preface

The word PhD carries a certain vagueness, a shapelessness that shifts from person to person, place to place. To some, it is an emblem of intellectual pursuit; to others, a mere abstraction. Even among those who embark on this journey or have completed it, its meaning takes on different shades. Within my own family, the interpretations vary widely. Some believe I am still a student, while others, like my grandmother, think I have ascended to the rank of a professor, specially chosen for an elite calling. Some jokingly suspect I am evading taxes, while others remain simply puzzled by what it is I do.

This space allows me to reflect on and clarify what the PhD means, at least to me. By sharing these reflections, I hope those I wish to thank will recognize their part in this journey and understand the significance it holds.

At the very beginning, the PhD was, for me, a gateway to intriguing, "unsolved" problems, a way to avoid work I found tedious and uninspiring. How naive and insightful that notion was, all at once. In the past, I often found myself caught up in trivial tasks, spending days stuck on them. During my time as a bachelor's student in Patra, my friends taught me two important lessons that changed how I saw things. I want to thank them for helping me realize, first, that not everything is worth your time and attention, but for the things that truly excite you, you should give your all. Second, they showed me that together, we are far more capable, powerful, and able to reach further than we ever could alone. On those premises, this excitement and romanticism got me to pursue a PhD. While it is true that a PhD opens the door to research and discovery, it does so in ways I had not anticipated. The truth I came to understand was far more profound.

The day I met Maarten, my supervisor, we arranged a meeting under COVID restrictions, taking a quiet walk around the university campus. It was a brief encounter, just enough to exchange a few words and ensure we'd seen each other in person before committing to what would become years of collaboration.

As a Greek, I value tranquility and an easy-going attitude, and when Maarten arrived with his summer-like ease, I noticed something familiar in it, a sense of balance and calmness that stood out. Time would prove me right.

Then there was Christos, my daily supervisor, who had been present at the football game where, a year before, I got a black eye during my first ten days in Leuven. He never missed the chance to tease me about it, even incorporating it into his sentence welcoming me to the group. Christos turned out to be the one who'd listen to my problems and somehow make them feel lighter before I even had to deal with them, probably without him even realizing it.

In those early PhD years, I enjoyed much freedom mixed with lighthearted moments, driven by an inner need to feel productive. Over time, this shifted. What began as outward fun, long vacations, trips and much of free time, but also accompanied with inner struggles transformed into outward challenges at work and a deep sense of inner fulfillment. Over the course of four years, Maarten, Christos, and I held roughly one hundred meetings. I stopped counting after number 53 in September 2022. I hold great gratitude for their time, their patient and their help in all of this. Somewhere in the process, I realized something important: it was never about Maarten or Christos guiding my path or driving my knowledge and career. The PhD journey was about me taking the wheel; bringing them the problems, the solutions, and the appropriate initiatives. It was never about becoming a good coder, mathematician or biomedical expert, it was all about thinking actively and using this thought to use my time into something productive. That ended up being, eventually, my definition of the term PhD and the phase of my life that evolved my thinking, processing, and dreaming capabilities the most.

Such a transformation became most apparent during one intense week in May 2024, while preparing a paper submission, that ultimately didn't make it. Yet, the progress made turned it into the most fulfilling academic week of my decade-long university journey. Exhausted but energized, I realized the journey had truly found its purpose. Much of this was thanks to Thomas, who stayed late on all of these days and responded to every idea or question, driven by a sense of "I want to" rather than "I have to", and Prof. Matthew Blaschko, who inspired me with his enthusiasm and fresh ideas. Together with Maarten and Christos, they made that week unforgettable.

Maarten's calm and laid-back attitude transformed into a sense of freedom, one that allowed me to explore, grow, and follow paths beyond what he might have personally chosen for me. I recognize this as the most important quality in a PhD advisor. While it can feel uncomfortable at times, as though you're left hanging, you eventually learn to swing. I hope these words convey more than just a simple thank you, with the continued collaboration between us serving

as proof of my appreciation.

At this point, I would like also to express my gratitude to Prof. Johan Suykens, my co-supervisor, for his invaluable feedback throughout the years, the nice collaboration in the SVM course and the great hallway conversations, and to Prof. Alexander Bertrand and Prof. Matthew Blaschko for their contributions to the supervisory committee both with meeting and detailed feedback. I would also like to thank Prof. Tinne Tuytelaars and Prof. Paul Pu Liang for honoring me with their participation in the examination committee and their effort to come with meaningful questions and contributions, and Prof. Joris De Schutter for chairing the defenses. Altogether, their insightful questions and feedback have been instrumental in shaping the thesis as it stands today.

Shifting to the personal side of the PhD, a different journey unfolds. The personal life of a PhD follows a similar path to the professional one, it's a rollercoaster. Some moments feel like pure freedom and others feel like well-crafted torture.

The person who has supported me in this journey with all her heart, and who truly deserves an award for patience and unwavering support, is Aigli. She has stood by me when I had to arrive hours late or cancel plans at the last minute, when I couldn't sleep because I was tossing and turning with ideas and stress, or when I carried my worries and schedule everywhere. She has helped me navigate these challenges without asking for her piece. I only hope I've been able to support her, even if just a fraction of the way she has supported me. On the brighter side, Aigli is also the one who knows how to bring joy and remind you of the simple pleasures in life. Whether through well-planned trips that showcased her impressive surfing skills, spontaneous road trips where she taught me how to survive being hungry in France after 10 pm or hiking under 35-degree heat at the southernmost point of Europe in August, Aigli has a unique ability to unlock that inner childlike happiness.

Next, I would like to mention my parents, my sister, and the rest of my family. Their unwavering support and belief in me, along with the faith they had long ago, have brought me to where I am today. I can't pinpoint a single moment, but their willingness to listen to my struggles, knowing that sometimes all I needed was an attentive ear, and their shared excitement and stress for the first paper that marked a milestone in their eyes, stand out. While conventional success might not be carried out, I'm certain that many more celebrations, like the first graduation in Patra, are yet to come, each carrying its own meaning of success in my life.

Another special part of this journey has been my flatmates through all these years, friends in the form of brothers, who I've had the chance to laugh, argue,

drink, and tease with. They listened to my problems and made them their own. From El Gianno, the survivor with his ability to integrate into any situation while always caring for others, to Andreas, the ultimate Erasmus, who shaped Maarten's impression of me with his constant ideas for another drink the night before an early morning meeting, to Periklis, the chief of contradictions, whose discussions halfway up the stairs would last for hours and who brought a unique vibe with his early morning music taste, to Manos, the hairy fairy, whose resilience to bullying, patience in urgent situations, disgust for the ordinary, and excitement over bad jokes I never understood, and along with Evelina, my comadre, they have all made Favierou and Geldenaaksebaan truly feel like home.

Then, I should bring up my childhood friends and beyond, those I cannot say good words if I know that they may read but those who have played a crucial role without participating in the daily academic life. It was them whose one phone call was enough to change their whole schedule and put me back in their lives the moment I set foot in Athens or Heraklion. Those who didn't care if the PhD was going well or not, were sure it would do and invited problems of such kind to be left back and enjoy the moment and the terrible jokes. It is simple moments that remind you that all these troubles of daily life are just something on top; you pass them to improve upon the directions you choose but you already made it as long as you have those simple moments.

Lastly, I want to talk about Biomed. Joran, the ravecaver with the best chimichurri in Leuven and his trusty steak thermometer, Elisabeth, the star of our office who takes bold steps in exploring the world, and Lore, with the reddest drunk face and the most responsible party-animal energy, have been the ultimate trio of officemates. With Joran, we're lucky that Elisabeth and Lore help keep us from getting lost in endless conversations unless it's about a new trip destination or some gossip, then work flies out of the window. Miguel, with his award-worthy mustache, legendary powernaps and bbqs, and readiness for breaks; Aggeliki, who I've proudly been her buddy, integrated well with her energy, photos from her village, and post-vacation sweets; Nefeli, the mythomatiac whose tales could fill a fiction novel; Kostas, whose laugh while teasing can lure the whole office into a break; Christos, who occasionally gives us with his goalkeeping skills and beer-pong precision while promoting group unity; Michiel and Cem, secretly playing hardcore techno from their laptops; Marta, the most dedicated and talented pizzamaker; Guido, the epitome of Italian style; Louis, with his remarkable calmness; Tim, always up for drunk philosophy on the way back from parties; Charles, inspiring us with his attitude and side projects; Thomas, whose availability, board game knowledge, and secret cycling streams in the office are unmatched; George, the enigmatic figure who keeps us guessing but always surprises; Phu, bringing smiles and good vibes; Helena, the

quiet yet caring presence of the office; Simon, with his big bets, great singing, and life-giving energy; Iustina and Teo, our proud new parents, sure to be an amazing duo; Anna and little Maarten, the Bonnie and Clyde of fresh ideas and funny discussions; Lennert, the most Greek-spirited Belgian, always ready to follow or spark any plan; the football team, which I've enjoyed playing with for years, and the volleyball team, with whom we've recently celebrated winning trophies for both sports. Together, along with all the Biomed members, they create a warm, welcoming atmosphere in the office, making it feel less like work and more like a place where you're encouraged to grow and thrive.

There's a line in a song that defines success by the number of glasses you place on your table, and I've been fortunate enough to have many people supporting me throughout these years. Without them, I wouldn't have been able to pass through this phase, evolving my way of thinking, shifting my life perspectives, and understanding what truly matters to me. A part of this thesis, and a part of what's yet to come, is dedicated to all these amazing people who, in one way or another, have been there for me. I aim at the finishing line to have made you all proud.

*Yours truly,*
*Konstantinos*

*Leuven,*
*January 2025*

# Abstract

Multimodal machine learning represents a pivotal field of research, leveraging diverse data sources to improve model performance, robustness, and applicability across a wide range of tasks. This dissertation investigates key challenges and advancements in multimodal learning, with a primary focus on multimodal fusion, and the optimization and generalization difficulties during training processes, collectively described as multimodal competition. By integrating multiple modalities such as images, text, audio, and physiological signals, the thesis addresses fundamental questions about the role of fusion points in a network, their effect on task performance and robustness, and strategies for achieving balanced training across all modalities. It explores how to enhance modality-specific contributions, mitigate competition, and improve information synergy in multimodal systems.

The first part of the dissertation examines multimodal fusion, emphasizing how architectural choices and training objectives affect task performance and robustness, especially under noisy or incomplete data conditions. It provides a comprehensive analysis of early, intermediate, and late fusion approaches, highlighting their unique advantages and trade-offs. The research introduces a coordinated representation architecture as a dynamic mechanism for selecting optimal fusion strategies, while also investigating the role of training objectives in enhancing model resilience to noise and missing modalities. This integrated exploration offers insights into designing robust multimodal systems that bridge the gap between controlled experimental designs and practical, real-world applicability.

The second part of the thesis addresses modality competition, a significant challenge in multimodal joint training where one modality dominates, impairing the contributions and learning of others. Through a gradient analysis, the root mechanism that allows competition to occur is identified. The work provides an in-depth discussion of existing balancing strategies, including methods that use unimodal performance to modulate gradient magnitudes and those that balance

models by incorporating unimodal losses. Observing performance inconsistencies in both approaches, this research proposes a new method, *Multi-Loss Balancing*, that integrates the strengths of both strategies. By combining unimodal losses with a more flexible modulation of both multimodal and unimodal gradients into one framework, the method achieves consistently improved performance in multimodal training.

The thesis extensively evaluates existing multimodal balancing methods, highlighting the persistent challenge of retaining unimodal information. To shed light on this issue, we propose an evaluation setup that compares new methods against unimodally pretrained approaches, such as ensembles and joint training with pretrained encoders. This setup highlights the benefits of maintaining unimodal information through pretraining and the potential for enhanced synergy achieved by multimodal training from scratch, providing a comprehensive basis for future assessments.

To address the challenge of balancing modality contributions while maintaining unique unimodal information, the thesis introduces the Multimodal Competition Regularizer (MCR). MCR quantifies unique information contributions from each modality using conditional mutual information, offering a principled and computationally efficient approach to evaluate and balance their roles. The proposed regularization loss function encourages both unique and shared information contributions but recognizes competing objectives where increasing the influence of one modality eventually reduces the contributions of others. These conflicts are addressed using game-theoretic strategies, with each modality forming a player competing to maximize its influence over the output. This adaptive approach enables MCR to regulate modality interactions effectively, preserving modality-specific insights while enhancing cross-modal synergies. Notably, MCR is the first method to significantly outperform all prior balancing techniques and unimodally pretrained alternatives, establishing a new benchmark in multimodal machine learning.

This thesis highlights the importance of effective training methodologies in narrowing the gap between the theoretical potential of multimodal learning and the practical improvements demonstrated. It shows that well-designed fusion strategies, like MCR, can enhance performance while leaving room for future advancements. Notably, these advancements can reduce reliance on large-scale data and models, as improvements in some cases are comparable to those achieved through scaling. The thesis concludes with optimism about the role of multimodal learning in shaping the next generation of artificial intelligence and its potential to drive meaningful advancements in the field.

In summary, this dissertation advances the field of multimodal learning by addressing key challenges and proposing novel solutions in fusion strategies

and modality competition. It introduces a dynamic fusion architecture to enhance both performance and robustness to faulty data and presents two innovative methods to tackle the optimization and generalization issues arising from multimodal competition. By framing joint multimodal training as a game where modalities compete for influence and resources, the work offers a new perspective on balancing contributions. Additionally, it proposes a structured benchmarking setup to guide future research and ensure meaningful progress in the field. Ultimately, the thesis highlights the significant benefits of balanced training strategies in unlocking the full potential of multimodal systems.

# Beknopte samenvatting

Multimodale machine learning combineert diverse gegevensbronnen om modelperformantie, robuustheid en toepasbaarheid te verbeteren bij uiteenlopende taken. Dit proefschrift richt zich op twee kernvragen: hoe kunnen modaliteiten optimaal worden gecombineerd, en hoe kunnen modellen effectief en evenwichtig worden getraind ondanks multimodale competitie? Het onderzoek verkent de rol van fusiepunten in neurale netwerken, hun impact op taakperformantie en robuustheid, en strategieën voor gebalanceerde training over alle modaliteiten. Door modaliteiten zoals afbeeldingen, tekst, audio en fysiologische signalen te combineren, worden interacties en synergieën binnen multimodale systemen geanalyseerd.

Het eerste deel van dit proefschrift behandelt multimodale fusie en onderzoekt hoe architecturale keuzes en trainingsdoelen de taakperformantie en robuustheid beïnvloeden, met name bij ruis of onvolledige gegevens. Het biedt een uitgebreide analyse van vroege, intermediaire en late fusiebenaderingen, met een nadruk op hun unieke voordelen en afwegingen. Dit deel introduceert een gecoördineerde representatie-architectuur als een dynamisch mechanisme om optimale fusiestrategieën te selecteren. Daarnaast wordt de rol van trainingsdoelen onderzocht bij het vergroten van de modelbestendigheid tegen ruis en ontbrekende modaliteiten. Deze geïntegreerde aanpak biedt waardevolle inzichten voor het ontwerpen van robuuste multimodale systemen die de kloof overbruggen tussen gecontroleerde experimentele opstellingen en praktische, realistische toepassingen.

Het tweede deel van dit proefschrift richt zich op modalitycompetitie, een belangrijke uitdaging bij gezamenlijke multimodale training, waarbij één modaliteit domineert en de bijdragen en het leren van andere modaliteiten belemmert. Door middel van een gradiëntanalyse wordt het onderliggende mechanisme dat competitie mogelijk maakt geïdentificeerd. Het werk biedt een diepgaande bespreking van bestaande balanceringsstrategieën, waaronder methoden die unimodale prestaties gebruiken om gradiëntgrootheden te

moduleren, en methoden die modellen balanceren door unimodale verliezen te integreren. Observatie van prestatie-inconsistenties in beide benaderingen heeft geleid tot de ontwikkeling van een nieuwe methode, Multi-Loss Balancing, die de sterke punten van beide strategieën combineert. Door unimodale verliezen te combineren met een flexibelere modulatie van zowel multimodale als unimodale gradiënten in één raamwerk, bereikt de methode consequent verbeterde prestaties in multimodale training.

Het proefschrift evalueert uitgebreid bestaande multimodale balanceringsmethoden en benadrukt de terugkerende uitdaging om unimodale informatie te behouden. Om deze kwestie te belichten, stellen we een evaluatie strategie voor die nieuwe methoden vergelijkt met unimodaal voorgetrainede benaderingen, zoals ensembles en gezamenlijke training met voorgetrainede encoders. Deze strategie benadrukt zowel de voordelen van het behouden van unimodale informatie door pretraining als de mogelijkheden voor verbeterde synergie via multimodale training vanaf nul, en biedt een uitgebreide basis voor toekomstige evaluaties.

Om het evenwicht tussen de bijdragen van verschillende modaliteiten te verbeteren en tegelijkertijd unieke unimodale informatie te behouden, introduceert dit proefschrift de Multimodal Competition Regularizer (MCR). MCR kwantificeert de unieke informatie bijdragen van elke modaliteit door gebruik te maken van conditionele wederzijdse informatie, wat zorgt voor een principiële en efficiënte manier om hun rollen te evalueren en in balans te brengen. De voorgestelde regularisatie lossfunctie bevordert zowel unieke als gedeelde informatie bijdragen, maar erkent concurrerende doelstellingen waarbij het vergroten van de invloed van één modaliteit uiteindelijk de bijdragen van andere modaliteiten vermindert. Deze conflicten worden aangepakt met strategieën uit speltheorie, waarbij elke modaliteit optreedt als een speler die probeert de invloed op de output te maximaliseren. Dankzij deze adaptieve aanpak kan MCR de interacties tussen modaliteiten effectief reguleren, unieke inzichten per modaliteit behouden en synergieën tussen modaliteiten versterken. Opmerkelijk is dat MCR de eerste methode is die alle eerdere balanceringsmethoden en unimodaal voorgetrainede alternatieven significant overtreft, waarmee het een nieuwe standaard zet in multimodale machine learning.

Dit proefschrift benadrukt het belang van effectieve trainingsmethoden om de kloof te verkleinen tussen het theoretische potentieel van multimodaal leren en de praktische verbeteringen die worden aangetoond. Het toont aan dat goed ontworpen fusie strategieën, zoals MCR, de prestaties kunnen verbeteren en ruimte bieden voor toekomstige vooruitgang. Deze vooruitgang kan met name de afhankelijkheid van grootschalige gegevens en modellen verminderen, aangezien verbeteringen in sommige gevallen vergelijkbaar zijn met die behaald door opschaling. Het proefschrift sluit af met optimisme over de rol van multimodaal

leren bij het vormgeven van de volgende generatie kunstmatige intelligentie en de potentiële betekenisvolle vooruitgang die het veld kan boeken.

Samenvattend draagt dit proefschrift bij aan het veld van multimodaal leren door belangrijke uitdagingen aan te pakken en nieuwe oplossingen te introduceren voor fusie strategieën en modaliteiten competitie. Het introduceert een dynamische fusie-architectuur die zowel de prestaties als de robuustheid verbetert bij defecte gegevens en biedt twee innovatieve methoden om de optimalisatie- en generalisatieproblemen als gevolg van multimodale competitie aan te pakken. Door gezamenlijke multimodale training te kaderen als een spel waarin modaliteiten strijden om invloed en middelen, biedt dit werk een nieuw perspectief op het balanceren van bijdragen. Bovendien wordt een gestructureerde benadering voorgesteld voor het opzetten van benchmarks, ter ondersteuning van toekomstig onderzoek en het waarborgen van betekenisvolle vooruitgang in het veld. Het proefschrift benadrukt uiteindelijk de aanzienlijke voordelen van uitgebalanceerde training strategieën bij het volledig benutten van het potentieel van multimodale systemen.

# List of Abbreviations

**Adam** Adaptive moment estimation optimizer. 43

**AGM** Adaptive Gradient Modulation. 112

**AL** Alignment Loss. 38

**BERT** Bidirectional Encoder Representations from Transformers. 17

**CA** Cross-Attention. 37

**CE** Cross-Entropy. 38

**CLIP** Contrastive Language–Image Pretraining. 16

**CLS** Class Token. xxv, 35

**CMI** Conditional Mutual Information. xxvii, 103

**CNN** Convolutional Neural Networks. 30

**CoRE** Coordinated Representation. 26

**CPC** Contrastive Predictive Coding. 17

**ECE** Expected Calibration Error. 82

**EEG** Electroencephalography. 34

**EOG** Electrooculography. 34

**FF** FeedForward network. 36

**FiLM** Feature-wise Linear Modulation. 80

**fMRI** Functional Magnetic Resonance Imaging. 28

**GPU** Graphical Process Unit. 43

**InfoNCE** Contrastive Loss used for learning representations. 38

**JSD** Jensen-Shannon Divergence. 108

**MCE** Multimodal Competition Error. 103

**MCR** Multimodal Competition Regularizer. 102

**MEG** Magnetoencephalography. 28

**MI** Mutual Information. 102

**MIPD** Mutual Information Perturbation Divergence. 107

**MLP** Multilayer Perceptron. 36, 80

**MMCosine** Multimodal Cosine Similarity. 112

**MS** Multi-Supervised Loss. 38

**MSLR** Modality-Specific Learning Rates. 112

**MTCNN** Multi-Task Cascaded Convolutional Neural Network. 79

**NNs** Neural Networks. 30

**NREM** Non-Rapid Eye Movement. 29

**OGM** On-the-fly Gradient Modulation. 112

**PMR** Prototypical Modal Rebalance. 112

**PSG** Polysomnography. 29

**ReLU** Rectified Linear Unit. 36

**REM** Rapid Eye Movement. 29

**RLHF** Reinforcement Learning from Human Feedback. 17

**RNN** Recurrent Neural Networks. 30

**SA** Self-Attention. 36

**SE** Squeeze-and-Excite. 78

**SHHS** Sleep Heart Health Study. 27, 33

**SSL** Self-Supervised Learning. 17

**STD** Standard Deviation. 43

**STFT** Short-Time Fourier Transform. 32, 34

**TF** Transformer. xxv, 35

# Contents

**Abstract** vii

**Beknopte samenvatting** xi

**List of Abbreviations** xvii

**Contents** xix

**List of Figures** xxiii

**List of Tables** xxix

**1 Introduction** 1
- 1.1 Philosophical Note . . . 3
- 1.2 Multimodal Machine Learning Landscape . . . 5
  - 1.2.1 Multimodal Data . . . 5
  - 1.2.2 Multimodal Tasks . . . 7
- 1.3 Motivation for Multimodal Learning . . . 8
  - 1.3.1 Modality-Specific Additional Information . . . 9
  - 1.3.2 Robustness and Calibration via Shared Information . . . 9
  - 1.3.3 Emergent Cross-Modal Insights . . . 10
- 1.4 Approaches to Multimodal Learning . . . 10
  - 1.4.1 Ensembles . . . 11
  - 1.4.2 Fusion Levels in Multimodal Networks . . . 12
  - 1.4.3 State-of-the-Art Multimodal Transformers . . . 15
- 1.5 Training Strategies for Multimodal Learning . . . 17
- 1.6 Open Research Questions to Address . . . 18
  - 1.6.1 Multimodal Representation and Fusion . . . 18
  - 1.6.2 Multimodal Competition . . . 19
- 1.7 Research Objectives and Overview . . . 19
  - 1.7.1 Research Objectives . . . 19

1.7.2 Chapter Overview . . . . . 21

I Designing Multimodal Models 23

2 Multimodal Application: Sleep Staging 25
2.1 Introduction . . . . . 27
2.1.1 Physiological Signals . . . . . 28
2.1.2 The task of Sleep Staging . . . . . 28
2.1.3 Automating Sleep Staging with Deep Learning . . . . . 30
2.1.4 Multimodal Sleep Staging . . . . . 31
2.1.5 Our Contribution . . . . . 32
2.2 Data & Methods . . . . . 33
2.2.1 Data . . . . . 33
2.2.2 Transformer Backbone . . . . . 34
2.2.3 The CoRe-Sleep Architecture . . . . . 36
2.2.4 Training Objectives . . . . . 38
2.2.5 Architecture Benchmarks (Early & Mid-Late Fusion) . . 40
2.2.6 Implementation . . . . . 43
2.2.7 Experiments . . . . . 43
2.3 Results & Discussion . . . . . 44
2.3.1 Multimodal . . . . . 44
2.3.2 Missing Modality . . . . . 44
2.3.3 Noisy Modalities . . . . . 47
2.3.4 Training with Incomplete Data . . . . . 47
2.4 Discussion and Conclusion . . . . . 49
2.4.1 What Matters Most in Multimodal Learning? . . . . . 50

II Multimodal Competition 53

3 Multimodal Competition: The Phenomenon 55
3.1 Introduction . . . . . 57
3.2 Gradient Interactions . . . . . 58
3.3 Review of Previous Balancing Strategies . . . . . 59
3.3.1 Gradient Balancing Methods . . . . . 60
3.3.2 Multi-Task Methods . . . . . 63
3.3.3 Additional Methods . . . . . 66
3.4 Conclusion . . . . . 68

4 Improving Multimodal Training with Multi-Loss Gradient Modulation 71
4.1 Introduction . . . . . 73
4.2 Multi-Loss Balancing Gradients . . . . . 75

4.2.1 Model Structure . . . 75
4.2.2 Balancing . . . 75
4.3 Experiments . . . 77
4.3.1 Datasets . . . 77
4.3.2 Models . . . 79
4.3.3 Experimental Results . . . 82
4.4 Coefficients Analysis . . . 84
4.5 Computational Complexity Analysis . . . 85
4.6 Discussion and Conclusion . . . 86

**5 Is there benefit in training multimodal models from scratch? 87**
5.1 Introduction . . . 89
5.2 Additional Methodology . . . 90
5.2.1 Unimodal Pretrained Baselines . . . 90
5.3 Experimental Setting . . . 91
5.3.1 Datasets . . . 91
5.3.2 Balancing Methods . . . 91
5.3.3 Models . . . 92
5.4 Results . . . 93
5.4.1 Effectiveness of Balancing methods . . . 93
5.4.2 Post-Hoc Multimodal Analysis . . . 95
5.5 Discussion . . . 98

**6 Regularizing Multimodal Competition 99**
6.1 Introduction . . . 101
6.2 Analysis of Multimodal Competition . . . 104
6.3 Multimodal Competition Regularizer . . . 105
6.3.1 Approximating MI Terms . . . 107
6.3.2 The Game of Multimodal Fusion . . . 109
6.3.3 Perturbations . . . 110
6.4 Experiments . . . 111
6.4.1 Synthetic Dataset . . . 112
6.4.2 Real-World Datasets . . . 113
6.4.3 Ablation Study . . . 116
6.4.4 Analysis of Multimodal Error . . . 117
6.5 Discussion . . . 119

**7 Conclusion 121**
7.1 Main Findings Part 1: CoRe-Sleep . . . 122
7.2 Main Findings Part 2: Multimodal Competition . . . 123
7.2.1 Understanding Modality Competition . . . 123
7.2.2 Methods to tackle Multimodal Competition . . . 123
7.2.3 Proposed Advancements in Balancing Approaches . . . 124

7.2.4 Empirical Evidence . . . 126
7.2.5 Key Insights and Limitations . . . 127
7.3 Future Work . . . 129
7.3.1 Exploring Strategies for the Multimodal Fusion Game . 129
7.3.2 Optimizing Synergy in MCR . . . 130
7.3.3 Investigating Multimodal Competition in Multimodal Translation . . . 130
7.3.4 Characterizing Multimodal Diversity & Training Challenges 131
7.3.5 Breaking the Egg-Chicken Cycle: Discovering and Balancing Modalities . . . 131
7.3.6 Expanding the Prerequisite Assumptions of Existing Methodology . . . 132
7.4 Ethical and Societal Implications . . . 132
7.5 Closing Note & Final Thoughts . . . 135

**A Appendix MLB** **137**
A.1 Calibration Error . . . 137
A.2 Implementation . . . 137
A.2.1 Balancing Hyperparameters . . . 138
A.2.2 Training Hyperparameters . . . 138
A.2.3 Dataset Splits . . . 139
A.3 Tabular Presentation of Results . . . 140

**B Appendix MCR** **141**
B.1 Experimental Setting . . . 141
B.1.1 Datasets . . . 141
B.1.2 Models: Backbone Unimodal Encoders . . . 142
B.1.3 Experimental Details . . . 143
B.1.4 MCR on M modalities . . . 144
B.1.5 Computational Speed and Memory Analysis . . . 145
B.1.6 Comparing Perturbation methods . . . 145
B.2 Proof of Supervised Contrastive Loss as Lower Bound . . . 146

**Bibliography** **149**

**Curriculum vitae** **171**

**List of Publications** **173**

**Acknowledgements** **175**

# List of Figures

1.1 Venn diagram illustrating the decomposition of information between two modalities, $X_1$ and $X_2$, highlighting regions of unique information (specific to each modality), shared information (common across both modalities), and emergent information (arising uniquely from the interaction between $X_1$ and $X_2$, indicated with the shaded areas). . . . . 9
1.2 Illustration of multimodal fusion points: (a) Early Fusion, where modalities are integrated at the input layer; (b) Late Fusion, combining modalities at the output layer; (c) Intermediate Fusion, merging modalities at middle network layers; and (d) Intermediate Fusion with Informational Exchange, where cross-modal interactions occur at earlier layers. The symbol $\oplus$ refers to an aggregation operation, such as summation or concatenation, used to combine information from the different modalities. . . . 13
2.1 A typical hypnogram over a full night of sleep (6-8 hours). The hypnogram shows the cyclical transition through NREM stages ($N_1$, $N_2$, $N_3$) and REM sleep, along with periods of wakefulness. Sleep cycles repeat every 90 minutes, with REM sleep increasing in duration as the night progresses and deep sleep ($N_3$) decreasing. 30
2.2 Schematic of a sequence-to-sequence model for sleep staging. The model processes raw input data (e.g., EEG) through multiple Epoch Encoders, which encode individual 30-second epochs into latent space representations. These latent representations are fed into a Sequence Encoder to capture temporal dependencies across multiple epochs. The model outputs sleep stage predictions such as wakefulness (W) and NREM stages (N1, N2, N3). The dashed lines indicate that the weights are shared among the Epoch Encoders. . . . . 31

2.3 In this figure the general layout of the CoRe-Sleep [90] architecture is shown. EEG and EOG data are processed by unimodal encoders to create unimodal representations and then fed to multimodal encoders for cross-attention. Supervised losses are calculated for each representation. . . . . 34
2.4 In this figure are depicted the specific components of CoRe-Sleep [90]architecture. More specifically, we depict the detailed processes of one of the unimodal and one of the multimodal encoders. Inner and outer sequences are processed by Transformer (TF) blocks, using a [CLS] embedding for summarization. Cross-attention between modalities creates coordinated representations. 35
2.5 **a)** Mid-Late fusion scheme (MS + AL): Each modality has its encoder, and information gets only merged before the multimodal predictor. The Mid-Late model refers to the model trained solely on the Multimodal loss, while the models incorporating MS and AL are explicitly identified. Similar to CoRe-Sleep, this Mid-Late fusion scheme can derive both unimodal and multimodal predictions in a single forward pass. **b)** Early fusion scheme (MS + AL): In the Early model, EEG and EOG embeddings are processed by a shared encoder to calculate the multimodal loss. The Early model is defined again without the MS and AL losses. Unlike the Mid-Late scheme, the Early model requires two forward passes to compute the unimodal representations. . 42
2.6 This figure provides an illustration of a patient from SHHS-1 whose EEG modality was highly corrupted in the second half of the recording. The top section of the figure displays the EEG (C4-A1) and EOG (L-R) signals, with a sudden increase in the standard deviation of the EEG signal indicating the presence of noise. The two rows below show the class predictions (hypnogram) of the unimodal EEG and EOG models and the multimodal EEG-EOG CoRe-Sleep model. Notably, in the highly noisy area, the EEG model performed poorly, while the multimodal model remained robust by leveraging the information provided by the EOG. . . . . 46

2.7 CoRe-Sleep was trained using a sub-group of 100 multimodal patients who had both EEG and EOG, as well as some unimodal patients who only had EEG or EOG: **a)** We add patients who provide only EEG and an equal number of patients who provide only EOG. We consistently observe an improvement in the performance of both multimodal and unimodal predictors. **b)** We add patients who provide only EEG. A decrease in EOG predictor performance is noted. **c)** We add patients who provide only EOG. Consistently, we notice a decrease in the modality performance that was not provided. All experiments were repeated three times with different dataset splits to obtain the variance shown. . . . 48
4.1 Categorization of state-of-the-art balancing methodology: (a) Gradient Balancing methods use estimates of unimodal performance to calculate coefficients ($k_a$ and $k_v$) and employ these to balance the multimodal loss. (b) Multi-Task methods incorporate unimodal classifiers into the model, each noted as CLS Head, to better estimate unimodal performance. The coefficients ($k_a$ and $k_v$) derived from comparing unimodal performance are used exclusively to balance the unimodal losses. (c) The proposed Multi-Loss Balancing method combines both strategies by incorporating unimodal classifiers for accurate unimodal performance estimation and balancing both multimodal and unimodal losses. . . . . 74
4.2 Comparing Balancing Coefficients ($k_i$, Eq. 4.7) and Performance Ratios ($r_i$, Eq. 4.5) across different $\alpha$ and $\beta$ settings. . . . . 78
4.3 Accuracy of models that differentiate by the backbone encoders (colours), the fusion strategies (Late with a linear and Mid with an MLP classifier), and the balancing techniques (x-axis). Across all datasets, results demonstrate that employing unimodal losses within the Multi-Loss framework and balancing them consistently yields the best performance. . . . . 81
4.4 ECE Comparison on Conformer CREMA-D. . . . . 82
4.5 Accuracy of models using FiLM [130], Gated [85] and TF fusion techniques on the ResNet and Conformer model applied to the CREMA-D dataset. . . . . 83
4.6 Comparison of balancing coefficients evolution of coefficients that amplify the contribution of each unimodal encoder's learning, smoothed over a 1k-point kernel. It illustrates that while all methods provide some balancing, only Multi-Loss Balancing (MLB) achieves self-muting when modalities have converged. . 84

5.1 Accuracy(%) difference compared to the Ensemble baseline for various methods across three datasets: CREMA-D, AVE, and UCF. Results are shown for two backbone models: Conformer (orange) and ResNet (blue). Positive values indicate an improvement over the Ensemble, while negative values represent a performance drop. Most existing methods fail to outperform the Ensemble, which does not utilize any multimodal learning. 94

5.2 Synergy vs. Routing Matrix for Multimodal and Unimodal Predictions. The matrix shows the overlap between correct and wrong predictions, highlighting emergent cross-modal information (Synergy) and the model's ability to leverage the best modality (Routing). Please note that the colours used here are not intended to convey any specific meaning. . . . . 95

5.3 Error comparison matrices across datasets and models, contrasting unimodal predictions with multimodal methods (Ensemble, Joint Training, AGM, MLB, and Uni-Pre Finetuned). Columns represent cases where both unimodal predictions fail or at least one succeeds. AGM and MLB excel at identifying emergent information in cases where unimodal models fail, while unimodally pretrained methods, such as Ensemble and Uni-Pre Finetuned, perform better when at least one unimodal prediction is correct. . . . . 97

6.1 (Left) Illustration of the conditional mutual information (CMI) terms, $\text{CMI}_1 : I(X_1; Y \mid X_2)$ and $\text{CMI}_2 : I(X_2; Y \mid X_1)$, representing the unique contributions ($U_1$ and $U_2$) of each modality to the target. The shared task-relevant information ($S$) between the modalities is defined as $I(X_1; X_2) - I(X_1; X_2 \mid Y)$. (Right) Accuracy as a function of the ratio between the unique information ($U_1$) from modality $X_1$ and the shared information ($S$) between the modalities. Synthetic data are generated as $X_1 = N_1 + Y, X_2 = N_1 + Y$ where $N_1, N_2$ are independent noise for each modality. We consider $S$ the percentage of the datapoints that both modalities have information about the label $Y$, $U_1$ and $U_2$ when only one has with the other modality equating to noise for those datapoints. In the experiment we keep $U_2$ constant while changing $U_1$ and $S$. As $U_1$ increases and $S$ decreases, accuracy deteriorates, reflecting intensified multimodal competition. Among the various methods, including Singleloss, Multiloss, Ensemble, unimodally pretrained and finetuned encoders (Uni-Pre Fine), OGM [129], AGM [96], and MLB [89]our regularization method MCR demonstrates a slower decline in accuracy. For further detals prease refer to Section 6.4.1 103

6.2 **Multimodal Competition Regularizer (MCR):** The diagram illustrates the MCR framework, which mitigates modality competition in multimodal learning. Raw data ($X_1$ and $X_2$) are encoded into latent representations ($Z_1$ and $Z_2$), which are then permuted to create $\tilde{Z}_1$ and $\tilde{Z}_2$ and the paired combinations. These combinations are passed through the Fusion Network to produce predicted outputs ($Y$, $\tilde{Y}_1$, $\tilde{Y}_2$). The comparison between predictions reveals each modality's contribution. For example, if $Y \approx \tilde{Y}_1$, it shows that $X_1$ has little impact, and the model relies on $X_2$. The MCR loss includes three components: $\mathcal{L}_{MIPD}$ maximize the Jensen-Shannon divergence (JSD) between task output and permuted modality predictions. $\mathcal{L}_{\text{Con}}$ aligns modality representations, while $\mathcal{L}_{\text{CEB}}$ penalizes task-irrelevant information by reconstructing back to the latent space. Further analysis for more than two modalities is in Appendix B.1.4. . . . . 106

6.3 This figure illustrates a key aspect of our training process, showing how competition strategies between modalities are applied. The gradient multiplier adjusts the video encoder's response to Audio Importance ($\text{Importance}_2$). When $k = 1$, the video encoder enhances $\text{Importance}_2$; at $k = 0$, it remains neutral, and at $k = -1$, it competes by reducing $\text{Importance}_2$ to prioritize its own ($\text{Importance}_1$). This reflects the principle that increasing the importance of one modality can reduce the importance of the other. . . . . 109

6.4 Accuracy(%) difference compared to the Ensemble baseline for various methods across three datasets: CREMA-D, AVE, and UCF101. Results are shown for two backbone models: Conformer (orange) and ResNet (blue). Positive values indicate an improvement over the Ensemble, while negative values represent a performance drop. Most existing methods fail to outperform the Ensemble, which does not utilize any multimodal learning. In contrast, our proposed method, MCR, consistently achieves positive accuracy differences, surpassing both the previous suggested methods and our own defined baselines. . . 115

6.5 Error comparison matrices across pairs of datasets and models, comparing unimodal predictions with multimodal models trained using various methods, including Ensemble, Joint Training, AGM, MLB, and MCR. Each column of the confusion matrix represents cases where both unimodal predictions are incorrect, where only one is correct, and where both are correct. The results highlight that MCR consistently performs well in cases where at least one unimodal prediction is correct, with ResNet-based models surpassing even the Ensemble in this category. Additionally, MLB demonstrates stronger performance in discovering emergent information, highlighting a current limitation of MCR in this aspect. . . . . 118

7.1 The figure illustrates performance trends for unimodal models, multimodal models without balancing methods, multimodal models using the current best method (MCR), and the theoretical best achievable performance. The results are shown on the CREMA-D dataset for two model scales, ResNet and Conformer. With dashed lines, we make four key observations: 1) Unimodal Baseline: Unimodal models often outperform multimodal baselines without balancing, questioning the value of extra modalities. 2) Current Balancing Impact: MCR improves multimodal performance, narrowing the gap to their potential. 3) Theoretical Gap: Significant room remains to reach the theoretical best multimodal performance. 4) Optimized Training: Better training objectives may achieve even larger gains in scaling models and data. . . . . 134

# List of Tables

2.1 The table compares the unimodal and multimodal performance of the Core, Early, and Mid-Late architectures on the SHHS-1 dataset. Systematically, AL and MS losses are added until CoreSleep achieves the best overall results. All models are trained on EEG & EOG and tested on EEG (middle), EOG (right) or both (left). We perform training three times with different dataset splits to report variance. Results from the literature on this dataset are added below the dashed line. Metrics used: Accuracy, Cohen's Kappa, Macro-F1. With bold are the best-performing models in each sub-block, and with blue bold, the best-performing model compared to all the rest. In the first column, we observe the comparison of multimodal performance on each model architecture and loss options. In contrast, in the second and third columns, we observe how they perform when one modality is missing. . . . . 41

2.2 Comparison of model performance (Accuracy, Cohen's $\kappa$ and Macro-F1) on subset of Noisy SHHS-1 Patients . . . . . 45

4.1 Binary classification accuracy of unimodal models and different multimodal balancing techniques on CMU-MOSEI dataset. . . 84

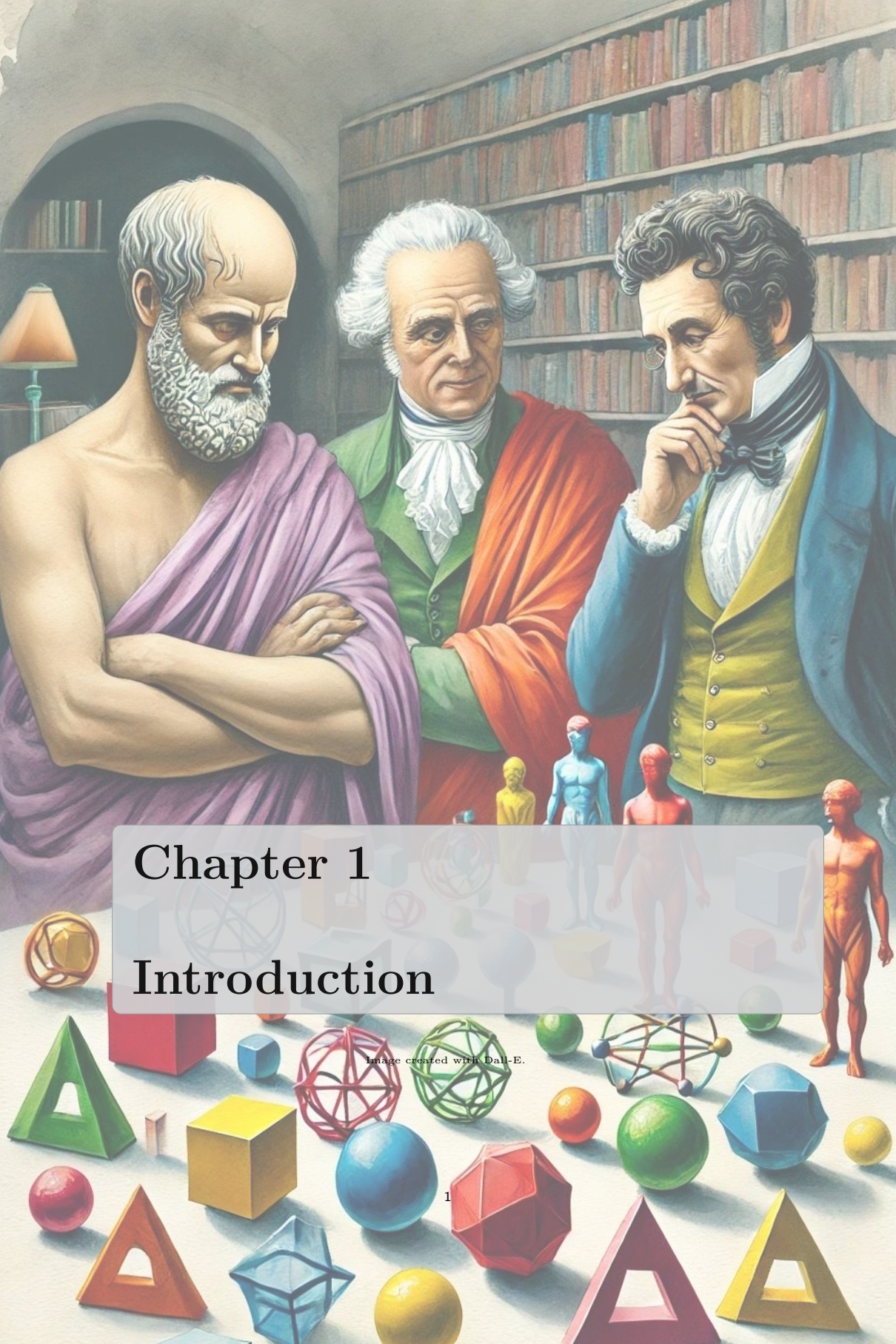
6.2 Ablation Study: Game Strategies – Comparison of model accuracy for different game strategies: Collaborative, Independent, and Greedy, using both ResNet and Conformer backbones. . . 116

6.1 This table compares the proposed MCR method with several baselines across the additional datasets: CMU-MOSEI, CMU-MOSI, and Something-Something. Results are shown for different modality combinations (e.g., V-A-T: Video-Audio-Text), using Transformer and Swin-TF backbones. MCR consistently outperforms other methods across all configurations. . . . . 116

6.3 Ablation Study: Regularization Components – Accuracy (%) of ResNet and Conformer models across datasets with different combinations of MCR components: $L_{MIPD}$, $L_{Con}$, and $L_{CEB}$. Results indicate that combining $L_{MIPD}$ and $L_{Con}$ is crucial for improvement, while $L_{CEB}$ does not benefit all models. . . . . 117

A.1 Hyperparameters for each dataset (CREMA-D, AVE, and UCF101), backbone encoder (ResNet and Conformer where applicable), balancing method (MSLR, MMCosine, OGM, PMR, AGM, MLB), and fusion method (Late-Linear, MLP, FiLM, Gated, and TF where applicable). . . . . 139
A.2 Accuracy presented in table format for each dataset (CREMA-D, AVE, and UCF101) and backbone encoder (ResNet and Conformer where applicable), balancing method (MMCosine, MSLR, OGM, PMR, AGM, MLB) using the Late-Linear fusion method. . . . . 140
A.3 Accuracy presented in table format for the different fusion methods (MLP, FiLM, Gated, TF) on the CREMA-D dataset. 140

B.1 Ablation study comparing different perturbation methods for MCR across multiple datasets. The table shows the performance of MCR when combined with various perturbation techniques, including noise in the input and latent space, Shapley values in the input space, and within-batch permutations. . . . . 146



# Chapter 1

# Introduction

Image created with Dall-E.

## 1.1 Philosophical Note

In the realm of artificial intelligence and deep learning, the ability to perceive and interpret complex data is essential. This capacity mirrors fundamental questions that have been explored for centuries: How do we integrate sensory information from multiple sources to form a cohesive understanding of the world? As technological advancements bring us closer to simulating human perception, exploring how various modalities, such as text, images, signals and sound, form cognition and interact becomes increasingly relevant. This thesis aims to provide a structured approach to identifying and enhancing the most critical aspects of multimodal deep learning, exploring how diverse sources of information can be effectively integrated to improve machine understanding and decision-making.

To appreciate the significance of multimodal learning, this PhD thesis begins by exploring its philosophical roots, highlighting how the integration of multiple forms of information has been fundamental to our understanding of perception, knowledge, and even reality itself.

The question of how we perceive and interpret the world has long been a central theme in philosophy. Philosophers have wondered how information from our various senses combines to create a unified experience, a process we often take for granted in everyday life. From Aristotle, with his notion of "common sense", to Merleau-Ponty, with his emphasis on "embodied perception", philosophers have considered how humans naturally combine information from multiple senses to form an understanding of the world. By drawing from these perspectives, we see that such an ability to integrate diverse inputs is fundamental to human cognition. Multimodal learning, in turn, seeks to emulate this natural capacity by developing artificial systems capable of combining information from varied sensory sources, simulating in part the remarkable human ability to derive unified meaning from complex data.

One of the earliest systematic explorations of how we integrate different forms of information comes from Aristotle. In his work *Περὶ Ψυχῆς (On the Soul)*, written around 350 BC, Aristotle provides a foundational framework for understanding how the senses contribute to a unified perception of reality. He identified five human senses: sight, hearing, smell, taste, and touch, each offering distinct yet complementary information about the external world. To explain how these diverse sensory inputs combine into a single cohesive experience, Aristotle introduced the concept of common sense (κοινὴ αἴσθησις), a central cognitive faculty responsible for integrating the outputs of individual senses.

This idea of a common integrating faculty laid the groundwork for understanding

perception as a unified process, where various sensory inputs do not exist in isolation but are synthesized cohesively. In modern multimodal learning, we see echoes of this Aristotelian vision: by combining visual, auditory, and other types of data, multimodal systems attempt to replicate this fusion of information to capture more complex and contextual relationships using all available sources.

Immanuel Kant, in 1781, further developed this concept of sensory integration in his book *Kritik der reinen Vernunft (Critique of Pure Reason)*. He argued that sensory experience alone is insufficient for understanding the world and supported that is shaped by the mind's inherent frameworks, where sensory information is actively organized using apriori categories like space, time, and causality. While prior knowledge is rarely directly incorporated into multimodal networks, it is implicitly present within the data, manifesting through the inherent biases and unique characteristics of each modality. For example, such structures are sometimes inherited in the data structures, such as in times series, which are represented in the form of sequential values. Kant's insights remind us that perception involves the merging of sensory information and the application of cognitive structures that make the integration meaningful, and there is a need to consider and discover such structures for better understanding.

In the 20th century, Maurice Merleau-Ponty, in his seminal work *Phenomenology of Perception* (1945), offered a phenomenological view on sensory integration, asserting that perception is fundamentally anchored in the body's active engagement with the world. While some view it as contrasting, we align with those who interpret it as an extension of Kant's arguments. Merleau-Ponty argues that our senses do not operate just based on some prior knowledge; rather, they work together and are grounded in bodily experience and interaction with our environment. He describes perception as inherently multisensory, where seeing, hearing, and touching are interconnected modes of engaging with the environment. Rather than being solely filtered through mental categories, sensory integration arises from the lived, relational context in which we perceive. Perception translates then to a process that is dynamic, contextual, and shaped by the particularities of each moment.

Merleau-Ponty's perspective on perception brings attention to a frequently neglected aspect of multimodal machine learning. Even if we incorporate Kant's a priori framework into our models and successfully create a unified Aristotelian "common sense", our models remain shaped by the trajectory of their learning experiences. In deep learning, this trajectory is influenced by choices that introduce bias into the process, such as the training methods. This thesis thoughtfully examines how the choice of training objectives impacts deep multimodal networks. Incorporating multiple modalities into a single model requires careful design of the model architecture and its priors, along with a tailored joint learning framework that aligns the learning experience and the

model's interaction with the world, with the underlying goals of multimodal learning, a focus that has often been overlooked.

## 1.2 Multimodal Machine Learning Landscape

Having explored the foundations of human perception, we now define multimodal learning within the context of machine learning. Machine learning traditionally involves optimizing a parameterized function $f$ to map inputs $X$ to targets $Y$, minimizing a loss function $\mathcal{L}(f(X), Y)$. Multimodal learning extends this by integrating inputs from multiple modalities $X = \{X_1, X_2, \ldots, X_M\}$, each offering unique and complementary perspectives. The primary goal is effectively integrating and utilising information from these additional sources to improve predictions, leveraging their combined strengths. This integration enables models to broaden task applicability and capitalize on cross-modal synergies. This section first examines what constitutes multimodal data, and then it categorizes multimodal tasks into broad groups, providing a structured overview of the field.

### 1.2.1 Multimodal Data

Multimodal data can be described as data from diverse sources characterized by both their heterogeneity and interconnectivity. To grasp these concepts fully, we begin with an intuitive understanding of the structural and informativeness heterogeneity between modalities and then proceed to their semantic interconnectivity.

#### Structural and Informational Heterogeneity

Structural heterogeneity refers to the inherent diversity in the representational format of each modality. Each data type possesses unique structural attributes that shape how it captures and conveys information while introducing modality-specific patterns. For example, images exhibit a spatial structure as they encode details through pixels arranged in a grid, emphasizing spatial relationships and visual patterns. In contrast, temporally ordered data, such as signals, inherently prioritize time-based information, capturing variations over time that facilitate the detection of patterns in rhythm, frequency, or sequence. These modality-specific correlational structures vary significantly, providing a rationale for integrating them within a joint multimodal learning framework.

Following the approach of Liang et al. [104], we analyze multimodal data along a spectrum of homogeneous and heterogeneous types. Homogeneous data types share similar structural properties but differ in content; for example, physiological signals from the heart and brain have temporal structures yet convey distinct physiological information, which can be jointly analyzed to assess a patient's health. In contrast, heterogeneous data types differ substantially in structure, such as when temporal signals are combined with a CT scan, which is represented as an image. At the homogeneous end of the spectrum, modality representations are easier to handle. However, as heterogeneity increases, the integration of these diverse modalities becomes more complex.

Informational heterogeneity refers specifically to differences in content between modalities rather than their structural format. It captures variations in the informational content each modality provides, which can be further decomposed by examining characteristics such as the independent signal-to-noise ratio, the frequency of information present within a modality and its relevance to a specific phenomenon or task. The more two modalities share similar information content, the closer we approach a multi-view scenario, and they offer less of the multimodal benefits. In contrast, as informational heterogeneity increases, fully utilizing all modalities becomes more advantageous.

### Interconnectivity

Despite the structure of the data given by each source of information, multimodal data often reveal interconnected relationships in the form of shared and complementary information. These relationships can be understood from both statistical and semantic perspectives, which together offer a holistic framework for exploring intermodal connections. In this distinction of modalities, we align with the interpretation of Liang et al. [104], which provides a comprehensive approach to understanding how different data sources relate to each other.

Interconnectivity in multimodal data refers to how different modalities are related. Statistical associations arise when data from separate modalities frequently appear together, revealing patterns that support joint modelling and alignment. Statistical dependence can highlight causal or temporal relationships, indicating how one modality might influence or be influenced by another, often requiring domain expertise. For example, in the classification of epilepsy, brain signals can indicate the onset of an event several seconds before it is visible in heart signals [18]. This temporal dependency between modalities exemplifies statistical dependence, where the sequence and timing of information across modalities provide crucial insights that a single modality may overlook. Semantic correspondence involves meaningful alignments between modalities, such as

associating visual objects with descriptive text. In contrast, semantic relations encompass more complex connections, including causal or temporal links that extend beyond direct equivalences.

Knowing the properties of connectivity between modalities can be particularly advantageous for multimodal modelling. An illustrative example is that of Visual Question Answering (VQA), where a model is provided with both a question and an image, with the answer often embedded within the visual content. This dependency can be exploited to effectively model the flow of information, creating the appropriate interactions between modalities. De Vries et al. [35] have shown that performance improves when the model processes the image with contextual awareness of the question. This example demonstrates that understanding the interconnectivity of modalities, even if it's just for a specific task, allows us to model them more effectively and gain substantial benefits.

### 1.2.2 Multimodal Tasks

Multimodal learning encompasses a wide range of tasks, many of which are beyond the scope of this thesis but deserve a brief discussion to illustrate the field's potential. These tasks leverage the complementary strengths of different modalities or enhance one modality based on another, enabling models to perform complex functions that would be difficult or unachievable solely with unimodal systems. In general, we categorize multimodal tasks into two types:

**Multimodal Translation:** In these tasks, information from one modality is transformed or translated into another. Multimodal translation involves using one or more modalities as input and producing one or more modalities as output. This may include applications such as image captioning (translating visual information into text) [182, 199] or text-to-speech synthesis (converting written language into audio) [179, 189] or text-to-image generation (prompting text as the description of an image)[146, 32]. These tasks require the model to understand and map information across distinct data structures, which often introduces challenges in preserving contextual and semantic nuances.

Leveraging modalities as supervisory feedback signals have been a driving force behind the recent surge of deep learning. Traditional supervised learning is constrained by the need for costly, labour-intensive labelling. Multimodal correspondence, however, offers a way to overcome these limitations and, along with scaling through advanced models and hardware, excel in several new tasks. In this thesis, we identify key issues in multimodal learning that arise even before expanding the complexity of the task by employing a modality as the supervision module instead of traditional labels. Due to the early stage of this

line of research and its significant computational demands, we have decided to focus on multimodal learning with traditional supervised labels, which remains highly relevant today. However, we view multimodal translation tasks as a promising direction and a natural extension of this work.

**Multimodal Fusion:** In multimodal fusion tasks, the model receives multiple modalities as inputs to combine them to achieve a target task, e.g. regression or classification. By integrating diverse data sources, fusion-based approaches can enhance performance and robustness. Examples explored in this thesis include emotion recognition using video, audio, and text modalities; action recognition with video, audio, and optical flow; and sleep stage classification using electroencephalography (EEG) and electrooculography (EOG). We expect each modality to contribute complementary information, enabling the model to make more accurate and robust predictions.

Each of these two categories can still be further decomposed into subcategories to accommodate specific cases [15]; for example, fusion can include co-learning where one modality aids training but does not participate during inference. Translation can be framed as alignment, where both modalities create a shared space to map correspondence. Nevertheless, we believe these two primary categories provide a clear structure for understanding and discerning different multimodal tasks. In the remainder of this thesis, we will focus primarily on multimodal fusion, where all input modalities contribute to achieving a predefined task supported by a set of labelled data points.

Moreover, we will use the term ***multimodal learning*** to refer to models and tasks that involve multiple modalities as input (including both fusion and translation) and incorporate training procedures that enable interaction between them. This is a key subset of multimodal machine learning, which is the primary focus of this thesis.

## 1.3 Motivation for Multimodal Learning

Multimodal learning enriches machine learning models by integrating multiple data sources, each providing distinct and complementary information about a phenomenon. These additional data sources offer several advantages: they deliver independent, modality-specific insights, complement information from other modalities, strengthen model robustness against faulty data, and reduce overall system uncertainty. Together, these properties unlock tasks and methods that not only emulate human capabilities but have the potential to surpass them, pushing the boundaries of what intelligent systems can achieve.

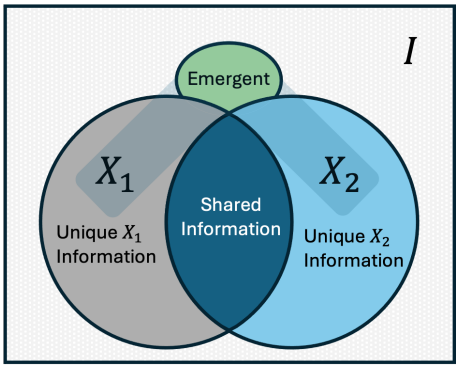



Figure 1.1: Venn diagram illustrating the decomposition of information between two modalities, $X_1$ and $X_2$, highlighting regions of unique information (specific to each modality), shared information (common across both modalities), and emergent information (arising uniquely from the interaction between $X_1$ and $X_2$, indicated with the shaded areas).

### 1.3.1 Modality-Specific Additional Information

The first direct motivation for multimodal learning lies in its ability to integrate additional, modality-specific information that each data source uniquely contributes. These are the areas of unique information for both modalities in the Venn diagram of Figure 1.1. For any given task, different modalities can capture distinct aspects of the event that may be inaccessible to others. A bird singing, for instance, can signal its presence even if it is hidden from the image view, while the sound of a church bell can convey positive or negative events or indicate the time. Such auditory cues supplement visual information, offering a richer, context-aware understanding that a single modality alone could miss. This wealth of additional information allows us to enhance performance on existing tasks and opens up possibilities for tackling new challenges that previously were beyond reach.

### 1.3.2 Robustness and Calibration via Shared Information

Another key advantage of multimodal learning lies in the shared information represented in the Venn diagram in Figure 1.1. At first glance, this shared information might seem redundant, as it is already present in each unimodal input. However, this overlap is crucial for building robust and less uncertain systems by ensuring more than once about the validity of the predictions.

An example of this that we will explore in this thesis is the scenario where sensors connected to EEG or EOG malfunction, producing noisy inputs. In such cases, each modality can help identify and filter out inconsistent or unreliable signals, as the shared information allows the system to cross-reference and validate against other modalities. This redundancy ensures the model remains accurate and stable, even when individual data sources encounter faults or interference. By aligning and reinforcing information across modalities, shared insights enhance the system's confidence and reliability, helping it to verify and correct potential errors from any single modality and ultimately leading to more consistent performance in varied and uncertain environments.

### 1.3.3 Emergent Cross-Modal Insights

In addition to modality-specific information, multimodal learning benefits from emergent insights that arise through cross-modal interactions. When modalities interact, they reveal interrelationships that contribute to the extraction of emergent information about the modalities. For example, in the VQA task, knowing the question in advance can guide the focus toward specific image details relevant to the answer. This highlights how jointly learning from multiple modalities can enable the emergence of insights. These temporal, spatial, or semantic interactions enable the model to capture nuances that would otherwise remain hidden within isolated modalities, as expressed in Figure 1.1.

Such integration can reduce uncertainty and enhance the model's predictive capability. By considering the relationships between modalities, we can capture redundant shared information and novel insights that emerge through their interaction. If the modalities have no shared or emergent information, their interaction does not provide additional predictive value. However, when such information exists, the model's understanding is enriched, leading to the discovery of novel multimodal synergies.

## 1.4 Approaches to Multimodal Learning

Motivated by the benefits that multimodal learning can contribute, we proceed by providing a timeline of advancements in this field, tracing its evolution from unimodal approaches to advanced multimodal fusion techniques. This overview will establish foundational knowledge to support the discussions in subsequent chapters of the thesis.

### 1.4.1 Ensembles

Ensemble learning has long been a foundational technique in machine learning, especially in unimodal contexts, where it combines multiple models to enhance predictive accuracy, reliability, and robustness. The core idea is that a diverse set of models can capture different local minima in the loss landscape, enabling the ensemble to achieve a more calibrated and accurate prediction than any single model could achieve on its own [37]. By leveraging the unique strengths of each model and offsetting their weaknesses, ensemble methods effectively reduce variance and bias, leading to more stable and generalizable predictions [138].

Ensemble methods like bagging, boosting, and stacking have been widely utilized to improve model performance by reducing variance and bias. Bagging, exemplified by Random Forests [22], trains multiple models on random subsets and combines their predictions while boosting, as in AdaBoost [153] and later methods like Gradient Boosting [48] and XGBoost [30], sequentially focuses on correcting errors from previous models. Stacking [195] combines predictions from various models using a meta-model to capture their strengths. In the deep learning era, ensemble techniques continue to show value, with deep ensembles providing reliable uncertainty estimates [92] and achieving success in applications like disease diagnosis [9], fraud detection [83], and image recognition [60]. The versatility of ensemble methods highlights their significant impact on machine learning tasks.

In the context of multimodal data, ensembles offer a powerful way to combine models that are each independently trained on different data modalities. By leveraging the unique strengths and specialized learning of each unimodal model, ensemble approaches capitalize on the excellence achieved within individual modalities. Such methods have proven effective across various tasks and modalities [50, 85, 88].

The separate training of unimodal models has both benefits and limitations. Unimodal training avoids many challenges of multimodal training that can disrupt learning by allowing each model to focus exclusively on its specific modality. Ensembles simplify optimization by focusing each time on single-modality objectives and then combining them without or minimal further training. In this thesis, we will avoid naming confusion, and Ensembles will consistently refer to models that have not undergone *any* further training with multimodal interactions. While Ensembles were one of the earliest methods to combine unimodal, modality-specific networks to harness the strengths of each, they have a key limitation. In most applications, we expect information from different modalities to be correlated, enabling the extraction of knowledge

that would otherwise be inaccessible. For example, the question asked in the VQA task discussed above should actively guide how the model filters and interprets the image. However, Ensembles lack any mechanism to capture these cross-modal interactions during training, resulting in suboptimal integration of complementary information. To address this, a wide range of research has focused on optimizing modality integration during training, which is frequently termed multimodal fusion.

### 1.4.2 Fusion Levels in Multimodal Networks

In the design of multimodal fusion methods, a primary consideration is determining the optimal point within the network at which information from different modalities should be integrated. This decision, often referred to as the fusion point or level, has been considered critical in how effectively the model can leverage multimodal, complementary information [145, 160]. Below, We examine the different stages of multimodal fusion:

#### Early or Data-Level Fusion

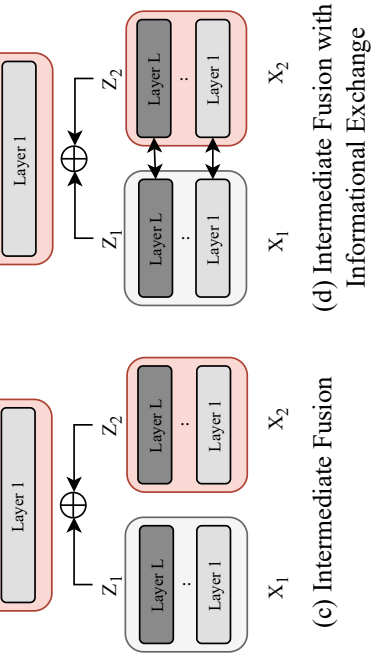
In early fusion, also known as data-level fusion, data from each modality is combined at the input stage by concatenating the raw input from multiple modalities into a unified representation. This fused input is then jointly processed by a parameterized model, making it particularly effective when strong interactions between modalities are expected for the given task [121, 163]. However, this approach poses challenges when modalities differ in structure or scale, as is often the case when combining data types like images and text, where the feature spaces may not align [15, 145]. The general form for early fusion models can be expressed as:

$$\text{Early-Fusion Model Family:} \quad f(x_1, x_2, \ldots, x_n) = g([x_1, x_2, \ldots, x_n]), \quad (1.1)$$

where $x_1, x_2, \ldots, x_n$ are the inputs from each modality, $[.,.]$ denotes the concatenation of features from each modality, and $g(\cdot)$ is a shared model that processes these fused features to produce the final output.

#### Intermediate or Mid-Level Fusion:

In intermediate fusion, also known as mid-level fusion, each modality is first processed individually through modality-specific encoders, which enables the model to extract unique features from each modality before integrating them.




(a) Early Fusion (b) Late Fusion (c) Intermediate Fusion (d) Intermediate Fusion with Informational Exchange

Figure 1.2: Illustration of multimodal fusion points: (a) Early Fusion, where modalities are integrated at the input layer; (b) Late Fusion, combining modalities at the output layer; (c) Intermediate Fusion, merging modalities at middle network layers; and (d) Intermediate Fusion with Informational Exchange, where cross-modal interactions occur at earlier layers. The symbol $\oplus$ refers to an aggregation operation, such as summation or concatenation, used to combine information from the different modalities.

This approach is the most frequently used method with choice of methods [121, 175, 72, 33, 211]. This approach is particularly advantageous as it mitigates issues related to feature dimensionality by transforming each modality into a common feature space. Additionally, it includes a network after the fusion point that processes and extracts any relevant cross-modal information, enhancing the model's ability to leverage complementary features from each modality. The general form of intermediate fusion models can be expressed as:

$$\text{Mid-Fusion Model Family:} \quad f(x_1, x_2, \ldots, x_n) = g(h_1(x_1), h_2(x_2), \ldots, h_n(x_n)) \tag{1.2}$$

where $x_1, x_2, \ldots, x_n \in X$ are the inputs from each modality. Each $h_i : X_i \to Z_i$ represents a modality-specific encoder, mapping input $x_i$ to a feature space $Z_i$. The shared function $g : Z_1 \times Z_2 \times \cdots \times Z_n \to Y$ then processes the combined features from all modalities to produce the final output.

#### Late or Decision-Level Fusion:

In late fusion, also referred to as decision-level fusion, each modality is processed independently through separate models until predictions are made. While notation for such models differs from different papers, we aim to clarify that in this work, we refer to late fusion methods as the models that fuse the predictions at the end with a simple non-parametric operation, e.g. summation. This approach is advantageous when the modalities contribute largely independent information or when each modality's prediction can stand alone. However, it may limit the potential for capturing fine-grained interactions between modalities compared to earlier fusion methods. While the structure of the model in late fusion may not differ compared to an ensemble model, it is nevertheless crucial to distinguish the two based on the way they are trained. In an ensemble model, only unimodal training is considered, while in late fusion, the model is trained under a joint multimodal objective. This difference brings interesting properties, augmented model abilities and subsequent problems. We will elaborate on them in Chapter 3.

#### Communicating Branches with Fusion Gates

Subsequent works recognized the potential benefits of enabling each modality branch to access information from the others. This cross-branch communication is often implemented through fusion gates, which facilitate selective information exchange either between modality-specific channels or during the final merging phase.

Here, we will discuss a few of these approaches. We begin with a widely adopted approach in this category, the Squeeze-and-Excite (SE) method [71, 86, 111]. SE gates reweight the feature representations of each modality by first compressing (squeezing) information from all modalities into a lower-dimensional space, reweighting patterns to emphasize shared features across modalities, and then expanding (exciting) the representations back to the original dimensional space for further processing. Other techniques aim to transfer information between branches directly, such as by element-wise summation [66], channel-wise convolutional exchanges [188], cross-attention modules [180, 98], or token exchanges in each layer [117]. Although a comprehensive examination of these alternatives is beyond our scope, we incorporate several of the most prominent approaches in our ablation studies throughout the thesis.

#### Separate Output Multimodal Models

In this category, networks produce separate outputs for each modality without merging them into a single unified representation or prediction. These models, also noted as coordinated representations [15], allow for distinct processing paths while still capturing correlations between modalities using some form of communication between the branches. This approach is particularly useful when each modality provides unique information with minimal emergent or complementary overlap, and direct fusion is not necessary. We will explore further this model architecture in Chapter 2, where we examine its properties and applications in detail.

### 1.4.3 State-of-the-Art Multimodal Transformers

In recent years, transformer-based models [180] have progressed multimodal learning by introducing a flexible architecture that reduces the need for modality-specific design interventions, enabling them to learn effectively from diverse data sources, including text [36, 70], vision [38], audio [12] and essentially any modality for which a sufficient amount of data can be collected. Transformers are highly parallelizable, allowing for efficient processing of large data volumes, and their attention mechanisms dynamically capture relationships across modalities by selectively focusing on relevant features while preserving modality-specific details. This flexibility enables transformers to handle a range of modalities, effectively managing synchronized and unaligned multimodal data, making them well-suited for complex integration tasks across various domains. We direct readers to Section 2.2.2 and to the original work by Vaswani et al. [180] for an in-depth explanation of transformer architecture and its foundational mechanisms.

Introducing transformers does not eliminate the need to carefully select the fusion point and the fusion method within the system's architecture. The network can still be designed with early, late, or intermediate fusion, with or without information exchange between unimodal branches. When information exchange is implemented, it typically relies on cross-attention mechanisms [98, 80, 97, 3] or on token exchange between modalities [117]. In networks designed with an early fusion approach, modality-specific learnable embeddings are often added to each modality's tokens [80]. More advanced methods further separate the fully-connected layers, assigning independent layers to each modality in a multi-expert structure [186]. The impact of these various techniques lies beyond the scope of this thesis; however, it remains an active area of machine learning research, with the potential to introduce valuable tools for multimodal learning. For a more comprehensive review, we refer readers to the survey by Xu et al. [200].

With the growing availability of large-scale multimodal datasets, transformers can now train on vast and varied data sources that enable them to capture complex cross-modal relationships. For example, datasets like OpenAI's CLIP dataset [143] and LAION-5B [154] provide extensive image-text pairs, facilitating models to learn visual-linguistic associations at an unprecedented scale. Similarly, YouTube-8M [1], and AudioSet [54] offer large collections of video and audio data, capturing synchronized temporal dependencies crucial for tasks involving sound and motion. In the biomedical domain, datasets such as MIMIC-CXR [79], which includes radiology images and associated reports, and SHHS [140] for sleep staging allow models to connect medical data from various modalities, supporting advancements in healthcare AI. Additionally, datasets like MSR-VTT [198], OpenKitchen [106], Something-Something [58], and HowTo100M [114] enable models to learn from instructional videos paired with descriptions, fostering improvements in video understanding and text alignment. These datasets have been pivotal in advancing state-of-the-art multimodal systems. In this thesis, some datasets, such as SHHS and Something-Something, are directly used for training, while we also leverage pretrained models whose weights were derived from training on other large-scale datasets.

The final component we aim to introduce here is the set of tasks that underpin the training of large multimodal models. As a central part of the thesis, we will address it separately in the next section.

## 1.5 Training Strategies for Multimodal Learning

While architectural differences in multimodal transformers may appear incremental, significant advancements have been driven by data availability and task selection. In this section, we will discuss the key tasks involved in training multimodal models.

Over the past decade, unimodal training has achieved remarkable success, with effective strategies ranging from supervised to self-supervised tasks. Both approaches have been widely explored for pretraining, with recent trends in self-supervised learning (SSL]) being especially promising for learning more generalizable feature representations and for leveraging large volumes of unlabeled data. Methods such as masked prediction [14, 8, 38], contrastive predictive coding (CPC) [124], and contrastive learning [67, 31, 61, 27] have been crucial for enabling models to learn meaningful representations from individual modalities without requiring labelled data. For instance, BERT's masked language modelling task [36] and wav2vec's CPC tasks [12] have enabled models to capture complex, modality-specific patterns in text and audio, respectively.

Building on the foundations of unimodal advancements, most large multimodal networks leverage a combination of SSL pretraining and a post-training stage involving multimodal fusion or translation finetuning to achieve robust performance across diverse tasks. Many models begin with unimodal SSL pretraining [98, 97, 3], while others incorporate multimodal SSL tasks designed to align representations across modalities. This alignment is often accomplished through contrastive learning objectives, where positive pairs with corresponding elements from two modalities are distinguished from non-corresponding or negative pairs. Models like CLIP [143] and ALIGN [77] employ such contrastive objectives to align visual and textual representations within a shared embedding space. Additionally, masked multimodal prediction tasks [110, 167, 95], where elements in one modality are masked and predicted based on cues from another, encourage the model to capture and utilize shared information across modalities.

Following SSL, a variety of supervised multimodal tasks with labels are used to finetune the models. These tasks include labels directly derived from the modalities, such as text in image captioning [199, 5] or audio in speech synthesis [178], as well as discrete values for classification [7], regression [69]. Object detection tasks [149, 148] to further enhance accuracy. Additionally, current state-of-the-art large models like LLaMA 3 [174, 42, 41] and Gemini [53] integrate Reinforcement Learning from Human Feedback (RLHF) [165, 125] to refine outputs by aligning them more closely with human preferences, where human feedback acts as the target objective. This combination of SSL-supervised fine-tuning and RLHF constitutes the foundational steps in training today's

large multimodal models.

## 1.6 Open Research Questions to Address

Despite significant advancements in multimodal learning, several key challenges persist, particularly in the areas of representation, fusion, and joint multimodal training. Even the most advanced multimodal models often lack a thorough examination of the impact of fusion method choices, leaving the effects of these decisions on performance and robustness largely unexplored. Additionally, these models frequently overlook the training imbalances between the modalities introduced by the various tasks, which can lead to an uneven emphasis on certain modalities, ultimately limiting the model's potential to fully leverage its multimodal inputs.

### 1.6.1 Multimodal Representation and Fusion

Effectively integrating diverse data modalities into a unified representation is a fundamental challenge in multimodal learning. This integration must balance modality-specific features with cross-modal relationships, preserving each modality's unique contributions while enabling supplementing interaction. Advances in unimodal models, cross-modal modules, and fusion modules have contributed significantly to this goal. Yet, selecting the optimal fusion strategy remains a complex decision, and its impact on model performance and robustness is still not fully understood.

Early fusion combines modalities at the input level, capturing low-level interactions but potentially diluting the unique, modality-specific information. In contrast, late fusion preserves modality-specific processing by merging information at the decision level yet risks missing important cross-modal relationships. Intermediate fusion is widely favoured, often assumed to balance modality-independent and correlated information. However, this assumption does not always hold in practice, and the balance can be difficult to achieve consistently across all modalities. Enabling models to adaptively select the optimal fusion strategy for a given task and dataset is essential for advancing the field and fully realizing the potential of multimodal learning.

### 1.6.2 Multimodal Competition

It is often expected that adding a new modality as a source of additional information should enhance a model's performance. However, in practice, this process presents significant challenges. Studies have shown that in some cases, multimodal models can perform worse than their strongest unimodal counterpart [187]. This phenomenon is attributed to multimodal competition, where one modality dominates the learning process, diminishing the contributions of other modalities. While the underlying causes of multimodal competition are not yet fully understood, it is often attributed to disparities in data quality, noise levels, feature dimensionality, and the complexity of each modality's relationship with the target, which can lead one modality to fit the target faster than others, ultimately causing the model to prioritize certain modalities over others due to these inherent biases.

Even in large multimodal models, where performance declines are less immediately apparent, this imbalance of the modalities has often been manifested as multimodal hallucinations, where models rely disproportionately on one modality, usually the text modality in vision-language models, and generate responses ungrounded in the actual visual input [190, 44, 63]. Such imbalance is part of most of the tasks that large multimodal models are trained on, ranging from simple supervised tasks to harder generative ones.

While most methods for training state-of-the-art models overlook this phenomenon, some prior works have been specifically designed to address training imbalances and enhance multimodal learning. In Chapter 3, we provide a detailed introduction to the majority of these approaches, along with their underlying assumptions about the problem. These methods predominantly include those that apply additional local unimodal losses to guide each modality individually, adjust the learning rate of each modality dynamically, or directly measure and optimize the contribution of each modality within the loss function. In this thesis, we highlight the limitations of these existing approaches and propose novel improvements that aim to more effectively mitigate multimodal competition and enhance model performance.

## 1.7 Research Objectives and Overview

### 1.7.1 Research Objectives

The general objective of this PhD thesis is to investigate specific critical aspects of multimodal machine learning models that enable and enhance effective

multimodal fusion, identify key bottlenecks, and propose strategies to address and improve them. More specifically, our objective is to identify the factors that prevent multimodal networks from fully exploiting available modalities and to design methods that enhance their integration, their learning process and ultimately, their performance. This thesis has three sub-objectives, each targeting specific open research questions identified in Section 1.6:

1. The first objective is to understand the impact of architectural choices in multimodal models on performance and robustness to noisy and missing data. Using the task of sleep staging, we observe that different architectural choices lead to varying multimodal performance levels and significantly differ in robustness. The model's ability to focus on relevant modalities while minimizing the influence of noise and missing modalities contributes to more stable and effective predictions.

   *Addressed by:* Introducing a novel multimodal architecture for the sleep staging task that enables dynamic fusion selection, complemented by tailored training objectives designed to enhance performance and robustness in the presence of missing or noisy modalities.

2. Building on the limitations identified in the first objective, the second objective analyzes why simple joint training approaches often fall short in multimodal fusion and examines how effectively current solutions address this challenge.

   *Addressed by*: Examining the gradient updates in multimodal joint training and aggregating, analyzing, and experimenting with most of the proposed solutions.

3. The third objective is to develop methods that build on existing multimodal balancing approaches, refining training objectives to further enhance multimodal learning.

   *Addressed by:* Proposing two novel methods. The first improves upon previous multimodal balancing approaches by introducing a framework that synchronizes the training progress of modalities, thereby enhancing overall performance and synergy. The second method employs information-theoretic techniques to measure unimodal contributions within a multimodal model, proposing that modalities engage in non-cooperative competition. This approach models regularization through game theory, improving the retention of unique unimodal information.

These objectives contribute to identifying the problem of training multimodal models and suggesting innovative solutions that take us one step closer to creating systems that not only exploit their unimodal understanding but also build upon it to reveal their multimodal supremacy.

### 1.7.2 Chapter Overview

This thesis is divided into two main parts. Following the introduction, Part 1 explores the multimodal task of sleep staging as an initial investigation into the development of multimodal models, highlighting key aspects of their design. Part 2 shifts focus to a critical component of multimodal models: training objectives. This part presents two comprehensive solutions and provides an in-depth reflection on how effectively these solutions and prior work address the challenges identified in multimodal learning.

Chapter 1 introduces the foundational concepts of multimodal learning, providing a concise overview of its evolution and current best practices in multimodal modelling. The chapter concludes by discussing the open research questions we address and our specific research objectives.

#### Part 1: Designing Multimodal Models

Chapter 2 explores the development of multimodal models, examining architectural choices, their impact, and training objectives, exploiting the task of sleep staging as a case study. The coordinated representation fusion scheme with separate outputs is proposed as a dynamic fusion module designed

to combine the benefits of early fusion, such as strong multimodal performance and cross-modal exploration, with the robustness advantages of late fusion, particularly in handling faulty data.

### Part 2: Multimodal Competition

Chapter 3 begins with a gradient analysis of multimodal models trained under a joint objective, highlighting that any multimodal network trained in this manner may exhibit modality imbalances. This is followed by a comprehensive analysis of previous works addressing multimodal competition.

Chapter 4 introduces a new approach to improve multimodal training by addressing modality imbalances through Multi-Loss Balancing (MLB) gradient modulation. While prior methods either lack accurate unimodal performance estimation or struggle to maintain balanced training, MLB combines these strategies to adaptively adjust learning rates across modalities. Applied across several multimodal datasets on action recognition, emotion recognition, sentiment analysis, and models, MLB consistently enhances multimodal performance by effectively mitigating multimodal competition.

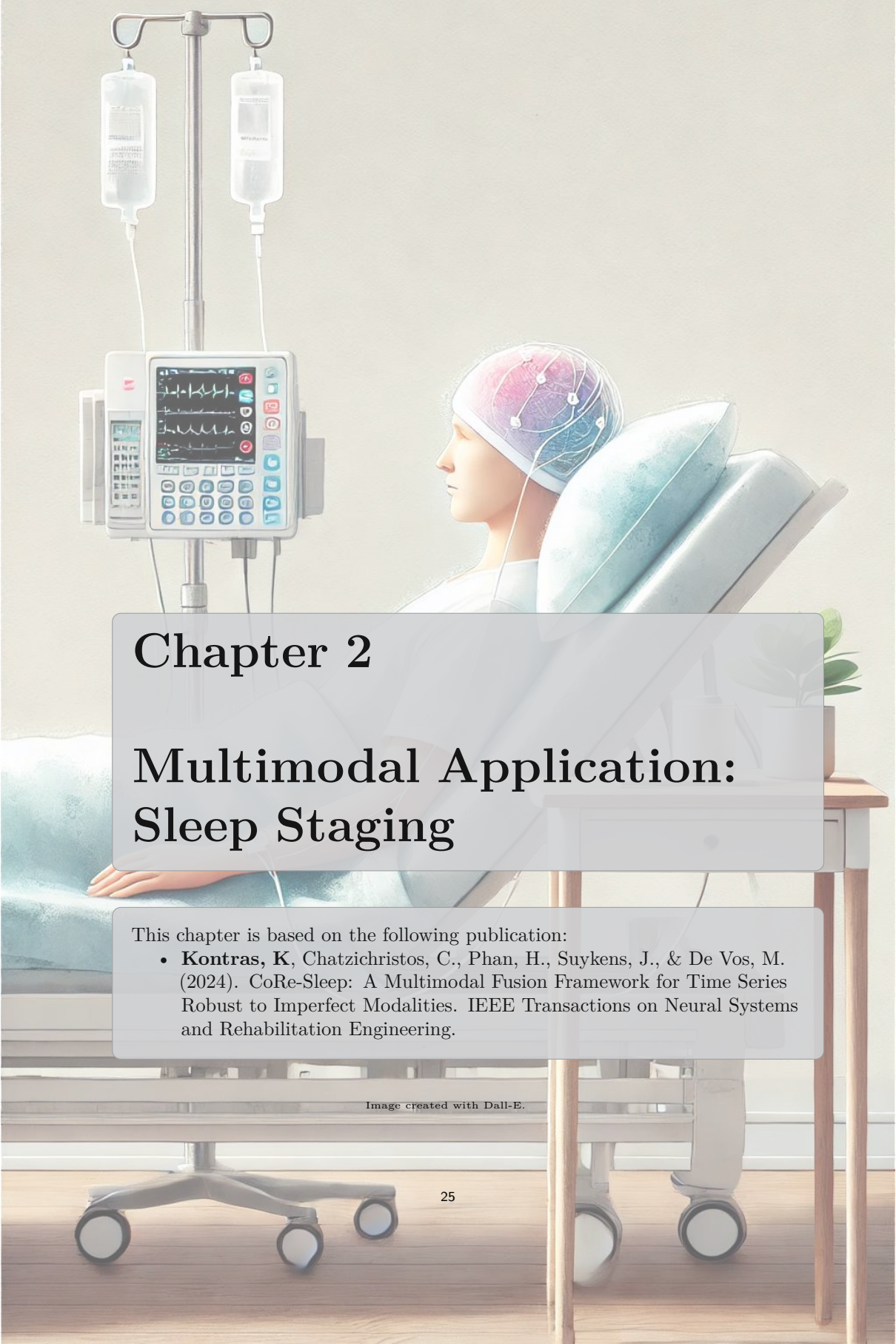
Chapter 5 assesses whether training multimodal models from scratch truly yields performance benefits over leveraging pretrained unimodal models. We examine state-of-the-art training methodologies to determine whether they have surpassed unimodal training and what specific advantages they offer. Our findings indicate that while current techniques succeed in improving the uncovering of unique cross-modal insights, they still face challenges in preserving and refining modality-specific knowledge.

Chapter 6 introduces an innovative approach to addressing the challenge of multimodal competition while preserving modality-specific contributions in multimodal training, where previous methods have often struggled. By framing multimodal competition as a game in which each modality strategically maximizes its influence on the final output, this approach leverages information theory and explainability techniques to optimize unique and shared information. The results demonstrate significant improvements in model performance across diverse datasets, modalities, and architectures, highlighting this method's potential to set new benchmarks in balanced multimodal learning.

Chapter 7 concludes the thesis by summarizing key advancements in multimodal learning and competition, discussing limitations, and outlining future work. Ethical and societal considerations are briefly addressed, followed by a reflection on the importance of enhancing multimodal learning techniques based on the insights gained.

# Part I

# Designing Multimodal Models



# Chapter 2

# Multimodal Application: Sleep Staging

This chapter is based on the following publication:

- **Kontras, K**, Chatzichristos, C., Phan, H., Suykens, J., & De Vos, M. (2024). CoRe-Sleep: A Multimodal Fusion Framework for Time Series Robust to Imperfect Modalities. IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Image created with Dall-E.

## Chapters Contribution

This chapter focuses on the development of the CoRe-Sleep architecture, a novel coordinated representation-based multimodal fusion model designed for the task of sleep staging. By integrating electroencephalography (EEG) and electrooculography (EOG) signals, CoRE-Sleep demonstrates enhanced performance and robustness under conditions of noisy or missing modalities. The chapter also provides a comprehensive comparison of early, mid, and late fusion approaches, highlighting the advantages of coordinated representations in leveraging cross-modal interactions. This work establishes the training objectives and a model framework for multimodal learning in clinical applications.

## 2.1 Introduction

In this chapter, we explore different multimodal model architectures for the task of sleep staging, focusing on revealing the properties of each, including performance and robustness. We enhance mainly two key aspects: 1) the architecture of the model and 2) the training objectives. This work is fundamental for identifying potential barriers to realizing the full benefits of multimodal learning. As the first contribution of this thesis, the current chapter highlights common steps followed for developing multimodal models. We examine fusion mechanisms and architecture design previously seen as the key components for multimodal excellence. We argue that training objectives affect multimodal performance more than architecture and fusion mechanisms. Nonetheless, fusion strategies and architecture design can still provide advantages that we will discuss in this chapter. Before the multimodal architectures, we will first look at the problem of sleep staging.

Sleep abnormalities can have severe health consequences. Automated sleep staging, i.e. labelling the sequence of sleep stages from the patient's physiological recordings, could simplify the diagnostic process. Previous work on automated sleep staging has achieved great results, mainly relying on the EEG signal. However, often, multiple sources of information are available beyond EEG. This can be particularly beneficial since other sources contain complementary information and can provide a solution when the EEG recordings are noisy or completely missing.

We propose CoRe-Sleep, a Coordinated Representation multimodal fusion network designed to improve the robustness of signal analysis in the presence of imperfect data. The model effectively handles noisy or missing segments of modalities, allowing training on incomplete data. Additionally, CoRe-Sleep demonstrates state-of-the-art performance when tested both multimodally and unimodally, utilizing a single model trained on Sleep Heart Health Study (SHHS), the largest publicly available study with sleep stage labels. Our results indicate that training on multimodal data significantly improves performance, even when evaluated on unimodal data, showcasing the flexibility and generalization capabilities of the model. This work bridges the gap between automated analysis tools and their practical clinical application.

The rest of the chapter is organized as follows: First, we introduce the physiological signals involved, the task of sleep staging, and a review of previous work on deep learning approaches for sleep staging. Then, we explain the details of the proposed method and the experimental setup, followed by a thorough presentation of the results. The chapter concludes with a discussion highlighting key insights gained from applying CoRe-Sleep and identifying the areas with

the most significant potential for future improvements in multimodal models.

### 2.1.1 Physiological Signals

The human brain comprises billions of neurons, which form networks by creating synapses among them. Each neuron releases neurotransmitters at synapses, acting as chemical signals that cause electrical changes in the receiving neuron. If this change is significant enough, it triggers an action potential or "spike." This spike is the electrical impulse neurons use to transmit information throughout the brain [81]. By measuring the collective electrical activity generated by these neuronal spikes, especially when averaged over specific brain regions, we can gain valuable insights into brain functionality, including consciousness, sensory perception, motor control or sleep stages.

There are various methods to measure brain activity, ranging from high-fidelity invasive techniques like intracranial Electroencephalography (EEG) to imaging modalities like magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI), however, the most accessible and cost-effective is the non-invasive EEG [122]. Non-invasive EEG offers a practical balance of ease of use and affordability and has a high temporal resolution, making it the most common modality choice for applications like sleep staging [74].

In addition to EEG, which records the brain's electrical activity, other physiological signals like electromyography (EMG) and electrooculography (EOG) can be used for comprehensive sleep analysis. EMG measures muscle activity by detecting electrical signals from skeletal muscles, distinguishing sleep stages through changes in muscle tone, such as the near-complete muscle atonia observed during REM sleep [159]. EOG records eye movements by measuring voltage changes resulting from the corneo-retinal potential difference as the eyes move, which is vital for identifying REM sleep characterized by rapid eye movements and for detecting sleep onset marked by slow rolling eye movements [17]. Together, EEG, EMG, and EOG form the core components of polysomnography (PSG), the gold standard for sleep staging, providing complementary data that enhance the accuracy of sleep stage classification [74].

### 2.1.2 The task of Sleep Staging

Sleep is a vital physiological process involving cyclic patterns of brain activity and bodily functions. It comprises two primary categories: Non-Rapid Eye Movement (NREM) sleep and Rapid Eye Movement (REM) sleep. Under the

American Academy of Sleep Medicine (AASM) guidelines, NREM sleep includes three stages (N1, N2, and N3), representing a progression from light to deep sleep. In contrast, REM sleep is associated with vivid dreaming and brain activity similar to wakefulness [74]. Understanding and accurately identifying these sleep phases is the task of sleep staging. This task is crucial for assessing sleep quality and diagnosing sleep disorders and various brain and mental health diseases [23]. Accurate sleep staging enables the identification of conditions such as insomnia, sleep apnea, Alzheimer's disease, and narcolepsy, facilitating appropriate interventions and enhancing our understanding of sleep's impact on overall health [28] [128].

A typical sleep staging process begins when a patient is invited to a sleep clinic to undergo an overnight polysomnography (PSG) assessment. The patient is asked to sleep in a controlled environment while multiple physiological signals are continuously recorded for 6 to 8 hours. The collected data is divided into 30 seconds, called epochs, and a trained sleep technician or clinician manually reviews each epoch. Using the AASM guidelines [74], the technician assigns each epoch to one of five categories: NREM stage 1 (N1), NREM stage 2 (N2), NREM stage 3 (N3), REM sleep, or wakefulness (W).

The manual review process involves analyzing the EEG to detect characteristic brainwave patterns, with slow-wave activity indicating deep sleep (N3) and higher-frequency activity observed during wakefulness and REM sleep, while consulting additional modalities when EEG is insufficient. Specifically, EOG provides information on eye movements, distinguishing between NREM and REM stages. Although manual scoring remains the clinical gold standard, it is labour-intensive and prone to variability across scorers, even when adhering to standardized protocols [151]. This variability and the time-consuming nature of the task underscore the need for automated sleep staging methods, which offer the potential for more consistent and efficient analysis.

After completing sleep staging annotation, the results are typically visualized using a hypnogram, a graphical representation of the sleep stages across the duration of the sleep study. A hypnogram plots time on the horizontal axis and sleep stages on the vertical axis, offering a clear overview of the patient's sleep pattern throughout the night. The graph shows how the patient cycles through different sleep stages, typically moving from lighter stages (N1 and N2) into deep sleep (N3) and then into REM sleep, repeating this cycle multiple times throughout the night. Each complete cycle generally lasts 90 to 120 minutes, with REM stages becoming longer as the night progresses [28]. Figure 2.1 depicts a typical healthy hypnogram.

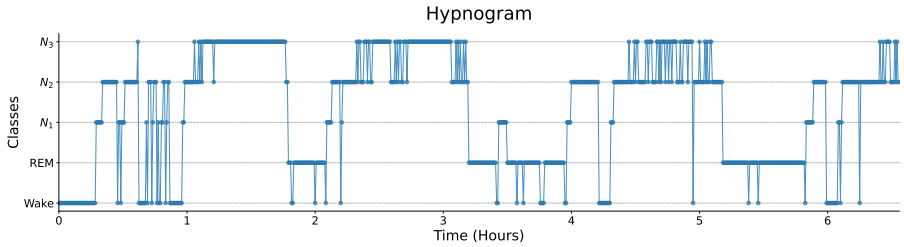



Figure 2.1: A typical hypnogram over a full night of sleep (6-8 hours). The hypnogram shows the cyclical transition through NREM stages ($N_1$, $N_2$, $N_3$) and REM sleep, along with periods of wakefulness. Sleep cycles repeat every 90 minutes, with REM sleep increasing in duration as the night progresses and deep sleep ($N_3$) decreasing.

### 2.1.3 Automating Sleep Staging with Deep Learning

Automating sleep staging is critical for reducing the labour and time required in manual scoring and minimizing the inconsistencies caused by inter- and intra-scorer variability [151]. This automation can streamline the analysis of large-scale sleep studies and improve diagnostic precision in clinical settings. Deep learning has emerged as the most promising approach to achieve this, as it eliminates the need for manual feature extraction by learning complex patterns directly from raw physiological signals.

Previous research on automating sleep staging has primarily focused on using a single modality, EEG. Initial approaches used handcrafted features [4, 100, 13, 47], based on prior knowledge about the characteristics of EEG. Recent methods employ Neural Networks (NNs) to automatically extract these features, for example with Convolutional NNs (CNN) [176, 169, 168], Recurrent NNs (RNN) [132, 133], Transformers (TF) [136, 168, 93, 139, 40] and more [131, 78].

An important factor in automated sleep staging has been the context window size, which refers to the sequence of consecutive 30-second epochs considered together during classification. Early works treated each epoch independently, assuming it was i.i.d., without considering temporal dependencies. However, it is now well established that expanding the context window significantly improves model performance by capturing the temporal structure of sleep stages. Figure 2.2 depicts such a model structure. At the time of our research, the top-performing models, XSleepNet [133], and SleepTransformer [136], utilized a context window of approximately 10 minutes. Recent advancements, such as L-SeqSleepNet [135], have explored even larger context windows, extending closer to the duration of a full sleep cycle (90-120 minutes), demonstrating

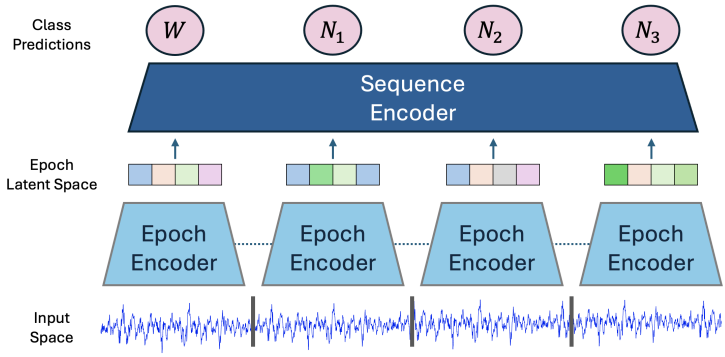



Figure 2.2: Schematic of a sequence-to-sequence model for sleep staging. The model processes raw input data (e.g., EEG) through multiple Epoch Encoders, which encode individual 30-second epochs into latent space representations. These latent representations are fed into a Sequence Encoder to capture temporal dependencies across multiple epochs. The model outputs sleep stage predictions such as wakefulness (W) and NREM stages (N1, N2, N3). The dashed lines indicate that the weights are shared among the Epoch Encoders.

further improvements in accuracy by capturing long-term dependencies across entire sleep cycles.

### 2.1.4 Multimodal Sleep Staging

Incorporating multiple physiological signals in sleep staging enhances classification accuracy by capturing complementary information that EEG alone may not provide. For instance, EOG adds critical insight into REM sleep detection through eye movements, improving accuracy, especially in REM stages. Multimodal setups also offer robustness in real-world scenarios, such as sensor detachment or noisy data, where missing or degraded signals can hinder unimodal approaches. By integrating multiple modalities, models can compensate for missing data from one source or refine decisions when signals are noisy.

#### Robustness

Missing and noisy data have been handled with imputation and denoising strategies. Imputing missing values has been widely explored in statistics either for a few missing values [24, 152] or for higher dimensional data [193, 164, 20,

94, 34, 26, 203]. Recently, imputation has been achieved by masking the missing data with a learnable embedding [126, 157], allowing the network to interpolate the values based on the remaining input. In the case of noisy samples, data are retained, but there might be ambiguity about the amount of helpful information they contain. Previous approaches facing noisy EEG include either an explicit [123, 19, 75] EEG-tailored denoising pre-phase or a more generic implicit one [16]. We will demonstrate that multimodality can face both missing and noisy modality issues elegantly.

#### Multimodal Fusion

Multimodal models can fuse the multiple input data in various ways. We have already covered most of the necessary background in the introduction, Chapter 1. Early fusion [10] refers to fusing data at the input level by projecting them onto a new dimensional space. This has been the most common approach in previous sleep staging research. Such casting on a new space can be challenging, especially if modalities differ significantly in dimensionality and sampling rate [145]. Late fusion, which combines modalities at the decision level, has been a widely adopted design approach [21] due to its simplicity. However, this comes at the cost of limiting multimodal collaboration. Finally, intermediate solutions include fusing the modalities before the classifier, which we will refer to in this chapter as Mid-Late fusion [29]. Despite being simplistic, this distinction based on the point of multimodal fusion seems to apply to the vast majority of multimodal architectures.

### 2.1.5 Our Contribution

We propose CoRe-Sleep, a multi-modal model for sleep staging. The model is based on a TF backbone encoder [180, 36], the coordinated representations multimodal fusion and is trained with a multi-task objective. CoRe-Sleep's TF encoder uses an 8-layer inner-outer block structure similar to SleepTransformer. The inner TF block processes the features extracted from Short-Time Fourier Transform (STFT) within a 30-second window. In contrast, the outer TF block maps interactions between the aggregated features of several sequential windows, 21 (10.5 minutes) following [133, 136]. Four TF blocks are used, one for each of each modality's unimodal and multimodal processing. In the multimodal blocks, pairwise cross-attention is added as a means of communication between modalities, similar to [98, 99, 204], to ensure the coordination of the modalities. The model is trained using multiple supervised losses for the unimodal and multimodal predictions and a self-supervised objective for aligning the unimodal representations, following previous works in vision and text [144, 98].

The architectural choices and the training objective of CoRe-Sleep led to the four benefits emerging:

- State-of-the-art results when trained and tested with EEG and EOG modalities on the most extensive public dataset, the SHHS.
- Robustness to missing modalities during inference; Improved performance compared to solely EEG-trained models and its unimodal equivalent when tested with only one modality. Training with both modalities benefits unimodal testing.
- Robustness to highly noisy data when present in one of the modalities.
- Ability to train even with missing-modality data, i.e., allowing for patient recordings that include only one modality.

The rest of the chapter is organised as follows. In Section 2.2, we present the dataset and methods used in CoRe-Sleep and introduce the experiments. In Section 2.3, we describe and discuss the results, comparing our method with previous state-of-the-art and additional benchmark models. Lastly, we conclude the chapter in Section 2.4.

## 2.2 Data & Methods

This section presents the architecture of the multimodal fusion network CoRe-Sleep and its multi-task objective. We begin by providing a brief overview of the largest publicly available dataset that is used in all our subsequent experiments.

### 2.2.1 Data

**SHHS** [140, 208]. Initially intended to determine cardiovascular and other consequences of sleep-disordered breathing, SHHS consists of two rounds of recordings. We experiment with data from the first round, recorded between 1995-1998 (SHHS-1). It contains 5.791 subjects aged from 39 to 90 and includes several modalities. We will use the EEG (C4-A1) sampled at 125Hz and EOG (L-R) sampled at 50Hz. Scoring was completed using the Rechtschaffen and Kales (R&K) guideline [147]. Following previous work, we merged stages N3 and N4 into the N3 stage, while Movement and Unscored labels were discarded.

A series of steps have been followed to preprocess the data. Initially, we load the signals and labels. Following previous works [133], we discard patients'

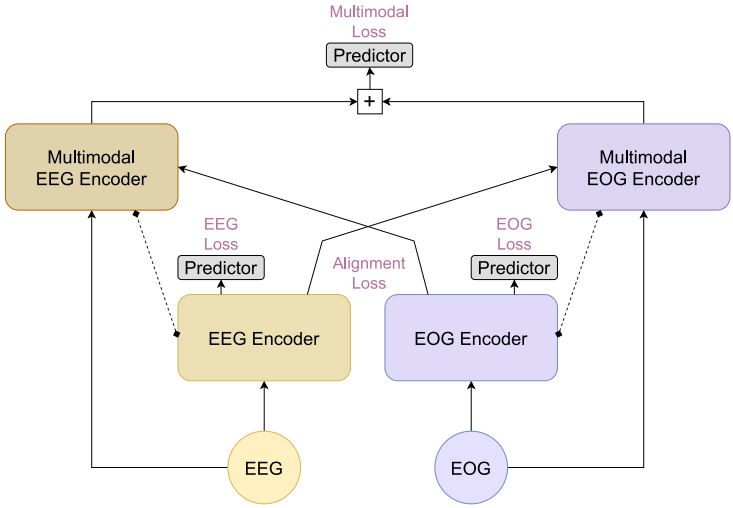



Figure 2.3: In this figure the general layout of the CoRe-Sleep [90] architecture is shown. EEG and EOG data are processed by unimodal encoders to create unimodal representations and then fed to multimodal encoders for cross-attention. Supervised losses are calculated for each representation.

recordings that do not contain windows with labels from each of the five classes. We also discard segments from the beginning and the end of the recording if the number of wake stage windows is greater than the windows of another stage. More specifically, we discard $\frac{\#wake - \#2^{nd} dominant}{2}$ from each side to balance the wake class, following previous works [136]. EEG and EOG are resampled at 100Hz and filtered with a bandpass filter on [0.3, 40] Hz and [0.3, 23] Hz. The signals are transformed with STFT using 2-second time bins with 1-second overlap and 256 points hamming window, resulting in 128-dimension features. Then, we create the 30-second windows without overlap with their corresponding label, using the majority policy when more than one class is presented within the 30 seconds. Finally, we split the patients of the dataset using a random split of 70-30% for the train-test set and keep 100 ($\sim$0.15%) patient recordings from the training as the validation set.

### 2.2.2 Transformer Backbone

The TF backbone is the TF encoder described in [180]. The encoder is responsible for processing the input sequence. The TF encoder consists of multiple layers, each with two components. The first is the multi-head self-

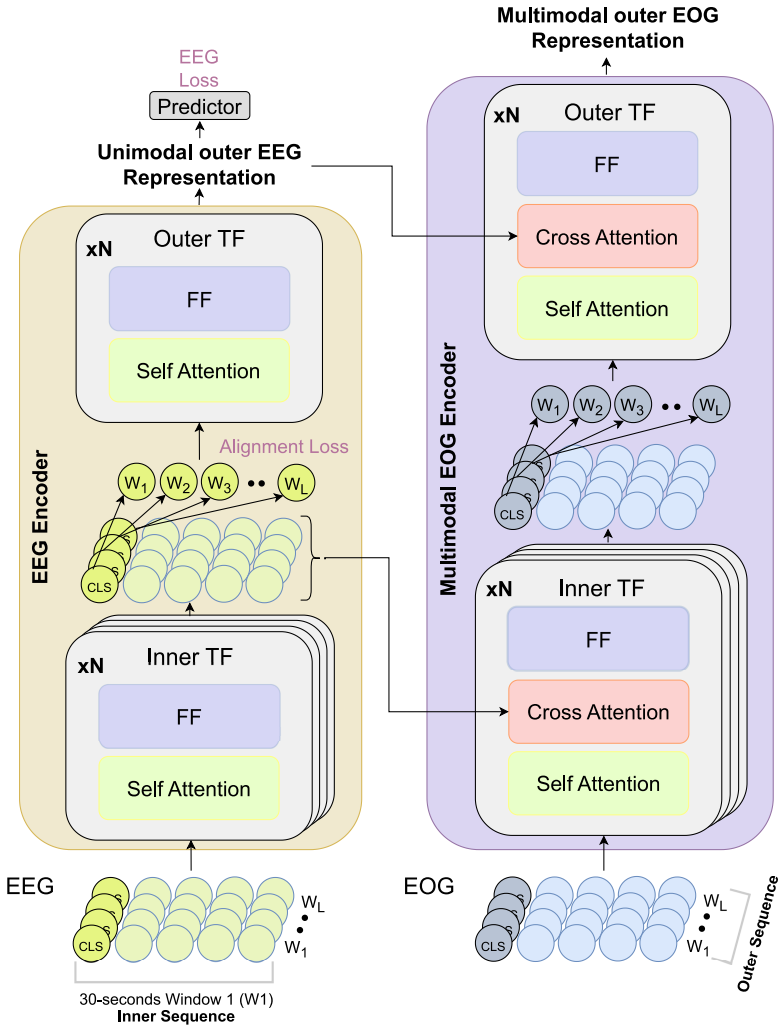



Figure 2.4: In this figure are depicted the specific components of CoRe-Sleep [90]architecture. More specifically, we depict the detailed processes of one of the unimodal and one of the multimodal encoders. Inner and outer sequences are processed by Transformer (TF) blocks, using a [CLS] embedding for summarization. Cross-attention between modalities creates coordinated representations.

attention (SA), which allows different parts of the input sequence to interact. The second is a fully connected feed-forward (FF) network, which applies a non-linear transformation to each position in the sequence. Both components have a residual connection and a (post-) layer normalization [11]. The equations describing such a TF encoder are:

$$att_h(X_m) = \text{softmax}(\frac{X_m W_h^K (X_m W_h^Q)^\intercal}{\sqrt{d_k}}) X_m W_h^V \tag{2.1}$$

$$SA(X_m) = \text{concat}[att_1(X_m), .., att_H(X_m)] W^O \tag{2.2}$$

$$Z_m = \text{layernorm}(X_m + SA(X_m)) \tag{2.3}$$

$$FF(Z_m) = max(0, Z_m W_1^F + b_1^F) W_2^F + b_2^F \tag{2.4}$$

$$output = \text{layernorm}(Z_m + FF(Z_m)) \tag{2.5}$$

Input modality $X_m \in \mathbb{R}^{sequence \times features}$ is a batch containing multiple sequences of features. The weights of the SA component have dimensions $W_h^K$ and $W_h^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_h^V \in \mathbb{R}^{d_{model} \times d_u}$, $W^O \in \mathbb{R}^{Hd_u \times d_{model}}$ and $d_{model} = 128$, $H = 8$, $d_k = 16$, $d_u = 128$. Briefly, we explain the role of each component in TF. The SA allows for measuring pairwise similarity between all the feature vectors in the sequence, re-weighting them to favour those with greater similarity. The utilization of multi-head facilitates the extraction of numerous interactions among different parts of the sequence. For the FF a two-layer multi-layer perception (MLP) with the Rectified Linear Unit (ReLU) [118] as activation function has been used, with weights $W_1^F \in \mathbb{R}^{d_{model}}$, $W_2^F \in \mathbb{R}^{d_{ff} \times d_{model}}$, biases $b_1^F \in \mathbb{R}^{d_{ff}}$, $b_2^F \in \mathbb{R}^{d_{model}}$ and dimensionality $d_{ff} = 1024$. The point-wise FF is responsible for processing each feature vector of the sequence independently of the others.

### 2.2.3 The CoRe-Sleep Architecture

The CoRe-Sleep architecture utilizes the TF backbone and the multimodal fusion of coordinated representations. Initially, we describe the encoder for a single modality, and subsequently, we proceed with the multimodal part of the network.

The architecture exploits the inner-outer framework, similar to [133, 169, 136]. Each modality has its unimodal encoder, which is an 8-layer TF encoder. The first 4-layers process the inner sequence, finding the interactions between the

features extracted from STFT within the 30-second windows. Afterwards, the inner sequence features are aggregated. The last 4-layers of the TF encoder process the outer sequence, i.e., the interactions between these aggregated features of neighbouring windows. We allow for a maximum of 21 sequential windows, following [133, 136], but recent works indicate that even greater numbers could be beneficial [135]. The Inner-Outer scheme is visualised in Figure 2.4.

CoRe-Sleep's unimodal encoder is similar to SleepTransformer [136] with two modifications:

1. It uses a learnable embedding, for the class [CLS], for aggregating inner-sequence features instead of a separate attention module. This allows more interaction via multi-head attention and multi-layer aggregation.

2. It uses learnable relative positional embeddings [156] instead of absolute sinusoidal ones. These allow the model to learn dependencies between positions in the feature sequence and are more efficient because they are incorporated into the attention mechanism.

These changes result in a small improvement. The unimodal encoder's output is utilized by a predictor, and together, they form the unimodal equivalent network, which is discussed in later sections.

To achieve multimodal fusion, we exploit coordinated representations. Coordinated representation is a mid-late fusion scheme, where each modality has a separate network stream with established communication between them. We first calculate the inner and outer sequence feature representations of each modality with the modality-specific unimodal encoders as described above. We then feed each modality's inner and outer representation to the corresponding multimodal encoder branches of the other modality, see Figure 2.3. The weights of SA and FF are shared among each modality's multimodal and unimodal encoders. The multimodal encoders have the same structure as the unimodal encoders with an additional cross-attention (CA) component to let one modality affect the representation of the other, similar to [180, 98, 99, 204, 175]. The CA component allows the pairwise similarities between the sequence steps of the two modalities $X_1, X_2$ to affect the importance by re-weighting one of them. The equations describing the CA component are:

$$\text{c-att}_h(X_1|X_2) = \text{softmax}(\frac{X_1 W_h^K (X_2 W_h^Q)^\intercal}{\sqrt{d_k}}) X_1 W_h^V \tag{2.6}$$

$$\text{CA}(X_1|X_2) = \text{cat}[\text{c-att}_1(X_1|X_2), .., \text{c-att}_H(X_1|X_2)] W^O \tag{2.7}$$

The matrices $W_h^K$, $W_h^Q$, $W_h^V$, $W^O$ have the same dimensionality as in the TF encoder described in 2.2.2. The CA component is placed between each layer's SA and FF blocks. Multiple sequential CAs can be used to extend CoRe-Sleep on more than two input modalities. The output of each multimodal encoder, which is dedicated to one modality, can be considered the representation of the corresponding modality grounded on (influenced by) the remaining modalities. The final representation based on all modalities is a summation of the different grounded multimodal representations. Such summation allows all predictors to have the same number of parameters and possibly share them among the predictors of each modality.

### 2.2.4 Training Objectives

The model jointly optimizes three distinct loss functions; see Figure 2.3. The two of them are supervised Cross-Entropy (CE) losses from the predictions of the unimodal and the multimodal network's output. The third is a loss of self-supervised alignment between the two modalities.

**Alignment Loss (AL)**: It exploits the fact that EEG and EOG signals are measured at the same time, allowing to cast this as a self-supervised task of predicting which 30-second windows of one modality corresponds to the other modality. This is motivated by literature on vision-text integration [143, 98, 99, 204], where images align with captions, and the task is to match them. CoRe-Sleep compares each window's unimodal representation with those of other modalities to predict matching pairs, with each representation being dedicated to a single window. Among various possible losses like InfoNCE [124] or Triplet loss [155], we opt for the approach in [143] due to its scalability and simplicity. We frame the problem as a double classification task, where each modality is tasked with predicting its position in the other modality.

**Multi-Supervised Loss (MS)**: CoRe-Sleep's architecture creates both unimodal and multimodal representations. The MS loss is the summation of supervised losses arising from placing additional predictors to each unimodal representation. It reduces the competing effects of each modality [73] by allowing the worse-performing modality to affect proportionally the total loss. MS loss forces the network to perform well with all modalities present and with each modality separately. In CoRe-Sleep, the multiple predictions come with minimal computational overhead.

Given a training set of patients' full night recordings, the goal is to identify the sleep label $y$ for each 30-second window. The patient recordings are segmented in windows to create the dataset $X = (\{X_1^1, ..X_M^1\}, .., \{X_1^L, ..X_M^L\})$ where $M$ denotes the number of modalities, in this work $M = 2$ and $L$ the total number

of windows in the dataset. Each $X_m^l \in \mathbb{R}^{T \times D}$ and $m = 1, .., M$ contains the frequency features with resolution $D = 128$ extracted from STFT where a smaller window of 2 seconds with a 1-second overlap is applied. It results in $T = 29$ feature vectors representing a 30-second window.

Sequential windows include useful information; therefore, we preserve such information by processing multiple neighbouring windows together. Following the inner-outer framework, we first process the within-window features $\{X_1^l, .., X_M^l\} \in \mathbb{R}^{M \times T \times D}$, aggregating information on the output of the inner $\in \mathbb{R}^{M \times 1 \times D}$. Subsequently, we map the interaction between neighbouring windows on the outer sequence. To achieve that, we take sequential windows of the same recording, allowing for a maximum of 21 sequential windows following [133, 136]. Each model $f$ trained in the experiments dedicates a subpart of its total parameters $\theta$ for each of the modalities $m$, denoted as $\theta_m$. Each such subpart of the network calculates the unimodal predictions used to calculate the MS loss. In conclusion, following [143], the alignment loss is obtained by summing up the CE losses, where the predictions are calculated using the outer product of pairwise unimodal representations, and the targets are set to the identity matrix $I$. Intuitively, AL tries to predict which window of one modality belongs to the other. The target of AL loss is the outer sequence length, so the AL loss has a different magnitude than the rest of the losses, therefore we include a scaling parameter $\lambda_A = 0.1$. We observed model performance is not sensitive to small changes on $\lambda_A$; however, this still might not be the optimal value.

The final training objective $L(X, y/f_\theta)$ is a sum of the previously mentioned:

$$L(X, y|f_\theta) = \text{CE}(f_\theta(X), y) + \text{MS}(X, y|f_\theta) + \text{AL}(X|f_\theta) \tag{2.8}$$

with

$$MS(X, y|f_\theta)) = \sum_{m=1}^{M} CE(f_{\theta_m}(_i), y) \tag{2.9}$$

and

$$AL(X|f_\theta)) = \lambda_A \sum_{m=1}^{M} \sum_{\substack{j=1 \\ m \neq j}}^{M} \text{CE}(f_{\theta_m}(X_m) f_{\theta_j}(X_j)^\intercal, I) \tag{2.10}$$

Each of the aforementioned losses consists of one or multiple CE terms. CE measures the difference between the predicted probability distribution outputted by the network and the ground-truth probability distribution. When applied along softmax on the logit outputs of a neural network CE is a logistic loss and is

Bayes consistent [210], meaning that a model that minimizes CE, essentially, will minimize the zero-one classification loss. The mean of the zero-one classification loss is accuracy. Therefore, minimizing CE loss will allow the model to maximize its accuracy on the classification task. The equation of CE loss in N multiclass classification with $p_i$ the predicted probability of each class is:

$$CE(y, p) = -\sum_{i}^{N}(y_i \cdot \log(p_i)) \tag{2.11}$$

### 2.2.5 Architecture Benchmarks (Early & Mid-Late Fusion)

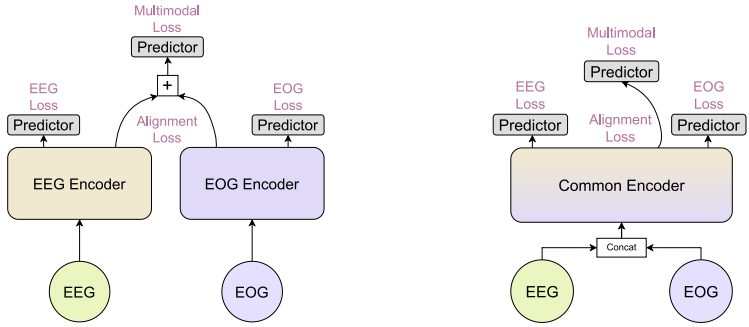
Here, we introduce two benchmark models, which, along with previous state-of-the-art literature, we compare with CoRe-Sleep. These models assist in illustrating the benefits of the multimodal fusion architecture and the training objectives. We take two of the uttermost cases, the Early and the Mid-Late fusion. As the name suggests, the former fuses multimodal input before encoding and the latter trains modality encoders separately and only fuses them before the predictor, see Figure 2.5.

**Early fusion** is a model structure with no modality-specific processing, and data are merged (by concatenation) before the model in the data level. Only preprocessing steps are distinctive between the two modalities. The same TF blocks process both modalities by considering the parts of the sequence, both inner and outer. That allows each modality to attend to each other and creates the fonts to identify better multimodal interactions. TF distinguishes the time relationships from the positional encoding. Similarly, to achieve the discrimination between the two modalities, we sum each input modality with a learnable embedding unique for every modality, similarly to [113]. We note that Early fusion requires an extra forward pass on each modality, considering only that as input each time separately to obtain the unimodal representations, hence reducing the inference speed. Those representations must be obtained for the AL and MS losses.

**Mid-Late fusion** is the architecture where each modality has its unimodal encoder, which produces the unimodal representations. By just summing them up, we get the multimodal as well. Incorporating additional classifiers on the unimodal representations enables us to calculate the MS loss. Again, we utilize the unimodal representations for the AL loss, but only those preceding the Outer TF of the unimodal encoder. The rationale behind this choice is to ensure that each representation contains information solely about the respective window without incorporating information from neighbouring ones that would make the

Table 2.1: The table compares the unimodal and multimodal performance of the Core, Early, and Mid-Late architectures on the SHHS-1 dataset. Systematically, AL and MS losses are added until Core-Sleep achieves the best overall results. All models are trained on EEG & EOG and tested on EEG (middle), EOG (right) or both (left). We perform training three times with different dataset splits to report variance. Results from the literature on this dataset are added below the dashed line. Metrics used: Accuracy, Cohen's Kappa, Macro-F1. With bold are the best-performing models in each sub-block, and with blue bold, the best-performing model compared to all the rest. In the first column, we observe the comparison of multimodal performance on each model architecture and loss options. In contrast, in the second and third columns, we observe how they perform when one modality is missing.

| Dataset: **SHHS-1** | Overall Metrics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Test Modalities | EEG-EOG | | | EEG | | | EOG | | |
| Model | Acc | $\kappa$ | MF1 | Acc | $\kappa$ | MF1 | Acc | $\kappa$ | MF1 |
| Early | 89.1 $\pm$0.0 | 0.847 $\pm$0.001 | **81.7** $\pm$0.3 | 84.5 $\pm$0.1 | 0.783 $\pm$0.001 | 75.9 $\pm$0.2 | 78.6 $\pm$0.2 | 0.696 $\pm$0.003 | 65.9 $\pm$0.4 |
| Mid-Late | 89.1 $\pm$0.1 | **0.848** $\pm$0.002 | 81.6 $\pm$0.2 | **86.3** $\pm$0.4 | **0.808** $\pm$0.006 | **78.4** $\pm$0.5 | **81.4** $\pm$0.2 | **0.737** $\pm$0.003 | **69.1** $\pm$0.2 |
| CoRe | 89.1 $\pm$0.0 | 0.847 $\pm$0.001 | 81.6 $\pm$0.3 | 84.0 $\pm$0.9 | 0.775 $\pm$0.012 | 75.0 $\pm$1.2 | 73.1 $\pm$1.2 | 0.616 $\pm$0.018 | 59.3 $\pm$1.1 |
| Early + **AL** | 89.2 $\pm$0.0 | 0.849 $\pm$0.000 | **81.9** $\pm$0.2 | 77.5 $\pm$5.4 | 0.678 $\pm$0.080 | 65.2 $\pm$8.1 | 73.1 $\pm$3.3 | 0.614 $\pm$0.049 | 58.7 $\pm$3.5 |
| Mid-Late + **AL** | 89.2 $\pm$0.1 | 0.848 $\pm$0.002 | 81.7 $\pm$0.3 | **86.5** $\pm$0.3 | **0.811** $\pm$0.004 | **78.7** $\pm$0.3 | **82.4** $\pm$0.6 | **0.751** $\pm$0.009 | **70.7** $\pm$1.1 |
| CoRe + **AL** | **89.3** $\pm$0.2 | **0.850** $\pm$0.002 | 81.8 $\pm$0.1 | 80.9 $\pm$0.4 | 0.731 $\pm$0.005 | 69.9 $\pm$0.4 | 77.5 $\pm$1.0 | 0.680 $\pm$0.016 | 63.1 $\pm$1.2 |
| Early + **MS** | **89.4** $\pm$0.1 | **0.851** $\pm$0.002 | **82.1** $\pm$0.3 | 81.7 $\pm$1.1 | 0.742 $\pm$0.015 | 65.8 $\pm$1.0 | 77.7 $\pm$1.5 | 0.686 $\pm$0.018 | 62.4 $\pm$1.2 |
| Mid-Late + **MS** | 89.2 $\pm$0.1 | 0.849 $\pm$0.002 | 81.6 $\pm$0.1 | 87.7 $\pm$0.2 | 0.828 $\pm$0.003 | 80.1 $\pm$0.2 | **84.9** $\pm$0.2 | **0.787** $\pm$0.002 | 74.4 $\pm$0.2 |
| CoRe + **MS** | **89.4** $\pm$0.1 | **0.851** $\pm$0.002 | 82.0 $\pm$0.3 | **87.9** $\pm$0.2 | **0.830** $\pm$0.003 | **80.3** $\pm$0.4 | **84.9** $\pm$0.1 | **0.787** $\pm$0.002 | **74.5** $\pm$0.2 |
| Early + **MS** + **AL** | **89.5** $\pm$0.1 | **0.853** $\pm$0.002 | **82.3** $\pm$0.2 | 87.1 $\pm$1.7 | 0.820 $\pm$0.021 | 79.7 $\pm$1.6 | 83.4 $\pm$3.0 | 0.770 $\pm$0.036 | 74.0 $\pm$2.3 |
| Mid-Late + **MS** + **AL** | 89.3 $\pm$0.1 | 0.851 $\pm$0.002 | 81.9 $\pm$0.1 | **88.0** $\pm$0.2 | 0.831 $\pm$0.003 | 80.4 $\pm$0.2 | **85.2** $\pm$0.1 | **0.792** $\pm$0.002 | 75.1 $\pm$0.1 |
| CoRe-Sleep (**MS** + **AL**) | **89.5** $\pm$0.1 | **0.853** $\pm$0.002 | **82.3** $\pm$0.3 | **88.2** $\pm$0.2 | **0.834** $\pm$0.003 | **80.8** $\pm$0.4 | **85.3** $\pm$0.1 | **0.792** $\pm$0.001 | **75.3** $\pm$0.3 |
| Unimodal Equivalent | - | - | - | 87.7 $\pm$0.2 | 0.828 $\pm$0.003 | 80.1 $\pm$0.4 | 85.1 $\pm$0.1 | 0.790 $\pm$0.001 | 75.0 $\pm$0.2 |
| XSleepNet [133] | 88.8 | 0.843 | 82.0 | 87.6 | 0.826 | 80.7 | - | - | - |
| SleePyCo [93] | - | - | - | 87.9 | 0.830 | 80.7 | - | - | - |
| SleepTransformer [136] | - | - | - | 87.7 | 0.828 | 80.1 | - | - | - |




(a) Mid-Late Fusion Architecture (b) Early Fusion Architecture

Figure 2.5: **a)** Mid-Late fusion scheme (MS + AL): Each modality has its encoder, and information gets only merged before the multimodal predictor. The Mid-Late model refers to the model trained solely on the Multimodal loss, while the models incorporating MS and AL are explicitly identified. Similar to CoRe-Sleep, this Mid-Late fusion scheme can derive both unimodal and multimodal predictions in a single forward pass. **b)** Early fusion scheme (MS + AL): In the Early model, EEG and EOG embeddings are processed by a shared encoder to calculate the multimodal loss. The Early model is defined again without the MS and AL losses. Unlike the Mid-Late scheme, the Early model requires two forward passes to compute the unimodal representations.

alignment a trivial task. Finally, unimodal and multimodal predictions can be acquired with a single forward pass.

On top of these benchmark models, we define CoRe, the model with the same architecture as CoRe-Sleep but trained without the additional training objectives MS and AL. CoRe follows the Mid-Late model of creating a separate encoder for each modality but with established communication between them. CoRe, Early and Mid-Late models have the same TF backbone and inner-outer logic in each block. The final multimodal representation is created by summing the two unimodal representations, allowing the dimensionality of the total representation to be the same and exploiting the same classifier when one modality is missing. A learnable class embedding [CLS] is utilised to aggregate inner-sequence information, following [36]. Each model is trained on the multimodal predictor CE loss and the two additional losses (MS and AL) are gradually added to show their benefit.

### 2.2.6 Implementation

All our models are implemented in PyTorch [127] and trained on a single GPU node. We use Adam [87] optimizer with the learning rate and the weight decay being 1e-4. We train with batch size 16 and outer sequence 21 [136], resulting in 336 sleep labels per batch. We validate our models every 400 optimization steps. We also use a cosine annealing [109] scheduler with a maximum learning rate of 0.03 and 20k warm-up optimization steps. We consider a model converged when it has not been improved in the last 100k optimization steps (approximately 9 epochs) and then apply early stopping.

Each unimodal and multimodal encoder has an inner and outer block of TF, each with 4 layers of post-normalization TF. Relative positional embedding [156] is only added to key vectors in attention components. Models have 128-dimensional input, 8 attention heads, 1024-dimensional MLP for FF, and 2-layer MLP for predictors, all with 0.3 dropout probability. Weights are shared among the SA and FF parts of the unimodal and multimodal encoders.

### 2.2.7 Experiments

To explore the robustness of the model to noisy modalities, a subset of SHHS-1, where one of the modalities is highly noisy, is extracted. Such noisy cases can be identified by thresholding the difference in STD of EEG and EOG modalities. The STD is evaluated in large chunks for 10 minutes to avoid picking artefacts. Only the patients that present corruption in over 40% of the recording are kept to avoid picking patients with noise segments at the beginning or end of a recording where sensors might not yet be attached or have been detached. This subsampling is applied three times, once for each dataset split with [17, 20, 11] patients having over 40% noisy EEG and [10, 4, 3] having over 40% noisy EOG on the different spits. Notably, EEG, the dominant modality for sleep staging, appears to contain more long noisy periods. We test the models on these sets of patients to evaluate robustness to noisy modalities.

CoRe-Sleep is also capable of training with modality-incomplete data. The motivation of this last experiment is to examine the behaviour of the model under such unimodal data additions. CoRe-Sleep is trained from scratch with an initial subset of 100 multimodal patients and a variable number of unimodal ones (either from EEG, EOG or both). Only on the initial 100 are the MS and AL losses calculated. When both unimodal modalities are added, they don't belong to the same patient. The experiment is repeated three times, gradually adding patients with EEG, EOG, or both modalities.

## 2.3 Results & Discussion

This section demonstrates the contribution of each component in multimodal fusion in standard, missing-modality, and noisy scenarios and analyses the behaviour of our model trained on incomplete multimodal data.

### 2.3.1 Multimodal

Results on training and testing CoRe-Sleep and the benchmark models (Early and Mid-Late) with multimodal input show that optimizing the multimodal fusion leads to outperforming previous state-of-the-art by a small margin, as shown in the first column of Table 2.1. All models, namely Early, Mid-Late and Core, performed similarly when no additional losses were added. However, adding the AL loss improved the performance of Early and Core, the models that include attention interactions between the two modalities. MS loss contributes consistently to all three models. We demonstrate that CoRe-Sleep performs similarly to Early ( when MS + AL are added), which is the model that allows the maximum interactions between modalities.

### 2.3.2 Missing Modality

One of the main advantages of CoRe-Sleep is its ability to handle missing modalities. In this second experiment, the already trained models are evaluated in scenarios where only one of the two modalities is present. The second and third columns of Table 2.1 display the missing modality scenario experimental results. Mid-Late fusion performs the best without additional losses. This is expected; the model with fewer interactions between modalities and each of its encoders is focused on a single modality, not to be affected as much by the absence of one modality. We note that when a modality is absent, the multimodal components of the network are not used; instead, only the unimodal encoders and classifiers are utilized.

We observe how the models behave when extra losses are included. Each model shows a distinct response to the addition of AL loss. In Mid-Late, it has a small contribution. In CoRe, when used without MS, it assists in balancing the performance between the two modalities. AL in the Early model deteriorates both the performances of EEG and EOG. However, when combined with MS, all models exhibit improved performance. Meanwhile, MS loss has a consistent effect on all models, indicating that each modality has to perform on its own as well. All models benefit from MS loss, which has a greater effect on CoRe

Table 2.2: Comparison of model performance (Accuracy, Cohen's $\kappa$ and Macro-F1) on subset of Noisy SHHS-1 Patients

| Dataset: Noisy SHHS-1 | Overall Metrics | | |
|---|---|---|---|
| Test Modalities | EEG-EOG Noisy | | |
| Model | Acc | $\kappa$ | MF1 |
| Unimodal Equivalent EEG | 56.1 $\pm$0.019 | 0.351 $\pm$0.029 | 43.7 $\pm$3.5 |
| Unimodal Equivalent EOG | 80.6 $\pm$2.1 | 0.722 $\pm$0.03 | 69.5 $\pm$2.0 |
| Early | 77.9 $\pm$2.9 | 0.669 $\pm$0.046 | 68.6 $\pm$4.4 |
| Mid-Late | 81.7 $\pm$0.3 | 0.702 $\pm$0.007 | 73.9 $\pm$0.4 |
| CoRe | 79.5 $\pm$1.4 | 0.706 $\pm$2.1 | 67.7 $\pm$1.8 |
| Early + **AL** | 81.1$\pm$1.2 | 0.730$\pm$0.017 | 70.4$\pm$1.1 |
| Mid-Late + **AL** | 82.2$\pm$0.7 | 0.746$\pm$0.009 | 70.2$\pm$0.7 |
| CoRe **AL** | 82.2$\pm$0.3 | 0.745$\pm$0.006 | 71.4$\pm$0.5 |
| Early + **MS** | 81.1$\pm$1.0 | 0.730$\pm$0.015 | 70.2$\pm$1.2 |
| Mid-Late + **MS** | 84.2$\pm$0.5 | 0.774$\pm$0.008 | 73.2$\pm$1.0 |
| CoRe + **MS** | 82.9$\pm$1.1 | 0.754$\pm$0.015 | 71.5$\pm$0.9 |
| Early + **MS** + **AL** | 83.1 $\pm$0.5 | 0.758 $\pm$0.007 | 72.2 $\pm$0.6 |
| Mid-Late + **MS** + **AL** | 83.8 $\pm$1.1 | 0.769 $\pm$0.016 | 73.0 $\pm$1.5 |
| CoRe-Sleep | 84.0 $\pm$1.2 | 0.771 $\pm$0.017 | 73.0 $\pm$0.7 |
| XSleepNet [133] | 75.5 $\pm$2.6 | 0.641 $\pm$0.04 | 61.7 $\pm$4.3 |

and Early. CoRe-Sleep's performance exceeds both in scenarios of missing EEG or EOG over all the other models. It even slightly outperforms its unimodal equivalent and the previous state-of-the-art models that have been trained solely on the present modality.

The contribution of this experiment is twofold. Firstly, it shows that the multimodal fusion and the multi-task objective can strengthen the model's ability to perform inference with a subset of the modalities given during training. This addition allows CoRe-Sleep to overcome issues such as sensor detachment, device deficiency or lack of all recording modalities. Secondly, the results show that, as CoRe-Sleep multimodal training surpasses its unimodal equivalent models for EEG and EOG, unimodal models, in general, can benefit from multimodal training.

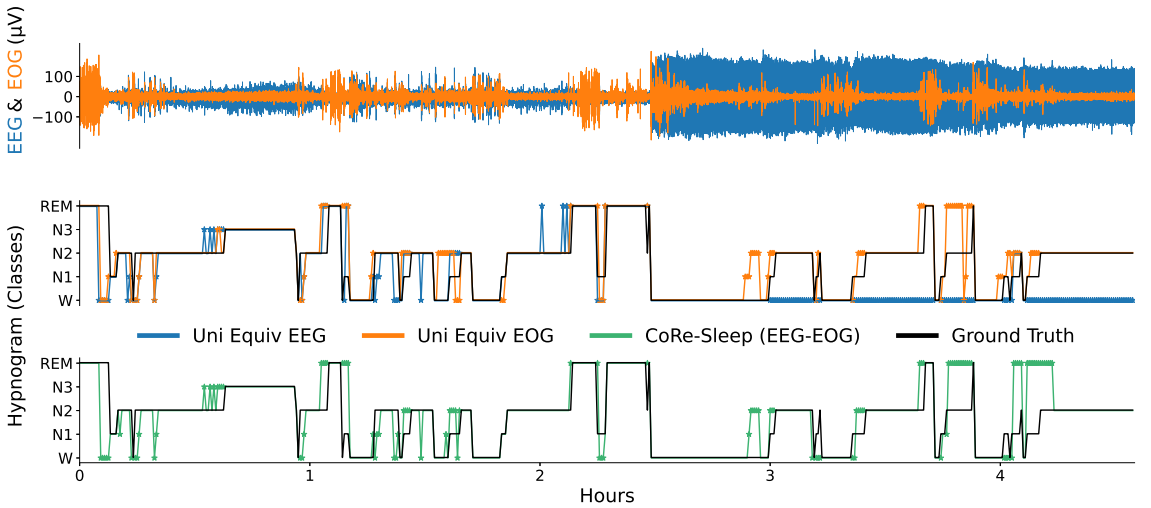



Figure 2.6: This figure provides an illustration of a patient from SHHS-1 whose EEG modality was highly corrupted in the second half of the recording. The top section of the figure displays the EEG (C4-A1) and EOG (L-R) signals, with a sudden increase in the standard deviation of the EEG signal indicating the presence of noise. The two rows below show the class predictions (hypnogram) of the unimodal EEG and EOG models and the multimodal EEG-EOG CoRe-Sleep model. Notably, in the highly noisy area, the EEG model performed poorly, while the multimodal model remained robust by leveraging the information provided by the EOG.

### 2.3.3 Noisy Modalities

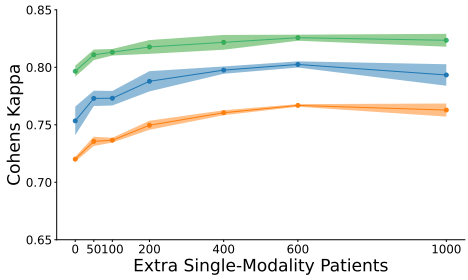
A patient in this subset is visualised in Figure 2.6. After almost two and a half hours, the EEG begins to behave irregularly, highly likely because the electrode is getting detached. The predictions of unimodal EEG in that area can be seen below in blue. When the recording is noisy, the predictions are wrong as continuous Wake is predicted. The CoRe-Sleep model is less affected by noise in the EEG data as it adapts its predictions based on information contained in the EOG. The predictions of unimodal EOG and CoRe-Sleep in the noisy area display great similarities.

Generally, testing on the subset of noisy patients decreases performance in all models. However, such imperfect data quality does occur in real life. Table 2.2 shows an overview of the performance of the noisy subset. Since EEG is more often the broken modality in the subset, its unimodal equivalent model performs substantially worse than the EOG one. Models, Early and CoRe, are underachieved in comparison to EOG only. On the contrary, Mid-Late shows some further robustness to noisy cases. The Mid-Late model has separate encoders for each modality, allowing one to remain uninfluenced by the noisy behaviour of the other. It's the same property that allows it to do well in the missing modality scenario. Adding AL and MS losses benefits all three models. Only in the case of Mid-Late do we see some declined behaviour when both are added compared with adding only the MS loss. CoRe-Sleep manages to perform close to the Mid-Late model, exhibiting robustness in the noisy modality scenario. The added value of using multimodal fusion is highlighted since the predictions based on noisy modality can be amended by the non-noisy ones. Finally, the literature benchmark evaluated on multimodal data, XSleepNet, is an early fusion approach and underperforms in this subset.

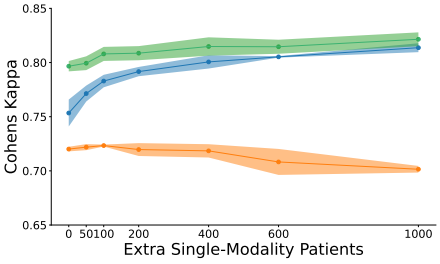
Evaluating a model solely on carefully screened data that excludes any noise or missing modalities, as is often the case with datasets, may not accurately reflect the model's full capabilities. Testing the models on missing or noisy modalities can give a further less biased impression of the models' abilities.

### 2.3.4 Training with Incomplete Data

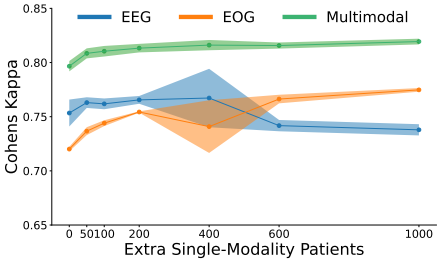
In the last experiment, the ability of CoRe-Sleep to be trained with the assistance of large sets of modality-incomplete data is examined. In Figure 2.7, three different training scenarios are presented. In each scenario, we train models with a different number of training patients. Every model has the same 100 multimodal patients, for which both MS and AL losses are calculated. Furthermore, unimodal patients are added gradually, either from both modalities




(a) Adding patients with only Unimodal EEG or Unimodal EOG




(b) Adding patients with Unimodal EEG




(c) Adding patients with Unimodal EOG

Figure 2.7: CoRe-Sleep was trained using a sub-group of 100 multimodal patients who had both EEG and EOG, as well as some unimodal patients who only had EEG or EOG: **a)** We add patients who provide only EEG and an equal number of patients who provide only EOG. We consistently observe an improvement in the performance of both multimodal and unimodal predictors. **b)** We add patients who provide only EEG. A decrease in EOG predictor performance is noted. **c)** We add patients who provide only EOG. Consistently, we notice a decrease in the modality performance that was not provided. All experiments were repeated three times with different dataset splits to obtain the variance shown.

or from solely EEG or EOG in Figures 2.7a, 2.7b and 2.7c respectively. For those additional patients, we calculate only the unimodal loss concerning the included modality. We emphasize that similar to the missing modalities when training with these additional patients, only the unimodal parts are utilized, while the multimodal parts and the CA components remain inactive. The generalization performance on the test set is measured for each step. In this experiment, parameter sharing among predictors facilitated faster and more stable convergence, particularly when data was limited in size.

Three major findings can be observed:

1. When the number of additional unimodal patients is close to the number of initial multimodal ones (in this case, 100), all three predictors increase their performance. This can be seen in all three experiments.
2. Adding the same number of unimodal patients from both modalities, without being the same patients, favours all three predictors even when adding unimodal ones outmatch by far the initial multimodal ones (see Figure 2.7a).
3. The performance of the other modality deteriorates when adding unimodal patients in much larger numbers than the initial multimodal ones, as shown in Figures 2.7b and 2.7c.

With this experiment, we show that patients' recordings with missing modalities can be included in the training and help to improve overall model performance. Adding unimodal data should be carefully exploited at a large scale (greater magnitudes than multimodal data), especially in cases where inference should come separately from both modalities. CoRe-Sleep can leverage all available data, making it more flexible for real-world scenarios.

## 2.4 Discussion and Conclusion

In this concluding part, we will first discuss the impact of this work on sleep staging and then proceed to discuss its impact on the field of multimodal models and fusion.

Regarding sleep staging, this work offers a new multimodal model architecture, CoRe-Sleep, leveraging coordinated representation architecture, transformer backbones and cross-attention modules to allow the branches of each modality to communicate with each other. CoRe-Sleep sets a new state-of-the-art surpassing by a small margin previous works, both when performing inference with one

modality missing or with both modalities available. At the same time, it remains robust to the case of having a noisy modality. We compare CoRe-Sleep with two other fundamental multimodal structures to reveal that it benefits both worlds when trained under the examined objectives. As noisy recordings or cases in which not all modalities are available cannot be avoided in real life, this work could be a significant step towards automating sleep staging in clinical settings.

The performance of our method, the benchmark, and some previous works have already reached human-level performance, with inter-agreement among doctors falling below $k \leq 0.9$. In future research, we aim to include more diverse test subjects from additional datasets and institutions, encompassing varied demographics and diseases, instead of the healthy subjects that SHHS largely includes. This will allow us to assess better the algorithms' ability to perform uniformly well across diverse scenarios, bringing them closer to their ultimate goal of seamless integration into the healthcare system.

Our study highlights two key insights regarding multimodal machine learning. Firstly, training models with multiple modalities can yield performance improvements, even when testing involves only one of the modalities. Contrary to common practice, training models only on the specific modalities present during testing may not always lead to optimal performance. Secondly, our findings suggest that including training data points that lack some of the modalities can benefit model performance. These insights lay a foundation for robust multimodal learning, allowing models to harness available data across a broad spectrum of applications fully.

### 2.4.1 What Matters Most in Multimodal Learning?

Our work shows that the multimodal performance differences between the architecture choices, CoRe, Early, and Late Fusion, are often minimal in specific scenarios. This raises an important question: how crucial is the fusion point or method when combining modalities?

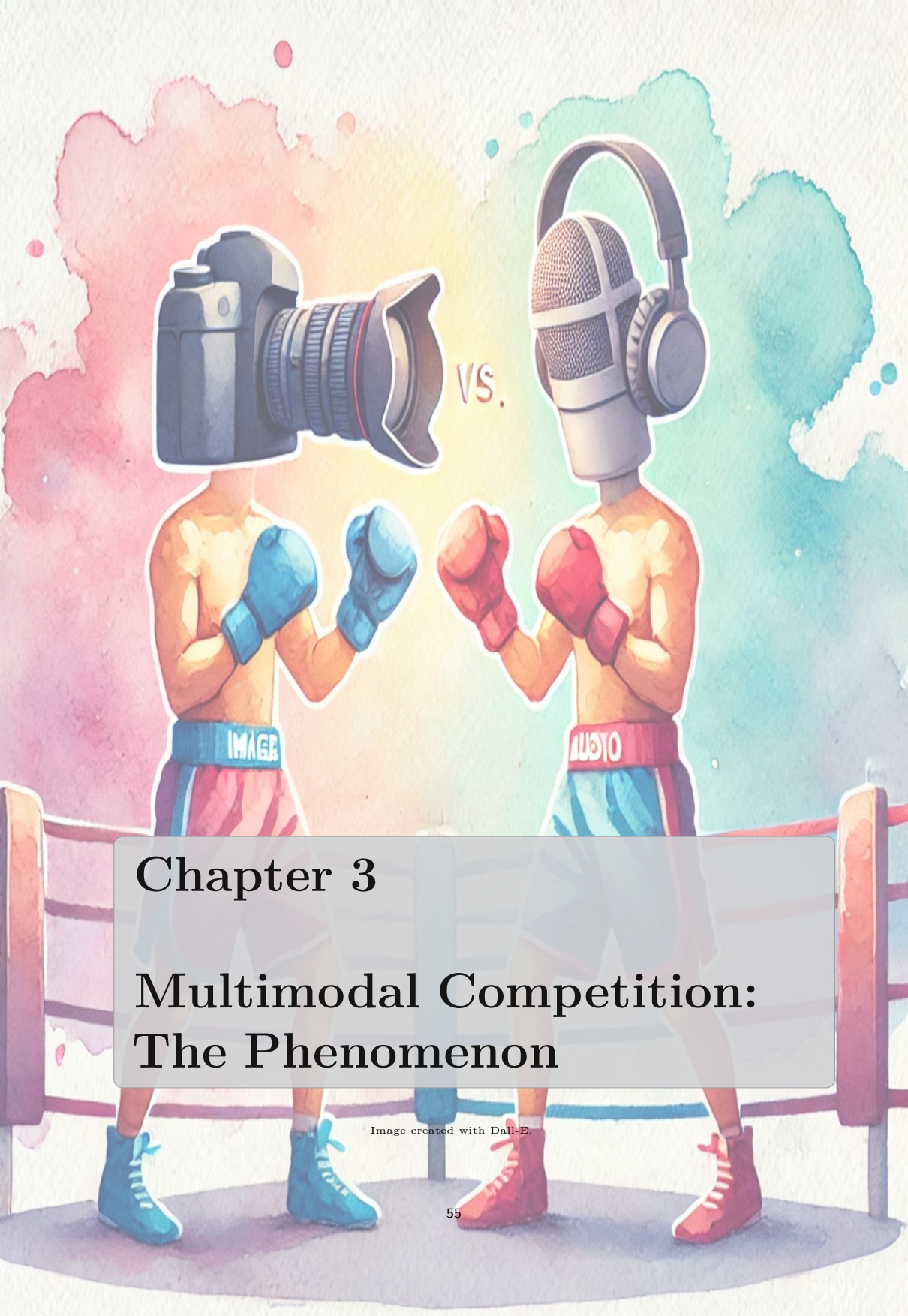
As seen in Table 2.1, models trained under the same objective show similar multimodal performance, suggesting that these architectural choices have a smaller impact on the final results. While they may still offer benefits in terms of robustness to noise or missing modalities (with Early Fusion performing notably worse in these cases, as shown in the second and third columns of Table 2.1 and Table 2.2), they do not appear to provide the significant leap in performance we expect from multimodal models.

Instead, our findings suggest that training objectives play a more pivotal role in

achieving better results. This leads to our second question: Could relying on a single joint task objective, such as cross-entropy, limit the potential performance gains in multimodal learning? The remainder of this thesis investigates this question, examining the challenges of joint multimodal training and proposing solutions to overcome them.

# Part II

# Multimodal Competition




# Chapter 3

# Multimodal Competition: The Phenomenon

Image created with Dall-E.

## Chapters Contribution

Chapter 3 focuses on the phenomenon of modality competition in multimodal learning, where one modality can dominate during joint training, limiting the contributions of others. Using a gradient-based analysis, the chapter uncovers the underlying mechanisms driving this competition and examines its impact on model performance and generalization. It also reviews existing balancing strategies, such as gradient modulation and multi-task learning approaches, outlining their strengths and limitations. This foundational work provides key insights for designing techniques that promote equitable contributions from all modalities, improving training of multimodal systems.

## 3.1 Introduction

In the previous chapter, we observed that varying fusion strategies, such as early fusion, late fusion, or coordinated representations, had only a modest effect on multimodal performance. In contrast, applying tailored training objectives led to significantly larger performance and robustness gains. This suggests that while the choice of fusion strategy plays a role, the design of effective training objectives is a more critical factor for improving robust and effective multimodal learning. In this chapter, we investigate the challenges of training multimodal models with a single joint task objective.

Several works [59, 137, 172, 2, 187] have highlighted optimization and generalization issues that arise when training models with additional modalities, often resulting in decreased performance. This outcome contradicts the expectation that additional information would aid model optimization and improve generalization. Wang et al. [187] encapsulate these findings, demonstrating across several datasets that multimodal models can experience performance drops to the extent that they underperform compared to their best unimodal counterparts. This surprising result emphasizes the complexity of multimodal learning and suggests that simply adding more modalities does not guarantee improved performance. Instead, careful consideration must be given to how these modalities interact during training to avoid negative interference.

Huang et al. [73] proceed by defining this issue as multimodal competition. They show that in joint training, particularly when they elaborate on late fusion models, there is a significant likelihood that a subset of modalities will dominate the learning process, leaving other modalities underexplored. Dominating modalities exhibit a higher correlation with the output, while the neglected modalities maintain a lower, relatively unchanged correlation throughout the entire training process. These observations highlight the need for a deeper understanding of the causes of multimodal competition.

In this chapter, we begin with a gradient analysis of multimodal models, illustrating why training under a joint objective makes them susceptible to multimodal competition. We show that gradient updates for one modality can be influenced by terms from another modality, potentially causing misaligned or diminished updates that disrupt effective learning. Subsequently, we provide a detailed review of prior studies that propose strategies to address this challenge, focusing on methods to mitigate multimodal competition and enhance training efficiency. This review establishes the groundwork for the following three chapters, which present experimental results aimed at improving training dynamics and evaluating the impact of various potential solutions.

## 3.2 Gradient Interactions

In this section, we investigate the training process of multimodal models and shed some light on the training characteristics that allow modality competition to occur. Our analysis centres on models trained with a single joint multimodal objective, and we hypothesize that any model, regardless of architecture, parameter size, fusion strategy, or additional loss terms, is susceptible to modality competition. Even when modalities do not interact in the forward pass, we show that they influence each other's parameters update in the backward pass, affecting the learning dynamics. This underscores the importance of carefully considering the training procedure and gradient estimation to develop effective multimodal models.

We consider a classification task with a dataset $D$ of $N$ i.i.d. datapoints, where each datapoint has two modalities $(X_1, X_2)$. These modalities are expressed as sets of varying numbers of features with different statistical properties, denoted as $(X, y) = (x_{1_i}, x_{2_i}, y_i)$ for $i = 1, \ldots, N$, where $x_{1_i} \in \mathbb{R}^{FD_1}$, $x_{2_i} \in \mathbb{R}^{FD_2}$, and $y_i \in 1, \ldots, M$ with M representing the number of classes. The goal is to determine a parameterized function $f(X; \theta)$ to predict y. Each modality is encoded using a unimodal encoder $f_m : X_m; \theta_m \rightarrow Z_m$, which transforms the input $X_m$ into a latent representation $Z_m$. The fusion network $f_c : [Z_1, \ldots, Z_M]; \theta_c \rightarrow y$ then combines these representations to predict the target. We consider models with a single output from both modalities. Such networks, regardless of their fusion operation (early, mid, late, decision), can be cast as:

$$f(X; \theta) = f_c(f_1(x_1; \theta_1), f_2(x_2; \theta_2); \theta_c) \tag{3.1}$$

As an illustration, if we choose to fuse on the decision level, then $f^c = \mathbf{I}$, representing the identity matrix, where $c$ denotes the network that combines in its input both modalities, and $\theta_c = \emptyset$. Alternatively, in the case of Late fusion, $f_c$ corresponds to a linear layer. The $f_1$ and $f_2$ are the unimodal encoders with $\theta_1$ and $\theta_2$ being parameters dedicated to each modality. The dimensionality of the feature vectors produced by the unimodal encoders can vary in size $\phi_v = f_1(x_1; \theta_1) \in \mathbb{R}^{\Phi_2}$ and $\phi_2 = f_2(x_2; \theta_2) \in \mathbb{R}^{\Phi_2}$.

We consider a Late fusion model with a linear classifier. We use concatenation noted as $[\cdot, \cdot]$ for fusion. The final model structure is summed as $f(\phi_1, \phi_2) = \mathbf{W}_o[\phi_1, \phi_2] + b_o = \mathbf{W}_{o_1}\phi_1 + \mathbf{W}_{o_2}\phi_2 + b_o$. The gradient of one of the parameters of the encoder of modality $X_1$ in that case would be:

$$\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial f(X; \theta)} \frac{\partial f(X; \theta)}{\partial \theta_1} = \frac{\partial L}{\partial f(X; \theta)} \mathbf{W}_{o_1} \frac{\partial f_1(x_1; \theta_1)}{\partial \theta_1} \tag{3.2}$$

We proceed to examine each term separately. The term $\mathbf{W}_{o_1} \frac{\partial f_1(x_1;\theta_1)}{\partial \theta_1}$ is only affected by the modality $X_1$ and the parameters related to that modality as expected. Using softmax cross entropy we can transform the term $\frac{\partial L}{\partial f(X;\theta)} = p - y$ where p is a vector with components:

$$p_i = \frac{e^{\mathbf{W}_{o_1}\phi_1 + \mathbf{W}_{o_2}\phi_2 + b_o}}{\sum_{j=1}^{n} e^{\mathbf{W}_{o_1}\phi_1 + \mathbf{W}_{o_2}\phi_2 + b_o}} = \frac{e^{\mathbf{W}_{o_1}\phi_1}}{Z} + \frac{e^{\mathbf{W}_{o_2}\phi_2}}{Z} + \frac{e^{b_o}}{Z} \tag{3.3}$$

Notably both the second term, $\frac{e^{\mathbf{W}_{o_2}\phi_2}}{Z}$, and the normalization factor $Z$ include contributions from both the modality $X_2$ and its corresponding weights, $\mathbf{W}_{o_2}$. As a result, the gradients of the parameters associated with the other modality, $X_1$, are indirectly influenced by the presence of $X_2$ and its parameters. This interaction between modalities can affect the learning dynamics between modalities during training.

A similar observation applies to any non-linear loss function (e.g., MSE, MAE, binary cross-entropy) and more complex model architectures, as the interplay between modalities will continue to influence the gradient flow and overall learning behaviour.

Multimodal models have traditionally focused on determining the stage at which multimodal fusion occurs, whether through early or late fusion strategies [90] or by allowing one modality to influence another early in the process [35]. This body of work has primarily concentrated on designing the forward pass architecture. However, we argue that the backward pass is equally, if not more important, specifically, how one modality's gradients influence another's parameters. Understanding and optimizing this interaction is crucial for improving the overall performance of multimodal models.

## 3.3 Review of Previous Balancing Strategies

Previous works have made numerous efforts to face the optimization problem introduced by the multimodal competition. We will proceed with a detailed analysis of some of these methods, their assumptions about the situation, and their limitations logically and computationally. While these methods can not necessarily be grouped, we will broadly categorise the Gradient Balancing and Multi-Task methods. Below, we will elaborate on methods that belong to these groups and others, presenting them in order from the simplest assumptions to the most intricate solutions for the problem.

#### Multimodal Cosine (MMCosine) [201]

Although it is one of the latest works in the field, we will begin our analysis with the Multi-Modal Cosine (MMCosine) method since it offers an orthogonal point of view compared to the rest of the methods.

Xu et al. introduce the MMCosine loss, which performs modality-wise L2 normalization on features and weights and relies on the cosine similarity metric to balance the optimization process between modalities. They constrain themselves on Late fusion and apply such a technique to remove the effect of the magnitude in the linear fusion operation of the network. They assume that what causes the imbalanced training lies in the magnitude difference of each $\mathbf{W}_{o_i}\phi_i$ term of equation $f(\phi_1, \phi_2) = \mathbf{W}_{o_1}\phi_1 + \mathbf{W}_{o_2}\phi_2 + b_o$. Therefore, they suggest using the scaled cosine similarity between the features and the weights, $\cos\theta_{1_j} = \frac{\mathbf{w}_{o_{1_j}} \cdot \phi_{1_i}}{\|\mathbf{w}_{o_{1_j}}\| \|\phi_{1_i}\|}$ to derive the prediction. The final suggested loss is:

$$L_{\text{mmcosine}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cdot(\cos\theta_{1_i} + \cos\theta_{2_i})}}{\sum_{j=1}^{n} e^{s\cdot(\cos\theta_{1_j} + \cos\theta_{2_j})}}, \tag{3.4}$$

where $s \in \mathbb{R}^1$. By doing so, MMCosine effectively removes radial variations in weights and features, which helps mitigate some aspects of imbalance. However, we argue that the problem persists, as the components that influence the gradient of one modality over the other are still present. These components continue to affect (possibly with a different magnitude) the gradients of the parameters of the other modality, thereby not alternating much the source of the problem itself.

### 3.3.1 Gradient Balancing Methods

#### Modality-Specific Learning Rates (MSLR) [202]

Introduced by Yao et al., Modality-Specific Learning Rates (MSLR) identify the problem of overfitting to one modality during multimodal training. They argue that when training two modalities together, the encoder for one modality might learn significantly faster, leading to overfitting, while the other modality is left under-optimized. This observation has been echoed by several other works, which similarly point to imbalances in the learning dynamics of different modalities as a key issue.

At the core of the problem, Yao et al. identify that the gradient magnitudes for each modality's encoder differ, with one modality dominating the training

process. This dominance hinders the other modality from contributing meaningfully to the learning process. To address this, they propose to adapt the learning rate of each encoder based on its recent performance on the training batch in each optimization step, ensuring that both modalities learn at an appropriate pace. MSLR focuses on late fusion models similar to MMCosine.

In MSLR, they use two biases separately for each modality, casting the final multimodal output as $f(\phi_1, \phi_2) = \mathbf{W}_{o_1}\phi_1 + b_{o_1} + \mathbf{W}_{o_2}\phi_2 + b_{o_2}$. Then they extract the unimodal performance by separating half of this last layer using $f_{uni_1}(\phi_1) = \mathbf{W}_{o_1}\phi_1 + b_{o_1}$ and $f_{uni_2}(\phi_2) = \mathbf{W}_{o_2}\phi_2 + b_{o_2}$. They use these estimations to find the performance of each modality separately on each optimization step $s_1$ and $s_2$ and then estimate the following ratios:

$$s_{1_i} = \sum_{k=1}^{C} \mathbf{1}_{k=y_i} \cdot \text{softmax}(\mathbf{W}_{o_1} \cdot \phi_1 + b_{o_1})_k = CE(\mathbf{W}_{o_1} \cdot \phi_1 + b_{o_1}, Y), \quad (3.5)$$

$$s_{2_i} = \sum_{k=1}^{C} \mathbf{1}_{k=y_i} \cdot \text{softmax}(\mathbf{W}_{o_2} \cdot \phi_2 + b_{o_2})_k = CE(\mathbf{W}_{o_1} \cdot \phi_1 + b_{o_1}, Y) \quad (3.6)$$

$$\rho_{1_i} = \frac{s_{1_i}}{\frac{1}{\text{Mem}} \sum_{i=1}^{\text{Mem}} s_{1_{(i-j)}}}, \quad \rho_{2_i} = \frac{s_{2_i}}{\frac{1}{\text{Mem}} \sum_{i=1}^{\text{Mem}} s_{2_{(i-j)}}}. \quad (3.7)$$

Instead of using a single global learning rate, MSLR adjusts the learning rate $\alpha$ separately for each modality at each optimization step $i$ using the estimated ratios $\rho_{1_i}$ and $\rho_{2_i}$. The learning rate of that step is then performed with:

$$\text{lr}_{1_i} = \rho_{1_i} \cdot \text{lr}_{1_i}, \quad \text{lr}_{2_i} = \rho_{2_i} \cdot \text{lr}_{2_i} \quad (3.8)$$

While MSLR recognizes that each modality's encoder can converge at a different speed and managing the different learning rates can compensate for it and mitigate this phenomenon, they do not solve the problem. Their two main limitations arise in their way of estimating unimodal performance and in their ratio. First, they use half of the last linear layer to access the unimodal performance, which casts doubts on whether it is a valid assessment. Second, although MSLR optimizes the learning rate for each modality, the interaction between modalities during gradient updates is not explicitly considered. This leaves room for further exploration in holistically handling the competition between modalities.

### On-the-fly Gradient Modulation (OGM) [129]

On-the-fly Gradient Modulation (OGM), introduced by Peng et al., addresses the problem of modality competition by modulating the gradients' magnitude of different modalities during the training process. This approach is motivated by the same observation as MSLR, that one modality often fits the problem faster, then overfits and leads to under-optimized representations in the other modality. It performs the same operation as MSLR, changing the learning rate or gradient magnitude at each optimization step. It improves upon MSLR since it considers the interaction of modalities in the ratio estimation.

Starting again from a Late fusion method, OGM measures unimodal contribution as MSLR through the unimodal performance of half of the final layer. Here, only one bias is used, and it gets halved in each unimodal estimator.

$$s_{1_i} = \sum_{k=1}^{M} \mathbf{1}_{k=y_i} \cdot \text{softmax}(\mathbf{W}_{o_1} \cdot \phi_{1_i} + b_{o_1})_k, \tag{3.9}$$

$$s_{2_i} = \sum_{k=1}^{M} \mathbf{1}_{k=y_i} \cdot \text{softmax}(\mathbf{W}_{o_2} \cdot \phi_{2_i} + b_{o_2})_k, \tag{3.10}$$

The authors propose monitoring the contribution of each modality to the overall learning objective by introducing a discrepancy ratio, denoted as $\rho_t^u$, which measures the relative performance of each modality at each training step. This ratio is calculated using the softmax outputs of each modality:

$$\rho_{1_t} = \frac{\sum_{i \in B_t} s_{2_i}}{\sum_{i \in B_t} s_{1_i}}, \quad \rho_{2_t} = \frac{1}{\rho_{1_t}}, \tag{3.11}$$

where $B_t$ are the datapoints within the mini-batch. Based on the calculated discrepancy ratio $\rho_t^u$, a modulation factor $k_t^u$ is applied to adjust the gradient of each modality:

$$k_{1_t} = \begin{cases} 1 - \tanh(\alpha \cdot \rho_{1_t}), & \text{if } \rho_{1_t} > 1, \\ 1, & \text{otherwise.} \end{cases}, \quad k_{2_t} = \begin{cases} 1 - \tanh(\alpha \cdot \rho_{2_t}), & \text{if } \rho_{2_t} > 1, \\ 1, & \text{otherwise.} \end{cases} \tag{3.12}$$

Here, $\alpha$ is a hyperparameter that controls the degree of modulation. The adjusted gradient is then used to update the model parameters for each modality:

$$\theta_{1_{t+1}} = \theta_{1_t} - \eta \cdot k_{1_t} \cdot g(\theta_{1_t}),$$

where $g(\theta_{1_t})$ is the stochastic gradient for the parameters of modality $X_1$ at time step $t$. This operation is equated with the change in the learning rate of the MSLR. By applying these modulation factors, OGM ensures that the dominant modality (with a higher $\rho$) has a reduced gradient, allowing the under-optimized modality to receive more attention during training. This results in a more balanced optimization process across modalities.

However, Peng et al. support that OGM alone can reduce the generalization ability of the model due to its dampening effect on the stochastic gradient noise, which has been shown to impact generalization positively. To address this, the authors introduce an adjustment to their method by injecting some Gaussian noise into the gradients, Generalization Enhancement (GE):

$$\theta_{1_{t+1}} = \theta_{1_t} - \eta \cdot (k_{1_t} \cdot g(\theta_{1_t}) + h(\theta_{1_t})),$$

where $h(\theta_{1_t}) \sim \mathcal{N}(0, \Sigma_{\text{sgd}}(\theta_{1_t}))$ represents the added Gaussian noise, ensuring that the stochastic gradient noise is maintained or even enhanced. While we show here the update step with SGD [150], similarly can be achieved with any other stochastic gradient optimizer.

OGM introduces a method that compares the training performance of each modality and slows down the learning of the faster modality to balance the training pace between modalities. Additionally, it incorporates gradient noise injection to enhance generalization. While this method improves upon simple joint objective training, it estimates unimodal performance by using only half of the final linear layer, which may serve as an unreliable estimator and restrict its application to Late fusion models. Furthermore, OGM only decelerates the learning rate, which can significantly prolong the training process. Despite these limitations, OGM offers a promising approach for addressing optimization imbalances in multimodal learning and has laid the foundation for subsequent work in the field.

### 3.3.2 Multi-Task Methods

Methods in this category go beyond the single joint loss objective, often incorporating unimodal local task losses applied specifically to the parameters of each unimodal encoder. Early works aimed at improving multimodal training, such as CentralNet [181], employed such losses to enhance performance. We refer to the approach that utilizes additional linear classifiers on the unimodal representations and applies losses directly to them, without any coefficient balancing, as **Multi-Loss** method. Below, we explore some prominent methods from this field.

### Prototypical Modal Rebalance (PMR) [43]

The Prototypical Modal Rebalance (PMR) method, introduced by Fan et al., improves upon OGM by addressing the unreliability of unimodal performance estimation during training. To tackle this, Fan et al. proposed a non-parametric prototype-based approach, where prototypes are defined as "representative embeddings of instances of a class." These prototypes are used to evaluate the unimodal performance and offer a more stable performance estimator that does not rely on additional model parameters. This allows for a more general application across different fusion strategies, improving upon the limitations of previous methods.

In PMR, each class is represented by a prototype, and classification is performed by computing the distances between the input features and the class prototypes. The prototypes are defined as the centroids of each class, and this distance-based classification is performed for each modality independently. For each modality $m$, the prototype $c_m^k$ for class $k$ is calculated as $c_m^k = \frac{1}{N_k} \sum_{i=1}^{N_k} \phi_{m_i}^k$ with $N_k$ is the number of instances in class $k$, and $\phi_{m_i}^k$ represents the feature embedding for modality $m$ of instance $i$ in class $k$. The classification is then performed by computing the distance between the feature embedding $\phi_{m_i}$ of an instance $i$ and the prototypes. This distance is used in a softmax classifier:

$$p_{\phi_m}(y = k|x_m) = \frac{\exp(-d(\phi_m, c_m^k))}{\sum_{k'} \exp(-d(\phi_m, c_m^{k'}))}, \tag{3.13}$$

where $d(\cdot, \cdot)$ denotes the distance function, in this case the Euclidean distance.

To monitor modality imbalance, PMR introduces the same ratio as in OGM, leveraging its unimodal prototype-based classifiers:

$$\rho_t = \frac{\sum_{i \in B_t} p_{\phi_0}(y_i|x_i^0)}{\sum_{i \in B_t} p_{\phi_1}(y_i|x_i^1)} \tag{3.14}$$

This imbalance metric, $\rho_t$, evaluates the degree of learning progress of each modality, helping dynamically adjust their learning. A notable difference between OGM and PMR is the inclusion of the unimodal Prototypical Cross-Entropy (PCE) losses, which contribute to the good performance of the classifiers. Additionally, PMR differs in how it estimates the coefficients that will affect the learning and how the framework exploits them. Instead of changing the learning rate of each encoder (or equally its gradient magnitude), it applies

these coefficients to the magnitude of the unimodal losses as follows:

$$\beta = 0, \quad \gamma = \text{clip}\,(0, \rho_t - 1, 1) \quad \text{if } \rho_t \geq 1 \tag{3.15}$$

$$\beta = \text{clip}\left(0, \frac{1}{\rho_t} - 1, 1\right), \quad \gamma = 0 \quad \text{if } \rho_t < 1 \tag{3.16}$$

$$L_{\text{PMR}} = L_{\text{CE}} + \alpha \cdot \left(\beta L^0_{\text{PCE}} + \gamma L^1_{\text{PCE}}\right) \tag{3.17}$$

where $\alpha$, $\beta$, and $\gamma$ are hyperparameters controlling the contribution of each loss component. This difference means that the coefficients balance only the contribution of the unimodal losses and not the multimodal. Subsequently in Chapter 4, we will show this is an important limitation. Moreover, we notice that if $\rho_t \geq 1$ then $\beta = 0$ or if $\rho_t < 1$ then $\gamma = 0$, meaning that on each optimization step, only one of the unimodal losses is activated. Experimentally, this and the prototype-based classifiers have caused us training instabilities.

### Gradient Blending (G-Blend) [187]

Gradient Blending (G-Blend) is one of the first works to report multimodal competition and propose a solution for training multi-modal networks. Focusing on the issues of overfitting and imbalance between modalities, G-Blend, like MSLR, PMR, and OGM, observes that networks tend to overfit to one modality while others generalize at different rates. G-Blend introduces a multi-task objective, similar to PMR, with coefficients that balance the unimodal losses, utilizing parameterized unimodal classifiers to address this.

The Overfitting-to-Generalization Ratio (OGR) is introduced to quantify how well the model is generalizing versus overfitting. OGR is computed as the ratio of overfitting to the generalization gain during training:

$$OGR = \left|\frac{\Delta O}{\Delta G}\right| = \left|\frac{(A^{\text{train}}_t - A^{\text{val}}_t) - (A^{\text{train}}_{t-1} - A^{\text{val}}_{t-1})}{A^{\text{val}}_t - A^{\text{val}}_{t-1}}\right| \tag{3.18}$$

where $A^{\text{train}}_t$ and $A^{\text{val}}_t$ represent training and validation accuracy at step $t$, respectively. A high OGR indicates overfitting, while a low OGR suggests better generalization. The blending weights $w_i$ for each modality are computed $w_i = \frac{1}{Z}\frac{G_i}{O_i^2}$ with $G_i$ is the gradient magnitude, $O_i^2$ is the squared overfitting measure, and $Z$ is a normalization constant ensuring $\sum_{i=1}^{M} w_i = 1$. The overall loss is then defined as:

$$L = w_1 L_{\text{uni}_1} + w_2 L_{\text{uni}_2} + w_{\text{multi}} L_{\text{multi}} \tag{3.19}$$

Wang et al. provide a theoretical optimum for the Overfitting-to-Generalization Ratio (OGR) but conclude that it is intractable in practice. Instead, they approximate OGR using accuracy on a separate validation set. This is the set that $A^{\text{val}}$ in Eq 3.18 is estimated on. This estimation is computationally expensive and can only be computed every few steps (typically at the end of an epoch). The need for these infrequent and costly computations, along with the requirement for additional validation data, has been a notable limitation of this method.

#### Unimodal Teachers [39]

Finally, we conclude this section on multi-task methods with one approach by Du et al. [39], which employs different types of local unimodal losses by comparing the features of pre-trained unimodal teacher models $\phi^*$ to the corresponding unimodal features within a multimodal model. This method offers no further balancing mechanisms and will not be discussed in detail. However, it serves as an example of alternative types of losses that can be used instead of the unimodal task loss. The loss function applied in this case is:

$$L_{\text{UMT}} := L_{CE} + \lambda \sum_{s \in [2]} \|\phi_s^*(X_i) - \phi_s(X_i)\|_2^2 \tag{3.20}$$

### 3.3.3 Additional Methods

This section explores methods that do not fit into the Gradient Balancing or Multi-Task methods but offer alternative approaches to multimodal learning challenges. We will briefly review their contributions and limitations.

#### Coordinated Representation Balancing [196]

Wu et al. [196] propose using unimodal training losses instead of a joint multimodal loss, keeping each modality separate within a coordinated network architecture while allowing them to exchange only context information. This design allows for a direct comparison of gradient magnitudes for each modality since the loss for each branch is influenced only by one modality, thereby avoiding the gradient interaction issues discussed in Section 3.2. Specifically, they measure the gradient magnitude of each modality and adjust training focus accordingly, not by altering the gradient itself as in the Gradient Balancing methods, but by incorporating phases of isolated training for the underperforming modalities. While gradient decomposition from each modality can offer useful insight into

the training state of each encoder, such an approach is intractable for most models. For this strategy to work, the model architecture must constrain itself to separate task losses for each modality rather than a joint loss, potentially limiting multimodal integration performance. Therefore, we proceed to compare this method with the previous approaches discussed.

### Adaptive Gradient Modulation (AGM) [96]

The Adaptive Gradient Modulation (AGM) method, proposes a different approach to addressing the challenge of modality competition in multimodal learning. AGM leverages a Shapley value-based attribution technique to isolate the contribution of each modality and adjust the gradient update accordingly. Specifically, AGM monitors the "mono-modal" response for each modality $f^1(X_1, X_2) = \frac{1}{2}\left[f(X_1, X_2) - f(0^{X_1}, X_2) + f(X_1, 0^{X_2})\right]$, representing the output of the model in the absence of modality competition.

This response is used to compute the modality-specific cross-entropy loss $s_m$, and its running average $\hat{s}_m$, which reflects the information each modality contributes. Similarly to MSLR, PMR and OGM, this method uses the estimation of the unimodal losses (this time on top of Shapley values) and their running means to estimate the modulation coefficients:

$$\kappa_m = \exp(-\alpha \cdot (r_m - \tau_m)), \tag{3.21}$$

where $r_m$ is the relative performance of modality $m$, $\alpha$ is a hyperparameter controlling the modulation intensity, and $\tau_m$ is the running average reference. AGM dynamically modulates the gradients of each mono-modal predictor similarly to OGM, ensuring that weaker modalities receive stronger gradient updates while dominant modalities are down-weighted.

Unlike earlier methods, such as OGM or MSLR, which are limited to late fusion models, AGM is model-agnostic and applies to various fusion strategies. AGM combines OGM's modality performance comparison with MSLR's running metrics to derive coefficients that account for interactions between modalities and incorporate information from previous steps. More importantly, AGM leverages Shapley values, making it one of the few methods not relying directly on unimodal performance to assess contribution. This avoids using an ill-posed metric for unimodal contribution, a topic we will elaborate on in Chapter 6.

Two limitations of this method are that it employs zero-masking Shapley values and is trained directly to maximize them. Shapley values can yield arbitrary or unreliable estimates, particularly in high-dimensional settings [112]. Moreover, these methods require input perturbation, which involves additional

forward passes, thereby increasing both computational and memory demands proportionally.

#### Maximizing Functional Entropy [52]

Gat et al. propose a method to reduce bias in multimodal classifiers using functional entropy. Their approach provides an information-theory-inspired method to measure unimodal contribution by independently assessing the model's output variance when noise is added to each modality. Specifically, they measure the sensitivity of the softmax prediction to Gaussian perturbations $Z$ of the input using the following function:

$$f(Z, X) = CE(p(y|Z), p_w(\cdot|X)), \tag{3.22}$$

where the random perturbation $Z$ is sampled from a Gaussian distribution with an expected value of $X$. This allows them to assess how sensitive the model's predictions are to changes in the input, assessing the distance of the model's output when noise is added to each modality independently. They introduce a regularization term based on the log-Sobolev inequality, linking functional entropy to functional Fisher information. By maximizing this Fisher information, the model encourages balanced attention across modalities, essentially pushing each modality to maximize its variance of the CE loss when noise is introduced.

While this method avoids directly manipulating gradient magnitudes and offers a more compact approach, it has two significant limitations. First, pushing for a high variance under noise contradicts the goal of having Lipschitz-smooth neural networks, making the model more sensitive to input noise and vulnerable to adversarial attacks. Second, introducing noisy examples for each modality increases computational demands, as each sample requires an additional forward pass to estimate the variance. This method will be further elaborated in Chapter 6.

## 3.4 Conclusion

In this chapter, we introduced the concept of modality competition, highlighting the generalization challenges multimodal models face. Through a gradient analysis of multimodal fusion networks trained under a single joint objective, we demonstrated that even in models without explicit interactions between modalities (e.g., late fusion), the training of each unimodal encoder is inherently influenced by all input modalities. This influence arises because the gradient

of the loss function integrates information from all modalities, resulting in two potential outcomes:

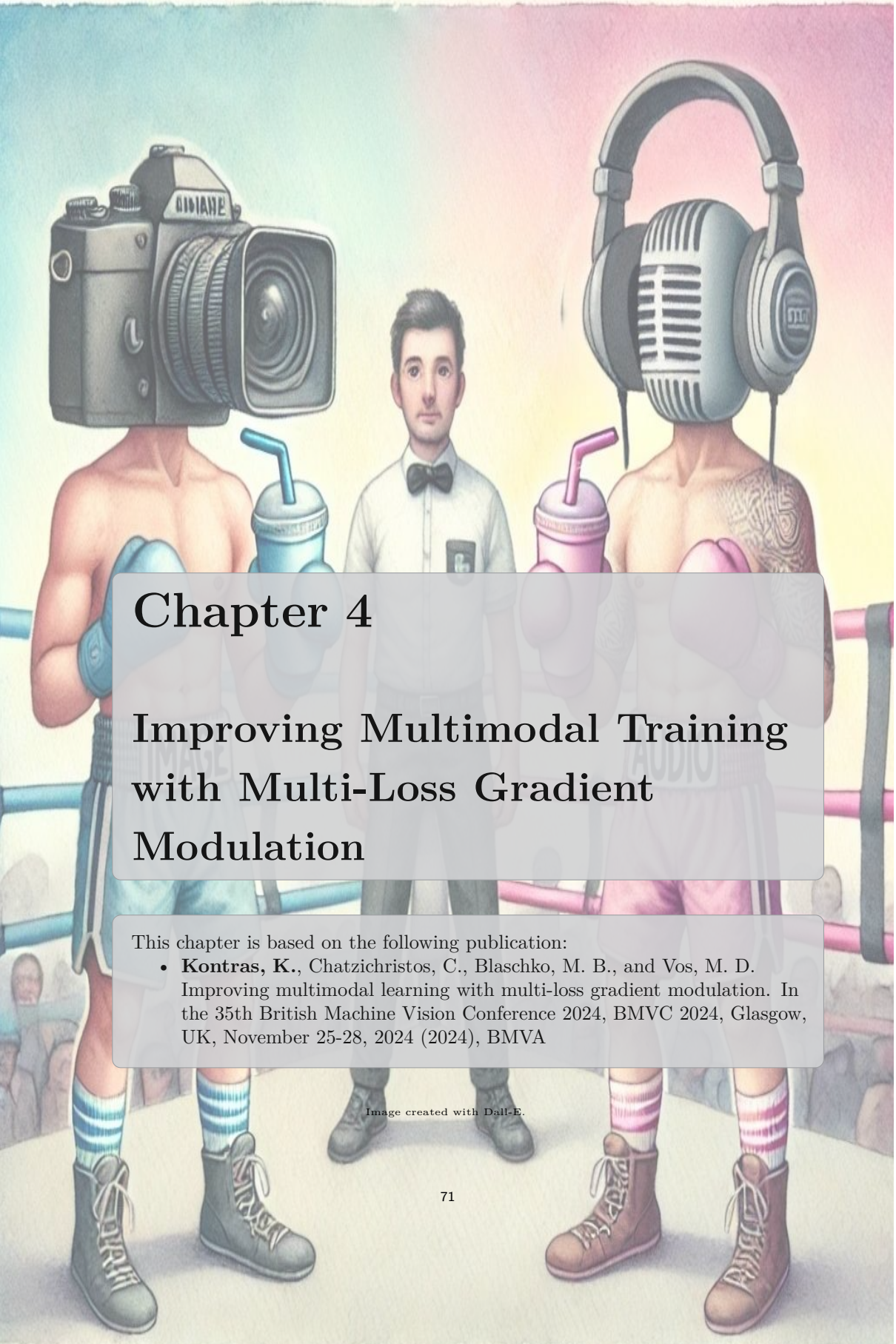
- Constructive Influence: Modalities filter out uncorrelated, modality-specific noise, enhancing their learning.
- Destructive Influence: Misdirected gradient updates hinder weaker modalities, potentially leading to vanishing gradients and overfitting to stronger modalities.

While the factors determining which behaviour prevails remain unclear, they motivate the exploration of methods that address this issue. Therefore, we provide an extensive review of prior multimodal balancing methods [1] along with their assumptions about the problem, which include:

- Feature Normalization: Methods that normalize features at the fusion point to equalize the contribution of the modalities.
- Adaptive Learning Rate Methods: Techniques that adjust encoder learning rates based on unimodal performance to synchronize the learning pace of modalities.
- Unimodal Task Loss Balancing: Strategies incorporating additional unimodal task losses guide multimodal learning.
- Explainability-Based Approaches: Methods that estimate unimodal contributions using explainability techniques and use either to balance losses or as a regularization term.

While previous works propose various methods to address the problem, we identify several intuitive and practical limitations. These include assumptions about the root cause of the problem (e.g., expecting equal modality contributions), methods for assessing unimodal contributions (e.g., based on unimodal performance), balancing strategies (e.g., in the adjustment of unimodal learning rates), and computational complexity. Moreover, inconsistencies in their reporting make it challenging to evaluate their effectiveness on the problem. In the following chapters, we will experimentally evaluate these methods in several datasets, analyzing their effectiveness in mitigating the issue of modality competition, and we will propose improvements to address their limitations.

[1] At the time of writing, additional works such as MMPareto [191] and an extension of OGM [192] have emerged. We encourage readers to explore these studies and stay updated on developments that may further advance the methods discussed in this thesis.



# Chapter 4

# Improving Multimodal Training with Multi-Loss Gradient Modulation

This chapter is based on the following publication:

- **Kontras, K.**, Chatzichristos, C., Blaschko, M. B., and Vos, M. D. Improving multimodal learning with multi-loss gradient modulation. In the 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024 (2024), BMVA

Image created with Dall-E.

## Chapters Contribution

Chapter 4 introduces the Multi-Loss Balancing (MLB) framework, a new approach to improve upon previous works on the challenges of balancing modality contributions during multimodal training. By integrating unimodal and multimodal gradients within a unified framework, MLB effectively mitigates performance inconsistencies observed in existing methods. The chapter provides a comprehensive evaluation across multiple datasets, encoders and fusion methods, demonstrating that MLB achieves superior performance by leveraging both unimodal losses and balancing gradient contributions dynamically. This work establishes MLB as a robust strategy for enhancing multimodal training and promoting equitable learning across modalities.

## 4.1 Introduction

In the previous chapter, we elaborated on the suggested methods to tackle multimodal competition and improve multimodal learning. In this chapter, we will showcase an improvement upon those previous methods. Works in the first category (Section 3.3.1) adjust the learning rate of unimodal encoders based on their estimated predictive performance [129, 202], but they do not explicitly train the model to achieve accurate unimodal predictors. Meanwhile, the methods in the second category (Section 3.3.2) employ additional loss functions derived from unimodal predictions, enabling more accurate estimation, but only balance the unimodal losses during multimodal learning [181, 187, 134, 43, 39], failing to address the imbalances introduced by the multimodal loss.

We aim to bridge both approaches by incorporating additional losses and simultaneously adjusting the learning rates. We improve upon previous work by introducing a multi-loss objective and further refining the balancing process, allowing it to dynamically adjust the learning pace of each modality in both directions, acceleration and deceleration, with the ability to phase out balancing effects upon convergence. We achieve superior results across three audio-video datasets: on CREMA-D, models with ResNet backbone encoders surpass the previous best by 1.9% to 12.4%, and Conformer backbone models deliver improvements ranging from 2.8% to 14.1% across different fusion methods. On AVE, improvements range from 2.7% to 7.7%, while on UCF101, gains reach up to 6.1%. Finally, we demonstrate on the CMU-MOSEI dataset, that our approach is valid when three modalities (audio, video, and text) are available. Our methodology provides the following improvements and novelties:

- Throughout the entire training process, we employ a multi-objective loss that not only ensures each encoder converges close to its unimodal optimum, similarly to the methods described in [181] and [187], but also facilitates accurate unimodal performance estimation. We apply balancing for both multimodal and unimodal losses, distinguishing our method from any prior efforts in these categories.
- The proposed balancing technique, inspired by [129], adaptively modifies the learning rates of encoders based on unimodal performance assessments. We enhance this strategy by enabling both acceleration and deceleration, tailored to the relative performance of the modalities.
- The balancing equations are designed such that when all unimodal encoders converge, the balancing naturally phases out, eliminating the need to predetermine explicitly which epoch should mark the end of balancing.

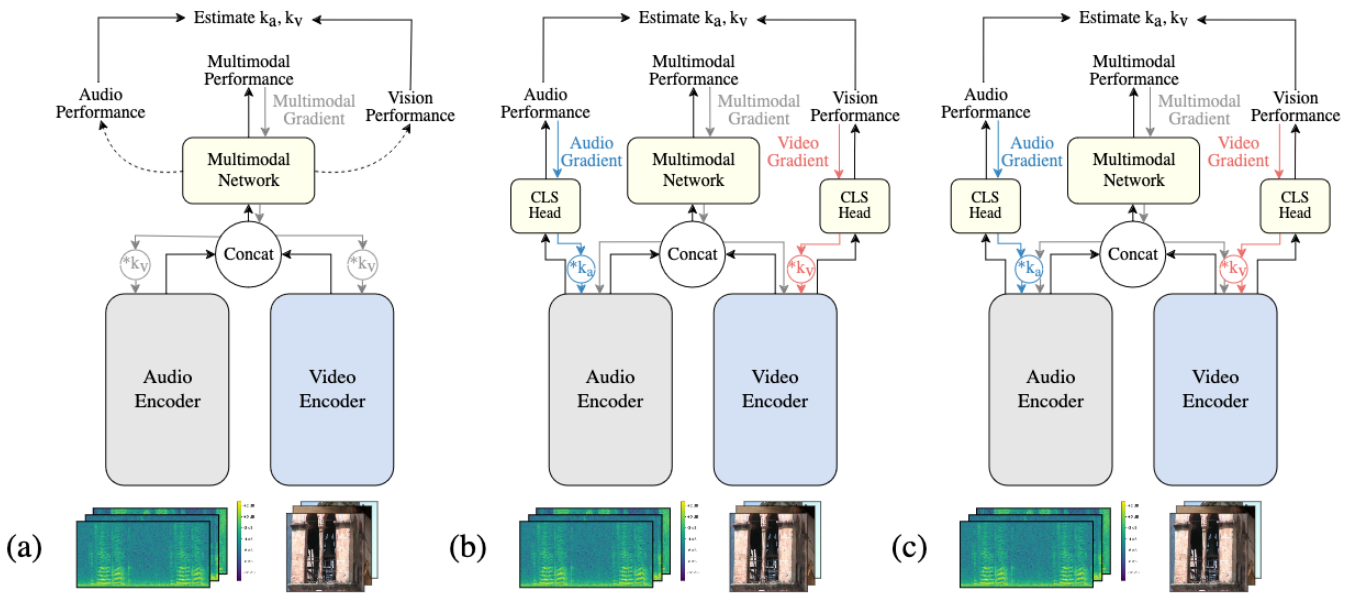



Figure 4.1: Categorization of state-of-the-art balancing methodology: (a) Gradient Balancing methods use estimates of unimodal performance to calculate coefficients ($k_a$ and $k_v$) and employ these to balance the multimodal loss. (b) Multi-Task methods incorporate unimodal classifiers into the model, each noted as CLS Head, to better estimate unimodal performance. The coefficients ($k_a$ and $k_v$) derived from comparing unimodal performance are used exclusively to balance the unimodal losses. (c) The proposed Multi-Loss Balancing method combines both strategies by incorporating unimodal classifiers for accurate unimodal performance estimation and balancing both multimodal and unimodal losses.

## 4.2 Multi-Loss Balancing Gradients

### 4.2.1 Model Structure

We have endeavoured to maintain our methodology as model-agnostic as possible, albeit with a few necessary assumptions. Our focus centres on multimodal models, wherein each modality is processed by a dedicated encoder. We aim to enhance the training efficiency of these encoders, thereby indirectly benefiting the entire network's performance. To elucidate, our models conform to a unified structural framework.

Given a dataset $D$ consisting of $N$ samples across $M$ modalities, denoted as $X = \{X_1, .., X_M\}$, each sharing a common ground truth label $Y$ with $C$ distinct classes, we employ networks structured as follows:

$$f(X;\theta) = f_v(f_1(X_1;\theta_1), .., f_M(X_M;\theta_M);\theta_v), \tag{4.1}$$

where $\theta$ denotes all the parameters of the network, $\theta_1, .., \theta_M$ the unimodal encoders', $f_1, .., f_M$ parameters and $\theta_v$ the common parameters of the common fusion function $f_v$.

### 4.2.2 Balancing

Our objective is to achieve synchronous convergence of the unimodal encoders, based on the hypothesis that this will prevent the model from overfitting to any single modality and may encourage the development of more synergistic behaviors.

To achieve this, following prior studies [43, 187, 129, 202], we dynamically adjust the learning rates of each unimodal encoder based on the comparative analysis of unimodal performances. We estimate balancing coefficients for each modality $i = 1, .., M$ as $k_i$ that indicate how much each modality should change the learning pace of its encoder. When $k_i > 1$, we accelerate the learning of the modality, while when $k_i < 1$, we decelerate it. The update rule using stochastic gradient descent (SGD), the initial learning rate $lr_{\text{base}}$ and the loss function $L$ would then be:

$$\Delta\theta_i = -lr_{\text{base}} \cdot k_i \cdot \nabla_{\theta_i} L(X, y, \theta). \tag{4.2}$$

We improve upon previous works [129] in three points. Firstly, we incorporate a multi-task objective with additional unimodal losses. Such an objective ensures a precise assessment of unimodal performance while aiding the convergence of the unimodal classifiers. Secondly, our method allows for both the acceleration and deceleration of learning across the modalities, without ever entirely halting their progress. Finally, while previous methods imposed a hard limit to cease balancing after a predetermined number of epochs, we employ a function that naturally reduces balancing effects as the performances of both modalities converge.

Our multi-loss objective $L$ is a summation of a cross-entropy ($CE$) loss for the multimodal predictions and similar cross-entropy losses for each unimodal prediction. This can be expressed as:

$$L = CE(f(X;\theta), y) + \sum_{i=1}^{M} CE(f_i(X_i;\theta_i), y). \tag{4.3}$$

To adaptively balance the learning of each unimodal encoder, we derive the coefficients $k_i$ of Eq. 4.2 based on modality performance. These coefficients are calculated as follows:

$$s_i = \sum_{j=1}^{N} \sum_{c=1}^{C} f_i(X_j;\theta_i) 1_{k=y_{c_j}}{}^{1}, \tag{4.4}$$

$$r_i = \frac{\frac{1}{M-1} \sum_{j=1, j\neq i}^{M} s_j}{s_i}, \tag{4.5}$$

$$\beta_i = \begin{cases} \beta_{max} & \text{if } r_i > 1, \\ 2 & \text{otherwise,} \end{cases} \tag{4.6}$$

$$k_i = 1 + (\beta_i - 1) \cdot \tanh(\alpha \cdot (r_i - 1)) \tag{4.7}$$

$s_i$ is the sum of correct class predictions by the unimodal encoder for modality $i$, $r_i$ the relative performance of each encoder compared to the average performance of the others, $\beta_i$ the max value of the coefficients and $k_i$ the final balancing coefficient for each modality, where $\alpha \in \mathbb{R}^+$ and $\beta_{max} \in \mathbb{R}^+$, $\beta_{max} \geq 1$ are predetermined hyperparameters. Finally, *tanh* is the hyperbolic tangent function, providing a smooth normalization of $r_i$.

The utilization of two distinct cases for $\beta_i$ arises from the varied behaviours we aim to achieve with the balancing method. When $r_i < 1$, it indicates that

[1] $1_{k=y_{c_j}}$: An indicator function that equals 1 if $k = y_{c_j}$, and 0 otherwise.

other modalities perform, on average, worse than modality $i$, necessitating a reduction in the learning rate of modality $i$ to slow its learning pace. To achieve this, we set $\beta_i = 2$, placing us in the portion of Figure 4.2 where $r_i \in [0, 1]$. The extent of the learning deceleration can vary and is influenced by the slope of the *tanh* function parameterized by $a$; a larger $a$ value results in more pronounced changes in the learning rate. Conversely, when $r_i > 1$, indicating that other modalities generally outperform modality $i$, we aim to increase the learning rate of modality $i$, requiring $k_i > 1$. The extent of the acceleration is more sensitive, as large values may lead to model divergence. To control this, we introduce the hyperparameter $\beta_{max}$, which sets the maximum permissible increase. For instance, if $\beta_{max} = 10$, we allow the learning rate to increase by up to 10 times based on the performance ratio $r_i$, as shown in Figure 4.2 when $r_i > 1$.

Extending the estimation of the balancing ratios allows for the consideration of different tasks without relying on softmax cross-entropy loss. To achieve this the value of $s_i$ is computed as the summation of the loss across all batch datapoints, and the ratio $r_i$ is estimated by inverting the original equation. For example, in the case of regression with Mean Squared Error (MSE) loss, the equations would be adapted as follows:

$$s_i = \sum_{j=1}^{N} (f_i(X_j; \theta_i) - y_{c_j})^2, \tag{4.8}$$

$$r_i = \frac{s_i}{\frac{1}{M-1} \sum_{j=1, j \neq i}^{M} s_j} \tag{4.9}$$

## 4.3 Experiments

### 4.3.1 Datasets

**CREMA-D [25]:** is an emotion recognition dataset with audio and video modalities. It features a diverse group of 91 actors, covering a wide range of ages, ethnicities, and genders. To ensure consistency, each actor is positioned at an equal distance from the camera, expressing six distinct emotions: Happy, Sad, Anger, Fear, Disgust, and Neutral. In alignment with methodologies from prior studies [129, 96, 43], video frames are sampled at 1 fps, selecting 3 consecutive frames, while audio segments are sampled at 22 kHz, capturing 3 seconds that correspond with the chosen video frames. Audio analysis utilizes a window size of 512 and a step size of 353 samples for Short-Time Fourier Transform (STFT),

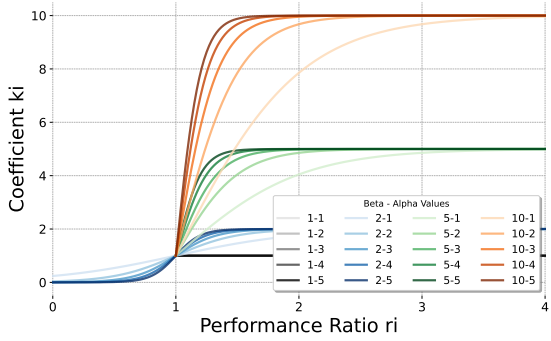



Figure 4.2: Comparing Balancing Coefficients ($k_i$, Eq. 4.7) and Performance Ratios ($r_i$, Eq. 4.5) across different $\alpha$ and $\beta$ settings.

creating log-Mel spectrograms. We create a second preprocessing version of this dataset aimed at the advanced pretrained models. In that version, our methodology aligns with Goncalves et al. [57], incorporating audio signals sampled at 16 kHz and utilizing pre-calculated facial features. We extract these facial features by first identifying bounding boxes in each video frame, then applying a pretrained EfficientNet-B2 [171] model for feature extraction. Unlike previous approaches [129, 96, 43], our dataset division follows Goncalves et al. [57], excluding actor overlap between training, validation, and test sets. We report the standard error (SE) across folds for consistency.

**AVE [173]:** contains 4143 videos across 28 event categories with a wide range such as frying food or playing the guitar, each with temporally labelled audio-visual events of at least 2 seconds. Following [43], video segments with event occurrences are sampled at 1 fps for 4 frames, with audio resampled at 16 kHz using CREMA-D's STFT settings. AVE provides predefined training, validation, and test splits. Our SE is derived from three random seeds on the same test set.

**UCF [161]:** features real-life action videos from YouTube in 101 action categories, expanding on UCF50. We select samples that include both video and audio modalities, narrowing our focus to 51 action categories. Data preparation mirrors that of the AVE dataset. For model evaluation, we utilize the 3-fold split offered [161], reporting the se across these folds.

**CMU-MOSEI [207]:** is a multimodal dataset designed for sentiment and emotion analysis. It comprises over 23k annotated video clips from various online platforms, featuring speakers expressing a range of opinions and emotions. The dataset includes three modalities: audio, video, and text (transcriptions).

Sentiment labels in CMU-MOSEI range from -3 (most negative) to 3 (most positive), framing the task as a regression problem. For our experiments, the predefined training, validation, and test sets are used, and SE is reported across 3 random seeds.

### 4.3.2 Models

In these experiments, we aim to show that our findings generalize across diverse unimodal encoders and multimodal fusion techniques, regardless of architecture or inherent data compatibility. This section details the principles guiding our choice of encoders and fusion methods for each dataset, ranging from baseline model configurations to the incorporation of larger, pre-trained models. We further explore various fusion strategies, from basic late linear fusion to more sophisticated techniques like fusion gates and Transformer-based approaches.

#### Models: Backbone Unimodal Encoders

In line with previous research [96, 129, 43, 201], our initial experiments adopt ResNet-18 [68] as the unimodal encoder for handling both video and audio modalities in the CREMA-D, AVE, and UCF datasets. These models are randomly initialized and incorporate adaptive pooling to accommodate diverse input dimensions. For the CMU-MOSEI dataset, we exploit a 5-layer Transformer [180] similar to previous works [102, 103].

We extend our investigation to include a larger, pre-trained set of unimodal encoders that are optimally suited for each specific modality. This selection process is designed to rigorously assess whether state-of-the-art models exhibit susceptibility to the same phenomena under investigation. We exploit these encoders on the CREMA-D dataset. Following [56] on CREMA-D, we deploy the first 12 layers of the Wav2Vec2 [12, 194] model with self-supervised pretrained weights for speech recognition, allowing the Wav2Vec2 model to be finetuned. For the video modality we extract the facing bounding boxes with the multi-task cascaded convolutional neural network (MTCNN) face detection algorithm [209] and afterwards the facial features of every available frame exploiting EfficientNet-B2 [171] as a frozen feature descriptor. The extracted audio features and pre-calculated facial features are further refined using a 5-layer Conformer [62], initialized from scratch. Although the model includes additional components, we refer to it as 'Conformer' for simplicity.

### Models: Multimodal Fusion

We extend our exploration by evaluating five distinct fusion methods that combine unimodal features from the encoders to generate the multimodal output. The simplest method, **Late-Linear**, applies a linear layer to the concatenated unimodal features. For **Mid-MLP**, we use a 2-layer Multilayer Perceptron (MLP) to assess how nonlinear feature combinations influence modality imbalances. Building on prior research, we experiment with more advanced strategies: **FiLM** (Feature-wise Linear Modulation) [130] and **Gated** mechanisms [85]. Additionally, we introduce a Transformer-based fusion method (**TF**) using a 2-layer Conformer [62], which incorporates both class and modality-independent tokens. This investigation aims to demonstrate the effectiveness of the proposed balancing methods across a wide array of fusion strategies.

### Previous Balancing Methods

A comprehensive overview of previous methods can be found in Chapter 3.3. Here, we briefly highlight the key approaches that are central to our experiments:

**Unimodal Training**: Each modality is trained independently without interaction between them.

**Joint Training**: All modalities are trained together under a single-loss objective, with no specific balancing mechanisms in place.

**Multi-Loss** [181]: Adds separate unimodal task losses in addition to the joint multimodal objective.

**MMCosine** [201]: Standardizes feature vectors and modality-specific weights to balance modality contributions.

**MSLR** [202]: Adjusts the learning rates of unimodal encoders dynamically, based on their validation performance.

**PMR** [43]: Similar to Multi-Loss but uses prototype classifiers and adjusts the weights of the unimodal losses.

**OGM** [129]: Extracts unimodal performance from the last layer, compares improvements, and modulates each encoder's learning rate.

**AGM** [96]: Utilizes zero-masking Shapley values to estimate the importance of each modality, modulating the task loss weights based on these values.

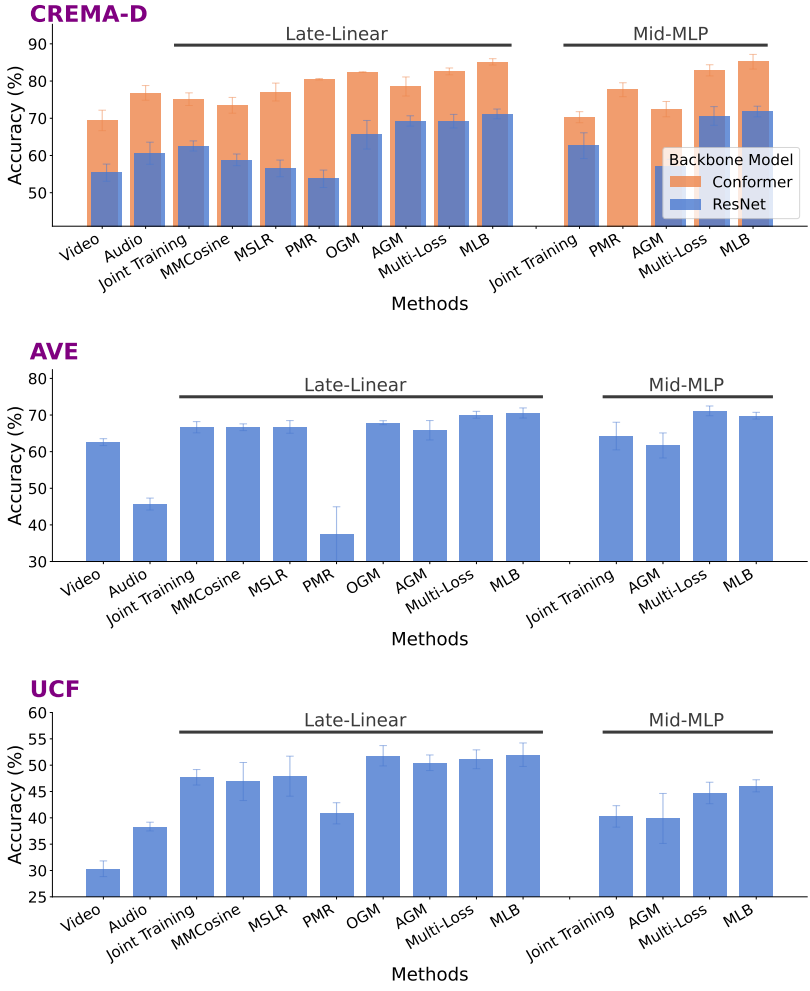



Figure 4.3: Accuracy of models that differentiate by the backbone encoders (colours), the fusion strategies (Late with a linear and Mid with an MLP classifier), and the balancing techniques (x-axis). Across all datasets, results demonstrate that employing unimodal losses within the Multi-Loss framework and balancing them consistently yields the best performance.

### 4.3.3 Experimental Results

In this section, we present results on the effectiveness of using different fusion strategies on multimodal training with several multimodal balancing techniques namely MMCosine, MSLR, OGM, AGM, PMR[2], Multi-Loss and our Multi-Loss Balancing (**MLB**). Hyperparameter tuning was conducted separately for each method, dataset, backbone encoder, and fusion strategy on the validation set. This ensured fairness with an equal number of trials for each configuration. The hyperparameters are available in Appendix A.2.1.

In the bar plots presented in Figure 4.3, we assess the previously mentioned models alongside suitable balancing methods across the three (video-audio) datasets. Results highlight that MLB consistently outperforms previous approaches, demonstrating its robustness across various settings. Our comparison with the simpler Multi-Loss strategy reveals that adding unimodal losses often suffices to adequately train the unimodal encoders. In most instances, balancing provides a further incremental benefit.

Furthermore, we use the Expected Calibration Error (ECE) [64] to assess whether multimodal training enhances model uncertainty awareness. The definition of the metric can be found in Appendix A.1. Figure 4.4 demonstrates that the MLB method, achieving the highest accuracy, also exhibits lower calibration error compared to other methods and to Multi-loss, highlighting the significance of balancing.

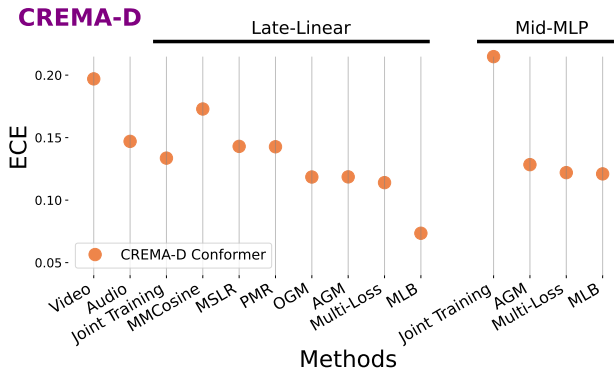



Figure 4.4: ECE Comparison on Conformer CREMA-D.

[2]The PMR method, as per the provided code, exhibits several instabilities, resulting in exploding values of the classifier prototypes. This instability predominantly affects models with ResNet encoders and certain nonlinear fusion methods. Despite these issues, we opt not to modify their implementation to avoid inadvertently influencing the method incorrectly.

To verify that the issue of overfitting on one of the modalities is not prevalent in the fusion method, we incorporate alternative fusion strategies, as mentioned in Section 4.3.2, FiLM, Gated and TF. Both the Multi-Loss and Multi-Loss Balancing employ additional linear classifiers to derive unimodal predictions and calculate the unimodal losses. Figure 4.5 supports our primary findings, indicating improvements in the learning process across models employing different fusion methods balanced with Multi-Loss and MLB. However, in some instances, we do not observe notable benefits from the addition of the balancing technique.

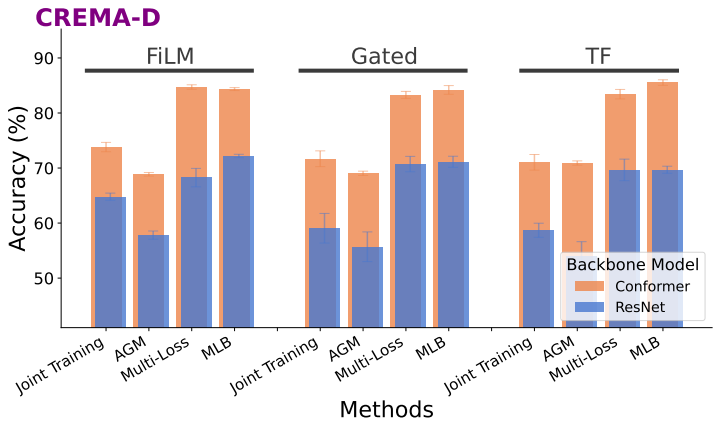



Figure 4.5: Accuracy of models using FiLM [130], Gated [85] and TF fusion techniques on the ResNet and Conformer model applied to the CREMA-D dataset.

To evaluate the capabilities of MLB on datasets with more than two modalities, the CMU-MOSEI dataset was employed, which includes three modalities: audio, video, and text. Pre-calculated features for each modality were used, leading to encoders of varying sizes. Although this approach may limit the results' bandwidth, Table 4.1 demonstrates that MLB performs effectively across multiple modalities. The sentiment prediction task, trained as a regression problem with continuous target values is further evaluated with accuracy, considering positive versus negative sentiment binary classification, following previous works. Notably, MLB, which has shown promising results in classification tasks on previous datasets, continues to perform well in this regression context. Since the task is regression the necessary modifications to the ratio estimation equations are provided below.

Table 4.1: Binary classification accuracy of unimodal models and different multimodal balancing techniques on CMU-MOSEI dataset.

| Unimodals | Video: $65.3_{\pm 0.2}$ | Audio: $63.5_{\pm 0.5}$ | Text: $80.2_{\pm 0.2}$ |
|---|---|---|---|
| Method | Late-Linear | MLP | TF |
| Joint Training | $80.7_{\pm 0.2}$ | $78.8_{\pm 0.1}$ | $80.5_{\pm 0.5}$ |
| Multi-Loss | $80.9_{\pm 0.2}$ | $80.1_{\pm 0.2}$ | $80.3_{\pm 0.1}$ |
| AGM [96] | $\mathbf{81.0}_{\pm \mathbf{0.1}}$ | $80.1_{\pm 0.2}$ | $80.1_{\pm 0.4}$ |
| MLB | $80.9_{\pm 0.1}$ | $\mathbf{80.7}_{\pm \mathbf{0.2}}$ | $\mathbf{81.0}_{\pm \mathbf{0.2}}$ |

## 4.4 Coefficients Analysis

Several methods, including MLB, estimate coefficients for each modality that are exploited to balance multimodal learning. In Figure 4.6, we compare the values of the $k_i$ coefficients for the different methods to illustrate the effect each one caused. This visualization demonstrates that MLB is the only one that, upon convergence, its coefficients are equalizing and approach value 1 phasing out the balancing when it's no longer necessary.

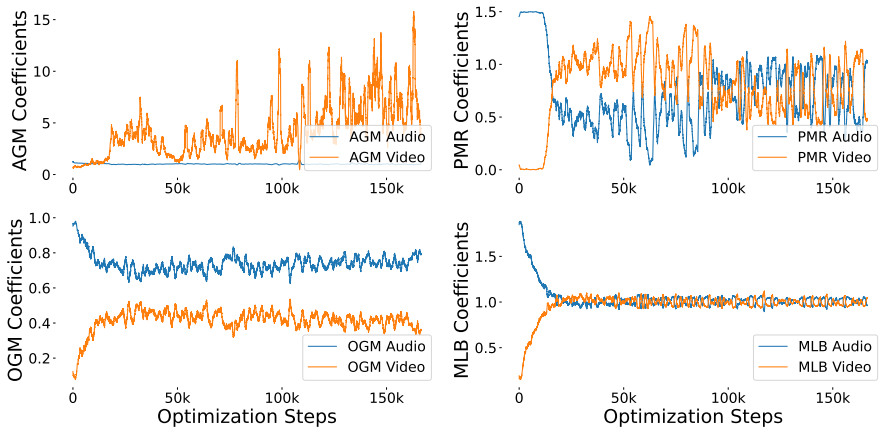



Figure 4.6: Comparison of balancing coefficients evolution of coefficients that amplify the contribution of each unimodal encoder's learning, smoothed over a 1k-point kernel. It illustrates that while all methods provide some balancing, only Multi-Loss Balancing (MLB) achieves self-muting when modalities have converged.

## 4.5 Computational Complexity Analysis

The computational complexity of the proposed balancing method, MLB, introduces negligible additional overhead compared to both training without any balancing and existing multimodal learning approaches. To quantify this, we analyze the complexity associated with the forward and backward passes of the multimodal encoders $f_1, f_2, \ldots, f_M$, with computational costs denoted by $C_{f_i}$ and $C_{b_i}$, respectively, alongside the common network $f_v$ with complexity $C_v$.

Methods classified under gradient balancing techniques typically do not introduce significant additional computation during network passes, rendering the balancing computations negligible. Approaches in the second category, which incorporate additional classification heads, do introduce some computational overhead. However, these heads usually consist of linear layers or prototype networks, whose computational costs are minimal compared to the overall network.

To illustrate this point, consider a ResNet-18 encoder used in our experiments, which requires approximately $1.8 \times N$ GFlops per forward pass for a batch size of N on the video modality. In contrast, a linear classifier mapping from a 512-dimensional feature space to $C$ classes involves only $512 \times N \times C$ flops, on the order of thousands of flops, rendering its cost negligible in comparison to the rest of the network.

Notable exceptions include methods like PMR, which necessitate prototype estimation at the end of each epoch, and the approach by Wang et al. [187], which computes balancing coefficients on a separate validation set each epoch, introducing additional overhead. The AGM method represents the most computationally expensive approach, as it requires $2^M - 1$ forward passes, resulting in a training complexity of $O((2^M - 1) \times (C_{f_i}) + C_{b_i})$.

In contrast, MLB adds only the computational cost of the linear classifiers, the estimation of additional losses, and the computation of balancing coefficients. These operations require only thousands of flops and thus are negligible relative to the overall model size. Consequently, in practical terms, these computations do not result in any noticeable increase in training time. Similarly, the memory overhead introduced by MLB remains minimal, comparable to the most efficient methods reviewed.

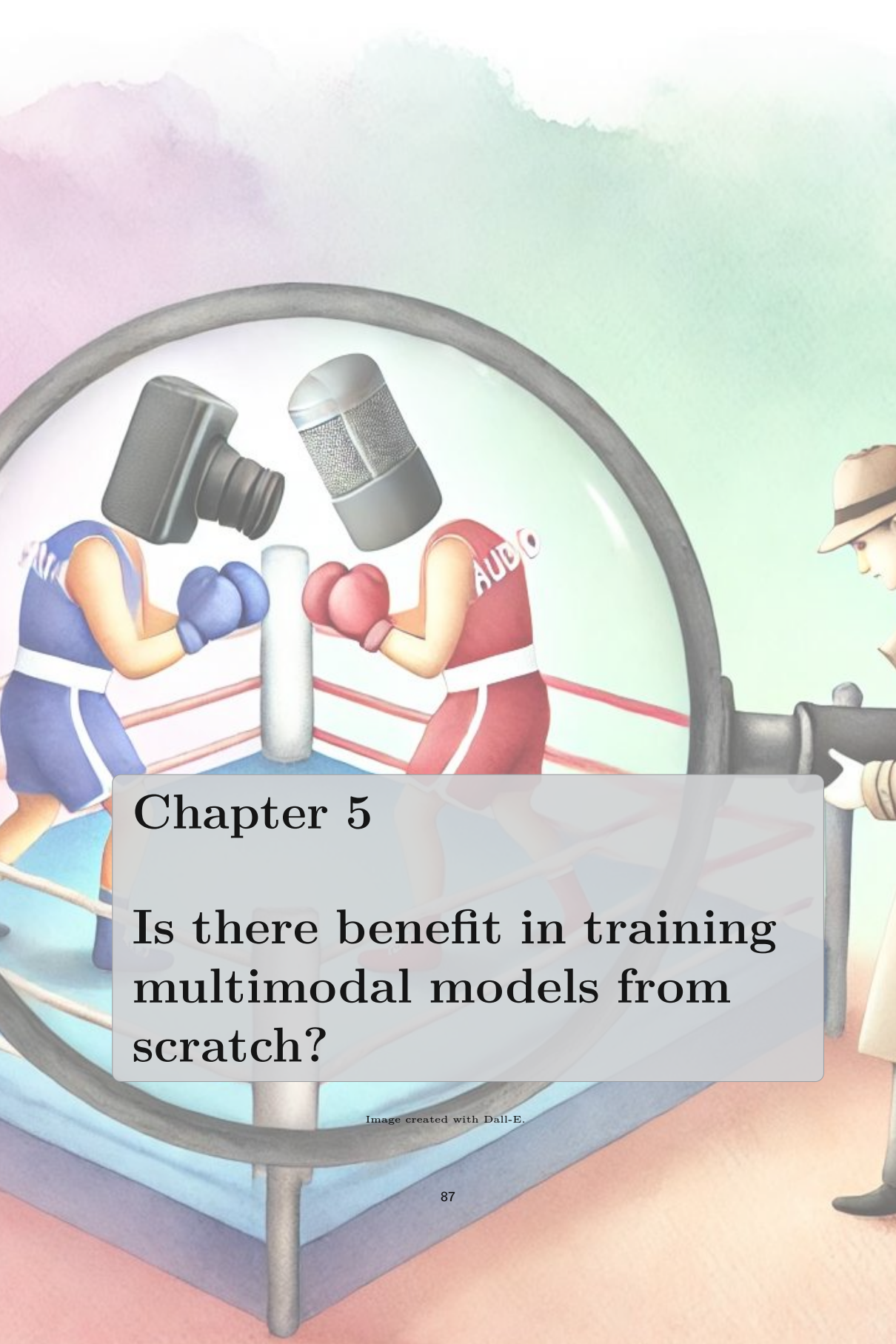
## 4.6 Discussion and Conclusion

In this chapter, we introduced a new method to effectively address the challenges of balancing multimodal training. We derived two key contributions by focusing on datasets that reveal generalization issues in multimodal learning. First, we demonstrated that existing methods often yield inconsistent performance improvements across diverse models and datasets. We refer readers to Appendix A.3, where results are presented in table format, and such inconsistencies are clear. This inconsistency underscores the gap for a balancing method for multimodal training that yields persistent improvements throughout different datasets, modalities, encoders and fusion methods.

Second, we propose the Multi-Loss Balancing (MLB) method, a refined framework designed to address these shortcomings. MLB integrates additional unimodal losses to enhance the training of individual encoders while introducing a dynamic balancing scheme that adjusts learning rates for each encoder. This scheme not only accelerates or decelerates their learning pace as required but also mutes the balancing interventions as the model approaches convergence, ensuring smoother integration into existing pipelines.

Our approach consistently demonstrated significant performance improvements through experiments on four diverse datasets. These results were achieved using both ResNet and Transformer-based backbones paired with a variety of fusion strategies, showcasing MLB's versatility and robustness across different multimodal settings.

This work represents a step forward in advancing multimodal training methodologies; however, it raises the question of to what extent the problem has truly been solved. The following chapter will explore these questions, reflecting on the broader implications of our contributions and assessing the progress made in bridging the gap between multimodal and unimodal training.



# Chapter 5

# Is there benefit in training multimodal models from scratch?

Image created with Dall-E.

## Chapters Contribution

Chapter 5 investigates the benefits and limitations of training multimodal models from scratch compared to leveraging unimodally pretrained baselines. The chapter evaluates the effectiveness of various balancing methods and demonstrates that training from scratch can unlock greater synergy between modalities, enabling improved performance in multimodal tasks. While unimodal pretraining provides a strong foundation, this work highlights the potential of multimodal training to achieve enhanced integration and collaboration across modalities. This work suggests a structured evaluation framework to better assess the trade-offs and potential of joint multimodal training strategies.

## 5.1 Introduction

In the previous chapters, we thoroughly reviewed existing multimodal balancing methods, evaluated their performance through extensive experiments, and proposed several improvements within our method. We established that these approaches mark a significant advancement over simple joint objective training. However, a natural question arises: can we conclusively argue that these methods are the optimal path forward, or is there still merit in adhering to unimodal training, which has been extensively refined and proven effective? While direct comparisons with individual modalities have shown the superiority of balancing approaches, we still question their superiority over methods that leverage the well-established unimodal training techniques and integrate them to combine the strengths of each modality effectively.

This chapter has two primary objectives. The first is to establish a rigorous experimental framework for evaluating multimodal balancing approaches, facilitating meaningful comparisons with both multimodal and unimodal alternatives and advancing the field through systematic assessment. The second objective is to investigate the benefits of individual balancing methods in greater depth, specifically examining whether observed improvements in accuracy stem from the emergence of new synergetic cross-modal interactions or from improved data routing that prioritizes the modality with the most relevant information. Although this chapter is procedural and does not introduce novel methodologies, it makes a significant contribution by providing a well-structured experimental protocol for the evaluation of future methods and by offering a detailed analysis of the mechanisms underlying the observed improvements.

Our results indicate that many of the previously proposed methods frequently struggle to outperform simpler baselines that leverage unimodal training strategies, such as ensembles or joint training models utilizing unimodally pretrained encoders. These findings are validated across multiple tasks, models, and datasets. Furthermore, by analyzing the distribution of common errors between multimodal approaches and unimodal models, we demonstrate that while some of the previously proposed methods, including our own, enhance multimodal synergy, they fall short in addressing data points correctly predicted by individual unimodal models. This observation underscores the limitations of current approaches and highlights that we remain far from achieving an optimal training strategy for multimodal fusion.

## 5.2 Additional Methodology

This chapter introduces supplementary methodologies that leverage unimodal training as a foundation for multimodal models. By establishing these pretrained configurations, we can evaluate whether multimodal training itself provides additional benefits beyond simply aggregating modality-specific knowledge. These setups, widely used in multimodal studies, provide efficient training pathways and often enhance robustness, interpretability, and transferability.

### 5.2.1 Unimodal Pretrained Baselines

To create a rigorous baseline for multimodal performance, we incorporate pretrained unimodal encoder weights into multimodal models. These pretrained baselines serve as benchmarks, offering stronger baselines to compare the effectiveness of complex multimodal balancing strategies. Specifically, we explore three setups, each with unique contributions to our understanding of multimodal performance:

1. **Ensemble**: Each unimodal model generates predictions independently, which are then combined, without additional training, to form the multimodal prediction. This method is widely adopted as an ensemble approach in multimodal applications [15, 163], where its simplicity and interpretability make it a strong baseline. Ensembles can reveal the extent to which modality-specific models capture task-relevant information individually and how their collective performance compares with joint training methods. It establishes a lower bound for assessing the added value of combined multimodal training.

2. **Pretrained Encoders (Frozen)**: Here, unimodal encoders are initialized with weights from unimodal training and kept frozen during multimodal training, allowing only the fusion layer to be learned. This setup called **Uni-Pre Frozen**, leverages previous findings suggesting that modality-specific knowledge, when maintained without further modification, can offer stable features for fusion, reducing overfitting risks [175]. By observing performance with frozen weights, we assess whether modality-specific representations alone can effectively support multimodal fusion.

3. **Pretrained Encoders (Finetuned)**: Each encoder is initialized with pretrained weights from unimodal training but is further optimized during multimodal training, referred to as **Uni-Pre Finetuned**. Finetuning allows the model to align modality-specific representations with the task's

multimodal requirements, potentially uncovering synergies that arise only when modalities are trained jointly [143, 73]. This configuration can enhance the model's ability to leverage complementary information from each modality, adapting the representations as needed to capture inter-modality correlations.

By examining these setups, we establish a foundation for future multimodal benchmarks, highlighting the strengths and limitations of each approach. These pretrained configurations offer a straightforward yet effective means of evaluating new multimodal methods against simpler baselines, ensuring that future innovations in multimodal training contribute meaningfully to the field.

## 5.3 Experimental Setting

### 5.3.1 Datasets

In this section, we describe the datasets used for evaluating the performance of the balancing methods in multimodal learning. These datasets span a range of tasks, including emotion recognition, action recognition, and audiovisual event classification, providing a comprehensive basis for assessing the effectiveness of the proposed techniques. We include a short description of each dataset and we refer to Section 4.3.1 for more information.

CREMA-D [25]: An emotion recognition dataset featuring 91 actors expressing 6 distinct emotions.

AVE [173]: A collection of videos with temporally aligned audio-visual events across 28 categories.

UCF [161]: An action recognition dataset consisting of real-life YouTube videos.

### 5.3.2 Balancing Methods

We provide a concise overview of the balancing methods employed in the upcoming experiments, most designed to enhance multimodal learning. For a detailed explanation and in-depth analysis of these techniques, please refer to Section 3.3. Below, we summarize the key approaches, laying the groundwork for the evaluations that follow.

**Unimodal Training**: Each modality is trained independently without interaction between them.

**Joint Training**: All modalities are trained together under a single-loss objective, with no specific balancing mechanisms in place.

**Multi-Loss** [181]: Adds separate unimodal task losses in addition to the joint multimodal objective.

**MMCosine** [201]: Standardizes feature vectors and modality-specific weights to balance modality contributions.

**MSLR** [202]: Adjusts the learning rates of unimodal encoders dynamically, based on their validation performance.

**PMR** [43]: Similar to Multi-Loss but uses prototype classifiers and adjusts the weights of the unimodal losses.

**OGM** [129]: Extracts unimodal performance from the last layer, compares improvements, and adjusts gradients by modulating each encoder's learning rate.

**AGM** [96]: Utilizes zero-masking Shapley values to estimate the importance of each modality, modulating the task loss weights based on these values.

**MLB** [89]: Combines Multi-Loss and gradient modulation techniques, adjusting the learning rates of encoders to improve unimodal performance and balance the contributions.

### 5.3.3 Models

The following models are employed to assess our balancing methods across different datasets and architectures. For full details on each model, refer to Section 4.3.2. Each setup explores various encoders and weight initializations to make sure that we cover several cases.

**ResNet** [68]: A baseline encoder for video and audio, used on CREMA-D, AVE, and UCF datasets with random weight initialization and adaptive pooling to accommodate varying input dimensions.

**Conformer (CREMA-D)**: Combines Wav2Vec [12] for audio, MMTC [209] for facial bounding boxed and EfficientNet-B2 [171] for facial features, with a 5-layer Conformer [62] refining the features from each modality.

**Conformer (AVE)**: Follows the CREMA-D model structure but uses HuBERT [70] pretrained on AudioSet [54] for the audio modality and ViViT [8] pretrained on Kinetics for video, with each modality refined as well by a 5-layer Conformer. Despite using different SSL-pretrained models, the architecture is still referred to as 'Conformer' for consistency.

For multimodal fusion, we focus on late-fusion, applying a linear layer to combine encoder outputs, thus allowing us to assess the impact of different balancing methods within a consistent fusion strategy. This streamlined approach is expected to yield insights that are applicable to other fusion methods as well.

## 5.4 Results

### 5.4.1 Effectiveness of Balancing methods

This section presents a comparative analysis of different balancing methods across three datasets to evaluate their impact on multimodal learning. The methods analyzed include Joint Training, MMCosine, MSLR, OGM, AGM, PMR, Multi-Loss, and MLB. Our goal is to understand the advantages of training multimodal models from scratch instead of leveraging pretrained unimodal models, including Ensemble, Uni-Pre Frozen, and Uni-Pre Finetuned.

In the bar plots presented in Figure 5.1, we assess the previously mentioned models alongside suitable balancing methods and baselines across three datasets: CREMA-D, AVE and UCF. The height of each bar indicates the accuracy difference achieved by the multimodal model compared to the Ensemble baseline, illustrating the accuracy attainable without any joint multimodal training. Drawing from these results, we make the following observations:

1. The Ensemble models outperform unimodal counterparts significantly, with margins from the best performing unimodal ranging from 6 to 14.4%. The inclusion of Ensemble serves as a critical benchmark which has often been overlooked in previous studies and a first indication of potential multimodal benefit.

2. The Joint Training model, which is trained from scratch with a single multimodal objective, exhibits a noticeable decline in performance across all datasets compared to the Ensemble, providing evidence of the generalization issues arising with multimodal training. In line with what previous works [187] and some of our previous experiments had noticed, in certain instances (CREMA-D Conformer), it even underperforms relative to the best unimodal model.

3. Only a few of the discussed balancing methods, including MLB, reach performance levels near that of the Ensemble, though they remain marginally inferior. This suggests that, despite promising approaches, previous methods struggle to establish effective training regimes for multimodal models.

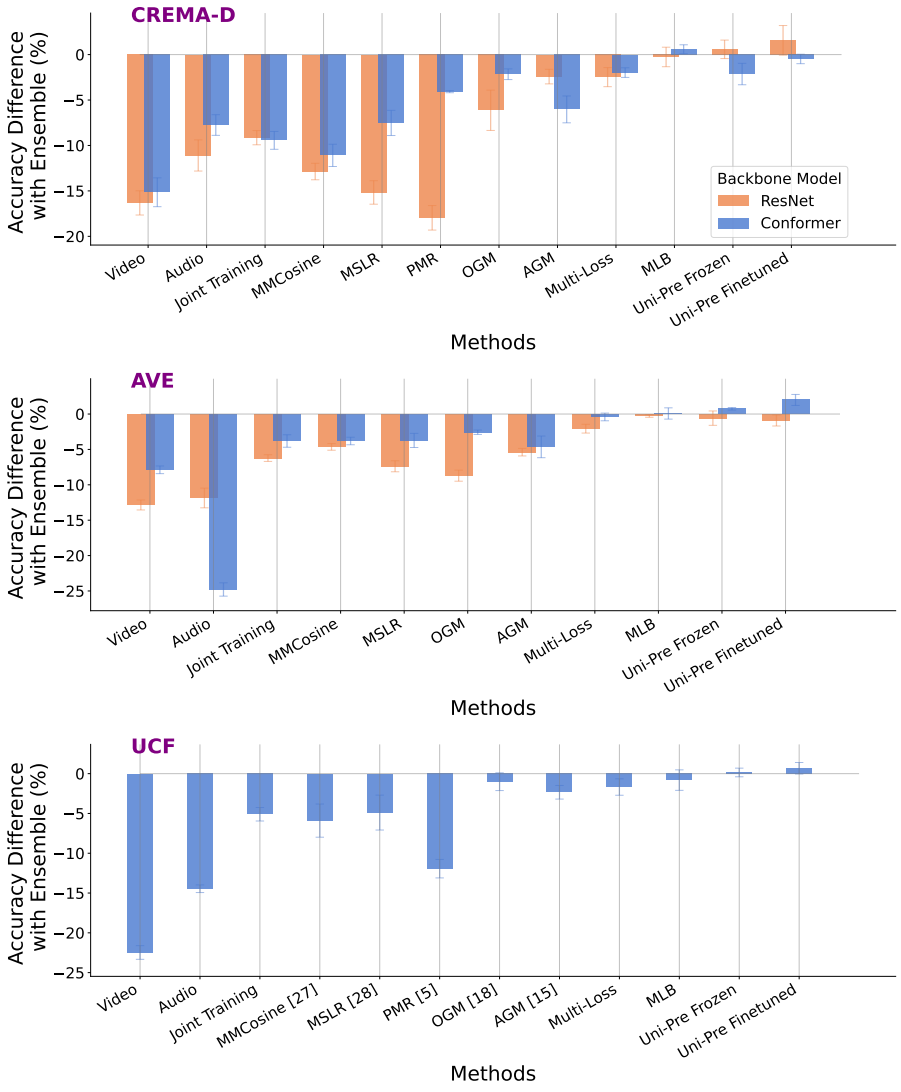



Figure 5.1: Accuracy(%) difference compared to the Ensemble baseline for various methods across three datasets: CREMA-D, AVE, and UCF. Results are shown for two backbone models: Conformer (orange) and ResNet (blue). Positive values indicate an improvement over the Ensemble, while negative values represent a performance drop. Most existing methods fail to outperform the Ensemble, which does not utilize any multimodal learning.

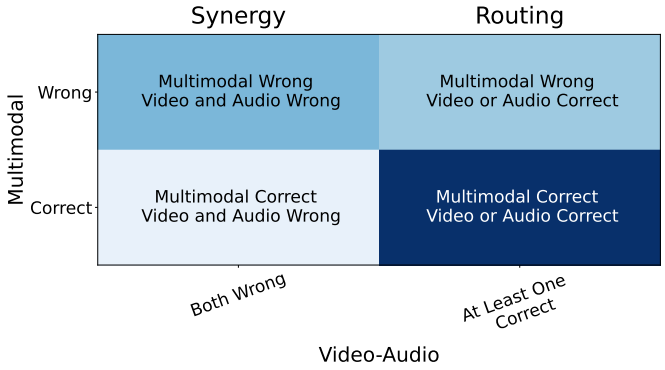



Figure 5.2: Synergy vs. Routing Matrix for Multimodal and Unimodal Predictions. The matrix shows the overlap between correct and wrong predictions, highlighting emergent cross-modal information (Synergy) and the model's ability to leverage the best modality (Routing). Please note that the colours used here are not intended to convey any specific meaning.

4. Among all methods evaluated, only those utilizing unimodally pretrained encoders surpass, in some cases, the Ensemble baseline, albeit by a modest margin ranging from 0.5 to 3%. This highlights that, in some cases, there might be potential advantages of jointly training multiple modalities that would be worth further investigation. We advocate including these two benchmarks (Uni-Pre Frozen and Finetuned), alongside the Ensemble method, as standard comparisons for any newly proposed multimodal approaches.

Throughout this thesis, we have closely analyzed the multimodal performance of various models, occasionally evaluating unimodal feature contributions within the multimodal framework. While several models, including Ensemble, MLB, and Uni-Pre approaches, exhibit strong performance, a deeper understanding of the distinct information each model captures is needed.

### 5.4.2 Post-Hoc Multimodal Analysis

This section investigates the information captured by various training regimes, emphasising understanding the synergistic effects of different approaches on predictive accuracy.

To elucidate the mechanisms by which these regimes leverage data, we analyze the overlap of correct and incorrect predictions across unimodal and multimodal models. Figure 5.2 presents an illustrative matrix that categorizes the prediction accuracy of each approach, highlighting four distinct outcome categories based on model combinations. To ease the analysis, we group scenarios in which at least one unimodal model yields a correct prediction, thus consolidating results to facilitate a more precise interpretation. For each dataset/model, we present three columns showcasing the Ensemble, the highest-performing scratch-trained model, and the model utilizing unimodally pretrained encoders. Analyzing the results of Figure 5.3, leads to several observations:

1. Emergent Information in Ensemble Models: The Ensemble model, which does not involve joint multimodal training, still manages to correctly identify some datapoints that all individual unimodal models misclassify. This suggests that combining modalities, even without training them together, can uncover emergent information not accessible by any single modality alone.

2. Improved Routing with Pretrained Encoders: Across different datasets and models, unimodally pretrained encoders consistently outperform other approaches on datapoints where at least one modality was already predicting correctly. This indicates that the Uni-Pre regimes excel at refining the routing of data points, preserving the information that each unimodal has acquired and leveraging the strengths of individual modalities more effectively. This finding aligns with the overall superior performance of these models.

3. Discovering New Information with Multimodal Training: For datapoints where all unimodal models fail, multimodal models trained from scratch show a distinct advantage over both pretrained encoders and the Ensemble model. The margin of improvement ranges from 0.5% to 1.6% over pretrained models and from 1.2% to 10.2% over the Ensemble. This suggests that while multimodal training may not fully capture all information from each modality, it can reveal new, emergent insights through joint learning.

4. MLB: The method we introduced in Chapter 4, MLB does not outperform AGM in discovering cross-modal emergent information; it performs comparably while also improving the routing of information between modalities. This balance of routing and synergy underscores MLB's value in multimodal learning.

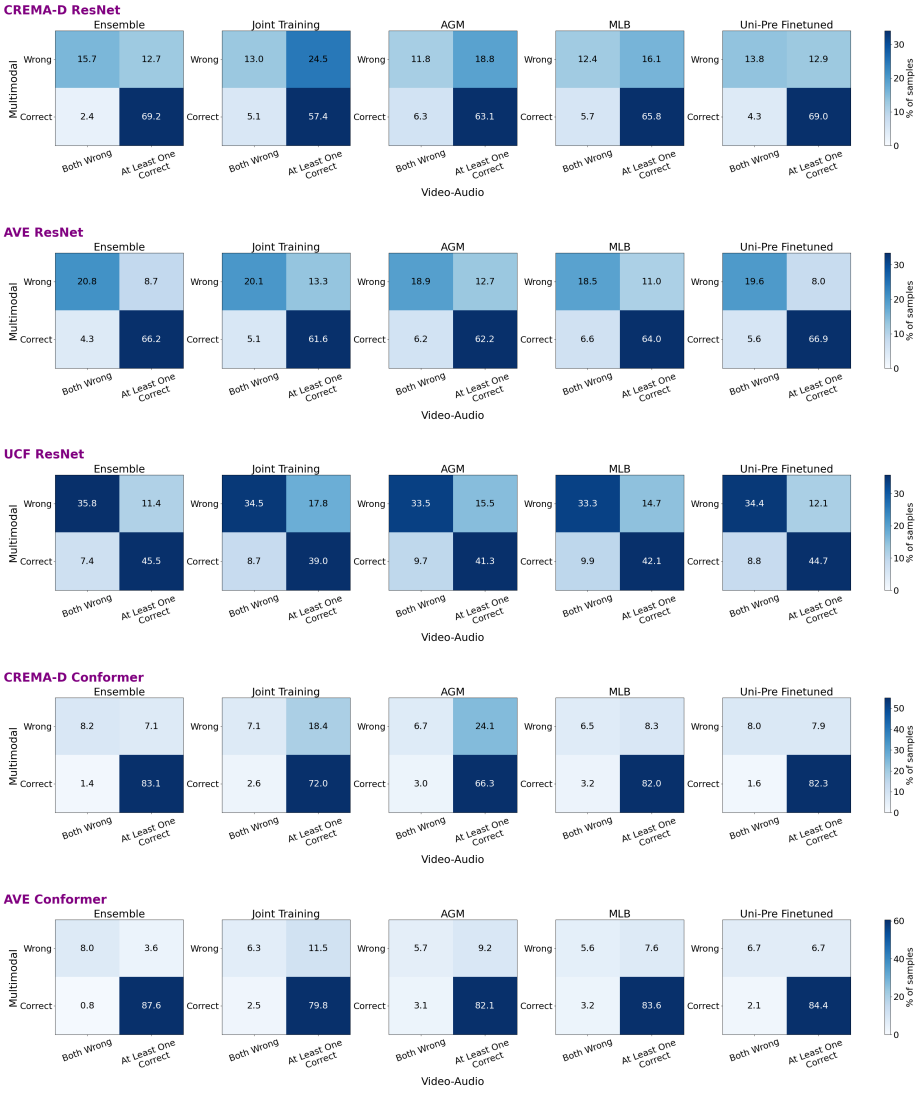


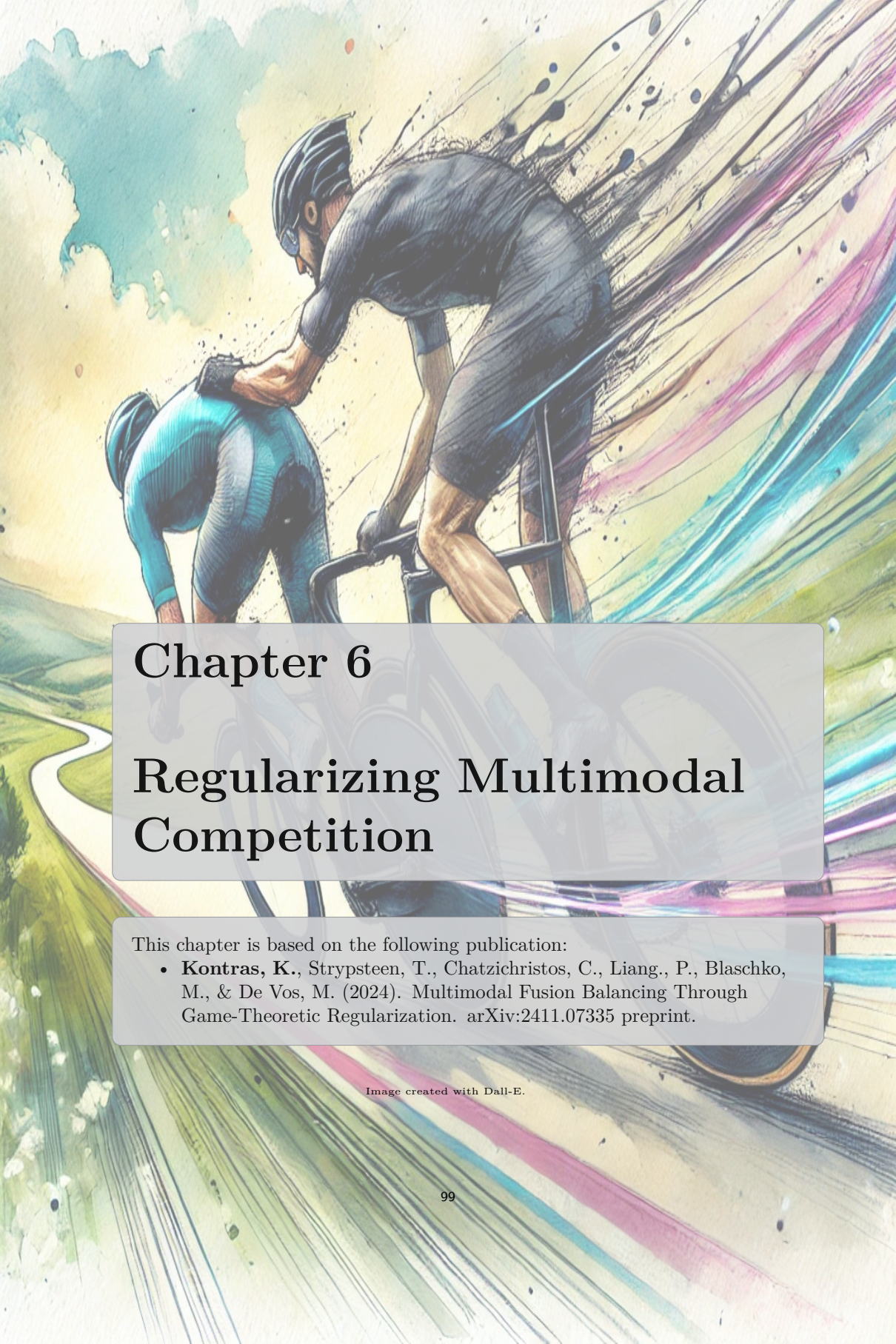

Figure 5.3: Error comparison matrices across datasets and models, contrasting unimodal predictions with multimodal methods (Ensemble, Joint Training, AGM, MLB, and Uni-Pre Finetuned). Columns represent cases where both unimodal predictions fail or at least one succeeds. AGM and MLB excel at identifying emergent information in cases where unimodal models fail, while unimodally pretrained methods, such as Ensemble and Uni-Pre Finetuned, perform better when at least one unimodal prediction is correct.

## 5.5 Discussion

In this chapter, we conducted a comprehensive evaluation of proposed balancing methods for enhancing joint multimodal training, comparing their performance against alternative approaches and strong baselines that utilize pretrained unimodal models. These baselines, whether used as ensembles without further training or with encoders that are frozen or finetuned, present a significant challenge to balancing methods, underscoring the need for further advancements to develop approaches capable of consistently surpassing these alternatives.

We also performed a post hoc analysis of errors to compare the performance of unimodal and multimodal models, leading to two important observations. First, multimodal balancing methods demonstrate improved synergetic behaviour by correctly predicting datapoints that neither unimodal model could handle independently. Second, ensemble-based and pretrained encoder methods excel at preserving correct predictions in cases where at least one unimodal model succeeds. However, even these methods fail to recover a substantial portion of such datapoints.

Using the post hoc error matrices we can estimate a theoretical upper bound up to which an improved multimodal model could achieve. We do so by hypothesizing that if multimodal training could achieve its goal it would at least be able to score correctly all of the routing datapoints, meaning the datapoints that unimodals score correctly, while maintaining the synergy that is currently achieved. Hopefully, synergy would also be improved expanding the bound even more. Estimating the theoretical performance upper bound as the performance that would have been achieved under perfect routing and current synergy, we observe that current methods remain significantly below this ideal performance level. This gap highlights the considerable room for improvement in multimodal fusion and motivates continued research into developing more effective training procedures that better harness the potential of multimodal learning.



# Chapter 6

# Regularizing Multimodal Competition

This chapter is based on the following publication:

- **Kontras, K.**, Strypsteen, T., Chatzichristos, C., Liang., P., Blaschko, M., & De Vos, M. (2024). Multimodal Fusion Balancing Through Game-Theoretic Regularization. arXiv:2411.07335 preprint.

Image created with Dall-E.

## Chapters Contribution

Chapter 6 introduces the Multimodal Competition Regularizer (MCR), a novel framework designed to advance the understanding and regulation of modality contributions in multimodal learning. MCR quantifies unique and shared information contributions from each modality using conditional mutual information and employs game-theoretic strategies to balance interactions between modalities. Extensive experiments demonstrate that MCR not only outperforms existing balancing techniques but also becomes the first method to surpass unimodal learning, setting a new benchmark for managing multimodal competition, and offering a principled and efficient solution for robust multimodal training.

## 6.1 Introduction

In the previous chapter, it was observed that multimodal balancing methods enhance the extraction of synergetic information and that synchronized learning progress among modalities leads to improved cross-modal interactions. However, these methods also exhibit challenges in preserving both the unique and shared information of individual modalities, which ultimately contributes to an increased overall error in the models trained with them compared to methods utilizing unimodally pretrained models.

While significant progress has been made with generalizable representations and larger datasets enabling solutions to complex tasks [90, 98, 117, 141, 142, 143, 175, 184, 205], multimodal models often fall short of their promise, underperforming relative to ensemble unimodal models or the strongest single modality [187, 202]. This outcome challenges the expectation that combining modalities should inherently enhance performance, assuming independent errors and above-chance predictive power [65].

These findings reinforce that current multimodal balancing methods fall short of their theoretical potential, underscoring the importance of developing more effective strategies. Modality competition, where dominant modalities suppress others during training [73], remains a critical obstacle, highlighting the necessity for approaches that dynamically monitor the unique contribution of each modality and balance them to fully leverage multimodal data.

In Chapter 3, several balancing strategies have been discussed that tackle this issue [39, 43, 49, 52, 76, 89, 96, 129, 134, 185, 187, 196, 202]. A central aspect of these methods is estimating each modality's contribution to the output. Most assume distributional independence between modalities on predicting the target, measuring contribution via unimodal performance [202, 89, 129, 43, 185]. Some methods bypass this assumption by estimating influence based on prediction differences between original and perturbed inputs [96, 76, 52]. Perturbations can take various forms, such as zeroing values [96], adding Gaussian noise [52], or using task-specific augmentations [76, 102]. These methods aim to amplify a modality's influence by increasing the impact of perturbations on the output. However, this also makes the network more sensitive to these changes (e.g., noise), risks becoming overly reliant on the perturbations, and struggles to scale when multiple perturbations are required. Additionally, increasing the contribution of one modality can be achieved by overshadowing others, leading to an imbalance that undermines overall performance and makes the objective counterproductive.

*Given these challenges, how can we design an efficient regularization method that addresses multimodal competition, ensuring balanced and effective learning across all modalities?*

In this chapter, we introduce the Multimodal Competition Regularizer (MCR), a loss component designed to promote the exploration of task-relevant information across all available modalities. By decomposing the joint mutual information (MI), we separately model shared and unique task-relevant information within the modalities. To efficiently capture the unique information from each modality, we employ a computationally inexpensive permutation-based approach. Our method maximizes the lower bounds of each MI term to encourage the network to learn both shared and unique information, while minimizing upper bounds on terms to suppress task-irrelevant information. We frame the problem in a game-theoretic setting, exploring strategies that involve both collaboration and competition among modalities to address the conflicting objectives that arise when increasing all modalities' contributions simultaneously. This approach allows their contributions to adapt dynamically, achieving balance during training.

We extensively evaluate MCR on synthetic datasets and several established real-world multimodal benchmarks, including action recognition on AVE [173] and UCF [161], emotion recognition on CREMA-D [25], human sentiment on CMU-MOSI [206], human emotions on CMU-MOSEI [207] and egocentric action recognition on Something-Something [58]. Our results demonstrate that MCR is the first balancing method to significantly improve supervised multimodal training over the ensemble baseline across a variety of datasets and models. The key contributions of this work are summarized as follows:

1. An analysis of multimodal competition, defining the error increase caused in multimodal training, while demonstrating in our results that most previous methods do no outperform simple baselines, such as unimodal ensembles.
2. A novel multimodal training strategy, MCR, designed to regularize multimodal competition, which includes:
   - Defining lower and upper bounds of the MI terms, encouraging the exploration of information across all modalities.
   - Introducing a game-theoretic perspective where modalities form the players that compete for training resources, assisting the regularization through balancing the corresponding MI terms.
   - Suggesting latent-space perturbations as an efficient way to estimate the lower bound of the CMI reducing the computational cost of multiple forward passes.

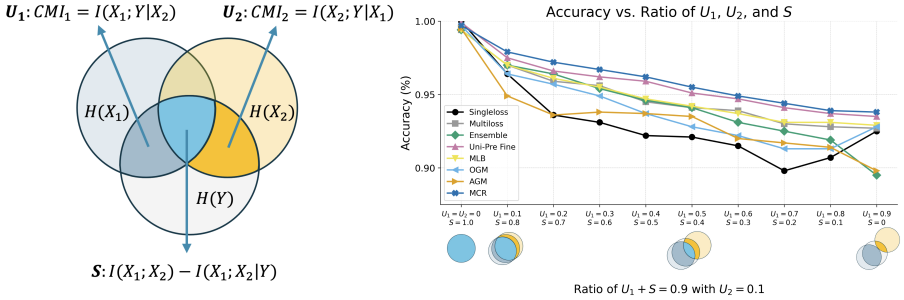



Figure 6.1: (Left) Illustration of the conditional mutual information (CMI) terms, $\text{CMI}_1 : I(X_1; Y \mid X_2)$ and $\text{CMI}_2 : I(X_2; Y \mid X_1)$, representing the unique contributions ($U_1$ and $U_2$) of each modality to the target. The shared task-relevant information ($S$) between the modalities is defined as $I(X_1; X_2) - I(X_1; X_2 \mid Y)$. (Right) Accuracy as a function of the ratio between the unique information ($U_1$) from modality $X_1$ and the shared information ($S$) between the modalities. Synthetic data are generated as $X_1 = N_1 + Y, X_2 = N_1 + Y$ where $N_1, N_2$ are independent noise for each modality. We consider $S$ the percentage of the datapoints that both modalities have information about the label $Y$, $U_1$ and $U_2$ when only one has with the other modality equating to noise for those datapoints. In the experiment we keep $U_2$ constant while changing $U_1$ and $S$. As $U_1$ increases and $S$ decreases, accuracy deteriorates, reflecting intensified multimodal competition. Among the various methods, including Singleloss, Multiloss, Ensemble, unimodally pretrained and finetuned encoders (Uni-Pre Fine), OGM [129], AGM [96], and MLB [89]our regularization method MCR demonstrates a slower decline in accuracy. For further detals prease refer to Section 6.4.1

The rest of this chapter is organized as follows. It begins with a detailed analysis of multimodal competition, outlining the limitations of supervised multimodal training and formalizing the impact of modality dominance on model generalization through the introduction of the Multimodal Competition Error (MCE). Next, the MCR method is presented, explaining its foundations in information theory, its game-theoretic perspective, and the permutation-based trick that enhances computational efficiency. The chapter then proceeds to describe the experimental settings and evaluations conducted on both synthetic and real-world datasets, showcasing MCR's performance relative to existing methods and baselines. An ablation study is included to explore the contributions of MCR's components and strategies. The chapter concludes with a discussion of the findings and their broader implications.

## 6.2 Analysis of Multimodal Competition

Consider a dataset of $N$ independent and identically distributed (i.i.d.) datapoints sampled from a distribution $\mathcal{D}$, where each datapoint has $M$ modalities $X = (X_1, \ldots, X_M)$ and a target $Y$. Our goal is to learn a parameterized function $f : X; \theta \rightarrow Y$, where $\theta$ denotes all the model's learnable parameters. The unimodal encoder for each modality is defined as $f_m : X_m; \theta_m \rightarrow Z_m$, encoding input $X_m$ into a latent representation $Z_m$. The fusion network $f_c : [Z_1, \ldots, Z_M]; \theta_c \rightarrow Y$ predicts $Y$ from the latent representations, as do the unimodal task heads $f_{c_m} : Z_m; \theta_{c_m} \rightarrow Y_m$. Model families are defined as,

Unimodal models for $m = [1, .., M]$ modalities:

$$\mathcal{F}_{u_m} : f_{u_m}\left(X_m; \theta_m, \theta_{c_m}\right) = f_{c_m}\left(f_m\left(X_m; \theta_m\right); \theta_{c_m}\right), \tag{6.1}$$

and for Multimodal models:

$$\mathcal{F} : f(X; \theta) = f_c\left(\left[f_1\left(X_1; \theta_1\right), .., f_M\left(X_M; \theta_M\right)\right]; \theta_c\right). \tag{6.2}$$

For simplicity, we continue our analysis with $M = 2$, focusing on models with two modalities.

**The limitation of supervised multimodal training**

In supervised learning with multimodal inputs $X_1$ and $X_2$, the objective is to learn representations $Z_1 = f_1(X_1; \theta_1)$ and $Z_2 = f_2(X_2; \theta_2)$ that, when combined through a fusion model $f_c([Z_1, Z_2])$, accurately predict a target $Y$. This can be expressed as minimizing the expected task loss or by maximizing the mutual information between the fused representation and the target:

$$(Z_1, Z_2) = \underset{\substack{Z_1 := f_1(X_1; \theta_1), \\ Z_2 := f_2(X_2; \theta_2)}}{\arg\max} I(f_c([Z_1, Z_2]); Y). \tag{6.3}$$

During training, the model prioritizes the more accessible or higher-quality modality, leading to over-reliance on one modality and limiting the effective use of all available information. As a result, the mutual information is dominated by the stronger modality, with $I(f_c([Z_1, Z_2]); Y) \approx I(Z_1; Y)$ and the conditional mutual information (CMI) $I(Z_2; Y \mid Z_1) \approx 0$, indicating that once $Z_1$ is learned, $Z_2$ adds little to predicting the target.

This issue extends beyond multimodal learning. In single-modality feature learning, models often prioritize dominant features, neglecting others that enhance generalization. Techniques like $l_1$ / $l_2$ penalties and dropout [120, 162] encourage balanced feature use and diverse pattern learning. However, adapting

these methods to multimodal learning is challenging; for instance, applying dropout individually to each modality [197] has proven limited, as shown afterward [129]. The key challenge remains effectively regulating modality interaction and competition, an open question in the field.

**Multimodal competition**

Multimodal competition occurs when the network primarily optimizes for one modality, leading to a decline in generalization. That modality reduces the training error by overfitting the data and limiting gradient feedback to the other modalities. Huang et al. [73] demonstrate that in late fusion models (i.e., $\theta_c = \emptyset$), each modality has a probability of causing multimodal competition. They define the phenomenon in terms of the correlation $\Gamma_m = \sum_{cl} \max[\langle X_m^{cl}, W^{cl} \rangle]^+$ of each input modality $m$ and the weights of the output layer for each corresponding class $cl$. If a modality dominates the competition, it causes the other modality to maintain its initial (before training) correlation levels. We proceed with this analysis by introducing the generalization error $\epsilon$ resulting from the effects of multimodal competition.

**Definition of Multimodal Competition Error (MCE):** Let $\Gamma_1$ and $\Gamma_2$ represent the correlations of modalities $X_1$ and $X_2$ with the output. If, at the end of training, $\Gamma_1 \gg \Gamma_2$, and modality $X_2$ has predictive power greater than randomness ($R(X_2) < R(\text{random})$), while making errors independent from those of $X_1$ ($\mathbb{E}[e_1 e_2] = 0$, where $e_1$ and $e_2$ are the errors of $X_1$ and $X_2$, respectively), then for a trained model $f$ and its optimal solution $f^*$, there exists a multimodal competition error $\epsilon$, such that:

$$\epsilon \geq R(f^*) - R_{\text{emp}}(f^*) \text{ and } \epsilon \leq R(f) - R_{\text{emp}}(f) \tag{6.4}$$

where $R(\cdot) = \mathbb{E}[L(Y, \cdot)]$ and $R_{\text{emp}}(\cdot)$ denote the generalization and empirical model risks. A higher $\epsilon$ indicates a stronger effect of multimodal competition, implying that the dominance of one modality significantly impacts the model's generalization. Transitioning from late-fusion models extends this definition by replacing correlation with MI, and handling multiple modality combinations in scenarios with more than two modalities.

## 6.3 Multimodal Competition Regularizer

In this section, we introduce $\mathcal{L}_{\text{MCR}}$, a set of loss components designed to address multimodal competition. These losses optimize the MI between the target $Y$ and the joint distribution of unimodal inputs $X = (X_1, X_2)$. While focused on two modalities, an extension to $M$ modalities is detailed in Appendix B.1.4. Each loss is motivated by the following MI decomposition:

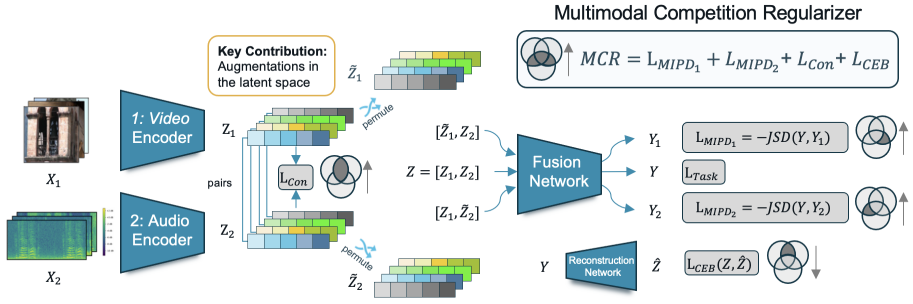



Figure 6.2: **Multimodal Competition Regularizer (MCR):** The diagram illustrates the MCR framework, which mitigates modality competition in multimodal learning. Raw data ($X_1$ and $X_2$) are encoded into latent representations ($Z_1$ and $Z_2$), which are then permuted to create $\tilde{Z}_1$ and $\tilde{Z}_2$ and the paired combinations. These combinations are passed through the Fusion Network to produce predicted outputs ($Y$, $\tilde{Y}_1$, $\tilde{Y}_2$). The comparison between predictions reveals each modality's contribution. For example, if $Y \approx \tilde{Y}_1$, it shows that $X_1$ has little impact, and the model relies on $X_2$. The MCR loss includes three components: $\mathcal{L}_{MIPD}$ maximize the Jensen-Shannon divergence (JSD) between task output and permuted modality predictions. $\mathcal{L}_{\text{Con}}$ aligns modality representations, while $\mathcal{L}_{\text{CEB}}$ penalizes task-irrelevant information by reconstructing back to the latent space. Further analysis for more than two modalities is in Appendix B.1.4.

$$I(X_1; X_2; Y) = \underbrace{I(X_1; Y \mid X_2) + I(X_2; Y \mid X_1)}_{\substack{\text{Task-Relevant Unique Information} \\ \text{of each modality} \sim \mathcal{L}_{\text{MIPD}}}} + \underbrace{I(X_1; X_2)}_{\substack{\text{Shared Information} \\ \mathcal{L}_{\text{Con}}}} - \underbrace{I(X_1; X_2 \mid Y)}_{\substack{\text{Task-Irrelevant} \\ \text{Shared Information} \\ \mathcal{L}_{\text{CEB}}}} . \tag{6.5}$$

This decomposition is illustrated by the Venn diagram in Figure 6.1. The CMIs $I(X_1; Y \mid X_2)$ and $I(X_2; Y \mid X_1)$ capture modality-specific information for predicting the target. Maximizing them with the Mutual Information Perturbed Difference (MIPD) loss, $\mathcal{L}_{\text{MIPD}}$, which assesses each modality's contribution via output variations under perturbations, encourages the extraction of task-relevant features. The third term, $I(X_1; X_2)$, quantifies shared information between modalities. Maximizing it with a contrastive loss, $\mathcal{L}_{\text{Con}}$, aligns representations and leverages their shared information effectively [77, 124, 143]. The final term, $I(X_1; X_2 \mid Y)$, represents task-irrelevant shared information. Penalizing it with $\mathcal{L}_{\text{CEB}}$ filters out irrelevant information, focusing the model on features relevant to the downstream task. For each term, we introduce a corresponding loss component, forming the proposed regularizer, which consists of three key losses:

**Regularization Loss:**

$$\mathcal{L}_{\text{MCR}} = \mathcal{L}_{\text{Con}} + \mathcal{L}_{\text{CEB}} + \mathcal{L}_{\text{MIPD}} \tag{6.6}$$

**Competition Terms:**

$$\mathcal{L}_{\text{MIPD}} = \lambda_M(\underset{\theta_1}{\mathcal{L}_{\text{MIPD}_1}} + k \cdot \underset{\theta_1}{\mathcal{L}_{\text{MIPD}_2}} + \underset{\theta_2}{\mathcal{L}_{\text{MIPD}_2}} + k \cdot \underset{\theta_2}{\mathcal{L}_{\text{MIPD}_1}}).^{1} \tag{6.7}$$

The $\mathcal{L}_{\text{MIPD}}$ contains conflicting terms, as increasing $\mathcal{L}_{\text{MIPD}_1}$, the importance of one modality can reduce the importance of the other $\mathcal{L}_{\text{MIPD}_2}$. To address this, we frame it as a game in which each modality encoder (with parameters $\theta_1$ and $\theta_2$) chooses a strategy based on hyperparameter $k$. The Lagrange multiplier $\lambda_m$ regulates the influence of $\mathcal{L}$MIPD on the total loss. Next, we detail each loss component and discuss modelling multimodal competition using game theory. The main parts of the method, as in Eq. 6.6 can been seen in Figure 6.2, the game strategies of Eq. 6.7 in Figure 6.3 and the complete algorithm is provided in Algorithm 1.

### 6.3.1 Approximating MI Terms

To approximate each CMI and capture the unique contribution of each modality, the Mutual Information Perturbed Difference (MIPD) serves as a surrogate function, measuring how input perturbations affect the model's output. By comparing predictions with and without these perturbations, MIPD estimates how much information each modality provides. If a modality is crucial, altering its input should significantly change the output, revealing its importance.

The perturbed version of modality $X_1$ is denoted as $\tilde{X}1$. We use permutations as our perturbation method, which will be detailed further in Section 6.3.3.

Estimating the CMI directly $I(X_1; Y \mid X_2) = H(Y \mid X_2) - H(Y \mid X_1, X_2)$ is typically intractable. Instead we use the MIPD, which is defined as:

$$\text{MIPD}(X_1; Y \mid X_2) = I(X_1; Y \mid X_2) - I(\tilde{X}_1; Y \mid X_2) \leq I(X_1; Y \mid X_2). \tag{6.8}$$

Using the entropy interpretation of the mutual information terms, each CMI can be expressed as the difference of the log probabilities with and without the

[1] The notation with the parameter set under the loss indicates that the loss is backpropagated exclusively to the underscored parameters.

perturbations:

$$\begin{aligned}\mathrm{MIPD}(X_1;Y \mid X_2) &= H(Y \mid X_2, \tilde{X}_1) - H(Y \mid X_2, X_1) \\ &= \mathop{\mathbb{E}}_{\substack{y\sim p(y)\\ x_1,x_2\sim p(x_1,x_2,y)}} \left[ - \mathop{\mathbb{E}}_{\tilde{x}_1\sim p(x_1)} [\log p(y \mid x_2, \tilde{x}_1)] + \log p(y \mid x_2, x_1) \right]. \end{aligned} \tag{6.9}$$

Moreover, we employ the Jensen-Shannon divergence (JSD) [105], which is symmetrically bounded, to avoid training instabilities leading to the following objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{MIPD}_{\mathrm{X}_1}} &= -\,\mathrm{MIPD}(X_1;Y \mid X_2) \\ &= - \mathop{\mathbb{E}}_{\substack{y\sim p(y)\\ x_1,x_2\sim p(x_1,x_2,y)\\ \tilde{x}_1\sim p(x_1)}} [\mathrm{JSD}(P(y \mid x_2, x_1), P(y \mid x_2, \tilde{x}_1))]\,. \end{aligned} \tag{6.10}$$

Next, the $I(X_1; X_2)$ MI term measures how much information the two modalities share, capturing the common patterns between the modalities and aligning the representations of these shared aspects. We exploit the available label information employing the supervised contrastive loss [84] $\mathcal{L}_{\mathrm{Con}}$. The term is defined as follows:

$$\mathcal{L}_{\mathrm{Con}}(X_1, X_2) = \mathop{\mathbb{E}}_{\substack{x_1,y\sim p(x_1,y)\\ x_2^+\sim p(x_2|y)\\ x_2^-\sim p(x_2|\neg y)^2}} \left[ \log \frac{\psi(x_1, x_2^+)}{\sum_k \psi(x_1, x_{2_k}^-)} \right], \tag{6.11}$$

where $\psi$ is the critic function, which, in our case, is the exponential dot product. Minimizing the $\mathcal{L}_{\mathrm{Con}}$, maximizes a lower bound on both the MI between the two modalities and the CMI terms:

$$I(X_2;Y|X_1) + I(X_1;Y|X_2) + 2 \cdot I(X_2;X_1) \geq \log N - \mathcal{L}_{\mathrm{Con}}^{Opt} \tag{6.12}$$

As $N$ increases the bounds become tighter, while the bound is not affected by the number of positive samples (same class datapoints). More details are provided in Appendix B.2.

Finally, $I(X_1; X_2 \mid Y)$ captures irrelevant shared information between modalities, and minimizing an upper bound on this ensures the model retains only task-relevant content. For this purpose, we exploit CEB [45], targeting superfluous information in multimodal representations via a reconstruction loss. A small reconstruction head, $h : Y; \theta_h \rightarrow Z = (Z_1, Z_2)$, predicts back into the latent

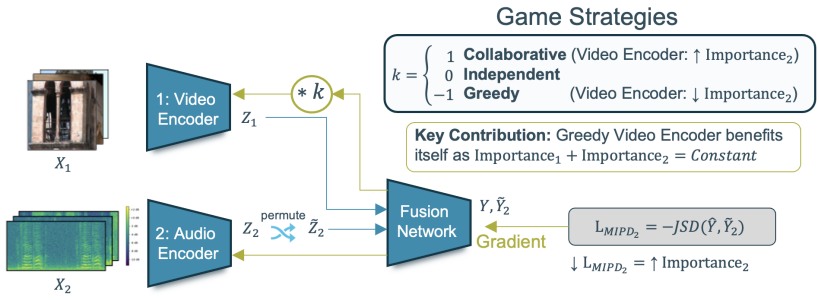



Figure 6.3: This figure illustrates a key aspect of our training process, showing how competition strategies between modalities are applied. The gradient multiplier adjusts the video encoder's response to Audio Importance ($\text{Importance}_2$). When $k = 1$, the video encoder enhances $\text{Importance}_2$; at $k = 0$, it remains neutral, and at $k = -1$, it competes by reducing $\text{Importance}_2$ to prioritize its own ($\text{Importance}_1$). This reflects the principle that increasing the importance of one modality can reduce the importance of the other.

space, effectively filtering out irrelevant content:

$$\mathcal{L}_{\text{CEB}}(X_1; X_2) = \mathop{\mathbb{E}}_{x_1, x_2, y \sim p(x_1, x_2, y)} \| [f_1(x_1), f_2(x_2)] - h(p(y \mid x_1, x_2)) \|^2 \tag{6.13}$$

Penalizing irrelevant information has been shown to enhance calibration and robustness [46], but it must be carefully evaluated, as it can introduce constraints that may hinder overall performance.

### 6.3.2 The Game of Multimodal Fusion

In this section, we present the game-theoretic approach used to balance the terms of the proposed multimodal training regularizer. The key insight is that the total contribution from all modalities, expressed through modality importance, remains constant. This implies that when one modality's contribution increases significantly, it may diminish the contributions of other modalities. This allows us to frame multimodal competition as a constant-sum game where each modality acts as a player. The $\mathcal{L}_{\text{MIPD}}$ corresponding to each modality represents the payoff function for that modality, where each modality aims to maximize its own contribution to the overall output. The players take concurrent actions, updating each modality during every optimization step. We constrain the players' actions by enforcing a fixed strategy throughout training. We explore three strategic actions for the modalities:

- **Collaborative Actions:** All modalities work together to increase each other's contributions. The $\mathcal{L}_{\text{MIPD}}$ terms are applied across all parameters, resulting in $\min_{\theta} \mathcal{L}_{\text{MIPD}}$ and $k = 1$ in Eq. 6.7.
- **Independent Actions:** Each modality focuses on maximizing its own contribution by optimizing solely its respective $\mathcal{L}_{\text{MIPD}}$ term, leading to $\min_{\theta_i} \mathcal{L}_{\text{MIPD}_{X_i}}$ and $k = 0$ in Eq. 6.7.
- **Greedy Actions:** Each modality seeks to maximize its own contribution by: 1) minimizing its own $\mathcal{L}_{\text{MIPD}}$ term, and 2) maximizing the $\mathcal{L}_{\text{MIPD}}$ terms of other modalities, resulting in a min-max strategy, $\min_{\theta_i} \max_{\theta_{\neg i}} \mathcal{L}_{\text{MIPD}_{X_i}}$ [3] and $k = -1$ in Eq. 6.7.

In Figure 6.3, we provide an example of the choices the video modality can make, illustrating how it can either assist, ignore, or diminish the importance of the audio modality. By default, we adopt the greedy actions strategy unless otherwise specified.

### 6.3.3 Perturbations

To measure the importance of modality $X_1$, we observe the difference in its predictive output when presented with $\{\tilde{X}_1, X_2\}$ compared to the normal input $\{X_1, X_2\}$.

Previous works have explored various perturbation techniques. Gat et al. [52] used additive noise to estimate functional entropy by maximizing output variance. Maximizing this variance can increase sensitivity to noise, undermining the goals of Lipschitz smoothness [6] and adversarial robustness [170]. Ji et al. [76] and Liang et al. [102] proposed task-specific augmentations, that may not be available in most tasks and modalities, while their method relies heavily on the success of these augmentations. Li et al. [96] employed zero-masking to estimate Shapley values and trained directly to maximize them. However, Shapley values can produce arbitrary or unreliable estimates, particularly in high-dimensional settings [112]. Additionally, each of these methods perturbs the input, requiring extra forward passes and proportionally increasing computational and memory demands.

To address the limitations of the suggested perturbations and face their computational need, we propose as the perturbation to use a within-batch permutation function $\sigma_e \sim \text{Uniform}(\mathcal{P})$, where $\mathcal{P}$ represents the set of all possible

[3] The notation $\neg i$ refers to the rest of the modalities except $i$.

permutations [51]. The permuted version of $X_1$ is then defined as $\tilde{X}_1 = \sigma_e(X_1)$. Permutations differ from previous methods by applying the transformation directly to the latent space, eliminating the need for multiple forward passes through the unimodal encoders, thereby reducing both computational and memory overhead.

An analysis of the computational benefits of applying perturbations to the latent space is provided in Appendix B.1.5, along with a comparison of perturbations in Appendix B.1.6.

**Algorithm 1** Multimodal Training with MCR

**Input:** Training dataset $D$ with modalities $X_1, X_2, \ldots, X_M$, labels $Y$, multimodal model $f \in \mathcal{F}$, initialized unimodal encoders $\theta_i$, reconstruction model $h$, $\lambda_{\text{uni}}, \lambda_{\text{M}}$ Lagrangian coefficients:

1: **for** each batch $(X_1, .., X_M, Y)$ of each epoch **do**
2: Compute $L_{task}(f(X_1, .., X_M), Y)$ and $L_{task}^{uni} = \lambda_{\text{uni}} \sum_{m=1}^{M} L_{task}(f_m^u(X_m), Y)$
3: Extract $(Z_1, .., Z_M)$ from $f(X_1, .., X_M)$
4: Assess the $L_{\text{Con}}$ with Eq. 6.11 and $L_{\text{CEB}}$ with Eq. 6.13
5: Sample $\sigma_e$ permutations and compute permuted pairs on the latent space $Z$
6: Pass each pair through the fusion model $f_c$ to get predictions $\tilde{Y}_m$ with modality $m$ permuted
7: Compute $L_{\text{MIPD}}$ using Eq. 6.7, 6.10 and choosing term $k$ through the corresponding strategy
8: Determine $L_{\text{MCR}}$ from Eq. 6.6
9: Update the model parameters based on $L_{total} = L_{task} + L_{uni} + L_{\text{MCR}}$
10: **end for**

## 6.4 Experiments

We perform extensive experiments on synthetic and real-world multimodal datasets to evaluate the effectiveness of our proposed method, MCR, against various baselines and previous methods:

**Unimodal Training**: Separate training for each modality.

**Ensemble**: Combines unimodal predictions without any further training.

**Joint Training**: Trains all modalities under a single-loss objective, without any balancing.

**Multi-Loss** [181]: Incorporates additional unimodal task losses alongside the joint objective.

**MMCosine** [201]: Equalizes modality contributions by standardizing features and weights.

**MSLR** [202]: Adapts unimodal encoders' learning rates based on validation performance.

**PMR** [43]: Similar to Multi-Loss, uses prototype classifiers and adjusts unimodal loss weights.

**OGM** [129]: Extracts unimodal performance from the last layer, compares improvements and modulates gradients by adjusting each encoder's learning rate.

**AGM** [96]: Uses zero-masking Shapley values as a predictor for the unimodal task losses and modulates them by adjusting their coefficients.

**MLB** [89]: Combines Multi-Loss for enhanced unimodal performance assessment and adjusts each encoder's learning rate.

**Uni-Pre Frozen**: Uses pre-trained unimodal encoders with frozen weights during joint training.

**Uni-Pre Finetuned**: Employs pre-trained unimodal encoders, fine-tuned during joint training.

### 6.4.1 Synthetic Dataset

We generate an artificial scenario where the mutual information between the label and the modalities varies, demonstrating the phenomenon of modality competition. Although many factors can contribute to this effect, a comprehensive exploration is beyond the scope of this work. Instead, we focus on the imbalance in modality informativeness to highlight the motivation behind our approach.

**Data:** We generate task-irrelevant information for each modality by sampling $N_1, N_2 \sim \mathcal{N}(0, I)$ and the 5-class label $Y$ from a uniform distribution $Y \sim$ Uniform(5). Using fixed transformations, similar to Liang et al.[102], each modality is converted into a high-dimensional vector. We relate both modalities to the label through a linear relationship: $X_1 = N_1 + Y$ and $X_2 = N_2 + Y$. Data

points are distributed such that either both modalities contain label information (Shared Information) or only one of the modalities (Unique Information). We vary the percentage of data points with shared and unique information to analyze how models perform under different conditions.

**Results**: Figure 6.1 shows the performance on synthetic data, comparing our method (MCR) with several baselines. As the shared information $S$ among the modalities decreases, and the unique information of one modality $U_1$ increases while the $U_2$ remains constant, we observe for all methods a performance drop. MCR maintains the highest accuracy across all combinations, demonstrating the slowest decline and highlighting its robustness to such imbalance.

### 6.4.2 Real-World Datasets

**Datasets:** We explore several real-world datasets, primarily with video, optical-flow, audio, and text modalities, that have either been identified as exhibiting significant imbalance among modalities or serve as standard multimodal benchmarks:

CREMA-D [25]: An emotion recognition dataset featuring 91 actors expressing 6 distinct emotions.

AVE [173]: A collection of videos with temporally aligned audio-visual events across 28 categories.

UCF [161]: An action recognition dataset consisting of real-life YouTube videos.

CMU-MOSEI [207]: Large-scale multimodal sentiment analysis dataset with 23k monologue clips.

CMU-MOSI [206]: Multimodal sentiment analysis dataset with over 2k YouTube video clips.

Something-Something (V2) [58]: Contains 220k clips of individuals performing 174 hand actions.

**Models:** We employ a variety of models and backbone encoders to examine the behaviour of both smaller-scale models trained from scratch and larger, more complex models pretrained with self-supervised learning (SSL). This combination allows us to demonstrate that our method is effective in both limited data scenarios without pretraining and in cases with ample data where the goal is fine-tuning. We utilize ResNet-18 and small-scale Transformers as alongside with state-of-the-art models like Conformer [56] Swin-TF [107].

Detailed model configurations and dataset descriptions both for the new datasets of this chapter and the ones presented in the previous ones are aggregated in

Appendix B.1.1. These choices aim to bridge the gap between theoretical work and practical application.

**Results:** In Figure 6.4 and in Table 6.1 we present the accuracy comparison among the different baseline methods on our datasets. In the figure, we provide accuracy as a comparison with that of Ensemble to highlight the difference between multimodal and unimodal training capabilities. We use both backbone encoders where we identified good SSL pretrained fit models: Conformer [56](in orange) and ResNet (in blue). Positive values indicate an improvement over the ensemble baseline, while negative values signify a decline in performance. Our main observations are:

a) Most previous methods and baselines fail to outperform the Ensemble, often showing negative accuracy differences, particularly on the CREMA-D and AVE datasets. These methods do not adequately address the modality competition problem, leading to a worse generalization error struggling to leverage the given multimodal data, regardless of the chosen backbone model.

b) Our method, MCR, stands out as the only approach that consistently achieves better across all datasets and backbone models. Notably, MCR outperforms consistently the Ensemble baseline across all models and datasets, demonstrating its effectiveness in balancing modality contributions under several different configurations.

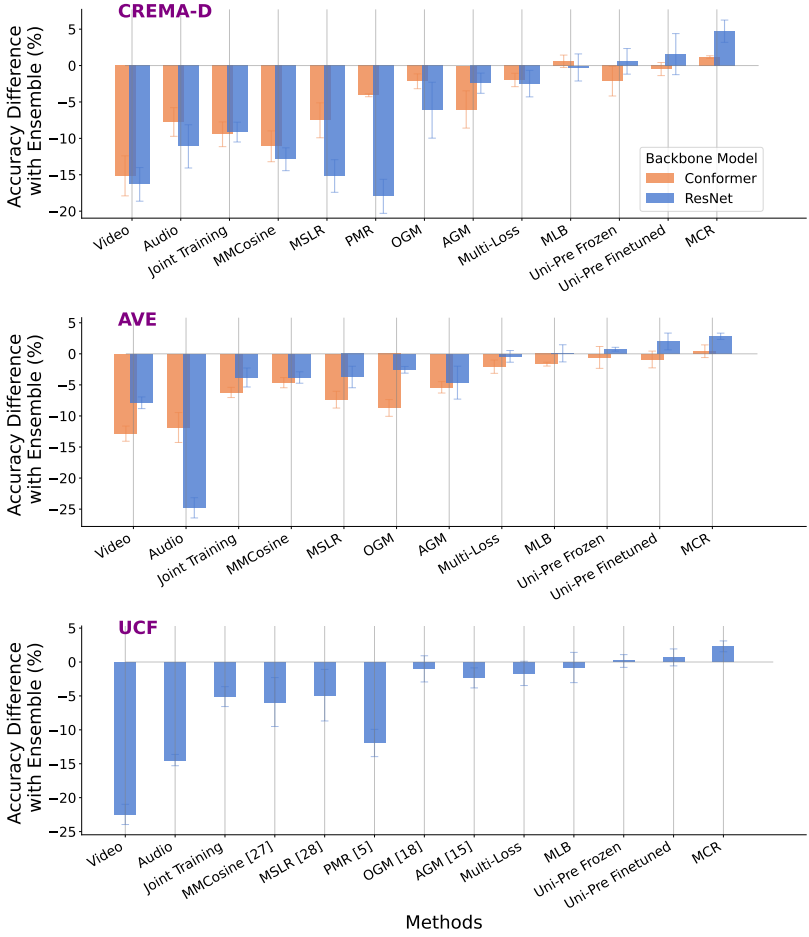



Figure 6.4: Accuracy(%) difference compared to the Ensemble baseline for various methods across three datasets: CREMA-D, AVE, and UCF101. Results are shown for two backbone models: Conformer (orange) and ResNet (blue). Positive values indicate an improvement over the Ensemble, while negative values represent a performance drop. Most existing methods fail to outperform the Ensemble, which does not utilize any multimodal learning. In contrast, our proposed method, MCR, consistently achieves positive accuracy differences, surpassing both the previous suggested methods and our own defined baselines.

Table 6.2: Ablation Study: Game Strategies – Comparison of model accuracy for different game strategies: Collaborative, Independent, and Greedy, using both ResNet and Conformer backbones.

| Model Setting | Collaborative $\min_{\theta} \mathcal{L}_{\text{MIPD}}$ | Independent $\min_{\theta_i} \mathcal{L}_{\text{MIPD}_{\text{X}_\text{i}}}$ | Greedy $\min_{\theta_i} \max_{\theta_{\neg i}} \mathcal{L}_{\text{MIPD}_{\text{X}_\text{i}}}$ |
|---|---|---|---|
| ResNet + MCR | $73.4_{\pm 3.0}$ | $76.0_{\pm 2.0}$ | $\mathbf{76.2}_{\pm \mathbf{1.7}}$ |
| Conformer + MCR | $82.9_{\pm 0.7}$ | $82.9_{\pm 1.4}$ | $\mathbf{85.7}_{\pm \mathbf{0.2}}$ |

### 6.4.3 Ablation Study

Table 6.1: This table compares the proposed MCR method with several baselines across the additional datasets: CMU-MOSEI, CMU-MOSI, and Something-Something. Results are shown for different modality combinations (e.g., V-A-T: Video-Audio-Text), using Transformer and Swin-TF backbones. MCR consistently outperforms other methods across all configurations.

| | Transformer | | | | Swin-TF |
|---|---|---|---|---|---|
| | MOSEI | | MOSI | | Something-Something |
| Method | V-T | V-A-T | V-T | V-A-T | V-OF |
| Unimodals | V: 63.6 A: 64.3 T: 79.9 | | V: 53.5 A: 54.9 T: 70.7 | | V:61.6 OF:50.6 |
| Ensemble | 77.4 | 76.6 | 69.8 | 43.3 | 64.8 |
| Joint Training | 80.5 | 81.2 | 74.7 | 73.9 | 56.3 |
| Multi-Loss | 80.8 | 79.9 | 75.3 | 72.0 | 59.6 |
| OGM [129] | 80.6 | - | 75.2 | - | 58.3 |
| AGM [96] | 79.5 | 80.6 | 75.3 | 75.2 | 56.9 |
| MLB [89] | 80.8 | 81.1 | 75.0 | 74.5 | 61.3 |
| Uni-Pre Frozen | 80.5 | 79.8 | 70.6 | 70.6 | 64.2 |
| Uni-Pre Finetuned | 80.7 | 80.1 | 73.8 | 71.2 | 61.9 |
| MCR | **81.2** | **81.6** | **76.1** | **77.9** | **65.0** |

**Strategy:** We perform an ablation study to test our hypothesis that framing multimodal competition regularization as a game benefits the model by avoiding destructive loss interactions in each backbone encoder. Table 6.2 compares the three strategies: Collaborative, Independent, and Greedy. The results show that, across the models allowing backbone encoders to maximize their own CMI term and concurrently minimizing the others (Greedy strategy) consistently yields the best performance. This result demonstrates that framing multimodal models as competing modalities using game-theoretic principles in the loss terms can be beneficial in balancing these loss terms.

**Loss Components:** Table 6.3 presents the model's performance comparison when different loss components are applied. The models utilize pretrained

Table 6.3: Ablation Study: Regularization Components – Accuracy (%) of ResNet and Conformer models across datasets with different combinations of MCR components: $L_{MIPD}$, $L_{Con}$, and $L_{CEB}$. Results indicate that combining $L_{MIPD}$ and $L_{Con}$ is crucial for improvement, while $L_{CEB}$ does not benefit all models.

| MCR Components | | | ResNet | | | Conformer | |
|---|---|---|---|---|---|---|---|
| $L_{MIPD}$ | $L_{Con}$ | $L_{CEB}$ | CREMA-D | AVE | UCF | CREMA-D | AVE |
| | | | 73.4±2.5 | 71.1±1.4 | 50.0±2.0 | 84.1±0.6 | 87.9±1.1 |
| ✓ | | | 74.1±2.9 | 72.1±0.5 | 49.4±1.9 | 83.9±1.8 | 87.8±1.7 |
| | ✓ | | 73.4±2.1 | 72.6±0.6 | 54.8±1.2 | 84.5±0.3 | 88.7±1.4 |
| ✓ | ✓ | | **76.2±1.7** | **73.3±0.5** | **55.1±0.6** | 84.5±0.3 | 88.7±1.4 |
| ✓ | ✓ | ✓ | 75.6±1.9 | 72.1±0.9 | 54.7±1.1 | **85.7±0.2** | **88.8±1.0** |

initialization: ResNet with unimodal pretraining and Conformer with SSL. Two key observations can be made:

1. The concurrent exploitation of both $L_{MIPD}$ and $L_{Con}$ is yielding consistent improvement. Exploiting them separately leads to smaller improvement for $L_{Con}$ and even to a decline for $L_{MIPD}$. Suggests that alignment in the latent space between the modalities is necessary for the permutations to be effective.

2. The $L_{CEB}$ term, which penalizes task-irrelevant information, improves the Conformer model's performance, likely due to its pretraining on large, unlabelled datasets that introduce such irrelevant information. In contrast, for the ResNet model, where pretraining already focuses on task-related information, the $L_{CEB}$ term does not provide additional benefits.

### 6.4.4 Analysis of Multimodal Error

In this section, we investigate the specific data points that MCR prioritizes during training. The framework is designed to balance and enhance the independent contributions of each modality through the CMI terms while strengthening their shared contributions via contrastive loss. However, no specific term explicitly addresses the synergetic, emergent information that arises from the interplay between modalities. The underlying expectation is that encouraging modalities to improve concurrently would naturally lead to such synergetic behaviour. Here, we conduct a post-hoc error analysis to examine the results and better understand the benefits that MCR provides.

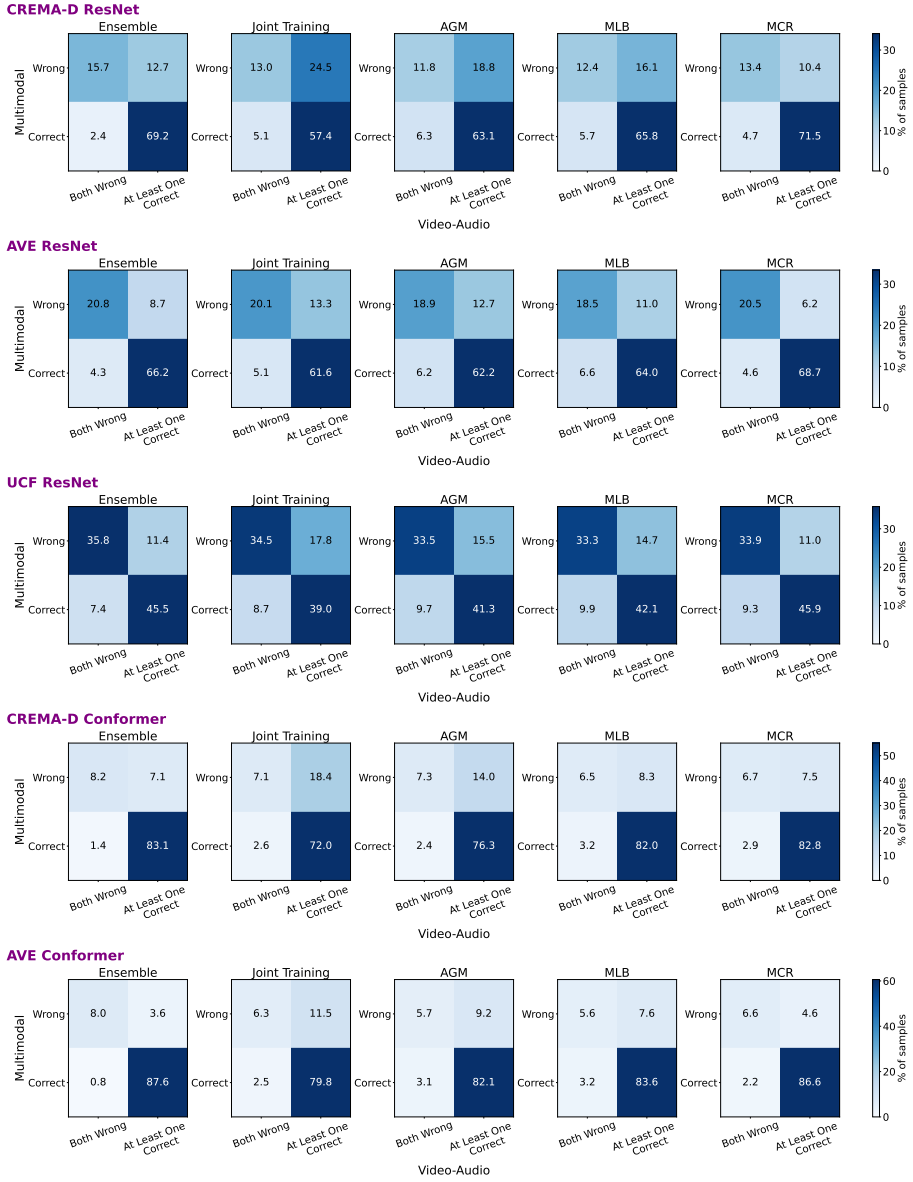



Figure 6.5: Error comparison matrices across pairs of datasets and models, comparing unimodal predictions with multimodal models trained using various methods, including Ensemble, Joint Training, AGM, MLB, and MCR. Each column of the confusion matrix represents cases where both unimodal predictions are incorrect, where only one is correct, and where both are correct. The results highlight that MCR consistently performs well in cases where at least one unimodal prediction is correct, with ResNet-based models surpassing even the Ensemble in this category. Additionally, MLB demonstrates stronger performance in discovering emergent information, highlighting a current limitation of MCR in this aspect.

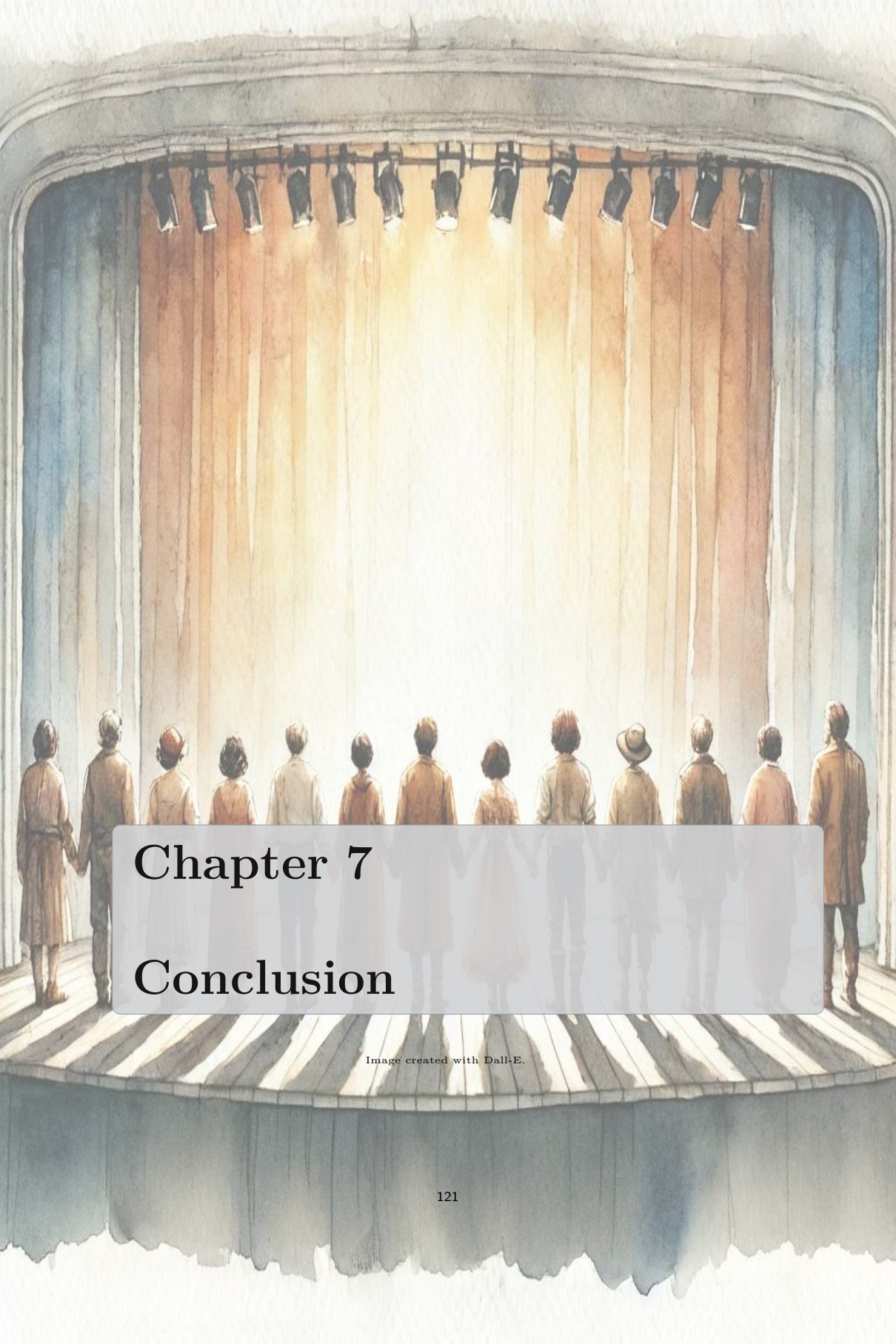
We analyze a matrix comparing the unimodal wrong and correct predictions with those from multimodal models trained using various methods, including Ensemble, Joint Training, AGM, MLB, and MCR. As shown in Figure 6.5, a consistent pattern emerges across different dataset/model pairs. MCR outperforms previous methods in cases where at least one modality independently makes the correct prediction, with the ResNet-based models even surpassing the Ensemble in this category. However, we observe that while MLB struggles to retain unimodal information effectively, it consistently performs better at discovering new emergent information. MCR improves upon this in certain instances by approaching the best-performing baselines but it generally shows a decline in that category. This analysis is encouraging for two reasons: it highlights a significant percentage of unimodal information that is not adequately leveraged by the multimodal model, and it indicates potential for further improving the discovery of emergent information, both of which are crucial for achieving performance gains that justify the effort required to collect, process, and compute the additional data.

## 6.5 Discussion

In this chapter, we propose a new multimodal balancing method to address multimodal competition, the phenomenon where certain modalities dominate the training process, leading to suboptimal performance. To tackle this issue, we formally define the concept of multimodal competition error (MCE) and aim to reduce it through our proposed method. We introduce the Multimodal Competition Regularizer (MCR), a novel framework inspired by information theory, which models multimodal learning as a game where each modality competes to maximize its contribution to the final output. MCR achieves this by efficiently identifying and computing lower and upper bounds to optimize both unique and shared task-relevant information for each modality. We conduct extensive experiments on both synthetic and real-world datasets, utilizing various modalities and backbone encoders, to demonstrate that MCR not only surpasses prior methods but also becomes the first approach to significantly outperform unimodally pretrained baselines. By achieving this, MCR fulfils the long-standing promise of multimodal fusion by improving beyond the ensemble baseline the joint multimodal performance.

At the same time, this method presents some limitations. While it improves the retention of datapoints correctly identified by unimodal models, it falls short of achieving the synergetic behaviour observed in some previous balancing methods. Furthermore, although MCR reduces the overall error, the theoretical potential achievable through multimodal fusion remains untapped, with possible gains

ranging from 4.6% to 11% as indicated by the post-hoc error analysis shown in Figure 6.5. These findings emphasize the significant room for improvement in multimodal fusion, presenting an exciting opportunity to push the boundaries of what can be achieved. Advancing methods to better retain unique unimodal information while simultaneously enhancing synergistic interactions among modalities could unlock unprecedented levels of performance. Such progress has the potential to bridge the current gap toward the theoretical limits of multimodal fusion, paving the way for models that fully exploit the richness of multimodal data and deliver transformative results.



# Chapter 7

# Conclusion

Image created with Dall-E.

## 7.1 Main Findings Part 1: CoRe-Sleep

In the initial part of this thesis, we investigated the impact of architectural choices and training objectives on multimodal models using sleep staging as a case study. Our first objective was to assess how different fusion strategies, early, late, and dynamic fusion points, affect the performance and robustness of multimodal systems. We observed that while the multimodal performance of various architectures often converges, robustness to data imperfections varies significantly across architectures. This finding underscores that while fusion mechanisms have traditionally been seen as central to multimodal performance, they primarily contribute to model resilience against issues such as noise or missing data.

The development of CoRe-Sleep, a Coordinated Representation model tailored for sleep staging. CoRe-Sleep enables flexible fusion, allowing the model to leverage multimodal data even when some modalities are unavailable. Compared to other standard multimodal architectures, CoRe-Sleep achieves state-of-the-art performance across scenarios, effectively balancing the benefits of early fusion's strong cross-modal integration and late fusion's robustness. This highlights how dynamic fusion architectures can bridge the gap between performance and robustness in complex multimodal tasks.

Our study also brought to light two insights for the broader field of multimodal learning. Firstly, training with multiple modalities can enhance overall performance even when not all are present during testing. This insight only challenges the conventional training approach with the modalities expected during evaluation, demonstrating that a more comprehensive approach to modality inclusion can yield stronger generalization. Secondly, including data points with missing modalities during training further strengthens the model's resilience, underscoring the importance of designing multimodal models that can effectively learn from incomplete datasets.

Ultimately, that work raised important questions about the role of training objectives in multimodal learning. We observed that different training objectives applied to any multimodal architecture notably impacted multimodal performance and robustness. This finding encouraged us to look further into multimodal optimization and the challenges of joint multimodal training.

## 7.2 Main Findings Part 2: Multimodal Competition

Multimodal learning promises to leverage complementary information from diverse modalities. Yet, it is often hindered by optimization and generalization difficulties during training, leading to only a subset of modalities being sufficiently trained, suppressing the contributions of others, and hence to suboptimal learning. This part of the thesis concludes our investigation into multimodal competition by summarizing its causes, briefly discussing the existing solutions, and reflecting on our proposed approaches to address its challenges.

### 7.2.1 Understanding Modality Competition

The optimization phenomenon of modality competition refers to the imbalance that arises during multimodal training, where certain modalities dominate the learning process, overshadowing others. We begin our analysis by showing that if trained under a joint multimodal objective with a non-linear loss, any model structure allows one modality to interfere with the learning updates of the parameters dedicated to the other modality. This allows for an uncontrolled affection of one modality to the learning of the other encoders. Such interference could potentially be constructive by allowing modalities to filter out modality-independent noise, however, in practice, we observe that they are usually destructive leading to misdirected and vanishing gradients for a subset of modalities, increasing the generalization error of the multimodal model.

### 7.2.2 Methods to tackle Multimodal Competition

In chapter 3, we introduced several prior methods to address this problem. Some methods assume that modality competition arises from the unequal magnitude of features at the fusion level, leading to imbalances during integration. To address this, they apply feature normalization techniques, standardizing modality-specific representations to equalize their contributions. Other approaches rest on the assumption that modalities have differing convergence rates in their respective parts of the network, which can create imbalances over time. To mitigate this, these methods propose measuring the unimodal contributions and dynamically adapting the learning rates of each encoder based on their performance, ensuring synchronized progress across modalities. Building on the same logic, other methods incorporate unimodal losses to guide multimodal training, ensuring that each unimodal encoder is well-trained and can contribute meaningfully to the fused representation. These approaches often include a balancing mechanism to weigh the unimodal losses, similar to the adaptive

learning rate techniques that adjust encoder updates based on unimodal contributions. By aligning the influence of unimodal losses with the overall training objective, this strategy seeks to enhance the effectiveness of multimodal learning, leveraging strong unimodal encoders. Our proposed method, Multi-Loss Balanced (MLB), combines the strengths of the latter two approaches by incorporating both unimodal losses and adaptive learning rate adjustments for each encoder. Finally, the last category includes methods that recognize that measuring unimodal performance may miss the importance of modalities that arise through correlations and suggests explainability methods to measure unimodal performance in a multimodal model and maximize each modality's contribution. Finally, our extension of such methods, Multimodal Competition Regularizer (MCR), recognizes that increasing the modality contribution of one modality might have adverse effects on the others. To mitigate this, MCR frames multimodal fusion through the lens of game theory, where each modality acts as a player striving to maximize its influence on the output, effectively exploring contributions from all modalities.

### 7.2.3 Proposed Advancements in Balancing Approaches

To address the challenges posed by modality competition and enhance multimodal learning, we propose two new methods: MLB [89] and MCR [91]. These approaches are designed to dynamically balance modality contributions while addressing key limitations in prior methods. MLB focuses on synchronizing learning dynamics through a combination of unimodal losses and adaptive balancing mechanisms, ensuring that all modalities contribute effectively to the joint objective. In contrast, MCR introduces a game-theoretic framework and mutual information decomposition to balance contributions by modelling modalities as players in a competitive dynamic.

More specifically, MLB addresses the limitations in prior works, which either apply adaptive balancing to the learning rates of unimodal encoders based on unimodal performance but without explicitly training for accurate performance estimation or exploit unimodal losses to improve estimation accuracy but only balance these losses. This oversight often leads to suboptimal results. MLB delivers consistent improvements across all datasets, backbone encoders, and fusion methods by combining these strategies and introducing enhancements to the balancing mechanism. It achieves this through three core components:

1. **Unimodal Loss Integration**: Additional unimodal losses are incorporated to guide the training of each modality-specific encoder, ensuring strong unimodal performance as a foundation for effective multimodal fusion.

2. **Dynamic Learning Rate Adjustments**: Learning rates for each encoder are dynamically adjusted based on the comparative metrics of unimodal performances. This improves upon prior works by balancing both unimodal and multimodal loss gradients, ensuring synchronized progress across modalities.

3. **Balancing Mechanism**: MLB includes a dynamic balancing mechanism that accelerates or decelerates updates as required to balance the contributions of all modalities. This mechanism phases out naturally as the encoders converge, eliminating the need for manual intervention or predefined stopping criteria.

Next, MCR takes a different approach by avoiding direct learning rate modifications and instead introducing additional loss components as regularization to the joint multimodal training objective. This method leverages unimodal losses while building upon explainability techniques to measure the importance of each modality for the output. Operating at the batch level, it provides direct access to measurable and differentiable terms, which can be penalized to guide the exploration of the optimal contribution from each modality. Recognizing that increasing the importance of one modality could negatively impact others, MCR frames the problem using a game-theoretic perspective. This approach facilitates a broader and more balanced exploration of modality contributions. MCR is the first method to significantly improve multimodal learning, surpassing the performance of baselines that rely on unimodally trained models. It accomplishes this through four distinct components:

1. **Mutual Information Decomposition**: Motivated by the maximization of mutual information, several terms are identified for shared and unique task-relevant information and are separately modelled to ensure that each modality's contributions are optimized without redundancy.

2. **Game-Theoretic Modeling**: We recognize the mutual information terms for each modality-independent information as one object to the other, and hence, we model them through game theory, proposing some early strategies from the game and offering a more exploratory training regime.

3. **Efficient Contribution Estimation**: MCR employs permutation-based techniques to estimate unique modality contributions. This approach reduces the computational overhead associated with mutual information calculations, translating MCR into a computationally feasible method even across large-scale applications.

### 7.2.4 Empirical Evidence

In chapters 4, 5, 6, experimentation was conducted on a variety of datasets of different sizes and modalities to evaluate the proposed methods and benchmark their performance.

In Chapter 4, we evaluated the performance of MLB across a diverse set of datasets, including AVE, CREMA-D, UCF, and MOSEI, encompassing video, audio, and text modalities. Our findings revealed that most existing methods struggle to consistently and significantly outperform models trained with a joint single-loss multimodal objective. Additionally, many of these methods face scalability issues, as they are primarily designed for scenarios involving two modalities and are not easily extendable to cases with three or more modalities. Even among the approaches capable of handling multiple modalities, their performance tends to vary inconsistently across different datasets and tasks. In contrast, MLB demonstrates robust and reliable improvements across all datasets and modalities, regardless of the scale or initialization of the backbone encoders. Whether using smaller-scale randomly initialized backbones like ResNet or larger self-supervised pretrained encoders, MLB consistently delivers superior performance. Moreover, its effectiveness extends across various fusion methods, ranging from simple linear and MLP approaches to more complex fusion gates and transformer layers. This versatility highlights MLB's adaptability and ability to address prior methods' limitations, making it a robust solution for diverse multimodal learning challenges.

In Chapter 5, we extend our baselines further by demonstrating that pretrained unimodal encoders represent a strong baseline that should not be overlooked. To evaluate this, we propose three benchmarks: an ensemble of unimodal models, and two models trained under a single-loss multimodal objective using pretrained encoders, either frozen or finetuned. Experiments reveal that MLB is the only balancing method capable of performing on par with these benchmarks, though it struggles to surpass them consistently. This raises an important question about the necessity of training multimodal models from scratch, particularly when unimodally pretrained encoders offer such a strong alternative.

Our analysis highlights a critical trade-off: while balancing methods, including MLB, often miss some unique unimodal information, they significantly enhance the discovery of emergent synergistic interactions between modalities. This trade-off ultimately reduces overall performance. These findings emphasize the need for methodologies that promote synergistic interactions and retain each modality's unique information to achieve a more balanced and effective multimodal learning framework.

Finally, in Chapter 6, we present the results of the MCR method, demonstrating

that incorporating information-theory-inspired regularization components and framing multimodal fusion as a game can significantly improve multimodal learning beyond unimodal baselines. To our knowledge, this is the first method to achieve such a performance increase. We further validate MCR's effectiveness through additional experiments on datasets like MOSI (evaluated with 2 and 3 modalities) and Something-Something. We utilise advanced models such as Swin-TF for Something-Something and additional large SSL pretrained models for AVE. A comprehensive ablation study of MCR's components provides valuable insights into the contributions of each part, promoting further refinement. MCR is the first method to fulfil the promise of multimodal fusion by demonstrating that integrating more modalities can consistently enhance model performance, marking an important milestone in the field.

### 7.2.5 Key Insights and Limitations

In this subsection, we aim to succinctly outline the key insights identified throughout this thesis, emphasizing the contributions and challenges addressed in the context of multimodal learning.

1. One of the key contributions of this thesis is the application and evaluation of a more flexible approach to multimodal modelling through coordinated representations, which avoid reliance on a predefined fusion point. We analyze the properties of various such models, examining their impact on multimodal performance and robustness, and conclude the implications of different architectural choices. Additionally, we emphasise the critical role of carefully designed training objectives as the foundation for robust and effective multimodal learning. This focus addresses some limitations of rigid architectural assumptions and highlights the importance of suitable training strategies.

2. We establish a foundation for proper empirical evaluation in multimodal learning by introducing rigorous benchmarks and well-defined experimental protocols. These frameworks enable consistent comparisons across diverse methods and datasets, providing deeper insights into the strengths and limitations of various approaches. This empirical groundwork is crucial in advancing the field toward more reliable and reproducible evaluations.

3. Our findings reveal that multimodal fusion is significantly more challenging than commonly perceived. This difficulty persists across a wide range of model architectures, from smaller-scale models to very large-scale ones, although the extent of the challenge varies with the scale. These results

underscore the need for deeper exploration and understanding of the underlying dynamics governing multimodal learning.

4. We demonstrate that previous methods have made only marginal progress in addressing the challenges of multimodal fusion. Despite the proliferation of balancing strategies, their impact has been limited, failing to outperform the unimodally pretrained baselines consistently. These findings highlight the pressing need for more effective solutions to advance the field.

5. In response, we propose methods that represent a significant step forward in improving multimodal fusion. Our contributions address modality competition through MLB, which enhances the synergy between modalities, and MCR, which excels in retaining unique unimodal information. In the latter, we introduce the concept of conflicting objectives between modalities during fusion, a practical framework that we believe can effectively unlock its full potential.

6. Our post hoc error analysis reveals a significant gap between current multimodal methods and their theoretical potential. Specifically, we observe that many data points correctly predicted by individual unimodal models are not captured by existing multimodal approaches, resulting in a high overall error rate. This suggests that multimodal models are still far from effectively leveraging the complementary information available across modalities. If a perfect multimodal fusion method were to combine the strengths of all unimodal and multimodal models, it could dramatically reduce errors and achieve a level of performance far beyond what current methods offer. This gap between current capabilities and theoretical potential is a strong motivation to continue exploring strategies for improving multimodal fusion.

7. We observe that the impact of multimodal training objectives is more pronounced when modalities are diverse. In tasks like sleep staging, where the modalities substantially overlap, the improvements from multimodal objectives were smaller compared to datasets with higher modality diversity, such as audio-video tasks. This indicates that the greater the diversity of the modalities in terms of the information they provide about the target, the more significant the competition becomes and the more it justifies focusing on modality-specific training objectives. While we have some first indications, a thoughtful examination of this remains a future direction for investigation.

8. While this thesis primarily focuses on supervised multimodal fusion, we believe that similar challenges and benefits will extend to other tasks, such as SSL and multimodal translation tasks, where multiple modalities are used as input to generate a new modality as the target (e.g., video captioning using both video and audio). The principles explored here, particularly

around modality competition and training objectives, could have valuable implications for improving performance in these tasks. This remains a limitation of the current work, and exploring these effects in different contexts is a promising direction for future research.

## 7.3 Future Work

### 7.3.1 Exploring Strategies for the Multimodal Fusion Game

A promising direction for future research lies in further exploring the game-theoretic dynamics of multimodal fusion. While our current work introduces the idea of framing modality interactions as a game [183, 158], fundamental questions remain unanswered. For instance, we still lack a clear understanding of whether this dynamic system can consistently solve the problem of modality competition and converge to a Nash equilibrium [119] or if it is inherently chaotic and sensitive to initialization or training dynamics [166, 177]. Investigating these aspects could provide critical insights into the stability and effectiveness of game-theoretic approaches in multimodal learning.

Another promising aspect in this direction involves exploring a broader range of strategy options for each modality. Currently, modalities are constrained to commit to a single strategy at the beginning of training, which remains fixed throughout the entire procedure. This static approach is likely suboptimal, as the optimal strategy for a modality may evolve throughout training due to changing dynamics and interactions. Enabling modalities to adapt their strategy dynamically during training could provide greater flexibility and improve overall performance. Furthermore, introducing more sophisticated strategic options beyond the simplistic binary choices (collaborative - "with you", or greedy - "against you") could enhance the interplay between modalities. By incorporating nuanced strategies, modalities could better preserve their unique contributions. Implementing a refined and dynamic strategy is a relatively straightforward enhancement that could play a key role in minimizing the loss of unimodal information, ultimately advancing the effectiveness of multimodal fusion.

Advancing this framework may enable better retention of the unique information from each modality, as the game-theoretic approach is primarily designed to emphasize this aspect. Our error comparison matrices indicate that enhancing the preservation of such information could result in performance gains ranging from 4.6% to 11%, depending on the dataset and task. These findings underscore the untapped potential of multimodal fusion and challenge us to master the rules and strategies of this complex game.

### 7.3.2 Optimizing Synergy in MCR

Our current best method, MCR, demonstrates significant advancements in addressing multimodal competition, but it still has two notable limitations. First, it does not fully retain all unimodal information, which we aim to address by refining the strategies within the multimodal fusion game, as explained in the previous Section 7.3.1. Second, it lacks the level of synergetic behaviour that other methods achieve.

To address the latter limitation, one of our future works involves incorporating specific terms into the loss function to encourage the model to enhance its synergetic behaviour explicitly. Inspiration for this approach is the work by Liang et al. [101], which introduces a quantification term defined as the difference between multimodal performance and that of an ensemble as a model with minimal cross-modality synergistic discovery capability. Incorporating such a term would directly motivate the model to exceed the information extracted by simple ensembles by effectively discovering and leveraging cross-modal synergies.

This future work step is straightforward. We plan to adapt and extend this idea within the MCR framework by integrating it into the loss function to drive further improvements. By doing so, we aim to enhance MCR's performance in cases requiring the discovery of emergent information. This enhancement provides a direct and practical pathway to advancing MCR's capabilities and contributing to the broader development of effective multimodal fusion methodologies.

### 7.3.3 Investigating Multimodal Competition in Multimodal Translation

While this thesis has extensively focused on supervised multimodal fusion tasks, a critical next step involves verifying our assumption that the phenomenon of multimodal competition also extends to multimodal translation tasks. These tasks can be framed as self-supervised learning objectives, focusing on either reconstruction or translation to a new modality. To highlight its significance, it's important to note that multimodal foundational models increasingly adopt masked reconstruction methods, where parts of the input are masked and predicted using information from multiple modalities [55]. Additionally, translation tasks, such as text-to-image generation or cross-modal retrieval, are used to fine-tune models, further enhancing their ability to understand and generate multimodal data. Exploring how multimodal competition influences these tasks could provide valuable insights and strategies for enhancing even the largest-scale multimodal models.

### 7.3.4 Characterizing Multimodal Diversity & Training Challenges

Throughout this thesis, we experimented with various datasets and modalities, observing that the degree of informational overlap between modalities significantly influenced training difficulties. This overlap can be assessed by comparing the individual correct predictions of each modality or by measuring the reduction in uncertainty through mutual information when including or excluding a modality. For example, EEG and EOG data in sleep staging are expected to capture similar information, both in terms of content and temporal coherence, leading to high performance even with simple joint training. In contrast, when working with more diverse modalities, such as audio and video, we encountered more pronounced training challenges, which we attribute to competition between modalities. However, it remains unclear how to best characterize and categorize the multimodal data to determine the necessary depth of training methods and guide decisions on model architecture, as well as the data magnitude required for optimal performance.

### 7.3.5 Breaking the Egg-Chicken Cycle: Discovering and Balancing Modalities

The challenge of balancing the contributions of different modalities in a multimodal system often mirrors the classic egg-chicken problem: when and how do we identify that there is a problem in the first place? Determining whether a modality is over- or under-contributing requires discovering the imbalance and assessing whether the contribution is close to an ideal balance. Motivated by questions and discussions with Prof. Liang, we highlight that as long as humans can interpret the contribution of each modality, there remains a pathway to identify discrepancies and correct them. However, beyond the point where human interpretability breaks down, assessing and ensuring an ideal contribution becomes significantly more difficult. This emphasizes the need for dynamic evaluation techniques that not only measure the impact of each modality but also align it with a human-understandable ideal. These techniques allow for iterative adjustments to maintain a balanced contribution, ensuring that the system fully exploits the complementary nature of modalities without over-reliance or neglect.

### 7.3.6 Expanding the Prerequisite Assumptions of Existing Methodology

This thesis considers modalities that meet several key assumptions: they can independently perform the given task, are paired, temporally aligned (typically measured simultaneously), and describe the same underlying event. While these are strong assumptions that may not hold universally in practice, they provide a solid foundation for achieving initial gains in performance. However, to extend the applicability of these methods to real-world scenarios, it is crucial to move beyond these constrained conditions and address cases where these assumptions do not apply. This expansion will enable the development of methodologies capable of functioning effectively in more diverse and less controlled environments.

## 7.4 Ethical and Societal Implications

Advancements in multimodal learning hold significant potential to impact various domains and aspects of society. Therefore, it is crucial to address the ethical and societal implications specifically tied to the integration of multiple modalities rather than those stemming purely from machine learning itself.

**Privacy and Data Sensitivity:** Multimodal systems often involve integrating sensitive information, such as biometrics or personal audio-visual data. Proper data governance frameworks and robust anonymization techniques are essential to protect user privacy when combining such data sources. For example, in a clinical decision support system, one medical centre might hold a patient's EEG recording while another has their medical imaging data. Integrating these modalities could improve diagnosis but risks exposing sensitive information if data is shared or centralized. Secure distributed frameworks, such as federated learning or encryption techniques, can enable models to learn from multimodal data without spreading or centralizing sensitive patient information, ensuring compliance with privacy regulations and maintaining trust.

**Resource Inequality:** Developing multimodal models often demands access to diverse and extensive datasets and substantial computational resources, which are not equally available to all researchers or organizations. This disparity risks widening the gap between well-resourced institutions and others, hindering equitable advancements in the field and limiting opportunities for broader contributions to multimodal research and innovation. While this is primarily a political concern, we emphasize the personal responsibility of researchers with access to such resources to actively seek and maintain collaborations with

universities and institutions worldwide, promoting inclusivity and equitable progress in the field.

**Interpretability and Accountability:** While multimodal models are often perceived as more challenging to interpret due to their complexity, they also present unique opportunities. For instance, modality translation can enable explanations of decisions through natural language, enhancing transparency and improving human interaction with the models. This capability bridges the gap between technical decision-making processes and user understanding, increasing trust and accountability.

**Bias Amplification:** Including more modalities introduces the risk of amplifying biases inherent in each modality, potentially compounding their effects when fused. For instance, in a multimodal system for job candidate evaluation, visual data (e.g., appearance) might carry biases related to gender or ethnicity. In contrast, textual data (e.g., resumes) might favour certain language styles associated with socioeconomic or educational backgrounds. When these modalities interact, they could disproportionately disadvantage candidates from underrepresented groups. Addressing these risks requires methods to identify and mitigate biases in the data selection process and within the model, ensuring balanced contributions from all modalities to promote fairness and inclusivity.

**Unintended Emergent Information and Complexity in Consent:** Combining different modalities can lead to the inference of new, emergent information inaccessible from individual data sources alone. For example, retailers collecting multimodal data such as purchase history (textual data), browsing patterns (behavioural data), and facial expressions during in-store or online interactions (visual data) could infer a customer's emotional state. This might then be used to strategically influence purchasing behaviour, raising ethical concerns. While users might consent to the collection of each modality, they may not fully understand how far the combination of this data can reach into their sensitive information. Restricting such practices and ensuring clear communication and transparency to users should be a high priority in policy-making for multimodal data.

**Broader Implications of Machine Learning on Society:** As we deviate from purely multimodal concerns, it is important to address a broader claim about machine learning and its societal implications. Current models and their physical realizations through robotics have reached a point where they can imitate and even surpass many human capabilities. Over the past few years, the release and rapid advancement of large language models have convinced even sceptics of this reality. While these developments provide tools designed to simplify and improve human life, the opposite effect is becoming evident: the quality of life for many is deteriorating. Instead of benefiting everyone, these

systems are often exploited to impose higher standards on individuals, pushing people to outperform machines rather than harnessing these tools for collective empowerment. With the exponential growth of their capabilities, this trajectory is unsustainable. It is imperative for those who understand these dynamics to advocate for solutions that decentralize the benefits of machine learning, ensuring that everyone, not just those who participated one way or the other in the creation of those systems, can rise above and share in this transformative power.

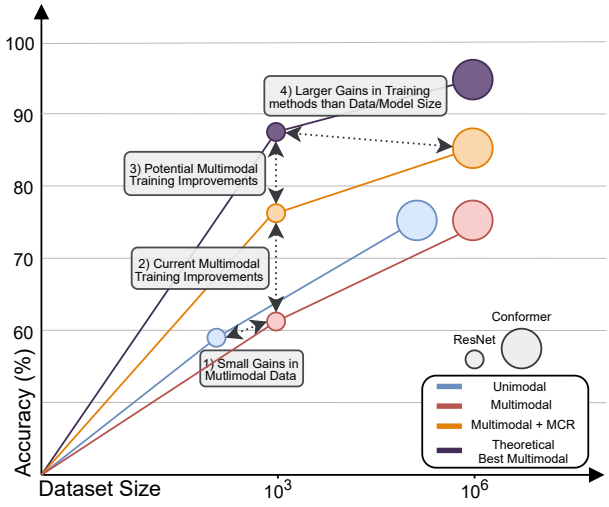



Figure 7.1: The figure illustrates performance trends for unimodal models, multimodal models without balancing methods, multimodal models using the current best method (MCR), and the theoretical best achievable performance. The results are shown on the CREMA-D dataset for two model scales, ResNet and Conformer. With dashed lines, we make four key observations: 1) Unimodal Baseline: Unimodal models often outperform multimodal baselines without balancing, questioning the value of extra modalities. 2) Current Balancing Impact: MCR improves multimodal performance, narrowing the gap to their potential. 3) Theoretical Gap: Significant room remains to reach the theoretical best multimodal performance. 4) Optimized Training: Better training objectives may achieve even larger gains in scaling models and data.

## 7.5 Closing Note & Final Thoughts

In this thesis, we explored critical challenges in multimodal machine learning, particularly in model architectures and training procedures. While there is a growing belief that these areas, which predominantly dominate academic research, are less significant compared to the scale of data and overparameterized models often prioritized by big tech, this work provides evidence to challenge that notion. While this may seem to apply to unimodal models, where scale often the last years emerges as the determining factor, this thesis demonstrates that for multimodal systems, effective training methodologies can lead to faster saturation of the performance gains than scale.

Current multimodal fusion models fail to fully exploit the available information. Our analysis shows that the theoretically optimal combination of unimodal and multimodal models achieves a smaller performance gap between simpler models (ResNets) and large pretrained models (Conformer). For CREMA-D and AVE, the error cases where all models fail are 12.4% and 18.5% for ResNets and 6.5% and 5.6% for Conformers, respectively. This suggests that effective multimodal training has the potential for significant improvements, even with smaller datasets, models and limited computational resources. It highlights that the gains in multimodal learning could be substantially steeper with well-designed fusion methodologies. In Figure 7.1, we illustrate these trend lines. By leveraging diverse datasets instead, we show how improved multimodal training can narrow the performance gap while significantly reducing the effort and cost associated with massive datasets. We remain optimistic that future endeavours will bridge this performance gap, delivering the message that ***diverse multimodal data are better than just more data*** if properly exploited. Finally, we are certain that multimodal learning will become a cornerstone of modern AI, and we are enthusiastic to witness how improved training methods, combined with large-scale data and models, could unlock unprecedented capabilities in understanding the multifaceted nature.

# Appendix A

# Appendix MLB

## A.1 Calibration Error

To estimate the calibration abilities of each balancing method and individual model, we measure the Expected Calibration Error (ECE) [64]. ECE measures the difference between average accuracy and average confidence for datapoints within a narrow confidence interval. We arrange predictions according to per-class confidence scores and categorize them into $K = 15$ bins $B_k$, each linked to a corresponding equally distant confidence interval. The lower the ECE, the better the model will be calibrated. With N being the total number of datapoints, ECE can be estimated as:

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \cdot \left|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\right|. \tag{A.1}$$

## A.2 Implementation

In this section, we provide a detailed account of our study's implementation steps and choices. These specifics are crucial for replicating our work and ensuring a fair comparison with previous methods.

### A.2.1 Balancing Hyperparameters

When tuning the hyperparameters of each method, we follow a consistent strategy. Most methods involve adjusting a few key hyperparameters. For example, AGM, PMR, and OGM focus on the $\alpha$ slope of the coefficients and the endpoint of the balancing process. MMCosine includes the magnitude increase scale $s$, while MSLR incorporates initial coefficients *init* used as priors and specifies the window of epochs for averaging performance $k_e$. MLB also has the $\alpha$ slope and introduces $\beta_{max}$, representing the maximum acceleration coefficient value. All hyperparameters are summarized in Table A.1. For AGM, OGM, and PMR, we did not observe any direct benefit from muting the balancing at a specific point, so it is not included in the table. For MLB, we noticed an improvement when $\beta_{max}$ was set higher than 1, indicating that acceleration was effective. However, values $\beta_{max} \geq 2$ did not show significant differences. We assume that in datasets with more severe overfitting on one of the modalities, this might not hold, and further acceleration might be needed.

We maintained uniform computational limits across all methods, conducting grid searches with equal points. Although we slightly favoured MSLR due to its multiple hyperparameters, this choice didn't significantly impact the results. Each method, backbone encoder, and fusion method underwent a separate search, leading to potentially differing hyperparameters in each experiment.

### A.2.2 Training Hyperparameters

For all datasets, we adopt a consistent strategy for defining training hyperparameters. Initially, we search for the learning rate and weight decay parameters when training each unimodal encoder individually with a linear classification head. Once we establish a suitable combination of these parameters, we apply them to the multimodal models. Interestingly, we find that models are not significantly sensitive to slight variations in these values. Consequently, we set our final values uniformly across all datasets: $(1e-3, 1e-4)$ for the ResNet encoders and $(5e-5, 1e-5)$ for Conformer encoders. We utilize the Adam optimizer [87] with default values for $(\beta_1, \beta_2)$ as $(0.9, 0.999)$ and $\epsilon$ set to $1e-07$. Batch sizes remain consistent across models using the same backbone encoder within each dataset but differ between the two encoder types. Specifically, we set the batch size to 32 for ResNet and adjusted it to 8 for Conformer models to accommodate our computational resources. Model validation occurs once at the end of each epoch. Training continues for a maximum of $1.5k$ epochs, with early stopping implemented at 300 epochs. Despite the large number of epochs allowed, we observed that nearly all runs converged before reaching 350 epochs. We employ a learning rate scheduler [109] based on cosine annealing

Table A.1: Hyperparameters for each dataset (CREMA-D, AVE, and UCF101), backbone encoder (ResNet and Conformer where applicable), balancing method (MSLR, MMCosine, OGM, PMR, AGM, MLB), and fusion method (Late-Linear, MLP, FiLM, Gated, and TF where applicable).

| Method | Fusion | Dataset | | | |
|---|---|---|---|---|---|
| | | CREMA-D | | AVE | UCF |
| | | ResNet | Conformer | ResNet | ResNet |
| MSLR | Late-Linear | $lr = [0.7, 1.3]$<br>$k_e = 20$ | $lr = [0.9, 1.1]$<br>$k_e = 20$ | $lr = [0.7, 1.3]$<br>$k_e = 10$ | $lr = [1.0, 1.0]$<br>$k_e = 5$ |
| MMCosine | Late-Linear | $s = 10.0$ | $s = 10$ | $s = 8$ | $s = 15$ |
| OGM | Late-Linear | $\alpha = 1.0$ | $\alpha = 1.0$ | $\alpha = 0.8$ | $\alpha = 1.0$ |
| PMR | Late-Linear | $\alpha = 2.0$ | $\alpha = 1.0$ | $\alpha = 1.5$ | $\alpha = 1.0$ |
| | MLP | - | $\alpha = 0.5$ | - | - |
| AGM | Late-Linear | $\alpha = 3.0$ | $\alpha = 3.0$ | $\alpha = 1.0$ | $\alpha = 1.5$ |
| | MLP | $\alpha - 1.0$ | $\alpha = 3.0$ | $\alpha = 2.0$ | $\alpha = 3.0$ |
| | FiLM | $\alpha = 1.5$ | $\alpha = 3.0$ | - | - |
| | Gated | $\alpha = 3.0$ | $\alpha = 1.5$ | - | - |
| | TF | $\alpha = 1.5$ | $\alpha = 1.5$ | - | - |
| MLB | Late-Linear | $\alpha = 5.0$ | $\alpha = 5.0$ | $\alpha = 2.0$ | $\alpha = 2.0$ |
| | MLP | $\alpha = 5.0$ | $\alpha = 3.0$ | $\alpha = 1.0$ | $\alpha = 4.0$ |
| | FiLM | $\alpha = 0.5$ | $\alpha = 4.0$ | - | - |
| | Gated | $\alpha = 5.0$ | $\alpha = 5.0$ | - | - |
| | TF | $\alpha = 5.0$ | $\alpha = 2.0$ | - | - |

with warm restarts, where $T_0 = 4$ and $T_{mult} = 2$, supplemented by a constant warming-up stage of 3 epochs. Finally, while the use of gradient clipping [115] could potentially aid in preventing exploding gradients during balancing, we opt not to utilize it in our models to ensure a fair comparison with previous methods, except for AGM [96], where we strictly follow their implementation.

### A.2.3 Dataset Splits

It is important to note that some of the previous methods introduced in the paper [96, 129, 201, 43], based on the official code they shared, seem to have used the test set as the validation set on CREMA-D, tuning model parameters directly on the test set. Additionally, the same actors were included in different sets, leading to further overfitting. Therefore, the observed performance difference in our results compared to the initially reported results of these papers is attributed to the fact that we validate the models with a separate validation set, comprising

Table A.2: Accuracy presented in table format for each dataset (CREMA-D, AVE, and UCF101) and backbone encoder (ResNet and Conformer where applicable), balancing method (MMCosine, MSLR, OGM, PMR, AGM, MLB) using the Late-Linear fusion method.

| Method | CREMA-D | | AVE | UCF |
|---|---|---|---|---|
| | ResNet | Conformer | ResNet | ResNet |
| Video | $55.4_{\pm 1.3}$ | $69.4_{\pm 1.6}$ | $62.6_{\pm 0.6}$ | $30.3_{\pm 0.9}$ |
| Audio | $60.6_{\pm 1.7}$ | $76.8_{\pm 1.2}$ | $45.7_{\pm 0.9}$ | $38.3_{\pm 0.5}$ |
| Joint Training | $62.6_{\pm 0.8}$ | $75.1_{\pm 1.0}$ | $66.7_{\pm 0.9}$ | $47.7_{\pm 0.9}$ |
| MMCosine | $58.8_{\pm 0.9}$ | $73.5_{\pm 1.2}$ | $66.7_{\pm 0.5}$ | $46.9_{\pm 2.1}$ |
| MSLR | $56.5_{\pm 1.3}$ | $77.1_{\pm 1.4}$ | $66.8_{\pm 1.0}$ | $47.9_{\pm 2.2}$ |
| OGM | $65.6_{\pm 2.2}$ | $82.4_{\pm 0.6}$ | $67.9_{\pm 0.3}$ | $51.8_{\pm 1.1}$ |
| PMR | $53.7_{\pm 1.3}$ | $80.5_{\pm 0.1}$ | $37.4_{\pm 4.3}$ | $40.9_{\pm 1.2}$ |
| AGM | $69.3_{\pm 0.8}$ | $78.5_{\pm 1.5}$ | $65.8_{\pm 1.6}$ | $50.5_{\pm 0.9}$ |
| Multi-Loss | $69.2_{\pm 1.0}$ | $82.6_{\pm 0.5}$ | $70.1_{\pm 0.5}$ | $51.1_{\pm 1.0}$ |
| MLB | $\mathbf{71.4}_{\pm \mathbf{1.1}}$ | $\mathbf{85.2}_{\pm \mathbf{0.5}}$ | $\mathbf{70.6}_{\pm \mathbf{0.8}}$ | $\mathbf{52.0}_{\pm \mathbf{1.3}}$ |

around 10% of the total data. We report results on the test set using 3-fold cross-validation with the splits provided by Goncalves et al. [56].

## A.3 Tabular Presentation of Results

The results from Figure 4.3 are also provided in table format to facilitate direct access and use by future research. Table A.2 presents the accuracy of each balancing method across various datasets and backbone encoders using Late-Linear fusion. Additionally, Table A.3 details the findings from Figure 4.5, illustrating the performance of different fusion methods on the CREMA-D dataset.

Table A.3: Accuracy presented in table format for the different fusion methods (MLP, FiLM, Gated, TF) on the CREMA-D dataset.

| Method | MLP | | FiLM | | Gated | | TF | |
|---|---|---|---|---|---|---|---|---|
| | ResNet | Conformer | ResNet | Conformer | ResNet | Conformer | ResNet | Conformer |
| Joint Training | $62.6_{\pm 0.8}$ | $75.1_{\pm 1.0}$ | $64.8_{\pm 0.6}$ | $73.8_{\pm 0.9}$ | $59.1_{\pm 2.7}$ | $71.7_{\pm 1.4}$ | $58.7_{\pm 1.3}$ | $71.0_{\pm 1.4}$ |
| AGM | $69.3_{\pm 0.8}$ | $78.5_{\pm 1.5}$ | $57.8_{\pm 0.8}$ | $68.9_{\pm 0.3}$ | $55.7_{\pm 2.7}$ | $69.0_{\pm 0.4}$ | $54.0_{\pm 2.6}$ | $70.9_{\pm 0.4}$ |
| Multi-Loss | $69.2_{\pm 1.0}$ | $82.6_{\pm 0.5}$ | $68.3_{\pm 1.7}$ | $84.7_{\pm 0.4}$ | $70.7_{\pm 1.4}$ | $83.3_{\pm 0.6}$ | $\mathbf{69.7}_{\pm \mathbf{2.0}}$ | $83.4_{\pm 0.9}$ |
| MLB | $\mathbf{71.4}_{\pm \mathbf{1.1}}$ | $\mathbf{85.2}_{\pm \mathbf{0.5}}$ | $\mathbf{72.2}_{\pm \mathbf{0.3}}$ | $\mathbf{84.4}_{\pm \mathbf{0.3}}$ | $\mathbf{71.1}_{\pm \mathbf{1.0}}$ | $\mathbf{84.2}_{\pm \mathbf{0.8}}$ | $\mathbf{69.7}_{\pm \mathbf{0.6}}$ | $\mathbf{85.5}_{\pm \mathbf{0.5}}$ |

# Appendix B

# Appendix MCR

## B.1 Experimental Setting

### B.1.1 Datasets

**CREMA-D [25]:** is an emotion recognition dataset with audio and video modalities. It features a diverse group of 91 actors, covering a wide range of ages, ethnicities, and genders. To ensure consistency, each actor is positioned at an equal distance from the camera, expressing six distinct emotions: Happy, Sad, Anger, Fear, Disgust, and Neutral. In alignment with methodologies from prior studies [129, 96, 43], video frames are sampled at 1 fps, selecting 3 consecutive frames, while audio segments are sampled at 22 kHz, capturing 3 seconds that correspond with the chosen video frames. Audio analysis utilizes a window size of 512 and a step size of 353 samples for Short-Time Fourier Transform (STFT), creating log-Mel spectrograms. For advanced models, our methodology aligns with Goncalves et al. [57], incorporating audio signals sampled at 16 kHz and utilizing pre-calculated facial features. Unlike previous approaches [129, 96, 43], our dataset division follows Goncalves et al. [57], excluding actor overlap between training, validation, and test sets. We report standard deviation across folds for consistency.

**AVE [173]:** contains 4143 videos across 28 event categories with a wide range such as frying food or playing guitar, each with temporally labelled audio-visual events of at least 2 seconds. Following [43], video segments where the event occurs are sampled at 1 fps for 4 frames, with audio resampled at 16 kHz using CREMA-D's STFT settings. AVE provides predefined training, validation, and

test splits. Our standard deviation is derived from three random seeds on the same test set.

**UCF101 [161]:** features real-life action videos from YouTube in 101 action categories, expanding on UCF50. We select samples that include both video and audio modalities, narrowing our focus to 51 action categories. Data preparation mirrors that of the AVE dataset. For model evaluation, we utilize the 3-fold split offered [161], reporting the standard deviation across these folds.

**Something-Something (V2)[58]:** presents 220,847 video clips where individuals execute 174 distinct, object-agnostic hand actions, spanning a wide range of simple hand movements without reliance on specific objects. This extensive collection markedly exceeds the data volume of prior datasets. In alignment with Radevski et al. [141], we integrate Optical Flow (OF) as the additional modality for capturing dynamic motions. We sample videos at 1 fps, retaining 16 frames per sample. Our data preparation adheres to the pipeline outlined in Radevski et al. [141]. Results are reported based on a single run.

**CMU-MOSI[206] and -MOSEI[207]:** datasets serve as benchmarks for multimodal sentiment and emotion analysis. MOSI comprises 2,199 video clips with video, audio and text modalities. MOSEI expands on this with around 10x more YouTube movie review clips from 1000 different speakers. Both datasets are annotated with sentiment scores (-3 to 3) and following previous works [175] we employ metrics such as 7-class accuracy, binary accuracy, F1 score, mean absolute error, and correlation with human annotations for model evaluation. We use for both datasets the aligned versions following [103].

### B.1.2 Models: Backbone Unimodal Encoders

In line with previous research [96, 129, 43, 201], our initial experiments adopt ResNet-18 [68] as the unimodal encoder for handling both video and audio modalities in the CREMA-D, AVE, and UCF datasets. These models are randomly initialized and incorporate adaptive pooling to accommodate diverse input dimensions. For the CMU-MOSEI dataset, we exploit a 5-layer Transformer [180] similar to previous works [102, 103].

We extend our investigation to include a larger, pre-trained set of unimodal encoders that are optimally suited for each specific modality. This selection process is designed to rigorously assess whether state-of-the-art models exhibit susceptibility to the same phenomena under investigation. We exploit these encoders on the CREMA-D dataset. Following [56] on CREMA-D, we deploy the first 12 layers of the Wav2Vec2 [12, 194] model with self-supervised pretrained weights for speech recognition, allowing the Wav2Vec2 model to be finetuned.

For the video modality we extract the facing bounding boxes with the multi-task cascaded convolutional neural network (MTCNN) face detection algorithm [209] and afterwards the facial features of every available frame exploiting EfficientNet-B2 [171] as a frozen feature descriptor. The extracted audio features and pre-calculated facial features are further refined using a 5-layer Conformer [62], initialized from scratch. Although the model includes additional components, we refer to it as 'Conformer' for simplicity.

For the AVE dataset, we use a similar architecture to CREMA-D, where each branch utilizes an advanced, pretrained model aligned with AVE data, followed by a 5-layer Conformer. For the video branch, we use the ViViT model [8] pretrained on Kinetics [82], and for the audio branch, the HuBERT [70] model pretrained on Audioset [54]. Ensure the audio pretraining included non-speech data was important for the pretraining to be beneficial. We also refer to this model as "Conformer," although it incorporates different large pretrained models in this instance.

In the case of the Something-Something dataset, our methodology builds upon the insights presented by Radevski et al. [141], which highlights the importance of modality-specific processing in multimodal tasks. For both video and optical flow data, we adopt the Swin Transformer [107] as the backbone of each modality. This state-of-the-art architecture excels in capturing hierarchical and spatiotemporal features through its shifted window attention mechanism.

### B.1.3 Experimental Details

In this section, we outline the necessary details to reproduce our experiments. Across all datasets, we follow a consistent procedure: we first determine an appropriate learning rate (lr) for the unimodal models by testing several candidates until finding one that works across both modalities. While this step could be avoided by exploiting parameter-specific learning rates, we expected stability implications which we aimed to avoid. For all the experiments of the same dataset/model pair, we use the same hyperparameters, except when fine-tuning pretrained encoders, where we apply a learning rate scaled one magnitude lower.

All models are optimized using Adam [87] with a cosine learning rate scheduler and a steady warm-up phase, except for the Something-Something dataset, where we use Adaw [108]. Early stopping is applied for all models, with maximum epochs set to 100 for ResNet and Transformer models, 50 for Conformer models, and 30 epochs in total for Swin Transformers without early stopping. Batch sizes are adjusted based on computational resources, with ResNets and Transformers both using a batch size of 32, Conformers using 8, and Swin-TF using 16.

These settings ensure balanced performance and efficient training across all experiments. We use different learning rates ($lr$) and weight decay ($wd$) values across experiments, tailored to each dataset and model. For ResNet models, we use $lr = 1e{-}3$ and $wd = 1e{-}4$ for CREMA-D, while both AVE and UCF use $lr = 1e{-}4$ and $wd = 1e{-}4$. For Transformer models, including MOSI and MOSEI on two and three modalities, the hyperparameters are consistent $lr = 1e{-}4$ and $wd = 1e{-}4$. Similarly, for Conformer models, we set $lr = 5e{-}5$ and $wd = 5e{-}6$ for CREMA-D, while AVE uses $lr = 1e{-}4$ and $wd = 1e{-}4$. Finally, for Swin-TF models trained on the Something-Something dataset, we configure $lr = 1e{-}4$ and $wd = 0.02$.

Each of the previous methods includes its own set of hyperparameters, typically just one, with some exceptions such as MSLR, which requires additional parameters. For each dataset/model combination, we conduct a brief hyperparameter search, ensuring an equitable number of trials across methods. Due to the extensive list of hyperparameters, we provide detailed configurations for each experiment in our GitHub repository.

### B.1.4 MCR on M modalities

In this section, we provide the analysis of MCR for M number of modalities. In that case, the total mutual information $I(X_1, \ldots, X_M; Y)$ can be decomposed into contributions from individual modalities and their subsets as:

$$I(X_1, .., X_M; Y) = \sum_{\mathcal{S} \subseteq \{X_1, .., X_M\}, \mathcal{S} \neq \emptyset} I(\mathcal{S}; Y \mid \{X_1, \ldots, X_M\} \setminus \mathcal{S}) \tag{B.1}$$

$$+ I(X_1, .., X_M) - I(X_1, .., X_M \mid Y), \tag{B.2}$$

where $\mathcal{S} \subseteq \{X_1, .., X_M\}$ represents a subset of all modalities, excluding the empty set ($\mathcal{S} \neq \emptyset$). The term $\{X_1, .., X_M\} \setminus \mathcal{S}$ denotes the complement of $\mathcal{S}$, capturing the set of modalities not included in $\mathcal{S}$. The mutual information $I(\mathcal{S}; Y \mid \{X_1, .., X_M\} \setminus \mathcal{S})$ quantifies the information shared between the subset $\mathcal{S}$ and the target variable $Y$, conditioned on the remaining modalities. This formulation ensures that all modalities, along with their combinations, are accounted for in the summation. It comprehensively captures interactions at every granularity, from individual modalities ($|\mathcal{S}| = 1$) to the full set of modalities ($|\mathcal{S}| = M$).

While we experimented with reducing the number of terms by considering only the cases where $|\mathcal{S}| = 1$, we observed a slight performance improvement when including all terms for three modalities. However, as the number of modalities

increases, it might be beneficial to sub-select and exclude certain terms to mitigate the computational burden and prevent an overflow of terms.

### B.1.5 Computational Speed and Memory Analysis

The computational load imposed by any sample that requires an additional pass can be divided into the encoders $f^{(1)}, f^{(2)}$ and the fusion network $f_c$. The encoders have a computational cost of $cost_{enc}$ per sample, and the fusion network has a cost of $cost_c$ per sample. Thus, the overall computational complexity is $\mathcal{O}(M * N * (cost_{enc} + cost_c))$, where $M$ is the times we draw noisy samples and $N$ is the batch size. If we now use permutation samples which can be directly drawn from the latent space, the additional computational complexity is reduced to $\mathcal{O}(N * cost_{enc} + M * N * cost_c))$. In most state-of-the-art models each modality encoder is significantly larger than the fusion network resulting in $cost_{enc} << cost_c$. In such networks, permutations can have almost negligible additional computations. The memory footprint follows a similar pattern.

### B.1.6 Comparing Perturbation methods

In this work, we explored three types of perturbation methods to analyze their impact on the performance of MCR: noisy perturbations, zero-masking, and permutations. Each method was applied in different spaces (input space and latent space) or within the batch structure to determine how effectively MCR can leverage these perturbations to enhance multimodal learning.

Table B.1 presents the results of an ablation study comparing the performance of MCR under these perturbation techniques across multiple datasets: CREMA-D, AVE, UCF, MOSEI, and MOSI. The methods include:

1. Noise in the Input Space: Adding noise directly to the input features of each modality, simulating realistic data corruption. For its implementation, we follow Gat et al. [52].

2. Shapley Input-Space Perturbations: Following the approach of Li et al. [96], Shapley zero-induced values are used to determine the importance of input modalities.

3. Noise in the Latent Space: Applying noise to the latent representations, encouraging robustness at the feature extraction level.

4. Within-Batch Permutations in the Latent Space: Permuting data points within the batch to disrupt alignment.

Table B.1: Ablation study comparing different perturbation methods for MCR across multiple datasets. The table shows the performance of MCR when combined with various perturbation techniques, including noise in the input and latent space, Shapley values in the input space, and within-batch permutations.

| Method | Datasets | | | | |
|---|---|---|---|---|---|
| | CREMA-D | AVE | UCF | MOSEI | MOSI |
| MCR with Noise Input-Space | $75.3_{\pm 2.9}$ | $72.1_{\pm 1.1}$ | $54.6_{\pm 0.8}$ | 80.9 | 76.1 |
| MCR with Shapley Input-Space | $73.6_{\pm 1.5}$ | $72.6_{\pm 0.9}$ | $\mathbf{55.5_{\pm 0.6}}$ | 81.2 | **76.2** |
| MCR with Noise Latent-Space | $73.6_{\pm 1.1}$ | $72.6_{\pm 0.4}$ | $54.5_{\pm 0.7}$ | **81.3** | 75.5 |
| MCR with Permutations | $\mathbf{76.1_{\pm 1.1}}$ | $\mathbf{73.3_{\pm 0.5}}$ | $55.2_{\pm 1.8}$ | 81.2 | **76.2** |

We observe that Shapley-based input-space perturbations show competitive performance, particularly in datasets like UCF, MOSI, and MOSEI, while noise-based methods (both input and latent spaces) achieve reasonable performance, they consistently underperform other techniques. While input-space perturbations could be a viable option, they significantly increase computational complexity, as they require an additional forward pass through the typically large unimodal encoders for each sample. This limitation, which we analyze in Appendix B.1.5, makes them less favourable as a practical solution. Finally, these findings support the choice of permutations as the preferred perturbation method, while suggesting that further exploration of alternative strategies could potentially lead to even greater improvements.

## B.2 Proof of Supervised Contrastive Loss as Lower Bound

We consider the supervised contrastive loss with $\psi$ being the critic function and we rewrite it as follows:

$$\mathcal{L}_{\text{Con}}(X_1, X_2) = \sum_{i \in \mathbb{I}} \frac{-1}{|\mathcal{P}_i|} \sum_{k \in \mathcal{P}_i} \left[ \log \frac{\psi(x_{1_i}, x_{2_k})}{\sum_{j \in \mathbb{I}} \psi(x_{1_i}, x_{2_j})} \right] \tag{B.3}$$

where $\mathcal{P}_i = \{p \in \mathbb{I} \mid y_p = y_i\}$. In supervised contrastive learning, the presence of multiple positive samples turns this into a multi-label problem, unlike traditional NCE methods [124], which typically assume only one positive sample. By taking the version of supervised contrastive learning with the expectation over the positives outside of the log, we can interpret each classification as an average of classifiers, with each classifier focusing on identifying one of the positive samples.

For each positive sample $p \sim \mathcal{P}i$, we aim to derive the optimal probability of correctly identifying that point, denoted $d = p$. This is done by sampling the point from the conditional distribution $p(x2_p \mid x_{1_i}, y_i)$ while sampling the remaining points from the proposal distribution $p(x_{2_l})$. This approach mirrors the technique used in InfoNCE [124] and leads to the following derivation:

$$p(d = p|X_2, x_{1_i}, y_i) = \frac{p(x_{2_p} \mid x_{1_i}, y_i) \prod_{l \in \mathcal{I}, l \neq p} p(x_{2_l})}{\sum_{j \in \mathcal{I}} p(x_{2_j} \mid x_{1_i}, y_i) \prod_{l \in \mathcal{I}, l \neq j} p(x_{2_l})} \tag{B.4}$$

$$= \frac{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}}{\sum_{j \in \mathcal{I}} \frac{p(x_{2_j}|x_{1_i}, y_i)}{p(x_{2_j})}} \tag{B.5}$$

The optimal value for the critic function $\psi$ in Equation B.5 is proportional to $\psi \propto \frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}$. The MI between the variables can be estimated as follows:

$$L_{\text{Con}}^{Opt} = - \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[ \mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[ \frac{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})}}{\frac{p(x_{2_p}|x_{1_i}, y_i)}{p(x_{2_p})} + \sum_{j \in \mathcal{I}, j \neq i} \frac{p(x_{2_j}|x_{1_i}, y_i)}{p(x_{2_j})}} \right] \right] \tag{B.6}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[ \mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[ 1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} \sum_{j \in \mathcal{I}, j \neq i} \frac{p(x_{2_j} \mid x_{1_i}, y_i)}{p(x_{2_j})} \right] \right] \tag{B.7}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[ \mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[ 1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} (N-1) \mathop{\mathbb{E}}_{j \in \mathcal{I}} \frac{p(x_{2_j} \mid x_{1_i}, y_i)}{p(x_{2_j})} \right] \right] \tag{B.8}$$

$$= \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[ \mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[ 1 + \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} (N-1) \right] \right] \tag{B.9}$$

$$\geq \mathop{\mathbb{E}}_{i \sim \mathcal{I}} \left[ \mathop{\mathbb{E}}_{i \sim \mathcal{P}_i} \log \left[ \frac{p(x_{2_p})}{p(x_{2_p} \mid x_{1_i}, y_i)} N \right] \right] \tag{B.10}$$

$$= \log N - I(X_2; X_1, Y), \text{ using MI properties [116, Chapter 6.3.4]} \tag{B.11}$$

$$= \log N - I(X_2; Y|X_1) - I(X_2; X_1) \tag{B.12}$$

Therefore, by taking both sides of the contrastive loss to predict $X_2$ from $X_1$ and $X_1$ from $X_2$ we derive to $I(X_2; Y|X_1) + I(X_1; Y|X_2) + 2 \cdot I(X_2; X_1) \geq \log N - L_{\text{Con}}^{Opt}$. This trivially also holds for other $\psi$ that obtain a worse(higher)

$L_{\text{Con}}$. Similarly to InfoNCE, the bound becomes more accurate as N increases, while due to the term $\frac{1}{|\mathcal{P}_i|}$ it is not affected by the number of positive pairs.

# Bibliography


[1] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2636–2645.

[2] Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T. K., Hori, C., Anderson, P., et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 7558–7567.

[3] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2024), NIPS '22, Curran Associates Inc.

[4] Alickovic, E., and Subasi, A. Ensemble svm method for automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement 67*, 6 (2018), 1258–1265.

[5] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, June 2018), IEEE Computer Society, pp. 6077–6086.

[6] Anil, C., Lucas, J., and Grosse, R. Sorting out lipschitz function approximation. In *International Conference on Machine Learning* (2019), PMLR, pp. 291–301.

[7] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)* (2015).

[8] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 6836–6846.

[9] Aruleba, R. T., Adekiya, T. A., Ayawei, N., Obaido, G., Aruleba, K., Mienye, I. D., Aruleba, I., and Ogbuokiri, B. Covid-19 diagnosis: A review of rapid antigen, rt-pcr and artificial intelligence methods. *Bioengineering 9*, 4 (2022), 153.

[10] Atrey, P., Hossain, M., El Saddik, A., and Kankanhalli, M. Multimodal fusion for multimedia analysis: A survey. *Multimedia Syst. 16* (11 2010), 345–379.

[11] Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[12] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems 33* (2020), 12449–12460.

[13] Bajaj, V., and Pachori, R. B. Automatic classification of sleep stages based on the time-frequency image of eeg signals. *Computer methods and programs in biomedicine 112*, 3 (2013), 320–328.

[14] Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A. G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., and Goldblum, M. A cookbook of self-supervised learning, 2023.

[15] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence 41*, 2 (2018), 423–443.

[16] Banville, H., Wood, S. U., Aimone, C., Engemann, D.-A., and Gramfort, A. Robust learning from corrupted eeg with dynamic spatial filtering. *NeuroImage 251* (2022), 118994.

[17] Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C., Vaughn, B. V., et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* (2012).

[18] Bhagubai, M., Vandecasteele, K., Swinnen, L., Macea, J., Chatzichristos, C., De Vos, M., and Van Paesschen, W. The power of ecg in semi-automated seizure detection in addition to two-channel behind-the-ear eeg. *Bioengineering 10*, 4 (2023).

[19] Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., and Robbins, K. A. The prep pipeline: standardized preprocessing for large-scale eeg analysis. *Frontiers in neuroinformatics 9* (2015), 16.

[20] Bischke, B., Helber, P., Koenig, F., Borth, D., and Dengel, A. Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)* (2018), IEEE, pp. 1–6.

[21] Breiman, L. Bagging predictors. *Machine learning 24* (1996), 123–140.

[22] Breiman, L. Random forests. *Machine learning 45* (2001), 5–32.

[23] Brzecka, A., Leszek, J., Ashraf, G. M., Ejma, M., Ávila-Rodriguez, M. F., Yarla, N. S., Tarasov, V. V., Chubarev, V. N., Samsonova, A. N., Barreto, G. E., et al. Sleep disorders associated with alzheimer's disease: a perspective. *Frontiers in neuroscience 12* (2018), 330.

[24] Buck, S. F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological) 22*, 2 (1960), 302–306.

[25] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing 5*, 4 (2014), 377–390.

[26] Cao, W., Wang, D., Li, J., Zhou, H., Li, L., and Li, Y. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems 31* (2018).

[27] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting

cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[28] Carskadon, M. A., Dement, W. C., et al. Normal human sleep: an overview. *Principles and practice of sleep medicine 4*, 1 (2005), 13–23.

[29] Chatzichristos, C., Van Eyndhoven, S., Kofidis, E., and Van Huffel, S. Chapter 10 - coupled tensor decompositions for data fusion. In *Tensors for Data Processing*, Y. Liu, Ed. Academic Press, 2022, pp. 341–370.

[30] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.

[31] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning (ICML)* (2020), 1597–1607.

[32] Cho, J., Lu, J., Schwenk, D., Hajishirzi, H., and Kembhavi, A. X-lxmert: Paint, caption and answer questions with multi-modal transformers. *arXiv preprint arXiv:2009.11278* (2020).

[33] Chun, S., Oh, S. J., De Rezende, R. S., Kalantidis, Y., and Larlus, D. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 8415–8424.

[34] Comas, A., Zhang, C., Feric, Z., Camps, O., and Yu, R. Learning disentangled representations of videos with missing data. *Advances in Neural Information Processing Systems 33* (2020), 3625–3635.

[35] De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. C. Modulating early visual processing by language. *Advances in Neural Information Processing Systems 30* (2017).

[36] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[37] Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (2000), Springer, pp. 1–15.

[38] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[39] Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., Yuan, Y., and Zhao, H. On uni-modal feature learning in supervised multi-modal learning. *arXiv preprint arXiv:2305.01233* (2023).

[40] Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C.-K., Li, X., and Guan, C. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 29* (2021), 809–818.

[41] et al., A. D. The llama 3 herd of models, 2024.

[42] et al., H. T. Llama 2: Open foundation and fine-tuned chat models, 2023.

[43] Fan, Y., Xu, W., Wang, H., Wang, J., and Guo, S. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 20029–20038.

[44] Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., and Soatto, S. Multi-Modal Hallucination Control by Visual Information Grounding . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, June 2024), IEEE Computer Society, pp. 14303–14312.

[45] Fischer, I. The conditional entropy bottleneck. *Entropy 22*, 9 (2020), 999.

[46] Fischer, I., and Alemi, A. A. Ceb improves model robustness. *Entropy 22*, 10 (2020), 1081.

[47] Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., and Dickhaus, H. Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier. *Computer methods and programs in biomedicine 108*, 1 (2012), 10–19.

[48] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[49] Fujimori, N., Endo, R., Kawai, Y., and Mochizuki, T. Modality-specific learning rate control for multimodal classification. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5* (2020), Springer, pp. 412–422.

[50] Gallo, I., Calefati, A., and Nawaz, S. Multimodal classification fusion in real-world scenarios. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)* (2017), vol. 5, IEEE, pp. 36–41.

[51] Gat, I., Schwartz, I., and Schwing, A. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems 34* (2021), 21630–21643.

[52] Gat, I., Schwartz, I., Schwing, A., and Hazan, T. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems 33* (2020), 3197–3208.

[53] Gemini Team, R. A. e. a. Gemini: A family of highly capable multimodal models, 2024.

[54] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2017), IEEE, pp. 776–780.

[55] Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., and Abbeel, P. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204* (2022).

[56] Goncalves, L., Leem, S.-G., Lin, W.-C., Sisman, B., and Busso, C. Versatile audio-visual learning for handling single and multi modalities in emotion regression and classification tasks. *arXiv preprint arXiv:2305.07216* (2023).

[57] Goncalves, L., Leem, S.-G., Lin, W.-C., Sisman, B., and Busso, C. Versatile audiovisual learning for handling single and multi modalities in emotion regression and classification tasks. *ArXiv e-prints (arXiv:2305.07216 )* (May 2023), 1–14.

[58] Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. The" something something"

video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 5842–5850.

[59] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6904–6913.

[60] Graham, C. A., Shamkhalichenar, H., Browning, V. E., Byrd, V. J., Liu, Y., Gutierrez-Wing, M. T., Novelo, N., Choi, J.-W., and Tiersch, T. R. A practical evaluation of machine learning for classification of ultrasound images of ovarian development in channel catfish (ictalurus punctatus). *Aquaculture 552* (2022), 738039.

[61] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS) 33* (2020), 21271–21284.

[62] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).

[63] Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models, 2024.

[64] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning* (2017), PMLR, pp. 1321–1330.

[65] Hansen, L. K., and Salamon, P. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence 12*, 10 (1990), 993–1001.

[66] Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision (ACCV)* (November 2016).

[67] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 9729–9738.

[68] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

[69] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1026–1034.

[70] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing 29* (2021), 3451–3460.

[71] Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141.

[72] Hu, R., Singh, A., Darrell, T., and Rohrbach, M. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 9992–10002.

[73] Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). *arXiv preprint arXiv:2203.12221* (2022).

[74] Iber, C. The aasm manual for the scoring of sleep and associated events: rules, terminology, and technical specification. *(No Title)* (2007).

[75] Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., and Gramfort, A. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage 159* (2017), 417–429.

[76] Ji, B., Zhang, T., Zou, Y., Hu, B., and Shen, S. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), pp. 6755–6764.

[77] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (2021), PMLR, pp. 4904–4916.

[78] Jia, Z., Lin, Y., Wang, J., Zhou, R., Ning, X., He, Y., and Zhao, Y. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *IJCAI* (2020), pp. 1324–1330.

[79] Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., Horng, S., Li, M. D., Shield, W. R., et al. Mimic-cxr: A large publicly available dataset of labeled chest radiographs. *Nature Scientific Data 6*, 1 (2019), 1–8.

[80] Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., and Uszkoreit, J. One model to learn them all, 2017.

[81] Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., Mack, S., et al. *Principles of neural science*, vol. 4. McGraw-hill New York, 2000.

[82] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[83] Khan, A. T., Cao, X., Li, S., Katsikis, V. N., Brajevic, I., and Stanimirovic, P. S. Fraud detection in publicly traded us firms using beetle antennae search: A machine learning approach. *Expert Systems with Applications 191* (2022), 116148.

[84] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems 33* (2020), 18661–18673.

[85] Kiela, D., Grave, E., Joulin, A., and Mikolov, T. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence* (2018), vol. 32.

[86] Kiela, D., Wang, C., and Cho, K. Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, Oct.-Nov. 2018), E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Association for Computational Linguistics, pp. 1466–1477.

[87] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[88] Kini, J., Fleischer, S., Dave, I., and Shah, M. Ensemble modeling for multimodal visual action recognition. *arXiv preprint arXiv:2308.05430* (2023).

[89] Kontras, K., Chatzichristos, C., Blaschko, M. B., and Vos, M. D. Improving multimodal learning with multi-loss gradient modulation. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024* (2024), BMVA.

[90] Kontras, K., Chatzichristos, C., Phan, H., Suykens, J., and De Vos, M. Core-sleep: A multimodal fusion framework for time series robust to imperfect modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024).

[91] Kontras, K., Strypsteen, T., Chatzichristos, C., Liang, P. P., Blaschko, M., and Vos, M. D. Multimodal fusion balancing through game-theoretic regularization, 2024.

[92] Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems 30* (2017).

[93] Lee, S., Yu, Y., Back, S., Seo, H., and Lee, K. Sleepyco: Automatic sleep scoring with feature pyramid and contrastive learning. *arXiv preprint arXiv:2209.09452* (2022).

[94] Lee, W., Lee, J., and Kim, Y. Contextual imputation with missing sequence of eeg signals using generative adversarial networks. *IEEE Access 9* (2021), 151753–151765.

[95] Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 07 (Apr. 2020), 11336–11344.

[96] Li, H., Li, X., Hu, P., Lei, Y., Li, C., and Zhou, Y. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 22214–22224.

[97] Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (2023), ICML'23, JMLR.org.

[98] Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086* (2022).

[99] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems 34* (2021), 9694–9705.

[100] Li, Y., Wong, K. M., and de Bruin, H. Electroencephalogram signals classification for sleep-state decision–a riemannian geometry approach. *IET signal processing 6*, 4 (2012), 288–299.

[101] Liang, P. P., Cheng, Y., Fan, X., Ling, C. K., Nie, S., Chen, R., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., Salakhutdinov, R. R., and Morency, L.-P. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems* (2023), A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Curran Associates, Inc., pp. 27351–27393.

[102] Liang, P. P., Deng, Z., Ma, M. Q., Zou, J. Y., Morency, L.-P., and Salakhutdinov, R. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems 36* (2024).

[103] Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M. A., Zhu, Y., et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502* (2021).

[104] Liang, P. P., Zadeh, A., and Morency, L.-P. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys 56*, 10 (2024), 1–42.

[105] Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory 37*, 1 (1991), 145–151.

[106] Liu, Y., Huang, G., Zhang, Z., and Van Gool, L. Openkitchen: Towards open-world kitchen activity recognition with massive real-world datasets. *arXiv preprint arXiv:2010.11849* (2020).

[107] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 10012–10022.

[108] Loshchilov, I. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[109] Loshchilov, I., and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[110] Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc.

[111] Lu, J., Yang, J., Batra, D., and Parikh, D. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2016), NIPS'16, Curran Associates Inc., p. 289–297.

[112] Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems 30* (2017).

[113] Ma, M., Ren, J., Zhao, L., Testuggine, D., and Peng, X. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18177–18186.

[114] Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 2630–2640.

[115] Mikolov, T., et al. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April 80*, 26 (2012).

[116] Murphy, K. P. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[117] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems 34* (2021), 14200–14213.

[118] Nair, V., and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), pp. 807–814.

[119] Nash, J. F. Non-cooperative games. In *The Foundations of Price Theory Vol 4*. Routledge, 2024, pp. 329–340.

[120] Ng, A. Y. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning* (2004), p. 78.

[121] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. Multimodal deep learning. In *ICML* (2011).

[122] Niedermeyer, E. *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2011.

[123] Nolan, H., Whelan, R., and Reilly, R. B. Faster: fully automated statistical thresholding for eeg artifact rejection. *Journal of neuroscience methods 192*, 1 (2010), 152–162.

[124] Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[125] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (2022), vol. 35, pp. 27730–27744.

[126] Pachal, S., and Achar, A. Sequence prediction under missing data : An rnn approach without imputation, 2022.

[127] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *NIPS-W* (2017).

[128] Patel, A. K., Reddy, V., and Araujo, J. F. Physiology, sleep stages. In *StatPearls [Internet]*. StatPearls Publishing, 2022.

[129] Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 8238–8247.

[130] Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence* (2018), vol. 32.

[131] Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., and Igel, C. U-sleep: resilient high-frequency sleep staging. *NPJ digital medicine* (2021), 1–12.

[132] Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 27*, 3 (2019), 400–410.

[133] Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., and De Vos, M. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[134] Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., and De Vos, M. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 9 (2021), 5903–5915.

[135] Phan, H., Lorenzen, K. P., Heremans, E., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., Baumert, M., Mikkelsen, K., and De Vos, M. L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging, 2023.

[136] Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., and De Vos, M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering 69*, 8 (2022), 2456–2467.

[137] Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042* (2018).

[138] Polikar, R. Ensemble learning. *Ensemble machine learning: Methods and applications* (2012), 1–34.

[139] Pradeepkumar, J., Anandakumar, M., Kugathasan, V., Suntharalingham, D., Kappel, S. L., De Silva, A. C., and Edussooriya, C. U. Towards interpretable sleep stage classification using cross-modal transformers. *arXiv preprint arXiv:2208.06991* (2022).

[140] Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., et al. The sleep heart health study: design, rationale, and methods. *Sleep* (1997), 1077–1085.

[141] Radevski, G., Grujicic, D., Blaschko, M., Moens, M.-F., and Tuytelaars, T. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 5213–5224.

[142] Radevski, G., Moens, M.-F., and Tuytelaars, T. Revisiting spatio-temporal layouts for compositional action recognition. *arXiv preprint arXiv:2111.01936* (2021).

[143] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763.

[144] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[145] Ramachandram, D., and Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine 34*, 6 (2017), 96–108.

[146] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International conference on machine learning* (2021), Pmlr, pp. 8821–8831.

[147] Rechtschaffen, A. A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. *Brain information service* (1968).

[148] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, June 2016), IEEE Computer Society, pp. 779–788.

[149] Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (2015), C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Curran Associates, Inc.

[150] Robbins, H., and Monro, S. A stochastic approximation method. *The annals of mathematical statistics* (1951), 400–407.

[151] Rosenberg, R. S., and Van Hout, S. The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. *Journal of clinical sleep medicine 9*, 1 (2013), 81–87.

[152] Rubin, D. B. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association* (1978), vol. 1, American Statistical Association Alexandria, VA, USA, pp. 20–34.

[153] Schapire, R. E. The strength of weak learnability. *Machine learning 5* (1990), 197–227.

[154] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

[155] Schultz, M., and Joachims, T. Learning a distance metric from relative comparisons. *Advances in neural information processing systems 16* (2003).

[156] Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155* (2018).

[157] Shen, Y., and Gao, M. Brain tumor segmentation on mri with missing modalities. In *International Conference on Information Processing in Medical Imaging* (2019), Springer, pp. 417–428.

[158] Shoham, Y., and Leyton-Brown, K. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[159] Siegel, J. M. Clues to the functions of mammalian sleep. *Nature 437*, 7063 (2005), 1264–1271.

[160] Snoek, C. G., Worring, M., and Smeulders, A. W. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia* (2005), pp. 399–402.

[161] Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[162] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks

from overfitting. *The journal of machine learning research 15*, 1 (2014), 1929–1958.

[163] Srivastava, N., and Salakhutdinov, R. R. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems* (2012), F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc.

[164] Stekhoven, D. J., and Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics 28*, 1 (2012), 112–118.

[165] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems* (2020), vol. 33, pp. 3008–3021.

[166] Strogatz, S. H. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (studies in nonlinearity). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering (Studies in Nonlinearity)* (2001).

[167] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2020).

[168] Supratak, A., Dong, H., Wu, C., and Guo, Y. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 25*, 11 (2017), 1998–2008.

[169] Supratak, A., and Guo, Y. Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2020), pp. 641–644.

[170] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[171] Tan, M., and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (2019), PMLR, pp. 6105–6114.

[172] Thomason, J., Gordon, D., and Bisk, Y. Shifting the baseline: Single modality performance on visual navigation & qa. *arXiv preprint arXiv:1811.00613* (2018).

[173] Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 247–263.

[174] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023.

[175] Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (2019), vol. 2019, NIH Public Access, p. 6558.

[176] Tsinalis, O., Matthews, P. M., Guo, Y., and Zafeiriou, S. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. *arXiv preprint arXiv:1610.01683* (2016).

[177] Turing, A. M. The chemical basis of morphogenesis. *Bulletin of mathematical biology 52* (1990), 153–197.

[178] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. In *Arxiv* (2016).

[179] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499 12* (2016).

[180] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[181] Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018), pp. 0–0.

[182] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3156–3164.

[183] Von Neumann, J., and Morgenstern, O. *Theory of games and economic behavior, 2nd rev.* Princeton university press, 1947.

[184] Wang, B., Yang, Y., Xu, X., Hanjalic, A., and Shen, H. T. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia* (2017), pp. 154–162.

[185] Wang, H., Luo, S., Hu, G., and Zhang, J. Gradient-guided modality decoupling for missing-modality robustness. *arXiv preprint arXiv:2402.16318* (2024).

[186] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022.

[187] Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 12695–12705.

[188] Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., and Huang, J. Deep multimodal fusion by channel exchanging. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 4835–4845.

[189] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[190] Wang, Z., Bingham, G., Yu, A., Le, Q., Luong, T., and Ghiasi, G. Haloquest: A visual hallucination dataset for advancing multimodal reasoning, 2024.

[191] Wei, Y., and Hu, D. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730* (2024).

[192] Wei, Y., Hu, D., Du, H., and Wen, J.-R. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[193] White, I. R., Royston, P., and Wood, A. M. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine 30*, 4 (2011), 377–399.

[194] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J.,

Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Online, Oct. 2020), Q. Liu and D. Schlangen, Eds., Association for Computational Linguistics, pp. 38–45.

[195] Wolpert, D. H. Stacked generalization. *Neural networks 5*, 2 (1992), 241–259.

[196] Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning* (2022), PMLR, pp. 24043–24055.

[197] Xiao, F., Lee, Y. J., Grauman, K., Malik, J., and Feichtenhofer, C. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740* (2020).

[198] Xu, J., Mei, T., Yao, T., Rui, Y., and Tao, D. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 5288–5296.

[199] Xu, K. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044* (2015).

[200] Xu, P., Zhu, X., and Clifton, D. A. Multimodal Learning With Transformers: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence 45*, 10 (Oct. 2023), 12113–12132.

[201] Xu, R., Feng, R., Zhang, S.-X., and Hu, D. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), IEEE, pp. 1–5.

[202] Yao, Y., and Mihalcea, R. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022* (Dublin, Ireland, May 2022), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Association for Computational Linguistics, pp. 1824–1834.

[203] Yoon, J., Zame, W. R., and van der Schaar, M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering 66*, 5 (2018), 1477–1490.

[204] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

[205] Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[206] Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[207] Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 2236–2246.

[208] Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association* (2018), 1351–1358.

[209] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters 23*, 10 (2016), 1499–1503.

[210] Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics 32*, 1 (2004), 56–85.

[211] Zhang, Z., Chen, J., Liu, X., Mai, W., and Cai, Q. 'what'and 'where'both matter: dual cross-modal graph convolutional networks for multimodal named entity recognition. *International Journal of Machine Learning and Cybernetics 15*, 6 (2024), 2399–2409.

# Curriculum vitae

Konstantinos Kontras originates from Heraklion, Creta in Greece. He obtained the B.Sc. and M.Eng. degree in electrical and computer engineering in 2018 at the University of Patras in Greece. In 2020, he obtained a postgraduate degree in artificial intelligence engineering at KU Leuven. Since 2020, he has been working towards a PhD in the STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics of the Electrical Engineering Department at KU Leuven. His research focused on enhancing multimodal fusion by refining model architectures and training objectives for deep learning networks.

# List of Publications

## Articles in internationally reviewed journals

- **Kontras, Konstantinos** and Chatzichristos, Christos and Phan, Huy and Suykens, Johan and De Vos, Maarten. "CoRe-Sleep: A Multimodal Fusion Framework for Time Series Robust to Imperfect Modalities." IEEE Transactions Journal of Neural System and Rehabilitation Engineering 2024;32:840-849. Doi: 10.1109/TNSRE.2024.3354388

## Articles in proceedings of international conferences

- **Kontras, Konstantinos** and Chatzichristos, Christos and Blaschko, Matthew and De Vos, Maarten. Improving multimodal learning with multi-loss gradient modulation. In 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024 (2024), BMVA

## Articles in Preprint

- **Kontras, Konstantinos** and Strypsteen, Thomas and Chatzichristos, Christos and Liang, Paul Pu and Blaschko, Matthew and De Vos, Maarten. "Multimodal Fusion Balancing Through Game-Theoretic Regularization" arXiv preprint arXiv:2411.07335.

# Acknowledgements

This work was supported in part by the Flemish Government through the Artificial Intelligence (AI) Research Program; in part by Fonds Weten- schappelijk Onderzoek (FWO) through AI for Data-Driven Personalized Medicine under Grant G0C9623N; in part by the Deep, Personalized Epileptic Seizure Detection under Grant G0D8321N.

# Statement on the Use of Generative AI

In accordance with KU Leuven's guidelines on the responsible use of Generative Artificial Intelligence (GenAI) in academic work, I confirm that any GenAI tools utilized in the preparation of this thesis, such as for language refinement or generating images, were employed responsibly and transparently. These tools were not used to generate substantive content, conduct data analysis, or replace critical academic judgment.

All content, including conceptual development, data analysis, and interpretations, is my own original work, reflecting academic rigor and integrity. Any specific instances where GenAI tools were employed have been appropriately documented to ensure transparency and adherence to ethical research practices.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING AND DATA ANALYTICS
Kasteelpark Arenberg 10
B-3001 Leuven
konstantinos.kontras@kuleuven.be
https://www.esat.kuleuven.be/stadius