UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Multimodal Representation Learning for Agent Perception and Action

Miguel Serras Vasco

Supervisor: Doctor Ana Maria Severino de Almeida e Paiva
Co-Supervisor: Doctor Francisco António Chaves Saraiva de Melo

Thesis approved in public session to obtain the PhD Degree in
Computer Science and Engineering

Jury final classification: Pass with Distinction and Honour

2023

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Multimodal Representation Learning for Agent Perception and Action

Miguel Serras Vasco

Supervisor: Doctor Ana Maria Severino de Almeida e Paiva
Co-Supervisor: Doctor Francisco António Chaves Saraiva de Melo

Thesis approved in public session to obtain the PhD Degree in
Computer Science and Engineering

Jury final classification: Pass with Distinction and Honour

Jury

Chairperson: Doctor Duarte Nuno Jardim Nunes, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Members of the Committee:

Doctor Frans Adriaan Oliehoek, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Netherlands
Doctor Louis-Philippe Morency, School of Computer Science, Carnegie Melon University, EUA
Doctor Francisco António Chaves Saraiva de Melo, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Doctor André Filipe Torres Martins, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Funding Institution

Fundação para a Ciência e a Tecnologia (FCT) - Portugal , grant SFRH/BD/139362/2018

2023

Acknowledgements

My research work, much of which is documented in this document, started on February 15, 2017, at precisely 11:00 AM. I had just arrived at Núcleo 7, a series of offices belonging to the GAIPS research group, located on the second floor of the Taguspark building of Instituto Superior Técnico, in Oeiras, Portugal. Anxious, but also quite excited, I was about to talk to Professor Ana Paiva about the possibility of doing research under her supervision. Since that day, much (if not all) of my life has changed.

There are no words to express the honor that has been being a Ph.D. student under the supervision of Ana Paiva and Francisco Melo. Ana Paiva has shown me nothing but kindness, encouragement, and guidance throughout my Ph.D. I will sincerely miss her wisdom, joyful nature and positive attitude in my everyday life, and I hope I can bring with me some part of that in my future endeavors. Francisco Melo was also a fundamental anchor during the hardships of the Ph.D., becoming a mentor in life, academia and, occasionally, in martial arts. More importantly, Francisco has become a friend that I will cherish for the rest of my life and whose advice I will always seek and rely upon. I would also like to give a warm appreciation to Professor Alberto Sardinha, Professor David Martins de Matos, Professor Tetsunari Inamura, and Professor Danica Kragic for their guidance at different moments of my Ph.D., which significantly contributed to the success of the overall project. I truly hope that the work present in this document has lived up to the expectations of all. I would also like to acknowledge the support of Fundação para a Ciência e Tecnologia (FCT-Portugal) through the fellowship SFRH/BD/139362/2018, which has allowed me to conduct research in the last four years.

It is also difficult for me to express the joy that I've felt being a part of the GAIPS research group. This has been my home for the past five years and I had the privilege to meet and interact with a wide range of wonderful people. From the old school members of GAIPS, I would like to thank Alexis Jacq for the crazy ideas, Daniel Tozadore for being my brother from Brazil, Elmira Yadolahi for putting up with me in India and *always* being nice to me, Hang Yin for being an infinite pool of knowledge and a cherished advisor, Kim Baraka for showing me that life is more than research, Luís Luz for our nights working together and for being my favourite photographer in the world, Patricia Alves-Oliveira for always being the most energetic and fun person in the room, Raul Paradeda for destroying me in Football every single time, Shruti Chandra for being the kindest person I've ever met, and Sofia Petisca for the company and Dim Sum in Macau. From the current members of GAIPS, I would like to thank Bernardo Esteves for being my favorite PhD student (despite not yet starting), Fábio Vital for the sake, Pocari Sweat and investing tips, Guilherme Varela for the interesting conversations about life, Henrique Fonseca for the friendly talks and being my doppelganger, Inês Batina for the intense football games and the help moving Pepper, Inês Lobo for the kindness and being the motor of our Social Media team, Jacopo Silvestrin per i meme e la morte è solo un'altra strada, quella che tutti noi dobbiamo

prendere, João Ribeiro for the companionship, José Rocha for the ancient wisdom, Manuel Guimarães for being the person with best musical taste in the lab, Maria José for the friendly conversations, Miguel Faria for missing out on a career in rapping, Pedro Fernandes for being the most successful artist in GAIPS and a close companion in Greece, Pedro Pinto Santos for teaching me a lot about research and for the nights working together, Ramona Merhej for the interesting conversations, Raquel Oliveira for the companionship, Ricardo Rodrigues for letting me hit him with a real sword, Samuel Gomes for being the John Travolta (Grease version) of GAIPS, and Silvia Tulli for the days in Montreal, the nights of gymnastic and for being a friend to me throughout these years. A special thanks to Sandra Sá, the *de facto* ruler of GAIPS, that, with a friendly smile and french music recommendations, has always found a way to solve all my administrative problems. I would also like to thank all the people I met during my visit to the National Institute of Informatics: Ana Salta, Artur Goncalves, Bastien Rigault, Florence Ho, Hai To, Rúben Geraldes, and Sandeep Hoonjan, for the friendship and karaoke nights in Tokyo, as well as Yoshiaki Mizuchi, Hiroki Yamada and Takashi Ogura, for the companionship and assistance during my internship (despite my terrible Japanese). Finally, I would like to thank all the people I met during my visit to KTH: Anastasiia Varava for enticing me to start a contraband business of Porto wine, Alfredo Ricchini for the interesting conversations about politics and the amazing beard, Emma Janson for being a friendly bridge between Sweden and Portugal, Federico Baldassare for the invitations to do rock climbing, Fethiye Irmak Dogan for the friendship and sheltering me from the cold, Giovanni Luca Marchetti for the nights in Baltimore and answering every possible question I ever had, Ignacio "Nacho" Torroba, for being my Spanish brother, João Carvalho for being a bastion of Portugal in a faraway country, Marco Moletta for being always kind to me as well as my rock brother, Michael Welle for the speed-running conversations and whisky nights, Vladislav Polianskii for the summer days in Lisbon (and for becoming a star of Portuguese television), and Yonk Shi, for being the coolest, chillest person I know. I am truly honored to have met all of you and the work presented in this document would not be possible without your direct, or indirect, contribution.

I would also like to thank the closest friends that I met during my Ph.D., which became a fundamental part of my life. I would like to thank Filipa Correia, that has become my best friend and a role model: decisive, courageous, and kind. Her friendship is something I treasure above most things and I hope that, despite our physical distance, we can always be close. I would like to thank Ali Kordia, who despite all struggles never failed to raise my spirits when I needed. Ali is a dear friend and someone I'm sure I can count on, wherever the future may take us. I would like to thank Ana Carrasco for keeping my sanity during my second visit to Japan. Ana was the perfect companionship for that adventure and I will cherish the memories of those crazy Shinjuku nights. I would like to thank Diogo Carvalho, a person whose intelligence and comedic timing never fail to surprise me. I cherish our many discussions, which have always allowed me to learn something new, something old and stories about senhor Paulo. I would like to thank Diogo Rato, a close friend that, to this day, I'm not entirely sure appreciates my company. Diogo is an incredible individual: hard-working, intelligent, kind, and someone that I look up to (and hope to work in the future). I would like to thank Petra Poklukar for just being *awesome*. Petra is the best teammate I could ever wish for, motivating me when I'm down and kind to my failures, and I sincerely hope we can continue working together no matter where we go. I would like to thank Rui Silva, which is probably the person that has taught me more during my Ph.D. Rui is everything I want to be as a researcher — methodical, focused, and efficient — but

also a kind, caring, and strong person, despite our irreconcilable differences in politics and football.

I would like to thank the love I receive from my family. I would like to thank Miguel Vasconcelos, Rui Martins, and Tomás Nunes for always supporting me and putting up with me (which, I know, is often a challenge). I would like to thank Maria Duarte, Mariana Lapa, Marta Mancelos, Marta Torre, and Teresa Araújo for all the kindness and love you have always given me. I know that, no matter where the future takes me, you will always be with me. I would like to thank Alberta Longhini for showing me a dream of a better future, with her by my side. I would like to thank Catarina, João, Beatriz, and Tomás Lobão for being the best family I could ever wish for.

Finally, I would like to thank my mom and dad, Lurdes Pedro and Fernando Vasco, with whom I have a debt I can never repay. Everything I am today is due to them, and everything I do is to try to make them proud.

Resumo

Neste trabalho investiga-se o problema de dotar agentes autónomos com mecanismos para aprender representações multimodais a partir de dados sensoriais e permitir que estes executem tarefas em condições de *disponibilidade perceptual parcial*, i.e., considerando diferentes subconjuntos de percepções disponíveis. Explora-se a aprendizagem de representações multimodais usando abordagens de aprendizagem supervisionada, não supervisionada e auto-supervisionada, bem como a utilização destas representações em cenários de aprendizagem por reforço com condições dinâmicas de disponibilidade perceptual em tempo de execução. No contexto da aprendizagem supervisionada de representações, a tese contribui com uma nova representação multimodal de acções humanas e um algoritmo de aprendizagem que permite que agentes considerem a informação contextual disponível nas demonstrações de ações, permitindo o seu reconhecimento eficiente. No contexto da aprendizagem não supervisionada de representações, investiga-se o fenómeno da *inferência cruzada de modalidades*—a estimativa de dados perceptuais em falta a partir de percepções disponíveis—contribuindo-se um novo modelo generativo multimodal hierárquico que atende aos requisitos de geração multimodal computacional. No contexto da aprendizagem auto-supervisionada de representações, a tese propõe um novo método baseado em aprendizagem por contraste multimodal que fornece um desempenho robusto em tarefas diversas com condições de disponibilidade percetual parcial em tempo de teste. Introduz-se também o problema de *transferência de políticas multimodais* em aprendizagem por reforço, onde um agente deve aprender e explorar políticas sobre diferentes subconjuntos de modalidades perceptuais, instanciando tal problema no contexto de jogos de Atari. Por fim, a tese extende as ideias de modelos perceptuais multimodais a cenários multi-agente, introduzindo-se o paradigma de execução *híbrida* para sistemas multiagente de aprendizagem por reforço, permitindo que agentes explorem informações partilhadas passivamente em tempo de execução para realizar tarefas cooperativas em cenários com qualquer nível possível de comunicação no ambiente.

Palavras-chave: Aprendizagem de Representações Multimodais; Modelação Generativa; Aprendizagem por Reforço Profunda; Aprendizagem Profunda; Aprendizagem Profunda Multimodal.

Abstract

In this thesis we address the problem of endowing agents with mechanisms to learn multimodal representations from sensory data and to allow the execution of tasks under *partial perceptual availability*, i.e., considering different subsets of available perceptions. We explore learning multimodal representations from supervised, unsupervised and self-supervised approaches and then to leverage such representations for reinforcement learning tasks under changing conditions of perceptual availability at execution time. In the context of supervised representation learning, we contribute a novel multimodal representation of human actions and a learning algorithm that enables agents to consider contextual information provided in action demonstrations, allowing sample-efficient recognition of human actions. In the context of unsupervised representation learning, we explore the *cross-modality inference* problem—the estimation of missing perceptual data from available perceptions—and contribute a novel hierarchical multimodal generative model that addresses the requirements of computational cross-modality generation. In the context of self-supervised representation learning, we propose a novel framework based on multimodal contrastive learning that provides robust performance to downstream tasks with missing modality information at test time. Furthermore, we introduce *multimodal policy transfer* in reinforcement learning, where an agent must learn and exploit policies over different subsets of input modalities and instantiate such problem in the context of Atari Games. Finally, we extend our ideas of multimodal perceptual models to multi-agent settings, and introduce the paradigm of *hybrid* execution for multi-agent reinforcement learning, allowing agents to perform cooperative tasks across all possible communication levels in the environment, while exploiting passively shared information at execution time.

Keywords: Multimodal Representation Learning; Generative Modelling; Deep Reinforcement Learning; Deep Learning; Multimodal Deep Learning.

Summary

Acknowledgements . . . i
Resumo . . . v
Abstract . . . vii
List of Tables . . . xiii
List of Figures . . . xv

1 Introduction 1
1.1 Research Question . . . 2
1.2 Thesis Approach . . . 3
1.3 Contributions of the Thesis . . . 4
1.3.1 Motion Concepts . . . 4
1.3.2 Multimodal Unsupervised Sensing (MUSE) Model . . . 4
1.3.3 Geometric Multimodal Contrastive (GMC) Learning . . . 5
1.3.4 Multimodal Transfer in Reinforcement Learning . . . 5
1.3.5 Hybrid Execution in MARL . . . 5
1.4 Summary of Contributions . . . 6
1.5 Structure of the Thesis . . . 7

2 Background 9
2.1 Representation Learning . . . 9
2.1.1 Unsupervised Representation Learning . . . 9
2.1.2 Self-Supervised Representation Learning . . . 11
2.2 Reinforcement Learning . . . 12
2.2.1 Multi-Agent Reinforcement Learning . . . 15

3 Related Work 19
3.1 Human Representation Learning . . . 19
3.2 Representation Learning of Human Actions . . . 20
3.2.1 Human Action Recognition with Deep Learning Models . . . 21
3.2.2 Multimodal Human Action Recognition . . . 22
3.3 Representation Learning with Variational Autoencoders . . . 23
3.3.1 Variational Autoencoders . . . 24
3.3.2 Hierarchical Variational Autoencoders . . . 24
3.3.3 Disentanglement in Variational Autoencoders . . . 25
3.3.4 Multimodal Variational Autoencoders . . . 26
3.4 Representations for Reinforcement Learning . . . 27
3.4.1 Abstractions in RL . . . 27
3.4.2 Adaptation in RL . . . 28
3.4.3 Representation Learning for Model-Based RL . . . 29

3.4.4 Representation Learning for Multi-Agent RL . . . 30

4 Learning Multimodal Representations of Human Actions 33

4.1 Multimodal Representation of an Action . . . 35
4.1.1 Setup . . . 35
4.1.2 Motion Prototype . . . 36
4.1.3 Motion Concepts . . . 37
4.2 Learning Motion Concepts from Demonstration . . . 38
4.2.1 Creating Motion Prototypes from Data . . . 38
4.2.2 Creating and Updating Motion Concepts . . . 38
4.2.3 Recognizing Motion Concepts . . . 39
4.3 Evaluation . . . 40
4.3.1 Experimental Setup . . . 41
4.3.2 Offline One-Shot Recognition (OSR) Task . . . 41
4.3.3 Online Tabula-Rasa Learning Task . . . 43
4.4 Concluding Remarks . . . 45

5 Learning Multimodal Representations for Effective Cross-Modality Inference 47

5.1 The Problem of Cross-Modality Inference (CMI) . . . 49
5.2 The Role of Hierarchy in Perception . . . 54
5.3 Multimodal Sensing (MUSE) model . . . 55
5.4 Multimodal Encoder . . . 56
5.4.1 Naive Encoder . . . 56
5.4.2 NEXUS Encoder . . . 57
5.4.3 ALMA Encoder . . . 59
5.5 Multimodal Handwritten Digits Dataset . . . 60
5.6 Evaluation . . . 61
5.6.1 Results . . . 62
5.7 Metrics for Cross-Modality Inference . . . 64
5.7.1 Multimodal Handwritten Digits . . . 66
5.8 Ablation Study . . . 67
5.9 Concluding Remarks . . . 68

6 Multimodal Contrastive Representation Learning 69

6.1 Geometric Misalignment . . . 70
6.2 Geometric Multimodal Contrastive Learning . . . 72
6.3 Experiments . . . 73
6.3.1 Overview of Delaunay Component Analysis (DCA) . . . 73
6.3.2 Unsupervised Representation Learning Task . . . 74
6.3.3 Supervised Learning Task . . . 76
6.3.4 Ablation studies . . . 79
6.4 Concluding Remarks . . . 80

7 Multimodal Transfer in Reinforcement Learning 81

7.1 Multimodal Transfer in Reinforcement Learning . . . 83
7.1.1 Learn a perceptual model of the world . . . 84
7.1.2 Learn to act in the world . . . 84
7.1.3 Policy transfer . . . 85

7.2 Multimodal Atari Games . . . 86
7.2.1 Pendulum Environment . . . 86
7.2.2 HYPERHOT environment . . . 87
7.3 Experimental Evaluation . . . 88
7.3.1 Pendulum . . . 89
7.3.2 HYPERHOT . . . 91
7.4 Concluding Remarks . . . 96

8 Multi-Agent Reinforcement Learning with Hybrid Execution 97
8.1 Hybrid Execution in Multi-Agent Reinforcement Learning . . . 99
8.2 Multi-Agent Observation Sharing with Communication Dropout (MARO) . 100
8.2.1 Predictive Model . . . 100
8.2.2 Training Scheme . . . 101
8.3 Evaluation . . . 102
8.3.1 Scenarios . . . 102
8.3.2 Baselines and Experimental Methodology . . . 103
8.3.3 Results . . . 104
8.3.4 Ablation Study . . . 105
8.4 Concluding Remarks . . . 107

9 Conclusions 109
9.1 Summary of Contributions . . . 110
9.2 Main Conclusions . . . 111
9.3 Future Work . . . 112

A Summary of Publications 131

B CMI Evaluation Metrics 133
B.1 Log-Likelihood Metrics . . . 133
B.2 Kullback-Leibler Distance . . . 133
B.3 Accuracy and Modality Distance . . . 134

C Ablation Study of MUSE 135
C.1 Hierarchical Design . . . 135
C.2 Standard Datasets . . . 137
C.3 MHD Dataset . . . 139

D Ablation Study of GMC 143
D.1 Temperature parameter . . . 143
D.2 Dimensionality of the low-level representations . . . 143
D.3 Dimensionality of the high-level latent representations . . . 144
D.4 Loss function . . . 144

E Additional Results of MARO 147
E.1 Main Experimental Results . . . 147
E.2 Ablation Study . . . 147
E.2.1 Switch Baseline . . . 147
E.3 Multi-agent Trajectory Prediction . . . 148

F Training Hyperparameters and Constants 165

G Network Architectures **169**

List of Tables

4.1 Action classes performed to evaluate motion concepts. 42
4.2 Accuracy in the one-shot recognition of motion concepts task. 43

5.1 Standard metrics for generative performance in the MNIST, CelebA and MNIST-SVHN datasets. 63
5.2 Cross-modality generation performance evaluation in the MNIST dataset. . 64
5.3 Evaluation of cross-modality generation in the MHD dataset. 67

6.1 Results of different multimodal representation methods in the MHD dataset. 75
6.2 Alignment scores A_{DCA} of the models in the MHD dataset. 76
6.3 Results of multimodal representation models in the CMU-MOSEI dataset. . 77
6.4 Results of multimodal representation models in the CMU-MOSI dataset. . . 78
6.5 Alignment scores A_{DCA} of the representation models in the CMU-MOSEI dataset. 79
6.6 Alignment scores A_{DCA} of the representation models in the CMU-MOSI dataset. 79

7.1 Zero-shot performance of cross-modality policy transfer in the Pendulum scenario. 90
7.2 Zero-shot performance of multimodal policy transfer in the pendulum scenario. 90
7.3 Zero-shot performance of cross-modality policy transfer in the Hyperhot scenario. 91
7.4 Zero-shot performance of multimodal policy transfer in the hyperhot scenario, with a random dataset. 92
7.5 Zero-shot performance of multimodal policy transfer in the hyperhot scenario, with a optimal dataset. 93

8.1 Average episodic returns and 95% bootstrapped confidence interval over five seeds in all scenarios. 104
8.2 Average episodic returns and 95% bootstrapped confidence interval for the ablated versions of MARO in all scenarios. 105

C.1 Results of the hierarchical evaluation of multimodal generative models accordingly to the alternative cross-modality inference metrics. 136
C.2 Results of the evaluation of multimodal generative models accordingly to the alternative cross-modality inference metrics in the MNIST and FashionMNIST dataset. 140
C.3 Evaluation of cross-modality generation in the MHD dataset for the MUSE ablation study. 141

D.1 GMC performance with different temperature values τ in the MHD dataset. 144
D.2 GMC alignment scores A_{DCA} with different temperature values τ in the MHD dataset. 144
D.3 GMC performance with different values of dimensionality of $\boldsymbol{h} \in \mathbb{R}^d$ in the MHD dataset. 145
D.4 GMC alignment scores A_{DCA} with different values of dimensionality of $\boldsymbol{h} \in \mathbb{R}^d$ in the MHD dataset. 145
D.5 GMC performance with different values of dimensionality of $\boldsymbol{z} \in \mathbb{R}^s$ in the MHD dataset. 145
D.6 GMC alignment scores A_{DCA} obtained with different values of dimensionality of $\boldsymbol{z} \in \mathbb{R}^s$ in the MHD dataset. 146
D.7 GMC performance with different loss functions in the MHD dataset. 146
D.8 GMC alignment scores A_{DCA} obtained with different loss functions in the MHD dataset. 146

F.1 Training hyperparameters employed in the evaluation of MUSE (Chapter 5) 166
F.2 Training hyperparameters of GMC (Chapter 6). 166
F.3 Constants employed in the multimodal Atari games (Chapter 7) 167
F.4 Hyperparameters employed in the training of generative models for the multimodal transfer in reinforcement learning scenarios (Chapter 7) 167
F.5 Hyperparameters employed in the training of reinforcement learning algorithms for the multimodal transfer in reinforcement learning scenarios (Chapter 7) 168
F.6 Hyperparameters for the RL controllers across all environments for hybrid execution in multi-agent reinforcement learning (Chapter 8). 168
F.7 Hyperparameters for the predictive model across all environments and algorithms for hybrid execution in multi-agent reinforcement learning (Chapter 8). 168

G.1 Modality-specific network architectures for the evaluation of MUSE 170
G.2 Multimodal network architectures for the evaluation of MUSE 170
G.3 Network architectures for the complementary metrics for cross-modality generative performance of MUSE. 171
G.4 Modality-specific network architectures for the multimodal Atari games scenarios. 172
G.5 Multimodal network architectures for the multimodal Atari games scenarios. 172

List of Figures

1.1 Depiction of the thesis research question. 3
1.2 Core challenges of multimodal machine learning. 7

2.1 Bayesian network of the variational autoencoder model. 10
2.2 Network architecture of the variational autoencoder model. 11
2.3 Self-supervised contrastive learning of data representations. 12
2.4 The closed-loop framework of reinforcement learning. 13
2.5 The closed-loop framework of MARL under partial-observability. 16

3.1 The Convergence-Divergence Zone framework. 20
3.2 The evolution of generative capabilities of variational autoencoder frameworks. 24
3.3 Architecture of of relevant works in representation learning for deep reinforcement learning agents. 29

4.1 Setup for learning multimodal action representation. 35
4.2 Diagram of the motion concept representation. 37
4.3 Setup employed to evaluate motion concepts. 40
4.4 Confusion Matrices in the one-shot recognition task. 43
4.5 Results of the online *tabula-rasa* learning task to evaluate motion concepts. 44

5.1 Architecture of the naive multimodal extension of the VAE model. 49
5.2 Architecture of the AVAE model. 50
5.3 Architecture of the JMVAE model. 51
5.4 Architecture of the MVAE model. 52
5.5 Architecture of the MMVAE model. 53
5.6 Cross-modality inference in the CDZ framework. 54
5.7 Architecture of the MUSE model. 55
5.8 Architecture of the naive joint-modality encoder. 57
5.9 Architecture of the NEXUS joint-modality encoder. 58
5.10 Architecture of the ALMA joint-modality encoder. 60
5.11 Image samples from the “Multimodal Handwritten Digits” dataset. 61
5.12 MNIST samples reconstructed from MUSE latent spaces. 64
5.13 Label-to-image generation in the MNIST and CelebA datasets. 65
5.14 Complementary metrics to evaluate computational cross-modality inference. 66

6.1 UMAP of complete and image latent representations of the MHD dataset. . 71
6.2 The *Geometric Multimodal Contrastive* (GMC) learning model. 72
6.3 UMAP of complete and sound latent representations of the MHD dataset . 76

7.1 Multimodal Transfer in Reinforcement Learning. 83
7.2 Learn a perceptual model of the world. 84
7.3 Learn to act in the world. 85
7.4 Multimodal Atari games. 86
7.5 Sound properties in the multimodal pendulum scenario. 87
7.6 Cross-Modality Policy Transfer. 89
7.7 Image generation from sound observations in the Hyperhot scenario. . . . 95

8.1 MARO approach for hybrid execution in MARL. 101
8.2 Experimental evaluation of MARO across different environments and algorithms. 105
8.3 Evaluation of the training scheme of MARO. 106
8.4 Evaluation of the predictive model of MARO. 107

B.1 Evaluation metrics for the performance of multimodal generative models. . . 134

C.1 Cross-modality image-to-label generation by hierarchical and single-level multimodal generative models. 136
C.2 Cross-modality label-to-image generation in the MNIST and FashionMNIST datasets. 138
C.3 Image reconstruction from MUSE latent spaces in the MNIST and Fashion-MNIST datasets. 139

E.1 Average episodic returns during training for different communication levels p, for all approaches and environments, using the IQL algorithm. 148
E.2 Average episodic returns during training for different communication levels p, for all approaches and environments, using the QMIX algorithm. 149
E.3 Average episodic returns during training for $p \sim \mathcal{U}(0, 1)$, for all approaches and environments, using the IQL algorithm. 150
E.4 Average episodic returns during training for $p \sim \mathcal{U}(0, 1)$, for all approaches and environments, using the QMIX algorithm. 151
E.5 Average episodic returns for different communication levels p at execution time, for all approaches across different environments, using the IQL algorithm.152
E.6 Average episodic returns for different communication levels p at execution time, for all approaches across different environments, using the QMIX algorithm. 153
E.7 Average episodic returns during training for different communication levels p, for MARO and the ablated versions across all environments, using the IQL algorithm. 154
E.8 Average episodic returns during training for different communication levels p, for MARO and the ablated versions across all environments, using the QMIX algorithm. 155
E.9 Average episodic returns during training for $p \sim \mathcal{U}(0, 1)$, for MARO and the ablated versions across all environments, using the IQL algorithm. 156
E.10 Average episodic returns during training for $p \sim \mathcal{U}(0, 1)$, for MARO and the ablated versions across environments, using the QMIX algorithm. 157
E.11 Average episodic returns for different communication levels p at execution time, for MARO and the ablated versions across different environments, using the IQL algorithm. 158

E.12 Average episodic returns for different communication levels p at execution time, for MARO and the ablated versions across different environments, using the QMIX algorithm. 159
E.13 Average episodic returns at execution time for $p \sim \mathcal{U}(0,1)$, for different sampling methods of the training scheme of MARO across environments, using the IQL algorithm. 160
E.14 Average episodic returns at execution time for $p \sim \mathcal{U}(0,1)$, for different sampling methods of the training scheme of MARO across environments, using the QMIX algorithm. 161
E.15 Average episodic returns for different communication levels p at execution time for MARO and the Switch baseline in the HearSee environment. 161
E.16 Trajectory prediction plots for the Spreadblindfold environment under the QMIX algorithm from the perspective of agent 0 (blue). 162
E.17 Trajectory prediction plots for the Spreadblindfold environment under the QMIX algorithm from the perspective of agent 1 (orange). 163
E.18 Trajectory prediction plots for the Spreadblindfold environment under the QMIX algorithm from the perspective of agent 2 (green). 164

Chapter 1

Introduction

"Knowledge begins with the senses, proceeds then to the understanding, and ends up with reason."
Immanuel Kant, *Critique of Pure Reason*

Our understanding of the world is fundamentally shaped by the information provided by our senses. For a moment, let us consider the scenario of a human walking across a field: in that very moment, light is piercing the retina of the human, exciting photoreceptor cells contained within and translating electromagnetic information into electrical pulses, allowing him to see where he is and where he is going [13]. Simultaneously, sound pervades the ear canal of the human, causing the eardrum to vibrate and resulting in the translation of mechanical information into electric signals, effectively allowing the human to hear his surroundings. The same environment provides chemical and pressure information, interpreted by specialized receptors in the nose, tongue and skin, allowing the human to smell, taste and touch the world. From such heterogeneous stimuli, captured by dedicated sensory organs, humans are able to probe the current state of their environment, considering their intrinsic perceptual limitations [29].

However, despite its partitioned origin, the human perceptual experience remains multimodal. The integration of single-modality signals in multiple convergence zones in the brain enables the creation of *multimodal representations*, whose conscious experience remains an open question [12, 30, 108]. The richness (and value) of such representations cannot be overstated: the interaction between the different modalities is responsible for the emergence of complex multimodal phenomena, whose response is often greater than the combined response to each individual stimuli [150]. Moreover, we humans employ multimodal representations to plan and execute tasks [86]. In particular, in conditions of *partial perceptual availability*, due to sensor malfunction (e.g., visual impairments) or the environment not providing some modality (e.g., absence of light in a dark room), humans leverage such multimodal representations to execute tasks (e.g., navigating a room), albeit with decreased performance [105, 149, 172]. The robust adaptation to changing conditions of perceptual availability is a powerful allure of multimodality.

Recently, the field of Artificial Intelligence (AI) has shown great interest in multimodal learning to exploit the benefits, and address the significant challenges, that arise when considering multiple heterogeneous sources of data [4, 5]. Simultaneously, deep reinforcement learning (RL) has shown great advance in allowing artificial agents, such as robots, to perform ever-so more complex tasks in real and virtual environments [2, 82, 141]. However, the full consideration of a multimodal perceptual experience for artificial agents remains largely unexplored: agents often not to take advantage of the powerful interaction between the different modalities that compose their perceptual input, limiting themselves to creating internal representations only from visual information [45, 126] or from the fixed fusion of different modalities [75, 100]. The disregard of the multimodal nature and dynamics of perceptual information can have significant consequences: the inability of the agent to perform tasks when modality-specific information is unavailable, when modality-specific information becomes degraded (such as visual information under low luminosity settings) or in the (frequent) case of sensory malfunction. If we wish to have artificial agents, such as service robots or autonomous vehicles, acting reliably in their environments they must be provided with mechanisms to overcome these issues. Moreover, the rise of privacy concerns regarding the acquisition of human data by artificial agents (e.g., from camera and microphone sensors) in real-world environments further motivates the need to develop solutions to allow agents to act under conditions of partial perceptual availability [32, 42].

1.1 Research Question

The prior discussion identifies significant challenges regarding both the learning and use of multimodal representations by artificial agents for the execution of tasks. Human multimodal representations are continuously learnt and updated since infancy, not in "*one great blooming, buzzing confusion*" [72] but through complex biological mechanisms that mimic supervised, unsupervised and reinforcement learning algorithms [35]. Furthermore, humans intuitively leverage such representations in the execution of tasks under partial perceptual availability. Addressing these issues for artificial agents leads to the following research question, highlighted in Fig. 1.1:

> **Research Question**: *How can we endow artificial agents with mechanisms to learn representations from multimodal observations provided by their environment and to leverage such representations in the execution of tasks under changing conditions of perceptual availability?*

Our research fills the gap between recent developments in multimodal machine learning and deep reinforcement learning, by extending typical single-modality agents with the ability to consider a broader perceptual space, composed of heterogeneous sources of data provided by their environment. We consider two successive objectives to address such gap. We initially develop computational mechanisms to endow artificial agents with the ability to learn representations from multimodal observations. Such mechanisms must be scalable to large number of modalities and robust to the possible absence of modality-specific

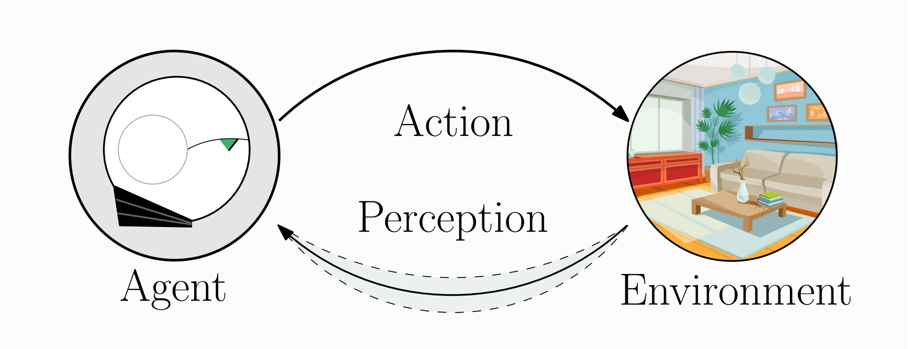


Figure 1.1: This thesis approaches the question of how an agent can perceive its environment to act robustly under changing conditions of perceptual availability.

information, be it due to changing perceptual conditions or due to sensor malfunction. Following such development, we address how to leverage such representation models in reinforcement learning scenarios, in order to provide agents with the ability to execute tasks robustly in scenarios of partial perceptual availability.

1.2 Thesis Approach

We start our work by highlighting the benefits of multimodal representations to execute tasks. In Chapter 4, we consider the task of recognizing human actions in a household environment. Humans interact with their environment in rich and diverse ways. In addition, human actions are often context-dependent: their interpretation depends not only on the motion performed, but also on what objects were employed in the action and where the action was performed. We contribute a novel probabilistic multimodal action representation that encodes both motion information and contextual information [166, 167]. We show two advantages in favor of a multimodal representation against considering only motion information. First, the addition of contextual information to the action representation improves the classification accuracy in an action recognition task, when trained on a limited number of samples. Second, contextual information reduces the ambiguity in recognizing action classes with similar motion patterns.

Following this work, we consider the scenario of learning multimodal representations from an arbitrary number of heterogeneous data sources without explicit feature engineering. Mimicking the human perceptual experience, we wish to translate the complex phenomena that arise from the interplay between different perceptual modalities to a computational setting. We address such challenge through two different approaches. First, in Chapter 5, we explore learning multimodal representations without an explicit supervision signal in the context of the *cross-modality* inference (CMI) problem, i.e., the generation of missing modality information from available perceptions. Taking inspiration from human perception, we argue in favor of considering *hierarchy* in the design of multimodal generative models to allow for effective cross-modality inference. To address the requirements of computational CMI, we propose a novel hierarchical multimodal generative model and our results show that the hierarchical architecture plays a fundamental part in allowing for effective cross-modal generation, regardless of the number and complexity of the target modalities [168, 169]. Second, in Chapter 6, we go beyond generative tasks and explore learning multimodal representations in a self-supervised setting to provide robust performance in downstream tasks with missing modality information at test time. We once again exploit hierarchy and propose a novel two-level contrastive learning framework that enforces the alignment of multimodal and modality-specific representations in a shared latent space [128]. The results show that our proposed framework outperforms other state-of-the-art multimodal representation models in downstream classification performance with missing modality information at test time.

After considering the process of learning multimodal representations, we address the question of leveraging such representations in reinforcement learning problems involving conditions of partial perceptual availability. We start in Chapter 7 by considering the single-agent case: we approach the problem of learning a policy that is transferable across different sets of perceptual modalities, allowing agents to perform tasks with partial perceptual availability at execution time, without requiring additional training. We propose a three-stage architecture that allows a reinforcement learning agent trained over a set of perceptual modalities (e.g., image) to execute its task on a distinct, and possibly disjoint, set of perceptual modalities (e.g., sound) [143]. The applicability of the proposed approach is evaluated in domains of increasing complexity, and the results show that the policies learned by our approach showcase a performance that is robust to the use of different subsets of modalities. Finally, in Chapter 8, we address the perceptual experience of multi-agent systems in regards to passively shared local information: we consider agents performing partially-observable cooperative tasks in environments with dynamic communication levels. We contribute a new approach that allows agents to take advantage of both a centralized training method and of shared information at execution time [139]. The results show that our approach consistently outperforms the baselines across all possible communication levels, allowing agents to exploit shared information during task execution.

In the next section, we outline the main contributions of this thesis.

1.3 Contributions of the Thesis

1.3.1 Motion Concepts [166, 167]

We contribute *Motion Concepts*, a novel probabilistic multimodal representation of human actions that considers motion information along with information regarding the objects interacted with and the location where the action is performed. Furthermore, we propose a novel online algorithm to learn such representations from demonstration data provided by the human user in a sample-efficient way. We evaluate our proposed representation in a virtual-reality household environment, considering both an offline one-shot recognition task and an online *tabula-rasa* learning task, where the agent must learn novel action classes from human demonstration and recognize previously observed action classes. The results highlight the importance of considering multimodal information for sample-efficient learning and recognition of human actions.

1.3.2 Multimodal Unsupervised Sensing (MUSE) Model [168, 169]

We contribute the *Multimodal Unsupervised Sensing* (MUSE) model, a multimodal hierarchical generative model that, inspired by the CDZ model of human perception [30], considers hierarchical representation levels: low-level modality-specific representations and a high-level multimodal representation. Within the MUSE framework, we discuss different solutions for merging modality information to encode multimodal representations. We evaluate MUSE in scenarios of increasing complexity including on a novel *Multimodal Handwritten Digits* dataset, a challenging scenario that considers the images, sounds, motion trajectory and label information pertaining to handwritten digits. We introduce complementary metrics to evaluate the cross-modality performance of multimodal generative models. The results show that MUSE outperforms other baselines in its ability to perform cross-modality inference, highlighting the benefits of considering hierarchical representation levels in the design of multimodal generative models.

1.3.3 Geometric Multimodal Contrastive (GMC) Learning [128]

We contribute *Geometric Multimodal Contrastive* (GMC), a simple multimodal contrastive learning framework that explicitly aligns the representations encoded from complete multimodal observations and those encoded from modality-specific observations. Following the MUSE model, we once again consider hierarchy in the design of GMC: initially we encode the complete observation and the partial observations into an low-level representation space, using *independent* base encoders; subsequently, we encode all low-level representations into a high-level, common, latent space using a *shared* projection encoder. Finally, we introduce a novel multimodal contrastive loss to enforce the alignment of the representations in the latent space, such that representations from similar observations are closer than the representations from distinct observations. We evaluate GMC against state-of-the-art multimodal representation models for classification tasks. The results show that GMC allows the execution of downstream tasks with missing modality information at test time, with minimal performance loss, outperforming the baselines. Moreover, we show that GMC can be easily integrated into current state-of-the-art models to improve their robustness to missing modality information.

1.3.4 Multimodal Transfer in Reinforcement Learning [143]

We contribute the novel problem of *multimodal transfer in reinforcement learning*, i.e., how can an agent learn a policy considering observations from a set of modalities and transfer that policy to scenarios where the environment provides observations from a different set of modalities, without requiring additional training. We propose a three-stage approach to allow agents to reuse policies under changing conditions of perceptual availability with minimum performance loss. To evaluate our approach we introduce the *Multimodal Atari Games* scenario, an extension of the Atari Games scenario where the agent is provided with both the image and the sound associated with the game state. We evaluate two different variations of the transfer problem: when all perceptual modalities available during policy training (*multimodal policy transfer*) and when the set of available modalities during policy training and evaluation are disjoint (*cross-modality policy transfer*). The results show that, in both scenarios, our agents are able to execute tasks under partial perceptual availability, with minimal performance loss. This conclusion holds with different multimodal generative models and reinforcement learning algorithms.

1.3.5 Hybrid Execution in Multi-Agent Reinforcement Learning [139]

We contribute the paradigm of *hybrid execution* for multi-agent systems, in which agents can passively share their local observations, according to an environment-specific communication matrix, to perform cooperative tasks with partial observability. Under hybrid execution, agents are expected to act in all possible communication levels, ranging from no communication (fully decentralized) to full communication between the agents (fully centralized). We formalize this setting introducing *hybrid partially observable Markov decision processes* (H-POMDP), a new class of multi-agent POMDPs that explicitly considers a communication process between the agents. To allow for hybrid execution, we contribute *multi-agent observation sharing with communication dropout* (MARO), an approach that combines an agent-specific autoregressive prediction model, that estimates non-shared information from past observations, and a training scheme that introduces communication dropout during training. To evaluate MARO, we contribute three novel environments for multi-agent systems that explicitly require sharing local information during execution to successfully perform cooperative tasks. The results show that MARO outperforms the baseline approaches, allowing agents to exploit passively shared information to successfully execute tasks under all possible communication levels.

1.4 Summary of Contributions

To summarize, the main contributions of this thesis are as follows:

1. **Motion Concepts**, a multimodal probabilistic representation for human actions and an online algorithm to learn representations from human demonstrations;
2. **MUSE**, a multimodal generative model to learn representations from an arbitrary number of heterogeneous data sources that leverages hierarchy to provide effective cross-model generation;
3. **GMC**, a multimodal contrastive representation model that aligns modality-specific representations in a shared latent space to provide robust performance to downstream classification tasks with missing modality information at test time;
4. **Multimodal Transfer in Reinforcement Learning**, an approach that leverages multimodal representations to allow agents to transfer task policies over different sets of perceptual modalities;
5. **Hybrid Execution in Multi-Agent Reinforcement Learning**, an approach that allows agents to exploit a centralized training scheme and passively shared observations during execution to perform cooperative tasks under all possible communication levels.

The main contributions of this thesis address core technical challenges in multimodal machine learning, depicted in Fig. 1.2, adapted from Baltrusaitis et al. [4]. Our first contribution considers the problem of **Representation** (**R**), *i.e.,* how to learn a low-dimensional representation of multiple heterogeneous perceptual modalities, and the problem of multimodal **Fusion** (**F**), *i.e.,* how to merge heterogeneous information from multiple sources of data. Following this, in our second contribution, we propose computational mechanisms to explore the rich interplay between the modalities in order to map information between them, such as to perform cross-modality inference, addressing a problem we denote by multimodal **Translation** (**T**). In our third contribution we address the problem of transferring knowledge between different sets of modalities for downstream tasks (such as classification tasks), which we denote by multimodal **Co-learning** (**CL**). Finally, in our fourth and fifth contributions we explore how to leverage representations for the robust actuation of (single and multiple) agents in scenarios with partial perceptual availability, framing such contributions in the **Co-Learning** (**CL**) challenge class.

We present a complete list of publications related to this thesis in Appendix A.

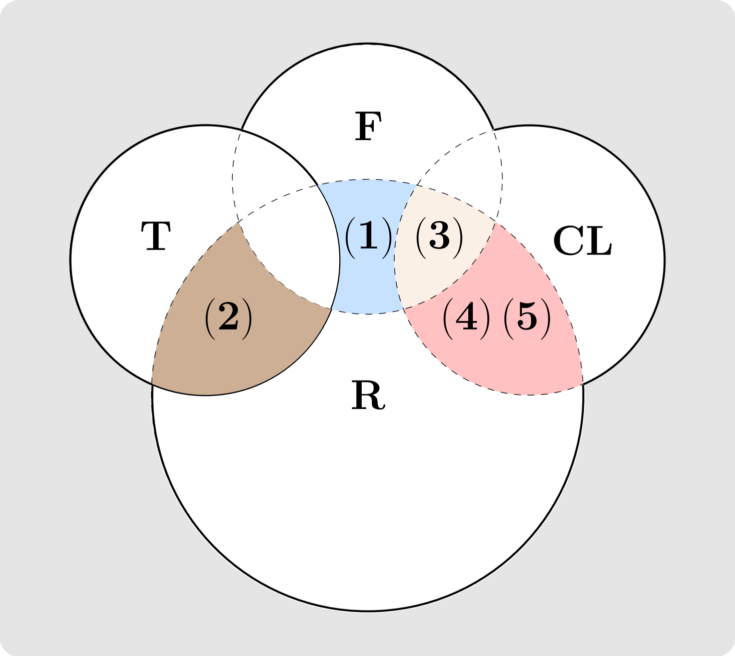


Figure 1.2: Core challenges of multimodal machine learning, introduced by Baltrusaitis et al. [4]: representation (**R**), fusion (**F**), translation (**T**) and co-learning (**CL**). We frame the expected contributions of this thesis according to such challenges: (**1**) motion concepts; (**2**) MUSE; (**3**) GMC; (**4**) multimodal transfer in reinforcement learning; (**5**) hybrid execution in multi-agent reinforcement learning.

1.5 Structure of the Thesis

The thesis is structured in 9 chapters, as follows,

> **Chapter 2** introduces relevant background on representation learning and reinforcement learning;
>
> **Chapter 3** discusses recent developments in biological representation learning, computational multimodal representation learning and representations for reinforcement learning agents;
>
> **Chapter 4** introduces the *motion concept* representation of human actions and a novel online algorithm to learn such representations from human demonstrations;
>
> **Chapter 5** discusses the computational *cross-modality inference* problem in light of recent developments in multimodal generative models and introduces the *Multimodal Unsupervised Sensing* (MUSE) model;
>
> **Chapter 6** introduces the *Geometric Contrastive Learning* (GMC) framework that provides robust performance to downstream classification tasks with missing modality information at test time;
>
> **Chapter 7** introduces the problem of *multimodal transfer in reinforcement learning* and proposes an approach to leverage multimodal representations in the execution of tasks with partial perceptual availability;

Chapter 8 introduces the paradigm of *hybrid execution* for multi-agent systems and proposes an approach to allow agents to exploit passively shared local information to perform cooperative tasks under all possible communication levels;

Chapter 9 revisits the contributions and conclusions of the thesis, and outlines potential future research directions.

Chapter 2

Background

This chapter introduces necessary notation and background information regarding two major components of this thesis, namely representation learning (Section 2.1) and reinforcement learning (Section 2.2).

2.1 Representation Learning

A fundamental component of the work presented in this thesis concerns the learning of representations of perceptual data. Representations are low-dimensional, descriptive features that encode relevant information for some downstream task regarding the original data (often high-dimensional) [8]. Let us consider a dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{D} \subset \mathcal{X} \subset \mathbb{R}^M$ of data *samples* $\boldsymbol{x}_i$ from an input space $\mathcal{X}$ that is embedded in a ambient space $\mathbb{R}^M$ of high-dimensionality M. We assume that the dataset $\mathcal{D}$ actually lies in a lower-dimensional manifold embedded in $\mathbb{R}^M$, in what is known as the *manifold hypothesis* [41]. The goal of representation learning is to learn a mapping $r : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^N$ from the input space to a latent space $\mathcal{Z}$, with $N \ll M$, that captures this low-dimensional manifold.

> **Definition 1** (Representation model) *A* representation model *is a map* $r : \mathcal{X} \to \mathcal{Z}$ *from input space* $\mathcal{X} \subset \mathbb{R}^M$ *to a lower-dimensional representation space* $\mathcal{Z} \subset \mathbb{R}^N$.

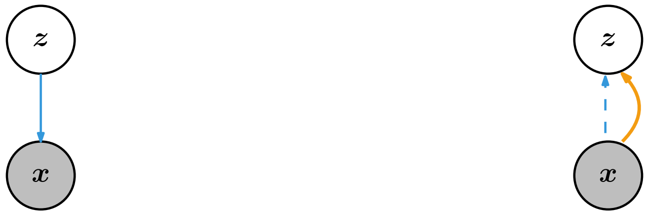
We aim at encoding data *representations* $\boldsymbol{z}_i$ that capture features of the original data relevant for some *downstream* task. Such tasks often require an additional mapping $g : \mathcal{Z} \to \mathcal{Y} \subset \mathbb{R}^T$ from the representation space to a target space $\mathcal{Y}$. The dimensionality of this target space can be smaller than the representation space ($\mathcal{Y} \ll \mathcal{Z}$), such as in the case of binary classification tasks, or larger than the representation space ($\mathcal{Y} \gg \mathcal{Z}$), such as in the case of image reconstruction tasks. We denote the map g as the *downstream* model.

> **Definition 2** (Downstream model) *A* downstream model *is a map* $g : \mathcal{Z} \to \mathcal{Y}$ *from the representation space* $\mathcal{Z} \subset \mathbb{R}^N$ *to a target domain space* $\mathcal{Y} \subset \mathbb{R}^T$.

In this thesis, we mainly focus on unsupervised (Section 2.1.1) and self-supervised (Section 2.1.2) learning approaches to learn data representations.

2.1.1 Unsupervised Representation Learning

In unsupervised learning settings, we aim at learning the (unknown) evidence distribution of the observed data $p(\boldsymbol{x})$. The data $\boldsymbol{x}$ is usually assumed to be generated trough a stochastic



(a) Likelihood distribution $p(\boldsymbol{x}|\boldsymbol{z})$. (b) Approximate posterior distribution $q(\boldsymbol{z}|\boldsymbol{x})$.

Figure 2.1: Bayesian network representation of the dependencies between the observed variable, $\boldsymbol{x}$, and the latent variable, $\boldsymbol{z}$ in a variational autoencoder model.

process mediated by some *latent* (unobserved) variable(s) $\boldsymbol{z} \in \mathcal{Z}$. The goal of *unsupervised* representation learning in to learn the representations $\boldsymbol{z}$ without an explicit supervision signal, such as the one provided in classification tasks.

Recently, deep generative models have shown great promise in learning representations of high-dimensional data [9, 138, 160]. One example of particular interest to this thesis is the Variational Auto-Encoder (VAE) model. Originally introduced by Kingma and Welling [79], the VAE explicitly computes a latent representation $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^N$ of data $\boldsymbol{x} \in X \subset \mathbb{R}^M$, such that $\boldsymbol{z}$ contains relevant information to (attempt to) reconstruct $\boldsymbol{x}$. A VAE can be represented as a Bayesian network, shown in Fig. 2.1. Formally, training a VAE model amounts to estimating the lower-bound of the evidence $p(\boldsymbol{x})$, resorting to a variational approach,

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z}) d\boldsymbol{z}, \tag{2.1}$$

where $p(\boldsymbol{z})$ is a pre-specified prior, often a unitary Gaussian distribution, and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is the likelihood distribution (Fig. 2.1a), parameterized by θ. The likelihood is often a Bernoulli distribution for binary data or a constant-variance Gaussian distribution for data in the unitary interval.

However, the integral in (2.1) is often intractable, hindering the optimization of the parameters θ of the model. To solve such issue, we introduce a family of proposal distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ (Fig. 2.1b), parameterized by ϕ, that attempts to estimate the true posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$, such that,

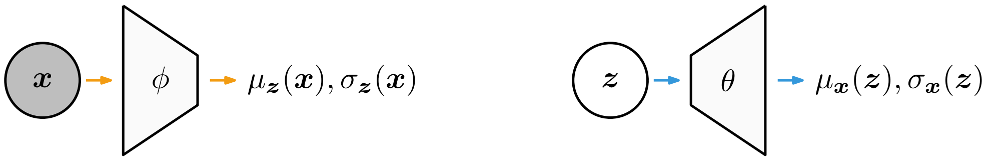
$$\log p(\boldsymbol{x}) = \log \int p_\theta(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z})\, \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} = \log \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}, \tag{2.2}$$

where we take the logarithm of the evidence probability. As the logarithm function is concave, we can estimate the lower bound of the evidence (ELBO) by applying Jensen's inequality [73],

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_\phi \parallel p), \tag{2.3}$$

where $D_{\mathrm{KL}}(q_\phi \parallel p)$ corresponds to the Kullback-Leibler divergence between the proposal distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the prior distribution of the latent variable $p(\boldsymbol{z})$, defined as,

$$D_{\mathrm{KL}}(q_\phi \parallel p) \triangleq \int q_\phi(\boldsymbol{z}|\boldsymbol{x})\, (\log q_\phi(\boldsymbol{z}|\boldsymbol{x}) - \log p(\boldsymbol{z})) d\boldsymbol{z}. \tag{2.4}$$


$\boldsymbol{x}$
ϕ
$\mu_{\boldsymbol{z}}(\boldsymbol{x}), \sigma_{\boldsymbol{z}}(\boldsymbol{x})$
$\boldsymbol{z}$
θ
$\mu_{\boldsymbol{x}}(\boldsymbol{z}), \sigma_{\boldsymbol{x}}(\boldsymbol{z})$

(a) Encoder $p_\theta(\boldsymbol{z}|\boldsymbol{x})$

(b) Decoder $q_\phi(\boldsymbol{x}|\boldsymbol{z})$

Figure 2.2: Architecture of a variational auto-encoder (VAE) with a Gaussian likelihood. Often the variance $\sigma_{\boldsymbol{x}}(\boldsymbol{z})$ is assumed to be constant.

It is possible to interpret the first term of the ELBO as a *reconstruction term*, accounting for how well p_θ is able to reconstruct $\boldsymbol{x}$ from a code $\boldsymbol{z}$ generated by q_ϕ. The second term can be interpreted as a *regularization term*, limiting the capacity of the latent representation to allow for the generation of novel samples of $\boldsymbol{x}$ by sampling the code $\boldsymbol{z}$ from the prior distribution $p(\boldsymbol{z})$. The training of a VAE attempts to select the model parameters θ and ϕ through gradient-based optimization, balancing the two terms of (2.3).

The diagram in Fig. 2.2 depicts the architecture of a VAE. In the VAE, the distributions q_ϕ and p_θ are usually instantiated as deep neural networks. For a given input $\boldsymbol{x}$, a first neural network — usually referred as the *encoder* — computes input-dependent mean $\mu_{\boldsymbol{z}}(\boldsymbol{x})$ and variance $\sigma_{\boldsymbol{z}}(\boldsymbol{x})$ that describes the approximate posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. Conversely, for continuous data $\boldsymbol{x}$, given a latent sample $\boldsymbol{z}$ a second neural network — usually referred as the *decoder* — computes a mean $\mu_{\boldsymbol{x}}(\boldsymbol{z})$ and variance $\sigma_{\boldsymbol{x}}(\boldsymbol{z})$ describing the likelihood distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. We can also interpret the encoder and decoder networks according to definition of representation and downstream maps, introduced in Section 2.1.

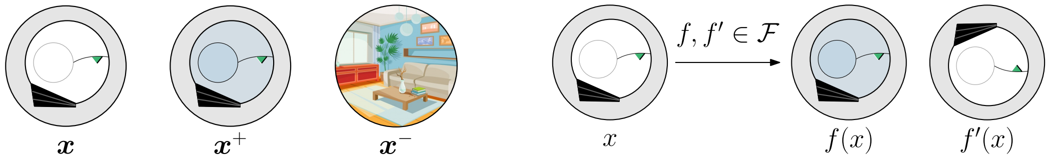
> **Definition 3** (Encoder and decoder) *An* encoder *network is a representation model $r = q_\phi : X \to \mathcal{Z}$, parameterized by ϕ, from data $\mathcal{X} \subset \mathbb{R}^M$ to a lower-dimensional representation space $\mathcal{Z} \subset \mathbb{R}^N$. A* decoder *network is a downstream model $g = p_\theta : \mathcal{Z} \to \mathcal{X}$, parameterized by θ, from representations $\mathcal{Z} \subset \mathbb{R}^N$ to a higher-dimensional data space $\mathcal{X} \subset \mathbb{R}^M$.*

2.1.2 Self-Supervised Representation Learning

In recent years, self-supervised representation learning approaches have shown remarkable performance, attaining results comparable to supervised approaches without explicit label information [178]. Of particular interest to this thesis, *contrastive* approaches aim at learning representation spaces where similar inputs are closer together and dissimilar inputs are far away, as depicted in Fig. 2.3a. Formally, this approach aims at learning a representation map $r : \mathcal{X} \to \mathcal{Z}$ by comparing inputs $\boldsymbol{x} \in \mathcal{X}$ according to some similarity score $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that $s(\boldsymbol{x}, \boldsymbol{x}^+) \geq s(\boldsymbol{x}, \boldsymbol{x}^-)$, with $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ similar and dissimilar inputs, respectively.

> **Definition 4** (Positive and negative pairs) *Given an anchor sample $\boldsymbol{x}$, a similar sample $\boldsymbol{x}^+$, and a dissimilar sample $\boldsymbol{x}^-$, the pair of inputs $\{(\boldsymbol{x}, \boldsymbol{x}^+)\}$ is denoted by* positive pair *and the pair of inputs $\{(\boldsymbol{x}, \boldsymbol{x}^-)\}$ is denoted by* negative pair.

Contrastive learning approaches fundamentally depend on the definition of similarity between different inputs. In a self-supervised setting, similarity is often defined by applying explicit *transformations* $f \in \mathcal{F}$, functions in a set of all transformations $\mathcal{F}$, over the



(a) Contrastive Pairs (b) Transformations

Figure 2.3: Self-supervised contrastive learning of data representations: (a) contrastive learning approaches aim at learning representations of data samples $\boldsymbol{x}$, where similar inputs $\{(\boldsymbol{x}, \boldsymbol{x}^+)\}$ are closer together and dissimilar pairs $\{(\boldsymbol{x}, \boldsymbol{x}^-)\}$ are far apart; (b) visual self-supervised approaches employ transformations $f \in \mathcal{F}$ to create similar pairs, without requiring explicit manual labeling.

input data $\boldsymbol{x}$. In modern visual contrastive learning, as depicted in Fig. 2.3b, image transformations over a data sample $\boldsymbol{x}$ such as cropping, scaling, color inversion are often used to define the similar samples $\boldsymbol{x}^+ = f(\boldsymbol{x})$, while different data points in the dataset are assumed to be negative samples $\boldsymbol{x}^-$ [71].

Of recent importance, the SimCLR framework learns representations of visual data by minimizing the contrast between differently augmented views of the same data sample [26]. The framework iteratively samples mini-batches $\mathcal{B} = \{\boldsymbol{x}_i\}_{i=1}^{B}$ and two different visual augmentations $f, f' \in \mathcal{F}$ are applied for for each sample $\boldsymbol{x}_i \in \mathcal{B}$, resulting in B positive pairs $\{(\boldsymbol{x}_i, \boldsymbol{x}_i^+)\}$:

$$\boldsymbol{x}_i = f(\boldsymbol{x}_i), \quad \boldsymbol{x}_i^+ = f'(\boldsymbol{x}_i), \tag{2.5}$$

and $B(B-1)$ negative pairs $\{(\boldsymbol{x}_i, \boldsymbol{x}_i^-)\}$. SimCLR employs a two-stage encoder function $r : X \to \mathcal{Z}$ to map the augmented data samples $\boldsymbol{x}$ to latent representations $\boldsymbol{z} \in \mathcal{Z}$. The similarity score s between two representations $\{(\boldsymbol{z}, \boldsymbol{z}')\}$ can be defined accordingly to:

$$s(\boldsymbol{z}, \boldsymbol{z}') = \exp(\mathrm{sim}(\boldsymbol{z}, \boldsymbol{z}')/\tau), \tag{2.6}$$

where $\mathrm{sim}(\boldsymbol{p}, \boldsymbol{q})$ is the cosine similarity between vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, scaled by the temperature parameter $\tau \in (0, \infty)$. The contrastive loss $\mathcal{L}_{\text{NT-XEnt}}$ can be computed accordingly:

$$\mathcal{L}_{\text{NT-XEnt}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = -\log \frac{s(\boldsymbol{z}_i, \boldsymbol{z}_j)}{\sum_{k=1}^{2B} \mathbb{1}_{[k \neq i]} s(\boldsymbol{z}_i, \boldsymbol{z}_k)}, \tag{2.7}$$

with the indicator function $\mathbb{1}_{[i \neq j]} = 1$ if $k \neq i$, and 0 otherwise. For contrastive approaches that rely on in-batch negatives samples, it becomes essential to employ large batch-sizes to learn meaningful representations, ensuring that the loss function can cover a diverse enough collection of negative samples.

2.2 Reinforcement Learning

A fundamental component of the work presented in this thesis concerns how an agent can learn to effectively perform sequential decision tasks under partial perceptual availability. To do so, we rely on *reinforcement learning* (RL), a computational framework that allows an agent to learn how to perform sequential decision-making tasks through trial-and-error interaction with its environment [154]. Reinforcement learning has been extensively addressed in literature, allowing agents to learn to perform complex tasks in virtual and real-world environments [2, 82, 141].

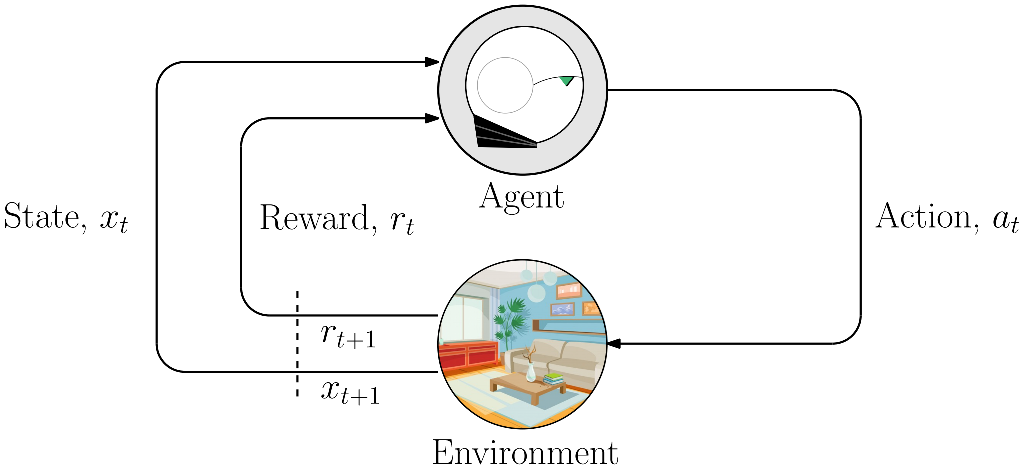


Figure 2.4: Overview of the reinforcement learning framework for sequential-decision making under uncertainty: at each time-step the agent receives the state of the environment x_t, performs an action a_t, receiving a reward r_t, and resulting in the transition of the environment to the next state x_{t+1}.

In the reinforcement learning framework, presented in Fig. 2.4, an agent interacts with a stochastic environment through a sequence of steps: at each time-step t, the agent is provided with the *state* of the environment x_t. Given state x_t, the agent performs an *action* a_t and receives an immediate *reward* r_t from the environment. As a consequence of such action, the environment transitions to state x_{t+1}. The long-term goal of the agent is to maximize the total reward it receives by carefully selecting its actions. A *policy* π designates a possible mapping from the perceived states to the actions of the agent. Reinforcement learning agents aim at learning an *optimal* policy, a policy which ensures that the agent collects as much reward as possible.

The reinforcement learning framework can be formalized as a *Markov decision process* (MDP) that describes a sequential decision problem under uncertainty with complete observability.

Definition 5 (Markov decision process) *A* Markov decision process $\mathcal{M}$ *is a tuple* $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P, r, \gamma)$*, where:*

- $\mathcal{X}$*, the* state space*, denotes the set of possible states of the environment;*
- $\mathcal{A}$*, the* action space*, denotes the set of possible actions that the agent is able to perform;*
- $\mathcal{P}$*, the* transition probabilities*, denotes the set of probability matrices that describe the dynamics of the environment where the agent operates. When the agent takes an action* $a \in \mathcal{A}$ *while in state* $x \in \mathcal{X}$*, the environment transitions to state* $y \in \mathcal{X}$ *with probability* $P(y \mid x, a)$*;*
- r*, the* reward function*, describes the process by which the agent receives instantaneous rewards from the environment as a function of its state and the action of the agent. When the agent takes an action* $a \in \mathcal{A}$ *while in state* $x \in \mathcal{X}$*, the agent receives an immediate expected reward* $r(x, a)$*;*

- $\gamma \in [0, 1)$, *the* discount factor, *sets the relative importance for the agent of present and future rewards.*

The reward function r, typically unknown to the agent, encodes the goal of the agent in its environment, whose dynamics are described by P, often unknown to the agent as well. The agent tries to learn a policy $\pi : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$, where $\pi(a \mid x)$ is the probability of performing action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$, that maximizes the expected discounted reward at each state $v_\pi(x)$,

$$v_\pi(x) \triangleq \mathbb{E}_{a_t \sim \pi(x_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \mid x_0 = x \right]. \tag{2.8}$$

Solving the MDP corresponds to computing an optimal policy $\pi^\star$ that verifies the following relation for all states $x \in \mathcal{X}$ and possible policies π,

$$v_{\pi^\star}(x) \geq v_\pi(x). \tag{2.9}$$

The optimal policy can be found from the optimal Q-function, defined recursively for every state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ as

$$Q^*(x, a) = r(x, a) + \gamma \sum_{y \in \mathcal{X}} P(y \mid x, a) \max_{a' \in \mathcal{A}} Q^*(y, a'). \tag{2.10}$$

Multiple methods can be used in computing this function [154], for example Q-learning [176]. Q-learning is an algorithm that allows an agent to learn the optimal action-value function $Q^\star$ in an online, incremental manner, from observed transitions (x_t, a_t, r_t, x_{t+1}). Q-learning converges to the optimal solution independently of the policy being followed, as long as all state-action pairs are visited by the agent infinitely often and the scalar step-size of the update α is sufficiently small [154]. The one-step Q-learning update is given by

$$Q(x_t, a_t) \leftarrow Q(x_t, a_t) + \alpha \left[r_t + \gamma \max_{a'} Q(x_{t+1}, a') - Q(x_t, a_t) \right], \tag{2.11}$$

where the the term $r_t + \max_{a'} Q(x_{t+1}, a')$ is often denoted by *target* of the update. The optimal policy, $\pi^\star$ can be directly computed from $Q^\star(x, a)$ by selecting the action that maximizes this function for each state $x \in \mathcal{X}$.

Q-learning and other reinforcement learning methods directly estimate a value function from the observed data, such as $Q^\star$, without requiring an explicit model of the environment, and thus are known as *value-based* methods. This is in contrast to *model-based* methods which require an explicit predictive model of the environment (P and r), learnt from data, before computing the value functions through, e.g., dynamic programming. Another class of methods directly optimize the policy from data are correspondingly named *policy-based* methods [84, 155].

More recently, deep learning research has addressed reinforcement learning problems, leading to new methods and extensions of classical algorithms [2]. The Deep Q-Network (DQN), a variant of the Q-learning algorithm, uses a deep neural network to parameterize an approximation of the Q-functions. To mitigate the effects of bootstrapping and off-policy sampling, which often causes Q-learning methods with function approximation to diverge [21], DQN proposes two techniques:

- *Target network*, an additional neural network with "fixed parameters" that estimates the value of the target value function (Eq. 2.11). The parameters of this network are infrequently replaced with the parameters of the latest approximation of the DQN neural network;
- *Experience replay*, a buffer that holds the history of the agent and that randomly samples transitions to update the Q-function estimate, thus reducing the correlation between samples and mitigating overfitting.

DQN has succeded in learning policies that beat Atari games [110]. However, DQN assumes discrete action spaces $\mathcal{A}$, which may not be suitable for control problems. For continuous action spaces, the Deep Deterministic Policy Gradient (DDPG), an actor-critic, policy gradient algorithm, has shown to perform well in complex control tasks [92].

2.2.1 Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) is the generalization of the RL framework for systems with more than one agent [189]. Extending the single-agent case, MARL can be formalized as a Markov game M_G, also known as *stochastic* game [93].

Definition 6 (Markov game) *A* Markov game *is a tuple* $M_G = ([n], \mathcal{X}, \mathcal{A}, \mathcal{P}, \boldsymbol{r}, \gamma)$, *where:*

- $[n] = \{1, \ldots, N\}$ *denotes the set of* $N > 1$ *agents;*
- $\mathcal{A} = \times^i \mathcal{A}^i$*, the joint action space, denotes the set of possible actions that all agents are able to perform, with* $\mathcal{A}^i$ *the finite set of actions of agent* i*;*
- $\mathcal{P}$*, the set of transition functions* $P(x'|x, \boldsymbol{a}) : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$*, denotes the probability of an transition from the current state* x *to the next state* x' *due to the joint action of all the agents* $\boldsymbol{a} \in \mathcal{A}$*;*
- $\boldsymbol{r}$ *denotes the set of reward functions* $\boldsymbol{r} = \{r^1, \ldots r^N\}$*, where* $r^i(x, \boldsymbol{a})$ *is the reward function that determines the immediate reward received by agent* $i \in [n]$.

At each step t, each agent $i \in [n]$ performs action a_t^i given the state of the environment $x_t \in \mathcal{X}$. The environment transitions to the next state x_{t+1}, as a function of the *joint* action selected by the agents $\boldsymbol{a}_t = \{a_t^1, \ldots, a_t^N\} \in \mathcal{A}$, rewarding each agent accordingly to $\boldsymbol{r}_t(x_t, \boldsymbol{a}_t)$. Similar to the single-agent case, the goal of each agent is to maximize the expected long-term reward, selecting actions accordingly to the policy $\pi^i : \mathcal{X} \times \mathcal{A}_i \to [0, 1]$. As the dynamics of the environment are a function of the joint action selected by all the agents, the agent-specific value-function $v^{\pi^i}(x)$ also becomes a function of the *joint* policy $\boldsymbol{\pi} := \prod_{i \in \mathcal{N}} \pi^i(a^i|x)$. We can define the agent-specific value-function $v^i_{\pi^i, \pi^{-i}}(x)$:

$$v^i_{\pi^i, \pi^{-i}}(x) \triangleq \mathbb{E}_{a_t^i \sim \pi^i(x_t), a_t^{-i} \sim \pi^{-i}(x_t)} \left[\sum_{t=0}^{\infty} \gamma^t r^i(x_t, \boldsymbol{a}_t) \mid x_0 = x \right], \tag{2.12}$$

where $-i$ represents the indices of all other agents besides agent i. Contrary to the single-agent case, the optimal performance of each agent is fundamentally dependent on the action-selection of all other agents. For infinite-horizon discounted Markov games, the most

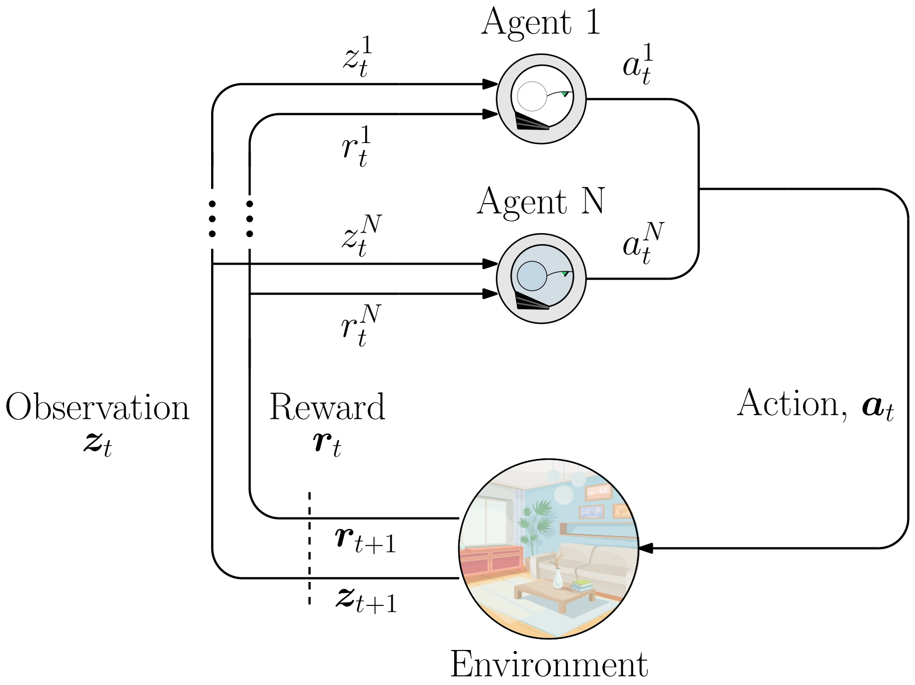


Figure 2.5: Overview of the multi-agent reinforcement learning (MARL) framework for sequential-decision making under partial-observability: at each time-step the agents receive a joint-observation of the environment $\boldsymbol{z}_t = \{z_t^1, \ldots, z_t^N\}$, and perform an joint-action $\boldsymbol{a}_t = \{a_t^1, \ldots, a_t^N\}$, receiving joint-rewards $\boldsymbol{r}_t = \{r_t^1, \ldots, r_t^N\}$.

common solution is a *Nash equilibrium* (NE) defined as a joint-policy $\boldsymbol{\pi}^\star = (\pi^{1,\star}, \ldots, \pi^{N,\star})$, such that:

$$v^i_{\pi^{i,\star},\pi^{-i,\star}}(x) \geq v^i_{\pi^i,\pi^{-i,\star}}(x), \quad \forall x \in \mathcal{X}, i \in [n] \tag{2.13}$$

In this thesis, we focus on *fully cooperative* Markov games. In this setting, all agents share a common reward function $r^1 = \ldots = r^N = r$. In particular, we consider fully-cooperative Markov games under *partial* observability: in this scenario, as shown in Fig. 2.5, the agents are not provided with the true state of the environment, x_t, but with an observation $\boldsymbol{z}_t \in \mathcal{Z}$ generated as a function of the true state and the joint-action previously performed. We formalize such scenario as a decentralized partially-observable MDP (Dec-POMDP) [117].

Definition 7 (Dec-POMDP) *A* decentralized partially-observable Markov decision problem *is a tuple* $\mathcal{D} = ([n], \mathcal{X}, \mathcal{Z}, \mathcal{O}, \mathcal{P}, \mathcal{A}, r, \gamma)$*, where:*

- $\mathcal{Z} = \times^i \mathcal{Z}^i$ *denotes the finite set of all possible observations of the agents in the environment;*

- $\mathcal{O}$*, the set of observation probability distributions* $O(\boldsymbol{z}|x', \boldsymbol{a}) : \mathcal{O} \times \mathcal{A} \times \mathcal{X}$*, denotes the probability of observing* $\boldsymbol{z} \in \mathcal{Z}$*, given the reached state* $x' \in \mathcal{X}$ *and taken joint action* $\boldsymbol{a} \in \mathcal{A}$.

MARL provides several benefits over the single-agent framework, such as the possibility to share experience amongst agents to accelerate learning: exchanging information through communication [195], in a teacher-learner setting [120], or by imitation [112]. However, MARL also introduces significant challenges: the exponential growth in computational complexity to RL algorithms such as Q-learning, due to the joint state-action space, and the non-stationarity in the policies that arise from the simultaneous learning of all agents in the system [16].

To mitigate the complexity of the joint state-action space, some approaches employ single-agent reinforcement learning algorithms to learn, independently across agents, *decentralized* policies. Under such policies, the agents are assumed to act only based on their individual observations. Despite not addressing the issue of the non-stationarity of the agents' policies, these approaches, such as independent Q-learning [157], can achieve reasonable performance in practice and are widely employed.

More recently, other approaches have been proposed that allow agents to learn in a centralized manner, avoiding the issues of coordination that arise with decentralized learning, and still execute their policies in a decentralized manner. One paramount example of such paradigm of *centralized training and decentralized execution* (CTDE) is the QMIX algorithm [132], which generalizes Q-learning to the CTDE setting. QMIX assumes that the agents share information during training time, in order to learn a value function of joint observations and actions. However, it constrains the learning process by requiring that the greedy evaluation of the joint value function Q_{tot} is the same as the greedy evaluation of each of the individual utility functions Q^i that compose it, i.e.,

$$\arg\max_u Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \begin{pmatrix} \arg\max_{a^1} Q^1(\tau^1, a^1) \\ \cdots \\ \arg\max_{a^N} Q^N(\tau^N, a^N) \end{pmatrix}, \tag{2.14}$$

where $\boldsymbol{\tau}$ is the joint action-observation history and τ^i is the action-observation history of the i-th agent. Such formulation can be achieved by imposing the monotonicity of the utility functions, i.e.,

$$\frac{\partial Q_{tot}}{\partial Q^n} \geq 0, \forall n \in \mathcal{N}. \tag{2.15}$$

This design choice allows the use of much more expressive families of functions, in comparison with Value Decomposition Networks (VDN) [153], which require Q_{tot} to be the sum of the individual Q^i. However, as they execute in a fully decentralized manner, the optimal policy obtainable by CTDE methods is still, in general, worse than an optimal centralized policy.

Chapter 3

Related Work

In this section, we present an overview of multimodal artificial intelligence, focused on works related to the contributions introduced by this thesis. In this thesis we wish to understand how artificial agents can learn multimodal representations of their environment to act under changing conditions of perceptual availability. As such, we start by discussing the case of human representation learning in Section 3.1. Following this, in Section 3.2 we discuss how computational systems can learn representation of human actions for classification tasks, a scenario we will use in Chapter 4 to motivate the need and highlight the benefits of multimodality. In Section 3.3, we discuss how deep neural networks can learn representations of high-dimensional data without supervision, in particular using deep variational autoencoders. Finally, in Section 3.4 we discuss literature in how representation learning plays a role in allowing reinforcement learning agents to act in their environments.

3.1 Human Representation Learning

In order to understand how autonomous agents can learn multimodal representations of their environment, we start by looking at the human case. Humans are provided with a complex cognitive framework that allows for multimodal perception. Several regions of the brain are responsible for the convergence of multimodal information, even in areas once thought to process only single-modality information [50]. These regions contain neurons that respond to stimuli from multiple sensory modalities [15, 17, 102, 104].

The *Convergence-Divergence Zone* (CDZ) framework is widely employed to explain the neural mechanisms of human perception [30, 108]. The CDZ model proposes two different sets of neuron ensembles, as depicted in Fig. 3.1: (i) lower-level ensembles in early sensory and motor cortices, responsible for the processing of modality-specific information; and (ii) higher-level ensembles in association cortices, responsible for the processing of multimodal information. According to the framework, the high-level (multimodal) neuron ensembles do not hold a composite version of the original perceptual information, but instead hold a record of the arrangement of the low-level neural ensemble activity generated by the perception of a given object [30]. The existence of higher-level multisensory convergence zones can be observed experimentally. The superior colliculus of the human midbrain contains multimodal neurons that respond to visual and auditory stimuli, in part responsible for the orienting behaviour of moving one's gaze towards the source of a sound [6, 39].

The CDZ framework also provides a graceful explanation of the human cross-modality inference process: the available perceptual information results in the activation of modality-specific low-level neural ensembles, whose activity patterns are forward-projected to the high-

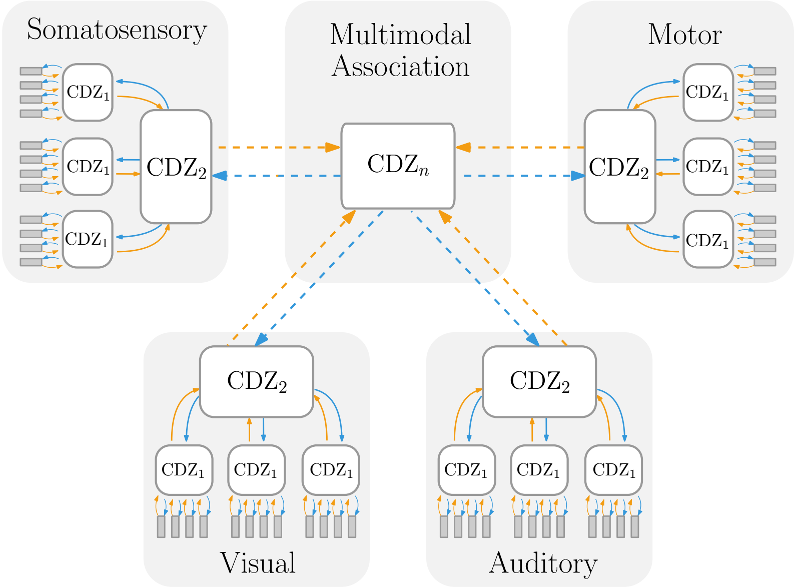


Figure 3.1: The CDZ framework, proposed by Damásio [30]. The model distinguishes between modality-specific neuron ensembles in early sensorimotor cortices and higher-order neural ensembles in multimodal association cortices. In the CDZ framework, information is propagated from the modality-specific cortices (orange arrows) to first-order CDZs, which, in turn, project back information (blue arrows) to the early cortical sites. Modality-specific information from low-order CDZs is propagated forward in order to encode a multimodal representation of the observed phenomena in higher-level association cortices (CDZ$_n$).

level multimodal ensembles. Subsequently, the high-level ensembles propagate information back to the modality-specific neural ensembles, inducing activity that is coherent with the (absent) perceptual phenomena. Such cross-modal activations have been observed experimentally as well. For example, the visual observation of lip movement (e.g., when observing a muted video clip) results in the retro-activation of early auditory cortices, whose activity pattern resembles that of the expected sound [10, 18]. Cross-modal activations have been observed for other sensory modalities, such as the activation of auditory and olfactory cortices by reading words with auditory or olfactory meaning [54, 76].

The activity of multimodal neurons resembles that of conceptual representations of external entities, such as other human beings [130]. These representational neurons respond to stimuli semantically coherent with a given entity regardless of the modality employed in its observation, be it photographs of the person, drawings of the person or even images of the name of the person. A similar multimodal behaviour has been recently observed in deep neural networks [53]. The parallel between biological and computational representation learning further motivates our work in multimodal representation learning for artificial agents, which we will discuss in the following sections.

3.2 Representation Learning of Human Actions

We start our discussion on learning representations in a computational setting by considering the scenario of human action recognition, which will be explored in Chapter 4 to motivate the need for multimodality and its benefits. The problem of learning representations of human actions, from complex ones such as running or cooking to more fine ones such as human gestures, have been extensively addressed in literature [28, 129]. Early approaches considered learning human action representations by considering only motion information. Xia et al. [180] proposed a novel view-invariant, histogram-based representation for human action recognition. The authors capture the dynamics of skeleton joint-angle information in the discretized 3D space, captured by a RGBD camera, clustering such dynamics into posture vocabularies. Actions are recognized by employing Hidden Markov Model (HMM) classifiers on sequences of vocabularies. The proposed system allows for real-time joint-information extraction, representation encoding and action classification. Similarly, Yang and Tian [182] employ a Naive Bayes Nearest Neighbor (NBNN) classifier in order to recognize action classes from video sequences. The authors propose a novel motion-based representation that considers the difference between joint positions in 3D space as a function of time. Moreover, the authors demonstrate that a low number of frames (15-20) is sufficient to attain reasonably accurate action recognition. To address the lack of interpretability of traditional action representations, Ofli et al. [115] propose an action representation based on the sequence of joints in the skeletal model of the human which, at each time instant, are considered to be the most informative. The informative criteria are based on predefined measures, such as the mean and variance of joint angle trajectories. The authors demonstrate the performance of their approach in cross-dataset recognition experiments, outperforming other motion-based approaches. Vemulapalli et al. [171] model an action as a curve in a Lie group manifold. The curve of each action is generated based on a skeletal-based representation that explicitly models the geometric relationships between various body parts using rotations and translations in 3D space. The authors experimentally demonstrate the performance of their approach for action recognition in multiple datasets. However all previous works require extensive manual feature engineering to extract human action representations suitable for downstream classification tasks.

3.2.1 Human Action Recognition with Deep Learning Models

Recently, deep-learning methods have also been proposed to address the problem of human action recognition with minimum manual feature engineering [188]. Ji et al. [74] considered the role of convolutional neural network (CNN) architectures for action recognition. The authors proposed a novel 3D CNN model, able to extract spatial and temporal features from video input, and the use of ensemble models to boost the performance of the method. The authors demonstrate the applicability of their approach in real-world surveillance videos. Other works explicitly model the temporal-dependence of human motion. Du et al. [37] proposed a hierarchical recurrent neural-network (RNN) framework for skeleton-based action recognition, in which the human skeleton is divided into five parts, according to the human's physical structure. Part-specific representations are fused hierarchically in higher-level components of the network. The authors demonstrate the effectiveness of the method, yet state the intrinsic difficulty in distinguishing action classes with similar joint-angle patterns. Simonyan and Zisserman [144] proposed a novel network architecture for action recognition in video that incorporates explicit spatial and temporal networks. Moreover, by decoupling spatial and temporal components, the authors are able to pretrain the spatial network on large-scale datasets, *e.g.* ImageNet, and subsequently fine-tune the network considering smaller-scale action recognition datasets. The results demonstrate that training the temporal network considering optical flow inputs instead of raw frames is preferable for small-scale datasets.

Similar to Ofli et al. [115], Liu et al. [94] proposed a novel class of Long-Short Temporal Memory (LSTM) network in order to automatically identify the most informative joints at each time step. The proposed architecture considers both joint-level and body-part attention mechanisms in order to focus, in each frame, on the most informative joints in the motion. Liu et al. [95] proposed a pipeline for view-invariant action recognition from video streams. The pipeline is composed of three steps: a sequence-based transform to cope with view variations, an image-based visualization method to represent joint-angle information and a CNN fusion model to extract features from joint-angle images. The authors evaluate the proposed method across multiple motion datasets, demonstrating the robustness of the method to partial occlusion. Song et al. [147] proposed an end-to-end deep LSTM model with spatio-temporal attention for action recognition. The spatial attention mechanism determines the informativeness of each joint-angle at the frame level. The temporal attention mechanism weights the importance of each frame during the complete motion.

While the prior deep-learning methods demonstrate impressive recognition results, they often require large training datasets to learn to perform classification tasks. Furthermore, by only considering motion information, the methods neglect the rich contextual background of an action, which is fundamental, for example, for the distinction of action classes with similar motion patterns.

3.2.2 Multimodal Human Action Recognition

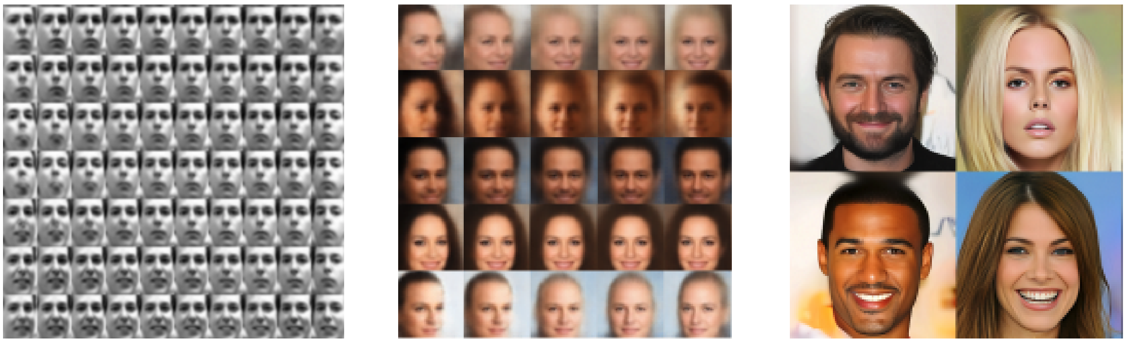
Other solutions explicitly consider multimodal information to learn representations of human actions. Yao et al. [183] proposed an “attributes and parts”-based representation of human actions, considering descriptive information (e.g., verbs associated with the action) and object information related to the action. However, this method uses only single image inputs and, as such, is unable to consider the dynamics of motion and contextual information present in human actions. Wang et al. [173] proposed features for depth data which capture human motion and human-object interaction data from a demonstration. These features, based on the local occupancy of skeleton joints during the motion, are transitional invariant and provide indirect object information. Furthermore, the authors introduce the notion of *actionlet*, a conjunction of features of a set of joints, and a data mining solution to identify relevant actionlets. An action class is then represented by a linear combination of actionlets, where the specific feature weights are learned employing a multiple kernel learning method. Recently, in the context of assistive robotics, Rodomagoulakis et al. [136] considered the fusion of information from image and sound modalities for action recognition. The authors employ pretrained acoustic models to compute features from sound information. Regarding image information, the authors compute features from a Bag of Visual Words (BoVW), clustered from trajectory information extracted from videos, allowing for classification with an SVM model. Information from single-modality features are merged into a multimodal representation as a linear combination of the (modality-specific) features. The authors evaluate their work on a novel dataset of multimodal interactions with a robotic platform.

The recent development of large-scale multimodal datasets for human action-recognition has spurred the development of novel methods that consider the fusion of multimodal information provided by the human user during actuation [24, 116]. Chen et al. [23] addressed the question of fusing depth image information and accelerometer signals provided by inertial body sensors for human action recognition. The authors consider merging information before classification (feature-level) or combining information from modality-specific classifiers (decision-level). The results attest to the improvement of accuracy in human action recognition when using multimodal information against using only single-modality information. Ahmad and Khan [1] proposed different strategies for multilevel multimodal fusion, i.e., across different depths of the network, of depth images and accelerometer information for action recognition. By employing CNNs, the proposed method is able to automatically extract features from the input modalities, which are fused throughout the network at different depths. The accuracy results of the multi-level fusion framework outperform single-level fusion state-of-the-art approaches. Proposed multimodal fusion approaches for action recognition are not limited to two modalities. Imran and Raman [70] proposed a novel approach for multimodal action recognition that considered the fusion of RGB video, 3D skeleton and accelerometer data. The authors propose a statistical feature fusion technique based on canonical correlation analysis (CCA), considering single-modality features extracted from CNNs and RNNs. The proposed method outperforms other state-of-the-art that consider only image and accelerometer information.

Current multimodal approaches attest to the importance of considering multiple sources of information for human action recognition. However, due to their dependence on large training datasets, current multimodal representation models are not suitable for sample-efficient recognition of human actions. Moreover, by only considering modalities intrinsic to the behavior of the demonstrator, the proposed methods neglect the rich contextual information of the action, such as the objects employed and location where the action is performed, which provides relevant information for its recognition. In Chapter 4, we propose to exploit the multimodal nature of human actions to provide sample-efficient representation learning for recognition tasks. In contrast with single-modality action representation methods, our multimodal approach allows for efficient distinction of action classes with similar motion patterns (*e.g.*, “Waving” and “Washing Window” actions). Our method also allows for the creation and recognition of action representations from a limited number of demonstrations, something that deep learning-based approaches struggle with. Finally, our method considers the context of the action demonstration in the learning process of a multimodal representation of human actions, effectively allowing for one-shot recognition.

3.3 Representation Learning with Variational Autoencoders

In this section, we discuss recent computational approaches that can be employed by artificial agents to learn representations of high-dimensional data without the need for manual feature engineering. *Deep Generative Models* (DGM) have been widely employed for representation learning, due to their ability to process complex modality information, such as images and sounds [9]. Several different classes of generative models have been proposed, each with unique trade-offs in architecture, training time and sampling efficiency. Early approaches considered energy-based solutions, which were prone to slow training and sampling [148]. Generative adversarial networks (GANs) have shown great promise in generating high-quality samples and scaling well to high-dimensional data [55]. However, their adversarial nature makes training GANs a difficult task. Recently, autoregressive models have also shown great promise in the generation of diverse, high-quality, samples. The diversity in the generation procedure comes at a cost of significant training and sampling time as well as memory requirements. In this section, we introduce recent works in representation learning considering variational autoencoder frameworks which are widely employed for representation learning, due to their low memory requirements, stable training



(a) Kingma and Welling [79] (b) Higgins et al. [66] (c) Vahdat and Kautz [164]

Figure 3.2: The evolution of generative capabilities of variational autoencoder frameworks (best viewed with zoom).

and fast sampling [80].

3.3.1 Variational Autoencoders

The variational autoencoder (VAE), introduced in Section 2.1.1, provides a straightforward method to learn the approximate distribution of single-modality data [79]. However, in more complex datasets, such as CelebA [96], the variational autoencoder tends to generate unrealistic and blurry samples [34], as shown in Fig. 3.2a. This behaviour is due to two fundamental reasons: the Gaussian likelihood in the training objective, that tends to average the reconstruction loss, and the use of a simple Gaussian prior, that attempts to naively regularize the latent space [194]. A higher capacity decoder model can be employed to minimize the role of the Gaussian likelihood yet that option often leads to posterior collapse, an issue where the decoder learns to disregard the information provided by the posterior distribution [62]. Other approaches attempt to model more complex prior distributions, for example, through normalizing flows, allowing the creation of complex distributions by sequentially mapping simple distributions through invertible functions [83]. The normalizing flow functions must be smooth, invertible, sufficiently expressive to model complex distributions, and computationally efficient. Rezende and Mohamed [135] introduced normalizing flows in the context of variational inference, to model more expressive approximate posterior distributions. The authors show that in an asymptotic regime with infinitesimal flows, the proposed framework is able to recover the true posterior distribution. Despite the intrinsic restrictions of normalizing flow functions, several approaches have been proposed to increase the representation power and computational efficiency of these methods [69, 81, 122].

3.3.2 Hierarchical Variational Autoencoders

Other works attempt to model complex priors by introducing hierarchical structures in traditional single-level VAEs. The vector-quantized variational autoencoder (VQ-VAE) employs an autoregressive model to learn such prior [165]. In this two-stage architecture, an autoencoder model initially learns a discrete latent space, defining a codebook for the training data. After learning the codebook, the authors fit an autoregressive model over the codebook, in order to generate data. Sønderby et al. [146] introduced the ladder variational autoencoder (LVAE), a generative model that considers a top-down structured inference model. The generative and inference networks share parameters, thus allowing to learn latent representations that consider data-specific bottom-up information and top-down prior information. The qualitative results reveal that LVAE is able to learn structured high-level representations in the MNIST dataset, allowing for efficient distinction between different digit classes in the latent space. Zhao et al. [193] addressed the limitations of LVAE in terms of representation efficiency and feature learning and proposed the Variational Ladder Autoencoder (VLAE). This model assumes that the abstraction level of the representation is correlated with the expressiveness of the neural network employed for inference and generative mapping. The authors show qualitatively in different natural image datasets that the model learns structured features at different representation levels, in line with the abstraction assumption employed. Recently, hierarchical VAEs have shown remarkable performance in image generation through careful design of neural architectures [164]. The nouveau VAE (NVAE) introduced several architectural improvements, such as depthwise convolutions in the generative model, and implementation techniques that aimed at stabilizing the training of very deep variational autoencoders, mitigating the effects of posterior collapse. The proposed architecture is the first VAE able to generate high-quality 256×256 pixel images, achieving state-of-the-art results on several natural image datasets, as shown in Fig. 3.2c. In our work, we explore hierarchy in order to generate high-quality samples from a multimodal representation, regardless of the intrinsic complexity of the target modality.

3.3.3 Disentanglement in Variational Autoencoders

Variational autoencoder frameworks have also been employed to learn interpretable representations. Accordingly to Bengio et al. [8], learning a factorized representation that accounts for the data generative factors of the world might provide artificial agents with mechanisms to attain a deeper understanding of their environment. The question of learning a *disentangled representation* from image data using VAEs was first introduced by Higgins et al. [66]. The authors introduced a constraint over the capacity of the latent representation, effectively controlling the regularization component of the evidence lower-bound of the model (Eq. 2.3) with a novel hyperparameter β such that

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \beta D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})). \tag{3.1}$$

In addition to the proposed framework, denoted by β-VAE, the authors also proposed a novel disentanglement metric to measure the independence and interpretability of the learnt representation. To do so, they sample sets of latent representations, fixing the value of a single latent index, and generate images from the sampled representations. Subsequently, the sets of images are encoded into latent spaces and the absolute linear difference between the inferred latent representations are provided to a linear classifier in order to predict the corresponding generative factor (index of the latent space) that was kept fixed. The disentanglement metric score is computed from the accuracy of this predictor. The results show that β-VAE outperforms the standard VAE in regards to learning a disentangled representation, without requiring any prior knowledge regarding intrinsic factors that were responsible for the generation of the data. However, by restricting the capacity of the latent representation, the approach introduces a trade-off between disentanglement and reconstruction quality. To mitigate the effects of such tradeoff, Burgess et al. [14] evaluated the original β-VAE from an information bottleneck perspective. The authors introduced a novel training objective with controlled capacity that allowed for better reconstruction fidelity,

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \gamma \left| D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})) - C \right| . \tag{3.2}$$

where the capacity C of the latent representation is linearly increased throughout the training. The methodology of linearly increasing the capacity of the representation during training is quite similar to the *warm-up* training procedure introduced by Sønderby et al. [146] to mitigate the effects of early *latent collapse*, i.e., a condition where latent variables would follow the prior distribution in the early stages of training and be stuck in a local minimum during the remaining training procedure. The results of Burgess et al. [14] on simple visual datasets show that (3.2) promotes the disentanglement of the latent representation, while minimizing the decrease in image reconstruction quality. To further identify the origin of disentanglement due the training of β-VAE, Chen et al. [25] proposed a novel training objective that explicitly considers the correlation between the individual latent codes. The authors proposed the Total Correlation VAE (β-TCVAE) and a new information-theoretic disentanglement metric which does not rely on pretrained classifiers, such as the one employed in the original work of Higgins et al. [66] that is generalizable to more complex, arbitrarily distributed, latent representations. The results show that β-TCVAE is more likely to learn disentangled representations than the original β-VAE, improving the robustness of the methods to random initialization.

The learning of disentangled representations using VAEs remains an open question, further motivated by the potential benefit of employing disentangled representations for sample-efficient learning in downstream tasks. However, the benefits and feasibility of learning disentangled representations without supervision were recently challenged by Locatello et al. [98]. The authors show that learning disentangled representations is fundamentally impossible for arbitrary generative models without introducing inductive biases. Furthermore, the authors performed a large-scale evaluation of more than 12,000 instances of generative models, considering different evaluation metrics and seven different datasets. The results clearly indicate that initialization factors (such as the seed) and training hyperparameters appear to have a more fundamental role in the disentanglement results than the choice of model and training objective. Furthermore, models with increased disentanglement scores did not provide increased sample-efficiency of learning downstream tasks. The authors conclude that future research in learning disentanglement representations should address the concrete benefits of having such property, while providing an extensive, and reproducible, evaluation of the proposed methods.

3.3.4 Multimodal Variational Autoencoders

All the approaches mentioned so far consider single-modality data. In this thesis, we explore learning joint-representations of multiple sources of heterogeneous data. VAEs have also been adapted to consider multimodal data. The initial multimodal extension of the Variational Autoencoder (VAE) model considered approximating the individual representations encoded from each modality and each subset of modalities [156, 184]. This approach has been extended for scenarios with 3 modalities [85] and considering pretrained representation models [159]. However, the necessity of instantiating individual neural-networks for each modality (and combination of modalities) hinders their applicability in scenarios with a larger number of modalities. To address this scalability issue, two approaches learn instead an *aggregated* multimodal representation, differing only on the function responsible for aggregating the individual modality representations. The Multimodal VAE (MVAE) employs a Product-of-Experts solution to generate the multimodal representation [179].

The Mixture-of-Experts Multimodal VAE (MMVAE) considers a Mixture-of-Experts to encode an *implicit* multimodal representation [142]. However both models have significant limitations: the MVAE is prone to overconfident expert prediction, learning a representation that neglects information from the lower-dimensional modalities, and MMVAE does not employ joint-modality information to training the latent representation. Moreover, all models parameterize an equal representation space for all modalities, regardless of their nature or complexity. In Section 5.1 we provide an in-depth discussion of different methods for multimodal representation learning using VAE frameworks, focusing on cross-modality generation.

In Chapter 5 we address the problem of learning a multimodal representation of an arbitrary number of modalities without supervision within the VAE framework. Contrasting with current models, we consider a hierarchical architecture to allow for efficient cross-modality inference able to generate high-quality, varied, and semantically correct data, regardless of the intrinsic complexity of the corresponding modality.

3.4 Representations for Reinforcement Learning

Reinforcement Learning (RL) agents have widely employed representation models to learn to act in their often complex environment. Such representations can encompass the structure of the environment (state representations) [90] or of the actions of the agent (action representations) [22]. In this section, we consider works that learn state representations from observations provided by the environment to the agent.

3.4.1 Abstractions in RL

Low-dimensional state representations have been widely employed to learn efficiently how to perform tasks in a given environment. The approach of *state abstraction* considers mapping the original, often-high-dimensional, state space into a lower-dimensional, compact, representation space, while preserving some properties of the original space [91]. Several approaches for state abstraction have been proposed. Bissimulation considers the grouping of states such that they are indistinguishable in terms of reward for any sequence of actions [87]. Bissimulation metrics have been proposed to measure the similarity between states [43, 52]. Another approach for state abstraction is to consider *homomorphisms* [133, 158], which define a map that matches not only equivalent states, but also equivalent actions in such states. However, classical state abstraction methods are not easily scalable to scenarios with large-scale state spaces or complex dynamics.

Recent works have proposed to extend state abstraction techniques to complex scenarios, employing modern neural networks. Gelada et al. [49] have proposed DeepMDP, a novel latent space model that introduces a low-dimensional latent representation of an MDP with theoretical guarantees regarding the quality of the approximated value function in comparison with the value function in the original MDP. In a partially-observable Markov decision problem (POMDP) setting, Zhang et al. [190] explored learning latent representations that encode local representations able to provide suitable gradient directions for policy improvement. The evaluation of both approaches attests to the potential of using latent representations of large-scale state spaces for sample-efficient RL training, regarding the performance of the agent in the task but also the sample-efficiency of the methods.

Differing from the above approaches, we can also consider the intrinsic partitioning of the MDP as a property of the states themselves. *Factored MDPs* consider that the global state of the environment is defined by a set of partial, weakly-dependent, state variables, and employ a dynamic Bayesian network [31] to represent the transition model of large structured MDPs [11]. Moreover, the framework assumes the rewards and transitions exhibit a certain level of conditional independence, i.e., the factorized reward and the transition functions involve only a small subset of partial state variables. Several works have shown that such factored approaches can effectively reduce the complexity of learning [27, 121].

In this thesis, we focus on learning representation models for RL agents able to perceive multimodal observations. Similar to the above approaches, we learn low-dimensional representations of the original, high-dimensional state space. However, differing from state abstraction approaches, in this thesis, we focus on the problem of learning global state representations from information provided by different modalities. Moreover, differing from factored MDPs, we do not factorize the reward and transition functions per modality (or combination of modalities) and instead assume that the information provided by each modality is able to recover the global state of the environment.

3.4.2 Adaptation in RL

Another perspective on representation learning for RL agents concerns the adaptation of the agent to similar domains [8]. The use of variational autoencoders for such a purpose was first introduced by Higgins et al. [65] to provide zero-shot policy transfer to a target domain. The proposed Disentangled Representation Learning Agent (DARLA), depicted in Fig. 3.3a, is trained accordingly to a multi-stage approach:

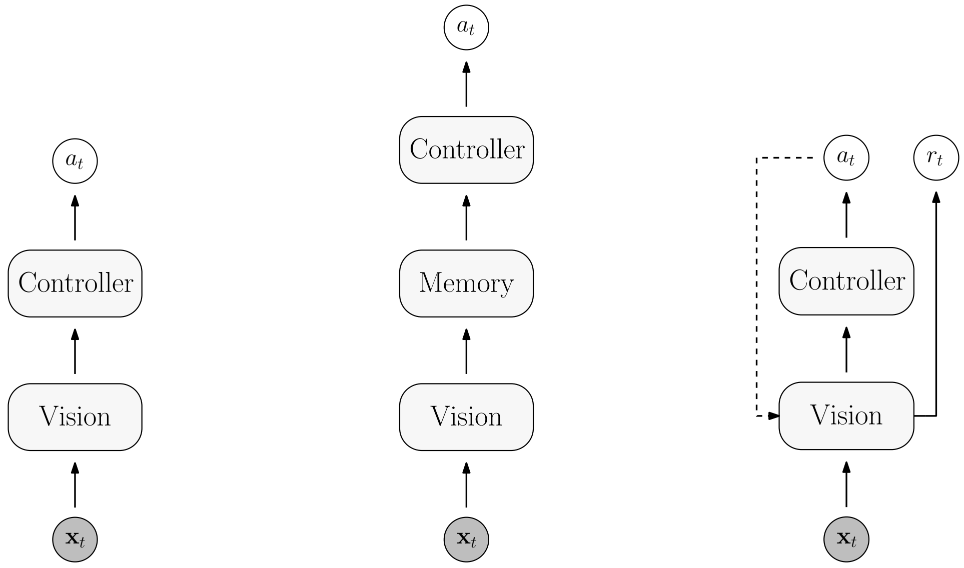
1. *Learn to see* - Train an unsupervised vision model, such as the β-VAE, to learn a disentangled representation from observations provided by the source environment, collected by the agent (often following a random policy);

2. *Learn to act* - Learn a controller policy on the source domain, employing a standard reinforcement learning algorithm over the previously trained latent representation (fixed at this point);

3. *Transfer* - Evaluate the source domain policy on a target domain without retraining the policy in the novel environment, i.e., in a zero-shot adaptation task.

The authors evaluate the proposed agent in simulation environments and in *sim-to-real* tasks, showing robustness to domain adaption and employing different reinforcement learning algorithms. Recently, another approach has exploited large-scale datasets and sequence models to promote adaptation to different environments. Reed et al. [134] introduced a generalist agent (GATO), that learns a single generalist policy for vision, language and control tasks, from multiple large-scale datasets. The authors employ a transformer-based architecture to learn the policy and show how their agent is able to generalize to out-of-distribution tasks. Moreover, the results show the important role of *scaling* up both number of parameters of the model and the amount of data to achieve such results.

In Chapter 7, we introduce the problem of *multimodal transfer in reinforcement learning*, where the agent must adapt to different perceptual conditions, and not to different domains. Under these conditions, at execution time, the agent may be provided with observations from a (possibly) different set of modalities, in comparison with the one provided while training the policy.



(a) Higgins et al. [65] (b) Ha and Schmidhuber [57] (c) Hafner et al. [58]

Figure 3.3: Simplified architecture of relevant works in representation learning for deep reinforcement learning agents: given the observation of the environment $\boldsymbol{x}_t$, the agents act accordingly to selected action a_t. The Dreamer model additionally also predicts the reward r_t associated with the current representation of the world.

3.4.3 Representation Learning for Model-Based RL

Another seminal work in representation learning for model-based RL concerns the *world model* framework, proposed by Ha and Schmidhuber [57]. The world model allows an agent to explicitly build a spatial and temporal representation of its environment before learning how to perform a task. The overall structure of world models are presented in Fig. 3.3b. Similarly to the DARLA agent, world model agents are sequentially trained:

1. Training the *Vision* module (V): In a first stage, the agent learns a spatial representation from observations collected from the environment, training a VAE to compress each image into the latent state (following a random policy);

2. Training the *Memory* module (M): Subsequently, the agent learns a temporal representation of the dynamics of the environment. The authors employ a RNN with a Mixture Density Network output layer (MDN-RNN) to predict the distribution of future latent states, given the current latent state and the action of the agent;

3. Training the *Controller* module (C): Finally, the agent learns to act in order to maximize the expected cumulative reward attained during a rollout in the environment. The authors employ a linear layer model for the controller, that outputs the action of the agent, given the current state of the world and the hidden state of the RNN memory.

By having a spatial and temporal representation of the environment, the agent is able to *hallucinate*: create a dream-like virtual environment by sampling latent states from MDN-RNN and generating the associated image using the VAE decoder. The authors show that a policy learnt from the real environment can be zero-shot transferred to the dream environment. Moreover the agent is also able to do the inverse: train a controller *inside the dream* environment and transfer such policy to the real environment. Several extensions of the world models frameworks have been proposed. Hafner et al. [58] proposed Dreamer, depicted in Fig. 3.3c, a novel world-model-based RL agent that learns to perform long-horizon tasks entirely in an hallucinated environment. The authors introduce a state value model to estimate rewards beyond the horizon of the imagined rollouts given the latent state of the vision module. The gradients are backpropagated in time through the dynamics of representation model, using the reparametrization trick to overcome the sampling operation of the latent representation. The results show that Dreamer outperforms prior methods on several long-horizon continuous control tasks regarding overall performance and data-efficiency. Recently, Hafner et al. [59] proposed Dreamer v2, an extension of the previous RL agent that employs a categorical latent space and introduces a learnt prior distribution on the vision module representation. The results show that the updated agent is able to achieve human-level performance on a benchmark of 55 Atari games by learning behaviors *inside the dream* of the agent — using information collected by the world model of the agent, without requiring interaction with the real environment.

However, all works in RL previously discussed assume that the agent is provided only, and always, with image information regarding its environment, and the adaptation task concerns the transfer of the policy learnt to different environments. These perceptual assumptions hinder the application of such methods in artificial agents, such as robots, which can be provided with multiple sensors to robustly probe the state of the environment. In Chapter 7, we present a novel adaptation problem for RL agents: how can an agent perform zero-shot transfer of a policy learnt on a set of perceptual modalities to settings where the environment provides a different set of perceptual observations. Following DARLA and world models approaches, we propose a multi-stage framework that effectively allows agents to transfer such policies across different modalities.

3.4.4 Representation Learning for Multi-Agent RL

Learning representations also play a role in multi-agent reinforcement learning, in particular in regards to addressing partial observability. Omidshafiei et al. [119] propose a decentralized MARL algorithm that uses RNNs to improve the observability of the agents, learning an implicit representation of the missing observations. Mao et al. [103] use an RNN to first compress the histories of the agents into embeddings, that are subsequently fed into deep Q-networks, helping to improve the observability of the agents. The commonly used paradigm of centralized training with decentralized execution also helps to reduce the inherent partial observability of the environment at training time, as every agent has access to the joint-observation of all agents [47, 48, 118, 132]. Under such paradigm, the calculation of value functions or policy gradients can exploit the centralization of information, thus alleviating partial observability.

Other approaches explore communication between the agents to alleviate the problem of partial observability in MARL [195]. Early works addressed communication in partially observable cooperative MARL tasks: Sukhbaatar et al. [152] share the outputs of the hidden layers of a shared neural network among the agents; Foerster et al. [47] explicitly learn the content of the messages transmitted between agents by following an end-to-end approach in which gradients are back-propagated through the communication variables. Recent works explore *how* [77, 114], *when* [68, 145], and *what* [47] should be communicated among the agents in order to foster cooperation. In Chapter 8 we assume that the agents have no control over when and with whom to communicate and, instead, should robustly perform under any type of communication policy. Moreover, we are not focused on learning the content of the messages being communicated (as in the work of Foerster et al. [47]), focusing our attention on "passive" communication by considering the sharing of local observations and actions among the agents.

Other works consider learning communication protocols robust to failures or missing information, by either limiting the variance of exchanged messages [191], or temporally smoothing information shared between agents [192], which is out of scope for the passive communication mechanism explored in this thesis. Kim et al. [78] propose a learning technique for MARL called message-dropout, which aims at: (i) effectively handling the increased input dimension in MARL with communication; and (ii) making learning robust against communication errors in the execution phase. Message-dropout drops the messages received from other agents independently at random during training before inputting them into the RL algorithm. In a similar fashion to message dropout, Wang et al. [174] propose a recurrent actor-critic algorithm for handling multi-agent coordination under partial observability with limited communication, showing that recurrency successfully contributes to robust performance when communication fails. In Chapter 8, we employ message-dropout with recurrent learners as a baseline, following both Kim et al. [78] and Wang et al. [174].

Several works have also explored ways to explicitly model information about the other agents, such as their actions and observations, based on local information available to the learning agent. The work of Papoudakis et al. [123] uses a recurrent neural network to predict the actions and observations of other agents in order to make better action selections in a centralized training with decentralized execution setting. At execution time, the agent then uses its learned model to make explicit predictions about other agents' observations and actions. Xie et al. [181] does similarly in a latent space. The work of He et al. [61] uses, instead, the other agents' observations to predict their actions, which assumes centralization will be available at execution time. In Chapter 8 we learn policies for multiple agents. To that end, we use a model of the observations and actions of the agents to make predictions about other agents and improve their performance in cooperative tasks.

Chapter 4

Learning Multimodal Representations of Human Actions

"Recognizing isn't at all like seeing; the two often don't even agree."
Sten Nadolny

Humans are able to interact with their environment in rich and diverse ways. This diversity arises not only from the large number of actions able to be performed in an environment but also from the variety of ways that a single action can be performed. Moreover, the same action can often be performed employing a wide range of objects. In addition, actions can also be context-dependent, whose meaning is shaped by the agents interacted with and by the environment where they are performed. The diversity in number, form and context of human actions hinders the goal of training an artificial agent that is readily able to recognize all possible actions performed by a human user. To address this issue, we can instead program agents to learn to recognize new human actions from demonstrations. However, it is unrealistic to assume that such learning will depend on large amounts of data, as required by many current action recognition algorithms. Instead, the agent should be able to learn and recognize novel actions from just a few demonstrations provided by the human.

In this chapter we motivate the need and highlight the benefits of multimodal representation learning in the context of sample-efficient action recognition, addressing the problem of learning to recognize human actions from few demonstrations provided by a user in a household environment. In contrast to conventional methods, we leverage information beyond motion data and consider the contextual background of the demonstration in order to encode *motion concepts*, a novel probabilistic multimodal representation of human actions. By considering contextual information, motion concepts allow the distinction between context-dependent actions with similar motion patterns and are effectively suitable for few-shot recognition tasks. Furthermore, we introduce a novel algorithm that allows the online learning and efficient recognition of motion concepts from demonstrations provided by the user. We evaluate motion concepts on two complementary tasks: an one-shot offline recognition task and in an online learning and recognition task with a human demonstrator. The results show that motion concepts outperforms single-modality motion-based representations, highlighting the benefits of considering multimodal representations for efficient recognition of human actions.

The main contributions of this chapter are three-fold:

- In Section 4.1, we contribute *motion concepts*, a novel multimodal representation of human actions. A motion concept encompasses a probabilistic description of the motion patterns observed, augmenting it with their contextual background information, namely the location of the action and the objects used during the demonstrations. Moreover, the motion concept takes into account information provided directly through interaction with a human and allows the agent to reason about the importance of each contextual feature for its recognition;

- In Section 4.2, we introduce the *Online Motion Concept Learning* (OMCL) algorithm, allowing the creation of new motion concepts through interaction with a human user. The algorithm is able to recognize motion concepts from a single demonstration and continuously update motion concepts as more demonstrations are provided;

- In Section 4.3, we evaluate motion concepts within a virtual-reality household environment. Initially, we instantiate an offline *one-shot* action recognition task (Section 4.3.2), revealing the importance of contextual information for the recognition of motion concepts built from a single training demonstration; Moreover, we evaluate the algorithm's ability to learn novel actions and recognize previously learnt actions in an online *tabula rasa* scenario (Section 4.3.3), i.e., a scenario in which the system, prior to the evaluation, is not trained on any data.

The work described in this chapter has been published in:

- **Miguel Vasco**, Francisco S. Melo, David Martins de Matos, Ana Paiva, and Tetsunari Inamura. *Online Motion Concept Learning: A Novel Algorithm for Sample-Efficient Learning and Recognition of Human Actions.* Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). 2019, pp. 2244–2246 [166];

- **Miguel Vasco**, Francisco S. Melo, David Martins de Matos, Ana Paiva, and Tetsunari Inamura. *Learning Multimodal Representations for Sample-efficient Recognition of Human Actions.* Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. 2019, pp. 4288–429 [167].

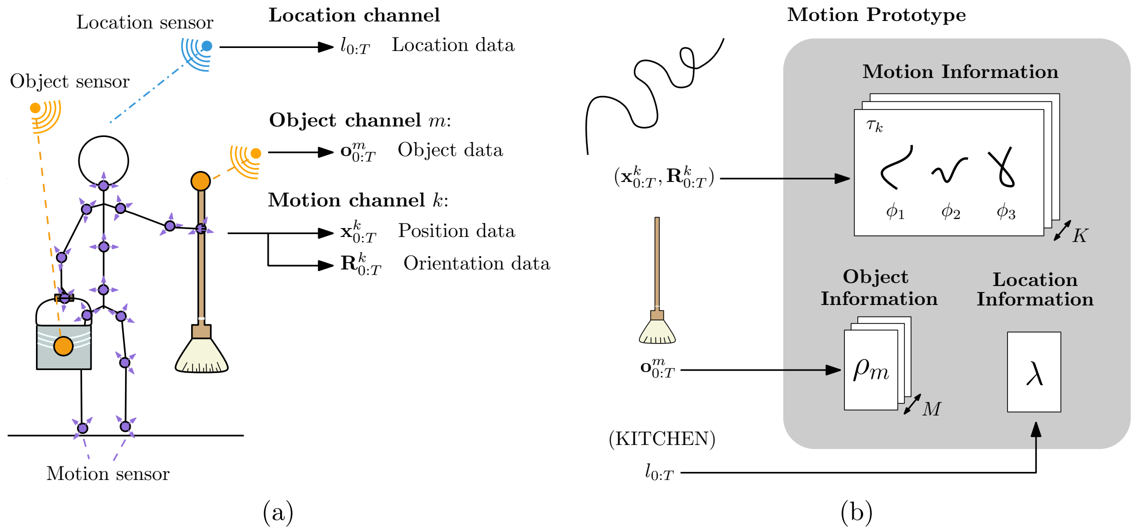


Figure 4.1: The multimodal nature of human action demonstrations: (a) we consider a learning from demonstration setup in which the environment is engineered with sensors providing information regarding the motion of the human, the objects interacted with and the location where the actions are performed; (b) from a single demonstration we extract motion (pose and orientation), object and location data to create a *motion prototype*.

4.1 Multimodal Representation of an Action

4.1.1 Setup

We consider the following setup for learning from demonstration: a human user demonstrates an action, which may involve interaction with objects in the environment. The environment comprises a number of locations of interest, and the human user may be in any of these locations at the time of demonstration.

We assume that the environment is engineered with a number of sensors, providing information regarding the *location* and *pose* of the human user as well as the *objects* that the user interacts with (see Fig. 4.1a for an illustration). The sensors act as input channels for the system, and as such we henceforth refer to sensors generally as a *channels*: a sensor deployed to provide object information is referred simply as an *object channel*, sensors deployed to track the human pose are referred as *motion channels* and the location of the user in the environment is provided by a dedicated sensing module, referred to as the *location channel*. A demonstration by a human user yields a number of data streams arising from the different input channels. In particular,

- Each motion channel k, $k = 1, \ldots, K$, provides two streams of length T, $\boldsymbol{x}^k_{0:T}$ and $\boldsymbol{R}^k_{0:T}$, where each $\boldsymbol{x}^k_t$ indicates the position of a body element (joint, limb) at time step t, with $t = 0, \ldots, T$, measured with respect to a common fixed world frame, and each $\boldsymbol{R}^k_t$ is a rotation matrix representing the orientation of that same body element at time step t;

- Each object channel m, $m = 1, \ldots, M$, provides one stream of length T, $\boldsymbol{o}^m_{0:T}$, where each individual observation $\boldsymbol{o}^m_t$ corresponds to a binary vector indicating the objects, from a predefined finite set of objects $\mathcal{O}$, that the user is interacting with at time step t, with $t = 0, \ldots, T$, according to channel m;

- Finally, the location module provides a stream of length T, $l_{0:T}$, where l_t indicates the location of the user at time step t. We assume that the location of the user takes values in a finite set $\mathcal{L}$ of possible locations.

To learn a general representation of an action (the *motion concept*), we start by taking the streams from the different input channels, provided by a single demonstration, and compile them into a unique, compact representation that we refer to as a *motion prototype*.

4.1.2 Motion Prototype

We introduce a low-level multimodal representation of a single demonstration of an action by a human user, which we denote by *motion prototype*. In particular, motion prototypes capture in a probabilistic manner motion, object and location information.

Definition 8 (Motion prototype) *A* motion prototype *is a tuple $P = (\boldsymbol{\tau}, \boldsymbol{\rho}, \lambda)$, where $(\boldsymbol{\rho}, \lambda)$ summarize the associated context information - namely object and location information - and $\boldsymbol{\tau}$ summarizes the motion observed in the demonstration:*

- *The object representation, $\boldsymbol{\rho} = \{\rho_m, m = 1, ..., M\}$, where M is the total number of object channels. For every object $o \in \mathcal{O}$,*

$$\rho_m(o) = \mathbb{P}\left[o^m_{t,o} = 1, t = 0, ..., T\right],$$

  where $o^m_{t,o}$ is a random variable indication whether object o was observed in object channel m at time step t. We assume that the observation of an object $o \in \mathcal{O}$ in channel m at any moment during the demonstration can be described as a Bernoulli random variable with parameter $\rho^m(o)$;

- *The location representation, λ, where, for every location $l \in \mathcal{L}$,*

$$\lambda(l) = \mathbb{P}\left[l_t = l, t = 0, ..., T\right],$$

  where l_t is a random variable indicating the location of the user at time step t. We assume that the location of the user during the demonstration can be described as a categorical distribution with parameters $\lambda(l)$, $l \in \mathcal{L}$;

- *The motion representation, $\boldsymbol{\tau} = \{\tau_k, k = 1, ..., K\}$, where K is the number of motion channels and each τ_k is a sequence of motion primitives $\{\phi_n, n = 1, ..., N\}$.*

Motion primitives are a concept widely explored to describe and represent animal and robot motion [46, 83]. In this work, a motion primitive ϕ_n is a probability distribution over the space of trajectories: given an arbitrary trajectory $(\boldsymbol{x}_{0:T}, \boldsymbol{R}_{0:T})$,

$$\phi_n\left(\boldsymbol{x}_{0:T}, \boldsymbol{R}_{0:T}\right) = \mathbb{P}\left[\mathrm{x}^n_{0:T} = \boldsymbol{x}_{0:T}, \mathrm{R}^n_{0:T} = \boldsymbol{R}_{0:T}\right].$$

For the purpose of learning and recognition, it is convenient to treat an action not as comprising a single trajectory $(\boldsymbol{x}_{0:T}, \boldsymbol{R}_{0:T})$ but, instead, as a sequence of smaller trajectories,

$$\left\{\left(\boldsymbol{x}_{0:t_1}, \boldsymbol{R}_{0:t_1}\right), \left(\boldsymbol{x}_{t_1:t_2}, \boldsymbol{R}_{t_1:t_2}\right), ..., \left(\boldsymbol{x}_{t_{N-1}:t_N}, \boldsymbol{R}_{t_{N-1}:t_N}\right)\right\},$$

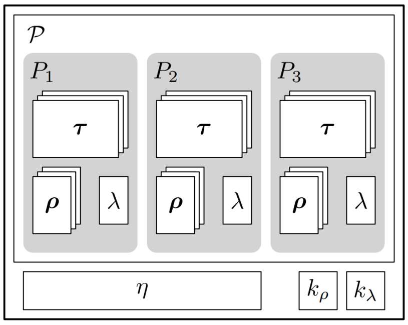


Figure 4.2: The schematic representation of a *motion concept*: a motion concept of action a is composed by a list of motion prototypes associated to that action class $\mathbf{P}$, a designation of the action class η and constants k_p and k_λ, that weight the importance of object and location information for the recognition of the action.

.

which are then encoded as a sequence of motion primitives $\{\phi_n, n = 1, ..., N\}$, each ϕ_n providing a compact description of $(\boldsymbol{x}_{t_{n-1}:t_n}, \boldsymbol{R}_{t_{n-1}:t_n})$. Each motion primitive ϕ_n is selected among a library Φ of available motion primitives to maximize the likelihood of the observed trajectory, i.e.,

$$\phi_n = \arg\max_{\phi \in \Phi} \phi \left(\boldsymbol{x}_{t_{n-1}:t_n}, \boldsymbol{R}_{t_{n-1}:t_n}\right). \tag{4.1}$$

Summarizing, a motor prototype compactly encodes a demonstration of an action in the form of a tuple $(\boldsymbol{\tau}, \boldsymbol{\rho}, \lambda)$, where $\boldsymbol{\tau}$ is a collection of trajectories (one for each motion channel), each represented as a sequence of motor primitives; $\boldsymbol{\rho}$ is a collection of probability distributions (one for each object channel), describing how the human interacted with each object in the environment; and λ is a probability distribution describing where the human was located during the demonstration (see Fig. 4.1b).

4.1.3 Motion Concepts

It is possible for a single action to be performed in multiple ways. A motion prototype, while providing a convenient representation for a single demonstration (and corresponding context), is insufficient to capture the diversity that a broader notion of *action* entails. As such, we extend motion prototypes and propose a *higher-level* multimodal representation of an action performed by human users, which we denote by *motion concept*, depicted in Fig. 4.2. A motion concept seeks to accommodate the different ways by which an action may be performed while, at the same time, encode distinctive aspects that are central in recognizing such action. For example, to distinguish actions such as *waving goodbye* and *washing a window*, it is important to note that the latter involves interaction with an object (such a sponge) while the first does not.

Definition 9 (Motion concept) *A* motion concept *of an action a is a tuple $\psi_a = (\mathbf{P}, \eta, k_\rho, k_\lambda)$ where,*

- $\boldsymbol{P} = \{P_1, ..., P_\ell\}$*, where each P_i is a motion prototype describing one possible way by which the action a can be performed;*

- *η is a designation (a name) provided by the user to refer to the action a —*

for example, it may consist of a label for a *or an utterance that corresponds to the spoken designation of* a*;*

- k_p *and* k_λ *are two constants used to weight the importance of object information and location information for the recognition of action* a*.*

In particular, the object and location weights k_p and k_λ play a fundamental role for the description of human actions. While some actions might be performed in a wide range of locations yet have specific object requirements, other household actions can be performed with a wide range of objects, yet in a specific location. The object and location weights allow the quantification of such variability in the context of an action.

4.2 Learning Motion Concepts from Demonstration

We now introduce the *Online Motion Concept Learning* (OMCL) algorithm, designed to construct motion concepts from demonstration data provided by a human user. The OMCL algorithm can roughly be understood as working on two different abstraction levels. At a lower level, OMCL takes the data from a single demonstration and constructs a motion prototype from such data. At a higher level, OMCL combines information from multiple demonstrations to build a motion concept that potentially contains multiple motion prototypes. We now detail OMCL at each of these two abstraction levels.

4.2.1 Creating Motion Prototypes from Data

Given a demonstration of action a, OMCL initially builds the corresponding motion prototype $P_a = (\boldsymbol{\tau}_a, \boldsymbol{\rho}_a, \lambda_a)$, considering the motion ($\boldsymbol{\tau}_a$), object ($\boldsymbol{\rho}_a$) and location (λ_a) information provided by the human.

Regarding the motion information, OMCL starts by segmenting the single trajectory of motion channel k $\left(\boldsymbol{x}^k_{0:T}, \boldsymbol{R}^k_{0:T}\right)$ into multiple sub-trajectories,

$$\{(\boldsymbol{x}^k_{0:t_1}, \boldsymbol{R}^k_{0:t_1}), \ldots, (\boldsymbol{x}^k_{t_{N-1}:t_N}, \boldsymbol{R}^k_{t_{N-1}:t_N})\} \tag{4.2}$$

which can be achieved using any segmentation method from the literature — OMCL is agnostic to the particular segmentation method used. OMCL uses online kernel density estimation[1] to build new motion primitives from sub-trajectory data. The list of segmented sub-trajectories (4.2) is used to update the previous library Φ of available motion primitives. Subsequently, each sub-trajectory $\left(\boldsymbol{x}^k_{t_{n-1}:t_n}, \boldsymbol{R}^k_{t_{n-1}:t_n}, 0 \leq n \leq N\right)$ is evaluated against the updated Φ and a primitive ϕ_n is selected accordingly to Eq. 4.1. The resulting sequence of motion primitives, $\{\phi^k_1, \ldots, \phi^k_N\}$, corresponds to the trajectory representation τ^k, previously described. Such procedure is repeated for all motion channels to build $\boldsymbol{\tau}_a$.

Regarding the object and location information provided by the demonstration, we use standard maximum likelihood estimation to compute the parameters $\rho_a(o), o \in \mathcal{O}$ and $\lambda_a(l), l \in \mathcal{L}$, from the data $\boldsymbol{o}_{0:T}$ and $l_{0:T}$, respectively.

4.2.2 Creating and Updating Motion Concepts

From a provided demonstration of action a, the OMCL algorithm builds a motion prototype $P_a = (\boldsymbol{\tau}_a, \boldsymbol{\rho}_a, \lambda_a)$. The motion prototype is then employed to create, or update, the motion

[1] In our implementation, we use the XOKDE++ algorithm from [44].

concept $\psi_a = (\mathbf{P}_a, \eta_a, k_{\rho,a}, k_{\lambda,a})$ associated with action a, depending if the demonstration is the first provided of that action class, or not.

If P_a is the first motion prototype of action a provided, we build a novel motion concept ψ_a by the following procedure:

- The motion prototype is added to the empty list of prototypes $\boldsymbol{P}_a$,

$$\boldsymbol{P}_a = \{P_a\}$$

- The importance weights $k_{\rho,a}, k_{\lambda,a}$ are initialized to predetermined values,

$$k_{\rho,a} = k_{\rho,0}, \quad k_{\lambda,a} = k_{\lambda,0}$$

- The user is queried for the designation η_a of the action.

If the motion prototype P_a concerns an action class a previously demonstrated, we update the respective motion concept ψ_a, as follows:

- The motion prototype is added to the list of prototypes $\mathbf{P}_a$,

$$\mathbf{P}_a = \{P_1, ..., P_a\}$$

- We update the values of the importance weights $k_{\rho,a}, k_{\lambda,a}$. If for the majority of $P_i = (\boldsymbol{\tau}_i, \boldsymbol{\rho}_i, \lambda_i) \in \boldsymbol{P}_a \setminus P_a$ previously contained in $\boldsymbol{P}_a$:

$$\arg\max_{o \in \mathcal{O}} \rho_{i,m}(o) = \arg\max_{o \in \mathcal{O}} \rho_{a,m}(o), \forall m \in M,$$

  we increase the value of $k_{\rho,a}$ by a percentage α_k of its current value. Otherwise, we decrease $k_{\rho,a}$ by a percentage α_k of its current value. We apply the same update rule for k_λ, now considering the location data λ_a available in the provided motion prototype.

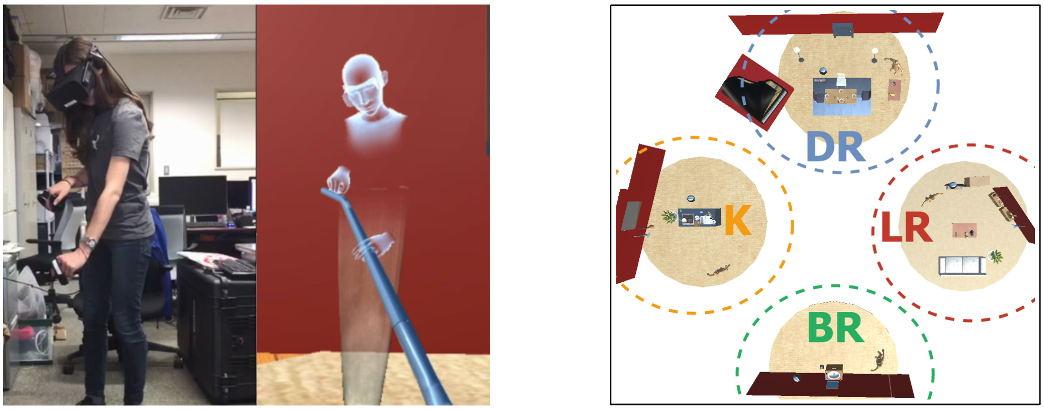
4.2.3 Recognizing Motion Concepts

The OMCL algorithm is also able to recognize previously observed actions and assess the novelty of previously unobserved actions. Given a motion prototype from an unknown action $P_* = (\boldsymbol{\tau}_*, \ \boldsymbol{\rho}_*, \ \lambda_*)$, OMCL compares P_* with the prototypes contained in every motion concept in the current library of motion concepts Ψ. The cost of assigning P_* to the motion concept $\psi_i \in \Psi$ is given by:

$$C(\psi_i, P_*) = \frac{1}{N_{\boldsymbol{P}_i}} \sum_{k=1}^{N_{\boldsymbol{P}_i}} C_M(\boldsymbol{\tau_k}, \boldsymbol{\tau_*}) + k_{\rho,i}\, C_O(\boldsymbol{\rho}_k, \boldsymbol{\rho}_*) + k_{\lambda,i}\, C_L(\lambda_k, \lambda_*) \tag{4.3}$$

where $N_{\boldsymbol{P}_i}$ refers to the number of motion prototypes currently associated with the motion concept ψ_i and,

- $C_M(\boldsymbol{\tau_i}, \boldsymbol{\tau}_*)$ is the *distance* between the sequence of motion primitives in $\boldsymbol{\tau}_*$ and the sequence of motion primitives $\boldsymbol{\tau_k}$ of motion prototype $P_k \in \mathbf{P}_i$ of ψ_i. To compute this distance, we use dynamic time warping [111] between the sequences of each motion channel, with 0-1 loss;



(a)

(b)

Figure 4.3: Evaluation setup employed to learn multimodal action representations: (a) the user interacts with the VR environment employing an Oculus Rift headset and hand motion controllers; (b) the household environment, composed of dining-Room (DR), kitchen (K), living-room (LR) and bathroom (BR) areas, each populated with location-specific and common objects.

- $C_O(\boldsymbol{\rho}_k, \boldsymbol{\rho}_*)$ is the *distance* between $\boldsymbol{\rho}_*$ and the collection of object probability distributions $\boldsymbol{\rho}_k$ of the motion prototype $P_k \in \boldsymbol{P}_i$ of ψ_i. The algorithm is agnostic to the type of metric used to compute the distance between distributions;
- $C_L(\lambda_k, \lambda_*)$ is the *distance* between λ_* and the location probability distribution λ_k of the motion prototype $P_k \in \boldsymbol{P}_i$ of ψ_i;
- $k_{\rho,i}, k_{\lambda,i}$ are the object and location information weights of ψ_i.

We compute the assignment cost $C(\psi_i, P_*)$ for all motion concepts $\psi_i \in \Psi$ and P_* is assigned to the motion concept with the lowest cost.

After the recognition procedure, we address the possible novelty of the action demonstration by considering the value of that assignment cost. As such, OMCL determines if P_* belongs to the assigned motion concept ψ_R or if it belongs to a novel action class. The decision takes into account the average cost C_W of assigning P_* to the motion concepts in $\Psi \setminus \psi_R$. If,

$$|C(\psi_R, P_*) - C_W| \geq \delta_C \, C(\psi_R, P_*), \tag{4.4}$$

the provisional assignment of P_* to ψ_R is confirmed, where δ_C is a user-defined threshold constant. Otherwise, if the assignment cost to the motion concept ψ_R is not significantly different from the average assignment cost to the other motion concepts, OMCL employs P_* to build a novel motion concept.

4.3 Evaluation

We evaluate the potential of using motion concepts to efficiently learn to recognize human actions from demonstration data, addressing the following two questions:

(i) Do motion concepts allow the recognition of human actions from a single demonstration?

(ii) Can we learn motion concepts online through interaction with a human user and distinguish between novel and previously observed action classes?

To address (i), in Section 4.3.2 we initially evaluate OMCL in an *offline* one-shot recognition task, showing the importance of considering the multimodal nature of the demonstrations, i.e. the contextual object and location information, in order to efficiently recognize action classes from a single demonstration. To address (ii), in Section 4.3.3, we evaluate OMCL in an *online tabula rasa* learning task, in which human users demonstrate novel and previously learnt action classes and the algorithm must distinguish the novelty of the demonstrations, starting from an empty database of motion concepts. We show that OMCL is able to successfully learn online novel action classes, leveraging the multimodal information in the demonstrations to access the novelty of the demonstrations provided.

4.3.1 Experimental Setup

The evaluation of the OMCL algorithm was performed in a virtual-reality (VR) environment. The user interacts with the VR environment using a Oculus Rift headset and hand motion controllers, as shown in Fig. 4.3a. Thus, in this setup, the number of motion channels K is equal to the number of object channels M, with $K = M = 3$.

We designed a virtual household environment composed of 4 different areas: *Kitchen* (K), *Living-Room* (LR), *Dining-Room* (DR) and *Bathroom* (BR), i.e., $\mathcal{L} = \{\mathrm{K}, \mathrm{LR}, \mathrm{DR}, \mathrm{BR}\}$, as shown in Fig. 4.3b. In addition, we populate the environment with objects of 23 classes, as shown in Table 4.1. Each area contains objects specific of that area (e.g. *Tooth-brush* is contained in the *Bathroom* area) as well as a number of common objects that can be found in multiple areas of the environment (e.g. *Cup* can be found in *Kitchen*, *Living-Room*, *Dining-Room*).

4.3.2 Offline One-Shot Recognition (OSR) Task

In the one-shot recognition (OSR) task we evaluate the performance of OMCL in recognizing motion concepts created from a single training demonstration. We asked 10 participants to perform, on the virtual household environment, two demonstrations of (randomly-ordered) 22 action classes, after a tutorial period of adaptation to the VR setting. Each action was recorded for 6 seconds, storing the motion data, from the VR headset and motion controllers, and the contextual data (object and location information) of the demonstration. We provided to the participants no information regarding which objects to use or where to perform the action. The complete list of action classes is presented in Table 4.1.

The action classes were chosen due to their simplicity, as complex manipulation of objects in a virtual environment is difficult, and the fact that participants could perform them stationary, minimizing the discomfort of locomotion in virtual space. Moreover, we selected actions with very similar motion patterns but distinct object and location contexts (e.g. *Wash Hands*/*Wash Plates* and *Wave*/*Wash Window*) and actions with highly variant motion patterns, object and location contexts (e.g. *Throw* action).

We optimize the values of the $(k_{\lambda,0}, k_{\rho,0})$ parameters of OMCL by parameter sweep, training with one random sample of each action class and evaluating the remaining samples in the training partition of the dataset. The training procedure is repeated 10 times per tuple of parameter values and the optimized values are selected based on the total accuracy of the model. The δ_C parameter is optimized following the same grid-search procedure: fixing the values of $(k_{\lambda,0}, k_{\rho,0})$ obtained previously, we build a motion concept from a single

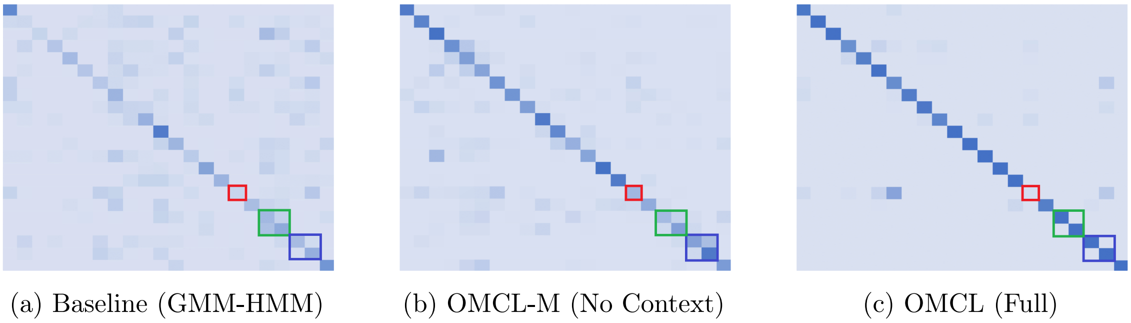
Table 4.1: Action classes performed in the Virtual-Reality environment, along with the most common objects used in the demonstrations and their most common locations in the household environment.

Action	Location	Objects
Bow	All	None
Comb hair	BR	Hairbrush
Cut	K	Knife, Apple, Banana, Pear
Drink	All	Mug, Glass, Bottle
Eat at Table	DR	Knife, Fork, Chopsticks
Fry	K	Frying Pan
High-Five	All	Hand
Hug	All	Body
Knock on door	LR	Door
Pet	DR, LR	Cat, Dog
Play Guitar	LR	Guitar
Play Piano	DR	Piano
Shake Hands	All	Hand
Stir Pot	K	Spoon, Pot
Sweep	K, LR	Broom
Throw	All	All
Vacuum clean	K, LR	Vacuum-cleaner
Wash Hands	BR	Soap
Wash Plates	K	Sponge, Dish
Wash Window	K	Sponge
Wave	All	None
Wring Sponge	K	Sponge

randomly-selected training sample of each action class. Subsequently, we evaluate the number of times OMCL assesses the test samples (provided without explicit class labels) as examples of the correct, corresponding, motion concept. The final parameter values are $k_{\lambda,0} = 0.005$, $k_{\rho,0} = 0.05$ and $\delta_C = 0.9$.

The performance of the OMCL algorithm is evaluated against a Gaussian Mixture-Hidden Markov Model (GMM-HMM) baseline, optimized through the same training procedure, yet resorting only to the motion data of the recorded actions. Using the total accuracy of the model as the selection criteria, the optimized number of hidden states in the model is $h_{\text{HMM}} = 16$ and the optimized number of components in the GMMs is $k_{\text{GMM}} = 3$. To evaluate the role of contextual information for action recognition, we include in the evaluation procedure a modified OMCL model (OMCL-M), in which we neglect the contribution of the contextual features (object and location information) to the recognition cost (4.3). In other words, the motion concepts in OMCL-M are built only considering motion data. In the OSR task, the recognition rates of the baseline algorithm, OMCL-M and OMCL algorithms in the test partition of the dataset are presented in Table 4.2. Moreover, their confusion matrices are presented in Fig. 4.4.

In the OSR task, the OMCL-M algorithm significantly outperforms the baseline algorithm, with an accuracy of 68.8 ± 19.7% against 37.6 ± 21.2%. This result validates the methodology of solving the recognition problem not through the direct comparison of



(a) Baseline (GMM-HMM) (b) OMCL-M (No Context) (c) OMCL (Full)

Figure 4.4: Confusion Matrices in the one-shot recognition task for the baseline and OMCL algorithms. We highlight the accuracy on the *Throw* action class (red), the accuracy on the *Wash Hands* and *Wash Dishes* classes (green) and the accuracy on the *Wash Window* and *Wave* classes (blue).

Table 4.2: Accuracy in the one-shot recognition task for the baseline, OMCL-M and OMCL algorithms. Higher is better.

Baseline (%)	OMCL-M (%)	OMCL (%)
37.6 ± 21.2	68.8 ± 19.7	90.5 ± 20.8

low-level joint data, which is prone to noise, measurement errors and variability, but through the comparison of motion primitives, previously learnt from data. However, the regular OMCL algorithm significant out-performs both methods (with $90.5 \pm 20.8\%$ accuracy rate), able to leverage the contextual information provided by the demonstrations.

Regarding their confusion matrices, presented in Fig. 4.4, the OMCL-M model (Fig. 4.4b) presents a significantly more diagonal matrix compared to the matrix of the baseline algorithm (Fig. 4.4a). Yet, the recognition of actions with similar motion patterns is still difficult, as both algorithms are not able to successfully distinguish between the *Wash Hands*/*Wash Dishes* actions (marked in green in Fig. 4.4) as well as between the *Wave*/*Wash-Window* actions (in blue in Fig. 4.4). The OMCL algorithm (Fig. 4.4c) shows a significant improvement in the recognition of the actions classes, indicated by the near-diagonal confusion matrix. Moreover, OMCL is able to distinguish between actions with similar motion patterns (*Wash Window*/*Wash Dishes*, *Wave*/*Wash-Window*) by considering the contextual data of the action (object and location). However, OMCL is still unable to recognize the action *Throw* (marked in red in Fig. 4.4) due to the similarity of its motion pattern to the class *High-Five* and the variance of objects used and locations where it can be performed. Indeed, the consideration of contextual information in the recognition of the *Throw* action seems to worsen the accuracy performance of the algorithm in comparison with the solely-motion-based version of OMCL. Yet, for the remaining action classes, their contextual information seems to play a fundamental role in the improvement of the recognition performance of OMCL.

4.3.3 Online Tabula-Rasa Learning Task

We evaluate the performance of OMCL on two different aspects: the recognition of previously observed action classes, whose motion concept was already created, and the identification of novel, previously unobserved, actions classes. The evaluation is performed on a *tabula*

(a) Total Interactions

Successful (%)	Unsuccessful (%)
80.4	19.6

(b) Novel Interactions

Successful (%)	Unsuccessful (%)
85.2	14.8

(c) Recognition Interactions

Successful (%)	Partial (%)	Unsuccessful (%)
71.7	18.3	10.0

Figure 4.5: Evaluation results of the online *tabula-rasa* learning task, considering a total of 168 interactions, corresponding to 14 actions demonstrations of 12 participants: (a) total accuracy of the interactions; (b) accuracy of 108 novel interactions, where the agent must identify new, previously unseen, action classes; c) accuracy of 60 recognition interactions, where the agent must identify the demonstrations as examples of previously observed action classes. For the latter subset of interactions we distinguish between successful, partial (where the agent initially recognizes the class of the demonstration yet subsequently evaluates it as a novel class) and unsuccessful interactions.

rasa scenario, i.e., a scenario in which the system, prior to the evaluation, is not trained on any data. We asked 12 participants to perform a sequence of demonstrations of 10 action classes. Each participant performs one demonstration of 5 action classes and two demonstration of the remaining 5 action classes. For each demonstration, OMCL processes the data streams and assesses its nature: a novel action class, previously unobserved, or a previously observed action class along with its denomination. The participant subsequently evaluates the assessment of the system and, accordingly to the participant's response, the system creates, or updates, the corresponding motion concept. The selection of the 10 classes and of the subset of classes with two demonstrations, along with the order of the actions to perform, are randomly selected from the classes presented in Table 4.1.

We evaluated 168 interactions corresponding to 14 action demonstrations of 12 participants, discarding the initial interaction of every participant which the system always recognizes as a novel motion concept. We define a *successful* interaction when the participant evaluates the system's assessment of the demonstration as correct, and *unsuccessful* otherwise. The overall results for the *Tabula Rasa* evaluation are presented in Fig. 4.5.

The results show that a majority of the interactions (80.4%) are successful. We can decompose the total 168 interactions into two different categories: 108 interactions due to the demonstration of a novel action class (*novel* interactions) and 60 interactions due to the performance of a previously observed action class (*recognition* interactions). The system is able to correctly evaluate the novelty of a action class previously not demonstrated in 85% of novel interactions. In 72% of recognition interactions, the system is also able to recognize previously observed action classes. In 18% of these interactions the system also recognizes the correct action class yet, due to significant differences in the motion pattern, location or objects interacted during the performance, it classifies the demonstration as an example of a novel action class. Nonetheless, the results highlight the potential of motion concepts for sample-efficient action recognition in the presence of a human demonstrator.

4.4 Concluding Remarks

In this chapter we presented motion concepts, a novel multimodal representation for human actions based on the kinematics of the action, the objects used during the action and the location where it was performed. Moreover, we presented a new algorithm to learn motion concepts from demonstration data provided by human users, and recognize previously created motion concepts. We evaluated the learning of motion concepts in an offline one-shot recognition task, which showed that our method allows action recognition from a single demonstration. Moreover we attested to the importance of leveraging contextual information to recognize actions with similar motion patterns. Finally, we evaluated motion concepts on an online tabula-rasa task, attesting the ability to learn and recognize novel action classes continuously, while interacting with a human user.

We have highlighted the potential of multimodal representations learning in terms of sample-efficiency for downstream tasks (such as classification) and the distinction of similar phenomena (in this case, action demonstrations) due to the complementary information provided by multiple sources of data. However, the richness of the multimodal motion concept representation is also a result of the careful human-design of its features, appropriate to the task at hand. Furthermore, the learning procedure requires explicit supervision from the human user, which might not be possible (or sensible) in some scenarios. Mimicking the human representation learning, we wish to learn multimodal representations from multiple sources of data without explicit supervision. In the next chapter, we discuss a novel multimodal generative model able to learn such representations from an arbitrary number of modalities, without supervision.

Chapter 5

Learning Multimodal Representations for Effective Cross-Modality Inference

"To read a poem consists of hearing it with our eyes"
Octavio Paz

In this chapter we address the problem of learning representations of high-dimensional multimodal data without supervision. In particular, we focus on the problem of *Cross-Modality Inference* (CMI), i.e., the generation of missing modality information from the available ones. For this goal, the representation should be *scalable*, able to account for an arbitrary number of modalities, and, to some extent, *agnostic* to the nature of these modalities. Complementary and redundant information provided by multiple modalities should be processed in a *compositional* manner to encode a common representation suitable for *cross-modality* generation, providing robustness under partial perceptual availability. Finally, the learning process should be *efficient*, without incurring in a significant computational cost.

To address the limitations brought upon by current approaches for cross-modality generation, we argue for considering *hierarchy* in the design of multimodal generative models. Inspired by human perception, we propose the *Multimodal Unsupervised Sensing* (MUSE) model, a novel generative model that considers a hierarchical relation between two sets of generative factors: low-level modality-specific factors, that encode information unique to each modality, and high-level multi-modal factors, that encode joint-modality information. Furthermore, we propose three different mechanisms to merge multimodal information and address the requirements of computational cross-modality inference. Our solution enables (i) the encoding of a single representation regardless of the number or nature of the available input modalities, thus being *scalable*, *agnostic* and *compositional*; (ii) a high-quality *cross-modality* generative process, regardless of the nature of the target modality, *efficiently* trained.

We evaluate MUSE against several distinct scenarios of increasing complexity, regarding the number and nature of the modalities involved. We evaluate MUSE in literature-standard multimodal datasets, showing that our model outperforms the baseline models in terms of generative performance, allowing for effective cross-modality inference regardless of the complexity of the target modality. Moreover, we argue that likelihood-based metrics commonly used are not completely adequate to evaluate the performance of the cross-modality generative process, advancing alternative metrics that may better capture such performance in class-based datasets. In addition, we introduce the Multimodal Handwritten Digit (MHD) dataset, a novel multimodal benchmark dataset composed of the image, sound, motion trajectory and label associated with handwritten digits. The results show that MUSE is the only model that is able to address simultaneously all requirements of computational cross-modality inference, highlighting the importance of considering hierarchy in the design of multimodal generative models.

The main contributions of this chapter are four-fold:

- In Section 5.1, we introduce the problem of computational *cross-modality inference* and discuss how current multimodal generative models fall short of addressing simultaneously all its requirements. In Section 5.2, we discuss the case of human representation learning and we argue for considering hierarchy in multimodal generative models;

- In Section 5.3, we propose the *Multimodal Unsupervised Sensing* (MUSE) model, a novel architecture that considers both low-level modality-specific representations, and high-level multimodal representations;

- In Section 5.5, we introduce the Multimodal Handwritten Digits (MHD) dataset, which includes images, trajectories, sounds and labels associated with handwritten digits;

- In Section 5.6, we evaluate MUSE against other state-of-the-art models across different literature standard scenarios. We advance potential complementary metrics to attest their performance in Section 5.7. The results show that MUSE outperforms all other baselines, being the only model that is able to address simultaneously all requirements of computational cross-modality inference.

The work described in this chapter has been published in:

- **Miguel Vasco**, Hang Yin, Francisco S. Melo, and Ana Paiva. *Leveraging hierarchy in multimodal generative models for effective cross-modality inference.* Neural Networks (2021 Special Issue on AI and Brain Science: Brain-inspired AI) 146, 2022, pp. 238–255 [168];

- **Miguel Vasco**, Hang Yin, Francisco S. Melo, and Ana Paiva. *How to Sense the World: Leveraging Hierarchy in Multimodal Perception for Robust Reinforcement Learning Agents.* Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). 2022, pp. 1301–1309 [169].

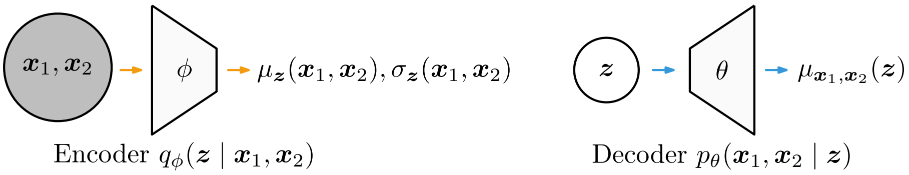

x_1, x_2
ϕ
$\mu_z(x_1, x_2), \sigma_z(x_1, x_2)$
Encoder $q_\phi(z \mid x_1, x_2)$
z
θ
$\mu_{x_1,x_2}(z)$
Decoder $p_\theta(x_1, x_2 \mid z)$

Figure 5.1: Architecture of the naive multimodal extension of the VAE model.

5.1 The Problem of Cross-Modality Inference (CMI)

The variational autoencoder (VAE) model, introduced in Section 2.1.1, is widely employed to learn low-dimensional representations of high-dimensional, single-modality data. However, the VAE framework can also be easily modified to process multimodal data. Consider a scenario where an agent has access to M different sensor channels and the environment provides multimodal information $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \in \mathcal{X}$, where $\mathbf{x}_m \in \mathcal{X}_m$ corresponds to the information provided by an input "modality" x_m and the multimodal input space $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_M$ is computed through the Cartesian product of the individual modality input spaces $\mathcal{X}_m$. Each modality may correspond to a different type of information (e.g., image, sound, force feedback), with a unique level of complexity and dimensionality.

Definition 10 (Multimodal generative model) *A multimodal generative model consists of a (encoder) representation map $r = q_\phi : \mathcal{X} \to \mathcal{Z}$, parameterized by ϕ, that encodes a compact representation $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^N$ of all modalities, and a (decoder) downstream map $g = p_\theta : \mathcal{Z} \to \mathcal{X}$, parameterized by θ, that allows the generation of high-dimensional data $\boldsymbol{x} \in \mathcal{X}$.*

The naive extension of the VAE model to multimodal input data mostly ignores the individual modalities $\boldsymbol{x}_m$, $m = \{1, \ldots, M\}$, and treats $\boldsymbol{x}$ in an aggregated manner, as a single input. Figure 5.1 depicts the network architecture of such approach, where,

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \prod_{m=1}^{M} p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z}), \tag{5.1}$$

and trivially leads to the loss,

$$\ell_{\mathrm{VAE}^\star}(\mathbf{x}) = D_{\mathrm{KL}}(q_\phi \parallel p) - \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\left[\sum_{m=1}^{M} \log p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z})\right], \tag{5.2}$$

where both the joint-modality encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the modality-specific decoders $p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z})$ are instantiated as neural networks. However, this simple solution ignores the possible decomposition of the input into distinct modalities and is unable to reconstruct $\boldsymbol{x}$ from partial inputs, i.e., to perform cross-modality inference.

One possible approach to enable the model to perform cross-modality inference — pioneered in the work of Yin et al. [184] — is to consider an architecture akin to that depicted in Fig. 5.2, denoted by Associative VAE (AVAE). The AVAE model trains a modality-specific encoder-decoder pair to learn, respectively, the distributions $p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z})$ and $q_{\phi m}(\boldsymbol{z}|\boldsymbol{x}_m)$, for $m = 1, \ldots, M$. These modality-specific models are combined by forcing the distributions over the latent space to *agree* for the same input. For two modalities, i.e., an input $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, the authors introduce an *association loss* term of the form,

$$\ell_{\mathrm{a}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = d\left(q_{\phi_1}(\cdot|\boldsymbol{x}_1), q_{\phi_2}(\cdot|\boldsymbol{x}_2)\right), \tag{5.3}$$

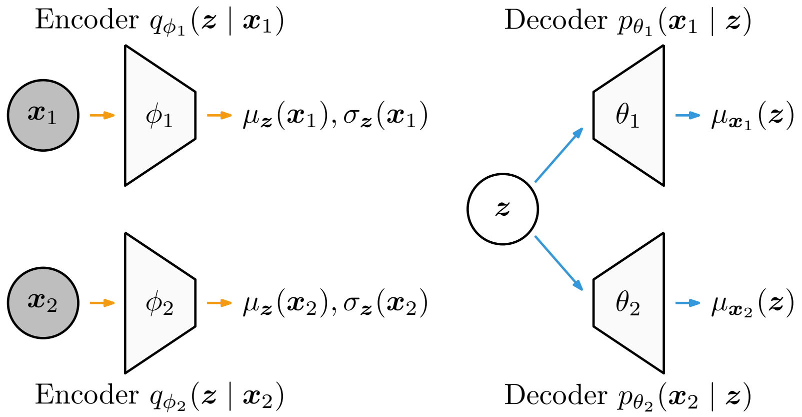

Encoder $q_{\phi_1}(\boldsymbol{z} \mid \boldsymbol{x}_1)$
$\boldsymbol{x}_1$
ϕ_1
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_1), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_1)$
$\boldsymbol{x}_2$
ϕ_2
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_2), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_2)$
Encoder $q_{\phi_2}(\boldsymbol{z} \mid \boldsymbol{x}_2)$
Decoder $p_{\theta_1}(\boldsymbol{x}_1 \mid \boldsymbol{z})$
θ_1
$\mu_{\boldsymbol{x}_1}(\boldsymbol{z})$
$\boldsymbol{z}$
θ_2
$\mu_{\boldsymbol{x}_2}(\boldsymbol{z})$
Decoder $p_{\theta_2}(\boldsymbol{x}_2 \mid \boldsymbol{z})$

Figure 5.2: Architecture of the AVAE model, supporting cross-modality inference, from Yin et al. [184]. (see main text for details).

where d is a distance metric between probability distributions[1]. Such enforcement requires that the latent distributions of each modality have the same dimensionality, regardless of the intrinsic complexity of the modality itself. The model in Fig. 5.2 is then trained as a single model, with a loss function,

$$\ell_{\text{AVAE}}(\boldsymbol{x}) = \ell_1(\boldsymbol{x}_1) + \ell_2(\boldsymbol{x}_2) + \lambda \ell_{\text{a}}(\boldsymbol{x}_1, \boldsymbol{x}_2), \tag{5.4}$$

where,

$$\ell_m(\boldsymbol{x}_m) = D_{\text{KL}}(q_{\phi_m} \parallel p) - \mathbb{E}_{\boldsymbol{z} \sim q_{\phi_m}(\cdot|\boldsymbol{x}_m)} \left[\log p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z})\right], \tag{5.5}$$

for $m = 1, 2$ and λ controls the relative importance of the association term. The trained model is now able, for example, to use the modality specific encoder q_{ϕ_1} to generate a latent vector $\boldsymbol{z}$ from the single modality input $\boldsymbol{x}_1$, and then use this latent vector $\boldsymbol{z}$ as an input for the decoder p_{θ_2} to generate the missing modality $\boldsymbol{x}_2$, successfully performing cross-modality inference.

The approach of Yin et al. [184] was limited to two modalities and unable to consider joint-modality information (having access to both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) for the generation process. Suzuki et al. [156] proposed an extension of the previous model to consider joint-modality information, the Joint-Modality VAE (JMVAE), depicted in Fig. 5.3. The architecture can be seen as a combination of multiple VAEs, one for every individual modality, and one "joint" encoder, $q_{\phi_J}(\boldsymbol{z}|\boldsymbol{x})$, for every possible combination of modalities. Similarly to the AVAE model, the JMVAE forces the modality-specific latent distributions and the joint-modality latent distribution to agree, following a joint-associative loss,

$$\ell_{\text{j-a}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = d\left(q_{\phi_1}(\cdot|\boldsymbol{x}_1), q_{\phi_J}(\cdot|\boldsymbol{x}_1, \boldsymbol{x}_2)\right) + d\left(q_{\phi_1}(\cdot|\boldsymbol{x}_1), q_{\phi_J}(\cdot|\boldsymbol{x}_1, \boldsymbol{x}_2)\right), \tag{5.6}$$

where d is once again a distance metric between probability distributions[2]. The training of the JMVAE model follows the loss function,

$$\ell_{\text{JMVAE}}(\boldsymbol{x}) = D_{\text{KL}}(q_{\phi_J} \parallel p) - \mathbb{E}_{\boldsymbol{z} \sim q_{\phi_J}(\cdot|\boldsymbol{x})} \left[\sum_{m=1}^{M} \log p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z})\right] + \alpha\, \ell_{\text{j-a}}(\boldsymbol{x}_1, \boldsymbol{x}_2). \tag{5.7}$$

[1] Yin et al. consider $d(p, q)$ to be the symmetric KL divergence between distributions p and q, defined as $d(p, q) = \text{KL}(p \parallel q) + \text{KL}(q \parallel p)$.

[2] Suzuki et al. consider $d(p, q)$ to be the KL divergence between p and q, defined as $d(p, q) = \text{KL}(p \parallel q)$.

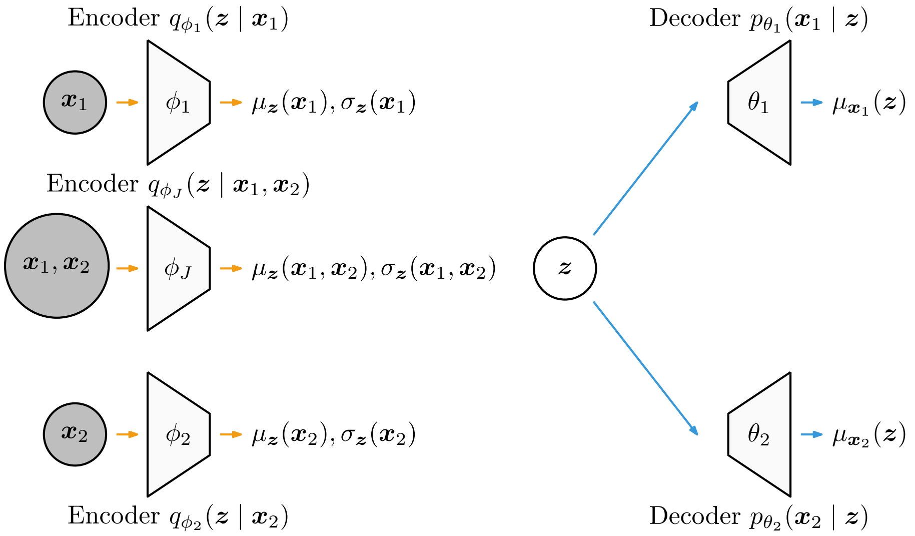

Encoder $q_{\phi_1}(\boldsymbol{z} \mid \boldsymbol{x}_1)$
$\boldsymbol{x}_1$
ϕ_1
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_1), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_1)$
Encoder $q_{\phi_J}(\boldsymbol{z} \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$
$\boldsymbol{x}_1, \boldsymbol{x}_2$
ϕ_J
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_1, \boldsymbol{x}_2), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$
$\boldsymbol{x}_2$
ϕ_2
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_2), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_2)$
Encoder $q_{\phi_2}(\boldsymbol{z} \mid \boldsymbol{x}_2)$
$\boldsymbol{z}$
Decoder $p_{\theta_1}(\boldsymbol{x}_1 \mid \boldsymbol{z})$
θ_1
$\mu_{\boldsymbol{x}_1}(\boldsymbol{z})$
θ_2
$\mu_{\boldsymbol{x}_2}(\boldsymbol{z})$
Decoder $p_{\theta_2}(\boldsymbol{x}_2 \mid \boldsymbol{z})$

Figure 5.3: Architecture of the JMVAE model, supporting cross-modality inference, from Suzuki et al. [156]. The model comprises a VAE for each of the two input modalities and a "joint" encoder, $q_{\phi_J}(\boldsymbol{z}|\boldsymbol{x}_1, \boldsymbol{x}_2)$, for the concatenation of the two.

This model, while not limited to two modalities, presents an obvious disadvantage: its dimension grows rapidly with the number of input modalities. To be more precise, as more modalities are made available to the model, JMVAE requires the instantiation of individual "sub-models" that account for every possible combination of modalities in order to associate their latent distribution to the joint-modality latent distribution. For example, in the case of three-modalities, an extension of JMVAE would require seven latent distributions, as shown in [85]: three modality-specific ones, three for the pairwise combinations, and a single one to account for all modalities

The models of Yin et al. [184] and Suzuki et al. [156] highlight one key difficulty in designing multimodal generative models:

(i) **Scalability**: The model must "merge" the information from the different input modalities, in order to perform cross-modality inference. As more modalities are considered in the model, such merging should be efficient, scaling gracefully with the number (and dimensionality) of the input modalities. The use of multiple "sub-models", as in the work of Suzuki et al. [156] does not scale well with the number of modalities and is, therefore, unsuited to deal with situations with a large number of modalities.

The two difficulties identified above are common to other approaches that consider multiple modalities [85, 101, 159, 162, 170]. To address the **scalability** design issue (ii), Wu and Goodman [179] proposed to employ an *implicit* joint-modality encoder, composed as some function f of single-modality distributions. In this work, the authors proposed a joint-modality encoder defined as the product-of-experts (POE) factorization of single-modality encoders with a prior-expert $q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \propto p(\boldsymbol{z}) \prod_{m=1}^{M} q_{\phi_m}(\boldsymbol{z}|\boldsymbol{x}_m)$, as shown in Fig. 5.4. Denoted by Multimodal VAE (MVAE), this approach is able to scale to arbitrarily large number of modalities without requiring the creation of specific "sub-models" to account for combinations of modalities. To perform cross-modality inference, the MVAE model requires a *sub-sampling* training scheme that considers ELBO terms for complete (joint)

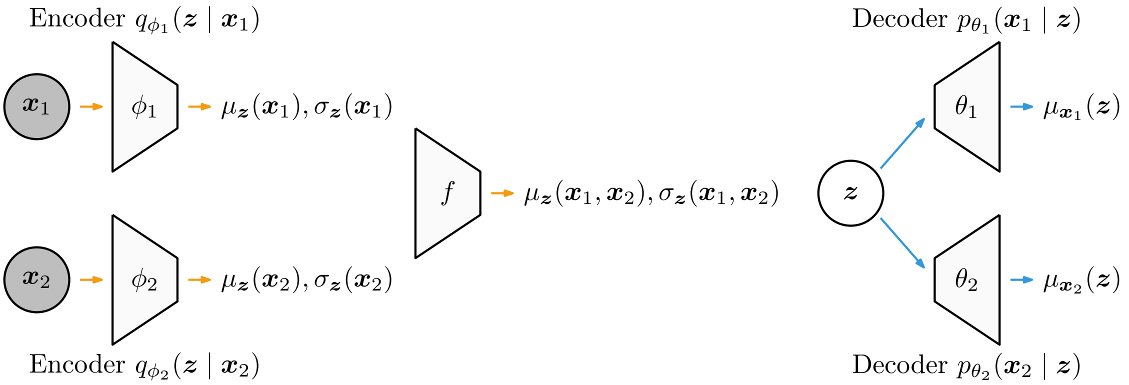

Encoder $q_{\phi_1}(\boldsymbol{z} \mid \boldsymbol{x}_1)$
$\boldsymbol{x}_1$
ϕ_1
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_1), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_1)$
f
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_1, \boldsymbol{x}_2), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$
$\boldsymbol{x}_2$
ϕ_2
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_2), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_2)$
Encoder $q_{\phi_2}(\boldsymbol{z} \mid \boldsymbol{x}_2)$
Decoder $p_{\theta_1}(\boldsymbol{x}_1 \mid \boldsymbol{z})$
$\boldsymbol{z}$
θ_1
$\mu_{\boldsymbol{x}_1}(\boldsymbol{z})$
θ_2
$\mu_{\boldsymbol{x}_2}(\boldsymbol{z})$
Decoder $p_{\theta_2}(\boldsymbol{x}_2 \mid \boldsymbol{z})$

Figure 5.4: Architecture for VAE supporting cross-modality inference with an *implicit* joint-modality encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x}_1, \boldsymbol{x}_2)$, composed of some function f of single-modality encoder distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x}_1)$, $q_\phi(\boldsymbol{z}|\boldsymbol{x}_2)$. The MVAE model proposed by Wu and Goodman [179] considers a product-of-experts (POE) factorization with a prior-expert $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z})q_\phi(\boldsymbol{z}|\boldsymbol{x}_1)q_\phi(\boldsymbol{z}|\boldsymbol{x}_2)$.

observations, for single-modality observations, and for partial observations (with randomly chosen subsets of modalities). For scenarios with two modalities, i.e., an input $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, this corresponds to a loss function,

$$\ell_{\text{MVAE}}(\boldsymbol{x}) = \ell_J(\boldsymbol{x}_1, \boldsymbol{x}_2) + \ell_1(\boldsymbol{x}_1) + \ell_2(\boldsymbol{x}_2), \tag{5.8}$$

where the joint-observation term is defined as,

$$\ell_J(\boldsymbol{x}) = D_{\text{KL}}(q_\phi \parallel p) - \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\cdot|\boldsymbol{x})} \left[\log p_{\theta_1}(\boldsymbol{x}_1|\boldsymbol{z}) + \log p_{\theta_2}(\boldsymbol{x}_2|\boldsymbol{z})\right], \tag{5.9}$$

and the partial ELBO terms are defined as,

$$\ell_m(\boldsymbol{x}) = D_{\text{KL}}(q_{\phi_m} \parallel p) - \mathbb{E}_{\boldsymbol{z} \sim q_{\phi(\cdot|\boldsymbol{x}_m)}} \left[\log p_{\theta_1}(\boldsymbol{x}_1|\boldsymbol{z}) + \log p_{\theta_2}(\boldsymbol{x}_2|\boldsymbol{z})\right]. \tag{5.10}$$

This training scheme presents an obvious disadvantage: the number of subsampling terms in the loss function grows rapidly with the number of input modalities, requiring multiple passages of the data through the model, and the associated gradient computations. In addition, the model presents another, less-obvious, disadvantage: developed for weakly-supervised learning scenarios, where joint-modality information may not be fully available during training, the POE solution is prone to overconfident expert prediction, often of the higher-dimensional modality (e.g. images) [142]. As such, the model is able to infer missing low-dimensional information (e.g. label) from high-dimensional modalities (e.g. image), yet struggle with the inverse inference process.

To address these issues, Shi et al. [142] proposed the Mixture-of-Experts MVAE (MMVAE), that employs an implicit mixture-of-experts (MOE) joint-modality encoder, $q_\Phi(\boldsymbol{z}|\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m \, q_{\phi_m}(\boldsymbol{z}|\boldsymbol{x}_m)$, where $\alpha_m = 1/M$, with the assumption that the available modalities are of similar complexity. For two modalities, i.e., $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, this results in a loss function,

$$\ell_{\text{MMVAE}}(\boldsymbol{x}) = \frac{1}{2} \sum_{m=1}^{2} \mathbb{E}_{z \sim q_\phi(\cdot|\boldsymbol{x}_m)} \log \left[\frac{p_{\theta_1}(\boldsymbol{x}_1|\boldsymbol{z}) p_{\theta_2}(\boldsymbol{x}_2|\boldsymbol{z}) p(\boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{x}_1) q_\Phi(\boldsymbol{z}|\boldsymbol{x}_2)}\right]. \tag{5.11}$$

As shown in Fig. 5.5, the MOE solution incurs on some computational overhead due to the necessity of computing M^2 passes over the single-modality decoders, as each modality

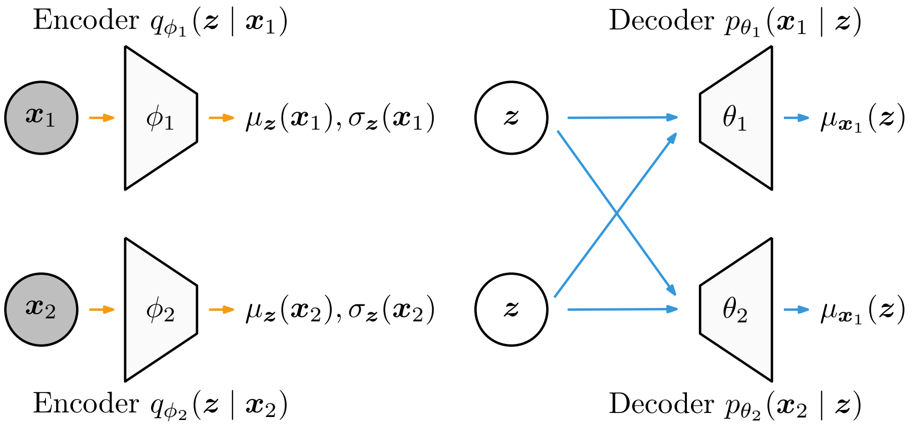

Encoder $q_{\phi_1}(\boldsymbol{z} \mid \boldsymbol{x}_1)$
$\boldsymbol{x}_1$
ϕ_1
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_1), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_1)$
$\boldsymbol{x}_2$
ϕ_2
$\mu_{\boldsymbol{z}}(\boldsymbol{x}_2), \sigma_{\boldsymbol{z}}(\boldsymbol{x}_2)$
Encoder $q_{\phi_2}(\boldsymbol{z} \mid \boldsymbol{x}_2)$
Decoder $p_{\theta_1}(\boldsymbol{x}_1 \mid \boldsymbol{z})$
$\boldsymbol{z}$
θ_1
$\mu_{\boldsymbol{x}_1}(\boldsymbol{z})$
$\boldsymbol{z}$
θ_2
$\mu_{\boldsymbol{x}_1}(\boldsymbol{z})$
Decoder $p_{\theta_2}(\boldsymbol{x}_2 \mid \boldsymbol{z})$

Figure 5.5: Architecture for VAE supporting cross-modality inference by Shi et al. [142]s. Employing an *implicit* joint-modality encoder, the MMVAE defines a mixture-of-experts (MOE) over the single-modality distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x}_1)$, $q_\phi(\boldsymbol{z}|\boldsymbol{x}_2)$.

provides samples from its own encoding distribution to be evaluated by all generative models. Thus, the training procedure does not consider joint-modality information to optimize ℓ_{MMVAE}, similarly to the case of Yin et al. [184]. The models of Wu and Goodman [179] and Shi et al. [142] highlight two more key difficulties in designing multimodal generative models with implicit joint-modality encoders:

(ii) **Effective Cross-Modality**: The model must be able to infer missing modality information from provided available information, regardless of the nature and complexity of both target and input modalities;

(iii) **Compositionality**: To perform cross-modality inference the model must be able to account for the information provided by *all* available modalities. As more modalities are made available, the model should be able to consider the redundant and complementary information they provide in order to encode a more adequate multimodal latent representation.

In this work, we address the issues brought up by computational approaches to the cross-modality inference process. Recently, Shi et al. [142] posited four criteria for the successful learning of a multimodal representation. The issues presented in this work can also be considered as desiderata for the learning process of multimodal generative models and, as such, we can naturally establish associations with those criteria: for example, issue (ii) shares the same concerns as the “Coherent Cross Generation” criteria. In this sense, the issues presented here can also be employed as evaluation criteria of the quality of multimodal generative models.

In the following sections, we contribute with a model that graciously scales with an arbitrary number of modalities, addressing the **scalability** (i) issue, and is able to consider all information provided to the model, regardless of the nature and complexity of the available modalities, addressing both the **effective cross-modality** (ii) and the **compositionality** (iii) issues. To do so, we leverage hierarchy in the design of multimodal generative models.

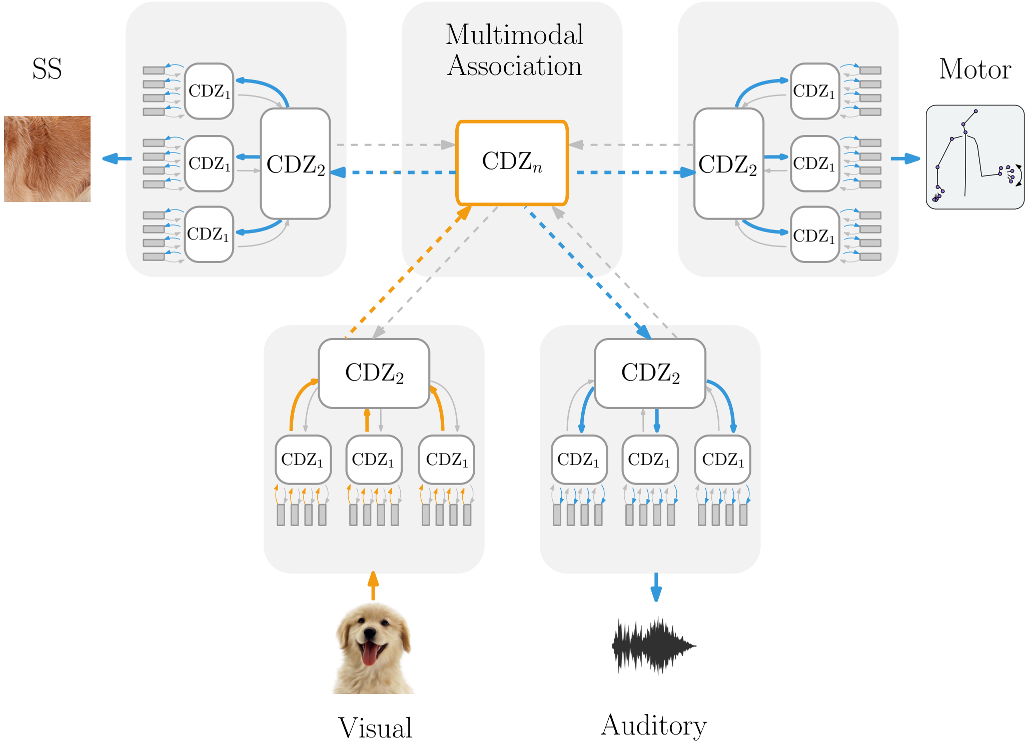
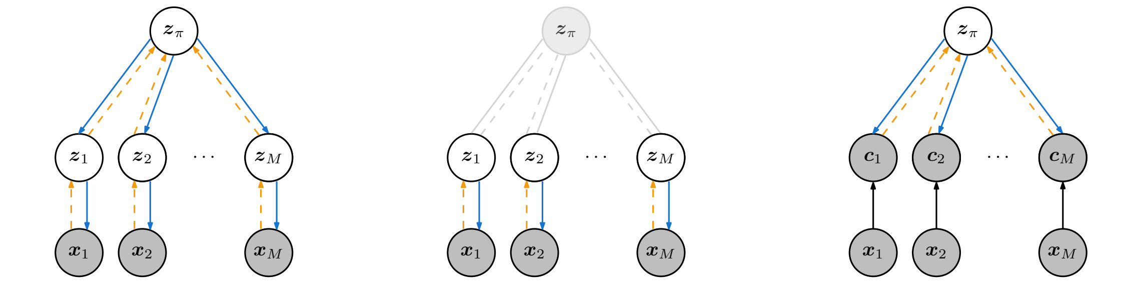


Figure 5.6: The cross-modality inference process in the CDZ framework, proposed by Damásio [30]: available information (e.g., image of a dog) is collected by the visual sensors of the human and forward processed in order to encode a multimodal representation, from which information is back-propagated to the remaining (absent) perceptual modalities.

5.2 The Role of Hierarchy in Perception

We start the argument for considering hierarchical generative models by highlighting the case of human representation learning. As introduced in Section 3.1, the *Convergence-Divergence Zone* (CDZ) framework is often employed to explain the biological mechanisms behind human perception and recognition [30, 108]. The CDZ model proposes an hierarchy of neuron ensembles, composed of multiple lower-level single-modality neurons and higher-level multimodal convergence zones. The CDZ model also provides a suitable explanation for the biological framework behind the human cross-modality inference process: as depicted in Fig. 5.6, the observed modality information (e.g. image) encodes a representation that is forward propagated until the multimodal convergence zones, from which it diverges into the modality-specific neural ensembles of absent modalities.

We now turn to the case of computational approaches to learn a multimodal representation, introduced in the previous section. The training of such models aims at learning a single compact latent representation $\boldsymbol{z}$ of all modalities. The difficulty in learning such representation cannot be overstated: the single representation $\boldsymbol{z}$, of limited capacity, must accommodate information responsible for reconstructing single modality data, regardless of their number or individual complexity, and for the interaction between the modalities, in order to allow for cross-modality inference. In stark contrast, the CDZ framework (Fig. 5.6) proposes to decouple the modality-specific generation and cross-modality inference problems: the lower-level representations specialize in encoding and generating modality-specific data



(a) Hierarchical architecture of MUSE.

(b) Modality-specific representations $\boldsymbol{z}_{1:M}$.

(c) Modality codes $\boldsymbol{c}_{1:M}$ and multimodal representation $\boldsymbol{z}_\pi$.

Figure 5.7: The MUSE model: (a) *encoder* (dashed, orange) and *decoder* (full, blue) networks of the model; (b) modality-specific representations $\boldsymbol{z}_{1:M}$, trained following the bottom-level loss $\ell_b(\boldsymbol{x}_{1:M})$ in Eq. 5.12; (c) multimodal representation $\boldsymbol{z}_\pi$ encoded from the modality codes $\boldsymbol{c}_{1:M}$, trained following the top-level loss $\ell_t(\boldsymbol{c}_{1:M})$ in Eq. 5.13.

while the top-level representation is responsible for propagating multimodal information to the lower-level neural ensembles.

5.3 Multimodal Sensing (MUSE) model

Inspired by human perception, we leverage hierarchy to learn a multimodal representation from information provided by the environment in an arbitrary number of different modalities, each of arbitrary nature and complexity. We propose the *Multimodal Sensing* (MUSE) model, described by the architecture in Fig. 5.7a. We design MUSE considering two sets of generative factors. On a bottom-level, we consider *modality-specific* factors, $\boldsymbol{z}_{1:M} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M\}$, unique to each modality and responsible for encoding and generating modality-specific data. At a top-level, we consider *multimodal* factors, $\boldsymbol{z}_\pi$, encoded from multimodal information, responsible for the generation of modality-specific latent samples.

The training of MUSE also takes a hierarchical approach: each set of generative factors is trained independently, yet simultaneously in the same data pass through the model. This allows the model to learn (a) modality-specific representations specialized in each modality, unconstrained by the multimodal reconstruction objective; and (b) a multimodal representation, specialized in the generation of coherent modality-specific latent samples.

Learning Modality-Specific Factors To train the *modality-specific* generative factors $\boldsymbol{z}_{1:M}$, we follow the standard VAE ELBO. We assume that each modality $\boldsymbol{x}_m$ is generated by a corresponding latent variable $\boldsymbol{z}_m$, of capacity/dimensionality appropriate to the complexity of the underlying modality. As seen in Fig. 5.7b, we learn a set of independent, single-modality, generative models $p(\boldsymbol{x}_{1:M}) = \prod_{m=1}^{M} p(\boldsymbol{x}_m)$, following a bottom-level loss,

$$\ell_b(\boldsymbol{x}_{1:M}) = \sum_{m=1}^{M} D_{\mathrm{KL}}(q_{\phi_m^b} \parallel p) - \mathbb{E}_{q_{\phi_m^b}(\boldsymbol{z}_m|\boldsymbol{x}_m)} \log p_{\theta_m^b}(\boldsymbol{x}_m \mid \boldsymbol{z}_m), \tag{5.12}$$

where the likelihoods $p_{\theta_m^b}(\boldsymbol{x}_m \mid \boldsymbol{z}_m)$ are parameterized by $\Theta^b = \{\theta_1^b, \ldots, \theta_M^b\}$ and the approximate single-modality posterior distributions $q_{\phi_m^b}(\boldsymbol{z}_m \mid \boldsymbol{x}_m)$ are parameterized by $\Phi^b = \{\phi_1^b, \ldots, \phi_M^b\}$, both instantiated as deep neural networks. In MUSE, and unlike its non-hierarchical counterparts, the capacity of the modality-specific latent spaces defined by the encoding $q_{\phi_m}(\boldsymbol{z}_m \mid \boldsymbol{x}_m)$ and prior distributions $p(\boldsymbol{z}_m)$, can be independently set for each modality, allowing for a more flexible design.

Learning Multimodal Factors At a higher-level, we merge the modality-specific information to learn *multimodal* factors, $\boldsymbol{z}_\pi$, that allow for joint and cross-modality generation. To propagate modality-specific information to the multimodal factors, we define modality *codes* $\boldsymbol{c}_{1:M} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}$, low-dimensional representations of the modality data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$. At training time, the modality codes are sampled from the modality-specific distributions, i.e., $\boldsymbol{c}_m \sim q_{\phi_m}(\boldsymbol{z}_m \mid \boldsymbol{x}_m)$. At test-time, the modality codes are computed as the expected value of the modality-specific distributions, i.e., $\boldsymbol{c}_m = \mathbb{E}\left[q_{\phi_m}(\boldsymbol{z}_m \mid \boldsymbol{x}_m)\right]$.

Conversely, in the generative process, $\boldsymbol{z}_\pi$ is responsible for generating the codes $\boldsymbol{c}_{1:M} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}$. As depicted in Fig. 5.7c, we learn a single multimodal decoding model $p(\boldsymbol{z}_\pi, \boldsymbol{c}_{1:M}) = p(\boldsymbol{z}_\pi) \prod_{m=1}^{M} p(\boldsymbol{c}_{1:M} \mid \boldsymbol{z}_\pi)$, following a top-level loss $\ell_t(\boldsymbol{c}_{1:M})$,

$$\ell_t(\boldsymbol{c}_{1:M}) = D_{\mathrm{KL}}(q_{\phi^t} \parallel p) - \sum_{m=1}^{M} \mathbb{E}_{q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})} \log p_{\theta^t_m}(\boldsymbol{c}_m \mid \boldsymbol{z}_\pi), \tag{5.13}$$

where the code likelihoods $p_{\theta^t_m}(\boldsymbol{c}_m \mid \boldsymbol{z}_\pi)$ are parameterized by $\Theta^t = \{\theta^t_1, \ldots, \theta^t_M\}$ and the approximate joint-modality posterior distributions $q_{\phi^t_m}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$ are parameterized by $\Phi^t = \{\phi^t_1, \ldots, \phi^t_M\}$, both instantiated as deep neural networks. The modality codes act as the link between the (bottom) modality-specific latent space and the (top) multimodal latent space, containing information suitable both to encode a compact multimodal representation but also to decode rich modality-specific information. Moreover, we stop the gradients of the top-level multimodal representation loss terms (Eq. 5.13) from propagating to the bottom-level computation graph by detaching the modality codes from the graph. We empirically found this solution to stabilize the training of the model.

5.4 Multimodal Encoder

We now turn to the question of how to define the multimodal joint proposal distribution $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$ in order to encode a representation able to tackle the remaining issues discussed in Section 5.1: how to learn such representation in a way that is (graciously) extendable to an arbitrary number of modalities (**scalability**) and that is able to consider the information provided by all available modalities (**compositionality**).

To do so, we present three different joint-modality encoding solutions, differing on the mechanism to merge modality-specific information and the corresponding training procedure: the *Naive* solution, concatenating the information from the modality-specific latent spaces; the *NEXUS* solution that merges modality-specific information through an aggregator function, requiring a dropout-like training procedure; the *ALMA* solution employs a Product-of-Expert (PoE) solution to encode the multimodal representation, requiring a approximation-based training scheme for the complete and partial multimodal distributions.

5.4.1 Naive Encoder

The *naive* joint-modality encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$ employs directly the information provided by the modality codes $\boldsymbol{c}_{1:M}$ in order to generate the multimodal representation, as shown in Fig. 5.8. We introduce a modality-data dropout masks $\boldsymbol{d}$, with dimensionality $|\boldsymbol{d}| = N$,

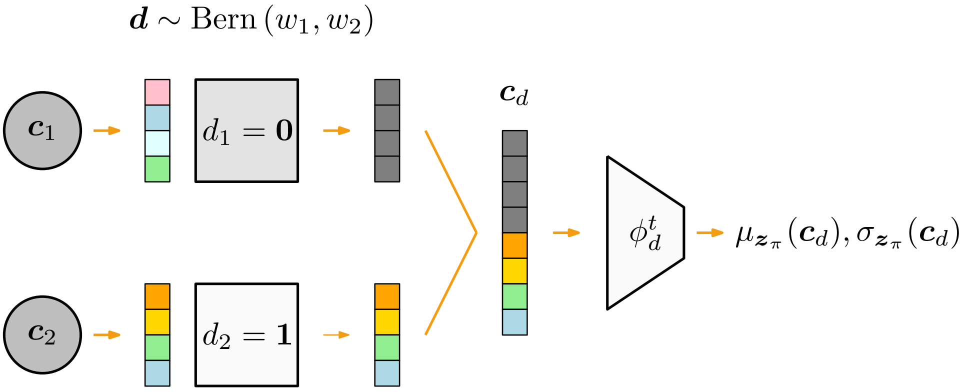

$\boldsymbol{d} \sim \text{Bern}\,(w_1, w_2)$
$\boldsymbol{c}_1$
$d_1 = \mathbf{0}$
$\boldsymbol{c}_2$
$d_2 = \mathbf{1}$
$\boldsymbol{c}_d$
ϕ_d^t
$\mu_{\boldsymbol{z}_\pi}(\boldsymbol{c}_d), \sigma_{\boldsymbol{z}_\pi}(\boldsymbol{c}_d)$

Figure 5.8: The *naive* joint-modality encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$, instantiated in a scenario with two modalities $\boldsymbol{x}_1, \boldsymbol{x}_2$: a concatenated version of modality-specific latent codes $\boldsymbol{c}_1, \boldsymbol{c}_2$ is employed to encode the multimodal representation $\boldsymbol{z}_\pi$. To provide robustness to missing modalities at execution time, we randomly drop modality codes accordingly to the dropout mask $\boldsymbol{d}$, zeroing out the selected components.

such that,

$$\boldsymbol{c}_d = \boldsymbol{d} \odot \boldsymbol{c}, \tag{5.14}$$

where $\boldsymbol{c} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$ corresponds to the list of modality codes. We effectively zero-out the selected components by considering that,

$$\boldsymbol{c}_i = \mathbf{0}, \text{ if } d_i = 1. \tag{5.15}$$

During training, for each datapoint, we sample $\mathbf{d}$ from a Bernoulli distribution,

$$\boldsymbol{d} \sim \text{Bern}\,(w_1, \ldots, w_N), \quad \text{with} \sum_{i=1}^{N} d_i \geq 1, \tag{5.16}$$

where the hyper-parameters $\boldsymbol{w}_{1:N}$ control the dropout probability of each modality representation. Moreover, we condition the mask sampling procedure in order to always allow (at least) one code to be non-zero. During execution, we directly zero out the non-available modalities. Finally, after dropout is applied, we concatenate the resulting codes to be used as input to the multimodal encoder. We can define the naive multimodal encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$ as,

$$q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) \triangleq q_{\phi^t}\left(\boldsymbol{z}_\pi \mid \boldsymbol{c}_d\right). \tag{5.17}$$

5.4.2 NEXUS Encoder

Beyond the naive concatenation function, we explore other solutions to merge the multimodal information in the modality codes. We propose the *NEXUS* encoder, depicted in Fig. 5.9. Following recent work in Graph Neural Networks [60], we approach the encoding process of multimodal data as a Directed Acyclical Graph (DAG), in which the nodes of the graph correspond to the modality codes $\boldsymbol{c}_{1:M}$ and the multimodal latent representation $\boldsymbol{z}_\pi$. Each modality code has a single directed edge towards the multimodal note $\boldsymbol{z}_\pi$. We can define the flow of information in the graph as,

$$\boldsymbol{z}_\pi \leftarrow f(\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_M), \tag{5.18}$$

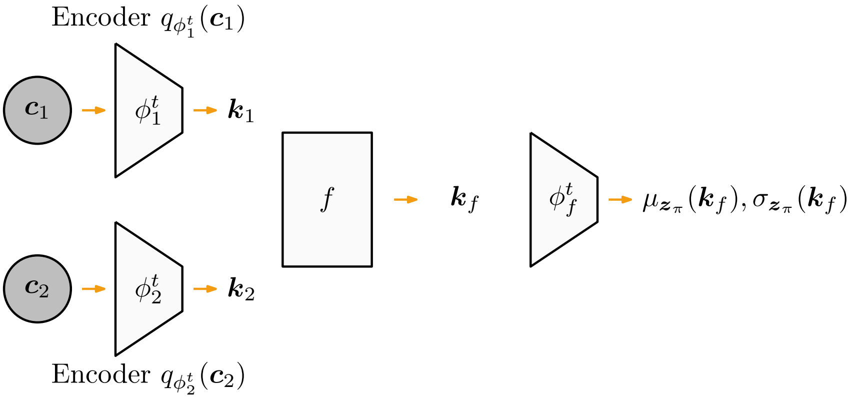

Encoder $q_{\phi_1^t}(\boldsymbol{c}_1)$
$\boldsymbol{c}_1$
ϕ_1^t
$\boldsymbol{k}_1$
f
$\boldsymbol{k}_f$
ϕ_f^t
$\mu_{\boldsymbol{z}_\pi}(\boldsymbol{k}_f), \sigma_{\boldsymbol{z}_\pi}(\boldsymbol{k}_f)$
$\boldsymbol{c}_2$
ϕ_2^t
$\boldsymbol{k}_2$
Encoder $q_{\phi_2^t}(\boldsymbol{c}_2)$

Figure 5.9: The *NEXUS* joint-modality encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$, instantiated in a scenario with two modalities $\boldsymbol{x}_1, \boldsymbol{x}_2$: the modality codes $\boldsymbol{c}_1, \boldsymbol{c}_2$ are initially mapped to a common-dimensionality representation $\boldsymbol{k}_1, \boldsymbol{k}_2 \in \mathbb{R}^k$. The representations are subsequently merged using an *aggregation* function f, resulting in a representation $\boldsymbol{k}_f$ which is encoded to generate the multimodal representation $\boldsymbol{z}_\pi$.

where we define an *aggregation* function $f^{(M)} : \{\boldsymbol{k}_1, \dots, \boldsymbol{k}_M\} \rightarrow \boldsymbol{k}_f \in \mathbb{R}^k$, responsible for aggregating the information provided by each modality. As the function requires that the samples provided are all of the same dimensionality, we process the modality codes using modality-specific *projector* networks $q_{\phi_m^t}(\boldsymbol{c}_m)$ to reduce all samples to common dimensionality d_k. Thus, we can define the NEXUS multimodal encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$ as,

$$q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) \triangleq q_{\phi^t}\left(\boldsymbol{z}_\pi \mid f\left(q_{\phi_1^t}(\boldsymbol{c}_1), q_{\phi_2^t}(\boldsymbol{c}_2), \dots, q_{\phi_M^t}(\boldsymbol{c}_M)\right)\right). \tag{5.19}$$

Several choices of an *aggregation* function can be employed, from simple concatenation to more complex recurrent networks, such as RNNs and LSTM. Empirically, we found a simple *mean* function to be is suitable for the procedure. By employing an *aggregation* function, we allow the multimodal encoding procedure to consider an arbitrary number of modalities, thus addressing the **scalability** issue. Moreover, the multimodal representation can be encoded considering information provided by all modalities (or any subset of available modalities), thus addressing the **compositionality** issue.

5.4.2.1 Forced Perceptual Dropout (FPD) Training Scheme

While the *aggregation* function is able to consider any subset of available modalities to encode the multimodal representation $\boldsymbol{z}_\pi$, by always providing all modalities during training, the model may lack robustness to missing modalities at test time and, as such, not be able to perform cross-modality inference. To address this issue, we propose a novel training scheme for the multimodal encoder which we denote by *Forced Perceptual Dropout* (FPD), whose pseudo-code is presented in Algorithm 1.

During training, we first determine if the dropout mechanism is to be applied for that data sample (line 4), by sampling from a Bernoulli distribution,

$$\mathbb{1}^d \sim \text{Bern}\,(\rho), \tag{5.20}$$

where the user-defined parameter ρ defines the probability of dropout occurring. If $\mathbb{1}^d = 1$, we sample the number of codes to drop (line 6), from an uniform distribution $U\{1, M-1\}$.

Algorithm 1 Forced Perceptual Dropout (FPD)

```
1: Input: Dropout parameter $\rho$; batch-size $N$; batch of modality codes $\boldsymbol{c}_{1:M} = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_M\}$
2: for $n = 1, 2, \dots, N$ do
3:     Define $\boldsymbol{c}_n = \boldsymbol{c}_{1:M,n}$
4:     Sample dropout indicator, $\mathbb{1}^d_n \sim \text{Bern}(\rho)$
5:     if $\mathbb{1}^d_n = 1$ then
6:         Sample number of codes do drop, $k \sim U\{1, M-1\}$
7:         Sample without replacement subset of codes $\boldsymbol{c}^d_n \in \boldsymbol{c}_n$, with $|\boldsymbol{c}^d_n| = k$
8:         Encode $\boldsymbol{z}_{\pi,n} \sim q_{\phi^t}(\cdot \mid \boldsymbol{c}^d_n)$
9:     else
10:        Encode $\boldsymbol{z}_{\pi,n} \sim q_{\phi^t}(\cdot \mid \boldsymbol{c}_n)$
```

Finally, given a complete set of modality codes $\boldsymbol{c}_{1:M} = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_M\}$, we define the smaller "available" subset $\boldsymbol{c}^d \in \boldsymbol{c}_{1:M}$ (line 7), by sampling codes from the complete set, without replacement. We define the multimodal encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$ accordingly to the following rule:

$$q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) = \begin{cases} q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}^d), & \text{if } \mathbb{1}^d = 1 \\ q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}), & \text{otherwise} \end{cases} \tag{5.21}$$

We repeat the FPD procedure for each sample in a training batch. Other multimodal models employ similar mechanisms to improve robustness to missing modalities, such as the subsampling training scheme of MVAE [179]. However, that scheme requires multiple forward passes of the whole batch of data through the model (and multiple gradient computations), proportional to the number of possible combinations of modalities, and, as such, is computationally intensive in scenarios with large number of modalities. On the other hand, FPD is applied in a single forward pass, lessening the computational overhead.

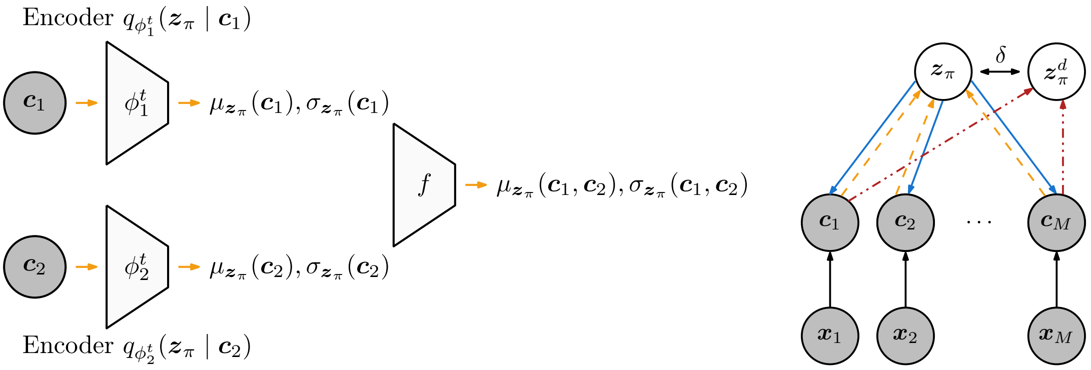
By employing FPD, we are forcing the model to learn to encode a multimodal representation, able to generate all modality-specific representations, despite not being given complete multimodal information. In this way, we explicitly account for cross-modality inference during training and promote the robustness of the model to missing-modality information at test-time.

5.4.3 ALMA Encoder

To mitigate the effect of propagating partial information throughout the network during training, we explore an additional solution to robustly encode the multimodal representation. We propose the *ALMA* joint-modality encoder, shown in Fig. 5.10a, where we approximate the joint-modality approximate posterior distribution $p(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$ with a product-of-experts (PoE) encoder solution [179],

$$p(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) = p(\boldsymbol{z}_\pi) \prod_{m=1}^{M} q(\boldsymbol{z}_\pi \mid \boldsymbol{c}_m). \tag{5.22}$$

This solution gracefully scales to an arbitrarily large number of modalities: assuming that both the prior and posteriors are Gaussian distributions, the product-of-experts distribution is itself a Gaussian distribution with mean $\mu = \left(\sum_m \mu_m T_m\right) \left(\sum_m T_m\right)^{-1}$ and covariance


Encoder $q_{\phi_1^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_1)$
$\boldsymbol{c}_1$
ϕ_1^t
$\mu_{\boldsymbol{z}_\pi}(\boldsymbol{c}_1), \sigma_{\boldsymbol{z}_\pi}(\boldsymbol{c}_1)$
f
$\mu_{\boldsymbol{z}_\pi}(\boldsymbol{c}_1, \boldsymbol{c}_2), \sigma_{\boldsymbol{z}_\pi}(\boldsymbol{c}_1, \boldsymbol{c}_2)$
$\boldsymbol{c}_2$
ϕ_2^t
$\mu_{\boldsymbol{z}_\pi}(\boldsymbol{c}_2), \sigma_{\boldsymbol{z}_\pi}(\boldsymbol{c}_2)$
Encoder $q_{\phi_2^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_2)$
$\boldsymbol{z}_\pi$
δ
$\boldsymbol{z}_\pi^d$
$\boldsymbol{c}_1$
$\boldsymbol{c}_2$
$\cdots$
$\boldsymbol{c}_M$
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_M$

(a) (b)

Figure 5.10: The *ALMA* joint-modality encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$: (a) the network architecture instantiated in a scenario with two modalities $\boldsymbol{x}_1, \boldsymbol{x}_2$; (b) we train the multimodal representation employing the Average Latent Multimodal Approximation (ALMA) training scheme.

$\Sigma = (\sum_m T_m)^{-1}$, where $T_m = \Sigma_m^{-1}$ and Σ_m is the covariance of the modality distribution $q(\boldsymbol{z}_\pi \mid \boldsymbol{c}_m)$.

However, the original PoE trained with the subsampling scheme proposed by Wu and Goodman [179] is prone to learn overconfident experts, hindering a robust performance of cross-modality generation across all target modalities [142]. To address this issue, we introduce a novel training scheme for the PoE encoder that we (also) name *Average Latent Multimodal Approximation* (ALMA), depicted in Fig. 5.7c. During training, we use a multimodal distribution for $\boldsymbol{z}_\pi$ that considers all modality codes, i.e., $\boldsymbol{z}_\pi \sim q(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})$. However, we also encode distributions with missing modality codes, one for every possible combination of modalities, yielding "partial-view latent variables" $\boldsymbol{z}_\pi^d \sim q(\boldsymbol{z}_\pi \mid \boldsymbol{c}^d)$. For example, in scenarios with two input modalities $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ we encode $D = 2$ partial multimodal distributions, corresponding to $\boldsymbol{z}_\pi^1 \sim q(\boldsymbol{z}_\pi \mid \boldsymbol{c}_1)$ and $\boldsymbol{z}_\pi^2 \sim q(\boldsymbol{z}_\pi \mid \boldsymbol{c}_2)$. To encode a multimodal representation robust to missing modalities, we force all these distributions to be similar by including additional loss terms, yielding,

$$\ell(\boldsymbol{x}_{1:M}, \boldsymbol{c}_{1:M}) = \ell_b(\boldsymbol{x}_{1:M}) + \ell_t(\boldsymbol{c}_{1:M}) + \frac{\delta}{D} \sum_{d=1}^{D} D_{\mathrm{KL}}^{\star}(q_\phi^t(\cdot \mid \boldsymbol{c}_{1:M}) \parallel q_\phi^t(\cdot \mid \boldsymbol{c}^d)), \tag{5.23}$$

where the parameter δ governs the impact of the approximation loss term and $D_{\mathrm{KL}}^{\star}$ is the symmetrical KL-divergence. The loss function in Eq. 5.23 contains three fundamental terms. The first term corresponds to the bottom-level loss, $\ell_b(\boldsymbol{x}_{1:M})$ in Eq. 5.12, allowing the model to learn a representation unique to each modality directly from data. The second term corresponds to the top-level loss, $\ell_t(\boldsymbol{c}_{1:M})$ in Eq. 5.13, allowing the model to learn a multimodal representation from the modality codes, employing a PoE solution. The third term addresses the overconfident expert phenomena in PoE by approximating the partial distributions to the complete multimodal distribution.

5.5 Multimodal Handwritten Digits Dataset

Cross-modality inference is well suited for phenomena that can be fully described resorting to information provided by distinct modalities. To provide a natural scenario to evaluate

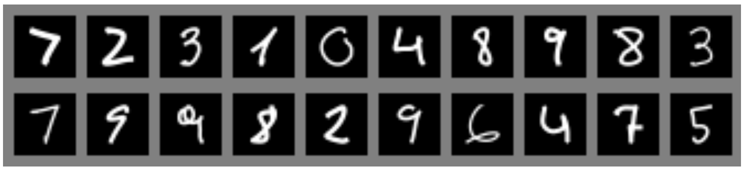


Figure 5.11: Image samples retrieved from the "Multimodal Handwritten Digits" dataset.

it and to address the lack of a benchmark dataset with a large number of modalities, we contribute the *Multimodal Handwritten Digits* (MHD) dataset, containing images, motion trajectories, sounds and labels associated with handwritten digits. The MHD dataset contains $6,000$ samples per digit class of images, trajectories, sounds and labels, partitioned in $50,000$ training and $10,000$ testing samples.

To generate the image and trajectory data, samples of which are presented in Fig. 5.11, we resort to the "UJI Char Pen 2" dataset [97], from which only one-stroke-formed digits are processed. To address the small number of digit samples presented in that dataset, we learn a probabilistic model of each character and re-sample with perturbations constrained in a kinematics feature space, following the procedure described in Yin et al. [184]. We generate $60,000$ samples of 28×28 grayscale images and 200-dimensional representations of the associated trajectories, corresponding to pairs of (x, y) positions for 100 time steps in the $\mathbb{R}^2$ drawing plane. Moreover, we normalize all trajectories to the unit interval.

To obtain the sound modality we extract from the "Speech Commands" dataset the samples belonging to the digit classes [175]. We process the original sound waves, with sample-rate of 16384 Hz, by truncating their duration to 1 second and construct a Mel Spectrogram representation, considering a 512 ms hopping window and 128 mel bins. This results in a 128×32 representation per audio sample, whose values we normalize to a 0-1 range. As the number of sounds per class is less than the required $6,000$, we divide proportionally the representations into training and testing partitions and associate each representation with an unique image-trajectory pair by sampling without replacement. We repeat the sampling procedure until all pairs have a corresponding sound representation associated.

5.6 Evaluation

We evaluate MUSE addressing the following two questions:

(i) What is the performance of MUSE against other baselines considering standard generative metrics?

(ii) Are standard metrics suitable to capture the cross-modality generative performance?

To address (i), in Section 5.6.1 we evaluate MUSE against other state-of-the-art variational autoencoder-based methods in benchmark datasets, using standard likelihood-based metrics. Our results lead us to (ii), in Section 5.7, where we take a closer look at the results and discuss whether current likelihood-based metrics adequately capture the cross-modality inference capabilities of generative models. Our discussion leads to a set of alternative metrics for classification scenarios. We provide both quantitative and qualitative results that attest that our model simultaneously allows and effective single-modal and cross-modality generative processes, regardless of the nature of the target modality.

Methodology We consider MUSE with the ALMA joint-encoder (Section 5.4.3), and leave a comparative study of the different encoder solutions proposed for Appendix C. We evaluate MUSE against two baseline models, presented in Section 5.1, that are agnostic to the nature of input modalities and can handle an arbitrary number of modalities: the MVAE [179] and MMVAE [142] models. We use the authors' publicly available code.[3] To train all models, we use the author's training loss functions, without importance-weighted sampling, and the suggested hyper-parameters.[4] We train MUSE with minimal hyper-parameter tuning: following literature standards we up-weight the reconstruction term of the lower-dimensional label modality. We assume that RGB and grayscale image distributions and latent variable prior distributions are Gaussian, while label and black-and-white image distributions are Bernoulli. Code, network architectures and hyper-parameters are available in Appendix.

Datasets The MNIST dataset is a two-modality scenario ($M = 2$), where $\boldsymbol{x}_1 \in \mathbb{R}^{1\times28\times28}$ is a grayscale image of a handwritten digit and $\boldsymbol{x}_2 \in \mathbb{R}^{10}$ is the associated label [88]. We use a total representation space of 64 dimensions for all models. For MUSE, we set the image-specific latent space $\boldsymbol{z}_1 \in \mathbb{R}^{50}$, the label-specific latent space $\boldsymbol{z}_2 \in \mathbb{R}^4$ and the multimodal latent space $\boldsymbol{z}_\pi \in \mathbb{R}^{10}$.

The CelebA dataset is significantly more challenging. It is a two-modality dataset ($M = 2$) of human face images $\boldsymbol{x}_1 \in \mathbb{R}^{3\times64\times64}$ and associated semantic attributes $\boldsymbol{x}_2 \in \mathbb{R}^{40}$. We use a total representation space of 100 dimensions for all models. For MUSE, we set the image-specific $\boldsymbol{z}_1 \in \mathbb{R}^{60}$, the attribute-specific latent space $\boldsymbol{z}_2 \in \mathbb{R}^{20}$ and the multimodal latent space $\boldsymbol{z}_\pi \in \mathbb{R}^{20}$.

The MNIST-SVHN scenario considers three different modalities ($M = 3$), where $\boldsymbol{x}_1 \in \mathbb{R}^{1\times28\times28}$ are grayscale images of handwritten digits from the MNIST dataset, $\boldsymbol{x}_2 \in \mathbb{R}^{3\times32\times32}$ are RGB images of house numbers from the SVHN dataset [113] and $\boldsymbol{x}_3 \in \mathbb{R}^{10}$ are the associated labels. We define a total representation space of 100 dimensions for all models. For MUSE, we set the "MNIST"-specific latent space $\boldsymbol{z}_1 \in \mathbb{R}^{40}$, a "SVHN"-specific latent space $\boldsymbol{z}_2 \in \mathbb{R}^{40}$, a label-specific latent space $\boldsymbol{z}_3 \in \mathbb{R}^4$, and a multimodal latent space $\boldsymbol{z}_\pi \in \mathbb{R}^{16}$.

Metrics We provide quantitative and qualitative assessment of MUSE as a multimodal generative model. For the qualitative assessment, we show examples of the images generated by the different models (more examples in Appendix C). For the quantitative assessment, we provide marginal, joint-modality and conditional log-likelihoods, averaged over 5 independently-seeded runs.

5.6.1 Results

Our results are summarized in Table 5.1. We can see that, in the two-modality scenarios, MUSE outperforms all baselines in terms of the image marginal likelihood, $\log p(\boldsymbol{x}_1)$. Regarding label marginal likelihood $\log p(\boldsymbol{x}_2)$, MUSE performs on par with MMVAE in the MNIST dataset, despite employing a representation space 16 times smaller. The reduced dimensionality of the label latent space in MUSE was a choice made for fair evaluation against the baselines: the hierarchical design of our model allows the adjustment of the

[3] The MVAE model is taken from `https://github.com/mhw32/multimodal-vae-public` and the MMVAE model is taken from `https://github.com/iffsid/mmvae`.

[4] We train all models for 100 epochs.

Table 5.1: Standard metrics for generative performance in the different datasets (best results in bold). We used 5000 importance samples for the MNIST dataset and 1000 importance samples for the CelebA dataset and the MNIST-SVHN scenario. All results averaged over 5 independent runs. Higher is better.

(a) MNIST

Model	$\log p(\boldsymbol{x}_1)$	$\log p(\boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_1, \boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$
MUSE	$\mathbf{-24.06 \pm 0.03}$	$\mathbf{-2.41 \pm 0.02}$	-34.72 ± 1.47	-33.97 ± 1.16	$-\ \mathbf{4.73 \pm 0.15}$
MVAE	-24.45 ± 0.14	$\mathbf{-2.38 \pm 0.01}$	$\mathbf{-25.46 \pm 0.28}$	$\mathbf{-29.23 \pm 1.00}$	$-\ 9.60 \pm 0.38$
MMVAE	-25.96 ± 0.11	-2.82 ± 0.07	-	-39.94 ± 0.48	-10.01 ± 0.23

(b) CelebA

Model	$\log p(\boldsymbol{x}_1)$	$\log p(\boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_1, \boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$
MUSE	$\mathbf{-163.63 \pm 0.13}$	-19.69 ± 0.09	-288.25 ± 3.70	$\mathbf{-394.65 \pm 2.22}$	$-\ \mathbf{47.48 \pm 0.46}$
MVAE	-165.88 ± 0.40	$\mathbf{-15.21 \pm 0.04}$	$\mathbf{-181.56 \pm 0.44}$	-471.50 ± 1.67	$-\ 72.36 \pm 0.53$
MMVAE	-548.11 ± 10.46	-28.43 ± 0.25	-	-787.24 ± 4.07	-108.50 ± 0.76

(c) MNIST-SVHN

Model	$\log p(\boldsymbol{x}_1)$	$\log p(\boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_1 \vert \boldsymbol{x}_2)$	$\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2, \boldsymbol{x}_3)$	$\log p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$
MUSE	$\mathbf{-24.23 \pm 0.02}$	$\mathbf{-36.03 \pm 0.04}$	$\mathbf{-39.93 \pm 1.45}$	$\mathbf{-37.66 \pm 1.42}$	$-\ 64.99 \pm 3.35$
MVAE	-24.34 ± 0.07	$\mathbf{-36.17 \pm 0.22}$	-43.44 ± 0.19	-40.96 ± 0.86	$-\ \mathbf{55.51 \pm 0.18}$
MMVAE	-28.21 ± 0.06	-41.33 ± 0.23	-54.84 ± 0.47	-	-166.14 ± 0.76

capacity of each modality-specific latent space to the inherent complexity of the modality. The results in the three-modality scenario again show that MUSE outperforms the baselines in single-modality generation $\log p(\boldsymbol{x}_1)$ and $\log p(\boldsymbol{x}_2)$.

Regarding the joint-modality log-likelihood, $\log p(\boldsymbol{x}_1, \boldsymbol{x}_2)$, we can understand the lower generative performance of MUSE on joint-modality encoding by observing the differences between bottom-level and top-level image reconstructions (Fig. 5.12). We see, for example, that the image sample corresponding to digit “3” reconstructed from the top multimodal representation (Fig. 5.12b) is more *prototypical* than the image reconstructed from the bottom modality-specific representation (Fig. 5.12a). Such abstraction is expected, since the top representation level merges the information from the different modalities, leading to the generation of coherent (but more prototypical) modality codes. It is such encoding that enables successful cross-modality inference, but implies that reconstructed samples from multimodal representation lose some variance for higher-dimensional modalities (e.g., image) while accentuating semantic features of the input (e.g., digit class).

Regarding the conditional log-likelihoods metrics $\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$ and $\log p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$, MUSE outperforms all other baselines in the CelebA dataset. We present image samples generated from label information from both MNIST and CelebA datasets in Fig 5.13. The quantitative results are aligned with the qualitative assessment from the observation of the generated images: MUSE is the only model able to generate high-quality, diverse, image samples, semantically coherent with the label information. MVAE generates high-quality but incoherent samples, due to the overconfident expert problem discussed in Shi et al. [142]. On the other hand, MMVAE generates coherent but low-quality samples, showing that the

(a) Image reconstruction from $\boldsymbol{z}_1$

(b) Image reconstruction from $\boldsymbol{z}_\pi$

Figure 5.12: MNIST samples reconstructed from (a) the modality-specific latent space $\boldsymbol{z}_1$; and (b) the multimodal latent space $\boldsymbol{z}_\pi$. We show, in the top row, the original image and, in the bottom row, the reconstructed image. We highlight examples where the abstraction provided by the multimodal latent space leads to the generation of more prototypical images compared with the modality-specific reconstructions.

default MoE solution struggles to learn a rich representation of both modalities. In the three-modality scenario, MUSE, in contrast with MMVAE, is able to consider the information provided by two modalities ($\boldsymbol{x}_2$ and $\boldsymbol{x}_3$) in its cross-modality generation process, indicated by the increase in log-likelihood performance from $\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$ to $\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2, \boldsymbol{x}_3)$.

Finally, we note that—for the MNIST dataset—the quantitative results for the label-to-image conditional log-likelihood, $\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$, are at odds with the qualitative assessment from the observation of the image samples (Fig. 5.13): MVAE seems to outperform MUSE in terms of $\log p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$, yet fails to generate coherent image samples from label information. Such contradictory evidence prompts a more in-depth discussion of standard metrics to evaluate cross-modality generative performance.

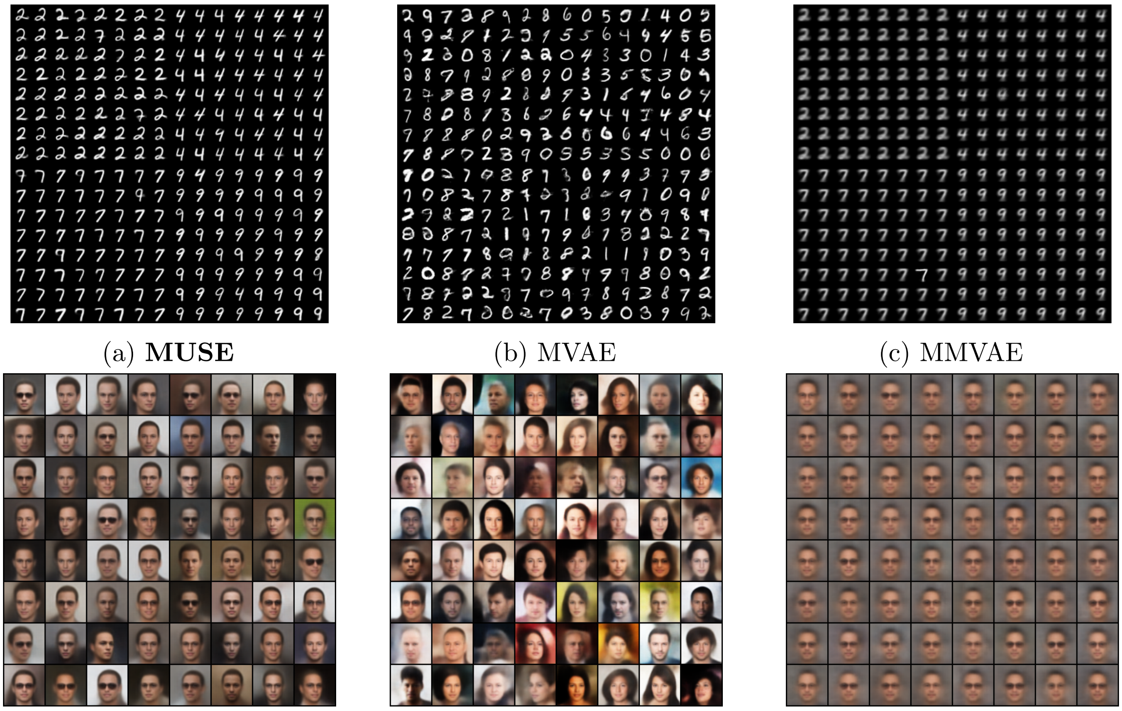
5.7 Metrics for Cross-Modality Inference

Our results in the MNIST dataset suggest a discrepancy between the conditional log-likelihood $\log p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ in Table 5.1, and actual cross-modality generative performance of the models, in terms of the images generated from label information, as seen in Fig. 5.13. We note that, with respect to this particular metric, *the distribution used as the metric for the evaluation is the one learned by the model being evaluated.* As such, if the model learns, say, a "multimodal" representation that disregards the label information, the conditional log-likelihood metric will actually evaluate a marginal probability. Such phenomenon can actually be observed by noting the similarity between the values of $\log p(\boldsymbol{x}_1)$ and $\log p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ for MVAE.

This observation motivates the need for alternative metrics to evaluate the cross-modality

Table 5.2: Evaluation of the cross-modality generation performance in the MNIST dataset, considering the proposed alternative metrics (KL distance, accuracy, modality distance). KL distance is normalized to the dimensionality of the image modality. Cross-modality accuracy is averaged over both target modalities. Modality distance is only computed for images generated from label inputs. All results are averaged over 5 independent runs.

Model	KL-distance	Cross-Modality Accuracy (%)	Modality Distance
MUSE	**2.79** ± **0.34**	**90.6** ± **7.7**	**261.6** ± **63.7**
MVAE	909.22 ± 349.59	54.3 ± 17.8	**225.3** ± **197.4**
MMVAE	4.71 ± 4.36	60.6 ± 37.6	516.5 ± 82.0



(a) **MUSE** (b) MVAE (c) MMVAE

(d) **MUSE** (e) MVAE (f) MMVAE

Figure 5.13: Image generation from label information: (top row) in the MNIST dataset considering $\boldsymbol{x}_2 = \{2, 4, 7, 9\}$, (bottom row) in the CelebA dataset considering $\boldsymbol{x}_2 = \{\text{Male}, \text{Eyeglasses}, \text{Black Hair}, \text{Receding Hairline}, \text{Goatee}\}$. The MUSE model is the only able to generate high-quality, coherent image samples from label information.

generative performance — namely, metrics that are more "impartial" to the model being evaluated. A possible metric could be a statistical distance, such as the KL divergence, between the output distributions of dataset samples belonging to each class and samples generated by cross-modality inference. The cross-modality distributions can be computed by the multimodal generative models and the dataset distributions are computed by an independent class-specific VAE model. We present the results of such evaluation in the first column of Table 5.2, which highlights that the MVAE baseline struggles to generate coherent image information from labels — as anticipated from the images in Fig. 5.13.

As another alternative metric, Shi et al. [142] proposed the use of *accuracy* to evaluate the cross-modality generation performance in classification scenarios. This metric evaluates the semantic coherence of the samples generated by cross-modality inference, using pre-trained modality-specific classifiers. While relevant, accuracy by itself does not evaluate if the generated samples are similar to those available in the original dataset. We propose to also evaluate the relative *quality* of the generated samples, employing class-and-modality-specific auto-encoders, in a metric that we name *modality distance*. For each class in the dataset, we encode representations of samples both from the dataset and generated by cross-modality inference, resulting in a distribution of real dataset representations $\mathcal{N}(\mu_r, \Sigma_r)$ and of generated representations $\mathcal{N}(\mu_g, \Sigma_g)$. The modality distance is then given by the Fréchet distance between the two distributions, averaged over all classes [64]. We present a depiction of the evaluation metrics in Fig. 5.14 and leave a more detailed discussion the metrics employed in this work for Appendix B.

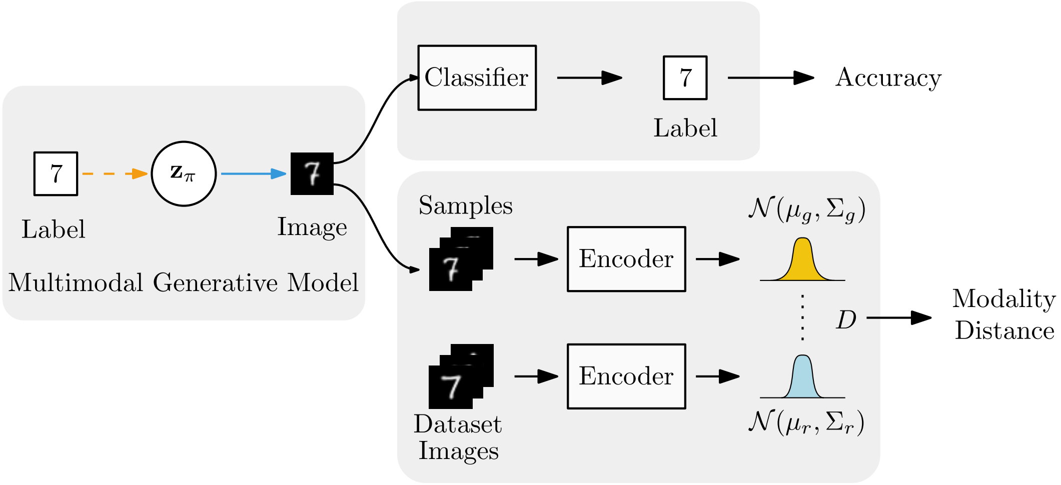


Figure 5.14: The proposed complementary metrics to evaluate the computational cross-modality inference process: we evaluate both the semantic coherence of the generated samples (*accuracy*), as well as their quality (*modality distance*) A robust cross-modality inference performance should provide samples in the high accuracy and low modality distance regime regardless of the complexity or nature of the target modality (best viewed with zoom).

In Table 5.2 we also include the performance of all methods in terms of accuracy and modality distance in the MNIST dataset. MUSE is the only model able to generate high-quality and semantically coherent samples through cross-modality inference: our model outperforms all other models in cross-modality accuracy and performs on par with MVAE in modality distance. MVAE is able to generate high-quality samples, yet struggle in generating samples that are coherent with the information provided to the model. On the other hand, MMVAE generates coherent, yet low-quality, samples.

5.7.1 Multimodal Handwritten Digits

In this final evaluation, we highlight the cross-modality inference performance of MUSE in the MHD dataset, which contains images $\boldsymbol{x}_i$, trajectories $\boldsymbol{x}_t$, sounds $\boldsymbol{x}_s$, and labels $\boldsymbol{x}_l$, associated with handwritten digits. In this challenging scenario, we show that MUSE is the only model able to perform cross-modality inference, regardless of the target modality and considering the information provided by any subset of available modalities.

Due to the high complexity of the sound modality, we pretrain a SigmaVAE model to learn a modality-specific representation of sound. We resort to the authors' optimal training scheme and consider a regularization hyperparameter $\beta = 10$ [137]. We employ the pretrained SigmaVAE model as the bottom-level sound-specific encoder and decoder. For a fair comparison, we evaluate our model against the hierarchical versions of the baselines, sharing the same network architectures and training hyperparameters of our own.

We evaluate the cross-modality generation performance for each target modality, as a function of the number of modalities provided to the models. The results are averaged over all combinations of provided modalities and 5 independent runs, and presented in Table 5.3.

The results show that the MUSE model outperforms the other baselines in accuracy and MFD across all target modalities, even in scenarios of large number of modalities, addressing the **scalability** issue. As the number of modalities provided to the model

Table 5.3: Evaluation of cross-modality generation in the MHD dataset, as a function of the number of the observed modalities provided to the models and the target CMI generated modality (I = Image, T = Trajectory, S = Sound, L = Label). Results over all combinations of input modalities and over 5 independent runs. Higher is better.

		Accuracy (%)		Modality Distance	
Model	Target	1 Modality	3 Modalities	1 Modality	3 Modalities
MUSE	I	**78.5 ± 13.9**	**98.7 ± 00.4**	**89.4 ± 23.6**	**100.2 ± 22.4**
	T	**73.9 ± 14.0**	**95.9 ± 00.9**	**265.0 ± 106.7**	**215.9 ± 90.1**
	S	**77.6 ± 09.7**	**93.1 ± 00.9**	**3354 ± 621**	**4105 ± 721**
	L	**72.0 ± 08.3**	**95.9 ± 01.1**	NA	NA
MVAE	I	28.6 ± 05.2	80.9 ± 07.2	228.4 ± 61.8	201.3 ± 45.2
	T	13.7 ± 04.6	17.8 ± 03.7	399.3 ± 179.1	391.0 ± 178.7
	S	33.6 ± 14.2	88.6 ± 09.7	6608 ± 1471	8133 ± 1751
	L	23.4 ± 13.4	39.9 ± 07.7	NA	NA
MMVAE	I	**66.1 ± 39.8**	–	236.9 ± 62.7	–
	T	**63.8 ± 38.1**	–	547.8 ± 235.4	–
	S	**70.4 ± 05.4**	–	14998 ± 1325	–
	L	**66.0 ± 39.6**	–	NA	NA

increases so does the accuracy of the respective cross-modality generated samples, addressing the **compositionality** issue. This is to be expected as the confidence of the model in generating samples of the correct class increases as more information is provided. Moreover, contrary to the MMVAE baseline model, MUSE is able to take advantage of this additional information provided by multiple modalities. The results also reveal that the MVAE baseline is able to generate trajectory and sound samples with low MFD at the cost of the accuracy of these lower dimensional modalities. Once again this is the result of the overconfident expert problem of this model. And while the hierarchical extension reduced this effect in the previous scenario, the same extension is less effective in a scenario where the differences in dimensionality of the modality-specific latent spaces is significantly greater.

Overall, the results show that the MUSE model outperforms the baselines by being the only model considered that is able to generate data with both high accuracy and low MFD, regardless of the target and available source modalities to the process, allowing for *effective* cross-modality inference.

5.8 Ablation Study

In Appendix C we perform a comparative study of the different joint-encoder methods proposed for MUSE, as well as of different architectural choices for our model, e.g., in regards to employing hierarchical representation spaces or not. The results show that MUSE with the ALMA encoder outperforms the other joint-encoder solutions in regards to high-quality, coherent cross-modality generation. Moreover, the hierarchical design of MUSE plays a fundamental role in allowing generation of high-quality samples through cross-modality inference.

5.9 Concluding Remarks

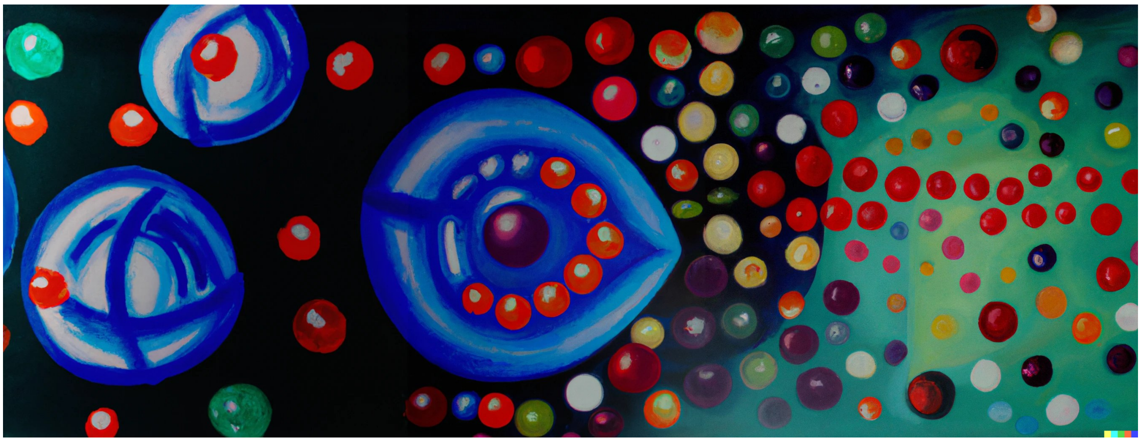
In this chapter, we discussed how to learn representations of an arbitrary number of heterogeneous data sources without supervision, employing a multimodal generative model. We have centered our discussion in the computational *cross-modality inference* problem. We have shown how current models fall short of addressing all the desiderata of cross-modality generation. Taking inspiration from human perception, we argued for considering *hierarchy* in the design of multimodal generative models. We presented MUSE, a novel architecture that considers distinct modality-specific and multimodal representation spaces. We proposed three different solutions to merge multimodal information while addressing the requirements of computational cross-modality inference. We evaluate extensively MUSE against relevant baselines in multiple scenarios of increasing complexity, in regards to the number and nature of the modalities involved. The results across all evaluations show that MUSE is the only model able to satisfy all conditions for effective cross-modality inference.

The MUSE model introduces a separation between modality-specific and multimodal representations, providing a *modular* environment to address other fundamental issues in the actuation of artificial agents, such as robots. Equipped with multiple sensors (prone to failing and upgrade), robots should be able to robustly combine data coming from new sensors with the previously available sensory channels, learning in a sample-efficient way how to encode a novel multimodal representation. Indeed, such plug-and-play multimodal sensory fusion represents a fundamental technical issue to be addressed in long-term research in AI [51]. In future work, we wish to investigate solutions that allow plug-and-play multimodal sensory fusion within the hierarchical framework of MUSE.

We have shown that the MUSE model is able to learn a multimodal representation robust to missing modality information, suitable for downstream generation tasks. However, it does so with a significant computational cost, due to the encoder-decoder architecture employed by the model. In the next chapter, we explore self-supervised learning approaches to achieve efficient multimodal representation learning, suitable to perform robustly a wider variety of downstream tasks with missing modality information at test time.

Chapter 6

Multimodal Contrastive Representation Learning



"The greater the contrast, the greater the potential."
Carl Jung

Good representations of multimodal data aim to *capture the joint semantics* from individual modalities necessary for performing a given downstream task. Additionally, in scenarios such as real-world classification and control, it is essential that the obtained representations are also *robust to missing modality* information [107, 161, 187]. To do so, the unique characteristics of each modality need to be processed accordingly and efficiently combined, which remains a challenging problem known as the *heterogeneity gap* [56].

An intuitive idea to mitigate the heterogeneity gap is to project heterogeneous data into a shared representation space such that the representations of complete and modality-specific representations are *aligned*. In Chapter 5, we explored how multimodal generative models can learn a shared representation space, to allow for efficient cross-modality inference. However, as we show in Section 6.1, these approaches often struggle to align complete and modality-specific representations, due to the demanding modality-specific reconstruction objective. This misalignment may lead to a poor performance in downstream tasks, under conditions of partial perceptual availability.

In this chapter, we introduce a novel multimodal representation learning framework that builds upon the simple idea of explicitly aligning modality-specific and complete representations in a latent representation space, without requiring an encoder-decoder architecture. Inspired by recent advances in visual contrastive representation learning [26], we contribute a novel multimodal contrastive loss that explicitly aligns modality-specific representations with the representations obtained from the corresponding complete observation. We instantiate our approach with the *Geometric Multimodal Contrastive* (GMC) representation learning framework. Following the architecture of the MUSE model, GMC also exploits hierarchy by considering a two-level representation space. GMC is instantiated as a collection of *modality-specific* base encoders, processing modality data into a *low-level* representation of fixed dimensionality, and a *shared* projection head, mapping the low-level representations into a *high-level* latent representation space, where the contrastive learning objective is applied. GMC can be scaled to an arbitrary number of modalities, and provides semantically rich representations that are robust to missing modality information.

We extensively evaluate GMC across a variety of challenging problems, such as learning representations without supervision signals (Section 6.3.2) and with auxiliary supervision signals (Section 6.3.3), for downstream classification tasks. We highlight how GMC can be integrated into existing models and show that GMC is able to achieve state-of-the-art classification performance with missing modality information. The main contributions of this chapter are three-fold:

- In Section 6.1, we discuss and empirically demonstrate the *geometric misalignment* of complete and modality-specific latent representations encoded by multimodal representation models;

- In Section 6.2, we propose the *Geometric Multimodal Contrastive* (GMC) representation learning framework, a novel approach that explicitly aligns modality-specific and complete representations of multimodal data. To perform such alignment, we introduce a novel multimodal contrastive loss, inspired by recent advances in visual contrastive representation learning;

- In Section 6.3, we evaluate our approach in two challenging learning scenarios: learning representations without supervision (Section 6.3.2) and with an auxiliary supervision signal (Section 6.3.3) for downstream classification tasks. We show that GMC is able to achieve state-of-the-art classification performance with missing modality information.

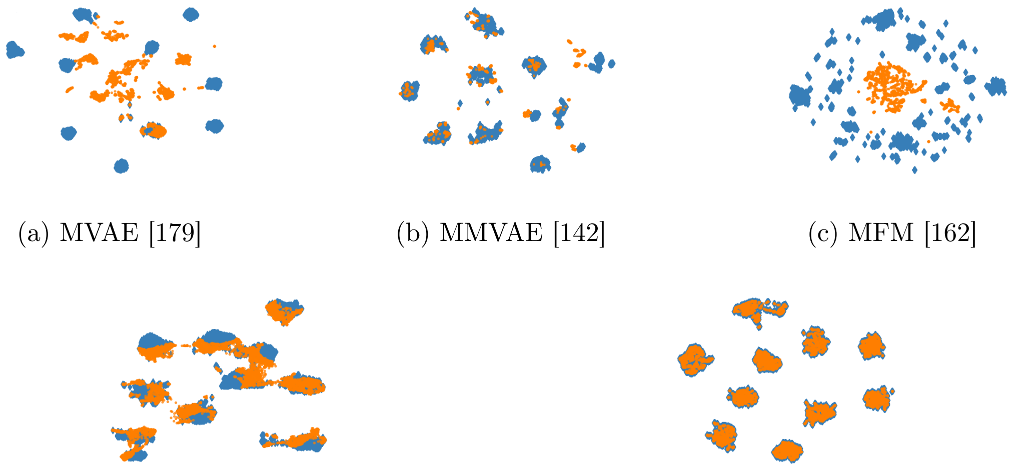
The work described in this chapter has been published in:

- Petra Poklukar*, **Miguel Vasco***, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. *Geometric Multimodal Contrastive Representation Learning.* Proceedings of the 39th International Conference on Machine Learning (ICML). 2022, pp. 17782–17800 [128].

6.1 Geometric Misalignment in Multimodal Representation Learning

We once again consider scenarios where information is provided in the form of tuples $\boldsymbol{x}_{1:M} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$, representing observations provided by M different modalities. We

* Shared first-authorship.



(a) MVAE [179]

(b) MMVAE [142]

(c) MFM [162]

(d) MUSE

(e) **GMC**

Figure 6.1: UMAP visualization of complete representations $\boldsymbol{z}_{1:4}$ (blue) and image representations $\boldsymbol{z}_1$ (orange) in a latent space $\boldsymbol{z} \in \mathbb{R}^{64}$ encoded from several multimodal representation learning models on the MHD dataset considered in Section 6.3.2. Only GMC is able to learn geometrically aligned *modality-specific* and *complete* representations.

refer to the tuples $\boldsymbol{x}_{1:M}$ consisting of all M modalities as *complete* observations and to the single observations $\boldsymbol{x}_m$ as *modality-specific*. The goal is to learn complete representations $\boldsymbol{z}_{1:M}$ of $\boldsymbol{x}_{1:M}$ and modality-specific representations $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_M\}$ of $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ that are:

i) informative, i.e., both $\boldsymbol{z}_{1:M}$ and any of $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_M\}$ contains relevant semantic information for some downstream task, and thus,

ii) robust to missing modalities during test time, i.e., the success of a subsequent downstream task is independent of whether the provided input is the complete representation $\boldsymbol{z}_{1:M}$ or any of the modality-specific representations $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_M\}$.

Prior work has demonstrated success in using complete multimodal representations $\boldsymbol{z}_{1:M}$ in a diverse set of applications, such as image generation [142, 179]. Intuitively, if complete representations $\boldsymbol{z}_{1:M}$ are sufficient to perform a downstream task, then learning modality-specific representations $\boldsymbol{z}_m$ that are geometrically aligned with $\boldsymbol{z}_{1:M}$ in the same representation space should ensure that $\boldsymbol{z}_m$ contain necessary information to perform the task even when $\boldsymbol{z}_{1:M}$ cannot be provided.

Therefore, we investigate the geometric alignment of $\boldsymbol{z}_{1:M}$ and each $\boldsymbol{z}_m$ on several multimodal datasets and state-of-the-art multimodal representation models. As an example, in Fig. 6.1, we visualize the representations encoded by multimodal representation models in the MHD dataset, presented in Section 5.5. We consider complete representations $\boldsymbol{z}_{1:M}$ (in blue) and image-specific representations $\boldsymbol{z}_m$ corresponding (in orange), highlighting that existing approaches produce geometrically misaligned representations. As we empirically show in Section 6.3, this misalignment is consistent across different learning scenarios and datasets, and can lead to a poor performance on downstream tasks.

To fulfill both prior objectives, we propose a novel approach that builds upon the simple idea of geometrically aligning modality-specific representations $\boldsymbol{z}_m$ with the corresponding

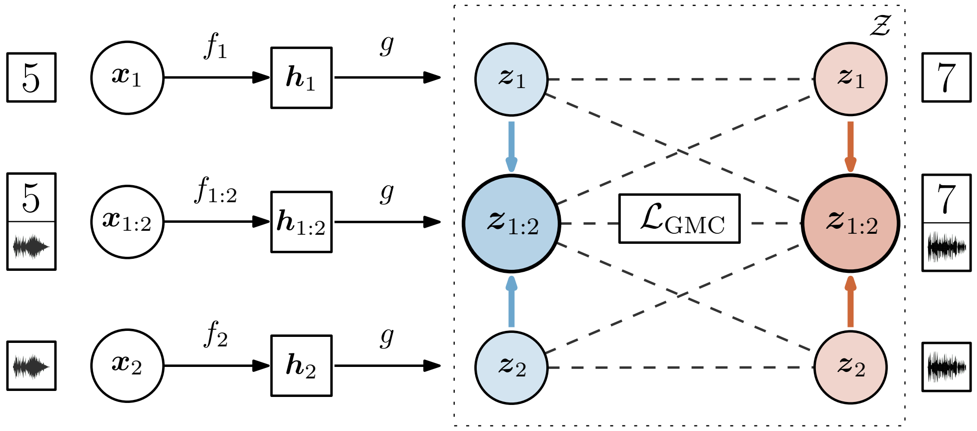


Figure 6.2: The *Geometric Multimodal Contrastive* (GMC) model to learn aligned modality-specific and complete representations of multimodal data: the *modality-specific* base encoders $f(\cdot)$ map observations to low-level representations $\boldsymbol{h}$ that are further projected with a *shared* projection head $g(\cdot)$ to a high-level latent representation space $\mathcal{Z}$, where we apply a novel multimodal contrastive loss $\mathcal{L}_{\text{GMC}}$.

complete representations $\boldsymbol{z}_{1:M}$ in a latent representation space, framing it as a contrastive learning problem.

6.2 Geometric Multimodal Contrastive Learning

We present the *Geometric Multimodal Contrastive* (GMC) framework, visualized in Figure 6.2, consisting of two main components:

- A collection of neural network *modality-specific* base encoders $f(\cdot) = \{f_{1:M}(\cdot)\} \cup \{f_1(\cdot), \dots, f_M(\cdot)\}$, where $f_{1:M}(\cdot)$ and $f_m(\cdot)$ take as input the complete $\boldsymbol{x}_{1:M}$ and modality-specific observations $\boldsymbol{x}_m$, respectively, and output *low-level* d-dimensional representations $\{\boldsymbol{h}_{1:M}, \boldsymbol{h}_1, \dots, \boldsymbol{h}_M\} \in \mathbb{R}^d$;

- A neural network *shared* projection head $g(\cdot)$ that maps the low-level representations given by the base encoders $f(\cdot)$ to the *high-level* latent representations $\{\boldsymbol{z}_{1:M}, \boldsymbol{z}_1, \dots, \boldsymbol{z}_M\} \in \mathbb{R}^s$ over which we apply the contrastive term. The projection head $g(\cdot)$ enables to encode the low-level representations in a shared representation space while preserving modality-specific semantics;

To promote the geometric alignment of complete $\boldsymbol{z}_{1:M}$ and modality-specific representations $\boldsymbol{z}_m$, we consider a contrastive prediction task where the goal is to identify $\boldsymbol{z}_m$ and its corresponding complete representation $\boldsymbol{z}_{1:M}$ in a given mini-batch. Let $\mathcal{B} = \{\boldsymbol{z}^i_{1:M}\}^B_{i=1} \subset g(f(X))$ be a mini-batch of B complete representations. Let (u, v) denote the cosine similarity among vectors u and v and let $\tau \in (0, \infty)$ be the temperature hyperparameter. We denote by,

$$s_{p,q}(i, j) = \exp((z^i_p, z^j_q)/\tau), \tag{6.1}$$

the similarity between representations z^i_p and z^j_q (modality-specific or complete) corresponding to the ith and jth samples from the mini-batch $\mathcal{B}$. For a given modality m, we define positive pairs as $(z_m^i, z_{1:M}^i)$ and $(z_{1:M}^i, z_m^i)$ for $i = 1, \ldots, B$ and treat the remaining pairs as negative ones. In particular, we denote by,

$$\Omega_{p,q}(i) = \sum_{j \neq i} s_{p,p}(i,j) + \sum_{j} s_{p,q}(i,j),$$

the sum of similarities among negative pairs that correspond to the positive pair (z_p^i, z_q^i). We define the contrastive loss for the positive pairs $(z_m^i, z_{1:M}^i)$ and $(z_{1:M}^i, z_m^i)$ as the sum

$$l_m(i) = -\log \frac{s_{m,1:M}(i,i)}{\Omega_{m,1:M}(i)} - \log \frac{s_{1:M,m}(i,i)}{\Omega_{1:M,m}(i)}.$$

Finally, we combine the loss terms for each modality $m = 1, \ldots, M$ and obtain the final training loss,

$$\mathcal{L}_{\mathrm{GMC}}(\mathcal{B}) = \sum_{m=1}^{M} \sum_{i=1}^{B} l_m(i). \tag{6.2}$$

As we only contrast single modality-specific representations to the complete ones, $\mathcal{L}_{\mathrm{GMC}}$ scales linearly to an arbitrary number of modalities. In Section 6.3, we show that $\mathcal{L}_{\mathrm{GMC}}$ can be added as an additional term to existing frameworks to improve their robustness to missing modalities. Moreover, we experimentally demonstrate that the architectures of the base encoders and shared projection head can be flexibly adjusted depending on the task.

6.3 Experiments

We evaluate the quality of the representations learned by GMC on two different scenarios:

- An *unsupervised learning* task, where we learn to encode multimodal representations without supervision on the Multimodal Handwritten Digits (MHD) dataset. We showcase the geometric alignment of representations and demonstrate the superior performance of GMC compared to the baselines on a downstream classification task with missing modalities at test time (Section 6.3.2);

- A *supervised learning* task, where we demonstrate the flexibility of GMC by integrating it into state-of-the-art representation learning models to provide robustness to missing modalities in challenging classification scenarios (Section 6.3.3);

In each corresponding Section, we describe the dataset, baselines, evaluation and training setup used. We report all hyperparameters in Appendix F. All results are averaged over 5 different randomly-seeded runs.

6.3.1 Overview of Delaunay Component Analysis (DCA)

To evaluate the geometric alignment of representations, we employ the recently proposed *Delaunay Component Analysis* (DCA) method, designed for general evaluation of representations [127]. DCA is based on the idea of comparing geometric and topological properties of an evaluation set of representations E with a reference set R, acting as an approximation of the true underlying manifold. The set E is considered to be well aligned with R if its global and local structure resembles well the one captured by R, i.e., the manifolds described by the two sets have similar number, structure and size of connected components of the graphs.

DCA approximates the manifolds described by R and E with a Delaunay neighbourhood graph and derives several scores reflecting their alignment. We consider three of them:

- *Network quality* $q \in [0, 1]$ which measures the overall geometric alignment of R and E in the connected components;

- *Precision* $\mathcal{P} \in [0, 1]$ which measures the proportion of points from E that are contained in geometrically well-aligned components;

- *Recall* $\mathcal{R} \in [0, 1]$ which measure the proportion of points from R, that are contained in geometrically well-aligned components.

To account for all three normalized scores, we define the alignment score A_{DCA} as their harmonic mean:

$$A_{\text{DCA}} = \begin{cases} \frac{3}{(1/\mathcal{P}+1/\mathcal{R}+1/q)}, & \text{if } \mathcal{P}, \mathcal{R}, q > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6.3}$$

In all experiments, we compute $A_{\text{DCA}} \in [0, 1]$ using complete representations $\boldsymbol{z}_{1:M}$ as the reference set R and modality-specific $\boldsymbol{z}_m$ as the evaluation set E, both obtained from samples of the test dataset. For a detailed description of the method and definition of the scores, please refer to Poklukar et al. [127].

6.3.2 Unsupervised Representation Learning Task

In this section, we evaluate both quantitatively and qualitatively how GMC is able to learn multimodal representations of a large number of modalities without an explicit supervision signal, suitable for downstream classification tasks with missing modality information.

Dataset

The MHD dataset is comprised of images (x_1), sounds (x_2), motion trajectories (x_3) and label information (x_4) related to handwriting digits. We collected $60,000$ 28×28 greyscale images per class as well as normalized 200-dimensional representations of trajectories and 128×32-dimensional representations of audio. The dataset is split into $50,000$ training and $10,000$ testing samples. More details regarding the dataset are found in Section 5.5.

Models

We consider several generation-based and fusion-based state-of-the-art multimodal representation methods: MVAE [179], MMVAE [142], MFM [162] and MUSE (Section 5). For a fair comparison, when possible, we employ the same encoder architectures and latent space dimensionality across all baseline models. For GMC, we employ the same modality-specific base encoders $f_m(\cdot)$ as the baselines with an additional base encoder $f_{1:4}(\cdot)$, taking complete observations as input. The shared projection head $g(\cdot)$ comprises of 3 fully-connected layers. We set the temperature $\tau = 0.1$ and consider 64-dimensional low-level and high-level shared representation spaces, i.e., $\boldsymbol{h}, \boldsymbol{z} \in \mathbb{R}^{64}$.

Table 6.1: Results of different multimodal representation methods in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy (%) results averaged over 5 independent runs. Higher is better.

Input	MVAE[7]	MMVAE	MFM	MUSE	**GMC**
Complete ($x_{1:4}$)	100.0 ± 0.00	99.81 ± 0.21	100.0 ± 0.00	$99.99 \pm 4\text{e}{-5}$	100.0 ± 0.00
Image (x_1)	77.94 ± 3.16	94.63 ± 2.61	34.66 ± 6.48	79.37 ± 2.75	$\mathbf{99.75 \pm 0.03}$
Sound (x_2)	61.75 ± 4.59	69.43 ± 26.43	10.07 ± 0.20	41.39 ± 0.18	$\mathbf{93.04 \pm 0.45}$
Trajectory (x_3)	10.03 ± 0.06	95.33 ± 2.56	25.61 ± 5.41	89.49 ± 2.44	$\mathbf{99.96 \pm 0.02}$
Label (x_4)	100.0 ± 0.00	87.99 ± 7.49	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00

Setup

We train all representation models for 100 epochs using a learning rate of 1e−3, employing the training schemes and hyperparameters suggested by the authors (when available). For GMC, we set the temperature $\tau = 0.1$. We follow the established evaluation in the literature using classification as a downstream task [142] and train a 10-class classifier neural network on complete representations $\boldsymbol{z}_{1:M} = g\left(\boldsymbol{h}_{1:M}(\boldsymbol{x}_{1:M})\right)$ from the training split. The classifier is trained for 50 epochs using a learning rate of 1e−3. We report the testing accuracy obtained when the classifier is provided with both complete $\boldsymbol{z}_{1:4}$ and modality-specific representations $\boldsymbol{z}_m$ as inputs.

Classification results

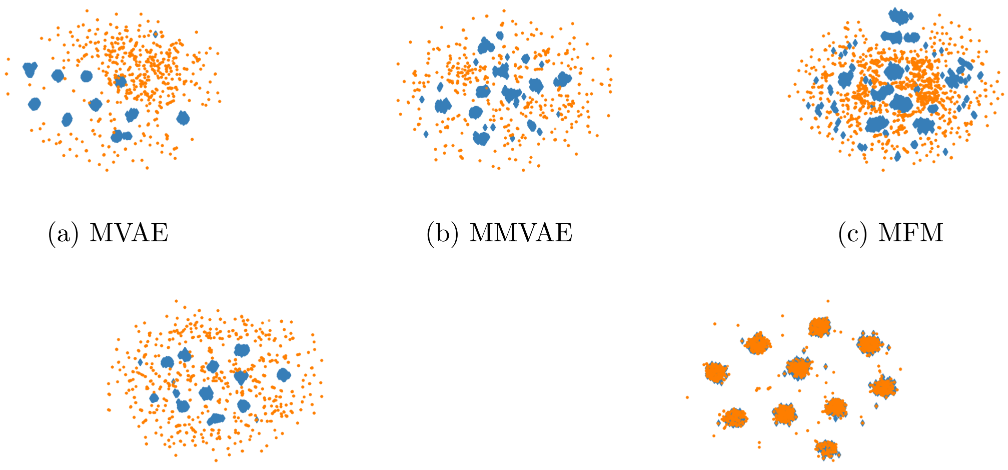
The classification results are shown in Table 6.1[1]. While all the models attain perfect accuracy on $x_{1:4}$ and x_4, we observe that GMC is the only model that successfully performs the task when given only image (x_1), sound (x_2) or trajectory (x_3) as input, significantly outperforming the baselines.

Alignment results

To validate that the superior performance of GMC originates from a better geometric alignment of representations, we evaluate the testing representations obtained from all the models using DCA. For each modality m, we compared the alignment of the evaluation set $E = \{\boldsymbol{z}_m\}$ and the reference set $R = \{\boldsymbol{z}_{1:4}\}$. The obtained alignment scores A_{DCA} are shown in Table 6.2 where we see that GMC outperforms all the considered baselines. For some cases, we observe the obtained representations are completely misaligned yielding $\mathcal{P} = \mathcal{R} = q = 0$. While some of the baselines are to some extend able to align $\boldsymbol{z}_1$ and/or $\boldsymbol{z}_4$ to $\boldsymbol{z}_{1:4}$, GMC is the only method that is able to align even the sound and trajectory representations, $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$, resulting in a superior classification performance.

We additionally validate the geometric alignment by visualizing 2-dimensional UMAP projections [106] of the representations z. In Fig. 6.1 we showed projections of $\boldsymbol{z}_{1:4}$ and image representations $\boldsymbol{z}_1$ obtained using the considered models. We clearly see that GMC not only correctly aligns $\boldsymbol{z}_{1:4}$ and $\boldsymbol{z}_1$ but also separates the representations in 10 clusters. Moreover, we can see that among the baselines only MMVAE and MUSE somewhat align

[1] Results for the MVAE model averaged over 3 randomly-seeded runs, as the training diverged in the remaining seeds.



(a) MVAE

(b) MMVAE

(c) MFM

(d) MUSE

(e) **GMC**

Figure 6.3: UMAP visualization of complete representations $\boldsymbol{z}_{1:4}$ (blue) and sound representations $\boldsymbol{z}_2$ (orange) obtained from several state-of-the-art multimodal representation learning models on the MHD dataset. Best viewed in color.

the representations which is on par with the quantitative results reported in Table 6.2. For MVAE and MFM, Fig. 6.1 visually supports the obtained alignment score. In Fig. 6.3, we show projections of complete $\boldsymbol{z}_{1:4}$ and sound representations $\boldsymbol{z}_2$ obtained using the considered models. Once again, the visual representation aligns with the quantitative results of accuracy (Table 6.1) and geometric alignment (Table 6.2): GMC is the only model able to learn sound-specific representations aligned with the complete representations. Note that points marked as outliers by DCA are omitted from the visualization.

6.3.3 Supervised Learning Task

In this section, we evaluate the flexibility of GMC by adjusting both the architecture of the model and its training procedure to receive an additional supervision signal that guides the learning process of complete representations. In this way, we demonstrate how GMC can be integrated into existing approaches to provide additional robustness to missing modalities at test time, with minimal computational cost.

Table 6.2: Alignment scores A_{DCA} of the models in the MHD dataset, evaluating the geometric alignment of complete representations $z_{1:4}$ and modality-specific ones $\{z_1, \dots z_4\}$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	MVAE[7]	MMVAE	MFM	MUSE	**GMC**
Complete ($\boldsymbol{z}_{1:4}$)	Image ($\boldsymbol{z}_1$)	0.01 ± 0.01	0.21 ± 0.29	0.00 ± 0.00	0.54 ± 0.44	$\mathbf{0.96 \pm 0.02}$
Complete ($\boldsymbol{z}_{1:4}$)	Sound ($\boldsymbol{z}_2$)	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.87 \pm 0.16}$
Complete ($\boldsymbol{z}_{1:4}$)	Trajectory ($\boldsymbol{z}_3$)	0.00 ± 0.00	0.01 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.86 \pm 0.05}$
Complete ($\boldsymbol{z}_{1:4}$)	Label ($\boldsymbol{z}_4$)	0.99 ± 0.01	0.74 ± 0.22	0.85 ± 0.06	0.93 ± 0.05	$\mathbf{1.00 \pm 0.00}$

Table 6.3: Results of different multimodal representation methods in the CMU-MOSEI dataset, in a classification task under complete and partial observations. Results averaged over 5 independent runs. Arrows indicate the direction of improvement.

Metric	Baseline	GMC
MAE (↓)	0.643 ± 0.019	$\mathbf{0.634 \pm 0.008}$
Cor (↑)	$\mathbf{0.664 \pm 0.004}$	0.653 ± 0.004
F1 (↑)	$\mathbf{0.809 \pm 0.003}$	0.798 ± 0.008
Acc (%, ↑)	$\mathbf{80.75 \pm 00.28}$	79.73 ± 00.69

(a) Complete Observations ($x_{1:3}$)

Metric	Baseline	**GMC**
MAE (↓)	0.805 ± 0.028	$\mathbf{0.712 \pm 0.015}$
Cor (↑)	0.427 ± 0.061	$\mathbf{0.590 \pm 0.013}$
F1 (↑)	0.713 ± 0.086	$\mathbf{0.779 \pm 0.005}$
Acc (%, ↑)	66.53 ± 09.86	$\mathbf{77.85 \pm 00.36}$

(b) Text Observations (x_1)

Metric	Baseline	**GMC**
MAE (↓)	0.873 ± 0.065	$\mathbf{0.837 \pm 0.008}$
Cor (↑)	0.090 ± 0.062	$\mathbf{0.256 \pm 0.007}$
F1 (↑)	0.622 ± 0.122	$\mathbf{0.676 \pm 0.015}$
Acc (%, ↑)	53.17 ± 09.47	$\mathbf{65.59 \pm 00.62}$

(c) Audio Observations (x_2)

Metric	Baseline	**GMC**
MAE (↓)	1.025 ± 0.164	$\mathbf{0.845 \pm 0.010}$
Cor (↑)	0.110 ± 0.060	$\mathbf{0.278 \pm 0.011}$
F1 (↑)	0.574 ± 0.095	$\mathbf{0.655 \pm 0.003}$
Acc (%, ↑)	44.33 ± 09.40	$\mathbf{65.02 \pm 00.28}$

(d) Video Observations (x_3)

Datasets

We employ the CMU-MOSI [186] and CMU-MOSEI [3], two popular datasets for sentiment analysis and emotion recognition with challenging temporal dynamics. Both datasets consist of textual (x_1), sound (x_2) and visual (x_3) modalities extracted from videos. CMU-MOSI consists of 2199 short monologue videos clips of subjects expressing opinions about various topics. CMU-MOSEI is an extension of CMU-MOSI dataset containing 23453 YouTube video clips of subjects expressing movie reviews. In both datasets, each video clip is annotated with labels in $[-3, 3]$, where -3 and 3 indicate strong negative and strongly positive sentiment scores, respectively. We employ the temporally-aligned version of these datasets: CMU-MOSEI consists of 18134 and 4643 training and testing samples, respectively, and CMU-MOSI consists of 1513 and 686 training and testing samples, respectively.

Models

We consider the Multimodal Transformer [163] which is the state-of-the-art model for classification on the CMU-MOSI and CMU-MOSEI datasets, which we will denote throughout this section as *baseline*. For GMC, we employ the same architecture for the joint-modality encoder $f_{1:3}(\cdot)$ as the Multimodal Transformer but remove the last classification layers. For the modality-specific base encoders $\{f_1(\cdot), f_2(\cdot), f_3(\cdot)\}$, we employ a simple GRU layer with 30 hidden units and a fully-connected layer. The shared projection head $g(\cdot)$ is comprised of a single fully connected layer. We consider 60-dimensional low-level and high-level representations $\boldsymbol{h}, \boldsymbol{z} \in \mathbb{R}^{60}$. In addition, we employ a simple classifier consisting of 2 linear layers over the complete representations $\boldsymbol{z}_{1:M}$ to provide the supervision signal to the model during training.

Table 6.4: Results of different multimodal representation models in the CMU-MOSI dataset, in a classification task under complete and partial observations. Results averaged over 5 independent runs. Arrows indicate the direction of improvement.

Metric	Baseline	GMC
MAE (↓)	1.033 ± 0.037	$\mathbf{1.010 \pm 0.070}$
Cor (↑)	0.642 ± 0.008	$\mathbf{0.649 \pm 0.019}$
F1 (↑)	0.770 ± 0.017	$\mathbf{0.776 \pm 0.023}$
Acc (%, ↑)	77.07 ± 01.67	$\mathbf{77.59 \pm 02.20}$

(a) Complete Observations ($x_{1:3}$)

Metric	Baseline	GMC
MAE (↓)	1.244 ± 0.100	$\mathbf{1.119 \pm 0.033}$
Cor (↑)	0.431 ± 0.208	$\mathbf{0.573 \pm 0.016}$
F1 (↑)	0.698 ± 0.053	$\mathbf{0.727 \pm 0.013}$
Acc (%, ↑)	66.28 ± 07.74	$\mathbf{72.32 \pm 0.013}$

(b) Text Observations (x_1)

Metric	Baseline	GMC
MAE (↓)	$\mathbf{1.431 \pm 0.025}$	1.434 ± 0.017
Cor (↑)	0.056 ± 0.071	$\mathbf{0.211 \pm 0.010}$
F1 (↑)	$\mathbf{0.588 \pm 0.076}$	0.570 ± 0.006
Acc (%, ↑)	47.20 ± 05.67	$\mathbf{55.91 \pm 01.11}$

(c) Audio Observations (x_2)

Metric	Baseline	GMC
MAE (↓)	$\mathbf{1.406 \pm 0.041}$	1.452 ± 0.035
Cor (↑)	0.021 ± 0.028	$\mathbf{0.176 \pm 0.028}$
F1 (↑)	$\mathbf{0.659 \pm 0.049}$	0.550 ± 0.015
Acc (%, ↑)	53.87 ± 05.77	$\mathbf{54.30 \pm 01.96}$

(d) Video Observations (x_3)

Setup

We follow the training scheme proposed by Tsai et al. [163] and train all models for 40 epochs with an decaying learning rate of 1e−3. For GMC, we consider a temperature $\tau = 0.3$. We evaluate the performance of representation models in sentiment analysis classification with missing modality information. We evaluate literature-standard metrics [162, 163] and report binary accuracy (Acc), mean absolute error (MAE), correlation (Cor) and F1 score (F1) of the predictions obtain on the test dataset.

Accuracy results

The results obtained on CMU-MOSEI are reported in Table 6.3. When using the complete observations $x_{1:3}$ as inputs, GMC achieves competitive performance with the baseline model indicating that the additional contrastive loss does not deteriorate significantly the model's capabilities (Table 6.3a). However, GMC significantly improves the robustness of the model to missing modalities at test time, as seen in Tables 6.3b, 6.3c and 6.3d where we use only individual modalities as inputs. While GMC consistently outperforms the baseline in all metrics, we observe the largest improvement on the F1 score and binary accuracy (Acc) where the baseline often performs worse than random. The results obtained on CMU-MOSI are reported in Table 6.4. We observe that is able to GMC improve the robustness of the model to the missing modalities at test time, as seen from Tables 6.4b, 6.4c and 6.4d where we use only individual modalities as inputs. However, such improvement is not equal to all modalities, as the baseline outperforms GMC on MAE and F1 scores for the audio (x_2) and video (x_3) modalities. We hypothesise that this behaviour is due to the intrinsic difficulty of forming good contrastive pairs in small-scale datasets [20], as the CMU-MOSI dataset only contains 1513 training samples.

Table 6.5: Alignment scores A_{DCA} of the representation models in the CMU-MOSEI dataset evaluating the geometric alignment of complete representations $\boldsymbol{z}_{1:4}$ and modality-specific ones $\{\boldsymbol{z}_1, \ldots \boldsymbol{z}_4\}$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	Baseline	**GMC (Ours)**
Complete ($\boldsymbol{z}_{1:3}$)	Text ($\boldsymbol{z}_1$)	0.50 ± 0.05	$\mathbf{0.95 \pm 0.01}$
Complete ($\boldsymbol{z}_{1:3}$)	Audio ($\boldsymbol{z}_2$)	0.41 ± 0.14	$\mathbf{0.86 \pm 0.04}$
Complete ($\boldsymbol{z}_{1:3}$)	Vision ($\boldsymbol{z}_3$)	0.50 ± 0.14	$\mathbf{0.92 \pm 0.02}$

Table 6.6: Alignment scores A_{DCA} of the models in the CMU-MOSI dataset, evaluating the geometric alignment of complete representations $\boldsymbol{z}_{1:4}$ and modality-specific ones $\{\boldsymbol{z}_1, \ldots \boldsymbol{z}_4\}$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	Baseline	**GMC (Ours)**
Complete ($\boldsymbol{z}_{1:3}$)	Text ($\boldsymbol{z}_1$)	0.54 ± 0.07	$\mathbf{0.93 \pm 0.02}$
Complete ($\boldsymbol{z}_{1:3}$)	Audio ($\boldsymbol{z}_2$)	0.14 ± 0.06	$\mathbf{0.75 \pm 0.05}$
Complete ($\boldsymbol{z}_{1:3}$)	Vision ($\boldsymbol{z}_3$)	0.36 ± 0.09	$\mathbf{0.85 \pm 0.04}$

Alignment results

We additionally evaluate the geometric alignment of the modality-specific representations $\boldsymbol{z}_m$ (comprising the set E) and complete representations $\boldsymbol{z}_{1:3}$ (comprising the set R). The resulting alignment scores A_{DCA}, reported in Table 6.5, supports the results shown in Table 6.3 and verifies that GMC significantly improves the geometric alignment compared to the baseline. In Table 6.6, we observe that GMC in the CMU-MOSI dataset improves the geometric alignment of the modality-specific representations $\boldsymbol{z}_m$ (comprising the set E) and complete representations $\boldsymbol{z}_{1:3}$ (comprising the set R) compared to the baseline, despite the small size of the dataset.

6.3.4 Ablation studies

In Appendix D we perform an ablation study on the hyperparameters of GMC, using the setup from Section 6.3.2 on the MHD dataset. In particular, we investigate: a) the robustness of the GMC framework when varying the temperature parameter τ; b) the performance of GMC when varying dimensionalities d and s of the low-level and high-level latent representation spaces, respectively; and c) the performance of GMC trained with a modified loss function that uses only complete observations as negative pairs. We report both classification results and DCA scores. The results show that GMC is robust to different experimental conditions both in terms of performance and geometric alignment of the representations.

6.4 Concluding Remarks

In this chapter, we addressed the problem of learning multimodal representations that are both semantically rich for a downstream tasks, and robust to missing modality information at test time. We contributed a novel Geometric Multimodal Contrastive (GMC) learning framework that geometrically aligns complete and modality-specific representations in a shared latent space. We have shown that GMC is able to achieve state-of-the-art performance with missing modality information across a wide range of different learning problems while being straightforward to integrate with existing state-of-the-art approaches. We believe that GMC broadens the range of possible applications of contrastive learning methods to multimodal scenarios and opens many future work directions, such as investigating the effect of modality-specific augmentations or usage of inherent low-level representations for modality-specific downstream tasks.

We have shown how to learn multimodal representations from an arbitrary number of perceptual modalities (with or without supervision), thus addressing the first half of our research question: *how can we endow artificial agents with mechanisms to learn representations from multimodal observations provided by their environment.* In the next chapter, we employ multimodal representation models as perception modules in the architecture of artificial agents. By learning a robust multimodal representation we wish to provide artificial agents with the ability to perform tasks under changing conditions of perceptual availability, such as missing modality information at test time.

Chapter 7

Multimodal Transfer in Reinforcement Learning

"If you can look, see. If you can see, notice."
José Saramago, *Blindness*

We now focus our attention on multimodal representation learning for the actuation of artificial autonomous agents. As discussed in Section 3.4, low-dimensional representations have been successfully exploited in reinforcement learning scenarios, providing agents with a compact description of the current state of their environment, suitable to guide policy learning. However, by design, such works assume that the agent is only able to perceive its environment through a single channel (often vision). This choice fundamentally limits the ability of the agent to probe the environment under changing conditions of perceptual availability, compromising its actuation. Furthermore, privacy concerns regarding human data acquisition by artificial agents, such as sound and video information, further motivates the need to develop computational methods to provide agents with the ability to robustly act in conditions of *partial perceptual availability*, i.e., when one or multiple modalities are not available at execution time.

In this chapter we investigate how an agent can leverage multimodal information to act robustly in its environment. We propose to exploit multimodal information provided by the environment to allow agents to execute tasks in scenarios of partial perceptual availability, with minimal performance loss. We introduce the novel problem of *multimodal transfer in reinforcement learning*, i.e., how can an agent learn a policy considering observations from a set of modalities and zero-shot transfer that policy to scenarios where the environment provides observations from a different set of modalities. Among others, we envision scenarios where RL agents are provided the ability of learning a visual policy (i.e., policy learned over image inputs), and then (re-)using such policy at test time when only sound inputs are available.

To achieve this, we contribute a three-stage approach which effectively allows an RL agent to learn robust policies over input modalities, achieving better out-of-the-box performance when compared to different baselines. We start by learning a multimodal latent space over the different input modalities that the agent has access to, employing a multimodal generative model. In the second step, the RL agent learns a policy directly on top of this latent space, while (possibly) only having access to a subset of the input modalities. Finally, the transfer occurs in the third step, where, at execution time, the agent reuses the same policy to perform the task, while having access to a different subset of modalities.

We instantiate the multimodal transfer in reinforcement learning problem in the context of Atari games, a literature-standard environment for RL research [7]. We extend such setting and introduce the *multimodal Atari games* scenario, where the RL agent is provided with both the image and sound of the game environment, thus providing a natural mechanism to access the *cross-modality policy transfer* performance of agents in conditions of partial perceptual availability. The results show that our agents are able to perform the task under such difficult settings. This is the case even when using different multimodal generative models and reinforcement learning algorithms [92, 110].The main contributions of this chapter are three-fold:

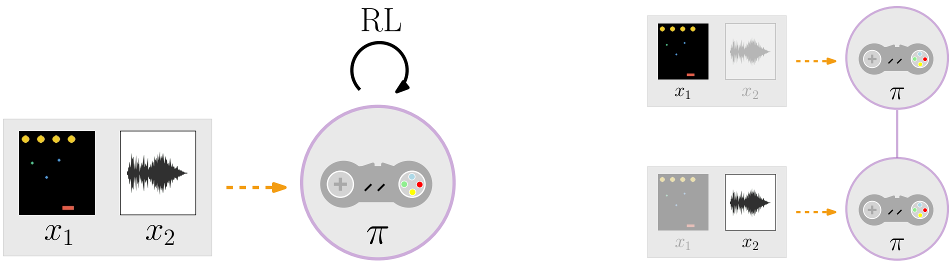
- In Section 7.1, we introduce the problem of *multimodal transfer in reinforcement learning* and propose a three stage approach to allow agents to execute tasks in conditions of partial perceptual availability, i.e., with missing modality information, at execution time, with minimal performance loss;

- In Section 7.2, we propose the *multimodal Atari games* scenario, a natural multimodal extension of Atari games for RL agents, in which the agents are provided with both the image and sounds of the game environment. We instantiate two scenarios of increasing complexity: a modified version of the Pendulum environment from OpenAI gym, with a simple sound source, and the novel HYPERHOT environment, a fast-paced *space invaders*-like game that assesses the performance of our approach in scenarios with more complex and realistic generation of sounds;

- In Section 7.3, we evaluate our approach in the *multimodal Atari games* scenario and show that our agent is able perform the task under conditions of changing perceptual availability, without requiring further training or fine-tuning of the previously learned policy to the available modalities at test-time.

The work described in this chapter has been published in:

- Rui Silva, **Miguel Vasco**, Francisco S. Melo, Ana Paiva, and Manuela Veloso. *Playing Games in the Dark: An Approach for Cross-Modality Transfer in Reinforcement Learning.* Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). 2020, pp. 1260–1268 [143].



(a) Training

(b) Execution

Figure 7.1: *Multimodal Transfer in Reinforcement Learning*: (a) a policy trained over a given subset of the agent's perceptual modalities (e.g., image and sound) must be able to (b) be transferred to every possible subset of input modalities (e.g., only image or sound), for the agent to perform the task under conditions of partial perceptual availability.

7.1 Multimodal Transfer in Reinforcement Learning

We consider an agent facing a sequential decision problem which can be described as a MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P, r, \gamma)$, as defined in Section 2.2. Slightly abusing our notation, the agent is endowed with a set $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$ of M different input modalities, which can be used to perceive the world and build a (possibly) partial observation of the current state $x \in \mathcal{X}$. Different modalities may provide more or less perceptual information than others. Some modalities may be redundant (i.e., provide the same perceptual information) or complement each other (i.e., jointly provide more information).

Our goal is for the agent to learn a policy while observing a subset of input modalities $\mathcal{X}_{\text{train}}$, and then use that same policy when observing a possibly different subset of modalities, $\mathcal{X}_{\text{test}}$, with as minimal performance degradation as possible, as depicted in Fig. 7.1. Inspired by recent work on representation learning for reinforcement learning agents [65], we propose to allow *cross-modality policy transfer* by following a three-stage pipeline:

1. *Learn a perceptual model of the world*

   Learn a multimodal latent space over the complete set of input modalities X. In particular, the agent learns either a *representation-downstream* model $\{r, g\}$, allowing for cross-modality inference, through unsupervised learning, or a representation model $\{r\}$, through self-supervised learning, that allows the encoding of multimodal data;

2. *Learn to act in the world*

   Reinforcement learning step to learn an optimal policy for the task described by MDP $\mathcal{M}$. Specifically, the agent is assumed to have access to a subset of input modalities $\mathcal{X}_{\text{train}}$, and the policy is trained over the latent space conditioned on such modalities subset $r(\mathcal{X}_{\text{train}})$;

3. *Policy transfer*

   The policy transfer occurs at execution time, when the agent may have access to a different subset of input modalities $\mathcal{X}_{\text{test}}$. We analyze the zero-shot transfer performance of this policy, i.e., we evaluate the performance of the agent on the new set of modalities without any fine-tuning or further training.

We now discuss each step in further detail.

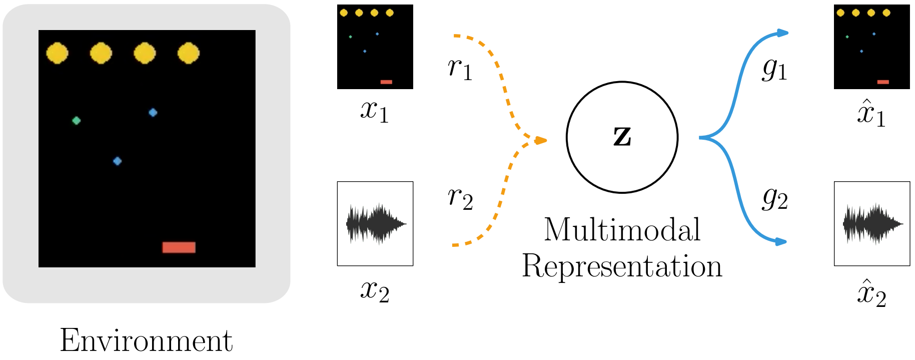


Figure 7.2: *Learning a perceptual model of the world*: Each time step of a game includes visual and sound components that are intrinsically coupled. This coupling can be encoded in a multimodal latent representation using the representation maps $r = \{r_1, r_2\}$, such that $r : \mathcal{X} \to \mathcal{Z}$. For multimodal generative models, such as MUSE, the latent representation can be used to infer the sound associated with a given image (or vice-versa), using the downstream maps $g = \{r_1, r_2\}$, such that $g : \mathcal{Z} \to \mathcal{X}$. For multimodal contrastive approaches, such as GMC, no such downstream map is learnt.

7.1.1 Learn a perceptual model of the world

Let $\mathcal{X}$ denote the Cartesian product of input modalities, $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M$. Intuitively, we can think of $\mathcal{X}$ as the complete perceptual space of the agent. Fig. 7.2 depicts an example on a game, where the agent can have access to two modalities, $\mathcal{X}_{\text{image}}$ and $\mathcal{X}_{\text{sound}}$, corresponding to visual and sound information. We write $\boldsymbol{x}$ to denote an element of $\mathcal{X}$. At each moment t, the agent may not have access to the complete perception $\boldsymbol{x}(t) \in \mathcal{X}$, but only to a partial view thereof. We are interested in learning a multimodal representation of the perceptions in $\mathcal{X}$. Revisiting the notation of Section 2.1, to encode such representation we resort to a set of representation maps $r = \{r_1, \ldots, r_L\}$, where each map r_ℓ takes the form $r_\ell : \text{proj}_\ell \to \mathcal{Z}$, where $\mathcal{Z}$ is a common multimodal latent space and proj_ℓ projects $\mathcal{X}$ to some subspace of K modalities, $\mathcal{X}_\ell = \mathcal{X}_{\ell_1} \times \mathcal{X}_{\ell_2} \times \ldots \times \mathcal{X}_{\ell_K}$. In Fig. 7.2 the set of representation maps r is used to compute a latent representation $\boldsymbol{z}$ from sound and image data.

To learn such mappings, we start by collecting a dataset of N examples of coupled sensory information:

$$\mathcal{D}(\mathcal{X}) = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\right\}.$$

We then follow an unsupervised learning approach, and train a multimodal representation model on dataset $\mathcal{D}(X)$ to learn the latent space over the agent's input modalities. For example, in the case of multimodal VAE models, such as MUSE, the representation maps in r correspond to the *encoder* networks while the *decoder* networks can be seen as a set of inverse latent mappings, $g = \{g_1, \ldots, g_L\}$ that allow for modality reconstruction and cross-modality generation. The collection of the initial data needed to generate $\mathcal{D}(\mathcal{X})$ can be easier or harder depending on the complexity of the task. In Section 7.3 we briefly discuss mechanisms to perform the data collection and their limitations.

7.1.2 Learn to act in the world

After learning a perceptual model of the world, the agent then learns how to perform the task. We follow a reinforcement learning approach to learn an optimal policy for the task

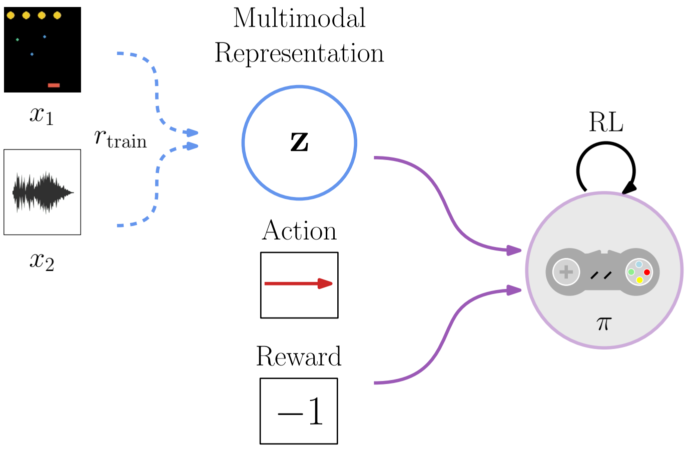


Figure 7.3: *Learn to act in the world*: To learn how to perform the task, we assume the agent has access to a subset of input modalities $\mathcal{X}_{\text{train}}$ (e.g. image and sound). We encode the available perceptions, using the frozen latent maps r_{train}, to generate the multimodal latent state $\boldsymbol{z}$. We employ the latent state, along with the recorded action and reward information, to learn a policy π that maps game states to actions. Note that the framework is agnostic to the type of RL algorithm employed to learn such mapping.

described by MDP $\mathcal{M}$. During this learning phase, we assume the agent may have access only to a subset of input modalities $\mathcal{X}_{\text{train}}$. As a result, during its interaction with the environment, the agent collects a sequence of triplets,

$$\left\{\left(\boldsymbol{x}_{\text{train}}^{(0)}, a^{(0)}, r^{(0)}\right), \left(\boldsymbol{x}_{\text{train}}^{(1)}, a^{(1)}, r^{(1)}\right), \dots\right\},$$

where $\boldsymbol{x}_{\text{train}}^{(t)}$, $a^{(t)}$, $r^{(t)}$ correspond to the perceptual observations, action executed, and rewards obtained at timestep t, respectively. However, as depicted in Fig. 7.3 our reinforcement learning agent does not use this sequence of triplets directly. Instead, it pre-processes the perceptual observations using the previously learned latent maps r to encode the multimodal latent state at each time step as $\boldsymbol{z}^{(t)} = r_{\text{train}}(\boldsymbol{x}_{\text{train}}^{(t)})$, where $r_{\text{train}} \in r$ maps $\mathcal{X}_{\text{train}} \to \mathcal{Z}$. In practice, the RL agent uses a sequence of triplets,

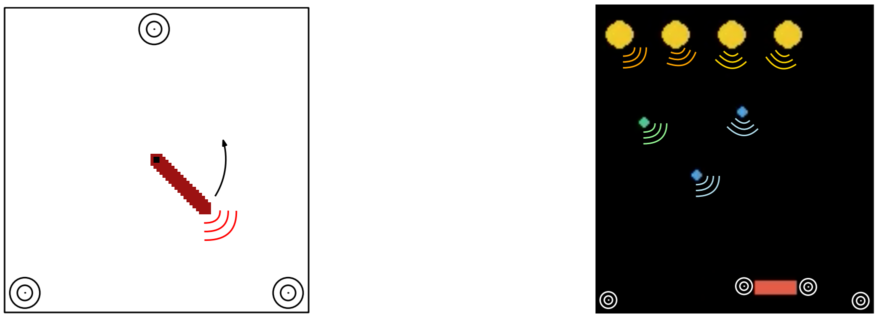
$$\left\{\left(\boldsymbol{z}^{(0)}, a^{(0)}, r^{(0)}\right), \left(\boldsymbol{z}^{(1)}, a^{(1)}, r^{(1)}\right), \dots\right\}$$

to learn a policy $\pi : \mathcal{Z} \to \mathcal{A}$, that maps the latent states to actions. Any continuous-state space reinforcement learning algorithm can be used to learn this policy π over the latent states. These latent states are encoded using the generative model trained in the previous section and, as such, the weights of this model are frozen during the RL training.

7.1.3 Policy transfer

Policy transfer occurs once the agent has learned how to perceive and act in the world. At this time, we assume the agent may now access a subset of input modalities $\mathcal{X}_{\text{test}}$, potentially different from $\mathcal{X}_{\text{train}}$, i.e., the set of modalities it used in learning the task policy π. As a result, during its interaction with the environment, at each time step t, the agent will now observe perceptual information $\boldsymbol{x}_{\text{test}}^{(t)}$.

In order to reuse the policy π, the agent starts by pre-processing this perceptual observation, again using the set of maps r previously trained, but now generating a latent



(a) (b)

Figure 7.4: The *multimodal Atari games* employed in this work, depicting the image and sound perceptual information provided to the agent, along with the sound receivers (concentric circles) of such information: (a) the Pendulum environment, a classic control problem, where the tip of the pendulum emits a frequency that is received by three microphones placed at the bottom left and right (bl, br) and middle top (mt); (b) the Hyperhot environment, a fast-paced Space-Invaders-like game where all enemies and bullets emit sounds that are received by four microphones at bottom left and right (bl, br) and paddle left and right (pl, pr).

state $\boldsymbol{z}^{(t)} = r_{\text{test}}(\boldsymbol{x}^{(t)}_{\text{test}})$ where $r_{\text{test}} \in r$ now maps $\mathcal{X}_{\text{test}}$ into $\mathcal{Z}$. Since policy π maps the latent space $\mathcal{Z}$ to the action space $\mathcal{A}$, it can now be used directly to select the optimal action at the new state $\boldsymbol{z}^{(t)}$. Effectively, the agent is reusing a policy π that was learned over a (possibly) different set of input modalities, with no additional training, i.e., zero-shot multimodal policy transfer.

7.2 Multimodal Atari Games

To evaluate the performance of our proposed cross-modality policy transfer approach, we consider the Atari games scenario, widely employed for the evaluation of deep reinforcement learning agents [38, 109, 140]. However, this scenario only provides visual information (image frames) of the environment to the agent. We propose the *multimodal Atari games* scenario, an extension to Atari games that allows the agent to receive information from the environment through additional modality channels. We instantiate such novel scenario in two environments of increasing complexity, not only in regards to the task but also in regards to the input modalities, depicted in Fig. 7.4. We start by considering a modified version of the Pendulum environment from OpenAI gym, with a simple sound source. Then, we consider hyperhot, a "Space Invaders"-like game that assesses the performance of our approach in scenarios with more complex and realistic generation of sounds.

7.2.1 Pendulum Environment

We consider a modified version of the Pendulum environment from OpenAI gym — a classic control problem, where the goal is to swing the pendulum up so it stays upright. We modify this environment so that the observations include both an image and a sound component. For the sound component, we assume that the tip of the pendulum emits a constant frequency f_0, which is received by a set of S sound receivers $\{\rho_1, \ldots, \rho_S\}$. Figure 7.4a depicts this scenario, where the pendulum and its sound are in red, and the sound receivers

(a) (b)

Figure 7.5: Different sound properties in the multimodal pendulum scenario. (a) Depicts the Doppler effect: as the sound source moves near (away from) the observer, the arrival time of the emitted waves decreases (increases), thus increasing (decreasing) the frequency. (b) Depicts how the amplitude of the sound decreases with the distance from the source. Fading semi-circles denote smaller intensities.

correspond to the concentric circles.

Formally, we let $\mathcal{X} = \mathcal{X}_{\text{image}} \times \mathcal{X}_{\text{sound}}$ denote the complete perceptual space of the agent. The visual input modality of the agent, $\mathcal{X}_{\text{image}}$, consists of the raw pixel-image observation of the environment. The sound input modality, $\mathcal{X}_{\text{sound}}$, consists of the frequency and amplitude received by each of the S microphones of the agent. Both image and sound observations may be stacked to account for the dynamics of the environment.

In this scenario, we assume a simple model for sound generation. Specifically, we assume that, at each timestep, the frequency f'_i heard by each sound receiver ρ_i follows the Doppler effect. The Doppler effect measures the change in frequency heard by an observer as it moves towards or away from the source. Slightly abusing our notation, we let $\boldsymbol{\rho}_i$ denote the position of sound receiver ρ_i and $\boldsymbol{e}$ the position of the sound emitter. Formally,

$$f'_i = \left(\frac{c + \dot{\boldsymbol{\rho}}_i \cdot (\boldsymbol{e} - \boldsymbol{\rho}_i)}{c - \dot{\boldsymbol{e}} \cdot (\boldsymbol{\rho}_i - \boldsymbol{e})} \right) f_0,$$

where c is the speed of sound and we use the dot notation to represent velocities. Figure 7.5a depicts the Doppler effect in the pendulum scenario. We then let the amplitude b_i heard by receiver ρ_i follow the inverse square law,

$$b_i = \frac{K}{\|\boldsymbol{e} - \boldsymbol{\rho}_i\|^2},$$

where K is a scaling constant. Figure 7.5b depicts the inverse square law applied to the pendulum scenario, showing how the amplitude of the sound generated decreases with the distance to the source.

7.2.2 Hyperhot environment

We consider the hyperhot scenario, a novel top-down shooter game scenario inspired by the "Space Invaders" Atari game[1], where the goal of the agent is to shoot the enemies above while avoiding their bullets, by moving left and right. Similarly to the Pendulum environment, in this scenario, the observations of the environment include both image and sound components. In hyperhot, however, the environmental sound is generated by multiple entities e_i emitting a predefined frequency $f_0^{(i)}$:

[1] We opted to use a custom environment implemented in pygame, since the *space invaders* environment in OpenAI gym does not provide access to game state, making it hard to generate simulated sounds.

- Left-side enemy units, e_0, and right-side enemy units, e_1, emit sounds with frequencies $f_0^{(0)}$ and $f_0^{(1)}$, respectively;
- Enemy bullets. e_2, emit sounds with frequency $f_0^{(2)}$;
- The agent's bullets, e_3, emit sounds with frequency $f_0^{(3)}$.

The sounds produced by these entities are received by a set of S sound receivers $\{\rho_1, \ldots, \rho_S\}$. Figure 7.4b depicts the scenario, where the yellow circles are the enemies; the green and blue bullets are friendly and enemy fire, respectively; the agent is in red; and the sound receivers correspond to the white circles. The agent is rewarded for shooting the enemies, with the following reward function:

$$r = \begin{cases} 10 & \text{if all enemies are killed, i.e., win;} \\ -1 & \text{if player is killed or time is up, i.e., lose;} \\ 0 & \text{otherwise.} \end{cases}$$

The environment resets whenever the agent collects a non-zero reward, be it due to winning or losing the game.

We let the perceptual space of the agent be as $\mathcal{X} = \mathcal{X}_{\text{image}} \times \mathcal{X}_{\text{sound}}$, with the visual input modality of the agent, $\mathcal{X}_{\text{vision}}$, consisting in the raw pixel-image observation of the environment. The sound, however, is generated in a more complex and realistic way. We model the sinusoidal wave of each sound-emitter e_i considering its specific frequency $f_0^{(i)}$ and amplitude $b_0^{(i)}$. At every frame, we take the sound waves of every emitter present in the screen, considering their distance to each sound receiver in S. The sound wave generated by emitter e_i is observed by receiver ρ_j as,

$$b^{(i)} = b_0^{(i)} \exp\left(-\delta \|\boldsymbol{e}_i - \boldsymbol{\rho}_j\|^2\right),$$

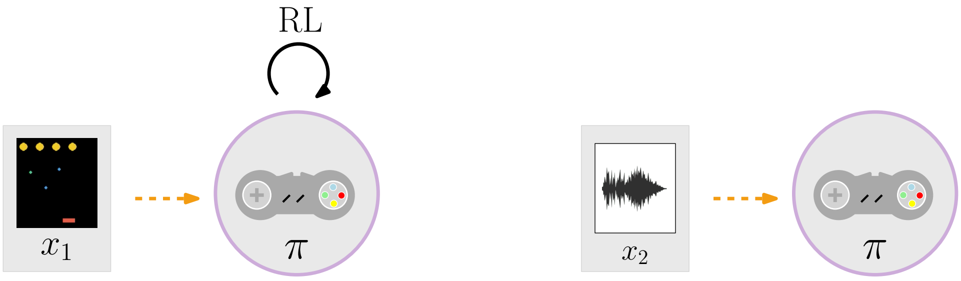
where δ is a scaling constant, $\boldsymbol{e}_i$ and $\boldsymbol{\rho}_j$ denote the positions of sound emitter e_i and sound receiver ρ_j, respectively. We generate each sinusoidal sound wave for a total of 1047 discrete time steps, considering an audio sample rate of 31 400 Hz and a video frame-rate of 30 fps. As such, each sinusoidal sound wave represents the sound heard for the duration of a single video-frame of the game (similarly to what is performed in Atari videogames). Finally, for each sound receiver, we sum all emitted waves and encode the amplitude values in 16-bit audio depth, considering a maximum amplitude value of a_M and a minimum value of $-a_M$.

7.3 Experimental Evaluation

We evaluate our approach for policy transfer across different sets of input modalities in multimodal Atari games, by answering the following questions:

(i) What is the performance our approach for *Cross-Modality Policy Transfer* (Fig. 7.6), i.e., settings where the agent employs the policy over observations of modalities that were not available during training, $X_{\text{train}} \cap X_{\text{test}} = \emptyset$;

(ii) What is the performance our approach for *Multimodal Policy Transfer* (Fig. 7.1), i.e., settings where the agent learns a policy over observations of the complete set of modalities and employs the same policy over observations from all possible subsets of modalities, $X_{\text{train}} = X,\ X_{\text{train}} \cap X_{\text{test}} \neq \emptyset$.



(a) Training

(b) Execution

Figure 7.6: *Cross-Modality Policy Transfer*: in this scenario the agent (a) learns a policy from image information and (b) reuses that very same policy at test-time when the environment only provides sound information, without further training or fine-tuning.

We evaluate the two different variants of policy transfer with specific goals: with (i), we aim at understanding the feasibility of executing policies with information provided by modalities unobserved during policy training, comparing the performance of our approach against non-transferable policies. With (ii), we aim at understanding the role of different perceptual models for acting under partial perceptual availability, comparing the performance of agents endowed with different state-of-the-art multimodal representation models across all possible perceptual conditions at execution time. As such, for the former evaluation, we employ an Associative VAE model [184] to learn a perceptual model of the world, while for the latter evaluation, we consider both GMC and the MUSE model, with ALMA encoder, due to their ability to encode a joint representation of multiple modalities. All training hyper-parameters are in Appendix F.

7.3.1 Pendulum

We start by evaluating the reinforcement learning agents in the multimodal pendulum scenario, for cross-modality and multimodal policy transfer. For this task, all perceptual models were trained over a dataset $\mathcal{D}(X)$ with N observations of images and sounds $\boldsymbol{x}^n = \left(\boldsymbol{x}^n_{\text{image}}, \boldsymbol{x}^n_{\text{sound}}\right)$, collected using a random controller. The random controller proved to be enough to cover the state space. Before training, the images were preprocessed to black and white and resized to 60×60 pixels. The sounds were normalized to the range $[0, 1]$, assuming the minimum and maximum values found in the M samples.

The agents learned how to perform the task using the DDPG algorithm. For the cross-modality policy transfer scenario, the environment only provided access to the image input modality, i.e., $X_{\text{train}} = X_{\text{image}}$. For the multimodal policy transfer scenario, the environment provided both image and sound input modalities, i.e., $X_{\text{train}} = X_{\text{image}} \times X_{\text{sound}}$. The actor and critic networks consisted of two fully connected layers of 256 neurons each. The replay buffer was initially filled with samples obtained using a controller based on the Ornstein-Uhlenbeck process, with the parameters suggested by Lillicrap et al. [92].

Cross-modality policy transfer

In this setting, we evaluated the performance of the policy trained when the agent only has access to the sound input modality, i.e., $X_{\text{test}} = X_{\text{sound}}$. Our approach is compared with two baselines:

Table 7.1: Average reward per episode, computed over 75 episodes, of different approaches for cross-modality policy transfer in the pendulum scenario. Results averaged over 10 randomly seeded runs. Higher is better.

Approach	**Training**	**Evaluation**	**Rewards** (↑)
Transfer Policy	Image	Sound	-2.00 ± 0.97
Sound Policy	Sound	Sound	-1.41 ± 0.91
Random Policy	Image	Image	-6.30 ± 0.29

Table 7.2: Average reward per episode, computed over 75 episodes, of agents with different perceptual models for multimodal policy transfer in the pendulum scenario, when provided with different input modalities during execution. Results averaged over 10 randomly seeded runs. Higher is better.

Input	**MVAE**	**MUSE**	**GMC**
Image, Sound	-1.17 ± 0.18	-0.94 ± 0.06	-0.94 ± 0.06
Image	-1.07 ± 0.17	-2.64 ± 0.82	-0.94 ± 0.06
Sound	-6.62 ± 0.24	-3.94 ± 0.68	$\mathbf{-0.96 \pm 0.08}$

- RANDOM policy baseline, which depicts the performance of an untrained agent. This effectively simulates the performance one would expect from a non-transferable policy trained over image inputs, and later tested over sound inputs;

- SOUND policy baseline, a DDPG agent trained directly over sound inputs (*i.e.* the sounds correspond to the states). Provides an estimate on the performance an agent trained directly over the test input modality may achieve.

Table 7.1 summarizes the transfer performance in terms of average reward observations throughout an episode of 300 frames. The results show that our approach provides the agent with an out-of-box performance improvement of over 300%, when compared to the untrained agent (non-transferable policy). It is also interesting to observe that the difference in performance between our agent and SOUND DDPG seems small, supporting our empirical observation that the transferred policy succeeds very often in the task: swinging the pole up[2].

Multimodal policy transfer

In this scenario, we evaluated the performance of the policy trained across multiple conditions of perceptual availability, i.e., when the agent has access to the image, sound or both input modalities. We evaluate the performance of MUSE and GMC as perceptual models against the MVAE model [179]. We use the same network architectures and training hyperparameters of the previous evaluation for all models, presented in Appendix F.

The results in Table 7.2 show that both MUSE and GMC outperform MVAE when executing the task with missing modality information. While the MVAE agent is unable

[2] We also note that the performance achieved by the SOUND DDPG agent is similar to that reported in the OpenAI gym leaderboard for the Pendulum scenario with state observations as the position and velocity of the pendulum.

Table 7.3: Average discounted reward per episode and win rate, computed over 100 episodes, of different approaches for cross-modality policy transfer in the HYPERHOT scenario. Results averaged over 10 randomly seeded runs. Higher is better.

Method	**Training**	**Evaluation**	**Reward** (↑)	**Win** (%, ↑)
Transfer Policy	Image	Sound	0.15 ± 0.16	36.1 ± 10.4
Image Policy	Image	Image	1.54 ± 0.20	75.0 ± 05.3
Sound Policy	Sound	Sound	0.10 ± 0.22	27.3 ± 21.4
Random Policy	Image	Image	-0.33 ± 0.16	08.3 ± 05.8

to act when provided only with sound inputs, both MUSE and GMC agents are better able to act in the same conditions. However, only GMC allows the agent to act under all conditions of perceptual availability, with no significant performance loss.

7.3.2 HYPERHOT

We now evaluate the reinforcement learning agents in the HYPERHOT scenario, for both cross-modality and multimodal policy transfer. To understand the role of the collected dataset to train the perceptual models, we collected a dataset $\mathcal{D}(X)$ of observations of images and sounds collected using a random controller. Before training, the images were preprocessed to black and white and resized to 80×80 pixels, and the sounds normalized to the range $[0, 1]$.

The agents learned how to play the game using the DQN algorithm. To learn the policy in the cross-modality policy transfer scenario, the environment only provided image observations, $X_{\text{train}} = X_{\text{image}}$, corresponding to the video game pixel-frames. For the multimodal policy transfer scenario, the agents are provided with both image and sound modalities to learn the policy, $X_{\text{train}} = X_{\text{image}} \times X_{\text{sound}}$. The policy and target networks consisted of two fully connected layers of 512 neurons each. We adopted a decaying ϵ-greedy policy.

Cross-modality policy transfer

In this setting, we evaluated the performance of the policy trained when the agent only has access to the sound input modality, i.e., $X_{\text{test}} = X_{\text{sound}}$. Our approach is compared with three baselines:

- IMAGE policy, a DQN agent trained directly over visual inputs;

- SOUND policy, an DQN agent trained directly over sound inputs;

- RANDOM policy, depicting tthe performance of an untrained agent.

Table 7.3 summarizes the transfer performance of the policy produced by our approaches, in terms of average discounted rewards and game win rates over 100 episodes.

Considering the results in Table 7.3, we observe:

- A considerable performance improvement of our approach over the untrained agent. The average discounted reward of the RANDOM baseline is negative, meaning this agent tends to get shot often, and rather quickly. This is in contrast with the

Table 7.4: Average discounted reward per episode (a) and win rate (b), computed over 100 episodes, of agents with different perceptual models for multimodal policy transfer in the Hyperhot scenario, when provided with different input modalities during execution. All perceptual models are trained with a dataset collected with a random controller. Results averaged over 10 randomly seeded runs. Higher is better.

(a) Reward (↑)

Input	**MVAE**	**MUSE**	**GMC**	**GMC-Recon**
Image, Sound	0.60 ± 0.13	0.52 ± 0.15	0.05 ± 0.14	0.86 ± 0.12
Image	0.47 ± 0.15	0.28 ± 0.17	0.03 ± 0.10	0.65 ± 0.16
Sound	-0.50 ± 0.14	0.36 ± 0.17	0.02 ± 0.12	0.40 ± 0.12

(b) Win-Rate (%, ↑)

Input	**MVAE**	**MUSE**	**GMC**	**GMC-Recon**
Image, Sound	48.3 ± 3.8	53.7 ± 5.6	26.4 ± 7.2	63.4 ± 5.1
Image	41.4 ± 5.3	39.8 ± 7.2	26.1 ± 4.6	57.2 ± 6.9
Sound	4.4 ± 4.3	43.2 ± 8.4	24.3 ± 5.4	45.9 ± 6.5

positive rewards achieved by our approaches. Moreover, the win rates achieved by our approaches surpass those of the untrained agent by 5-fold.

- A performance comparable to that of the agent trained directly on the sound, Sound policy. In fact, the average discounted rewards achieved by our agent are slightly higher. However, the Sound agent followed the same DQN architecture and number of training steps used in our approach. It is plausible that with further parameter tuning, the Sound agent could achieve better performances.

- The approach that could fine-tune to the most informative perceptual modality, Image policy, achieved the highest performances. While achieving lower performances, our approach is the only able to perform cross-modality policy transfer, that is, being able to reuse a policy trained on a different modality. One may argue that this trade-off is worthwhile.

Multimodal policy transfer

In this setting, we evaluate the performance of the policy executed across multiple conditions of perceptual availability, i.e., when the agent has access to image, sound, or both modalities. Once again, we evaluate the performance of MUSE and GMC as perceptual models against the MVAE model [179]. Additionally, we introduce a reconstruction-based GMC model (GMC-Recon), similar to Section 6.3.3, where, in addition to original the contrastive loss, we force the model to reconstruct the observations of the agent from the joint-modality latent representation.

In Table 7.4 we present the discounted rewards, averaged over 100 episodes, collected by agents endowed with the different perceptual models across all settings of perceptual availability at execution time. The results in both game scenarios show that both MUSE and GMC-Recon outperform the baseline MVAE model in being able to act when only provided with sound inputs, evidence of the *overconfident experts* issue of the MVAE model

Table 7.5: Average discounted reward per episode (a) and win rate (b), computed over 100 episodes, of agents with different perceptual models for multimodal policy transfer in the Hyperhot scenario, when provided with different input modalities during execution. All perceptual models are trained with a dataset collected with a pretrained controller. Results averaged over 10 randomly seeded runs. Higher is better.

(a) Reward (↑)

Input	**MVAE**	**MUSE**	**GMC**	**GMC-Recon**
Image, Sound	0.75 ± 0.26	0.46 ± 0.10	-0.08 ± 0.13	1.02 ± 0.21
Image	0.66 ± 0.20	0.24 ± 0.18	-0.12 ± 0.05	0.98 ± 0.22
Sound	-0.50 ± 0.11	0.01 ± 0.23	-0.02 ± 0.14	0.48 ± 0.16

(b) Win-Rate (%, ↑)

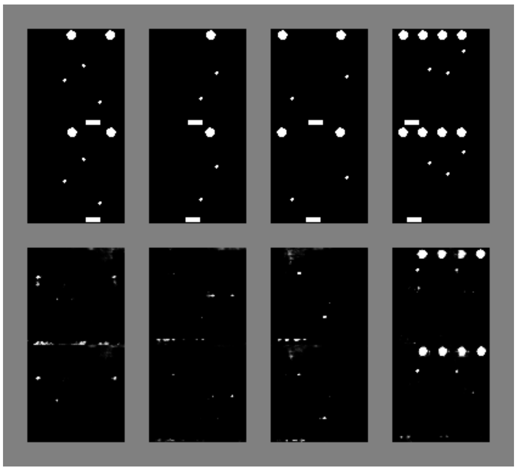
Input	**MVAE**	**MUSE**	**GMC**	**GMC-Recon**
Image, Sound	48.4 ± 6.4	51.3 ± 3.4	18.1 ± 4.9	59.1 ± 7.1
Image	45.2 ± 5.8	37.3 ± 9.3	17.6 ± 3.1	61.5 ± 6.8
Sound	4.3 ± 3.7	30.8 ± 11.3	21.8 ± 5.5	43.4 ± 7.0

discussed in Section 5. Moreover, we note that the default version of GMC is unable to encode a suitable representation for downstream policy learning, evidenced by the low performance of the method even when provided with complete observations at execution time. We plan of exploring the limits of self-supervised representation learning approaches for RL tasks in future work.

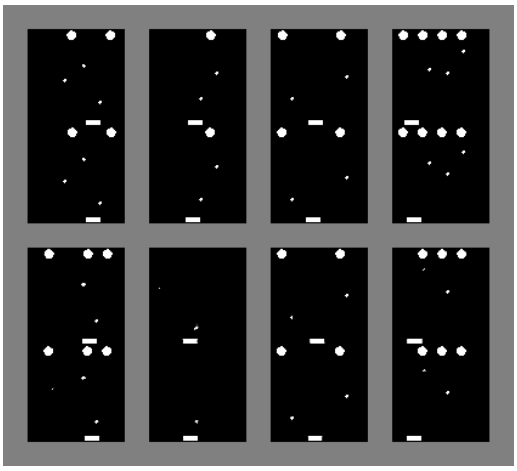
To understand the role of data collection for training the perceptual model, we collected an additional dataset $\mathcal{D}^*(X)$ of observations of images and sounds using a pretrained controller (DQN). We repeat the previous experiment and present the results in Table 7.5. The results show that the performance of the agents is not significantly affected by the policy used to collect the data for training the representation models: GMC-Recon and MUSE are still the only models suitable to learn a policy to be transferred across multiple settings of perceptual availability, with minimal performance loss.

To better understand the relationship between the performance of the different agents and the quality of the multimodal representation learnt by their perceptual models, we present in Fig. 7.7 samples of images generated from sound inputs in the Hyperhot scenario. The results suggest a relation between the performance of the agent when executing a task with different available modalities and the performance of the representation models in generating missing information (e.g. images) from available inputs (e.g. sounds). The MVAE baseline agent fails at generating coherent image information from sound input, supporting its inability to act when provided only with sound observations highlighted in Table 7.4. On the other hand, both the MUSE and GMC agents are able to generate coherent image information from sound observations, supporting their ability to act when provided only with sound observations, as shown in Table 7.4. However, both models still aren't able to perfectly reconstruct image information only from sound observations, in regards to the number of enemies and projectiles, supporting the overall lower performance of the agents when provided only with sound information, in comparison with performance when provided with the complete set of modalities.

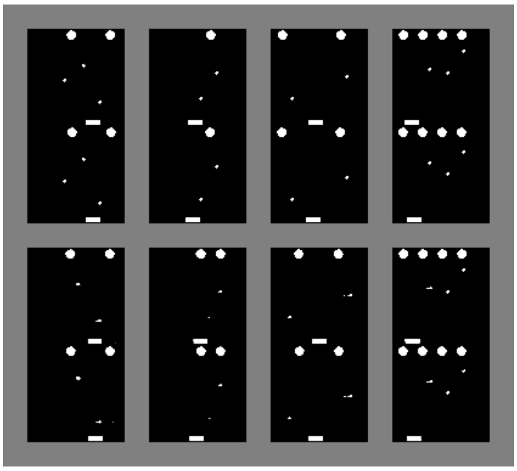
Finally, we highlight that the GMC-Recon agent (both in Table 7.4 and Table 7.5), which trained its policy over both image and sound inputs, consistently outperforms the "Sound Policy" agent of Table 7.3, which trained its policy over sound inputs, at performing the task when *only sound is available* at execution time. By providing multimodal information as input to the policy (indirectly through the latent representation), the agent is able to explore the complete state space of the environment better than when the policy is provided with single modality information. This effect is particularly relevant in scenarios such as Hyperhot in which, contrary to the multimodal pendulum, sound is not as suitable as the image for decision-making: for example, the frequency emitted by each enemy on a given side of the screen is not unique which hinders the identification of the number of enemies present on each side, a phenomena we can observe in the CMI reconstructions of Figure 7.7. This result highlights one of the perks of considering multimodal perception when designing reinforcement learning agents.



(a) MVAE



(b) MUSE



(c) GMC-Recon

Figure 7.7: Image generation from sound observations in the Hyperhot scenario: (top) real image observation from the game; (bottom) generated image from sound observation provided by the game (best viewed with zoom).

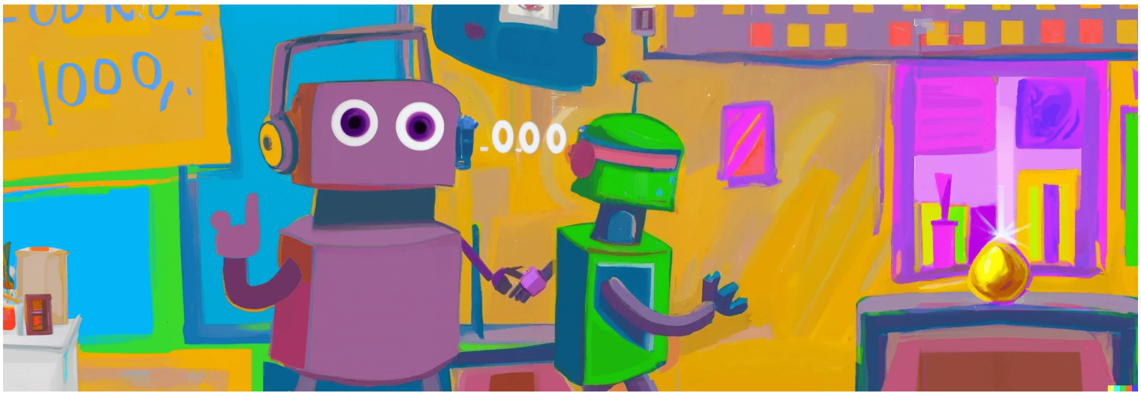
7.4 Concluding Remarks

In this chapter we explored the use of multimodal latent representations to capture information provided by different sensory modalities, in order to allow agents to learn and reuse policies over different conditions of perceptual availability. To this end, we formalized the *multimodal transfer in reinforcement learning* problem, and contributed a three-stage approach that effectively allows RL agents to learn policies that can be transfered across different sets of available input modalities. The results show that our proposed approach effectively enables an agent to learn and exploit policies over different subsets of input modalities, regardless of whether the modalities were available at policy training time. This sets our work apart from existing ideas in the literature. For example, DARLA follows a similar three-stages architecture to allow RL agents to learn policies that are robust to some shifts in the original domains [65]. However, that approach implicitly assumes that the source and target domains are characterized by similar inputs, such as raw observations of a camera. Our work allows agents to transfer policies across different input modalities. We assess the applicability and efficacy of our transfer approach in different domains of increasing complexity. We extended well-known scenarios in the reinforcement learning literature to include both image and sound observations. The results show that the policies learned by our approach were robust to these different input modalities.

In the next chapter, we extend the notion of acting with partial perceptual availability to multi-agent systems. We consider scenarios where multiple agents perform partial-observable cooperative tasks and are able to passively share their local observations. In particular, we focus on the execution of tasks across all possible levels of communication, ranging from settings in which no communication is allowed up to full communication between the agents.

Chapter 8

Multi-Agent Reinforcement Learning with Hybrid Execution



"I get by with a little help from my friends."
Ringo Star*, The Beatles*

Throughout this thesis, we have considered scenarios in which agents perceive and act upon their environment under partial perceptual availability. In Section 7, we have shown how to robustly perform single-agent reinforcement learning tasks under different levels of perceptual availability, in regards with different sets of available *modality-specific* observations. In this section, we naturally extend the previous scenario to consider multi-agent systems that, similarly to the single-agent case, execute tasks under different levels of perceptual availability, in regards to different sets of shared *agent-specific* observations.

Recently, deep multi-agent reinforcement learning (MARL) has been successfully applied to tasks such as game-playing [124], traffic light control [177], and energy management [40]. Despite these successes, the multi-agent setting remains significantly more challenging than the single-agent counterpart, in part due to the the exponential growth in the state/action space, and to environmental constraints, both in perception and actuation [19]. As a way to deal with these issues, existing methods aim to learn decentralized policies that allow the agents to act based on local perceptions and partial information about the intentions of other agents. *Centralized training with decentralized execution* methods take advantage of the fact that additional information, available only at training time, can be used to learn decentralized policies that alleviate the need for communication [118, 132, 47].

However, the assumption that agents cannot communicate at execution time is often too restrictive for a great number of real-world application domains, such as robotics and autonomous driving [67, 185]. In this chapter our objective is to develop agents that are able to exploit the benefits of centralized training while, simultaneously, being able to take advantage of passively-shared information at execution time. We introduce the paradigm of *hybrid* execution, in which agents act in scenarios with any possible communication level, ranging from no communication (fully decentralized) to full communication between the agents (fully centralized). In particular, we focus on multi-agent cooperative tasks in which the sharing of local information (observations and actions of the agents) is *critical* to their successful execution. To formalize our setting, we start by defining *hybrid partially observable Markov decision process* (H-POMDP), a new class of multi-agent POMDPs that explicitly considers a communication process between the agents. Our goal is to find a method that allows an agent to solve H-POMDPs regardless of the communication process the agent encounters at execution time. To allow for hybrid execution, we propose to employ an autoregressive model that explicitly predicts non-shared information from past observations. In addition, we propose a training scheme that introduces missing communication during training. We denote our coupled approach by *multi-agent observation sharing with communication dropout* (MARO).

We evaluate the performance of MARO across different communication levels, in different MARL benchmark environments and using multiple RL algorithms. Furthermore, we introduce three novel MARL environments that explicitly require communication during execution to successfully perform cooperative tasks, currently missing in literature. Finally, we perform an ablation study that highlights the importance of both the predictive model and the training scheme to the overall performance of MARO. The results show that our method consistently outperforms the baselines, allowing agents to exploit passively shared information during execution and perform tasks under all possible communication levels.

The main contributions of this chapter are three-fold:

- In Section 8.1, we propose the setting of *hybrid execution* in MARL, in which agents perform partially-observable cooperative tasks under all possible communication levels, ranging from no communication (fully decentralized) to full communication between the agents (fully centralized). We formalize our setting with *hybrid partially observable Markov decision process* (H-POMDP);

- In Section 8.2, we propose the *multi-agent local observation sharing under communication dropout* (MARO), an approach that combines an autoregressive predictive model of the observations of the agents and a novel dropout-based training scheme;

- In Section 8.3, we evaluate MARO in different benchmark and novel environments, using different RL algorithms, showing that our approach consistently allows agents to act with different communication levels.

The work described in this chapter has been published in:

- P. P. Santos*, D. S. Carvalho*, **M. Vasco***, A. Sardinha, P. A. Santos, A. Paiva, & F. S. Melo. *Centralized Training with Hybrid Execution in Multi-Agent Reinforcement Learning.* In: arXiv preprint arXiv:2210.06274, 2022 (Under review).

* Shared first-authorship.

8.1 Hybrid Execution in Multi-Agent Reinforcement Learning

A fully cooperative multi-agent system with Markovian dynamics can be modelled as a decentralized partially observable Markov decision process (Dec-POMDP) [117], introduced in Section 2.2.1. Fully decentralized approaches to MARL directly apply standard single-agent RL algorithms for learning each agents' policy π_i in a decentralized manner. In Independent Q-learning (IQL) [157], each agent treats other agents as being part of the environment, ignoring influences of other agents' observations and actions. More recently, under the paradigm of centralized training with decentralized execution, QMIX [132] aims at learning decentralized policies with centralization at training time while fostering cooperation among the agents. Finally, if we know that all agents can share their local observations among themselves at execution time, we can use either of the two approaches to learn fully centralized policies.

None of the aforementioned methods assume that, at certain time-steps, agents may have access to observations of other agents. Therefore, decentralized agents are unable to take advantage of the additional information that they may receive from other agents at execution time, and centralized agents are unable to act when the sharing of information fails. In this work, we introduce *hybrid execution* in MARL, a setting in which agents act regardless of the communication level in the environment, while taking advantage of additional information they may receive during execution. To formalize this setting, we define *hybrid partially observable Markov decision process* (H-POMDPs), a new class of multi-agent POMDPs that explicitly considers a specific communication process C among the agents.

Definition 11 (H-POMDP) *An hybrid partially observable Markov decision process* (H-POMDP) *is defined as a tuple* $([n], \mathcal{X}, \mathcal{A}, \mathcal{P}, r, \gamma, \mathcal{Z}, \mathcal{O}, C)$ *where:*

- $[n] = \{1, \ldots, n\}$ *is the set of indexes of* n *agents;*
- $\mathcal{X}$ *is the set of states of the environment;*
- $\mathcal{A} = \times_i \mathcal{A}_i$ *is the set of joint actions, where* $\mathcal{A}_i$ *is the set of individual actions of agent* i*;*
- $\mathcal{P}$ *is the set of probability distributions over next states in* $\mathcal{X}$*, one for each state and action in* $\mathcal{X} \times \mathcal{A}$*;*
- $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *maps states and actions to expected rewards;*
- $\gamma \in [0, 1[$ *is a discount factor;*
- $\mathcal{Z} = \times_i \mathcal{Z}_i$ *is the set of joint observations, where* $\mathcal{Z}_i$ *is the set of local observations of agent* i*;*
- $\mathcal{O}$ *is the set of probability distributions over joint observations in* $\mathcal{Z}$*, one for each state and action in* $\mathcal{X} \times \mathcal{A}$*;*
- C*, a* $n \times n$ *communication matrix such that* $[C]_{i,j} = p_{i,j}$ *is the probability that, at a certain time step, agent* i *has access to the local observation of agent* j *in* $\mathcal{Z}_j$*.*

H-POMDPs generalize both the notion of decentralized execution and of centralized execution in MARL. Specifically, for a given Dec-POMDP, we can consider C the identity matrix to capture fully decentralized execution and C a matrix of ones to capture fully centralized execution.

In our setting we assume that, at execution time, agents will face an H-POMDP with an unknown communication matrix C, sampled from a set $\mathcal{C}$ according to an unknown probability distribution μ. The performance of the agent is measured as $J_\mu(\pi) = \mathbb{E}_{C\sim\mu}\left[J(\pi; C)\right]$, where $J(\pi; C)$ denotes the expected discounted cumulative reward under an H-POMDP with communication matrix C. At training time, agents may have access to the fully centralized H-POMDP. Therefore, the setting we consider is one of *centralized training with hybrid execution*, subject to an unknown communication process at execution time.

We note here that every H-POMDP has a corresponding Dec-POMDP, which can be obtained by adequately changing the observation space $\mathcal{Z}$ and the set of emission probability distributions $\mathcal{O}$. Consequently, any reinforcement learning method can be trained to solve a specific H-POMDP, with a specific communication matrix C, by solving the corresponding Dec-POMDP. However, we seek to find a method that takes explicit advantage of the characteristics of hybrid execution to be able to act on H-POMDPs *regardless of the matrix* C that models the communication process at execution time. To the best of our knowledge, currently there exists no method that is able to solve such problem.

8.2 Multi-Agent Observation Sharing with Communication Dropout (MARO)

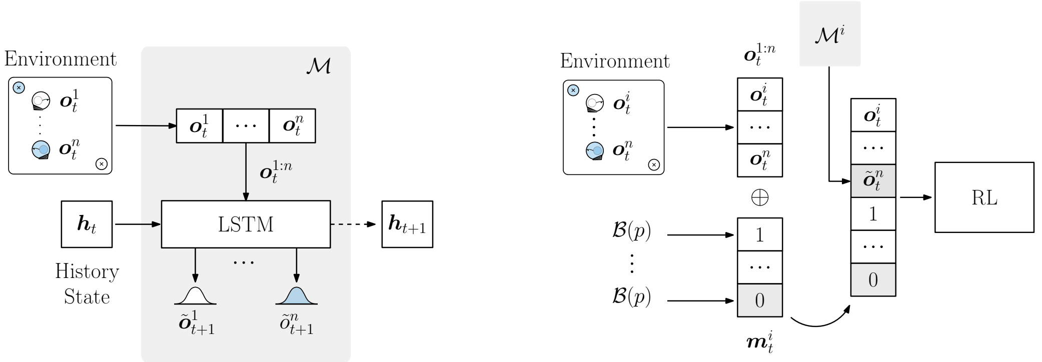
While acting on an H-POMDP, agents may not have access to the complete joint-observation due to partial perceptual availability. We propose MARO, a novel approach to exploit shared information and overcome communication failures during task execution. MARO is composed of two elements: an autoregressive predictive model, that learns a joint representation of the agents observations and estimates missing observations from previous ones, and a training scheme for the RL controllers, that simulates faulty communication at training time.

8.2.1 Predictive Model

The predictive model $\mathcal{M}$, depicted in Fig. 8.1a, is used to estimate the local observations of all agents $\boldsymbol{o}_t^{1:n} = \{\boldsymbol{o}_t^1, \ldots, \boldsymbol{o}_t^n\}$, where $\boldsymbol{o}_t^i$ corresponds to the observation of the i-th agent, in order to overcome missing observations during execution. Thus, we learn a transition model $p(\boldsymbol{o}_{t+1}^{1:n} \mid \boldsymbol{o}_t^{1:n}, \boldsymbol{h}_t)$, where $\boldsymbol{o}_t^{1:n}$ is the current observations of the agents and $\boldsymbol{h}_t$ is a *joint-history* representation, that implicitly encodes information regarding the policy of the agents. The transition model is trained in order to predict the next-step observations of the agents, $\boldsymbol{o}_{t+1}$. We instantiate $p_\theta(\boldsymbol{o}_{t+1}^{1:n} \mid \boldsymbol{o}_t^{1:n}, \boldsymbol{h}_t)$ as an LSTM, parameterized by θ, with:

$$p_\theta(\boldsymbol{o}_{t+1}^{1:n} \mid \boldsymbol{o}_t^{1:n}, \boldsymbol{h}_t) = \prod_{i=1}^{n} p_\theta(\boldsymbol{o}_{t+1}^{i} \mid \boldsymbol{o}_t^{1:n}, \boldsymbol{h}_t), \tag{8.1}$$

where $p_\theta(\boldsymbol{o}_{t+1}^{i} \mid \boldsymbol{o}_t^{1:n}, \boldsymbol{h}_t)$ is the Gaussian distribution of the predicted observations of the i-th agent. We train the predictive model and RL controllers simultaneously: we consider single-step observation transitions $(\boldsymbol{o}_t^{1:n}, \boldsymbol{o}_{t+1}^{1:n})$ and evaluate the negative log-likelihood of the target next-step observation $\boldsymbol{o}_{t+1}^{1:n}$, given the estimated next-step observation distribution



(a) Predictive Model.

(b) Training Scheme.

Figure 8.1: MARO for hybrid execution: (a) an autoregressive predictive model $\mathcal{M}$ to estimate the next-step observations of all agents, $p(o_{t+1}^{1:n} \mid o_t^{1:n}, h_t)$, given the previous ones, $o_t^{1:n}$, and a joint-history representation, h_t; (b) a training scheme for RL controllers, that randomly drops agent observations following the communication masks m_t^i, sampled accordingly to the communication level in the environment p.

$p_\theta(\cdot \mid \boldsymbol{o}_t^{1:n}, \boldsymbol{h}_t)$:

$$\mathcal{L}_{\mathcal{M}}(\boldsymbol{o}_t^{1:n}, \boldsymbol{o}_{t+1}^{1:n}) = -\sum_{i=1}^{n} \log p_\theta(\boldsymbol{o}_{t+1}^i \mid \boldsymbol{o}_t^{1:n}, \boldsymbol{h}_t). \tag{8.2}$$

8.2.2 Training Scheme

We also introduce an RL training scheme, depicted in Fig. 8.1b, which simulates the communication process at execution time and is agnostic to the type of algorithm. We setup all RL controllers to receive as input the joint observation $\boldsymbol{o}_t^{1:n}$. However, in our setting, some observations may not be shared at execution time. To overcome such issue, we employ the predictive model to estimate the non-shared observations $\tilde{\boldsymbol{o}}_t^j$, with $j \in [n]$. We also setup the controllers to receive as additional input *communication masks*, $\boldsymbol{m}_t$, binary vectors that indicate the real and predicted components of $\boldsymbol{o}_t^{1:n}$. These masks implicitly provide the recurrent controllers with information regarding the communication level in the environment, allowing them to measure the uncertainty regarding the input.

During centralized training, methods typically assume that each agent has access to the local observations of all other agents. Instead of using such information to train the agents, we instead propose to explicitly simulate the communication conditions of execution time: we randomly *dropout* agent observations. At the beginning of each episode, we sample a communication level $p \sim \mathcal{U}(0, 1)$.[1] Given p, we build at each time-step a communication mask $m_t^i = \{0, 1\}^n$ for each agent $i \in [n]$, with $m_t^i[i] = 1$. We sample the communication masks from independent Bernoulli distributions $m_t^i[j] \sim \mathcal{B}(p)$, for $j \in [n] \setminus i$. At execution time, we extract the observation masks m_t^i directly from the environment, accordingly to the communication faults experienced by each agent during execution.

The communication mask indicates which components of the agent-specific joint-observation $\boldsymbol{o}_t^{1:n,i}$ are dropped. Specifically, we use the real observation $\boldsymbol{o}_t^k$ if $m_t^i[k] = 1$ and use the estimated observation $\tilde{\boldsymbol{o}}_t^k$ otherwise. We provide each agent with an independent

[1] In the absence of prior information regarding the communication level in the environment, in the training scheme we sample symmetrical communication matrices C, such that $p_{i,j} = p_{j,i} = p$ and $p_{i,i} = 1$.

instance of the predictive model $\mathcal{M}^i$, which updates the estimated joint-observations in the perspective of the agent $\tilde{\boldsymbol{o}}_t^{1:n,i} = \{\tilde{\boldsymbol{o}}_t^{1,i}, \ldots, \tilde{\boldsymbol{o}}_t^{n,i}\}$ and maintains an agent-specific joint-history representation, h_t^i.

8.3 Evaluation

We evaluate our approach for hybrid execution by answering the following questions:

(i) What is the performance of MARO in multi-agent cooperative tasks with partial observability, considering unknown levels of communication at execution time?

(ii) What is the importance of the training scheme for hybrid execution, and considering different dropout schemes?

(iii) What are the benefits of the predictive model for hybrid execution?

To address (i), we evaluate in Section 8.3.3 our approach against other relevant baselines and considering multiple RL algorithms. The results show that MARO outperforms the baselines, allowing the execution of tasks across multiple communication levels. Regarding (ii), we perform in Section 8.3.4.1 an ablation study of the training scheme, highlighting the importance of simulating the communication process at execution time during training. We address (iii) in Section 8.3.4.2, highlighting the benefits of the predictive model, both in terms of training sample efficiency and in allowing centralized execution agents to exploit shared information across multiple communication levels.

8.3.1 Scenarios

We focus our evaluation on multi-agent cooperative environments. As discussed in Papoudakis et al. [124], the main challenges in current MARL benchmark scenarios involve coordination, large action spaces, sparse rewards and non-stationarity. As such, in these tasks the sharing of local information *is not critical* to their successful execution (as we show in Section 8.3.3) Thus, in order to evaluate the role of passively sharing information amongst agents in MARL, we propose three environments adapted from Lowe et al. [99]:

- **SpreadBlindfold**: The environment consists of three agents and three designated landmarks in a 2D map. At the start of each episode both the position of the agents and of the landmarks are randomly generated. The goal of the agents is to cover all the landmarks while avoiding collisions: agents are (globally) rewarded considering how far the closest agent is to each landmark (sum of the minimum distances) and are (locally) penalized if they collide with other agents. In contrast with the original Simple Spread environment, the agent's observation only includes the position and velocity of the agent itself and the relative position of all landmarks. Through communication, the agents can access the position and velocities of the other agents.

- **HearSee**: The environment consists of two heterogeneous agents and a single landmark in a 2D map. At the start of each episode both the position of the agents and of the landmark are randomly generated. The goal of the agents is to cooperate in order for both of them to cover the landmark: agents are (globally) rewarded considering how far each agent is to the landmark (sum of the distances). In this scenario, one of the agents ("Hear" agent) is provided with the absolute position of the landmark in its observation. However, it does not have access to its own position. The other agent ("See" agent) always has access the position and velocities of both agents in its observation, yet does not have access to the position of the landmark. Through communication, the agents have access to both their positions and the position of the landmark in order to complete the task.

- **SpreadXY**: The environment consists of two heterogeneous agents and two designated landmarks in a 2D map. At the start of each episode both the position of the agents and of the landmarks are randomly generated. The goal of the agents is to cover all the landmarks while avoiding collisions: agents are (globally) rewarded considering how far the closest agent is to each landmark (sum of the minimum distances) and are (locally) penalized if they collide with other agents. Differently from SpreadBlindfold, one of the agents has access to the X position and velocity of both agents, while the other agent has access to the Y position and velocity of both agents. Both agents observe as well the absolute position of all landmarks. Through communication, the agents can access the complete position and velocities of the other agents and cover the landmarks.

In addition to the proposed environments, we evaluate our approach in standard MARL benchmark scenarios, in particular in the SpeakerListener environment [99]. Finally, we consider H-POMDPs with communication matrices such that each agent i can always access its own local observation, i.e., $p_{i,i} = 1$, and the communication matrix is symmetric between agents i and j, i.e., $p_{i,j} = p_{j,i}$. Moreover, we use the same $p_{i,j} = p$ for all pairs of different agents i, j. Therefore, we also use p to unambiguously denote the *communication level* of a given H-POMDP. Additionally, we note that in all the environments the previous action taken by agent i, a^i_{t-1}, is included in its local observation o^i_t.

8.3.2 Baselines and Experimental Methodology

We compare MARO against different baselines that correspond to different levels of information-sharing between the agents. We consider two "extreme" cases:

- **Observation** (Obs.): Agents only have access to their own observations and are unable to communicate with other agents, corresponding to standard MARL algorithms designed for decentralized execution;

- **Joint-Observation** (J. obs.): Agents always have access to the observations of all agents, corresponding to standard MARL algorithms designed for centralized execution. This baseline is unable to perform when communication fails and can be seen as an upper bound.

To the best of our knowledge, there exists no method developed specifically for the problem of executing with faulty communication with unknown dynamics to serve as a direct comparison to MARO. As such, we modify the model proposed by Kim et al. [78], and repurpose it as a baseline:

- **Message-Dropout** (MD): Agents train with communication failing half of the times (fixed $p = 0.5$), without communication masks and without the predictive model.

We employ the same controller networks across all evaluations. The networks include recurrent layers to mitigate the effects of partial observability [174]. We consider two different MARL algorithms: QMIX and Independent Q-Learning (IQL). We follow the

Table 8.1: Average episodic returns and 95% bootstrapped confidence interval over five seeds for MARO and other methods in all scenarios. J.obs corresponds to an upper-level performance bound and Obs corresponds to a lower-level performance bound. Higher is better.

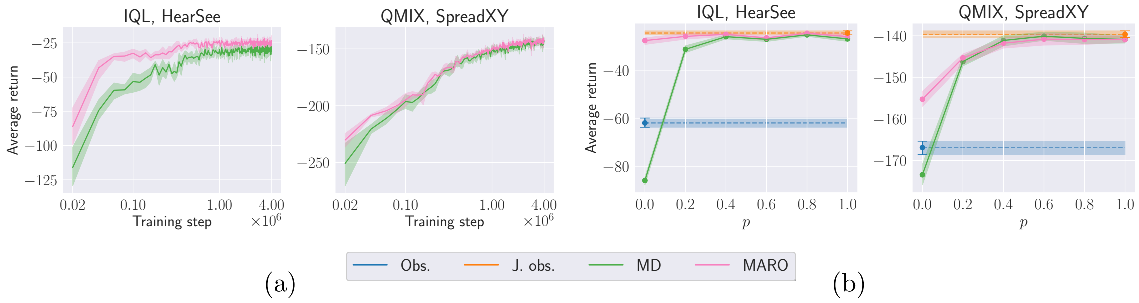
	IQL				**QMIX**			
Environment	**Obs.**	**J. obs.**	**MD**	**MARO**	**Obs.**	**J. obs.**	**MD**	**MARO**
SpreadXY	-173.2 (-0.8,+1.1)	-140.3 (-0.8,+0.8)	**-150.6** (-1.1,+1.6)	**-148.7** (-0.7,+0.7)	-166.9 (-1.7,+1.5)	-139.7 (-0.8,+0.9)	-147.1 (-0.5,+0.5)	**-144.2** (-0.5,+0.6)
SpreadBlindfold	-432.2 (-5.9,+6.0)	-403.4 (-5.4,+3.9)	**-402.7** (-1.0,+1.0)	**-405.3** (-4.3,3.4)	-418.1 (-3.3,+3.5)	-376.0 (-2.9,+3.3)	**-405.7** (-5.0,+5.0)	**-401.2** (-6.4,+6.4)
HearSee	-61.9 (-1.8,+2.1)	-24.5 (-0.8,+0.9)	-37.1 (-0.6,+0.6)	**-25.9** (-0.3,+0.3)	-54.9 (-1.5,+1.3)	-24.1 (-0.7,+0.9)	**-25.4** (-0.4,+0.4)	**-25.1** (-0.3,+0.4)
SpeakerListener	-31.7 (-1.2,+1.4)	-25.4 (-1.1,+1.1)	-27.8 (-0.3,+0.3)	**-25.6** (-0.5,+0.6)	-26.1 (-1.3,+1.2)	-25.1 (-1.1,+1.1)	**-23.2** (-1.0,+1.2)	**-23.3** (-1.1,+1.3)

training hyperparameters suggested by Papoudakis et al. [124]; we train all models for 4M steps, performing 5 training runs for each experimental setting and 50 evaluation rollouts for each training run. We assume that $p = 1$ at $t = 0$ for the MD and MARO algorithms. The performance of the Obs. and J. Obs. agents is evaluated by aggregating evaluation rollouts with $p = 0$ and $p = 1$, respectively. The other algorithms are evaluated for p sampled from a discretized uniform distribution. If the communication level is not explicitly referred, then the values correspond to the average performance across all communication levels. We refer to Appendix F for a complete description of the experimental methodology, including hyperparameters of the predictive model and the RL controllers.

8.3.3 Results

We present the main evaluation results in Table 8.1. For each environment, RL algorithm and method, we present the values of the accumulated rewards obtained. The values that are not significantly different than the highest are presented in bold. The results show that MARO consistently performs equal or better than the MD baseline across all scenarios and algorithms. MARO is able to exploit the information provided by the other agents, in contrast with the fully decentralized approaches (Obs.). Moreover, MARO is often able to achieve performances comparable to the fully centralized agent (J. Obs.), which executes with full communication, despite acting under partial perceptual availability, i.e., with faulty communication. We also note here that information sharing between agents may not lead to performance gains across all scenarios. For instance, in the SpeakerListener scenario, where the speaker agent can actively share the goal position to the listener agent, decentralized QMIX (Obs.) is able to perform competitively in comparison to centralized QMIX (J. Obs.), without requiring the passive sharing of local observations.

In Figure 8.2, we highlight the training curves and the performance of the approaches for different communication levels p for the HearSee and SpreadXY environments (more in Appendix E.1). The results show that MARO outperforms the MD baseline in terms of sample efficiency, with a bigger jump-start in the initial stages of the training (Fig. 8.2a), and overall performance across all communication levels (Fig. 8.2b). Additionally, MARO significantly outperforms the Obs. baseline in settings with no communication ($p = 0$); MD struggles to act in the same setting, performing worse than the fully decentralized baseline. Moreover, the performance of MARO improves as the level of communication in the environment increases, showing that our model is able to efficiently make use of all



(a) (b)

Figure 8.2: (a) Average episodic returns during training with 95% confidence interval for MARO and MD; (b) Average episodic returns during training with 95% confidence interval for different communication levels at execution time for all approaches.

provided information. Appendix E.1 includes all training curves and performance results.

8.3.4 Ablation Study

We perform an ablation study on the two components of MARO to study their impact on the overall performance of the method. We introduce ablated versions of MARO along two different axes: (i) training with our proposed scheme (**Train ✓**), against training with fixed $p = 1$ (**Train ✗**); and (ii) using the predictive model to estimate missing observations (**Pred ✓**) against a naive dropout approach that replaces missing observations with zeros (**Pred ✗**). Table 8.2 shows the results of the ablation study across environments and algorithms. We discuss the results along each axis separately.

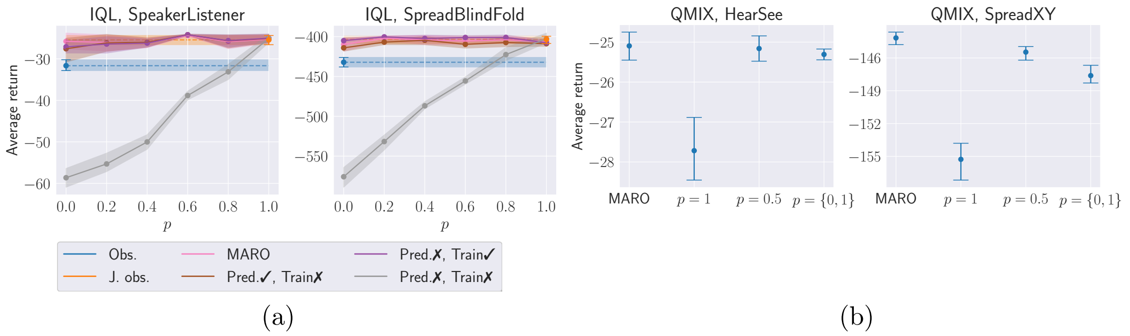
8.3.4.1 Training Scheme

The results in Table 8.2 highlight the importance of the training scheme for hybrid execution: introducing our proposed training scheme (**Pred ✗, Train ✓**) over the fully-ablated version of MARO (**Pred ✗, Train ✗**) results in a significant performance improvement, while removing the training scheme from MARO (**Pred ✓, Train ✗**) reduces the performance of our approach.

In Figure 8.3a we evaluate the impact of the training scheme in the performance across different communication levels, shown for the SpeakerListener and SpreadBlindfold

Table 8.2: Average episodic returns and 95% bootstrapped confidence interval over five seeds for the ablated versions of MARO in all scenarios. Higher is better.

	IQL				QMIX			
	MARO	**Pred. ✓**	**Pred. ✗**	**Pred. ✗**	**MARO**	**Pred. ✓**	**Pred. ✗**	**Pred. ✗**
Environment		**Train ✗**	**Train ✓**	**Train ✗**		**Train ✗**	**Train ✓**	**Train ✗**
SpreadXY	-148.7 (-0.7,+0.7)	-158.1 (-0.6,+0.6)	**-146.4** (-0.6,+0.7)	-197.4 (-1.7,+1.8)	**-144.2** (-0.5,+0.6)	-155.3 (-1.5,+1.9)	**-144.4** (-1.1,+1.7)	-197.8 (-1.7,+1.7)
SpreadBlindfold	**-405.3** (-4.3,3.4)	-408.6 (-3.0,+3.0)	**-402.9** (-1.8,+2.1)	-479.2 (-2.0,+2.4)	**-401.2** (-6.4,+6.4)	**-393.8** (-5.2,+4.8)	-407.7 (-5.9,+6.9)	-479.6 (-3.9,+4.3)
HearSee	**-25.9** (-0.3,+0.3)	-27.7 (-0.5,+0.5)	**-26.1** (-0.3,+0.3)	-78.0 (-4.2,2.6)	**-25.1** (-0.3,+0.4)	-27.7 (-0.8,+0.7)	**-25.1** (-0.3,+0.3)	-79.4 (-3.4,+3.4)
SpeakerListener	**-25.6** (-0.5,+0.6)	**-25.8** (-0.7,+0.7)	**-25.8** (-0.7,+0.7)	-43.5 (-0.9,+0.9)	**-23.3** (-1.1,+1.3)	**-23.5** (-1.2,+1.4)	**-23.3** (-1.2,+1.4)	-39.0 (-1.5,+1.7)



(a)

(b)

Figure 8.3: (a) Average episodic returns with 95% confidence interval for different communication levels at execution time for the ablated versions of MARO; (b) Average episodic returns with 95% confidence interval for different sampling schemes.

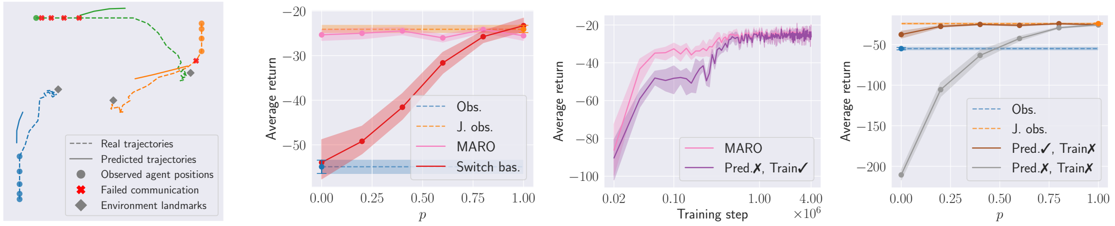
environments (additional results in Appendix E.2). The results reveal that MARO is able to perform well across the whole communication spectrum. The results also show that while the fully-ablated version of our approach struggles to act with any communication level other than $p = 1$, adding MARO's training scheme alone results in a method that is able to perform close to optimally across the entire spectrum of p. Finally, removing the training scheme from MARO results in a performance drop, especially for $p = 0$. In summary, the training scheme is beneficial for different values of p without sacrificing performance for $p = 1$.

We also evaluate the impact of different sampling strategies of p during training, beyond our proposed communication sampling scheme during training, $p \sim \mathcal{U}(0, 1)$. In Figure 8.3b, we compare our sampling scheme in the HearSee and SpreadXY environments (more in Appendix E.2) against three other sampling approaches: a categorical distribution over the communication level extremes, $p \sim \mathcal{U}\{0, 1\}$; a fixed sampling scheme with $p = 0.5$, which corresponds to the training approach of [78] with additional observation masks; and a fixed sampling scheme with $p = 1$, without observation masks. The results highlight the importance of simulating faulty communication during the training phase, as the fully centralized training scheme ($p = 1$) is outperformed by all other approaches. In Appendix E.2, we present the full results of this evaluation, showing the impact of the sampling scheme in other scenarios and RL algorithms.

8.3.4.2 Predictive Model

The results in Table 8.2 also highlight the significant improvement in the performance of MARO when employing the predictive model to estimate the missing observations. Specifically, adding the predictive model (**Pred ✓, Train ✗**) to the fully ablated version of MARO (**Pred ✗, Train ✗**) results in a significant performance increase. Removing the predictive component from MARO (**Pred ✗, Train ✓**) still results in a competitive algorithm without a significant performance drop in the majority of scenarios. However, there are significant advantages to employing the predictive model:

- The predictive model improves the sample efficiency of MARO. In Figure 8.4c, we show the training curves for MARO, with and without the predictive model. The results show that employing the predictive model results in a significant jump-start in terms of sample efficiency. The predictive model provides the RL controllers with an estimate of the missing agent trajectories, instead of replacing the missing input



(a) S.Blindfold, QMIX. (b) HearSee, QMIX. (c) HearSee, IQL. (d) HearSee, QMIX.

Figure 8.4: Evaluation of the predictive model of MARO: (a) Trajectory estimation from the perspective of the blue agent, using its agent-specific predictive model; (b) Average episodic returns with 95% confidence interval for different communication levels at execution time of MARO against the Switch baseline; (c) Average episodic returns during training with 95% bootstrapped confidence interval for MARO and an ablated version without the predictive model; (d) Average episodic returns with 95% confidence interval for different communication levels at execution time of ablated versions of MARO, showing the ability of the predictive model for zero-shot execution.

with zeros, resulting in improved performance during the initial stages of the training. This improvement is consistent across several environments and algorithms, as shown in Appendix E.2;

- The predictive model provides robustness to centralized execution methods when performing tasks in settings with potential faulty communication, despite never being trained to execute in such conditions. In other words, the predictive model allows for zero-shot multi-agent execution with respect to the communication failures. In Figure 8.4d, we show that employing the predictive model to predict missing information at execution time allows a standard centralized method to perform the task with minimum performance loss (**Pred ✗, Train ✗**);

- The predictive model is able to perform accurate agent modelling with faulty communication, providing an interpretable insight into the decision-making process of the agents. In Figure 8.4a, we show the predicted trajectories of all agents in the perspective of the blue agent, which are close to the real trajectories performed by all agents (more in Appendix E.3).

We can also assess the correctness of the predictions made by the predictive model by evaluating the performance of MARO against a *Switch* baseline. In this baseline, the agents choose actions using two controllers, selected accordingly to the level of communication in the environment at each timestep: one that uses the joint observation at the timesteps it is available (similar to the Joint Observation baseline), and one that uses only the local observation, otherwise (similar to the Observation baseline). We show the results of the comparison in performance between MARO and the Switch baseline in Fig. 8.4b. The results reveal that MARO is able to exploit the predicted observations for communication levels $p < 1$, outperforming the Switch baseline in different communication settings.

8.4 Concluding Remarks

In this chapter we explored how multi-agent reinforcement learning agents can exploit shared information to perform partially-observable cooperative tasks under partial perceptual availability, i.e., with different communication levels at execution time. We introduced the paradigm of hybrid execution, where agents are expected to act across all possible communication levels, and formalized the setting with H-POMDPs. We contributed with MARO, an approach that combines an autoregressive predictive model and a dropout-based training scheme to allow for hybrid executions. Furthermore, we introduced three novel cooperative scenarios for MARL that explicitly require information sharing during execution. The results show that MARO allows agents to exploit shared information to successfully execute tasks under all possible communication levels, outperforming the baseline approaches.

We have shown how artificial agents, in single and multi-agent settings, can be designed in order to perform tasks under partial perceptual availability, thus addressing the second part of our research question: *how to leverage representations in the execution of tasks under changing conditions of perceptual availability?*. However, by design, we have assumed that the observations collected by the agent accurately reflect the state of the environment. In more realistic scenarios, perceptual information can become degraded, due to changing perceptual conditions, hindering its representational utility. Moreover, complex artificial agents, such as robots, equipped with with multiple sensors are targets for adversarial sensory attacks, which may provide incorrect information regarding the state of the world and compromise the safe actuation of the agent. In future work, we plan on addressing how agents can autonomously evaluate the quality of its perceptual experience, in order to be resilient to adversarial attacks.

Chapter 9

Conclusions

"Enquanto houver estrada pra andar, a gente vai continuar."
Jorge Palma

In this thesis we addressed fundamental issues in multimodal representation learning for the perception and agency of artificial agents. Motivated by the urgency of addressing such issues, in this thesis we investigated the following research question:

> **Research Question**: *How can we endow artificial agents with mechanisms to learn representations from multimodal observations provided by their environment and to leverage such representations in the execution of tasks under changing conditions of perceptual availability?*

Throughout the thesis, we have highlighted several advantages of *multimodality* for the perception and actuation of artificial agents. With *motion concepts* (**contribution 1**), we have shown how multimodality allows sample-efficient learning of representations for downstream classification tasks. We have also shown with *multimodal transfer in reinforcement learning* (**contribution 4**) that agents able to learn to represent multimodal information provided by their environment can outperform agents that perceive, train and execute tasks over single-modality information. Finally, we have shown how the benefits of multimodality can be exploited in multi-agent systems with *hybrid execution* (**contribution 5**), in which we have shown how a perceptual model that aggregates observations of different agents during training can be used to predict missing observations at execution time, and improve the sample-efficiency of downstream RL controllers.

To answer *how can we endow artificial agents with mechanisms to learn representations from multimodal observations provided by their environment*, inspired by human perception we exploited *hierarchy* in the design of multimodal representation models. With MUSE (**contribution 2**), we have shown the fundamental role of hierarchy for high-quality and coherent cross-modality generation. We explored hierarchy, once again, with GMC (**contribution 3**) and showed how it can be applied to a variety of downstream tasks, providing robust classification and control performance with missing modality information at test time. We have also validated the benefits of hierarchy in *multimodal transfer in reinforcement learning* (**contribution 4**), where we have shown that agents endowed with hierarchical perceptual models outperform agents with single-level perceptual models in transferring policies across different sets of available modalities.

Finally, to answer "*how to leverage multimodal representations in the execution of tasks*". we proposed different methods that allow single-agent and multi-agent systems to act, with minimum performance loss, *under changing conditions of perceptual availability.* We first defined the problem of *multimodal transfer in reinforcement learning* (**contribution 4**) and contributed a three-stage approach that allows RL agents to train and execute policies across different sets of modalities, without fine-tuning the policy to the new perceptual conditions of the environment. Secondly, we defined the paradigm of *hybrid execution* in multi-agent reinforcement learning and proposed an approach that allows agents to exploit multiple observations, passively-shared among the agents, to perform collaborative tasks under different communication levels in the environment.

9.1 Summary of Contributions

The main contributions of this thesis were as follows:

1. **Motion Concepts** [166, 167], a multimodal probabilistic representation for human actions and an algorithm to learn representations from human demonstration;

2. **MUSE** [168, 169], a multimodal generative model to learn representations from an arbitrary number of heterogeneous data sources, leveraging hierarchy to perform effective cross-modality inference;

3. **GMC** [128], a multimodal contrastive representation model that aligns multimodal representations from an arbitrary number of heterogeneous data sources, providing robust downstream classification and control performance with missing modality information;

4. **Multimodal Transfer Reinforcement Learning** [143], an approach that considers multimodal representations to allow agents to transfer task policies across different sets of perceptual modalities available at execution time;

5. **Hybrid Execution in Multi-Agent Reinforcement Learning** [139], an approach that allows agents to exploit passively-shared information at execution time in order to perform cooperative tasks under any possible communication level.

We present a complete list of publications related to this thesis in Appendix A.

9.2 Main Conclusions

In light of the contributions above, the main conclusions of this thesis are:

- We have shown the benefits of learning multimodal representations for sample-efficient task execution. For example, in Chapter 4 we have shown that by accounting for multiple channels of complementary information, motion concepts effectively allow for few shot recognition of human actions, often from a single demonstration. Moreover, the redundancy of the information provided by the different modality channels allows the distinction of action classes with similar motion patterns, a difficult task for single (motion) modality-based representations;

- We have shown how *hierarchy* plays a fundamental role in learning rich multimodal representations. In Chapter 5, we shown how the hierarchical design of MUSE allows for high-quality and coherent cross-modality generation, regardless of its intrinsic complexity, from available modality information. Once again, in Chapter 6 we revisit hierarchy in the design of *GMC*, and show how it outperforms other state-of-the-art multimodal representations models across a wide range of scenarios, in downstream classification tasks. Moreover, we show quantitatively and qualitatively the improved alignment of multimodal representations encoded by GMC in comparison with other models. Overall, the results show that considering hierarchical latent variable models improves the performance of multimodal generative models. Indeed, current state-of-the-art multimodal generative models, such as diffusion models [131], employ hierarchical architectures;

- We have shown how multimodal representations can be employed by autonomous agents to *act under changing conditions of perceptual availability*. In Chapter 7, we have considered the single-agent scenario and evaluated two variations of the multimodal policy transfer problem: *cross-modality* policy transfer, where the modalities at execution time are distinct from the ones available during policy training, and *multimodal* policy transfer, where the agent only has access to a partial set of modalities at execution time, in comparison with the complete set of modality information available during policy training. For both classes of multimodal policy transfer problems, we have shown that our agents are able to robustly perform tasks under partial perceptual availability. Moreover, we have shown how controllers trained over a multimodal representation can outperform a controller trained directly over observations of a single modality;

- We have shown how the ideas of multimodal representation learning can be extended to *multi-agent systems* performing cooperative tasks under partial observability. In Chapter 8, we have shown how a perceptual model that is able to aggregate information provided by the observations of different agents at training time can be used both to predict missing observations and to mitigate the decrease in performance of the agents, due to lack of communication at execution time. In fact, the results show that our approach allows agents to perform cooperative tasks across all possible communication levels, consistently outperforming the baseline approaches, and highlights the increased sample efficiency of RL algorithms, when the agents employ a perceptual model to estimate missing information during training.

9.3 Future Work

We now discuss potential extensions of the ideas introduced in this thesis.

Compositional-GMC

In Chapter 6, we introduced GMC, a self-supervised approach for multimodal representation learning that explicitly aligns single-modality and multimodal representations. GMC allows the execution of downstream classification and control tasks, with complete observations or with single-modality observations. However, the current method disregards information jointly provided by subsets of multiple modalities. In complex scenarios with large number of modalities, such information should be exploited to improve the performance of the model in the downstream task, in particular in scenarios where modalities provide complementary information regarding the observed phenomena. A natural next step would be to tackle such limitation and extend GMC with the ability to consider subset of available modalities by taking inspiration from current approaches in self-supervised computer vision, such as masked autoencoders [63], and contrast over different subsets of available modalities at training time.

Robustness to Adversarial Perceptual Attacks

In Chapter 7, we introduced the problem of *multimodal transfer reinforcement learning*, where agents transfer policies across different sets of perceptual modalities at execution time. However, we implicitly have assumed that the observations that the agent collects are representative of the state of the environment. In real-world scenarios, modality-specific observations can become degraded, due to adverse perceptual conditions, or even become under attack, providing information that does not correspond to the actual state of the world. It would be interesting to consider how can an agent leverage multimodal information to reason over the perceptual quality of the modality-specific information it attains from the world. In particular, it could be interesting to extend our architectures with predictive components that allow our agents to model the uncertainty over current observations given previous modality-specific observations, following approaches such as world models [57]. Addressing such problem will provide further resilience to the actuation of multimodal reinforcement learning agents.

Evaluation of the Perceptual Robustness of Real-World Agents

Recently, Du et al. [36] have shown how robots, endowed with image and sound sensors, can leverage sound information to perform tasks in moments of visual occlusion through ad-hoc methods. Following this line of research, it could be interesting to evaluate extensions of *multimodal transfer in reinforcement learning* to real-world scenarios, where robots are able to leverage multimodal sensory information to perform tasks, without requiring explicit supervision over which signals to consider at each time-step. This will require further research on the quality of perceptual information provided by real-world sensors, which was not considered in the virtual environments explored in this thesis.

◇ ◇ ◇

In this thesis we addressed how to *leverage multimodal representations* to allow artificial agents to execute tasks when provided with *partial perceptual information*, with minimum performance loss. We explored computational methods to learn representations from multiple heterogeneous sources of data and to endow reinforcement learning agents with robust performance to missing perceptual information at test time.

The expected widespread use of autonomous agents in our near-future societies raises new technical challenges regarding the perception and actuation of such agents. Both virtual (e.g. personal assistants) and physical (e.g. social robots, autonomous vehicles) agents will require perceptual capabilities beyond single-modality observations, allowing them to effectively perceive the world around them through complementary and (often) redundant channels. Moreover, the actions of these agents will be scrutinized under strict security regulations, requiring the development of actuation mechanisms robust to changing environmental conditions, beyond the controlled laboratory setting [89, 33]. However, this is not an issue exclusively reserved for the future: perceptual and actuation issues of real-world artificial agents are already impacting the lives of human beings [151].

We hope that the contributions of this thesis can spark the discussion on the requirements of such real-world applications, while simultaneously opening the door to additional research on the application of multimodal representation learning methods for the safe and robust actuation of artificial agents.

Bibliography

[1] Z. Ahmad and N. Khan. Human Action Recognition Using Deep Multilevel Multimodal Fusion of Depth and Inertial Sensors. *IEEE Sensors Journal*, 20(3):1445–1455, 2020.

[2] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. A Brief Survey of Deep Reinforcement Learning. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[3] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. 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*, pages 2236–2246. ACL, 2018.

[4] T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2019.

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Challenges and applications in multimodal machine learning. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition*, volume 21, pages 17–48. ACM, 2018.

[6] A. H. Bell, M. A. Meredith, A. J. Van Opstal, and D. P. Munoz. Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. *Journal of Neurophysiology*, 93(6):3659–3673, 2005.

[7] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[8] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 2013.

[9] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 2022.

[10] M. Bourguignon, M. Baart, E. C. Kapnoula, and N. Molinaro. Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech. *Journal of Neuroscience*, 40(5):1053–1065, 2020.

[11] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11(1): 1–94, 1999.

[12] R. E. Briscoe. Multisensory Processing and Perceptual Consciousness: Part I. *Philosophy Compass*, 11(2):121–133, 2016.

[13] V. Bruce and P. Green. *Visual perception: Physiology, psychology and ecology, 2nd ed.* Visual perception: Physiology, psychology and ecology, 2nd ed. Lawrence Erlbaum Associates, Inc, 1990.

[14] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-VAE. *CoRR*, abs/1804.03599, 2018.

[15] H. Burianová, L. Marstaller, P. Sowman, G. Tesan, A. N. Rich, M. Williams, G. Savage, and B. W. Johnson. Multimodal functional imaging of motor imagery using a novel paradigm. *NeuroImage*, 71:50–58, 2013.

[16] L. Buşoniu, R. Babuška, and B. De Schutter. Multi-agent Reinforcement Learning: An Overview. In *Innovations in Multi-Agent Systems and Applications - 1*, Studies in Computational Intelligence, pages 183–221. Springer, 2010.

[17] G. A. Calvert. Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex (New York, N.Y.: 1991)*, 11(12):1110–1123, 2001.

[18] G. A. Calvert, E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. R. Williams, P. K. McGuire, P. W. R. Woodruff, S. D. Iversen, and A. S. David. Activation of Auditory Cortex During Silent Lipreading. *Science*, 276(5312):593–596, 1997.

[19] L. Canese, G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Re, and S. Spanò. Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. *Applied Sciences*, 11(11):4948, 2021.

[20] Y.-H. Cao and J. Wu. Rethinking Self-Supervised Learning: Small is Beautiful. *CoRR*, abs/2103.13559, 2021.

[21] D. Carvalho, F. S. Melo, and P. Santos. A New Convergent Variant of Q-Learning with Linear Function Approximation. In *Advances in Neural Information Processing Systems*, volume 33, pages 19412–19421. Curran Associates, Inc., 2020.

[22] Y. Chandak, G. Theocharous, J. Kostas, S. Jordan, and P. Thomas. Learning Action Representations for Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 941–950. PMLR, 2019.

[23] C. Chen, R. Jafari, and N. Kehtarnavaz. Improving Human Action Recognition Using Fusion of Depth Camera and Inertial Sensors. *IEEE Transactions on Human-Machine Systems*, 45(1):51–61, 2015.

[24] C. Chen, R. Jafari, and N. Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 168–172, 2015.

[25] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[27] X. Chen, J. Hu, L. Li, and L. Wang. Efficient Reinforcement Learning in Factored MDPs with Application to Constrained RL. In *Proceedings of the International Conference on Learning Representations*, 2021.

[28] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles. Advances in Human Action Recognition: A Survey. *CoRR*, abs/1501.05964, 2015.

[29] M. A. Cohen, D. C. Dennett, and N. Kanwisher. What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences*, 20(5):324–335, 2016.

[30] A. Damásio. Time-Locked Multiregional Retroactivation: A Systems-Level Proposal for the Neural Substrates of Recall and Recognition. *Cognition*, 33(1-2):25–62, 1989.

[31] T. Dean and K. Kanazawa. A Model for Reasoning About Persistence and Causation. *Computational Intelligence*, 5(2):142–150, 1989.

[32] T. Denning, C. Matuszek, K. Koscher, J. R. Smith, and T. Kohno. A Spotlight on Security and Privacy Risks with Future Household Robots: Attacks and Lessons. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 105–114. Association for Computing Machinery, 2009.

[33] V. Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way.* Springer Cham, 2019.

[34] A. Dosovitskiy and T. Brox. Generating Images with Perceptual Similarity Metrics based on Deep Networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[35] K. Doya. What Are the Computations of the Cerebellum, the Basal Ganglia and the Cerebral Cortex? *Neural Networks: The Official Journal of the International Neural Network Society*, 12(7-8):961–974, 1999.

[36] M. Du, O. Y. Lee, S. Nair, and C. Finn. Play it by Ear: Learning Skills amidst Occlusion through Audio-Visual Imitation Learning. *CoRR*, abs/2205.14850, 2022.

[37] Y. Du, W. Wang, and L. Wang. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.

[38] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. Go-Explore: a New Approach for Hard-Exploration Problems. *CoRR*, abs/1901.10995, 2021.

[39] S. B. Edwards, C. L. Ginsburgh, C. K. Henkel, and B. E. Stein. Sources of Subcortical Projections to the Superior Colliculus in the Cat. *The Journal of Comparative Neurology*, 184(2):309–329, 1979.

[40] X. Fang, J. Wang, G. Song, Y. Han, Q. Zhao, and Z. Cao. Multi-Agent Reinforcement Learning Approach for Residential Microgrid Energy Scheduling. *Energies*, 13(1):123, 2020.

[41] C. Fefferman, S. Mitter, and H. Narayanan. Testing the Manifold Hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[42] F. E. Fernandes, G. Yang, H. M. Do, and W. Sheng. Detection of Privacy-Sensitive Situations for Social Robots in Smart Homes. In *Proceedings of the IEEE International Conference on Automation Science and Engineering*, pages 727–732, 2016.

[43] N. Ferns, P. Panangaden, and D. Precup. Metrics for Finite Markov Decision Processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 162–169. AUAI Press, 2004.

[44] J. Ferreira, D. M. de Matos, and R. Ribeiro. Fast and Extensible Online Multivariate Kernel Density Estimation. *CoRR*, abs/1606.02608, 2016.

[45] C. Finn and S. Levine. Deep Visual Foresight for Planning Robot Motion. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2786–2793. IEEE Press, 2017.

[46] T. Flash and B. Hochner. Motor Primitives in Vertebrates and Invertebrates. *Current Opinion in Neurobiology*, 15(6):660–666, 2005.

[47] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[48] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. Counterfactual Multi-Agent Policy Gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18, pages 2974–2982. AAAI Press, 2018.

[49] C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare. DeepMDP: Learning Continuous Latent Space Models for Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2170–2179. PMLR, 2019.

[50] A. A. Ghazanfar and C. E. Schroeder. Is Neocortex Essentially Multisensory? *Trends in Cognitive Sciences*, 10(6):278–285, 2006.

[51] Y. Gil and B. Selman. A 20-Year Community Roadmap for Artificial Intelligence Research in the US. *CoRR*, abs/1908.02624, 2019.

[52] R. Givan, T. Dean, and M. Greig. Equivalence Notions and Model Minimization in Markov Decision Processes. *Artificial Intelligence*, 147(1):163–223, 2003.

[53] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3):e30, 2021.

[54] J. González, A. Barros-Loscertales, F. Pulvermüller, V. Meseguer, A. Sanjuán, V. Belloch, and C. Avila. Reading Cinnamon Activates Olfactory Brain Regions. *NeuroImage*, 32(2):906–912, 2006.

[55] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[56] W. Guo, J. Wang, and S. Wang. Deep Multimodal Representation Learning: A Survey. *IEEE Access*, 7:63373–63394, 2019.

[57] D. Ha and J. Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[58] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *Proceedings of the International Conference on Learning Representations*, 2020.

[59] D. Hafner, T. P. Lillicrap, M. Norouzi, and J. Ba. Mastering Atari with Discrete World Models. In *Proceedings of the International Conference on Learning Representations*, 2021.

[60] W. Hamilton, Z. Ying, and J. Leskovec. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[61] H. He, J. Boyd-Graber, K. Kwok, and I. I. I. Hal Daumé. Opponent Modeling in Deep Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1804–1813. PMLR, 2016.

[62] J. He, D. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. In *Proceedings of the International Conference on Learning Representations*, 2019.

[63] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[64] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[65] I. Higgins, A. Pal, A. Rusu, L. Matthey, C. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner. DARLA: improving zero-shot transfer in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML'17*, pages 1480–1490. JMLR, 2017.

[66] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *Proceedings of the International Conference on Learning Representations*, 2022.

[67] F. Ho, A. Salta, R. Geraldes, A. Goncalves, M. Cavazza, and H. Prendinger. Multi-Agent Path Finding for UAV Traffic Management. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19,

pages 131–139. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[68] G. Hu, Y. Zhu, D. Zhao, M. Zhao, and J. Hao. Event-Triggered Multi-agent Reinforcement Learning with Communication under Limited-bandwidth Constraint. *CoRR*, abs/2010.04978, 2020.

[69] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural Autoregressive Flows. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

[70] J. Imran and B. Raman. Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):189–208, 2020.

[71] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9(1):2, 2021.

[72] W. James. *The Principles of Psychology, Vol I.* The principles of psychology, Vol I. Henry Holt and Co, 1890.

[73] J. L. W. V. Jensen. Sur Les Fonctions Convexes Et Les Inégalités Entre Les Valeurs Moyennes. *Acta Mathematica*, 30(none):175–193, 1906.

[74] S. Ji, W. Xu, M. Yang, and K. Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[75] J. Kelly and G. S. Sukhatme. Visual-Inertial Sensor Fusion: Localization, Mapping and Sensor-to-Sensor Self-calibration. *The International Journal of Robotics Research*, 30(1):56–79, 2011.

[76] M. Kiefer, E.-J. Sim, B. Herrnberger, J. Grothe, and K. Hoenig. The Sound of Concepts: Four Markers for a Link Between Auditory and Conceptual Brain Systems. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 28(47):12224–12230, 2008.

[77] D. Kim, S. Moon, D. Hostallero, W. J. Kang, T. Lee, K. Son, and Y. Yi. Learning to Schedule Communication in Multi-agent Reinforcement Learning. *CoRR*, abs/1902.01554, 2019.

[78] W. Kim, M. Cho, and Y. Sung. Message-dropout: an efficient training method for multi-agent deep reinforcement learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI'19, pages 6079–6086. AAAI Press, 2019.

[79] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *CoRR*, abs/1312.6114, 2014.

[80] D. P. Kingma and M. Welling. An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[81] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved Variational Inference with Inverse Autoregressive Flow. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[82] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022.

[83] I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021.

[84] V. Konda and J. Tsitsiklis. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[85] T. Korthals, D. Rudolph, J. Leitner, M. Hesse, and U. Rückert. Multi-Modal Generative Models for Learning Epistemic Active Sensing. In *Proceedings of the International Conference on Robotics and Automation*, pages 3319–3325, 2019.

[86] W. M. Land, D. Volchenkov, B. E. Bläsing, and T. Schack. From Action Representation to Action Execution: Exploring the Links Between Cognitive and Biomechanical Levels of Motor Control. *Frontiers in Computational Neuroscience*, 7:127, 2013.

[87] K. G. Larsen and A. Skou. Bisimulation Through Probabilistic Testing. *Information and Computation*, 94(1):1–28, 1991.

[88] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[89] R. Leenes and F. Lucivero. Laws on robots, laws by robots, laws in robots: Regulating robot behaviour by design. *Law, Innovation and Technology*, 6(2):193–220, 2014.

[90] T. Lesort, N. Díaz-Rodríguez, J.-F. I. Goudou, and D. Filliat. State Representation Learning for Control: An Overview. *Neural Networks*, 108:379–392, 2018.

[91] L. Li, T. Walsh, and M. Littman. Towards a Unified Theory of State Abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, 2006.

[92] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous Control with Deep Reinforcement Learning. *CoRR*, abs/1509.02971, 2019.

[93] M. L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, ICML'94, pages 157–163. Morgan Kaufmann Publishers Inc., 1994.

[94] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. Kot. Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks. *IEEE Transactions on Image Processing*, 27:1586–1599, 2018.

[95] M. Liu, H. Liu, and C. Chen. Enhanced Skeleton Visualization for View Invariant Human Action Recognition. *Pattern Recognition*, 68:346–362, 2017.

[96] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[97] D. Llorens, F. Prat, A. Marzal, J. M. Vilar, M. J. Castro, J. C. Amengual, S. Barrachina, A. Castellanos, S. España, J. A. Gómez, J. Gorbe, A. Gordo, V. Palazón, G. Peris, R. Ramos-Garijo, and F. Zamora. The UJIpenchars Database: a Pen-Based Database of Isolated Handwritten Characters. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 2008.

[98] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

[99] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6382–6393. Curran Associates Inc., 2017.

[100] R. C. Luo and C. C. Lai. Multisensor Fusion-Based Concurrent Environment Mapping and Moving Object Detection for Intelligent Service Robotics. *IEEE Transactions on Industrial Electronics*, 61(8):4043–4051, 2014.

[101] S. Ma, D. Mcduff, and Y. Song. Unpaired Image-to-Speech Synthesis With Multimodal Information Bottleneck. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7597–7606, 2019.

[102] K. Man, J. Kaplan, H. Damásio, and A. Damásio. Neural Convergence and Divergence in the Mammalian Cerebral Cortex: From Experimental Neuroanatomy to Functional Neuroimaging. *The Journal of Comparative Neurology*, 521(18):4097–4111, 2013.

[103] W. Mao, K. Zhang, E. Miehling, and T. Başar. Information State Embedding in Partially Observable Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 59th IEEE Conference on Decision and Control*, pages 6124–6131, 2020.

[104] L. Marstaller and H. Burianová. The Multisensory Perception of Co-Speech Gestures – a Review and Meta-Analysis of Neuroimaging Studies. *Journal of Neurolinguistics*, 30:69–77, 2014.

[105] D. Maurer, T. Pathman, and C. J. Mondloch. The Shape of Boubas: Sound-Shape Correspondences in Toddlers and Adults. *Developmental Science*, 9(3):316–322, 2006.

[106] L. McInnes, J. Healy, N. Saul, and L. Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[107] C. Meo and P. Lanillos. Multimodal VAE Active Inference Controller. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2693–2699, 2021.

[108] K. Meyer and A. Damásio. Convergence and Divergence in a Neural Architecture for Recognition and Memory. *Trends in Neurosciences*, 32(7):376–382, 2009.

[109] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing Atari with Deep Reinforcement Learning. *CoRR*, abs/1312.5602, 2013.

[110] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-Level Control Through Deep Reinforcement Learning. *Nature*, 518(7540): 529–533, 2015.

[111] M. Müller. Dynamic Time Warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer, 2007.

[112] K. K. Ndousse, D. Eck, S. Levine, and N. Jaques. Emergent Social Learning via Multi-agent Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7991–8004. PMLR, 2021.

[113] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *Proceedings of the NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[114] Y. Niu, R. Paleja, and M. Gombolay. Multi-Agent Graph-Attention Communication and Teaming. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pages 964–973. International Foundation for Autonomous Agents and Multiagent Systems, 2021.

[115] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 8–13, 2012.

[116] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60, 2013.

[117] F. A. Oliehoek and C. Amato. *A Concise Introduction to Decentralized POMDPs*. Springer Publishing Company, Incorporated, 1st edition, 2016.

[118] F. A. Oliehoek, M. T. J. Spaan, and N. Vlassis. Optimal and Approximate Q-value Functions for Decentralized Pomdps. *Journal of Artificial Intelligence Research*, 32 (1):289–353, 2008.

[119] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep Decentralized Multi-Task Multi-Agent Reinforcement Learning Under Partial Observability. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML'17*, pages 2681–2690. JMLR, 2017.

[120] S. Omidshafiei, D.-K. Kim, M. Liu, G. Tesauro, M. Riemer, C. Amato, M. Campbell, and J. P. How. Learning to Teach in Cooperative Multiagent Reinforcement Learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI'19, pages 6128–6136. AAAI Press, 2019.

[121] I. Osband and B. V. Roy. Near-optimal Reinforcement Learning in Factored Mdps. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 1 of *NIPS'14*, pages 604–612. MIT Press, 2014.

[122] G. Papamakarios, T. Pavlakou, and I. Murray. Masked Autoregressive Flow for Density Estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 2335–2344. Curran Associates Inc., 2017.

[123] G. Papoudakis, F. Christianos, and S. V. Albrecht. Agent Modelling under Partial Observability for Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2021.

[124] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. *CoRR*, abs/2006.07869, 2021.

[125] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8026–8037. Curran Associates Inc., 2019.

[126] L. Pinto, D. Gandhi, Y. Han, Y.-L. Park, and A. Gupta. The Curious Robot: Learning Visual Representations via Physical Interactions. In *Proceedings of the European Conference on Computer Vision*, Lecture Notes in Computer Science, pages 3–18. Springer International Publishing, 2016.

[127] P. Poklukar, V. Polianskii, A. Varava, F. T. Pokorny, and D. K. Jensfelt. Delaunay Component Analysis for Evaluation of Data Representations. In *Proceedings of the International Conference on Learning Representations*, 2022.

[128] P. Poklukar*, M. Vasco*, H. Yin, F. S. Melo, A. Paiva, and D. Kragic. Geometric Multimodal Contrastive Representation Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17782–17800. PMLR, 2022.

[129] R. Poppe. A Survey on Vision-Based Human Action Recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[130] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant Visual Representation by Single Neurons in the Human Brain. *Nature*, 435(7045):1102–1107, 2005.

[131] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR*, abs/2204.06125, 2022.

[132] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4295–4304. PMLR, 2018.

[133] B. Ravindran and A. G. Barto. Relativized options: choosing the right transformation. In *Proceedings of the Twentieth International Conference on Machine Learning*, ICML'03, pages 608–615, 2003.

[134] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. A Generalist Agent. *CoRR*, abs/2205.06175, 2022.

[135] D. J. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, volume 37 of *ICML'15*, pages 1530–1538. JMLR, 2015.

[136] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos. Multimodal Human Action Recognition in Assistive Human-robot Interaction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2702–2706, 2016.

[137] O. Rybkin, K. Daniilidis, and S. Levine. Simple and Effective VAE Training with Calibrated Decoders. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9179–9189. PMLR, 2021.

[138] R. Salakhutdinov. Learning Deep Generative Models. *Annual Review of Statistics and Its Application*, 2(1):361–385, 2015.

[139] P. P. Santos*, D. S. Carvalho*, M. Vasco*, A. Sardinha, P. A. Santos, A. Paiva, and F. S. Melo. Centralized Training with Hybrid Execution in Multi-Agent Reinforcement Learning. *CoRR*, abs/2210.06274, 2022.

[140] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature*, 588(7839):604–609, 2020.

[141] K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao. A Survey of Deep Reinforcement Learning in Video Games. *CoRR*, abs/1912.10944, 2019.

[142] Y. Shi, N. Siddharth, B. Paige, and P. H. S. Torr. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 15718–15729. Curran Associates Inc., 2019.

[143] R. Silva, M. Vasco, F. S. Melo, A. Paiva, and M. Veloso. Playing Games in the Dark: An Approach for Cross-Modality Transfer in Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1260–1268. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

[144] K. Simonyan and A. Zisserman. Two-stream Convolutional /Networks for Action Recognition in Videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 1 of *NIPS'14*, pages 568–576. MIT Press, 2014.

[145] A. Singh, T. Jain, and S. Sukhbaatar. Individualized Controlled Continuous Communication Model for Multiagent Cooperative and Competitive Tasks. In *Proceedings of the International Conference on Learning Representations*, 2019.

[146] C. K. Sønderby, T. Raiko, L. Maalø e, S. K. Sønderby, and O. Winther. Ladder Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[147] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4263–4270. AAAI Press, 2017.

[148] Y. Song and D. P. Kingma. How to Train Your Energy-Based Models. *CoRR*, abs/2101.03288, 2021.

[149] C. Spence. Crossmodal Correspondences: A Tutorial Review. *Attention, Perception & Psychophysics*, 73(4):971–995, 2011.

[150] T. R. Stanford and B. E. Stein. Superadditivity in Multisensory Integration: Putting the Computation in Context. *Neuroreport*, 18(8):787–792, 2007.

[151] J. Stilgoe. Who Killed Elaine Herzberg? In *Who's Driving Innovation? New Technologies and the Collaborative State*, pages 1–6. Springer International Publishing, 2020.

[152] S. Sukhbaatar, A. Szlam, and R. Fergus. Learning Multiagent Communication with Backpropagation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[153] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[154] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* A Bradford Book, 2018.

[155] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[156] M. Suzuki, K. Nakayama, and Y. Matsuo. Joint Multimodal Learning with Deep Generative Models. *CoRR*, abs/1611.01891, 2016.

[157] M. Tan. Multi-agent reinforcement learning: independent versus cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, ICML'93, pages 330–337. Morgan Kaufmann Publishers Inc., 1993.

[158] J. Taylor, D. Precup, and P. Panagaden. Bounding Performance Loss in Approximate MDP Homomorphisms. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[159] Y. Tian and J. Engel. Latent Translation: Crossing Modalities by Bridging Generative Models. *CoRR*, abs/1902.08261, 2019.

[160] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Information Fusion*, 64:131–148, 2020.

[161] J.-F. Tremblay, T. Manderson, A. Noca, G. Dudek, and D. Meger. Multimodal Dynamics Modeling for Off-Road Autonomous Vehicles. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1796–1802, 2021.

[162] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov. Learning Factorized Multimodal Representations. In *Proceedings of the International Conference on Learning Representations*, 2018.

[163] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569. Association for Computational Linguistics, 2019.

[164] A. Vahdat and J. Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679. Curran Associates, Inc., 2020.

[165] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[166] M. Vasco, F. Melo, D. Martins de Matos, A. Paiva, and T. Inamura. Online Motion Concept Learning: A Novel Algorithm for Sample-Efficient Learning and Recognition of Human Actions. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 2244–2246. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[167] M. Vasco, F. S. Melo, D. M. d. Matos, A. Paiva, and T. Inamura. Learning Multimodal Representations for Sample-efficient Recognition of Human Actions. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4288–4293, 2019.

[168] M. Vasco, H. Yin, F. S. Melo, and A. Paiva. Leveraging Hierarchy in Multimodal Generative Models for Effective Cross-Modality Inference. *Neural Networks*, 146: 238–255, 2022.

[169] M. Vasco, H. Yin, F. S. Melo, and A. Paiva. How to Sense the World: Leveraging Hierarchy in Multimodal Perception for Robust Reinforcement Learning Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '22, pages 1301–1309. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

[170] R. Vedantam, I. Fischer, J. Huang, and K. Murphy. Generative Models of Visually Grounded Imagination. In *Proceedings of the International Conference on Learning Representations*, 2018.

[171] R. Vemulapalli, F. Arrate, and R. Chellappa. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014.

[172] P. Walker, J. G. Bremner, U. Mason, J. Spring, K. Mattock, A. Slater, and S. P. Johnson. Preverbal Infants' Sensitivity to Synaesthetic Cross-Modality Correspondences. *Psychological Science*, 21(1):21–25, 2010.

[173] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012.

[174] R. E. Wang, M. Everett, and J. P. How. R-MADDPG for Partially Observable Environments and Limited Communication. *CoRR*, abs/2002.06684, 2020.

[175] P. Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *CoRR*, abs/1804.03209, 2018.

[176] C. J. C. H. Watkins. *Learning from Delayed Rewards.* PhD Thesis, Cambridge University,, 1989.

[177] H. Wei, G. Zheng, V. Gayah, and Z. Li. A Survey on Traffic Signal Control Methods. *CoRR*, abs/1904.08117, 2020.

[178] L. Weng. Self-Supervised Representation Learning. *lilianweng.github.io*, 2019.

[179] M. Wu and N. Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[180] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27, 2012.

[181] A. Xie, D. Losey, R. Tolsma, C. Finn, and D. Sadigh. Learning Latent Representations to Influence Multi-Agent Interaction. In *Proceedings of the Conference on Robot Learning*, pages 575–588. PMLR, 2021.

[182] X. Yang and Y. L. Tian. EigenJoints-based action recognition using Naïve-Bayes-Nearest-Neighbor. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 14–19, 2012.

[183] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the International Conference on Computer Vision*, pages 1331–1338, 2011.

[184] H. Yin, F. S. Melo, A. Billard, and A. Paiva. Associate latent encodings in learning from demonstrations. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 3848–3854. AAAI Press, 2017.

[185] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access*, 8:58443–58469, 2020.

[186] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[187] M. Zambelli, A. Cully, and Y. Demiris. Multimodal representation models for prediction and control from partial information. *Robotics and Autonomous Systems*, 123:103312, 2020.

[188] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors*, 19(5), 2019.

[189] K. Zhang, Z. Yang, and T. Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In *Handbook of Reinforcement Learning and Control*, Studies in Systems, Decision and Control, pages 321–384. Springer International Publishing, 2021.

[190] M. Zhang, S. Vikram, L. Smith, P. Abbeel, M. Johnson, and S. Levine. SOLAR: Deep Structured Representations for Model-Based Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7444–7453. PMLR, 2019.

[191] S. Q. Zhang, Q. Zhang, and J. Lin. Efficient Communication in Multi-Agent Reinforcement Learning via Variance Based Control. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[192] S. Q. Zhang, Q. Zhang, and J. Lin. Succinct and Robust Multi-Agent Communication With Temporal Message Control. In *Advances in Neural Information Processing Systems*, volume 33, pages 17271–17282. Curran Associates, Inc., 2020.

[193] S. Zhao, J. Song, and S. Ermon. Learning Hierarchical Features from Deep Generative Models. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4091–4099. PMLR, 2017.

[194] S. Zhao, J. Song, and S. Ermon. Towards Deeper Understanding of Variational Autoencoding Models. *CoRR*, abs/1702.08658, 2017.

[195] C. Zhu, M. Dastani, and S. Wang. A Survey of Multi-Agent Reinforcement Learning with Communication. *CoRR*, abs/2203.08975, 2022.

Appendix A

Summary of Publications

The work presented in this dissertation directly resulted in 1 journal article (Neural Networks), 4 conference proceedings publications (ICML, AAMAS, IROS), 2 refereed extended abstracts and 2 preprints.

Journal Articles

- **Miguel Vasco**, Hang Yin, Francisco S. Melo, and Ana Paiva. Leveraging Hierarchy in Multimodal Generative Models for Effective Cross-modality Inference. *Neural Networks (2021 Special Issue on AI and Brain Science: Brain-inspired AI)*, 146: 238–255, 2022;

Conference Articles

- Petra Poklukar*, **Miguel Vasco***, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric Multimodal Contrastive Representation Learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 17782–17800, PMLR, 2022;

- **Miguel Vasco**, Hang Yin, Francisco S. Melo, and Ana Paiva. How to Sense the World: Leveraging Hierarchy in Multimodal Perception for Robust Reinforcement Learning Agents. In *Proceedings of 21st International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1301–1309, 2022;

- Rui Silva, **Miguel Vasco**, Francisco S. Melo, Ana Paiva, and Manuela Veloso. Playing Games in the Dark: An Approach for Cross-Modality Transfer in Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1260–1268, 2020;

- **Miguel Vasco**, Francisco S. Melo, David Martins de Matos, Ana Paiva, and Tetsunari Inamura. Learning Multimodal Representations for Sample-efficient Recognition of Human Actions. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4288–429, 2019;

* Shared-first authorship.

Referred Extended Abstracts

- **Miguel Vasco**. Multimodal Representation Learning for Robotic Cross-Modality Policy Transfer. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 2225–2227. 2020;

- **Miguel Vasco**, Francisco S. Melo, David Martins de Matos, Ana Paiva, and Tetsunari Inamura. Online Motion Concept Learning: A Novel Algorithm for Sample-Efficient Learning and Recognition of Human Actions. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 2244–2246, 2019;

Preprints

- Pedro P. Santos*, Diogo S. Carvalho*, **Miguel Vasco***, Alberto Sardinha, Pedro A. Santos, Ana Paiva, & Francisco S. Melo. Centralized Training with Hybrid Execution in Multi-Agent Reinforcement Learning. *CoRR*, abs:2210.06274, 2022 (Under review);

- **Miguel Vasco**, Francisco S. Melo, and Ana Paiva. MHVAE: a Human-inspired Deep Hierarchical Generative Model for Multimodal Representation Learning. *CoRR*, abs:2006.02991, 2020;

* Shared first-authorship.

Appendix B

Cross-Modality Inference Evaluation Metrics

In this section we present a detailed description of the evaluation metrics employed in this work.

B.1 Log-Likelihood Metrics

In Section 5.6.1, we evaluate the generative performance of MUSE following standard log-likelihood metrics. As depicted in Fig. B.1a, we employ the model to compute a distribution and, also, to evaluate such distribution. We start by estimating the marginal log-likelihoods $\log p(\mathbf{x}_i)$ through importance-weighted sampling, following the standard single-modality evidence lower-bound,

$$\log p(\boldsymbol{x}_i) \leq \mathbb{E}_{\boldsymbol{z}_i^{1:N} \sim q_{\phi_i^b}(\cdot|\boldsymbol{x}_i)} \left[\log \sum_{n=1}^{N} \left(\frac{p_{\theta_i^b}(\boldsymbol{x}_i \mid \boldsymbol{z}_i^n)\, p(\boldsymbol{z}_i^n)}{q_{\phi_i^b}(\boldsymbol{z}_i^n \mid \boldsymbol{x}_i)} \right) \right]. \tag{B.1}$$

To evaluate the joint-modality log-likelihood $\log p(\boldsymbol{x}_{i:M})$, we compute a importance-weighted estimate using the standard multimodal evidence lower-bound,

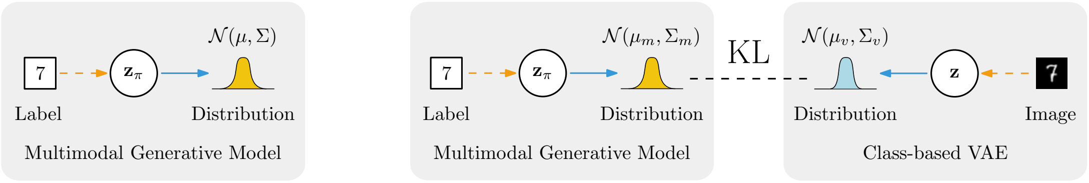
$$\log p(\boldsymbol{x}_{i:M}) \leq \mathbb{E}_{\boldsymbol{z}_\pi^{1:N} \sim q_\phi(\cdot|\boldsymbol{x}_{i:M})} \left[\log \sum_{n=1}^{N} \left(\frac{p(\boldsymbol{z}_\pi^n) \prod_i^M p_{\theta_i}(\boldsymbol{x}_i \mid \boldsymbol{z}_\pi^n)}{q_\phi(\boldsymbol{z}_\pi^n \mid \boldsymbol{x}_{i:M})} \right) \right], \tag{B.2}$$

where the posterior distribution $q_\phi(\boldsymbol{z}_\pi^n \mid \boldsymbol{x}_{i:M})$ accounts for the combination of the bottom-level decoders $q_{\phi_{i:M}^b}(\boldsymbol{c}_{i:M} \mid \boldsymbol{x}_{i:M})$ and top-level decoders $q_{\phi_{i:M}^t}(\boldsymbol{z}_\pi^n \mid \boldsymbol{c}_{i:M})$. Finally, following Shi et al. [142], we resort to the corresponding single variational posterior $q(\boldsymbol{z}_\pi \mid \boldsymbol{x}_{\neg i})$ to compute the conditional log-likelihood $\log p(\boldsymbol{x}_i \mid \boldsymbol{x}_{\neg j})$,

$$\log p(\boldsymbol{x}_i \mid \boldsymbol{x}_{\neg i}) \leq \mathbb{E}_{\boldsymbol{z}_\pi^{1:N} \sim q_\phi(\cdot|\boldsymbol{x}_{\neg j})} \left[\log \sum_{n=1}^{N} \left(\frac{p_{\theta_i}(\boldsymbol{x}_i \mid \boldsymbol{z}_\pi^n)\, p(\boldsymbol{z}_\pi^n)}{q_\phi(\boldsymbol{z}_\pi^n \mid \boldsymbol{x}_{\neg i})} \right) \right], \tag{B.3}$$

B.2 Kullback-Leibler Distance

In Section 5.7, we evaluate the cross-modality generative performance of MUSE employing the Kullback-Leibler Distance (KL-distance) metric.


$\mathcal{N}(\mu, \Sigma)$
7
$\mathbf{z}_\pi$
Label
Distribution
Multimodal Generative Model
$\mathcal{N}(\mu_m, \Sigma_m)$
KL
$\mathcal{N}(\mu_v, \Sigma_v)$
7
$\mathbf{z}_\pi$
$\mathbf{z}$
Label
Distribution
Distribution
Image
Multimodal Generative Model
Class-based VAE

(a) Log-likelihood

(b) KL-distance

Figure B.1: Standard and proposed evaluation metrics for the performance of multimodal generative models: (a) the log-likelihood and (b) KL-distance metrics. Note that in (a) *the distribution used as the metric for the evaluation is the one learned by the model being evaluated.* (Best viewed with zoom).

As depicted in Fig. B.1b, for each class $k \in [0, K]$ in the dataset, we employ a class-specific, pretrained, variational autoencoder to encode a distribution of class-specific images $\mathcal{N}(\mu_v^k, \Sigma_v^k)$. We compare such distribution with the distribution of class-specific images, generated from label information using MUSE $\mathcal{N}(\mu_m^k, \Sigma_m^k)$. The KL-distance metric for each model is computed accordingly to,

$$KL = \frac{1}{K} \sum_{k=0}^{K} D_{\mathrm{KL}} \left(\mathcal{N}(\mu_m^k, \Sigma_m^k) \mid \mathcal{N}(\mu_v^k, \Sigma_v^k) \right), \tag{B.4}$$

where we assume that the distributions are multivariate Gaussians with diagonal covariance matrices.

B.3 Accuracy and Modality Distance

In addition to Kullback-Leibler Distance, we employ two different complementary metrics to evaluate the cross-modality generation performance of MUSE that account for the semantic coherence of the samples generated, as well as their relative quality to the samples of the dataset.

Accuracy, originally proposed by Shi *et al.* [142], evaluates the semantic coherence of the samples generated by cross-modal inference, using pre-trained modality-specific classifiers.

To evaluate the relative *quality* of the generated samples, we propose the *modality distance* (D) metric. For each class $k \in [0, K]$ in the dataset, we encode representations of samples both from the dataset and generated by cross-modal inference, resulting in a distribution of real dataset representations $\mathcal{N}(\mu_r^k, \Sigma_r^k)$ and of generated representations $\mathcal{N}(\mu_g^k, \Sigma_g^k)$. The encoding procedure of both distributions is performed by pretrained class-and-modality-specific auto-encoders. The modality distance is then given by the Frechét distance between the two distributions, averaged over all classes [64]:

$$D = \frac{1}{K} \sum_{k} \left\| \mu_r^k - \mu_g^k \right\|^2 + \mathrm{Tr} \left(\Sigma_r^k + \Sigma_g^k - 2 \left(\Sigma_r^k \Sigma_g^k \right)^{1/2} \right). \tag{B.5}$$

Appendix C

Ablation Study of MUSE

In this Section, we present additional evaluations performed on the MUSE model, considering the scenarios presented in Chapter 5. We introduce two additional variations of the MUSE model, considering the proposed joint-modality encoder solutions:

- **MUSE** - The original MUSE model employing the POE encoder introduced in Section 5.4.3 and the ALMA training scheme.
- **MUSE-0** - The MUSE model employing the naive encoder introduced in Section 5.4.1;
- **MUSE-σ** - The MUSE model employing the Nexus encoder introduced in Section 5.4.2 and the FPD training scheme (with $\rho = 0.1$ selected empirically);

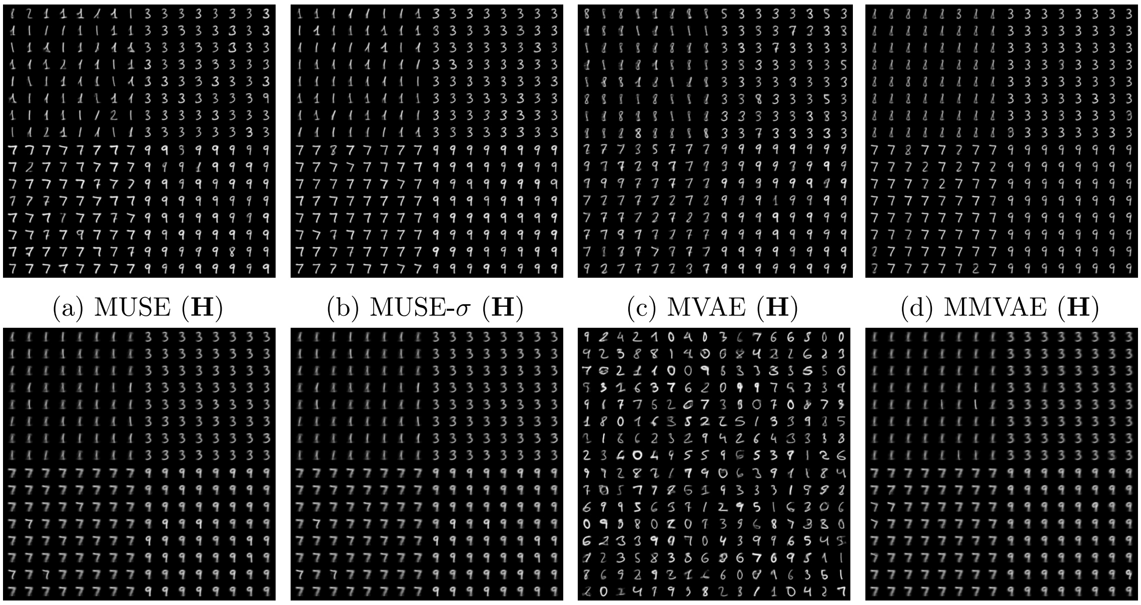
C.1 Hierarchical Design

In a preliminary evaluation, we appraise the role of hierarchy for cross-modal inference. We consider a subset of the MHD dataset, concerning only the image, $\mathbf{x}_i$, and label, $\mathbf{x}_l$, associated with handwritten digits. We show that MUSE is able to generate high-quality image samples through CMI, outperforming all other baselines.

We implement non-hierarchical versions of the MUSE models where we input the modality observations directly into the joint-modality encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{x}_{1:M})$. Moreover, to evaluate the potential of hierarchical architectures regardless of the base model, we also extend the baseline models with two representation levels following the MUSE architecture. The hierarchical version of MVAE employs a top-level POE multimodal encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) \propto \prod_{m=1}^{M} q_{\phi^t_m}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_m)$. The hierarchical version of MMVAE employs a top-level MOE multimodal encoder $q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) = \sum_{m=1}^{M} \frac{1}{M} q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_m)$.

We evaluate all models considering the complementary metrics for cross-modal inference presented in Section 5.7. The quantitative results concerning the cross-modal *accuracy* and *modality distance* metrics are presented in Table C.1 and image samples resulting from cross-modal generation are shown in Fig. C.1. All numerical results are averaged over 5 independently-seeded runs.

The results attest that MUSE is the only model able to perform CMI with both high accuracy and low image modality distance. For MUSE the extension to a hierarchical architecture results in a significant decrease on the modality distance metric, resulting in higher-quality samples, as shown in Fig. C.1a. This decrease demonstrates the benefit of considering hierarchical representation levels in the architecture of multimodal generative models: the top multimodal representation learns a representation able to generate coherent



(a) MUSE (**H**) (b) MUSE-σ (**H**) (c) MVAE (**H**) (d) MMVAE (**H**)

(e) MUSE (**S**) (f) MUSE-σ (**S**) (g) MVAE (**S**) (h) MMVAE (**S**)

Figure C.1: Cross-modality images from available label information $\boldsymbol{x}_l = \{\text{“0”}, \text{“4”}, \text{“7”}, \text{“9”}\}$, generated by hierarchical (a-d) and single-level (e-h) multimodal generative models. MUSE is the only model able perform CMI and generate samples with high *accuracy* and low *modality distance* (best viewed with zoom).

modality-specific latent samples, of lower dimension and complexity than the modality data itself. The modality-specific generators interpret these latent samples in order to generate high quality modality data. Without hierarchy, the same multimodal representation must be able to encode and generate the modality-data itself, a more complex task than the former. As shown in Figure C.1 and Table C.1, while the accuracy of the generated data is not significantly affected by hierarchy, the image modality distance decreases significantly with the hierarchical extension. Between the different encoding solutions, we can observe that the Alma solution (MUSE), being based on a POE encoder, is able to learn representation suitable to generate higher-quality samples in comparison with the other solutions. However, the increased generation quality comes at a cost of decreased accuracy. Quite surprisingly,

Table C.1: Results of the hierarchical evaluation of cross-modal accuracy (higher is better), averaged over both modalities, and cross-modal image modality distance (lower is better). Results averaged over 5 independent runs.

(a) Hierarchical models

Model	Accuracy (%)	Image MD
MUSE	90.6 ± 06.7	138.2 ± 35.4
MUSE-0	98.2 ± 01.8	198.0 ± 40.8
MUSE-σ	93.4 ± 05.3	202.4 ± 44.5
MVAE	62.9 ± 31.5	237.0 ± 41.1
MMVAE	91.8 ± 03.3	286.9 ± 59.8

(b) Single-level models

Model	Accuracy (%)	Image MD
MUSE	88.3 ± 08.7	247.4 ± 68.8
MUSE-0	93.0 ± 06.9	272.9 ± 65.8
MUSE-σ	96.5 ± 03.6	260.7 ± 54.6
MVAE	16.5 ± 04.1	150.4 ± 42.2
MMVAE	91.9 ± 09.0	217.0 ± 75.5

the naive concatenation solution (MUSE-0) appears to outperform the more complex nexus solution (MUSE-σ), with a higher average accuracy and lower average image modality distance. We hypothesize that this behavior is due to the low number of modalities ($M = 2$) of the scenario. In fact, the results in Appendix C.3 attest to this fact: MUSE-0 struggles to scale to scenarios with higher number of modalities.

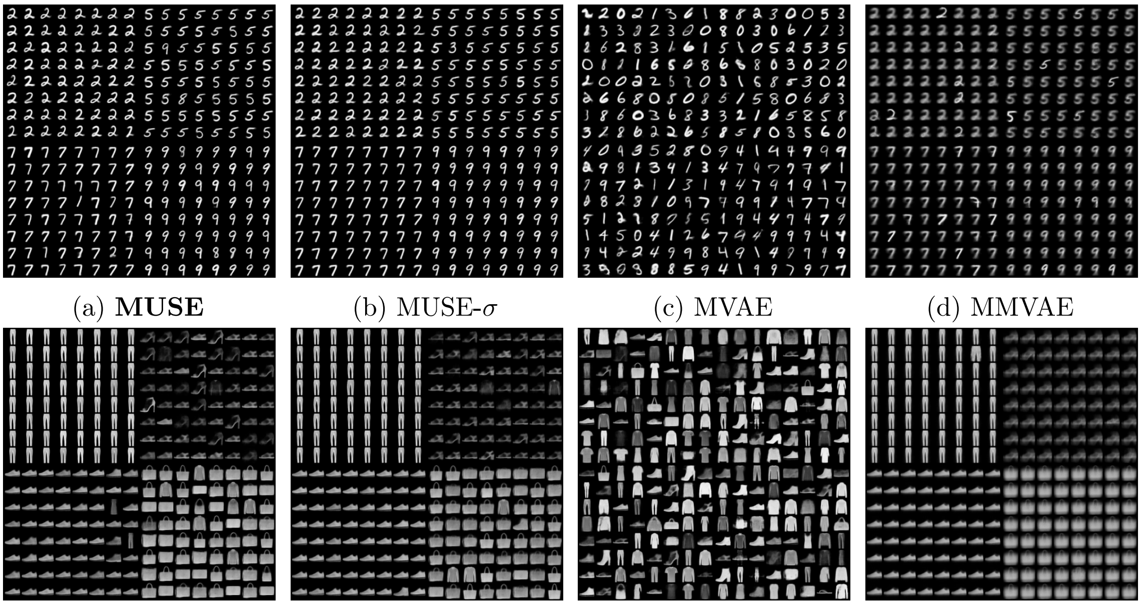
Regarding the MMVAE baseline, a direct comparison of Fig. C.1d and Fig. C.1h shows that the hierarchical extension of the MMVAE is able to generate higher-quality image samples than the single-level version. Such visual inspection seems contrary to the results regarding image modality distance in Table C.1. This seemingly contradiction is due to the computation of the modality distance score: the blurriness in the batches of images generated by MMVAE are interpreted by the auto-encoders as variability. With the hierarchical extension, the images become much higher-quality and the lack of variability becomes evident, as seen in Fig. C.1d. Thus, the image modality distance score increases on average. Nonetheless, despite the hierarchical extension, the generated samples still present high modality distance, suggesting that the MoE solution employed by the model struggles to learn a representation of the modalities.

For the MVAE baseline, the same hierarchical extension results in a significant increase in the accuracy metric, but also in the image modality distance score. The latter increase is explained by the overconfident expert problem of the single-level MVAE: the model learns a multimodal representation that disregards the information provided by the lower-dimensional modality in scenarios with modalities of distinct complexities (784-dimensional images against 10-dimensional labels). As shown in Fig. C.1g, the MVAE model learns a representation that is able to generate high-quality images (low image modality distance) at the cost of low accuracy. By extending MVAE with hierarchy, the imbalance between the modalities decreases, as the difference in dimensionality of their modality-specific representation spaces is smaller than the difference in dimensionality of the original data. The generation procedure of the the hierarchical version MVAE loses the quality provided by the overconfident expert but gains in accuracy, as shown in Fig. C.1c. However, the high variance of accuracy reveals that, despite the hierarchical extension, the CMI generation procedure is still not robust across all target modalities.

C.2 Standard Datasets

We attest the qualitative and quantitative performance of MUSE on two literature-standard datasets: MNIST and FashionMNIST. We evaluate MUSE against the MVAE and MMVAE baselines, regarding single-modality reconstruction accuracy, joint-modality reconstruction accuracy, cross-modal generation accuracy and cross-modal quality score, in addition to standard likelihood based metrics. We employ the same model architectures and training hyperparameters of the previous evaluation.

The quantitative results regarding accuracy and image modality distance for the MNIST and FashionMNIST dataset are presented in Table C.2. All results are averaged over 5 independent runs. In addition, we present image samples generated from label information in Fig. C.2. The results on both datasets show that MUSE is the only model able to encode a multimodal representation capable of generating modality data in the high accuracy and low modality distance regimes, regardless of given single-modality or joint-modality observations. Regarding the different joint-modality encoding schemes, the ALMA solution (MUSE) once again outperforms the other modalities in terms of image modality distance, while both the other two encoding solutions appear to outperform the former on accuracy.



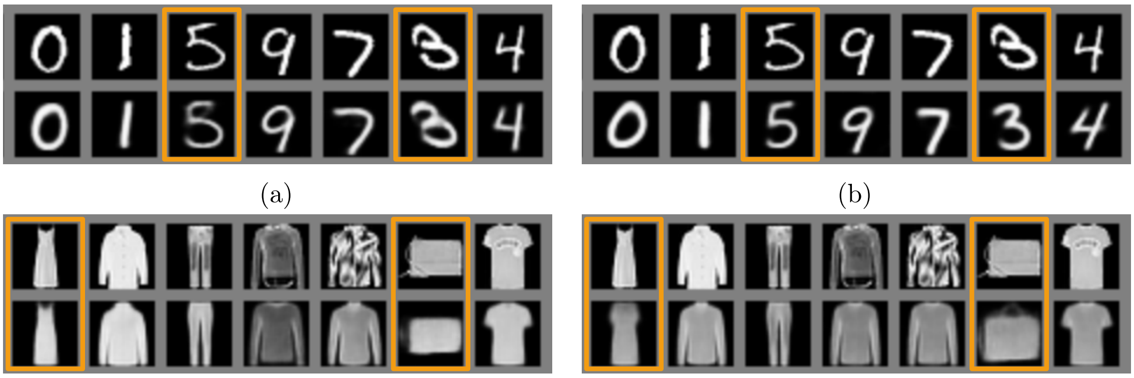
(a) **MUSE** (b) MUSE-σ (c) MVAE (d) MMVAE

(e) **MUSE** (f) MUSE-σ (g) MVAE (h) MMVAE

Figure C.2: Cross-modality image samples considering the MNIST dataset, from labels $\boldsymbol{x}_l = \{$"2", "5", "7", "9"$\}$ (a-d), and considering the FashionMNIST dataset, from labels $\boldsymbol{x}_l = \{$"Trouser", "Sandal", "Sneaker", "Bag"$\}$ (e-h). The MUSE model is the only model able perform CMI and generate samples with high *accuracy* and low *modality distance* (best viewed with zoom).

The results of the MVAE baseline reveal that the PoE solution employed suffers once again from overconfident expert prediction issue: the model is unable to generate semantically coherent modality data (low cross-modal accuracy), as seen in Fig. C.2. Moreover, the minor increase in accuracy from single-modality observations to joint-modality observations suggests that the model is unable to consider the information provided by the two modalities, hinting once again to the overconfident expert issue. Finally, the MMVAE model is also able to generate semantically-coherent modality data, even outperforming MUSE in the FashionMNIST dataset. It does so, however, at the cost of the quality of the generated images, as visually shown in Fig. C.2, having the highest image modality distance results in both datasets.

We can further understand the impact of the hierarchical configuration of MUSE on the generation of modality-specific information by investigating its reconstruction process. Modality-specific information can be reconstructed directly at the lower-level, modality-specific, representation space or be reconstructed from the top-level, multimodal, representation space. In Figure C.3 we present images reconstructed from both representation spaces, using the MUSE-σ model. The samples reveal that the images reconstructed from the contextual multimodal representation $\boldsymbol{z}_\pi$ are more *prototypical* than the samples reconstructed from the modality-specific representation $\boldsymbol{z}_i$. We can understand such abstraction given the hierarchical nature of the representation spaces in MUSE: the modality-specific representations encode an abstraction of high-dimensional modality information, generating low-dimensional codes. Such low-dimensional codes are encoded and generated by the top multimodal representation space, which learns to generate coherent modality-specific codes, providing another layer of abstraction. The abstraction provided by the genera-



(a)

(b)

(c)

(d)

Figure C.3: Images reconstructed from the modality-specific latent space $\boldsymbol{z}_1$ (a, c) and multimodal latent space $\boldsymbol{z}_\pi$ (b, d) of MUSE-σ, presenting the original image data (top row) and the reconstructed data (bottom row). We highlight samples (in orange) where the abstraction provided by the multimodal latent space (b, d) allows the generation of more prototypical information, in comparison with the modality-specific reconstructions (a, c).

tion procedure from the multimodal representation provides a significant advantage to MUSE in comparison with non-hierarchical models: the samples generated accentuate the features that unequivocally define the observed phenomena (in this case, digit class correspondence). This allows MUSE to sample coherent modality-information regardless of the target modality and even in scenarios with a high number of modalities.

C.3 MHD Dataset

We present the results of the different MUSE versions in the cross-modality generation task introduced in Section 5.7.1. We evaluate the cross-modality generation performance for each target modality, as a function of the number of modalities provided to the model. The results are averaged over all possible combinations of provided modalities and 5 independent runs, as shown in Table C.3.

The results show once again that the MUSE models outperform the other baselines in accuracy and modality distance across all target modalities. The original MUSE model, in particular, outperforms all other models in modality distance score, able to generate information regardless of the complexity of the target modality, and performs on par with the other MUSE-based variants regarding accuracy. The results also show a clear distinction between the naive MUSE-0 version and the Nexus MUSE-σ version: the simple concatenation mechanism to merge multimodal information is unsuitable to generate semantically coherent data when only provided with a single modality, hinting at the lack of scalability of the MUSE-0 version.

Table C.2: Results of evaluation in the (a) MNIST and (b) FashionMNIST datasets, regarding the cross-modal generation accuracy and cross-modal generated image modality distance (MD) for different multimodal generative models. Results averaged over 5 independent runs.

(a) MNIST

Model	Single-Modality Accuracy (%)	Joint-Modality Accuracy (%)	Cross-Modal Accuracy (%)	Cross-Modal Image MD
MUSE	92.44 ± 05.94	98.79 ± 01.76	91.91 ± 06.21	278.1 ± 72.4
MUSE-0	97.48 ± 02.54	99.96 ± 00.08	97.38 ± 02.67	365.5 ± 85.9
MUSE-σ	97.03 ± 02.97	99.98 ± 00.02	97.19 ± 02.88	372.7 ± 91.9
MVAE	96.10 ± 03.61	97.09 ± 02.82	13.48 ± 03.31	285.1 ± 61.7
MMVAE	72.39 ± 28.75	-	68.06 ± 30.83	493.2 ± 134.3

(b) FashionMNIST

Model	Single-Modality Accuracy (%)	Joint-Modality Accuracy (%)	Cross-Modal Accuracy (%)	Cross-Modal Image MD
MUSE	77.67 ± 20.73	87.70 ± 10.34	69.60 ± 08.44	97.9 ± 27.8
MUSE-0	84.14 ± 15.87	89.49 ± 10.64	74.59 ± 03.22	120.7 ± 34.1
MUSE-σ	82.00 ± 18.01	88.59 ± 11.43	74.09 ± 02.96	120.2 ± 34.1
MVAE	85.24 ± 14.47	87.79 ± 11.99	20.91 ± 07.95	112.8 ± 40.2
MMVAE	85.46 ± 14.50	-	83.46 ± 06.54	133.4 ± 52.7

Table C.3: Evaluation in the MHD dataset, as a function of the number of the observed modalities provided to the models and the target cross-modal generated modality (I = Image, T = Trajectory, S = Sound, L = Label). Results averaged over all possible combinations of input modalities and over 5 independent runs.

		Accuracy (%)		Modality Distance	
Model	Target	1 Modality	3 Modalities	1 Modality	3 Modalities
MUSE	I	78.5 ± 13.9	98.7 ± 00.4	89.4 ± 23.6	100.2 ± 22.4
	T	73.9 ± 14.0	95.9 ± 00.9	265.0 ± 106.7	215.9 ± 90.1
	S	77.6 ± 09.7	93.1 ± 00.9	3354 ± 621	4105 ± 721
	L	72.0 ± 08.3	95.9 ± 01.1	NA	NA
MUSE-0	I	49.0 ± 39.3	99.3 ± 00.4	265.2 ± 144.6	94.0 ± 19.2
	T	46.1 ± 37.3	75.1 ± 06.8	625.4 ± 341.9	539.0 ± 220.1
	S	64.8 ± 10.1	60.0 ± 07.3	15452 ± 1152	15778 ± 1196
	L	67.8 ± 04.3	99.2 ± 00.3	NA	NA
MUSE-σ	I	84.9 ± 12.4	99.0 ± 00.2	203.9 ± 86.3	76.6 ± 02.8
	T	81.0 ± 10.4	93.8 ± 01.6	618.3 ± 264.0	444.4 ± 36.7
	S	77.2 ± 08.9	94.4 ± 04.2	22393 ± 1659	20239 ± 3411
	L	82.0 ± 06.6	96.9 ± 00.5	NA	NA
MVAE	I	28.6 ± 05.2	80.9 ± 07.2	228.4 ± 61.8	201.3 ± 45.2
	T	13.7 ± 04.6	17.8 ± 03.7	399.3 ± 179.1	391.0 ± 178.7
	S	33.6 ± 14.2	88.6 ± 09.7	6608 ± 1471	8133 ± 1751
	L	23.4 ± 13.4	39.9 ± 07.7	NA	NA
MMVAE	I	66.1 ± 39.8	–	236.9 ± 62.7	–
	T	63.8 ± 38.1	–	547.8 ± 235.4	–
	S	70.4 ± 05.4	–	14998 ± 1325	–
	L	66.0 ± 39.6	–	NA	NA

Appendix D

Ablation Study of GMC

We perform a thorough ablation study on the hyperparameters of GMC using the setup from Section 6.3.2 on the MHD dataset. In particular, we investigate:

1. the robustness of the GMC framework when varying the temperature parameter τ;

2. the performance of GMC with different dimensionalities of the low-level representations $\boldsymbol{h} \in \mathbb{R}^d$;

3. the performance of GMC with different dimensionalities of the high-level, shared latent representations $\boldsymbol{z} \in \mathbb{R}^s$;

4. the performance of GMC with a modified loss $\mathcal{L}^*_{\text{GMC}}$ that only uses complete observations as negative pairs.

In all experiments we report both classification results and alignment scores.

D.1 Temperature parameter

We study the performance of GMC when varying $\tau \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ (see Equation 6.1). We present the classification results and alignment scores in Table D.1 and Table D.2, respectively. We observe that classification results are rather robust to different values of temperature, while increasing the temperature seems to have slightly negative effect on the geometry of the representations. For example, in Table D.2, we observe that for $\tau = 0.5$ the trajectory representations $\boldsymbol{z}_3$ are worse aligned with $\boldsymbol{z}_{1:4}$.

D.2 Dimensionality of the low-level representations

We vary the dimension of the low-level representations space $d = \{32, 64, 128\}$ and present the resulting classification results and alignment scores in Table D.3 and Table D.4, respectively. The differences in classification results across different dimensions are covered by the margin of error, indicating the robustness of GMC to different sizes of the low-level representations. We observe similar stability of the DCA scores in Table D.2 with minor variations in the geometric alignment for the sound modality z_2 which benefits from the larger low-level representation space.

Table D.1: Performance of GMC with different temperature values τ (Equation 6.1) in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$\tau = 0.05$	$\tau = 0.1$ (Default)	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.5$
Complete Observations	99.99 ± 0.01	100.00 ± 0.00	99.99 ± 0.01	$99.97 \pm 3\text{e}{-5}$	99.96 ± 0.01
Image Observations	99.78 ± 0.02	99.75 ± 0.03	99.84 ± 0.03	99.80 ± 0.04	99.89 ± 0.03
Sound Observations	93.55 ± 0.22	93.04 ± 0.45	91.98 ± 0.29	91.87 ± 0.58	95.01 ± 0.38
Trajectory Observations	99.94 ± 0.01	99.96 ± 0.02	99.97 ± 0.02	99.96 ± 0.01	99.80 ± 0.20
Label Observations	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

D.3 Dimensionality of the high-level latent representations

We repeat a similar evaluation for the dimension of the latent space $s = \{32, 64, 128\}$ and present the classification and alignment scores in Table D.5 and Table D.6, respectively. We observe that GMC is robust to changes in s both in terms of performance and geometric alignment.

D.4 Loss function

We consider an ablated version of the loss function, $\mathcal{L}^*_{\text{GMC}}$, that considers only complete-observations as negative pairs, i.e. $\Omega^*(i) = \sum_{i \neq j} s_{1:M,1:M}(i,j)$ for $j = 1, \ldots, B$ where B is the size of the mini-batch. Due to the symmetry of negative pairs in this setting, we only consider positive pairs $(z^i_m, z^i_{1:M})$. We present the classification results and alignment scores in Table D.7 and Table D.8, respectively. The results in Table D.7 highlight the importance of the contrasting the complete representations to learn a robust representation suitable for downstream tasks as we observe minimal variation in classification accuracy when considering different loss. However, we observe worse geometric alignment when using $\mathcal{L}^*_{\text{MC}}$ loss during training of GMC. This suggests that contrasting among individual modalities is beneficial for geometrical alignment of the representations.

Table D.2: Alignment scores A_{DCA} obtained on GMC representations when trained with different temperature values τ (Equation equation 6.1) in the MHD dataset, evaluating the geometric alignment of complete representations $\boldsymbol{z}_{1:4}$ and modality-specific ones $\boldsymbol{z}_1, \ldots \boldsymbol{z}_4$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	$\tau = 0.05$	$\tau = 0.1$ (Default)	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.5$
Complete ($\boldsymbol{z}_{1:4}$)	Image ($\boldsymbol{z}_1$)	0.96 ± 0.02	0.96 ± 0.02	0.93 ± 0.01	0.92 ± 0.00	0.89 ± 0.02
Complete ($\boldsymbol{z}_{1:4}$)	Sound ($\boldsymbol{z}_2$)	0.95 ± 0.02	0.87 ± 0.16	0.96 ± 0.02	0.99 ± 0.00	0.87 ± 0.04
Complete ($\boldsymbol{z}_{1:4}$)	Trajectory ($\boldsymbol{z}_3$)	0.96 ± 0.02	0.86 ± 0.05	0.90 ± 0.03	0.92 ± 0.00	0.64 ± 0.11
Complete ($\boldsymbol{z}_{1:4}$)	Label ($\boldsymbol{z}_4$)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.02

Table D.3: Performance of GMC with different values of low-level representation dimensionality $\boldsymbol{h} \in \mathbb{R}^d$ in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete Observations	99.99 ± 0.01	100.00 ± 0.00	99.99 ± 0.01
Image Observations	99.75 ± 0.04	99.75 ± 0.03	99.72 ± 0.07
Sound Observations	93.31 ± 0.41	93.04 ± 0.45	93.34 ± 0.51
Trajectory Observations	99.96 ± 0.01	99.96 ± 0.02	99.96 ± 0.01
Label Observations	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table D.4: Alignment scores A_{DCA} obtained on GMC representations when varying the dimension of low-level representations $\boldsymbol{h} \in \mathbb{R}^d$ in the MHD dataset, evaluating the geometric alignment of complete representations $\boldsymbol{z}_{1:4}$ and modality-specific ones $\boldsymbol{z}_1, \dots \boldsymbol{z}_4$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete ($\boldsymbol{z}_{1:4}$)	Image ($\boldsymbol{z}_1$)	0.91 ± 0.04	0.96 ± 0.02	0.92 ± 0.04
Complete ($\boldsymbol{z}_{1:4}$)	Sound ($\boldsymbol{z}_2$)	0.77 ± 0.17	0.87 ± 0.16	0.96 ± 0.04
Complete ($\boldsymbol{z}_{1:4}$)	Trajectory ($\boldsymbol{z}_3$)	0.86 ± 0.04	0.86 ± 0.05	0.86 ± 0.07
Complete ($\boldsymbol{z}_{1:4}$)	Label ($\boldsymbol{z}_4$)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Table D.5: Performance of GMC with different values of the dimensionality of the high-level latent representation $\boldsymbol{z} \in \mathbb{R}^s$ in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete Observations	99.99 ± 0.01	100.00 ± 0.00	99.99 ± 0.01
Image Observations	99.75 ± 0.04	99.75 ± 0.03	99.72 ± 0.07
Sound Observations	93.31 ± 0.41	93.04 ± 0.45	93.34 ± 0.51
Trajectory Observations	99.96 ± 0.01	99.96 ± 0.02	99.96 ± 0.01
Label Observations	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table D.6: Alignment scores A_{DCA} obtained on GMC representations when varying the dimension of the high-level latent representations $\boldsymbol{z} \in \mathbb{R}^d$ in the MHD dataset, evaluating the geometric alignment of complete representations $\boldsymbol{z}_{1:4}$ and modality-specific ones $\boldsymbol{z}_1, \dots \boldsymbol{z}_4$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete ($\boldsymbol{z}_{1:4}$)	Image ($\boldsymbol{z}_1$)	0.93 ± 0.03	0.96 ± 0.02	0.91 ± 0.03
Complete ($\boldsymbol{z}_{1:4}$)	Sound ($\boldsymbol{z}_2$)	0.89 ± 0.01	0.87 ± 0.16	0.86 ± 0.19
Complete ($\boldsymbol{z}_{1:4}$)	Trajectory ($\boldsymbol{z}_3$)	0.81 ± 0.03	0.86 ± 0.05	0.88 ± 0.06
Complete ($\boldsymbol{z}_{1:4}$)	Label ($\boldsymbol{z}_4$)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Table D.7: Performance of GMC with different loss functions in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$\mathcal{L}_{\text{MC}}$ (Default)	$\mathcal{L}^*_{\text{MC}}$
Complete Observations	100.00 ± 0.00	99.97 ± 0.02
Image Observations	99.75 ± 0.03	99.87 ± 0.01
Sound Observations	93.04 ± 0.45	92.79 ± 0.24
Trajectory Observations	99.96 ± 0.02	99.98 ± 0.01
Label Observations	100.00 ± 0.00	100.00 ± 0.00

Table D.8: Alignment scores A_{DCA} obtained on GMC representations when trained different loss functions in the MHD dataset, evaluating the geometric alignment of complete representations $\boldsymbol{z}_{1:4}$ and modality-specific ones $\boldsymbol{z}_1, \dots \boldsymbol{z}_4$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	$\mathcal{L}_{\text{MC}}$ (Default)	$\mathcal{L}^*_{\text{MC}}$
Complete ($\boldsymbol{z}_{1:4}$)	Image ($\boldsymbol{z}_1$)	0.96 ± 0.02	0.80 ± 0.02
Complete ($\boldsymbol{z}_{1:4}$)	Sound ($\boldsymbol{z}_2$)	0.87 ± 0.16	0.27 ± 0.14
Complete ($\boldsymbol{z}_{1:4}$)	Trajectory ($\boldsymbol{z}_3$)	0.86 ± 0.05	0.86 ± 0.03
Complete ($\boldsymbol{z}_{1:4}$)	Label ($\boldsymbol{z}_4$)	1.00 ± 0.00	0.24 ± 0.10

Appendix E

Additional Results of MARO

In this section, we display the complete experimental results of Section 8.3. Our main results are presented in Sec. E.1; the results of our ablation study are presented in Sec. E.2. In Sec. E.2.1, we display the results for the *Switch* baseline. In Sec. E.3 we display a set of figures that illustrates the predictions made by the predictive model. In all plots, alongside scalar mean values, we report the 95% bootstrapped confidence interval.

E.1 Main Experimental Results

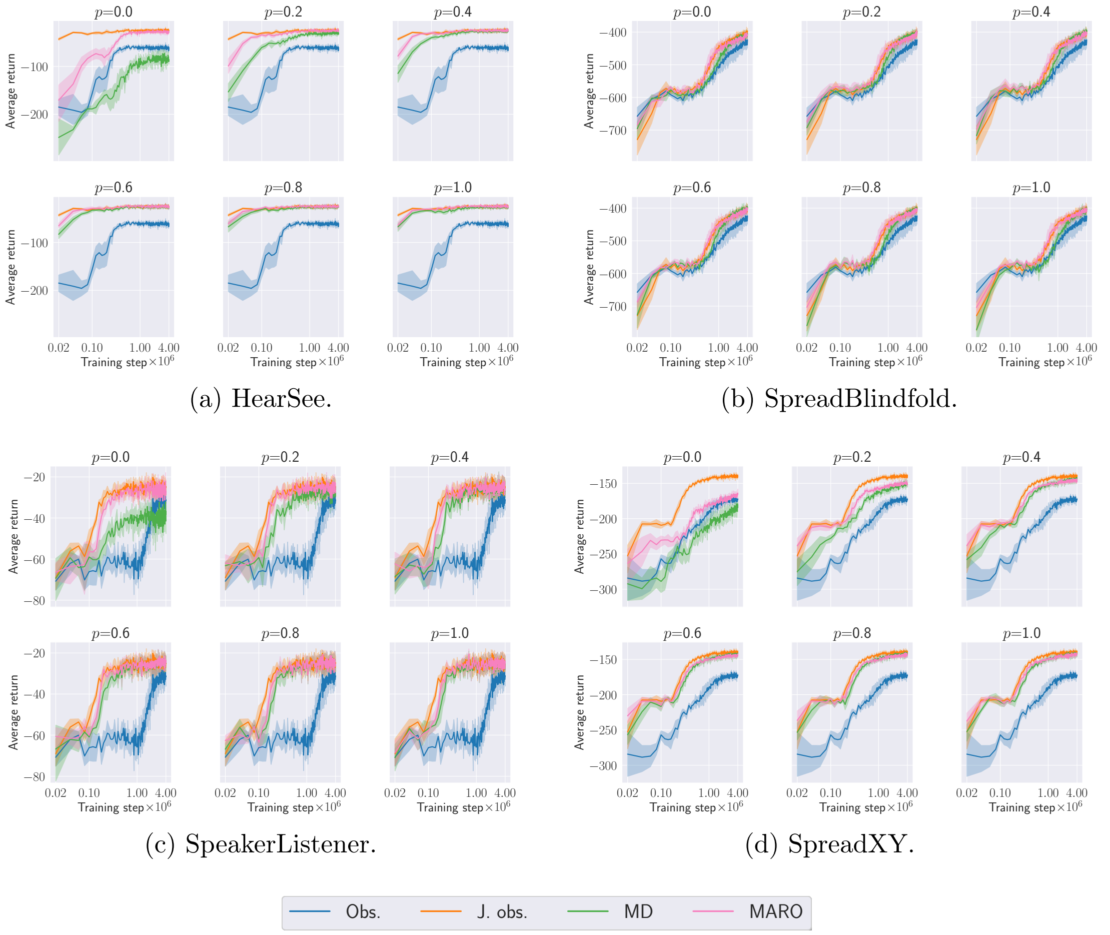
e present the complete experimental results of all approaches across all environments and algorithms: in Figures. E.1 and E.2 we show the performance in all environments of all approaches during training, considering different communication levels p, using the IQL and QMIX algorithms, respectively; in Figures. E.3 and E.4 we show the performance in all environments of all approaches during training, considering $p \sim \mathcal{U}(0,1)$, using the IQL and QMIX algorithms, respectively; in Figures. E.5 and E.6 we show the performance in all environments of all approaches during execution, considering different communication levels p, using the IQL and QMIX algorithms, respectively.

E.2 Ablation Study

In this section, we present the complete experimental results of the ablation study of MARO across all environments and algorithms: in Figures. E.7 and E.8 we show the performance in all environments of MARO and the ablated versions during training, considering different communication levels p, using the IQL and QMIX algorithms, respectively; in Figures. E.9 and E.10 we show the performance in all environments of MARO and the ablated versions during training, considering $p \sim \mathcal{U}(0,1)$, using the IQL and QMIX algorithms, respectively; in Figures. E.11 and E.11 we show the performance in all environments of MARO and the ablated versions during execution, considering different communication levels p, using the IQL and QMIX algorithms, respectively; in Figures. E.13 and E.14 we show the performance in all environments of different sampling methods for the training scheme of MARO, using the IQL and QMIX algorithms, respectively.

E.2.1 Switch Baseline

In Figure E.15 we present the experimental results that compare MARO against the *Switch* baseline, introduced in Section 8.3.4.2, in the HearSee environment. The Switch baseline



(a) HearSee.

(b) SpreadBlindfold.

(c) SpeakerListener.

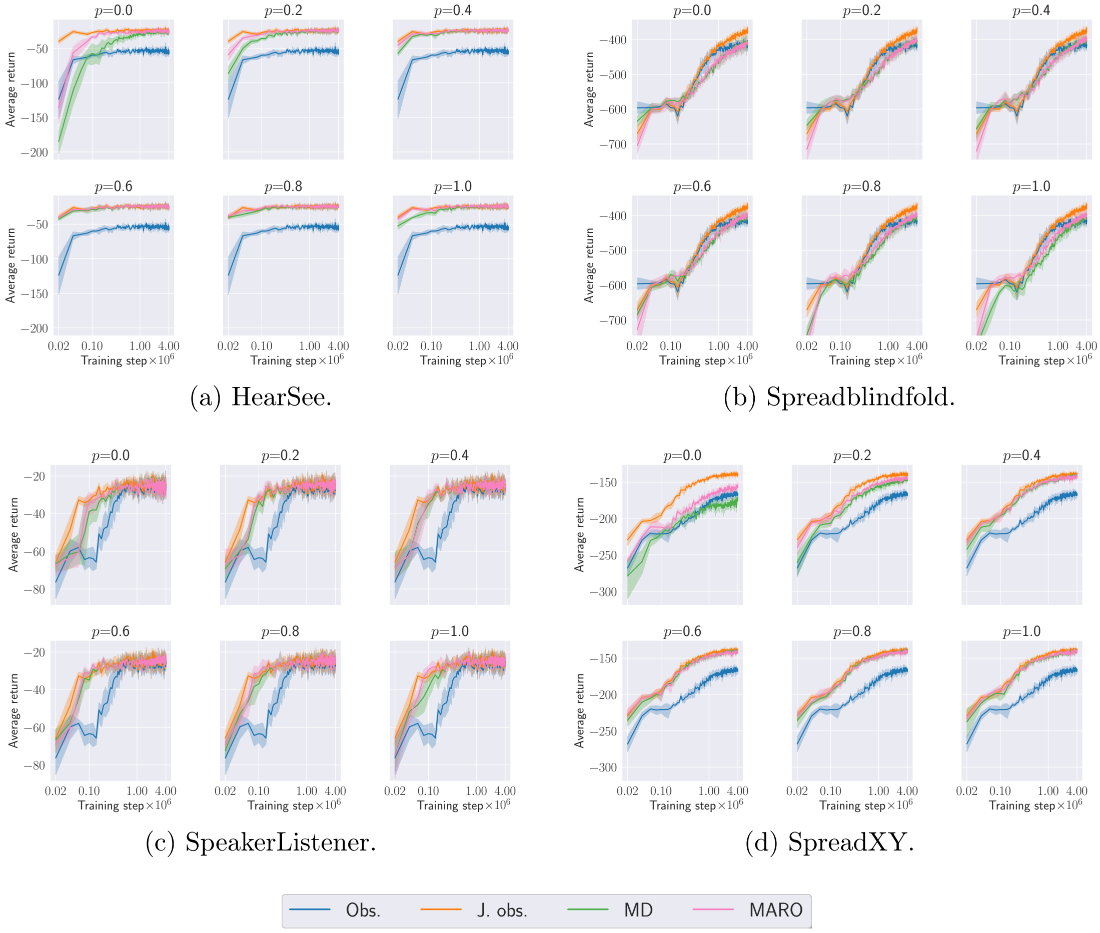
(d) SpreadXY.

Figure E.1: Average episodic returns during training with 95% bootstrapped confidence interval for different communication levels p, for all approaches and environments, using the IQL algorithm.

selects actions using two controllers: one that receives the joint observation, used when communication is allowed, and another that receives only the local observation, used otherwise.

E.3 Multi-agent Trajectory Prediction

We display, in Figures. E.16, E.17 and E.18, an illustration of the trajectory predictions made by the predictive model from the perspective of each of the agents. The plots are computed, at each timestep and from the perspective of each agent, by computing the estimated trajectories of all agents for the next 4 timesteps. The 4-step ahead predictions are entirely computed using estimated quantities, i.e., real observations are not incorporated into the predictions and the predictive model works in a fully auto-regressive manner.



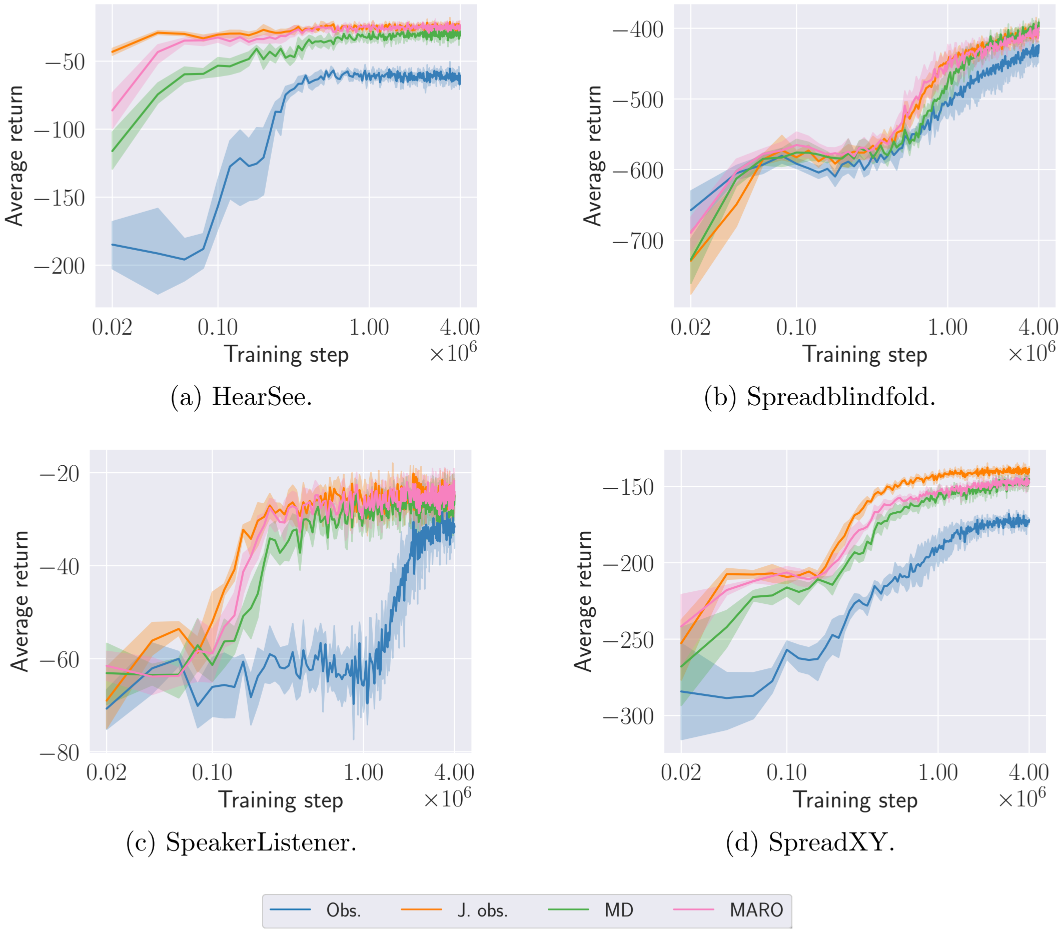
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.2: Average episodic returns during training with 95% bootstrapped confidence interval for different communication levels p, for all approaches and environments, using the QMIX algorithm.



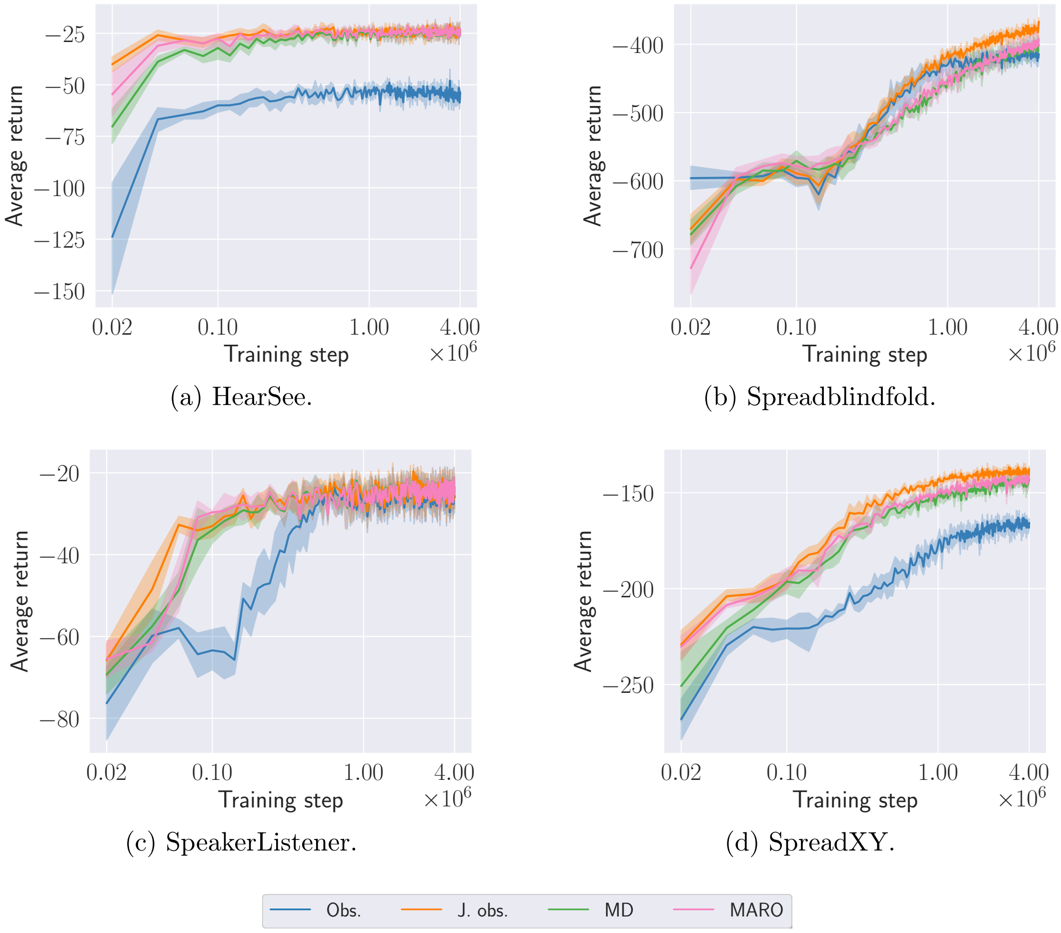
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.3: Average episodic returns during training with 95% bootstrapped confidence interval for $p \sim \mathcal{U}(0, 1)$, for all approaches and environments, using the IQL algorithm.



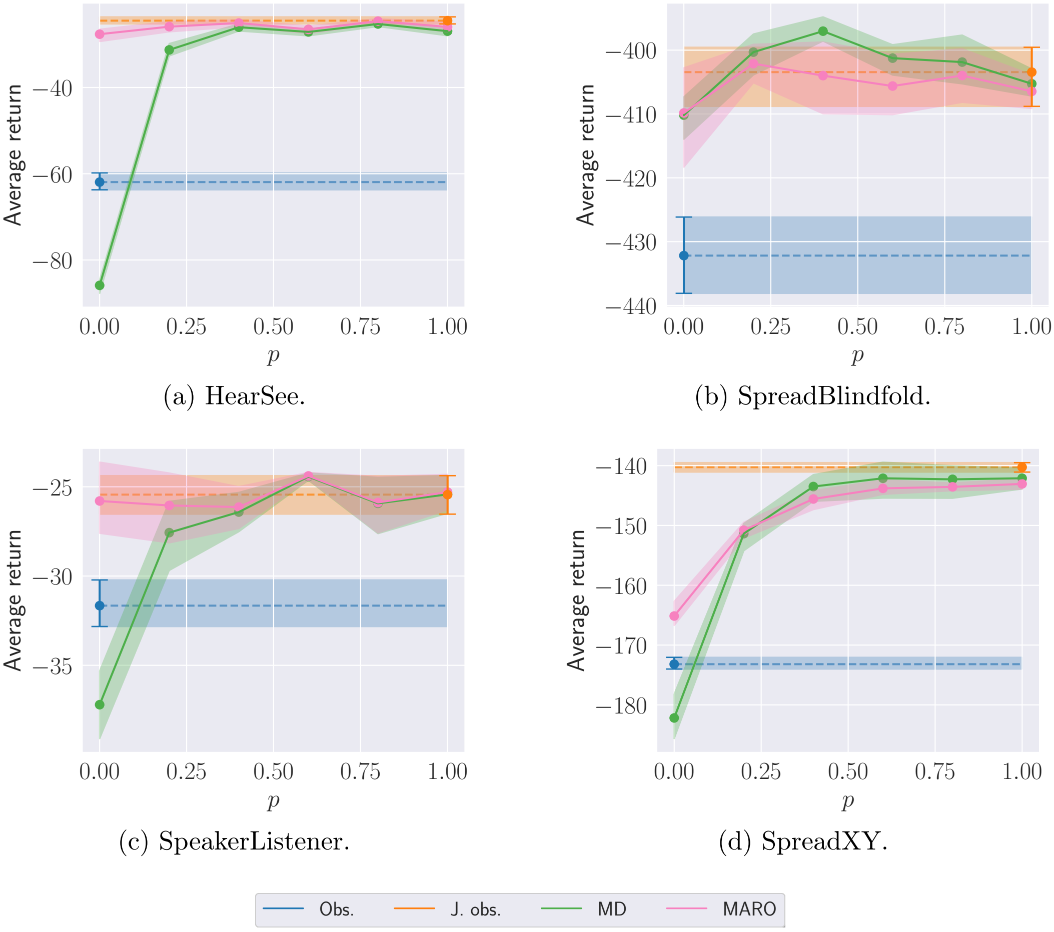
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.4: Average episodic returns during training with 95% bootstrapped confidence interval for $p \sim \mathcal{U}(0, 1)$, for all approaches and environments, using the QMIX algorithm.



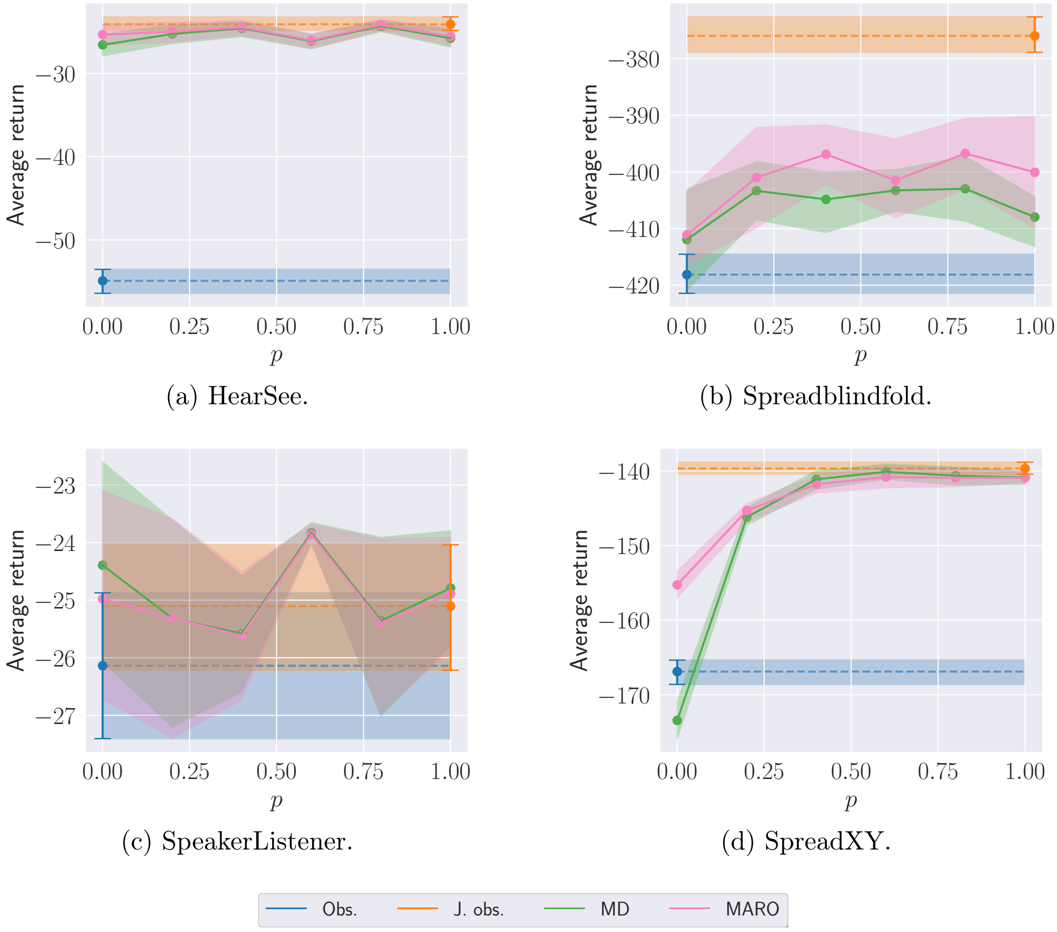
(a) HearSee.

(b) SpreadBlindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.5: Average episodic returns with 95% bootstrapped confidence interval for different communication levels p at execution time, for all approaches across different environments, using the IQL algorithm.



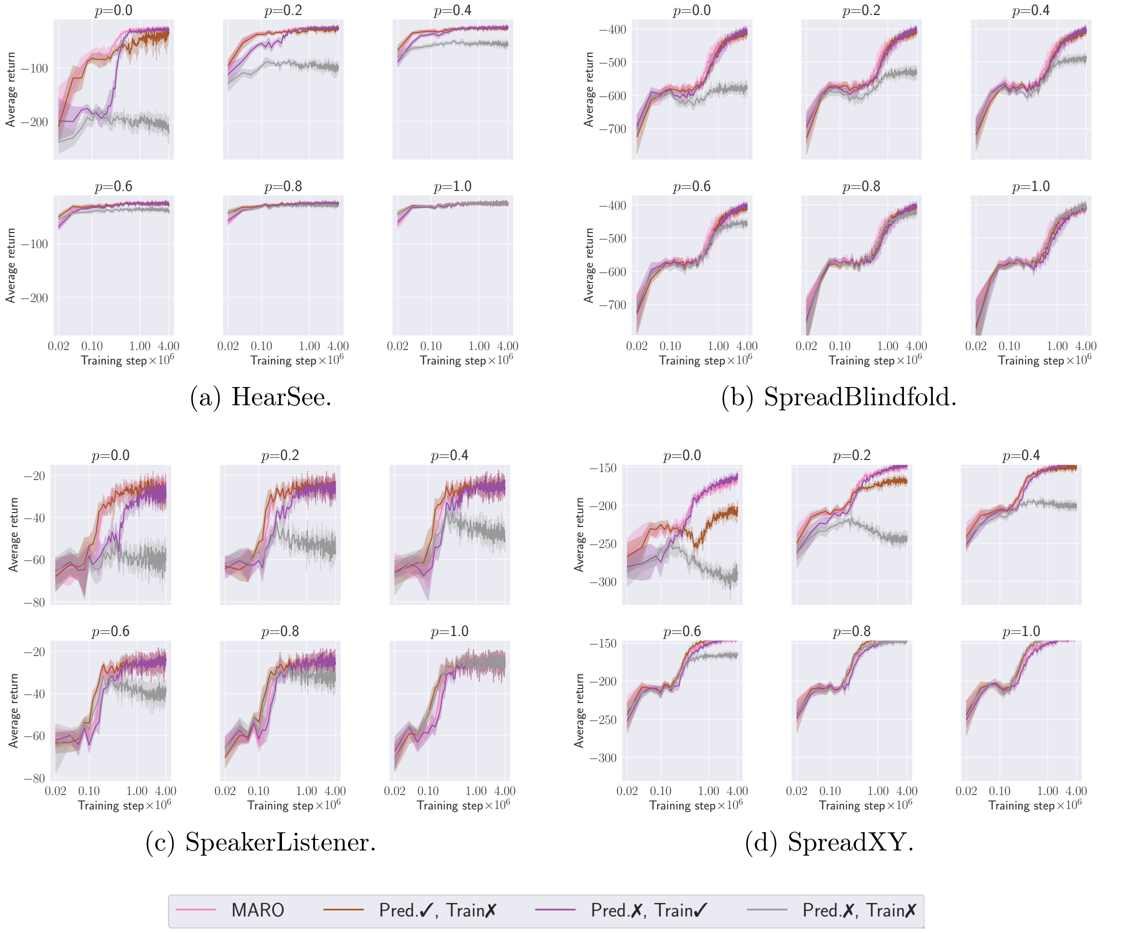
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.6: Average episodic returns with 95% bootstrapped confidence interval for different communication levels p at execution time, for all approaches across different environments, using the QMIX algorithm.



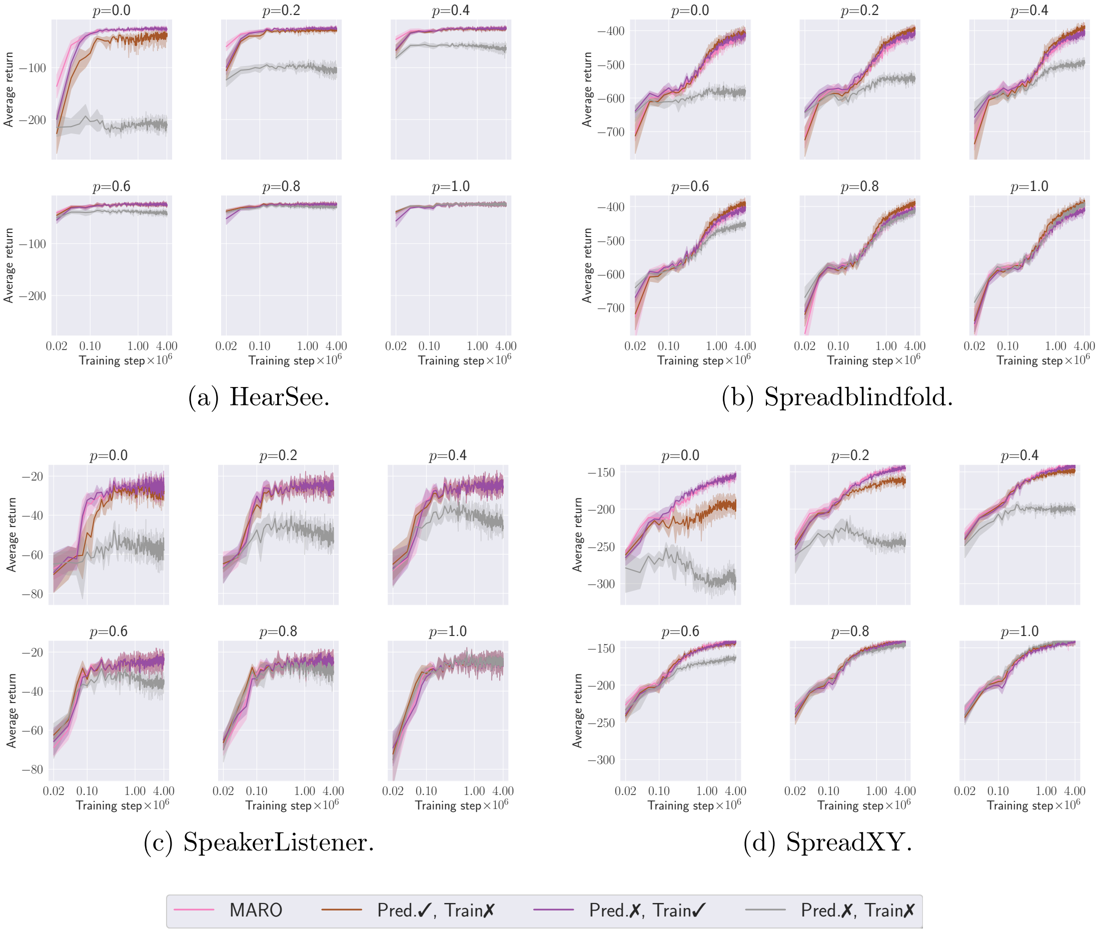
(a) HearSee.

(b) SpreadBlindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.7: Average episodic returns during training with 95% bootstrapped confidence interval for different communication levels p, for MARO and the ablated versions across all environments, using the IQL algorithm.



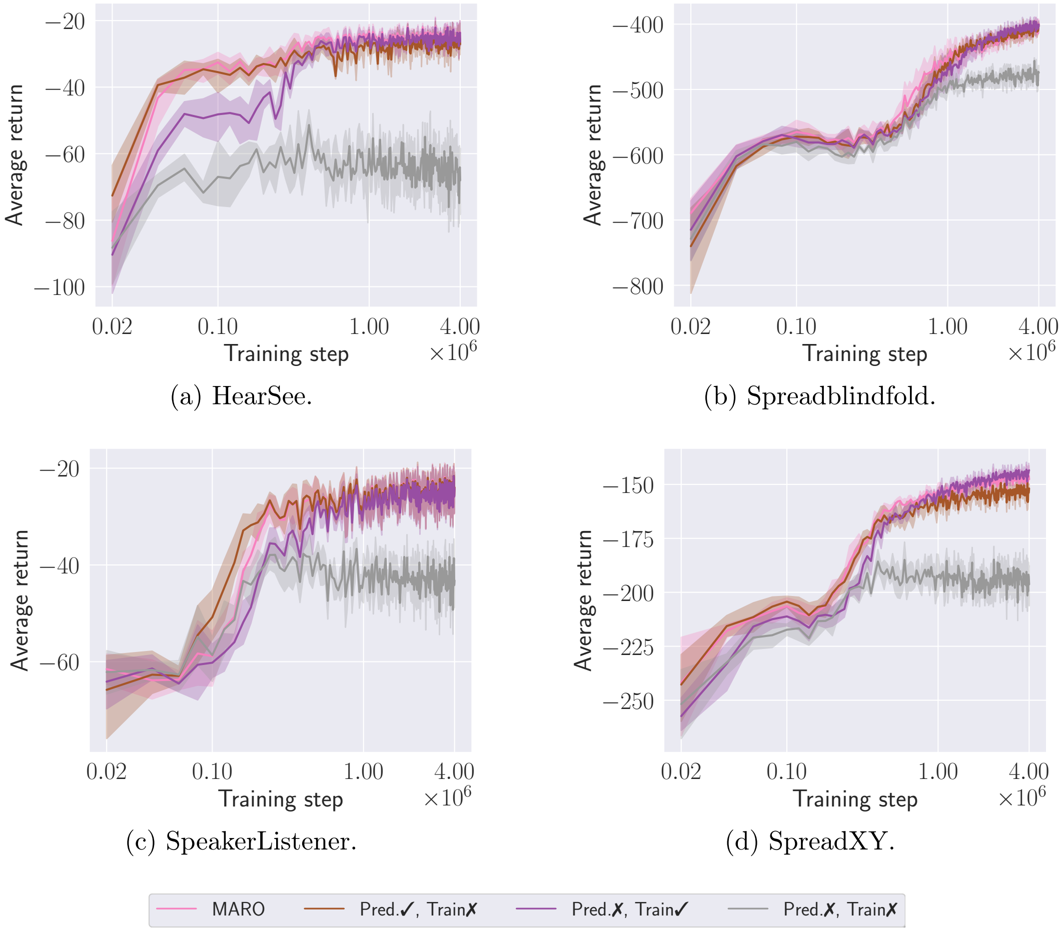
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.8: Average episodic returns during training with 95% bootstrapped confidence interval for different communication levels p, for MARO and the ablated versions across all environments, using the QMIX algorithm.



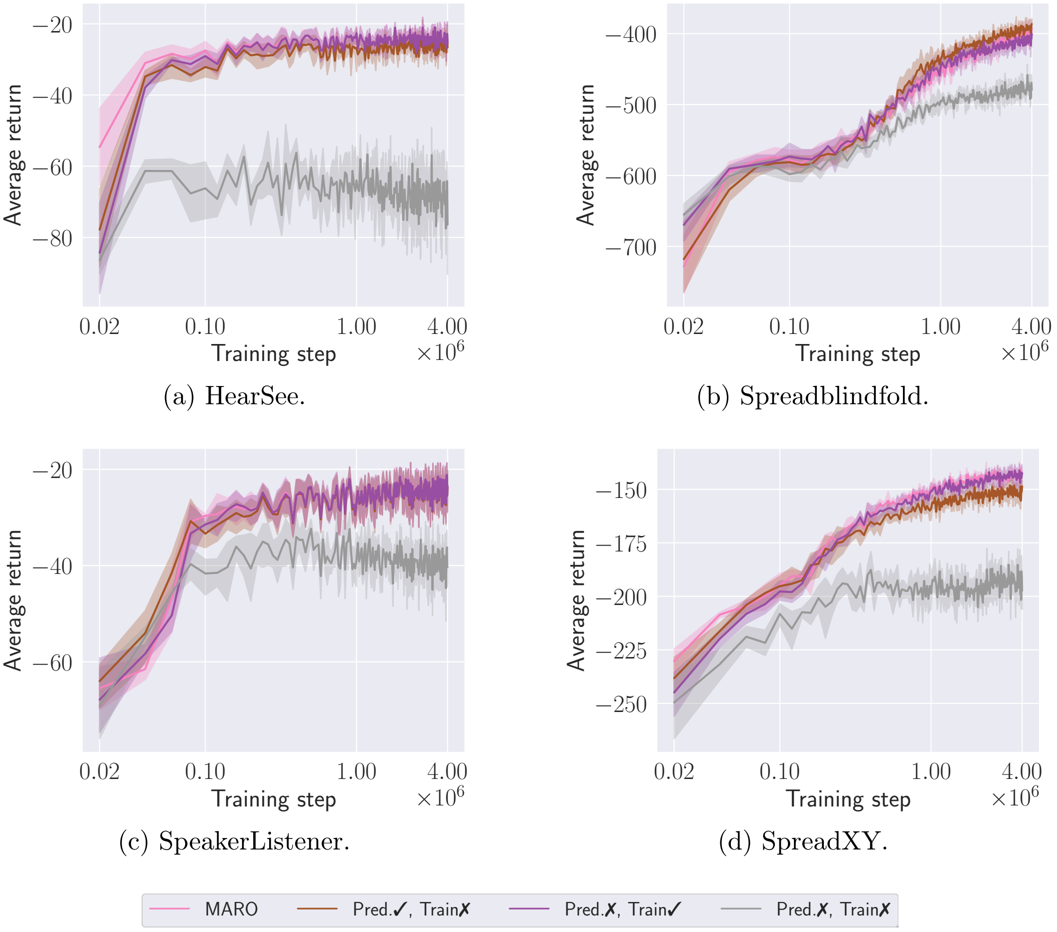
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.9: Average episodic returns during training with 95% bootstrapped confidence interval for $p \sim \mathcal{U}(0,1)$, for MARO and the ablated versions across all environments, using the IQL algorithm.



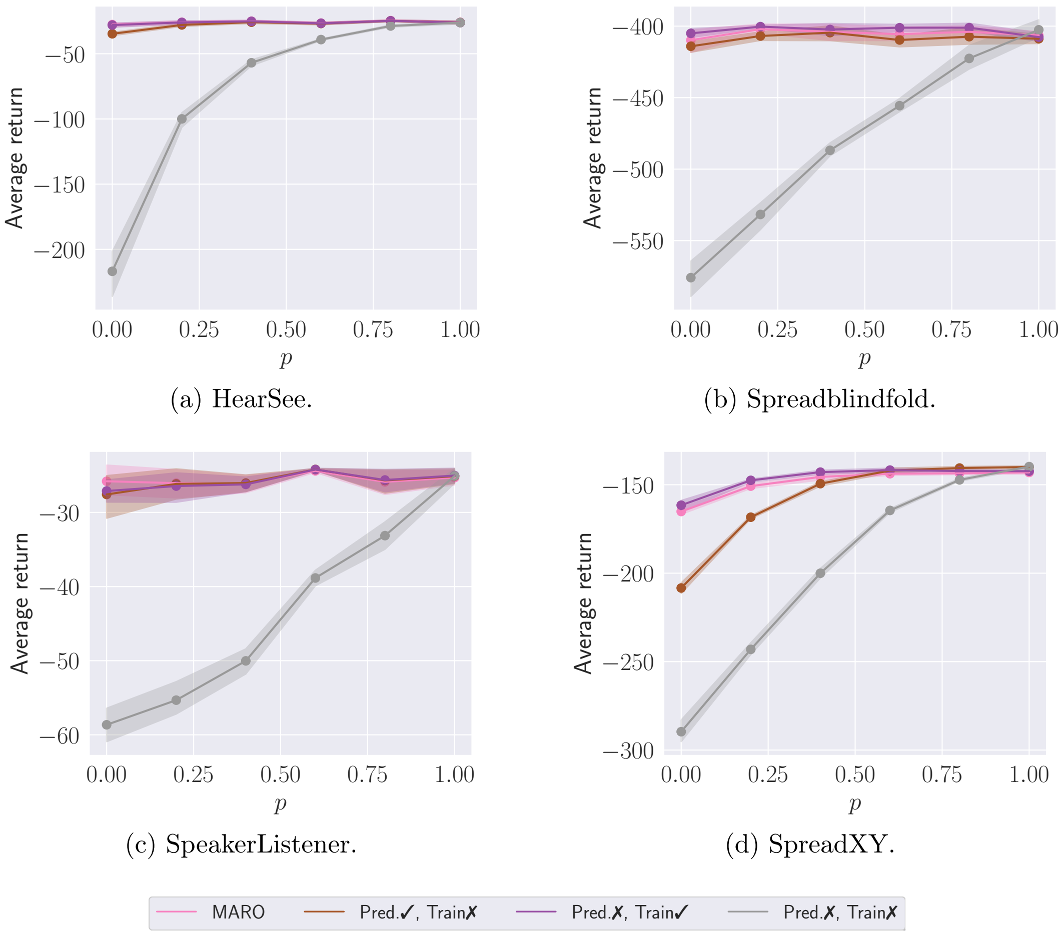
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.10: Average episodic returns during training with 95% bootstrapped confidence interval for $p \sim \mathcal{U}(0, 1)$, for MARO and the ablated versions across environments, using the QMIX algorithm.



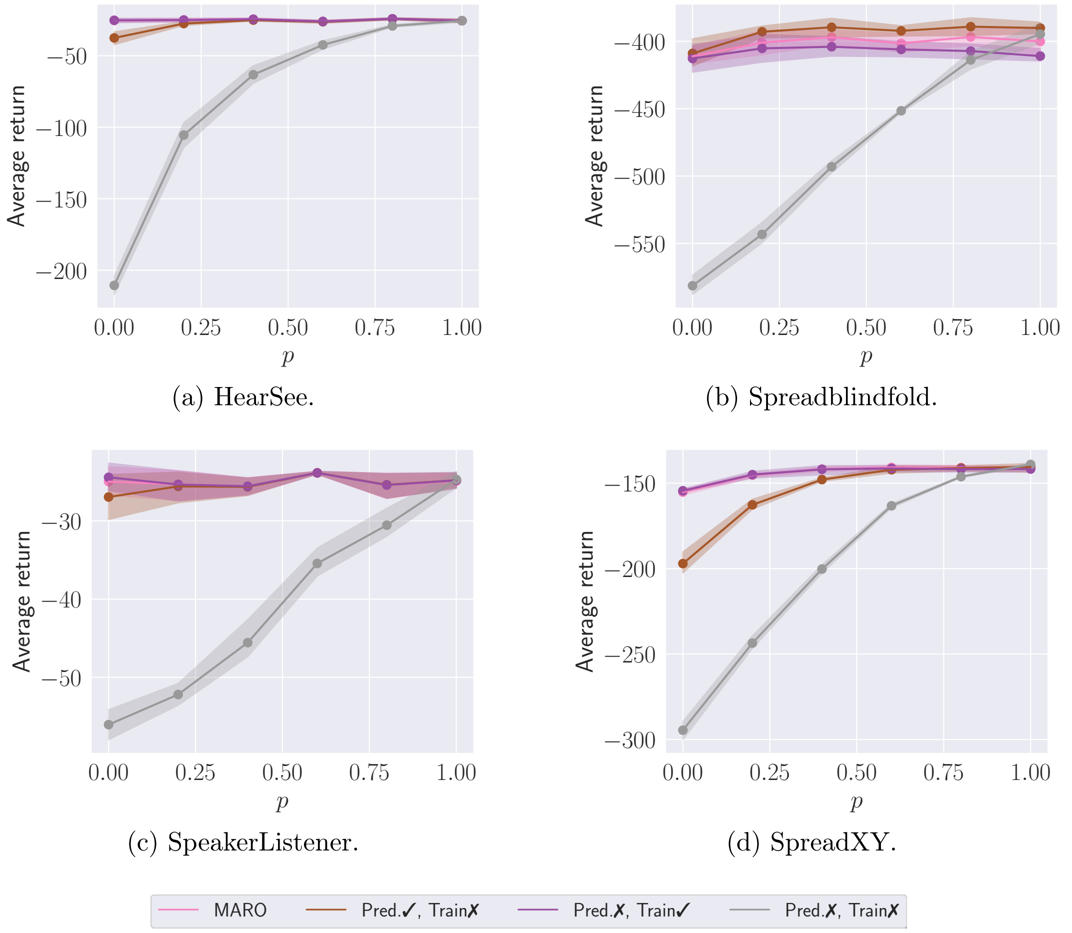
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.11: Average episodic returns with 95% bootstrapped confidence interval for different communication levels p at execution time, for MARO and the ablated versions across different environments, using the IQL algorithm.



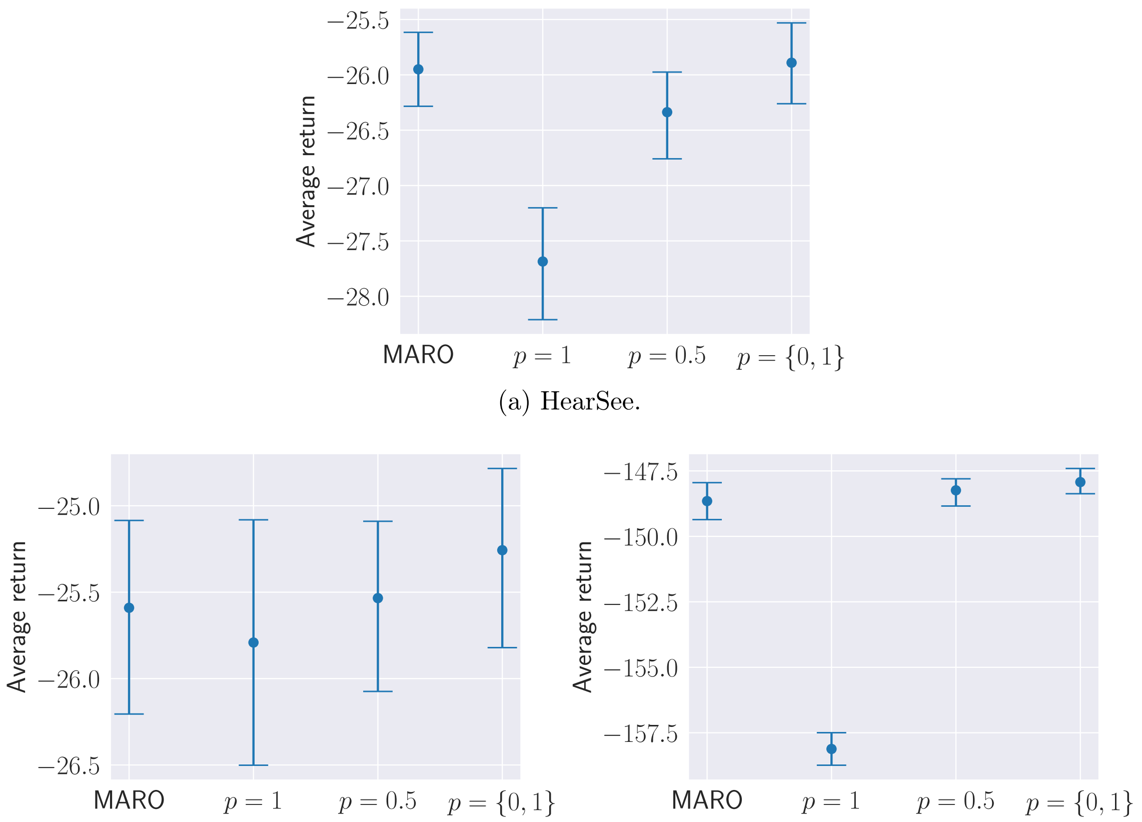
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.12: Average episodic returns with 95% bootstrapped confidence interval for different communication levels p at execution time, for MARO and the ablated versions across different environments, using the QMIX algorithm.

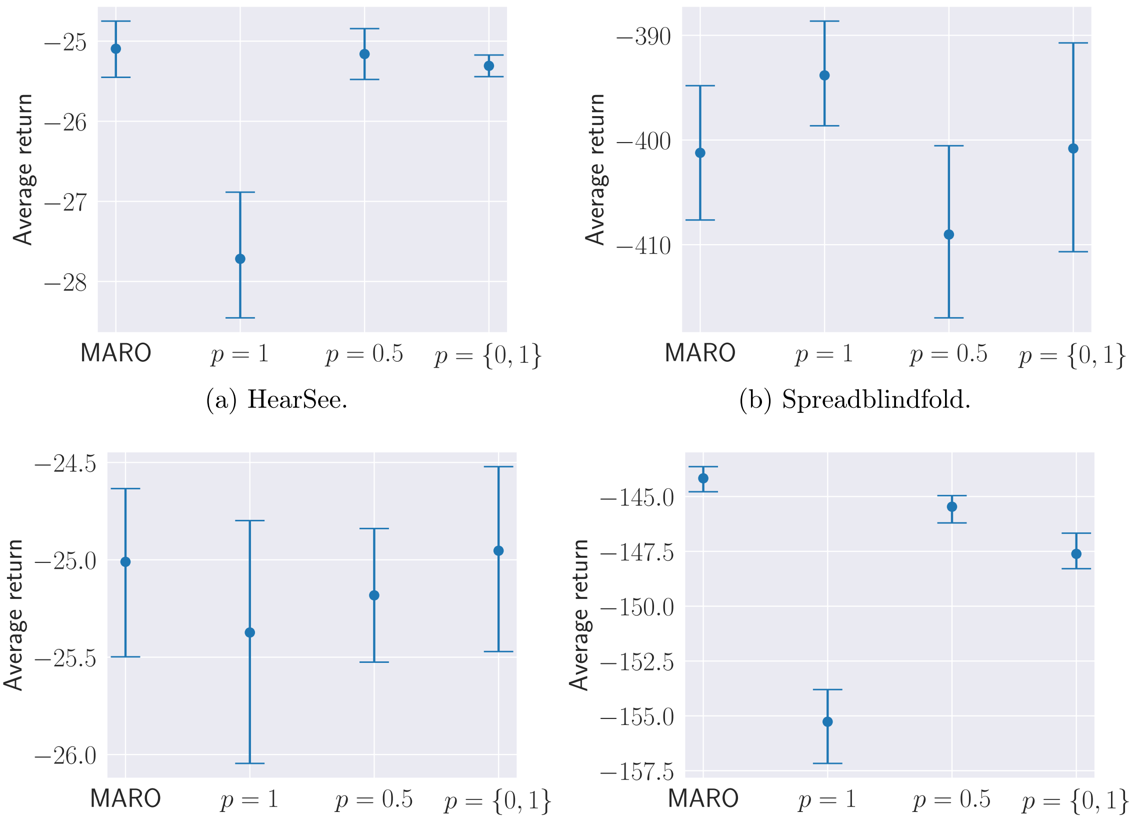


(a) HearSee.

(b) SpeakerListener.

(c) SpreadXY.

Figure E.13: Average episodic returns at execution time with 95% bootstrapped confidence interval for $p \sim \mathcal{U}(0, 1)$, for different sampling methods of the training scheme of MARO across environments, using the IQL algorithm.



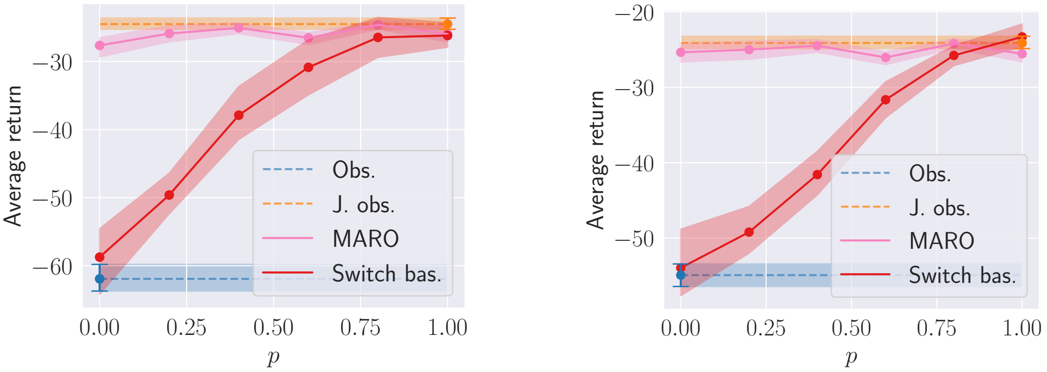
(a) HearSee.

(b) Spreadblindfold.

(c) SpeakerListener.

(d) SpreadXY.

Figure E.14: Average episodic returns at execution time with 95% bootstrapped confidence interval for $p \sim \mathcal{U}(0,1)$, for different sampling methods of the training scheme of MARO across environments, using the QMIX algorithm.



(a) IQL.

(b) QMIX.

Figure E.15: Average episodic returns with 95% bootstrapped confidence interval for different communication levels p at execution time for MARO and the Switch baseline in the HearSee environment.

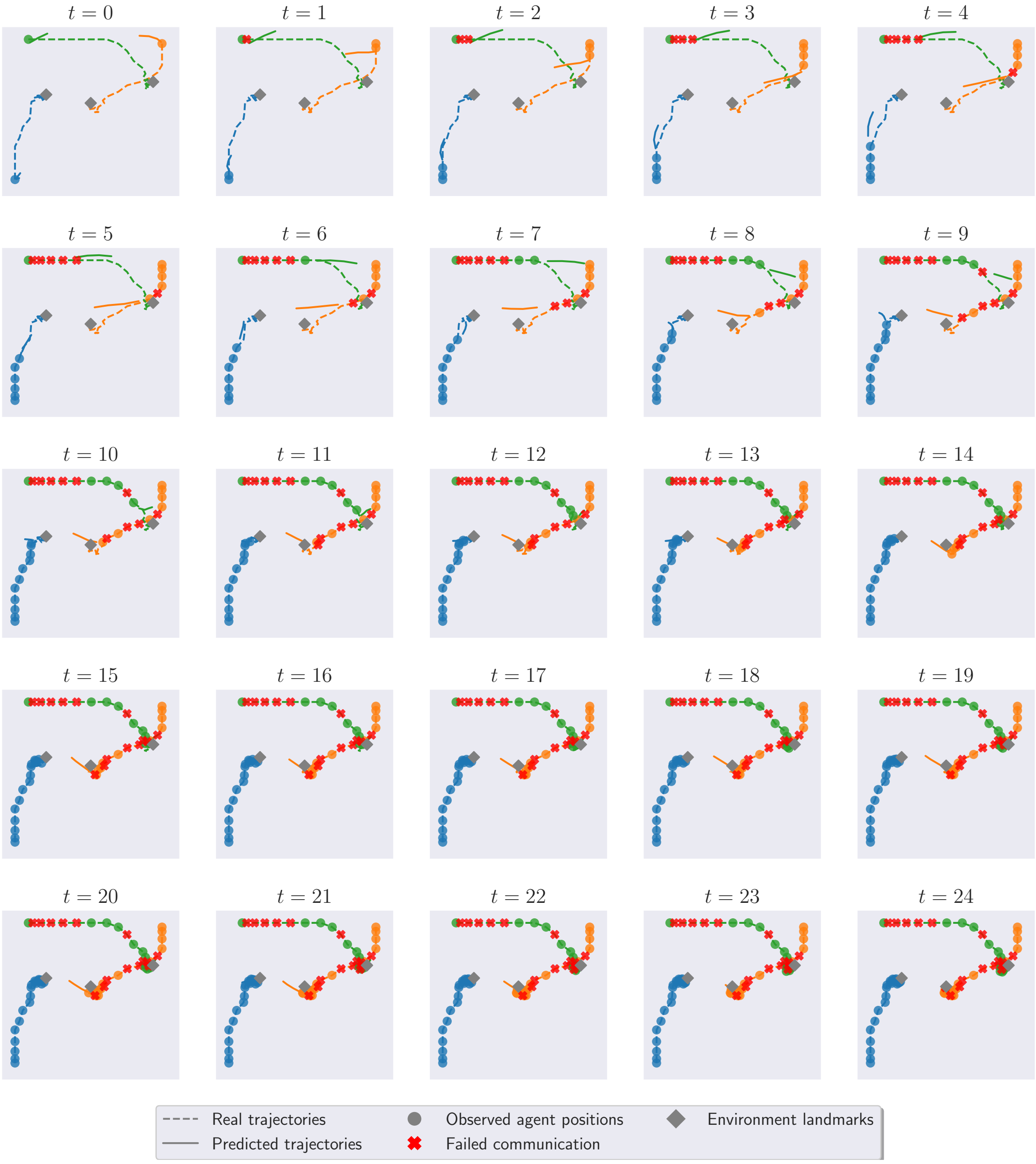


Figure E.16: Trajectory prediction plots for the Spreadblindfold environment under the QMIX algorithm from the perspective of agent 0 (blue).

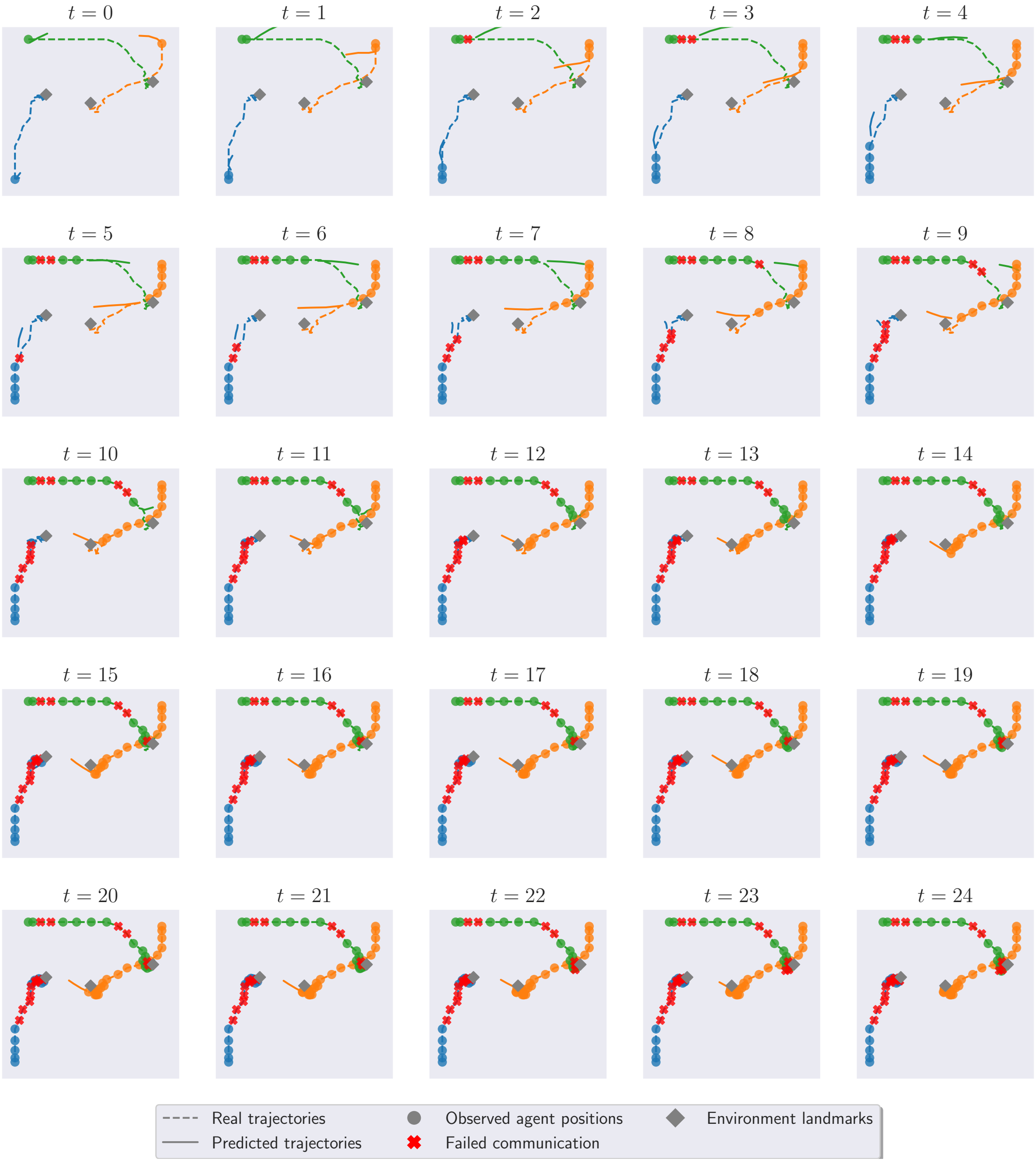


Figure E.17: Trajectory prediction plots for the Spreadblindfold environment under the QMIX algorithm from the perspective of agent 1 (orange).

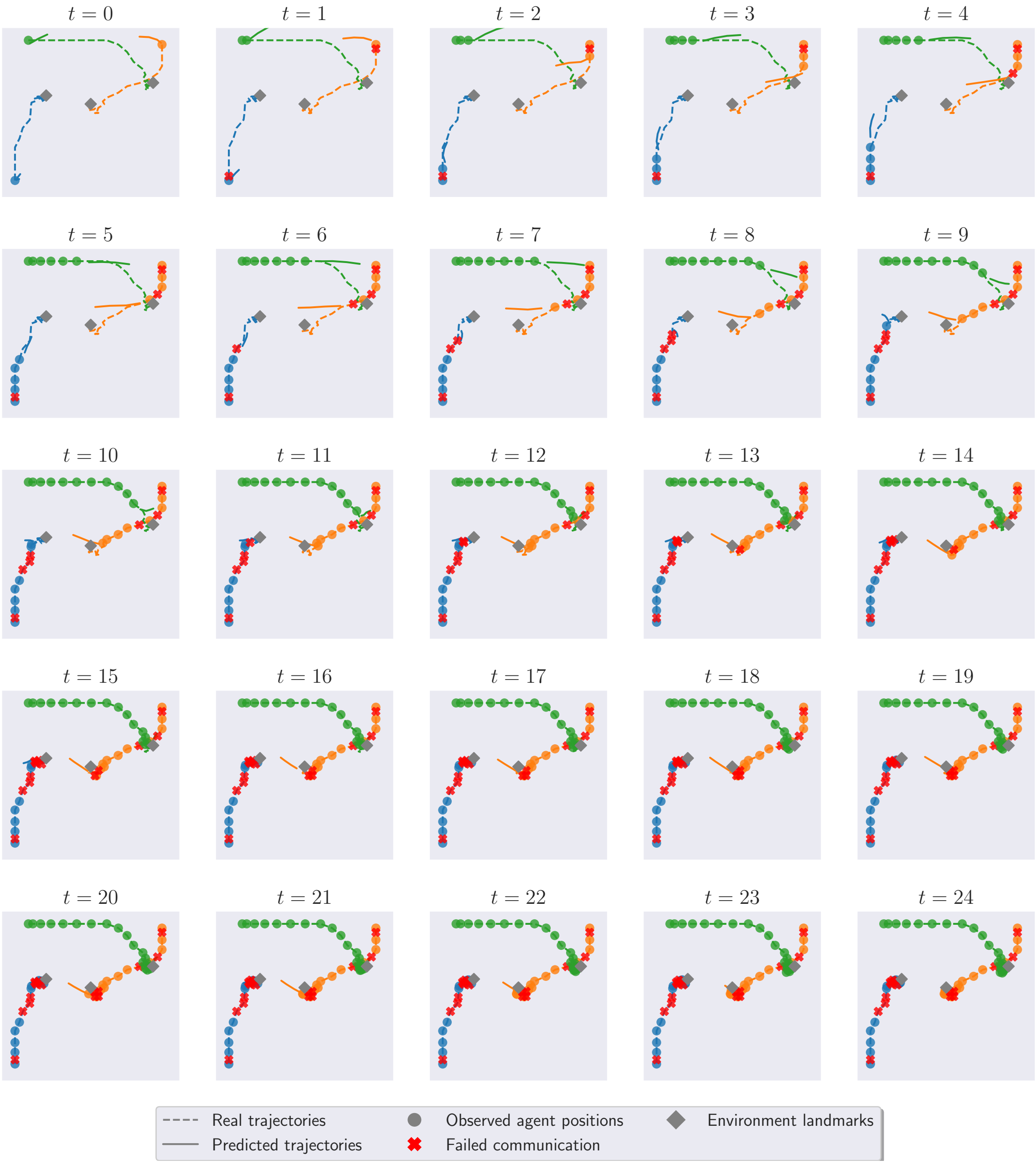


Figure E.18: Trajectory prediction plots for the Spreadblindfold environment under the QMIX algorithm from the perspective of agent 2 (green).

Appendix F

Training Hyperparameters and Constants

Multimodal Representation Learning for Efficient Cross-Modal Inference (Chapter 5)

In this section we describe the training hyperparameters employed for the evaluation of MUSE presented in Chapter 5. We recover the total loss of MUSE (Eq. 4),

$$\begin{aligned}\ell(\boldsymbol{x}_{1:M}, \boldsymbol{c}_{1:M}) = {} & \sum_{m=1}^{M} \alpha_m D_{\mathrm{KL}}(q_{\phi^b_m}(\boldsymbol{z}_m \mid \boldsymbol{x}_m) \parallel p(\boldsymbol{z}_m)) - \mathbb{E}_{q_{\phi^b_m}(\boldsymbol{z}_m \mid \boldsymbol{x}_m)} \lambda_m \log p_{\theta^b_m}(\boldsymbol{x}_m \mid \boldsymbol{z}_m) \\ & + \beta D_{\mathrm{KL}}(q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) \parallel p(\boldsymbol{z}_\pi)) - \sum_{m=1}^{M} \gamma_m \, \mathbb{E}_{q_{\phi^t}(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M})} \log p_{\theta^t_m}(\boldsymbol{c}_m \mid \boldsymbol{z}_\pi) \\ & + \frac{\delta}{D} \sum_{d=1}^{D} D^{\star}_{\mathrm{KL}}(q^t_\phi(\boldsymbol{z}_\pi \mid \boldsymbol{c}_{1:M}) \parallel q^t_\phi(\boldsymbol{z}_\pi \mid \boldsymbol{c}^d)), \qquad \text{(F.1)}\end{aligned}$$

where the modality-specific hyperparameters λ_m, α_m and γ_m control the modality data reconstruction, modality-specific distribution regularization and modality code reconstruction objectives, respectively. The hyperparameter β controls the regularization of the multimodal latent distribution. We present the training hyperparameters employed in the evaluation of MUSE in Table F.1.

Geometric Multimodal Contrastive Learning (Chapter 6)

In this section we describe the training hyperparameters employed in the evaluation of GMC presented in Chapter 6. We present the hyperparameters for the unsupervised learning scenario in Table F.2a. In Table F.2b we present the hyperparameters employed in the training of GMC for the supervised learning scenario.

Multimodal Transfer in Reinforcement Learning (Chapter 7)

In this section we describe the training hyperparameters employed in the evaluation of the multimodal transfer in reinforcement learning pipeline presented in Chapter 7. We present

Table F.1: Training hyperparameters employed in the evaluation of MUSE (Chapter 5)

(a) MNIST

Parameter	Value
Training Epochs	200
Learning Rate	10^{-3}
Batch-size	64
Optimizer	Adam
λ_1	1.0
λ_2	50.0
$\alpha_1 = \alpha_2$	1.0
$\gamma_1 = \gamma_2$	10.0
β	1.0
δ	1.0

(b) CelebA

Parameter	Value
Training Epochs	100
Learning Rate	10^{-4}
Batch-size	64
Optimizer	Adam
λ_1	1.0
λ_2	50.0
$\alpha_1 = \alpha_2$	1.0
$\gamma_1 = \gamma_2$	10.0
β	1.0
δ	1.0

(c) MNIST-SVHN

Parameter	Value
Training Epochs	100
Learning Rate	10^{-3}
Batch-size	64
Optimizer	Adam
$\lambda_1 = \lambda_2$	1.0
λ_3	50.0
$\alpha_1 = \alpha_2 = \alpha_3$	1.0
$\gamma_1 = \gamma_2 = \gamma_3$	10.0
β	1.0
δ	1.0

(d) MHD

Parameter	Value
Training Epochs	100
Learning Rate	10^{-3}
Batch-size	64
Optimizer	Adam
$\lambda_i = \lambda_s$	1.0
$\lambda_t = \lambda_l$	50.0
$\alpha_i = \alpha_s = \alpha_t = \alpha_l$	1.0
$\gamma_i = \gamma_s = \gamma_t = \gamma_l$	10.0
β	1.0
δ	1.0

Table F.2: Training hyperparameters of GMC (Chapter 6).

(a) Unsupervised

Parameter	Value
Low-level size d	64
High-level latent size s	64
Model training epochs	100
Classifier training epochs	50
Learning rate	1e−3
Batch size B	64
Temperature τ	0.1

(b) Supervised

Parameter	Value
Low-level size d	60
High-level latent size s	60
Model training epochs	40
Learning rate	1e−3 (Decay)
Batch size B	40
Temperature τ	0.3

the scenario constants for the instantiation of the multimodal Atari games in Table F.3. In Table F.4 we present the hyperparameters employed in the training of generative models for the multimodal Atari games scenarios. Finally, in Table F.5 we present the hyperparameters employed in the training of reinforcement learning algorithms for the multimodal Atari game scenarios.

Hybrid Execution in Multi-Agent Reinforcement Learning (Chapter 8)

In this section we describe the training hyperparameters for training MARO, presented in Chapter 8. We consider two MARL algorithms: IQL and QMIX. We employ the same LSTM-based controller networks across all evaluations. We follow the hyperparameters suggested by Papoudakis et al. [124]; we train all models for 4M steps, performing 5 training runs for each experimental setting and 50 evaluation rollouts for each training run. We assume that $p = 1$ at $t = 0$ for the MD and MARO algorithms. The performance of the Obs. and J. Obs. baselines are evaluated by aggregating evaluation rollouts with $p = 0$ and $p = 1$, respectively. The other algorithms are evaluated for p sampled from a discretized uniform distribution. We display our training hyperparameters for the RL controllers and the predictive model in Table F.6 and Table F.7, respectively.

We developed our code in a Python environment using the EPyMARL framework [124] and PyTorch [125]. The computational code is available in Github.

Table F.3: Constants employed in the multimodal Atari games (Chapter 7)

(a) Pendulum

Parameter	Value
f_0	440.0 Hz
K	1.0
c	20.0
Sound Receivers	$\{lb, rb, mt\}$
Frame Stack	2

(b) Hyperhot

Parameter	Value
$f_0^0, f_0^1, f_0^2, f_0^3$	$(261, 329, 392, 466)$ Hz
$a_0^0, a_0^1, a_0^2, a_0^3, a_M$	1.0
δ	0.025
c	20.0
Sound Receivers	$\{lb, rb, pl, pr\}$
Frame Stack	2

Table F.4: Hyperparameters employed in the training of generative models for the multimodal transfer in reinforcement learning scenarios (Chapter 7)

(a) AVAE (Pendulum)

Parameter	Value
Latent Space	10
Epochs	500
Learning Rate	10^{-3}
Batch-size	128
Optimizer	Adam
$\lambda_{\text{image}} = \lambda_{\text{sound}} = \beta = \alpha$	1.0
M	20000

(b) MUSE (Pendulum)

Parameter	Value
Latent Space $\lvert\mathbf{z}_\pi\rvert$	10
Latent Space $\lvert\mathbf{z}_{\text{image}}\rvert, \lvert\mathbf{z}_{\text{sound}}\rvert$	$\{16, 8\}$
Epochs	500
Learning Rate	10^{-3}
Batch-size	128
Optimizer	Adam
λ_{image}	1.0
λ_{sound}	100.0
$\alpha_{\text{image}} = \alpha_{\text{sound}} = \beta = \delta$	1.0
$\gamma_{\text{image}} = \gamma_{\text{sound}}$	10.0
M	20000

(c) AVAE (Hyperhot)

Parameter	Value
Latent Space	40
Epochs	250
Learning Rate	10^{-3}
Batch-size	128
Optimizer	Adam
λ_{image}	0.02
λ_{sound}	0.015
β	$1e-5$
α	0.05
M	32000

(d) MUSE (Hyperhot)

Parameter	Value
Latent Space $\lvert\mathbf{z}_\pi\rvert$	40
Latent Space $\lvert\mathbf{z}_{\text{image}}\rvert = \lvert\mathbf{z}_{\text{sound}}\rvert$	64
Epochs	250
Learning Rate	10^{-3}
Batch-size	128
Optimizer	Adam
λ_{image}	0.02
λ_{sound}	0.015
$\alpha_{\text{image}} = \alpha_{\text{sound}} = \beta = \delta$	$1e-5$
M	32000

Table F.5: Hyperparameters employed in the training of reinforcement learning algorithms for the multimodal transfer in reinforcement learning scenarios (Chapter 7)

(a) DDPG (Pendulum)

Parameter	Value
Batch-size	128
Learning Rate Actor	10^{-4}
Learning Rate Critic	10^{-3}
γ	0.99
Max Episode Length	300 frames
Replay Buffer	25000
Max Frames	150000
τ	$1e-3$

(b) DQN (Hyperhot)

Parameter	Value
Batch-size	128
Learning Rate	10^{-5}
γ	0.99
Max Episode Length	450 frames
Replay Buffer	350000
Max Frames	1750000

Table F.6: Hyperparameters for the RL controllers across all environments for hybrid execution in multi-agent reinforcement learning (Chapter 8).

(a) IQL

Hidden dimension	128
Learning rate	0.0005
Reward standardisation	True
Network type	GRU
Evaluation epsilon	0.0
Epsilon anneal	$500,000$
Target update	200

(b) QMIX

Hidden dimension	128
Learning rate	0.0005
Reward standardisation	True
Network type	GRU
Evaluation epsilon	0.0
Epsilon anneal	$50,000$
Target update	200

Table F.7: Hyperparameters for the predictive model across all environments and algorithms for hybrid execution in multi-agent reinforcement learning (Chapter 8).

Hidden dimension	128
Learning rate	0.001
Grad clip	1.0

Appendix G

Network Architectures

Multimodal Representation Learning for Efficient Cross-Modal Inference (Chapter 5)

In this section we describe the network architectures employed in the evaluation of MUSE, presented in Chapter 5. In Table G.1, we present the modality-specific, bottom-level networks. In Table G.2, we present the top-level networks, specific for each evaluation. Finally, in Table G.3, we present the network architecture of the additional components of the complementary cross-modality evaluation metrics.

Multimodal Transfer in Reinforcement Learning (Chapter 7)

In this section we describe the network architectures employed in the evaluation of the multimodal transfer in reinforcement learning approach presented in Chapter 7. In Table G.4, we present the modality-specific networks for the evaluation in the Pendulum and Hyperhot scenarios. In Table G.5, we present the multimodal networks, specific for each multimodal Atari game.

Table G.1: Modality-specific network architectures for the evaluation of MUSE (best viewed with zoom).

(a) Image (MNIST, MHD)

Encoder	**Decoder**
Input $\mathbb{R}^{1+28+28}$	Input $\mathbb{R}^{D}$
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 512 + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 6272 + Swish
FC, 512 + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish
FC, D, FC, D	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Sigmoid

(b) Label (MNIST, MHD)

Encoder	**Decoder**
Input $\mathbb{R}^{10}$	Input $\mathbb{R}^{D}$
FC, 64 + ReLU	FC, 64 + ReLU
FC, 64 + ReLU	FC, 64 + ReLU
FC, D, FC, D	FC, 64 + ReLU
-	FC, 10 + Log Softmax

(c) Image (CelebA)

Encoder	**Decoder**
Input $\mathbb{R}^{1+64+64}$	Input $\mathbb{R}^{D}$
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 6400 + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Batchnorm + Swish	Transposed Convolutional, 4x4 kernel, 1 stride, 0 padding + Batchnorm + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Batchnorm + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Batchnorm + Swish
Convolutional, 4x4 kernel, 1 stride, 0 padding + Batchnorm + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Batchnorm + Swish
FC, 512 + Swish + Dropout ($p = 0.1$)	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Sigmoid
FC, D, FC, D	-

(d) Attribute (CelebA), Trajectory (MHD)

Encoder	**Decoder**
Input $\mathbb{R}^{40}$	Input $\mathbb{R}^{D}$
FC, 512 + Batchnorm + Swish	FC, 512 + Batchnorm
FC, 512 + Batchnorm + Swish	FC, 512 + Batchnorm
FC, D, FC, D	FC, 512 + Batchnorm
-	FC, 40 + Sigmoid

(e) SVHN image (SVHN-MNIST)

Encoder	**Decoder**
Input $\mathbb{R}^{1+64+64}$	Input $\mathbb{R}^{D}$
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	Transposed Convolutional, 4x4 kernel, 1 stride, 0 padding + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish
FC, 1024 + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Sigmoid
FC, 512 + Swish	-
FC, D, FC, D	-

Table G.2: Multimodal network architectures, where $D_m = |\boldsymbol{z}_m|$ and $D_\pi = |\boldsymbol{z}_\pi|$, for the evaluation of MUSE (best viewed with zoom).

(a) MNIST

Encoder	**Decoder**
Input $\mathbb{R}^{D_m}$	Input $\mathbb{R}^{D_\pi}$
FC, 128 + ReLU	FC, 128 + ReLU
FC, 128 + ReLU	FC, 128 + ReLU
FC, D_π, FC, D_π	FC, D$_m$

(b) CelebA

Encoder	**Decoder**
Input $\mathbb{R}^{D_m}$	Input $\mathbb{R}^{D_\pi}$
FC, 128 + ReLU	FC, 128 + ReLU
FC, 128 + ReLU	FC, 128 + ReLU
FC, D_π, FC, D_π	FC, D_m

(c) MNIST-SVHN, MHD

Encoder	**Decoder**
Input $\mathbb{R}^{D_m}$	Input $\mathbb{R}^{D_\pi}$
FC, 512 + ReLU	FC, 512 + ReLU
FC, 512 + ReLU	FC, 512 + ReLU
FC, 512 + ReLU	FC, 512 + ReLU
FC, D_π, FC, D_π	FC, D_m

Table G.3: Network architectures for the complementary metrics for cross-modality generative performance of MUSE (best viewed with zoom).

(a) MNIST VAE, with $D = 64$

Encoder	**Decoder**
Input $\mathbb{R}^{1+28+28}$	Input $\mathbb{R}^{D}$
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 512 + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 6272 + Swish
FC, 512 + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish
FC, D, FC, D	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Sigmoid

(b) MNIST Autoencoder, with $B = 64$

Encoder	**Decoder**
Input $\mathbb{R}^{1+28+28}$	Input $\mathbb{R}^{B}$
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 512 + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 6272 + Swish
FC, 512 + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish
FC, B	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Sigmoid

(c) MNIST Classifier

Classifier
Input $\mathbb{R}^{1+28+28}$
Convolutional, 5x5 kernel, 1 stride, 0 padding + ReLU + Dropout($p = 0.2$) + MaxPool(2,2)
Convolutional, 5x5 kernel, 1 stride, 0 padding + ReLU + Dropout($p = 0.2$) + MaxPool(2,2)
FC, 128 + Dropout($p = 0.2$)
FC, 64 + Dropout($p = 0.2$)
FC, 10

Table G.4: Modality-specific network architectures for the multimodal Atari games scenarios (best viewed with zoom).

(a) Image (Pendulum)

Encoder	**Decoder**
Input $\mathbb{R}^{1+60+60}$	Input $\mathbb{R}^{D}$
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 128 + Swish
Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish	FC, 14400 + Swish
FC, 128 + Swish	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Swish
FC, D, FC, D	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + Sigmoid

(b) Sound (Pendulum)

Encoder	**Decoder**
Input $\mathbb{R}^{S}$	Input $\mathbb{R}^{D}$
FC, 50 + BatchNorm + ReLU	FC, 50 + BatchNorm + ReLU
FC, 50 + BatchNorm + ReLU	FC, 50 + BatchNorm + ReLU
FC, D, FC, D	FC, S + Sigmoid

(c) Image (Hyperhot)

Encoder	**Decoder**
Input $\mathbb{R}^{1+80+80}$	Input $\mathbb{R}^{D}$
Convolutional, 8x8 kernel, 2 stride, 1 padding + ReLU	FC, 512 + ReLU
Convolutional, 4x4 kernel, 2 stride, 1 padding + ReLU	FC, 4096 + ReLU
Convolutional, 3x3 kernel, 1 stride, 0 padding + ReLU	Transposed Convolutional, 3x3 kernel, 1 stride, 0 padding + ReLU
FC, 512 + ReLU	Transposed Convolutional, 4x4 kernel, 2 stride, 1 padding + ReLU
FFC, D, FC, D	Transposed Convolutional, 8x8 kernel, 4 stride, 2 padding + Sigmoid

(d) Sound (Hyperhot)

Encoder	**Decoder**
Input $\mathbb{R}^{S}$	Input $\mathbb{R}^{D}$
FC, 512 + Batchnorm + ReLU	FC, 512 + Batchnorm + ReLU
FC, 512 + Batchnorm + ReLU	FC, 512 + Batchnorm + ReLU
FC, D, FC, D	FC, S + Sigmoid

Table G.5: Multimodal network architectures, where $D_m = |\boldsymbol{z}_m|$ and $D_\pi = |\boldsymbol{z}_\pi|$, for the multimodal Atari games scenarios (best viewed with zoom).

(a) Pendulum

Encoder	**Decoder**
Input $\mathbb{R}^{D_m}$	Input $\mathbb{R}^{D_\pi}$
FC, 256 + ReLU	FC, 256 + ReLU
FC, 256 + ReLU	FC, 256 + ReLU
FC, 256 + ReLU	FC, 256 + ReLU
FC, D_π, FC, D_π	FC, D_m

(b) Hyperhot

Encoder	**Decoder**
Input $\mathbb{R}^{D_m}$	Input $\mathbb{R}^{D_\pi}$
FC, 512 + ReLU	FC, 512 + ReLU
FC, 512 + ReLU	FC, 512 + ReLU
FC, 512 + ReLU	FC, 512 + ReLU
FC, D_π, FC, D_π	FC, D_m