Causal Inductive Biases for Learning Structured World Models



Anson Lei

Mansfield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2025

To my beloved parents — for your unwavering support.
獻給我親愛的父母 — 謝謝您們一直以來無私的支持。

Acknowledgements

I have been incredibly fortunate to have the support of many people during my DPhil. To list them all would likely take more space than the thesis itself, so I must limit myself to acknowledging some of the key individuals and groups who have made this journey possible. First and foremost, I extend my heartfelt gratitude to my wonderful supervisors, Ingmar and Bernhard, for being an infinite source of inspiration and, perhaps more importantly, for always having more faith in me than I do myself. Your endless enthusiasm and support are what made this DPhil possible. I would also like to thank my tutor from my Undergraduate years, Pawan Mudigonda, whose encouragement and insights 'tricked' me into pursuing a DPhil. I am grateful to all my colleagues at the A2I and at MPI-IS. It has been an inspiring and humbling experience to work alongside this group of immensely talented and kind-hearted individuals. Special thanks go to my collaborators, Hsiao-Ru, Florent, Markus, and Frederik, for the countless hours of insightful conversations fuelled by our shared passion for research (and coffee). My thanks go to my friends Tonny and Enson, who began their PhD journeys around the same time I did, for their camaraderie and companionship over the past four years. I would also like to thank my childhood friends, Matthew and Tony, for periodically, and at times forcefully, dragging me out of my academic bubble. I could not have asked for better buddies to grow up with. I must express the deepest gratitude to my parents, for their love, guidance, and unwavering support, from the day I announced, at the age of three, my dream of 'becoming a scientist' to the final page of this thesis. Finally, to my dear Ceri: thank you for being by my side for the past eight years, and for talking me into bringing Lyra home. Not a single day has gone by without my being grateful that I found the courage to speak to you on Norham Gardens. As I expressed in my Master's thesis, and I mean it even more now, I am truly lucky to have you in my life.

Abstract

World Models have emerged as a powerful paradigm for building autonomous agents: by learning a predictive model of the environment, agents can plan and make decisions based on imagined future scenarios. Despite their success, existing approaches to learning world models largely rely on dense, monolithic architectures that do not reflect the underlying structure of the environment. This thesis explores the role of explicit causal structures in building *adaptable*, *robust*, and *interpretable* world models. We propose three complementary approaches, *VCD*, which jointly learns latent variables and sparse causal dependencies to induce disentangled representations; *COMET*, which induces independent, reusable mechanisms via competition-of-experts; and *SPARTAN*, a Transformer-like architecture that discovers local interaction graphs using sparse attention. Across synthetic and real-world domains, we demonstrate that our methods are able to capture the underlying structure of the environment, and that compared to monolithic baselines, structured world models exhibit improved adaptability to distribution shifts, robustness to irrelevant distractors, as well as enhanced interpretability. In addition, we present two case studies that showcase the versatility of the SPARTAN architecture beyond world models, demonstrating its potential as a general-purpose tool for interpretability in reinforcement learning and large language models.

Contents

1 Introduction **1**

1.1 Contributions and Outline of the Thesis 5

1.1.1 Causal Representation Learning for World Models 6

1.1.2 Learning Independent Mechanisms via Competition 7

1.1.3 Sparse Transformers for Learning Local Dependencies 7

1.1.4 Sparse Transformers Beyond World Models 8

1.2 Publications 9

2 Background **11**

2.1 World Models 12

2.1.1 Representation Models 13

2.1.2 Dynamics Models 15

2.1.3 Policy Learning and Planning 16

2.2 Structure and Causality 17

2.3 Related Works 20

2.3.1 Representation Learning 20

2.3.2 Learning Independent Mechanisms 22

2.3.3 Structure Discovery 24

2.3.4 Causal Structure as a Prior 25

2.4 Summary 27

3 Variational Causal Dynamics **28**

3.1 Background 30

3.2 Variational Causal Dynamics 33

3.2.1 Causal dynamics model 34

3.2.2 Recurrent module 36

3.2.3 Training 36

3.2.4 Adaptation 37

3.2.5 Identifiability of Causal Graph and Causal Representations . 38

3.3 Experiments 41

3.3.1 Experiment Setup 42

3.3.2 Prediction Performance 43

3.3.3 Causal discovery and Disentanglement 45
3.3.4 Adaptation Efficiency 46
3.4 Conclusion and Discussions 47

4 Learning Independent Mechanisms via Competition 49
4.1 COMET: COmpetitive Mechanisms for Efficient Transfer 52
4.1.1 Problem Setup and Model Architecture 52
4.1.2 Phase 1: Learning Reusable Mechanisms via Competition . 54
4.1.3 Phase 2: Learning to Compose Mechanisms 56
4.2 Experiments 57
4.2.1 Experiment Setup 57
4.2.2 Disentanglement of Mechanisms 58
4.2.3 Adaptation Efficiency 61
4.3 Conclusion and Discussions 62

5 Learning Local Dependency Structures with Sparse Transformers 65
5.1 Background 67
5.2 Sparse Transformer World Models 70
5.2.1 Learnable Sparse Connections 71
5.2.2 Interventions 72
5.2.3 Training 72
5.3 Experiments 74
5.3.1 Experiment Setup 74
5.3.2 Graph Learning 75
5.3.3 Robustness and Adaptation Efficiency 78
5.3.4 Relations to Related Approaches 79
5.4 Conclusion 81

6 Sparse Transformers Beyond World Models 83
6.1 Direct Advantage Estimation with Sparse Transformers 84
6.1.1 Background and Approach 85
6.1.2 Experiments 88
6.1.3 Discussion 90
6.2 Sparse Language Models for Mechanistic Interpretability 92
6.2.1 Background and Approach 94
6.2.2 Experiments 96
6.2.3 Discussion 100
6.3 Conclusion 100

7 Conclusion **104**
7.1 Contributions 104
7.2 Limitations and Future Work 108
7.3 Final Remarks 110

Appendices

A Appendix for VCD **112**
A.1 Derivations 112
A.2 Implementation Details 113
A.3 Experiment Details 114
A.4 Qualitative Exploration 115
A.5 Further Experimental Results 117

B Appendix for COMET **121**
B.1 Implementation Details 121
B.2 Experiment Details 124
B.3 Ablation Studies 127
B.4 Large Data Limit Behaviour 128

C Appendix for SPARTAN **130**
C.1 Implementation Details 130
C.2 Datasets 131
C.3 Further Experimental Results 134

D Appendix for Interpretability Experiments **135**
D.1 Details for Direct Advantage Estimation 135
D.2 Details for Mechanistic Interpretability 136

References **139**

1 Introduction

World models [1]—learnt models that simulate the dynamics of the environment—have emerged as a key component that enables autonomous agents to understand and operate in the world. Accurate world models enable agents to *imagine* future states and evaluate hypothetical action plans without potentially risky and costly real-world interactions. Equipped with this capability of mental simulation, agents can make intelligent decisions by, for example, explicitly planning their actions in the world model [e.g. 2] or learning policies via reinforcement learning using imagined, synthetic experiences [e.g. 3]. This paradigm of world model based methods have gained significant traction in recent years, and have achieved state-of-the-art performance across various domains such as game-playing [3, 4] and robotics [5]. While the challenge of effectively acquiring accurate, high-fidelity world models from sensory inputs remains, recent advances in generative modelling for complex, high-dimensional observations (e.g., VQ-VAE [6], DC-AE [7]), in tandem with performant sequence prediction architectures (e.g. Transformer [8], State-space models [9]), have led to world models that are capable of consistent predictions over longer time horizons and in increasingly complex environments. Leveraging these architectural components, recent breakthroughs in large-scale foundation world models in the domains of autonomous driving [10], game generation [11, 12], and physical modelling [13] dramatically demonstrate the scope and potential of this

approach. Taken together, the conceptual appeal of world models, coupled with their empirical successes in a variety of domains, has cemented their role as a central tenet of artificial intelligence. Yet, despite their predictive power, a critical question remains regarding the structure of the knowledge these models acquire.

In parallel to the developments within machine learning, this idea of an internal world model that encodes observations and makes predictions echoes with theories in cognitive neuroscience which posit that the human mind functions as a *predictive machine* [14, 15] — a generative and predictive model of the world. However, while analogies can be drawn between current world model approaches and human cognition, the way in which knowledge is represented appears to be fundamentally different. For humans, it is conjectured that the internal representation of the world is highly *structured* [16]: when presented with observations, our understanding of the environment is mediated by discrete entities such as objects and their properties, how they interact with each other, and how the situation at hand relates to past experiences. Typical world models, in stark contrast, often rely on *monolithic* architectures that encode observations into a single latent embedding and make predictions using a single dynamics model. To illustrate this discrepancy with an intuitive example, imagine observing a busy traffic intersection: while a learnt world model is capable of achieving high prediction accuracy by mapping the entire scene into a single latent embedding, humans naturally parse observations into distinct vehicles, pedestrians, and traffic signs, each interacting with other entities. Furthermore, we decompose our understanding of the dynamics into independent rules and mechanisms such as speed limits, stopping at red lights, and yielding to pedestrians. The juxtaposition of these two perspectives, *structured* versus *monolithic*, underscores several fundamental shortcomings of current world modelling approaches:

- **Adaptability.** When confronting changes in the environment dynamics, monolithic models often require retraining or fine-tuning of the entire model. It has been argued that the flexibility afforded by human cognition stems from our ability to organise knowledge into generalisable and composable

components that can be selectively reconfigured [16]. Returning to the traffic example, humans can rapidly adapt to driving on the opposite side of the road by only updating a small subset of the underlying rules, while the majority of our understanding can be effectively reused. It is this ability to adapt to changes via modular updates that enables us to efficiently transfer knowledge across different contexts, a capability that is lacking in current world models.

- **Robustness.** In a similar vein, monolithic world models are prone to learning spurious correlations. Current state-of-the-art models are sensitive to perturbations in irrelevant features: empirical evidence shows that, for instance, in the autonomous driving domain, removing irrelevant vehicles from the scene can lead to significant changes in the models' predictions [17]. Structurally representing the correct relationship between entities and their interactions, therefore, plays a crucial role in improving robustness.

- **Interpretability.** Interpretability — the ability to understand and explain the decisions made by a model — is an essential desideratum for building safe and trustworthy AI systems, offering a means to diagnose model behaviours as well as derive insights about the underlying environment's dynamics. To date, however, most world model architectures are effectively black-boxes that are opaque to interpretation. Structured models, by virtue of explicitly representing the underlying entities and mechanisms, hold the potential for enhancing interpretability.

The overarching goal of this thesis is to develop practical algorithms that learn structured world models from observations, and to demonstrate their efficacy over standard, monolithic world models in terms of *adaptability*, *robustness*, and *interpretability*. To this end, insights from *causality* offer a conceptual framework for incorporating structures into models: the *Independent Causal Mechanism* (ICM) Principle [18, 19], which posits that the data generating process is composed of a set of causal mechanisms that do not inform each other and can be independently intervened upon, provides the guiding principles for building structured world models.

Complementing this, the *Sparse Mechanism Shift* (SMS) hypothesis [20, 21] further suggests that naturally occurring changes in the environment can often be attributed to sparse and local changes in the underlying causal mechanisms, implying that many causal mechanisms remain *invariant* across different environments. Collectively, these insights indicate that causality provides a principled and natural framework for organising knowledge about the world: the ICM principle directly motivates a modular architecture that explicitly captures interpretable, independent causal mechanisms, while the SMS hypothesis provides the basis for the potential of improved robustness and adaptability.

The central question remains: *how can we learn causally structured world models from data without explicit supervision?* In order to operationalise this causal framework, we draw on recent innovations in the emerging field of *causal machine learning*, which aims to integrate causal principles into machine learning systems, such as differentiable causal discovery [e.g. 22] and causal representation learning [20]. Building on these ideas, this thesis investigates the unique challenges and opportunities presented by the context of acquiring world models from complex environments. Here, we highlight three distinct, but intricately linked, key challenges:

- **Representation Learning.** World models need to operate on complex observations such as images, necessitating the learning of compact abstractions on which predictions can be made. However, discovering causal structures requires identifying semantically meaningful, disentangled abstractions, e.g. causal variables such as object properties. Learning these abstractions in a principled manner is a critical open challenge.

- **Mechanism Learning.** Analogous to the need for meaningful representations for observations, as posited by the ICM principle, a causally structured world model also requires disentangling the independent mechanisms that govern the dynamics of the environments, such as rules and interactions. Identifying and parameterising these mechanisms remains an underexplored line of inquiry.

- **Structure Discovery.** Finally, a structured world model equipped with representations for both state and mechanism must be able to reason about how these components are bound to each other. This task is akin to the problem of causal discovery, where the goal is to identify the causal graph structure that explains the observations. In the particular context of world models, however, the classical graph learning framework must be significantly extended to account for the complexity of real-life environments, such as the need for flexible, time-varying graph structures.

This thesis proposes novel algorithms that address each of these facets of the problem, and incorporate these causal principles as *soft inductive biases* that guide the learning process of world models. This can be understood through the framework of Bayesian inference, where a structural prior is placed over the space of possible world models. In particular, the architectural and regularisation techniques developed in this work bias the model towards structured representations and mechanisms that are consistent with the ICM principle and the SMS hypothesis. Throughout the thesis, we empirically demonstrate the effectiveness of this prior through a series of applications ranging from simple simulation-based environments to complex, real-world datasets. The next section outlines the overall structure of the thesis and gives a high-level overview of the proposed methods.

1.1 Contributions and Outline of the Thesis

We begin in Chapter 2, which provides a comprehensive review of the relevant literature, covering foundational approaches and recent breakthroughs in world modelling, including existing methods for leveraging structure in world models. The chapter also covers key concepts in causality and provides an overview of current approaches, broadly organised around the three core challenges of representation learning, mechanism learning, and structure discovery. Subsequently, Chapters 3, 4, and 5, which are designed to be self-contained, present the main contributions of this thesis, namely the novel algorithms developed over the course of the

DPhil: VCD (Chapter 3), COMET (Chapter 4), and SPARTAN (Chapter 5). In Chapter 6, we explore the potential for extending the scope of the sparse transformer architecture, SPARTAN, developed in Chapter 5, to applications beyond world models, including reinforcement learning and mechanistic interpretability. Finally, in Chapter 7, we conclude the thesis with a summary of the main findings, a critical discussion of the limitations of the work, and highlight potential avenues for future research directions. Below, we provide a brief overview of the approaches and the contributions of each chapter.

1.1.1 Causal Representation Learning for World Models

In Chapter 3, we explore the feasibility of extending existing latent-state world models with causal structures by instantiating the dynamics model as a causal graphical model. Here, the principal challenge is one of *representation learning*: discovering causal structures requires access to causally meaningful variables. How can we learn a latent representation that is amenable to causal factorisation? Our key insight is that it is precisely because only meaningful representations afford sparse causal relationships, *jointly learning latent representations and performing causal discovery in the latent space* induces the emergence of disentangled causal representations. We present Variational Causal Dynamics (VCD), a structured world model that exploits the invariance of causal mechanisms across environments to achieve fast and modular adaptation. By causally factorising a dynamics model, VCD is able to identify reusable components across different environments. This is achieved by combining causal discovery and variational inference to learn a latent representation and dynamics model jointly in an unsupervised manner. Specifically, we optimise the evidence lower bound jointly over a representation model and a dynamics model structured as a causal graphical model. In evaluations on simulated environments with state and image observations, we show that VCD is able to successfully identify causal variables and to discover consistent causal structures across different environments. Moreover, given a small number of observations in a previously unseen, intervened environment, VCD is able to identify the sparse

changes in the dynamics and to adapt efficiently. In so doing, VCD significantly extends the capabilities of current latent world models while also comparing favourably in terms of prediction accuracy. This serves as a first proof-of-concept for how a structured world model can facilitate efficient and interpretable adaptation.

1.1.2 Learning Independent Mechanisms via Competition

In Chapter 4, we turn our attention to the problem of *mechanism learning*. Here, the central idea is that the dynamics of an environment can be decomposed into reusable, versatile, and generalisable mechanisms such as interaction primitives and rules. Explicitly capturing this mechanistic structure requires both architectural and algorithmic innovations. To this end, we present COmpetitive Mechanisms for Efficient Transfer (COMET), a modular world model that leverages reusable, independent mechanisms across different environments. COMET is trained on multiple environments with varying dynamics via a two-step process: competition and composition. This enables the model to recognise and learn transferable mechanisms. Specifically, in the competition phase, COMET is trained with a winner-takes-all gradient allocation, encouraging the emergence of independent mechanisms. These are then reused in the composition phase, where COMET learns to re-compose learnt mechanisms in ways that capture the dynamics of intervened environments. In so doing, COMET explicitly reuses prior knowledge, enabling efficient and interpretable adaptation. We evaluate COMET in environments with image-based observations. COMET captures recognisable mechanisms without supervision, a capability lacking in competitive baselines. Moreover, we show that COMET is able to adapt to new environments with varying numbers of objects with improved sample efficiency compared to more conventional finetuning approaches.

1.1.3 Sparse Transformers for Learning Local Dependencies

Causal structures play a central role in world models that flexibly adapt to changes in the environment. Chapter 5 addresses the challenge of *learning structures*, i.e. identifying which entity influences which other entities in a scene. Here, the context

of learning world models with complex dynamics presents unique challenges, as the interactions between entities can be temporally sparse and context-dependent, rendering the classical approach of learning a fixed graph structure to explain the entire dataset insufficient. While recent works motivate the framework of *local* causal graphs for dynamics modelling [e.g. 23–25], in this chapter, we demonstrate that accurately capturing these relationships in realistic settings remains challenging for the current state-of-the-art. To remedy this shortcoming, we postulate that sparsity is a critical ingredient for the discovery of such local causal structures. To this end, we present the SPARse TrANsformer World model (SPARTAN), a Transformer-based world model that learns local causal structures between entities in a scene. By applying sparsity regularisation on the attention pattern between object-factored tokens, SPARTAN identifies sparse local causal models that accurately predict future object states. Furthermore, we extend our model to capture sparse interventions with unknown targets on the dynamics of the environment. This results in a highly interpretable world model that can efficiently adapt to changes. Empirically, we evaluate SPARTAN against the current state-of-the-art in object-centric world models on observation-based environments and demonstrate that our model can learn accurate local causal graphs and achieve significantly improved few-shot adaptation to changes in the dynamics of the environment, as well as robustness against removing irrelevant distractors. To the best of our knowledge, SPARTAN is the first model that is able to learn local causal structures on large-scale, real-life datasets such as the Waymo Open Motion Dataset [17].

1.1.4 Sparse Transformers Beyond World Models

Besides its capability to learn local causal structures and perform accurate predictions on large-scale, real-life datasets, the SPARTAN architecture can also be used as a slot-in replacement for the standard Transformer architecture. Owing to the ubiquity of Transformers as a backbone for various applications, SPARTAN holds potential for improving interpretability across a diverse range of domains. In Chapter 6, we present two case studies that explore how SPARTAN can improve

interpretability: advantage estimation for *reinforcement learning* and mechanistic interpretability for *large language models*. In the context of reinforcement learning, we explore using SPARTAN as interpretable function approximators for policy learning. Specifically, building on recent works that elucidate the connection between the *advantage function* and the causal formulation of reinforcement learning [26], we explore how SPARTAN can be employed to learn interpretable representations of the advantage function, thereby enhancing the interpretability of derived policies. For mechanistic interpretability, we investigate the potential of using the SPARTAN architecture as language models that utilise sparse attention patterns. In this investigation, we empirically demonstrate that the sparsity regularisation scheme developed in SPARTAN is effective as a post-training technique for standard large language models, improving the interpretability of language models while maintaining their performance at the scale of up to 1 billion parameters.

1.2 Publications

The list of publications resulting from this thesis is as follows:

- Chapter 3: A. Lei, B. Schölkopf, and I. Posner. "Variational causal dynamics: Discovering modular world models from interventions". In *Transactions on Machine Learning Research (2023)* [27]; A shorter workshop version is presented as A. Lei, B. Schölkopf, and I. Posner. "Causal discovery for modular world models". In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (2022)* [28].

- Chapter 4: A. Lei, F. Nolte, B. Schölkopf, I. Posner. "Modular World Models with Competitive Independent Mechanisms". In *Causal and Object-Centric Representations for Robotics Workshop at CVPR (2024)* [29]; A longer preprint version is available as A. Lei, F. Nolte, B. Schölkopf, I. Posner. "Compete and compose: Learning independent mechanisms for modular world models". In *arXiv preprint arXiv:2404.15109, (2024)* [30].

- Chapter 5: A. Lei, B. Schölkopf, I. Posner. "SPARTAN: A sparse transformer learning local causation". In *Advances in Neural Information Processing Systems, (2025)* [31]

- Chapter 6: H. Pan, A. Lei, I. Posner, B. Schölkopf. "Learning What Matters: Endogenous Representation through Advantage Estimation and Sparse Attention". In *The Causal Reinforcement Learning Workshop at Reinforcement Learning Conference (2025)* [32]; The work on mechanistic interpretability is not published at the time of writing.

The work done in Chapter 6 was carried out during a research exchange at the Max Planck Institute for Intelligent Systems, Tübingen, Germany, as part of the ELLIS programme. In cases where the work is carried out collaboratively, a *statement of authorship* is included at the end of the respective chapters.

2
Background

The previous chapter established the core motivation of this thesis: to develop world models that reflect the underlying causal structure of the environment for improved adaptability, robustness, and interpretability. To contextualise and ground our core contributions, this chapter provides a review of the relevant background and literature. Our review is structured to first establish the state-of-the-art in world modelling and then introduce the causal framework that offers a path beyond its limitations. Sec. 2.1 provides a comprehensive overview of world models, tracing their conceptual foundations as well as current architectural trends. In particular, we will review the architectural advances in representation and dynamics models that have enabled their success. In Sec. 2.2, we provide a brief exposition of the causality framework that will form the conceptual backbone of the rest of the thesis, and discuss the key aspect of this framework that motivates our approach. Finally, in Sec. 2.3, we contextualise our contributions within the landscape of recent research that also aims to learn structured models, both in the context of world models and in the broader machine learning literature. This review of related works is organised around the three pillars of our proposed methods: *learning representations*, *learning independent mechanisms*, and *structure discovery*.

2.1 World Models

The premise that an autonomous agent requires an internal model of its environment to make intelligent decisions has deep roots across control theory, reinforcement learning, and cognitive neuroscience. Early works in model predictive control laid the groundwork for how agents can leverage predictive models to plan their actions optimally [33]. These ideas have been further developed in the context of model-based reinforcement learning [34], which takes a more holistic view of the planning-acting-learning loop and explores how agents can *learn* a model of the environment from experience, demonstrating the potential for improved sample efficiency over model-free approaches [35]. This has led to early seminal works such as PILCO [36], which uses a Gaussian process to learn a dynamics model that guides planning and exploration, serving as a prototype for this paradigm. In parallel, the perspective of maintaining an internal model finds compelling analogues in cognitive neuroscience. Prominent theories such as the predictive processing framework [37], the Bayesian brain hypothesis [38], and active inference [15] have posited that the brain functions as a generative model of the world, continuously predicting sensory inputs and updating its internal beliefs. These theories shed light on neurobiologically plausible mechanisms for how agents might learn and leverage world models for both perception and action. Taken together, these convergent lines of research established the conceptual foundations for modern world models, providing principles for how they can be learnt from experience and leveraged for decision-making.

The advent of deep learning has provided a powerful mechanism for implementing world models for complex environments with high-dimensional observations. In their seminal work, Ha and Schmidhuber [1] successfully used a combination of variational autoencoders and recurrent neural networks to learn accurate world models from raw pixel observations, enabling agents to learn effective policies entirely within *imagined* trajectories without directly interacting with the world. This crucial ability to derive actions from simulated experience has been a key driver of the

development of world models, particularly in the context of embodied agents, where interaction data is scarce and safety is of paramount concern.

This initial success has motivated a plethora of subsequent works that extend the capabilities of world models and explore their applications in various domains. An influential family of works is the Dreamer series [3, 39, 40] which advocates for an end-to-end approach to learning world models directly from observations, achieving state-of-the-art performance in terms of sample efficiency across a wide range of game-like environments. The success of world models in simulated environments has motivated their application to a variety of domains such as robotics [5] and autonomous driving [41], highlighting their versatility and potential for real-world impact. Following recent trends in large-scale foundation models, recent endeavours have also explored the scalability of world models, leading to the development of high-fidelity, long-horizon, foundation-scale world models such as the Genie series for game generation [11, 12, 42], UniSim for robotics [43], and the GAIA series for autonomous driving [10]. In sum, world models have emerged as a versatile and rapidly evolving paradigm for building data-efficient and highly performant autonomous agents. Below, we will provide a comprehensive review of the key components of standard world models, while deferring the discussion of approaches that explicitly capture structures, such as object-centric models, to Sec. 2.3.

2.1.1 Representation Models

Pre-trained Representations Since it is often infeasible and inefficient to learn dynamics directly in the high-dimensional observation space, most world models architectures are composed of two main components (Fig. 2.1): a *representation model* that encodes high-dimensional observations into a compact latent space, and a *dynamics model* that predicts the evolution of these latent representations over time given the agent's actions. From the outset, methods for learning latent representations for world models have been heavily influenced by advances in unsupervised representation learning, in particular in the field of computer vision. Early works such as Ha and Schmidhuber [1] employed variational autoencoders

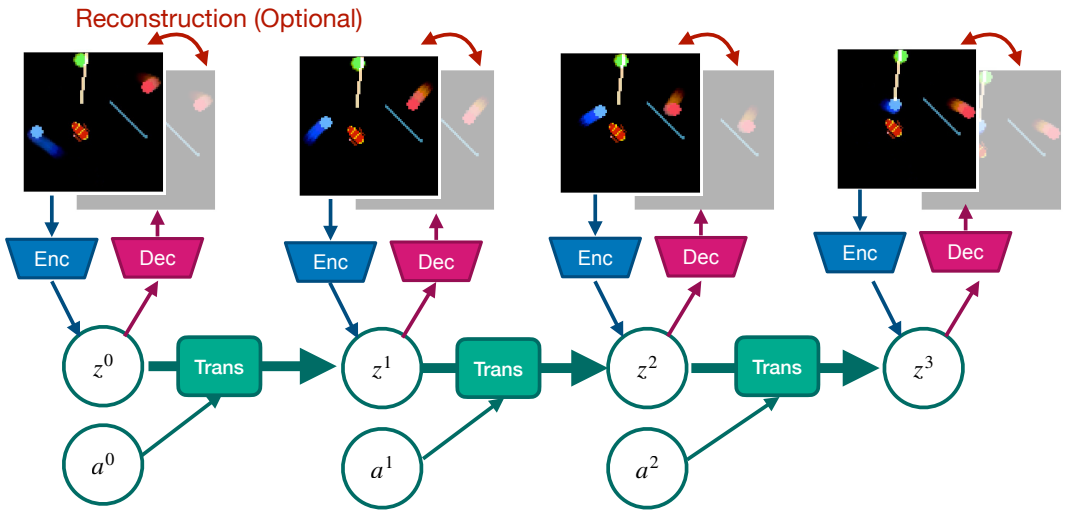


Figure 2.1: An illustration of a typical world model architecture. The representation model encodes observations, such as images, into a compact latent space. The dynamics model predicts the evolution of these latent representations over time given the agent's actions. These predictions can optionally be decoded back into the observation space.

(VAEs) [44] to learn a probabilistic mapping from high-dimensional observations to a lower-dimensional latent space by learning to reconstruct images. To improve the quality of the learnt representations, subsequent works have explored more sophisticated generative models. For example, IRIS [4, 45] utilises VQ-VAEs [6]; MWM [46] uses Masked AutoEncoders [47] with a vision transformer (ViT) [48] backbone. Other works, such as DinoWM [49] and LaDi-WM [50], further leverage the capabilities of large-scale pre-trained vision representation models [51] to extract high-quality and semantically rich features from images.

Jointly Learnt Representations. An orthogonal line of research has argued that representation learning should be integrated with the learning of the dynamics model instead of the aforementioned decoupled approach. The rationale is that joint learning can lead to representations that are more attuned to task-relevant features, which enables accurate predictions. This is often achieved by optimising a variational lower bound on the likelihood of the observed data, which ensures that the representation is informative enough to reconstruct the original image and predict future states (see also Sec. 3.1). Prominent examples include E2C [52, 53],

RSSM [54] and Dreamer [39]. Subsequent iterations of Dreamer [3, 40] develop this approach further by incorporating discrete latent representations, which improve the stability of long-term predictions. A parallel strand of research [e.g. 55, 56] has argued that the need to reconstruct images can be overly constraining and may lead to representations that capture irrelevant details, and instead advocates for reconstruction-free objectives such as contrastive predictive coding [57]. In a similar vein, various auxiliary objectives have been proposed to explicitly encourage robust representations that focus their expressive power on task-relevant features. This includes bisimulation metrics [58, 59], prototypical representations [60], and endogenous representations [61] (see also Sec. 6.1.1). To this end, our proposed method, VCD (Chapter 3), builds on the joint learning paradigm and uses the discovery of latent causal structure as a form of auxiliary objective.

2.1.2 Dynamics Models

After the observations are encoded into a latent space, the dynamics model predicts the evolution of these latent embeddings over time given the agent's actions. Early works such as E2C [52, 53] and SOLAR [62] employed locally linear dynamics models that simplified downstream planning and control. However, for complex and partially observable environments, more expressive models are required to capture the intricate dynamics. Recurrent architectures such as LSTM and GRU have been widely used to model dynamics since the early days of world models [1, 39], and have remained a staple in many subsequent works [3, 5, 40] due to their ability to capture temporal dependencies. More recently, the transformer architecture [8], originally developed for natural language processing, has emerged as an ubiquitous tool for sequence modelling. By casting dynamics prediction as modelling sequences of latent states and actions, recent works such as IRIS [4, 45], TWM [63], and STORM [64] successfully applied transformers to learn dynamics models that achieve state-of-the-art performance in terms of long-horizon predictions. Complementary approaches are inspired by the success of denoising diffusion probabilistic models, which generate plausible samples via a diffusion process, in the context of image

generation. Here, the idea is to learn a generative model over latent trajectories in order to improve long-term predictions by avoiding compounding errors that are inherent in autoregressive approaches. Notable examples in this category include PolyGRAD [65], IRASim [66], and LaDi-WM [50]. In contrast to these approaches, this thesis will develop dynamics models that explicitly capture the underlying causal structure of the environment.

2.1.3 Policy Learning and Planning

A significant body of research has explored how world models can be used to derive behaviours. While the scope of this thesis is primarily on learning world models, for completeness, we will briefly review the main approaches to policy learning and planning using world models. Given a world model, model predictive control methods such as the cross entropy method [67] can be used to iteratively optimise over action sequences to maximise expected rewards or to reach pre-specified goals. Examples of this approach include PlaNet [54] and PETS [68]. These approaches can be made more efficient if the dynamics model has certain structural properties: local linear models [52, 62] allow the use of simple and efficient LQR controllers; differentiable models [69] enable gradient-based optimisation of action sequences; where the dynamics model is a diffusion model, planning can be performed by conditional sampling [70, 71]. In more challenging, long-horizon settings where the planning approach becomes computationally expensive, the Dreamer family [39] and its variants leverage otherwise model-free reinforcement learning algorithms such as soft-actor critic [72] to learn policies in the learnt latent space. Subsequent works such as TD-MPC [2, 73] explored hybrid approaches that combine planning and policy learning. Beyond learning policies for reward maximisation, world models also serve as a natural framework for exploration by providing intrinsic rewards that guide the agent towards novel states [23, 74, 75].

2.2 Structure and Causality

The ability to adapt flexibly and efficiently to novel situations is one of the most distinctive and compelling features of the human mind. It has been conjectured that this capability stems from our ability to not only learn internal models of the world, but also encode generalisable, structural relationships in the environment [16], allowing for selective and targeted adaptation. Despite the aforementioned advances in world models, most existing methods learn *monolithic* models that do not capture structures necessary for systematic adaptation and generalisation. Building on this motivation, we now turn to causality, which offers a principled framework for encoding structure. To this end, causality provides a principled framework for reasoning about structures and generalisation, and has attracted significant interest in the machine learning community in recent years. Richens and Everitt [76], for example, provide a theoretical account of how understanding causal relationships leads to generalisable and robust agents. In this section, we will provide an overview of the key concepts in causal modelling as well as a discussion on how these ideas can foster the desired properties of adaptability, robustness, and interpretability.

We will ground our discussion in the framework of causal graphical models (CGM) [77], on which most of the methods presented in this thesis are based.[1] A key concept in causality is the notion of the *independent causal mechanisms* principle (ICM) [20], which posits that the data-generating process can be decomposed into a set of autonomous modules that do not inform or influence each other. CGMs provide an instantiation of this principle by factorising the joint distribution of a set of random variables into a product of conditional distributions, each corresponding to a causal mechanism. Specifically, a CGM is defined as a set of random variables $\{X_1, \ldots, X_d\}$, their joint distribution P_X, and a directed acyclic graph (DAG), $\mathcal{G} = (X, E)$, where each edge $(i, j) \in E$ implies that X_i is a direct cause of X_j. The

[1] We refer the reader to [19] for a more comprehensive treatment of alternative formulations such as structural causal models.

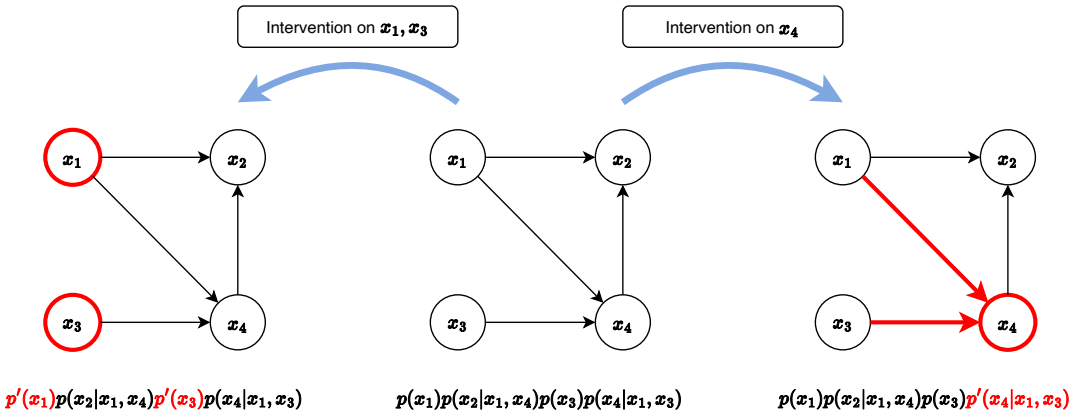


Figure 2.2: An illustration of interventions in a CGM and their corresponding factorised distributions. Interventions are modelled as replacing the conditional distribution of the intervened variable (here illustrated as red lines).

joint distribution admits a causal factorisation such that

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i|PA_i), \tag{2.1}$$

where PA_i is the set of parents to the variable X_i in the graph. Each of the conditional distributions can be considered as an independent *causal mechanism*. Crucially, unlike standard probabilistic graphical models, CGMs support the notion of *interventions*, i.e., local changes in the causal distribution. Formally, an intervention on the variable X_i is modelled as replacing the conditional distribution $p(x_i|PA_i)$ while leaving the other mechanisms unchanged. Given the set of intervention targets $I \subset X$, the interventional distribution can be written as

$$p'(x_1, \ldots, x_d) = \prod_{i \notin I} p(x_i|PA_i) \prod_{i \in I} p'(x_i|PA_i), \tag{2.2}$$

where $p'(\cdot|\cdot)$ is the new conditional distribution corresponding to the intervention.[2] A counterpart of the ICM principle, the *Sparse Mechanism Shift* (SMS) hypothesis [20], posits that naturally occurring distribution shifts tend to correspond to sparse interventions in a causal model when factorized as (2.1), suggesting that only a small number of causal mechanisms are affected by the shift. Causal mechanisms thus tend to be *invariant* across environments [18, 78, 79].

[2]In the causal literature, this replacement is often referred to as a *soft intervention* as a generalisation of a *hard intervention* which sets the variable to a constant value.

To illustrate this idea of sparse distribution shifts intuitively, consider the example of modelling the joint distribution over whether a person carries an umbrella, U, and whether it is raining, R [21]. A causal model that reflects the underlying causal structure, namely that rain causes people to carry umbrellas, would factorise the joint distribution as $p(U, R) = p(R)p(U|R)$. In this setting, a natural distribution shift might be a change in the weather patterns, which can be accounted for by adapting the marginal distribution $p(R)$ while leaving the conditional distribution $p(U|R)$ unchanged. Similarly, a change in behavioural patterns in umbrella usage would correspond to a change only in $p(U|R)$. In contrast, a non-causal factorisation $p(U, R) = p(U)p(R|U)$ would not afford such targeted adaptation, since a change in either the weather or umbrella usage would require adapting both $p(U)$ and $p(R|U)$. This simple example demonstrates how causal mechanisms tend to be resilient to naturally occurring distribution shifts and afford selective adaptation.

The notion of sparse distribution shifts underpins the conceptual motivation of this thesis: learning world models that are composed of independent causal mechanisms can facilitate modular adaptation to environment changes. By the same token, the invariance of causal mechanisms across environments can also improve robustness in cases where the intervention targets are not related to the task at hand, such is the case when removing distractors in the scene [17]. In addition, the explicit representation of graph structures enhances the interpretability of the learnt models by providing insights into the relationships between variables, and diagnostic cues in case the model is inaccurate. In summary, causality provides not only a conceptual lens but also concrete modelling principles that guide the structured world models developed in this thesis. We note that the causal framework also supports counterfactual reasoning, see for example [25, 80, 81], but exploring this aspect lies beyond the scope of this thesis. In the next section, we will discuss the challenges of learning structured models from observations and situate our contributions within the broader landscape of related works.

2.3 Related Works

To date, many approaches have explored ways to incorporate structure into machine learning models. The methods proposed in this thesis take inspiration from and build upon a wide range of prior works in the context of world models, causal machine learning, and in the broader machine learning literature. In this section, we will situate our contributions within this landscape of related works, organising our discussion around three core challenges in learning structured world models: *learning representations*, *learning independent mechanisms*, and *structure discovery*, which broadly correspond to Chapters 3, 4, and 5 respectively. Finally, we provide a discussion on how the approaches developed in this thesis can be viewed within a Bayesian framework as a form of structured prior, and how this perspective contextualises our contributions within the broader landscape of research on inductive biases and priors in machine learning.

2.3.1 Representation Learning

Choosing a suitable representation model that maps observations into latent representations is a significant design choice in any world model. Unlike the methods reviewed in Sec. 2.1 which rely on monolithic representations that encode observations into a single unstructured latent vector, discovering causal structure requires high-level, semantically meaningful *variables*. This idea of learning disentangled latent representations that explicitly capture the underlying generative factors has been a long-standing goal in machine learning [82–84]. Yet, learning disentangled representations without supervision remains a major challenge to date. The original VAE framework [44] and subsequent works, such as β-VAE [85] and FactorVAE [86], induce disentanglement by enforcing independence amongst latent dimensions, which have successfully disentangled factors of variation in simple synthetic datasets. However, the success of these methods is limited to simple datasets with strong statistical independence between factors of variation.

Inspired by the intuition that natural images can be factorised into discrete objects, more recent approaches in *object-centric generative modelling* learn to

decompose scenes into sets of latent slots, each intended to capture an individual object or part. This is primarily achieved through enforcing architectural constraints that establish correspondences between latent slots and pixels, and learning the generative distribution over slots either autoregressively [87, 88] or via iterative refinement [89, 90], which relaxes the independence requirement. Extensions to these works further integrate temporal dynamics for video-based segmentation [91], support multiple input views [92], contrastive loss formulations [93], and further disentangle the positional information from properties of objects [94–96]. The success of object-centric representation learning has inspired a class of world models with dynamics models that act on object slots [97–100], demonstrating improved prediction accuracy and generalisation performance. The methods proposed in Chapter 4 and Chapter 5 belong to this family of work, with a focus on learning and independent mechanisms that act on these object-centric representations.

The drawback of the object-centric approach is that it often requires hand-crafted heuristic inductive biases based on our subjective notion of what good abstractions should be. To this end, works in the field of non-linear Independent Component Analysis [101] and Causal Representation Learning [20] take a more theoretically grounded perspective to understanding how and when disentangled representations emerge. A key result in this space is that disentanglement is, in general, *impossible* without any inductive biases on both the models and the dataset, as there are infinitely many ways to represent observations [102]. Subsequent works investigate the necessary conditions for disentanglement to occur through theoretical identifiability results that elucidate how different inductive biases can constrain the solution. Prior works investigate a wide range of potential inductive biases, which can be broadly categorised into two groups. The first group considers constraints on the model class, such as enforcing additive structures on the decoders [103], using specific object-centric encoders and decoders [104, 105], and enforcing block-diagonal Hessian structures that limit the interactions between latent dimensions [106]. The second group of works investigates cases where extra information about the data is

available. This includes weak supervision provided as contrastive pairs where only a sparse subset of the underlying factors of variation are changed [107–109].

In the causal modelling framework, extra information can also be formulated as intervention targets, i.e., labels for factors of variation that exhibit different dynamics in each environment. Known intervention targets [110, 111], unknown binary interventions [112], or access to multi-task or multi-view data [113, 114] can be used as training signals for disentanglement. Assumptions on the dynamics of the latent factors, such as sparse interactions between factors [115] and known equivariances [116], together with model architectures that reflect these structures, can also lead to disentanglement. In Chapter 3, we bridge the gap between these theoretical insights and the practical constraints inherent to the context of world models. In particular, drawing from the sparse interaction and sparse intervention framework proposed in [115], we propose the VCD architecture, demonstrating that these inductive biases can be effectively realised in a practical world model to achieve improved adaptation efficiency.

2.3.2 Learning Independent Mechanisms

We now turn our attention to the dynamics model that predicts the evolution of latent representations over time. Analogous to how observations can be factorised into factors of variations or objects, the dynamics of an environment can also be decomposed into a set of *independent mechanisms* that govern how these factors evolve over time. Humans naturally understand the evolution of the environment through reusable rule-like mechanisms, i.e. *how* objects interact: objects collide, blocks stack, balls roll, fluids flow, and so on. Such organisation of knowledge into high-level, recomposable abstract concepts has been argued to be a key enabler of systematic generalisation [117]. This has inspired the development of architectures that explicitly decompose dynamics into a collection of independently acting mechanisms, each focusing on a different aspect of the environment's dynamics. SLDS [118] uses a variational approach to learn to pick different dynamics models conditioned on state, but is limited to linear transitions. A series

of works subsequently investigates more expressive architectural instantiations of the independent mechanism idea. RIMs [119] use independent and sparsely interacting recurrent units as dynamics models. SCOFF [120] proposes an architecture that contains learnable modules that resemble the structure of object-oriented programming languages. VIM [121] uses a variational approach to disentangle mechanisms and objects in the setting where objects move independently of each other. NPS [122] learns a set of independently parameterised mechanisms and binds objects to these mechanisms via attention. These works provide a proof-of-concept that showcases the potential of parameterising dynamics as independent mechanisms, demonstrating improved generalisation and interpretability in simple environments.

In parallel, similar architectural motifs have gained significant traction in the broader machine learning literature, in the form of Mixture-of-Expert models [123, 124] in large-scale frontier models [125–127]. Originally designed to improve the scalability of large models by activating only a subset of the model at inference time, these architectures consist of a set of expert modules and a router that selects which experts to activate for each input. Empirical evidence [126, 128, 129] shows that these modules often specialise in specific tasks, i.e. different tasks such as coding, mathematics, and legal reasoning tend to activate different experts. This emergent specialisation suggests that, to some extent, given the right architectural inductive biases, large-scale models also learn to decompose knowledge into independently acting modules.

Yet, a principal challenge remains: in the context of world modelling, architectural biases alone rarely lead to the emergence of disentangled mechanisms that correspond to meaningful physical interactions. In Chapter 4, we empirically demonstrate that architectural bias alone is insufficient to recover disentangled mechanisms in complex environments. To this end, we draw inspiration from prior works in the context of learning independent causal mechanisms. Parascandolo et al. [130] first propose a competition of experts algorithm that induces the emergence of mechanisms that explain transformations in images. This idea of utilising a competitive training scheme on modular model architectures has subsequently

been applied to diverse settings such as lifelong learning [131, 132], generative models [133], and object-centric scene composition [134]. Building on this line of work, our proposed method employs a competitive training scheme as an inductive bias for disentangling modes of interaction in the setting of world model learning.

2.3.3 Structure Discovery

Another key challenge in learning structured dynamics models is discovering the structural relationships between entities. In this space, a significant body of works leverages object-centric representations and captures interactions between objects via various forms of message passing schemes. Examples include C-SWM [97], RFM [135], and OP3 [100], which use graph neural networks to explicitly model interactions. More recently, SlotFormer [98] combines object-centric representations with a transformer-based dynamics model to achieve state-of-the-art performance in terms of long-term predictions. However, these approaches rely on fully-connected graph structures, which do not fully exploit the sparsity of the underlying interactions between objects: in many physical systems, objects and entities only interact with a small subset of other objects. As discussed in the previous section on the role of causality, a key aspect of the arguments put forward by this thesis is that such *sparse* and *causal* interactions are key to enabling robustness and generalisation.

To this end, we take inspiration from the causal discovery literature, which aims to recover sparse causal graphs from data, i.e. the set of causal parents for each variable in Eq. 2.2. Here, approaches can be categorised as constraint-based [e.g. 136], which use statistical independence tests to rule out incompatible graph structures, and score-based [e.g. 137], which search for optimal graph structures based on a scoring criterion.[3] In this space, an emergent class of methods formulates causal discovery as a continuous optimisation problem [e.g. 22, 138, 139], which serves as the backbone of the proposed methods in this thesis as they can be readily integrated into the gradient-based learning framework of world models. Specifically, in Chapter 3, we draw from the Differentiable *Causal Discovery from Interventional*

[3]The reader is referred to [19] for a detailed review of causal discovery methods.

data (DCDI) framework [22], and leverage its sparsity regularisation as an inductive bias for learning disentangled representations.

However, using a fixed causal graph often fails to accommodate realistic scenarios where the interaction patterns are highly time-varying: objects may move in and out of each other's proximity, and the nature of their interaction can change over time. This motivates the development of the *local causal model* framework [23–25], which infers state-dependent interaction graphs and demonstrates improved interpretability, sample efficiency, and robustness in simple state-based environments. The CoDA family of works [25, 140] uses local causal graphs to enable the generation of counterfactual data for data augmentation. Yet, estimating a local causal model from data remains challenging. Several approaches to learning local causal structures have been proposed. CAI [141] uses conditional mutual information to estimate the local influence an agent has on the environment. ELDEN [23] uses sparsity-regularised partial derivatives between states based on a learnt dynamics model to infer local causal graphs, while FCDL [24] uses sparsity regularisation to learn a quantised codebook of possible causal connections. However, these methods often fall short in terms of scalability and versatility: for example, ELDEN requires scalar state representations, and FCDL can only operate on a fixed number of objects. To remedy this, in Chapter 5, we propose SPARTAN, which performs local causal discovery using a modified transformer architecture that reliably infers local causal relationships while inheriting the flexibility and scalability of standard transformer models.

2.3.4 Causal Structure as a Prior

One general perspective on learning world models, and machine learning more broadly, is that of Bayesian inference, in which learning can be viewed as inferring a posterior distribution over the space of possible world models given the observed data. In this framework, inductive biases can be viewed as *priors* that express assumptions about the structure of the world and guide learning towards more robust and generalisable models. Through this lens, the methods developed in this thesis

can be interpreted as imposing an *explicit* prior for causal structure: namely, that a good world model should admit latent representations whose dynamics decompose into a small collection of reusable, sparsely interacting, and independent mechanisms.

Viewed in this way, a preference for causal structure can also be interpreted as a bias towards simplicity and modularity. This naturally connects to Occam's razor and to the minimum description length (MDL) principle [142], which suggests that models that explain the data more compactly are more likely to generalise well. This relationship between compression and causal structure is further explored in work on algorithmic causality [143, 144], and offers a useful theoretical perspective on the methods developed in this thesis.

A natural question, however, is whether such a prior for causal structure must be imposed *explicitly*. Recent work [145, 146], aimed at explaining the success of modern large-scale models, argues that the training process itself may induce an implicit bias towards simple and compressible solutions. For example, Saxe et al. [147] describe how the training dynamics of neural networks can naturally give rise to competitive dynamics similar to those explicitly encouraged in Chapter 4. Throughout this thesis, we provide empirical evidence that, in the context of world model learning, making this structural bias explicit leads to improved adaptability, robustness, and interpretability, and that these properties are not reliably recovered from the implicit biases of standard architectures and training procedures alone.

In a similar vein, recent work has also investigated the use of large-scale pre-trained foundation models as a form of prior for learning world models. These approaches leverage the rich representations and knowledge encoded in these models by using them as structured latent spaces for downstream world model learning [49, 50]. Unlike these approaches, which rely on broad statistical biases encoded in the pre-trained model, this thesis focuses on encoding *targeted* structure that reveal specific causal mechanisms of the environment, which can be more directly leveraged for systematic adaptation and interpretation. It is important to note, however, that these approaches are complementary: hybrid approaches that combine pre-trained

representations with the explicit structural biases developed in this thesis offers a promising direction for future research.

2.4 Summary

In this chapter, we reviewed the key components of world models, namely representation learning and dynamics modelling, and discussed how the causality framework provides a principled approach to learning structured world models that address the shortcomings of existing methods, namely the lack of adaptability, robustness, and interpretability. We provide a comprehensive review of related works that share our goal of incorporating structure into world models and situate our contributions within this landscape of prior works.

3
Variational Causal Dynamics

To recapitulate the overarching aim of this thesis, we seek to build world models that reflect the causal structure of the environment, and demonstrate that these structures can be leveraged to facilitate efficient modular adaptation. In this chapter, we begin by building on existing latent state-space world models and incorporating a causal structure in the dynamics model. To see how causal structures lead to sample-efficient adaptation to changes in the environment, recall the structure of a CGM (with interventions) from Chapter 2:

$$p'(x_1, \ldots, x_d) = \prod_{i \notin I} p(x_i | PA_i) \prod_{i \in I} p'(x_i | PA_i), \tag{3.1}$$

where I is the set of intervention targets, and PA_i are the parents of the variable x_i in the causal graph. Within this framework, distribution shifts can be modelled as interventions on these causal mechanisms, i.e. replacing the corresponding conditional distributions. The SMS hypothesis suggests that naturally occurring distribution shifts tend to manifest as *sparse* interventions, implying that many causal mechanisms remain invariant across domains. As such, given observations from a new environment, a causally factored world model can facilitate efficient adaptation by identifying the sparse set of intervened mechanisms and re-learning only those, while reusing the invariant mechanisms.

The key challenge here is: *how can we learn a world model with a causal structure from high-dimensional observations?* In the context of world model learning, we typically only have access to sequences of high-dimensional observations, such as images, from potentially multiple environments with different dynamics. Mapping this back to the CGM frameworks, this means that the world model needs to learn the *graph structure* (i.e. the causal relationships between variables), the *intervention targets* (i.e. how each environment differs from each other), the *causal mechanisms* (i.e. the conditional distributions), and more fundamentally, the *causal variables* (i.e. inferring the state from observations). While causal discovery methods [22, 77, 136, 138] can be leveraged to learn the graph structure and mechanisms, to date, these methods require access to the abstract causal variables, which are not typically available in the context of world model learning.

In order to benefit from the structure of causal models and the ability to represent high-dimensional observations, we propose *Variational Causal Dynamics* (VCD), which combines causal discovery with variational inference. Specifically, we train a latent state-space model with a structural dynamics model using variational inference and sparsity regularisation from causal discovery. An illustration of the overall approach is shown in Fig. 3.1. By jointly learning a latent representation of the observations and a dynamics model, VCD learns a causally factorised world model that can modularly adapt to different environments. The key intuition behind our approach is that, since sparse structures can only be discovered on semantically meaningful variables, training the representation and the causal discovery module in an end-to-end manner acts as an inductive bias that encourages well-disentangled representations. By leveraging the learnt causal structure, VCD is able to identify the sparse mechanism changes in the environment and re-learn *only* the intervened mechanisms. Empirically, we demonstrate that VCD is able to learn disentangled representations and identify the causal structures that correspond to the underlying generative rules. In so doing, VCD is able to adapt efficiently to changes in the environment, significantly outperforming monolithic baselines.

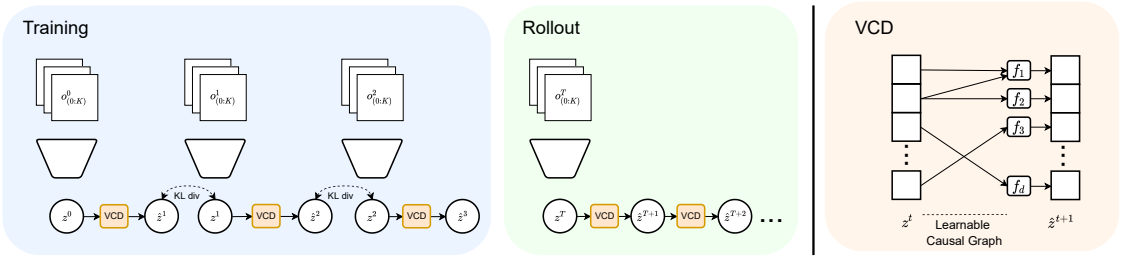


Figure 3.1: Left: The general architecture of VCD. The dynamics model is trained using observation sequences from multiple environments by minimising the KL divergence between the predicted state distribution and the encoded posterior state distribution. Rollouts in the latent space can be performed by recursively applying the learnt dynamics model. Right: The structure of the causal dynamics model. Each dimension of the latent space is treated as a causal variable. Predictions are made using only the causal parents of each variable, according to a *learnt* causal graph.

3.1 Background

This section describes the prerequisite definitions and formulations for our proposed method. Specifically, we give an overview of recurrent state space models (RSSM) [54] and differentiable causal discovery methods, and motivate our approach that builds upon these to learn causal world models.

Latent state-space models In a complex environment with high-dimensional observations, such as images, learning a compact latent state space that captures the dynamics of the environment has been shown to be more computationally efficient than learning predictions directly in the observation space [54, 148]. Given a dataset of sequences $\{(o^{0:T}, a_i^{0:T})\}_{i=0}^{N}$[1], with observations o^t and actions a^t at discrete timesteps t, a generative model of the observations can be defined using latent states $z^{0:T}$ as

$$p(o^{0:T}, a^{0:T}) = \int \prod_{t=0}^{T} p_\theta(o^t|z^t)p(a^t|z^t)p_\theta(z^t|z^{t-1}, a^{t-1})dz^{0:T}, \tag{3.2}$$

where $p_\theta(o^t|z^t)$ and $p_\theta(z^t|z^{t-1}, a^{t-1})$ are the observation model and the dynamics model respectively. For simplicity, we do not learn the policy term $p(a^t|z^t)$ and

[1] In the current work, we focus on environments without rewards. However, the proposed method can be readily extended to include reward prediction.

omit it throughout this paper as it is constant with respect to the parameter θ. The variational evidence lower bound can be written as

$$\begin{aligned}\mathbb{ELBO}(\theta,\phi) = &\sum_{t=0}^{T} \mathbb{E}_{q_\phi(z^t|o^t)}\Big[\log(p_\theta(o^t|z^t))\Big] \\ &- \mathbb{E}_{q_\phi(z^{t-1}|o^{t-1})}\Big[\mathbb{KL}[q_\phi(z^t|o^t)||p_\theta(z^t|z^{t-1},a^{t-1})]\Big],\end{aligned} \tag{3.3}$$

where $q_\phi(z^t|o^t)$ is a learnable approximate posterior of the observations.[2] See Appendix A.1 for the derivation. All conditional distributions are modelled as diagonal Gaussian distributions with the mean and variance parameterised by neural networks. The model parameters can be learnt by maximising the ELBO using the reparameterisation trick [44].

Whilst this formulation of the latent state-space model has the capacity to capture complex scene dynamics, in practice, fully stochastic models struggle to make long-horizon predictions. In order to alleviate this problem, RSSM employs a flexible dynamics model parameterised as a fully connected recurrent neural network, where the transition probability is split into a stochastic part and a deterministic recurrent part,

$$z^t \sim p_\theta(z^t|h^t), \quad h^t = f_\theta(h^{t-1}, z^{t-1}, a^{t-1}), \tag{3.4}$$

where $f(\cdot)$ is instantiated as a GRU [149] and h^t is the associated hidden state. Intuitively, this provides a path through which information can be passed on over multiple timesteps.

Despite their success, these models cannot reason about changes in the environment. Specifically, they are unable to structurally utilise prior knowledge learnt from different environments under distribution shifts. To this end, we argue that imposing a *causal* structure on the dynamics model equips the learning agent with the ability to adapt to changes in a modular and efficient manner.

[2]The approximate posterior, which only depends on the current observation, is a simplification of the full posterior, which is conditioned on all observations. This is consistent with prior works and is found to be sufficient for the dynamical systems in the experiment. However, the algorithm presented in this work can be readily adapted to use the full posterior instead.

Differentiable causal discovery In order to learn a causal model in the context of world models, we draw inspiration from recent advances in causal discovery, which aim to learn causal structures from data. We focus on methods that formulate causal discovery as a continuous optimisation problem [22, 138, 150] as these can be naturally incorporated into the variational inference framework. Furthermore, since the causal variables are *learnt* in our model, meaning that we do not have access to what each variable corresponds apriori, a key desiderata for the causal discovery algorithm is the ability to learn causal graphs from unknown intervention targets. In this work, we follow the formulation in Differentiable Causal Discovery with Interventional data (DCDI) [22], which optimises a continuously parameterised probabilistic belief over graph structures and intervention targets.

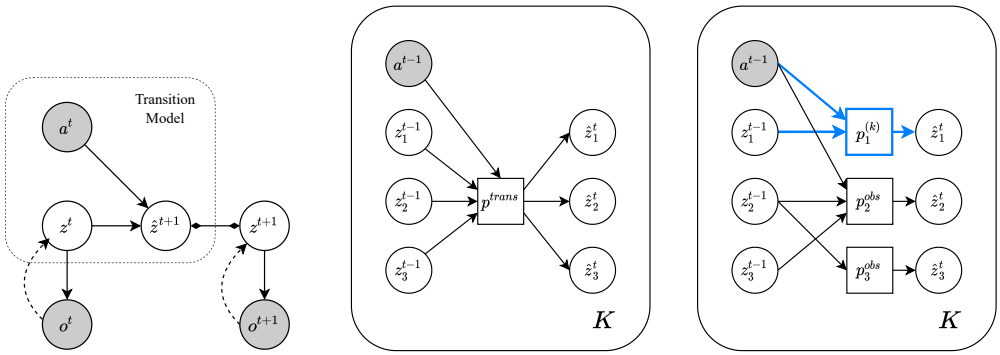
Given samples from an observed data distribution $P_X^{(0)}$ and K intervened distributions $P_X^{(k)}$, DCDI optimises a probabilistic belief over causal graphs $\mathcal{G}$ and intervention targets $\mathcal{I}$. Specifically, the distribution over $\mathcal{G}$ and $\mathcal{I}$ is parameterised by random binary matrices where each entry follows an independent Bernoulli distribution, and encodes the existence of a causal edge or an intervention. Under this parameterisation, causal discovery can be formulated as maximising the expected data log likelihood with sparsity regularisation,[3]

$$L(\theta, \alpha, \beta) = \mathbb{E}_{\alpha,\beta}\left[\sum_{k=1}^{K} \log[p_\theta^{(k)}(x_1^k, \ldots, x_d^k; \mathcal{G}, \mathcal{I})] - \lambda_G|\mathcal{G}| - \lambda_I|\mathcal{I}|\right], \tag{3.5}$$

where $p^{(k)}$ is the data likelihood under causal graph $\mathcal{G}$ and intervention targets $\mathcal{I}$ in the kth environment. Each conditional distribution is parameterised as a feedforward neural network with parameter θ; α and β are the Bernoulli parameters for $\mathcal{G}$ and $\mathcal{I}$, and $\lambda_{G,I}$ are hyperparameters to control sparsity. This training objective can be optimised by estimating gradients through the expectation using the Gumbel-Softmax trick [151].

In the context of learning world models from high-dimensional observations, the drawback of this approach, and of causal discovery methods in general, is that

[3] In the original DCDI framework, this is also subject to an acyclicity constraint. This is not necessary in the context of our work as we assume there are no instantaneous causal effects (i.e., within a timestep).



(a) Latent state space model **(b)** Typical dynamics model **(c)** VCD dynamicsmodel

Figure 3.2: (a) The structure of latent state space models in one timestep. Solid lines denote the generative process of the environment, dashed lines denote approximate posteriors, shaded nodes denote observed variables, and the bi-directional arrow denotes the KL divergence in the ELBO term. In this work, we propose a novel architecture for the dynamics model. (b) The structure of latent dynamics models where the probability distribution is modelled as a fully connected neural network. (c) The causal dynamics model in VCD. Each individual variable has a separate conditional distribution p_i. Each mechanism is conditioned on a subset of the previous state and action. The blue lines highlight the intervened mechanism which is specific to the environment k, as opposed to the mechanisms corresponding to the black lines, which are shared across environments.

it requires access to semantically meaningful causal variables, much like classical AI required symbols in terms of which algorithms could be formulated. In the next section, we present a method to perform causal discovery and learn causally meaningful representations jointly.

3.2 Variational Causal Dynamics

In this section, we present Variational Causal Dynamics (VCD). Similar to causal discovery with interventional data, VCD learns from action-observation sequences from an undisturbed environment, $(o_{(0)}^{0:T}, a_{(0)}^{0:T})$, and K intervened environments, $\{(o_{(k)}^{0:T}, a_{(k)}^{0:T})\}_{k=1}^{K}$. Throughout this chapter, we assume that 1. changes between the environments are due to 'soft' interventions where the structure of the causal graphs remains the same but individual parameters of the mechanisms change, 2. there are no instantaneous causal dependencies in each timestep, and 3. the observation function does not change across environments, i.e. the appearance of the entities remains the same such at we can use the same encoder. The general approach of

VCD follows the latent state-space model framework (Fig.3.2a), where a latent representation of the observations is jointly learnt with a dynamics model by maximising the ELBO. In contrast to existing approaches that parameterise the transition probability $p(z^t|z^{t-1}, a^{t-1})$ as a feedforward neural network (Fig.3.2b), VCD learns a *structured* dynamics model by utilising inductive biases from causal discovery. Importantly, our approach is grounded in the hypothesis that causal structures can only be discovered on meaningfully disentangled representations of the system. We therefore argue that training a representation and a dynamics model jointly to optimise a causal discovery objective can lead to a latent representation that affords sparse dynamics models and is therefore semantically meaningful. We discuss this approach in relation to existing theoretical works on identifiability in Section 3.2.5. Taking the view that changes in the dynamics can be attributed to sparse causal interventions, we posit that a causally factorised dynamics model facilitates modular adaptation to new environments by explicitly leveraging the invariance of causal mechanisms. We propose an algorithm for adaptation in Section 3.2.4.

3.2.1 Causal dynamics model

In order to perform causal discovery on the transition dynamics, the dynamics model in VCD is designed to mimic the structure of a CGM (Eq. 3.1). In the following paragraphs, we motivate and describe the three main architectural features of VCD that enable causal discovery: *independent mechanisms*, *sparse causal dependencies*, and *sparse interventions*. Access to the causal graph $\mathcal{G}$ and the intervention targets for each environment $\mathcal{I}_k$ is assumed throughout this subsection. We describe the method of learning these jointly with the model in Section 3.2.3.

Independent mechanisms In contrast to a fully-connected dynamics model, the transition probability is factorised into individual conditional distributions as

$$p(z^t|z^{t-1}, a^{t-1}) = \prod_i^d p_i(z_i^t|z^{t-1}, a^{t-1}), \tag{3.6}$$

where d is the dimension of the latent space, to be set as a hyperparameter. z_i denotes the ith dimension of the latent state, and each conditional distribution p_i

is a one-dimensional normal distribution with mean and variance given by *separate* neural networks. This factorisation and separation of parameters can be seen as an instantiation of the ICM principle. This explicit modularity of the model structure enables the notion of interventions, where individual conditional distributions are locally changed without affecting the other mechanisms.

Sparse causal dependencies Following the structure of a CGM, we condition each variable only on its causal parents according to the learnable causal graph $\mathcal{G}$, rather than the full state. Given a graph $\mathcal{G}$, we define the binary adjacency mask $M^{\mathcal{G}}$ where the entry $M^{\mathcal{G}}_{ij}$ is 1 if and only if $[z^{t-1}, a^{t-1}]_i$ is a causal parent of z^t_j. This is consistent with the intuition that, in physical systems, states interact with each other in a sparse manner [119], and actions tend to have a direct effect on only a subset of the states. Under this parameterisation, the causal transition probability can be written as

$$p(z^t|z^{t-1}, a^{t-1}) = \prod_i^d p_i(z^t_i|M^{\mathcal{G}}_i \odot [z^{t-1}, a^{t-1}]), \tag{3.7}$$

where $\odot$ denotes the element-wise product, $[\cdot, \cdot]$ denotes the concatenation of vectors, and $M^{\mathcal{G}}_i$ is the binary mask that selects only the causal parents of z_i under the graph $\mathcal{G}$.

Sparse interventions Following the SMS hypothesis [20], we assume that changes in distributions across the K intervened environments are due to sparse interventions in the ground truth causal generative process. In order to incorporate sparse interventions in VCD, given the set of learnt intervention targets $\mathcal{I}_k$ for each environment, we define the binary intervention mask $R^{\mathcal{I}}$ where the entry $R^{\mathcal{I}}_{ki}$ is 1 if and only if the variable z_i is in the set of intervention targets in environment k. For each variable z_i, $R^{\mathcal{I}}_{ki}$ acts as a switch between reusing a shared observational model and an environment-specific interventional model. The full interventional causal model of the transition probability in the environment k can be written as

$$p^k(z^t|z^{t-1}, a^{t-1}) = \prod_i^d p^{(0)}_i(z^t_i|M^{\mathcal{G}}_i \odot [z^{t-1}, a^{t-1}])^{1-R^{\mathcal{I}}_{ki}} p^{(k)}_i(z^t_i|M^{\mathcal{G}}_i \odot [z^{t-1}, a^{t-1}])^{R^{\mathcal{I}}_{ki}}, \tag{3.8}$$

where $p^{(0)}$ is the observational mechanism that is shared across all environments and $p^{(k)}$ is the intervened distribution specific to environment k. Intuitively, this model ensures that all environments reuse the same conditional distributions p^{obs} unless it is deemed that a particular mechanism has been intervened on (i.e. $R^{\mathcal{I}}_{ki} = 1$). This modular adaptation between environments facilitates structural transfer of knowledge between different environments. Note that the structure is that of a causal graphical model. The overall architecture of the VCD dynamics model is summarised in Fig. 3.2c.

3.2.2 Recurrent module

Similar to RSSM [54], we augment the model with a deterministic recurrent path to enable long-term predictions. To ensure that each conditional distribution only has access to the causal parents, in the same way that each conditional distribution is modelled by a separate network, each conditional distribution keeps a separate recurrent unit and a corresponding hidden activation:

$$z_i^t \sim p_i(z_i^t | h_i^t), \quad h_i^t = f_i(h_i^{t-1}, M_i^{\mathcal{G}} \odot [z_i^{t-1}, a_i^{t-1}]), \tag{3.9}$$

where f_i is a recurrent module specific to the variable z_i, instantiated as a GRU [149].

3.2.3 Training

The task of model training is to determine the model parameters θ, the approximate posterior parameters ϕ, as well as the causal graph $\mathcal{G}$ and the intervention targets $\mathcal{I}$. These can be jointly trained in a way similar to DCDI. We parameterise the belief over the causal adjacency matrix $M^{\mathcal{G}}$ as a random binary matrix. Each entry $M^{\mathcal{G}}_{ij}$ follows a Bernoulli distribution with success probability $\sigma(\alpha_{ij})$, where α_{ij} is a scalar parameter and $\sigma(\cdot)$ is the sigmoid function. Similarly, a random binary matrix $R^{\mathcal{I}}$ is parameterised using the scalar variable β_{ki} for each entry. Unlike DCDI, due to the existence of latent variables, we cannot directly maximise the data likelihood. Instead, we maximise the expected ELBO across all environments

over causal graphs and intervention masks,

$$L(\theta, \phi, \alpha, \beta) = \sum_{k \in [0,1,\dots,K]} \mathbb{E}_{\mathcal{G},\mathcal{I}} \Big[\mathbb{ELBO}(o^{0:T}_{(k)}, a^{0:T}_{(k)}; \theta, \phi, \mathcal{G}, \mathcal{I}) - \lambda_G |\mathcal{G}| - \lambda_I |\mathcal{I}| \Big], \quad (3.10)$$

where

$$\begin{aligned} \mathbb{ELBO}(o^{0:T}_{(k)}, a^{0:T}_{(k)}; \theta, \phi, \mathcal{G}, \mathcal{I}) = & \sum_{t=0}^{T} \mathbb{E}_{q_\phi(z^t|o^t)} \Big[p_\theta(o^t|z^t) \Big] \\ & - \mathbb{E}_{q_\phi(z^{t-1}|o^{t-1})} \Big[\mathbb{KL}[q_\phi(z^t|o^t) || p^{(k)}_\theta(z^t|z^{t-1}, a^{t-1})] \Big]. \end{aligned} \quad (3.11)$$

$p^{(k)}_\theta(z^t|z^{t-1}, a^{t-1})$ is further factorised as in Equation (3.8). The gradients through the outer expectation and the expectation term in ELBO are estimated using the straight-through Gumbel-max trick [151] and the reparameterisation trick [44], respectively. Note that the training process can also be, in theory, implemented as a L1 regularisation on the causal graph and the intervention targets, where the binary masks are replaced with continuous values. However, this is found to be less performant in early experiments, as well as in previous work on differentiable causal discovery [22]. One explanation is that the binary masks provide a way to explicitly set some parameters to be exactly zero, whereas with continuous masks, the model can adjust edge weights to be small but non-zero, which can be counterbalanced by increasing the weights of the subsequent layers. Hence, in this work, and for the rest of the thesis, we opt to use the sampled binary mask approach for training.

3.2.4 Adaptation

Owing to the modular nature of the dynamics model, VCD can naturally adapt to new, unseen environments by jointly inferring the intervention targets and the new model parameters for the intervened mechanisms. Specifically, the dynamics model in a new environment can be written as

$$p^{new}(z^t|z^{t-1}, a^{t-1}) = \prod_i p^{(0)}_i(z^t_i | M^{\mathcal{G}}_i \odot [z^{t-1}, a^{t-1}])^{1-R'_i} p'_i(z^t_i | M^{\mathcal{G}}_i \odot [z^{t-1}, a^{t-1}])^{R'_i}, \quad (3.12)$$

where R' is the intervention mask for the new environment and p'_i is the environment-specific distribution. The parameters of $p^{(0)}_i$ are carried over from the training

environments, meaning that the model reuses trained mechanisms unless the variable is intervened on in the new environment. Under the SMS hypothesis, only a small subset of the mechanisms needs to be adapted in a new environment. As such, by leveraging past experience from previous environments, VCD can adapt more quickly to environment change. We empirically verify this in Section 3.3.4.

One way to implement modular transfer is to train VCD on trajectories in the new environment while reusing the trained parameters for the causal graph and the mechanisms in the undisturbed environment. This can be done by identifying the set of intervention targets in the new environment I' and learning the model parameters for the corresponding conditional distributions θ'. These can be jointly trained in an analogous way to Eq. 3.10,

$$L^{adapt}(\theta', \beta') = \mathbb{E}_{\mathcal{G},\mathcal{I}'}\Big[\mathbb{ELBO}(o^{new}_{0:T}, a^{new}_{0:T}; \theta', \phi, \mathcal{G}, \mathcal{I}') - \lambda_I|\mathcal{I}'|\Big], \tag{3.13}$$

where β' are the Bernoulli parameters for the distribution over intervention targets. Under this framework, VCD has the capability to identify the sparse mechanism changes in a new environment and learn the transition mechanisms only for the intervened variables. In Section 3.3, we show that VCD can identify sparse interventions from a small amount of data and is able to refit the world model with improved data efficiency.

3.2.5 Identifiability of Causal Graph and Causal Representations

VCD's ability to adapt in a modular fashion hinges on the discovery of the correct, or at least approximately correct, causal structure as well as on a causally meaningful representation space. The conditions under which such structures can be learnt from data are studied under the term *identifiability*. While we make no formal claims about the identifiability of causal graphs and representation in the thesis, our approach is grounded in recent theoretical results that make use of the sparsity of causal relationships [22, 115]. In this subsection, we foreground the relationship between these results and our approach.

In the DCDI framework, Brouillard et al. [22] show how the correct causal graph can be identified in a setting similar to ours. Given observed causal variables X in k different environments with intervention targets $\mathcal{I}^*$ and ground-truth causal graph $\mathcal{G}^*$, the following score function is proposed,

$$\mathcal{S}(\mathcal{G}, \mathcal{I}) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X\ p^{(k)}} \log f^{(k)}(X; \mathcal{G}, \mathcal{I}, \phi) - \lambda_{\mathcal{G}}|\mathcal{G}| - \lambda_{\mathcal{I}}|\mathcal{I}|, \tag{3.14}$$

where $f^{(k)}$ is the likelihood of the data under the model in environment k, parameterised by ϕ, and other definitions are the same as eq.3.10. The main theoretical results in [22] show that with small enough $\lambda_{\mathcal{G}}$ and $\lambda_{\mathcal{I}}$, maximising this score function will result in the correct intervention targets and a causal graph that is $\mathcal{I}^*$-*markov equivalent* to the ground-truth graph. In other words, with a suitable choice of hyperparameters, maximising the score function will lead to the correct causal graph and intervention targets. The main idea of this theorem is that if the model learns the underlying distribution perfectly, any learnt graph that is not equivalent to the ground-truth graph is necessarily less sparse and hence has a lower score. VCD uses the same score function as the training objective, with the log likelihood replaced with the evidence log lower bound since we do not have access to the causal variables.

To this end, Lachapelle et al. [115] consider the identifiability of causal variables from high-dimensional observations, i.e. whether an encoder can learn latent representations that correspond to causally meaningful variables. Given a dataset of observation-action sequences, $\{o^t, a^t\}_{t=0}^{T}$ generated by the dynamics model:

$$o^t = \mathbf{f}(z^t), \quad z^t = \prod_i p(z_i^t | z^{<t}, a^{<t}), \tag{3.15}$$

$$p(z_i^t | z^{<t}, a^{<t}) = h_i(z_i^t) \exp\{\mathbf{T}_i(z_i^t)^T \boldsymbol{\lambda}_i(G_i^z \odot z^{<t}, G_i^a \odot a^{<t}) - \psi_i(z^{<t}, a^{-t})\}, \tag{3.16}$$

where $\mathbf{f}$ is an invertible function that maps the latent causal variables to high-dimensional observations. The theory holds conditional distributions within the exponential family, where $\mathbf{T}$ is the sufficient statistics, $\boldsymbol{\lambda}$ is a transition function that outputs the natural parameters of the distribution, h and ψ are the base measure and the normalising constant, respectively. G^z and G^a are the binary masks corresponding to the causal graph of the observed system, analogous to

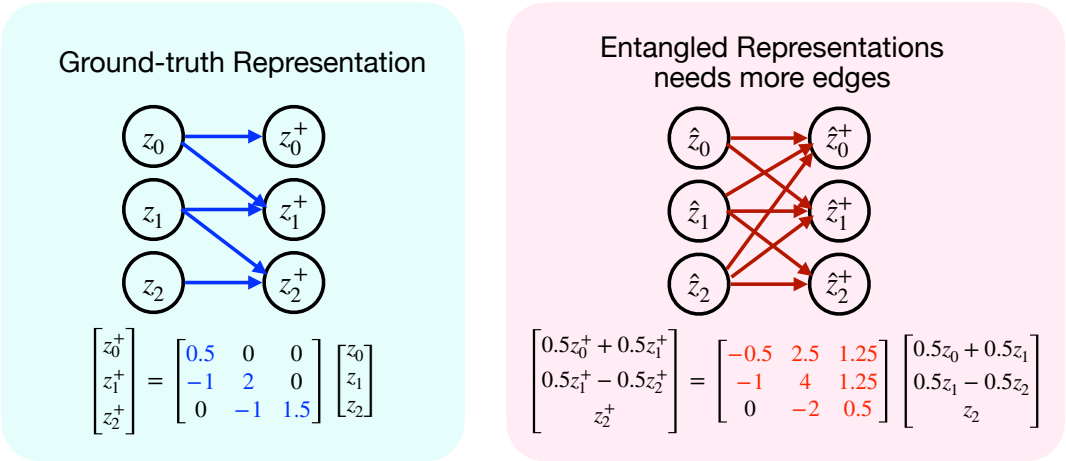


Figure 3.3: A simple linear system to illustrate the sparsity-induced identifiability results in [115]. The main idea is in most cases, entangling the variables will require more edges in the learnt causal graph. Therefore, applying a sparsity regularization will encourage disentangled representations.

the learnt causal masks for VCD. In this framework, the learnable parameters are $\theta := (\mathbf{f}, \boldsymbol{\lambda}, G)$. The main theorem states that, if the distribution induced by the learnt model $\hat{\theta}$ matches the ground-truth model θ, then under certain technical conditions concerning the structure of the problem, $||\hat{G}||_0 \leq ||G||_0$ implies that θ and $\hat{\theta}$ are equivalent up to permutations. In other words, if the learnt model perfectly captures the data distribution with a causal graph that is at least as sparse as the ground-truth causal graph, then the learnt causal variables must be completely disentangled. This suggests that by forcing the causal graph to be as sparse as possible while ensuring that the model captures the data distribution, the learnt representation will be disentangled.

To illustrate this identifiability result, we present a simple numerical example in Fig. 3.3. Consider a simple linear system with three causal variables $z_{1,2,3}$. The dynamics of the system are governed by a linear transition function that follows a causal graph with five edges (Fig. 3.3, left). If, for example, a representation model learns an entangled representation such that z_1 and z_2 are mixed into latent variables $\hat{z}_1 = 0.5z_1 + 0.5z_2$, $\hat{z}_2 = 0.5z_1 - 0.5z_2$, then in order to still consistently predict the next state, the learnt dynamics model must require more edges in the causal graph (Fig. 3.3, right). Analogously, entangled representations will also require

more interventions to explain the changes in distributions across environments. For example, if in the above system, z_1 is intervened on, then in the entangled representation, both $\hat{z}_1$ and $\hat{z}_2$ will be affected, and hence both variables will need to be adapted to explain the change in distribution. As such, the only way to learn a transition function and intervention targets that are as sparse as the ground-truth is to learn a disentangled representation, up to scaling and permutation. Note that there are counterexamples where this is not the case, for example, if the ground-truth transition function is fully-connected, then any representation will have the same number of edges. The reader is referred to [115] for a rigorous treatment of the criteria under which the theorem holds.

In sum, these results show that sparsity plays a central role in the identification of both the causal graph and a causal representation. With this in mind, the architecture of VCD is designed to afford a sparse dynamics model that mimics the structure of a CGM. Whilst, in practice, the condition of perfectly matched distributions might not hold, this general principle of sparsity regularisation forms the basis of our approach. In Sec. 3.3, we empirically show that VCD is able to learn a disentangled latent representation and identify most of the causal relationships.

3.3 Experiments

In this section, we demonstrate that VCD is able to learn from multiple environments with different dynamics by learning a causal world model that explicitly captures the changes between the environments as interventions. We evaluate, qualitatively and quantitatively, the learnt representations and causal structures and show that VCD is able to capture the dynamics of the scene using sparse dependencies between states, as well as identify sparse changes between the environments. Moreover, we illustrate that, in an unseen environment, VCD can leverage past experience to

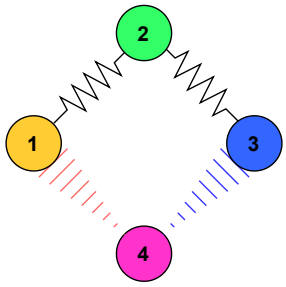


Figure 3.4: The configuration of the multi-body environment. The red and the blue lines represent an attractive and repulsive force respectively.

perform modular adaptation, resulting in significantly improved data efficiency over the baselines.

3.3.1 Experiment Setup

Datasets

We evaluate VCD on a simulated dataset of a 2-D multi-body system, which contains four particles that affect each other via a spring or an electrostatic-like force. The configuration of the environment is shown in Fig. 3.4. The action $a \in \mathbb{R}^2$ is an external force that applies to particle 4. The environment is designed such that there is an unambiguous ground-truth causal graph between the causal variables and well-defined changes in the dynamics. We consider changes such as strengthening or removing one of the springs, increasing or decreasing the mass of a particle, or constraining the position of a particle along the x or the y axis. Each of these changes can be considered as an intervention on one or more causal variables. See Appendix A.3 for further details of the environment and the full list of interventions. We evaluate VCD in two experiments:

1. **Mixed-state**, where the observation is given by applying an affine transformation to the ground-truth positions of the particles.

2. **Image**, where the observation is given by rendering the environment as an RGB image.

The mixed-state experiment serves as a sanity check to verify that VCD is able to learn a disentangled representation and discover the causal structure of the environment. The image experiment is more challenging as the model has to learn a representation from high-dimensional observations.

Baselines

We compare the performance of VCD against RSSM [54], a state-of-the-art latent world model that served as inspiration for VCD. As RSSM does not support learning from multiple environments, we consider two adaptations of RSSM with different levels of knowledge transfer between environments:

1. **RSSM**, where one dynamics model is trained over all environments, i.e., maximum parameter sharing across environments.

2. **MultiRSSM**, where individual dynamics models are trained on each environment, with shared encoders and decoders. This corresponds to the case where no knowledge about scene dynamics is transferred across environments, i.e., each model is a local expert.

Compared to these two extremes of knowledge sharing, VCD captures environment-specific behaviours whilst reusing invariant mechanisms.

3.3.2 Prediction Performance

VCD and the baselines are trained on a dataset composed of 2000 trajectories from each of the undisturbed environment and five intervened environments. The models are evaluated on trajectories from a validation set drawn from the training environments. Fig. 3.5 shows the rollout error for each of the models with state and image observations. Unlike VCD and MultiRSSM, RSSM can only capture the average dynamics of the environments due to parameter sharing. As such, it is not able to capture environment-specific behaviours. This is reflected in the prediction accuracy of the model. In the mixed-state experiment, VCD is able to identify changes between the environments and performs as well as MultiRSSM in terms of prediction error. A similar trend is shown in the image space, where VCD outperforms the baselines. [4] Qualitatively, Fig. 3.6 shows sample rollouts in the image space from RSSM and VCD, where the environment has been intervened such that the yellow particle is constrained to move vertically. Qualitatively, the yellow particle in the VCD rollout stays along a vertical line, whereas RSSM fails to capture this environment-specific constraint.

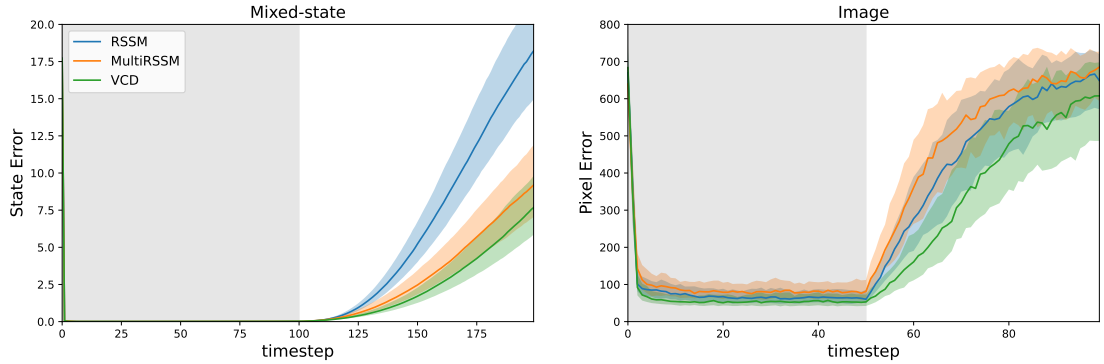


Figure 3.5: The rollout error as measured on validation trajectories. The models receive observations on each timestep for the first half of each trajectory (shaded), after which the model is rolled out based on latent predictions. The reported error in the mixed-state environment (left) and the image environment (right) are the squared error in the ground-truth state space and the squared error in pixel values, respectively. VCD outperforms the baselines in both modes of observation.

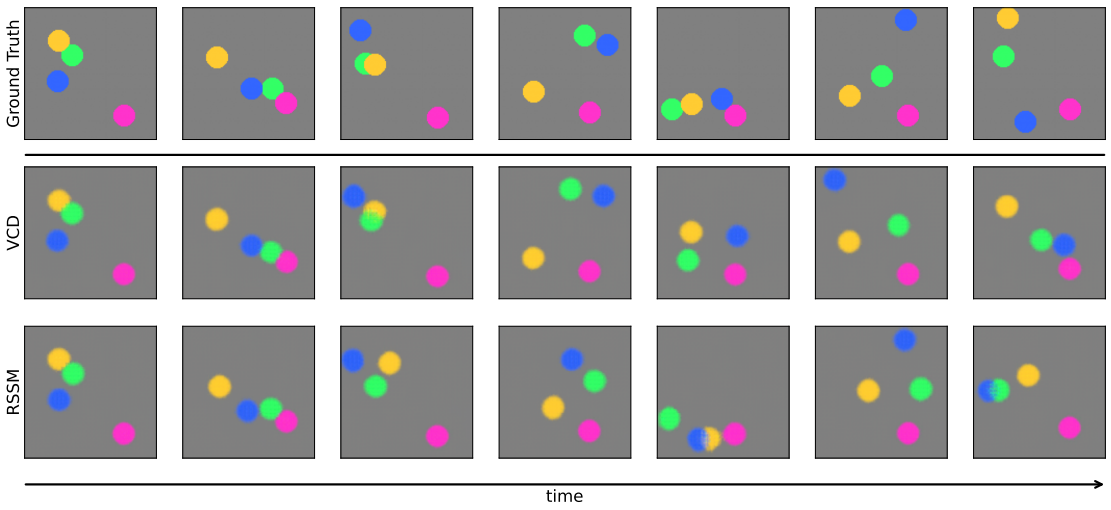


Figure 3.6: Sample rollouts from an intervened environment. VCD successfully captures the constraint on the yellow particle (vertical movement only) over a long time horizon.

Table 3.1: The modified MCC scores for RSSM and VCD. Higher score means better disentanglement.

Expriment	RSSM	VCD
Mixed-state	0.728	**0.975**
Image	0.715	**0.908**

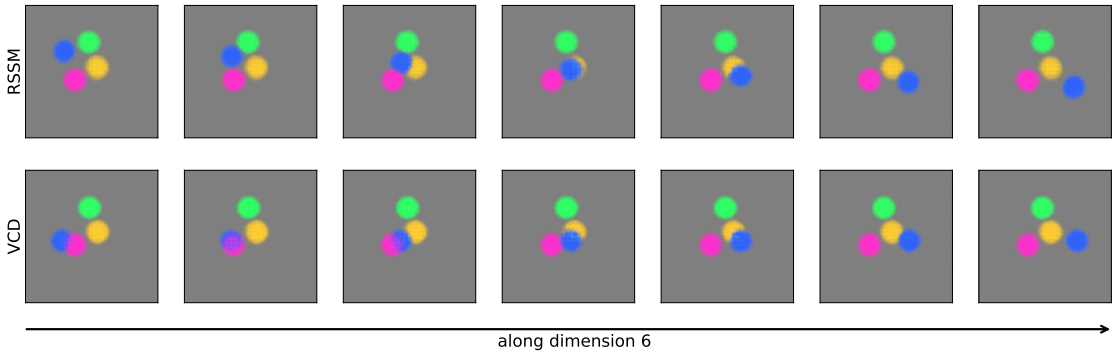


Figure 3.7: Reconstructed images from points sampled along dimension 6 of the latent space. Compared to the latent space learnt in RSSM, VCD is able to learn an axis-aligned representation where the blue particle moves horizontally. We provide more examples in Appendix A.4.

3.3.3 Causal discovery and Disentanglement

One of the key hypotheses of this work is that jointly learning a representation and a dynamics model using causal discovery leads to causally meaningful representations. Here, we examine the quality of the learnt latent space and the causal structure of the dynamics model. Fig. 3.7 shows the image reconstructions of points drawn from a straight line along dimension 6 of the latent space. Since both models are initialised with the same encoder, this provides a qualitative intuition as to how the causal discovery inductive bias shapes the latent space. Whilst this dimension of the latent space in RSSM is able to encode only the blue particle, VCD is able to learn an *axis-aligned* coordinate (x coordinate) of the blue particle with respect to the environment. Note that the motion of bouncing off the boundaries is only separable in the x, y frame, implying that the dynamics in axis-aligned coordinates are sparser, which is consistent with our framework. In table 3.1, we provide the mean absolute correlation coefficient score[5] for the representations for both RSSM and VCD, which measures the quality of representation disentanglement by comparing the learnt latent space with the ground-truth states. We see that in both cases, training with

[4]We note that in the image space, MultiRSSM does not perform well compared to RSSM and VCD. We hypothesise that this is due to the fact that the environment-specific dynamics models are only trained on data from one environment each, compared to RSSM, which has access to data from all environments. VCD does not suffer from this due to modular parameter sharing.

[5]Due to the fact that the ground-truth state dimension (8) is less than the latent state dimension (16), we modify the MCC score by calculating the mean of the top 8 scores out of the 16, effectively ignoring the latent variables that do not capture the state information.

Table 3.2: Sparsity of learnt causal graphs and intervention targets. We compare the learnt graph with the ground truth causal graph by mapping each ground-truth state dimension to a latent dimension by picking the dimension with the highest correlation score.

Observation		# of edges	Correct Edges	Missed Edges	False Pos.
Mixed state	graph	42/288	19	7	8
	intv.	18/80	10	0	5
Image	graph	73/288	22	4	13
	intv.	31/80	10	0	17

the sparsity regularisation in VCD achieves a higher MCC score, which corroborates that VCD learns a more disentangled representation. In App. A.4, we provide an extensive qualitative analysis of the learnt representations in both experiments, further demonstrating the efficacy of VCD in terms of representation learning.

By considering the correlation between each dimension of the learnt representation and the ground-truth states, it is possible to compare the learnt causal graph with the ground-truth (table 3.2). As desired, the learnt causal graphs in both the mixed-state and the image environment are found to be sparse and encode most of the causal edges of the graph. Taken together with the prediction accuracy of the model, this means that VCD is able to retain causal edges that are important for predicting the dynamics of the environment while pruning out spurious edges.

3.3.4 Adaptation Efficiency

We provide empirical evidence that VCD can adapt to a new environment with less data compared to RSSM and MultiRSSM by reusing learnt mechanisms in a modular fashion. We collect datasets of different sizes in a previously unseen intervened environment where particle 1 is constrained horizontally. RSSM and MultiRSSM adapt to the new environment by optimizing the ELBO (Eq. 3.3), with the difference that MultiRSSM instantiates a new dynamics model randomly and RSSM initialises the dynamics model using pre-trained parameters, akin to typical fine-tuning strategies. Following Eq. 3.13, VCD performs adaptation by jointly estimating the intervention targets and the parameters of the intervened

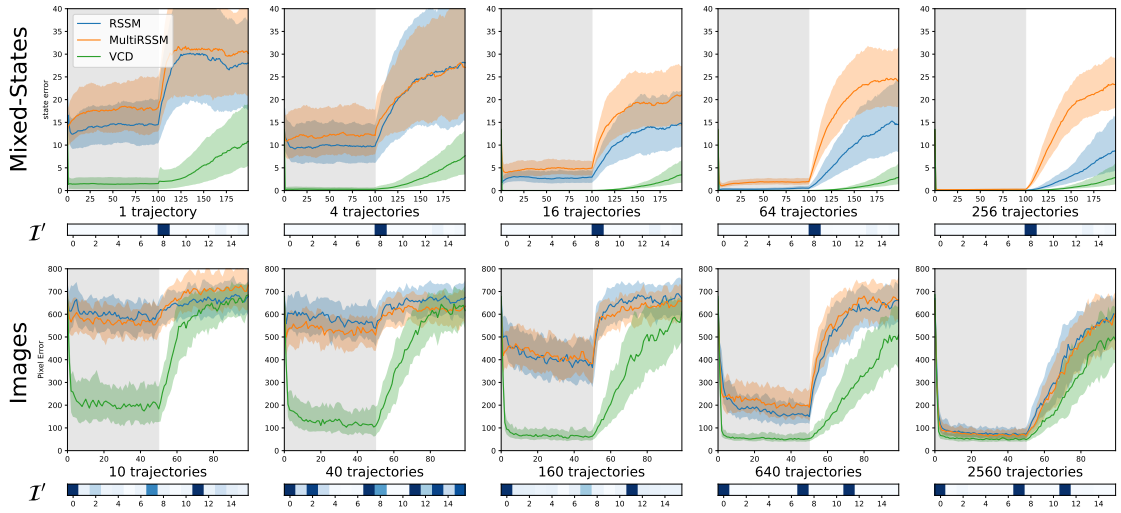


Figure 3.8: Rollout errors in the mixed-state and image experiments, where the models are trained on datasets of varying sizes. The models receive observations for the first half of the trajectory (shaded), and perform latent space prediction for the rest. VCD significantly outperforms the other models with little data. The bar below each plot shows the learnt intervention targets in the new environment. VCD is able to reuse most of the previously learnt mechanisms, as indicated by the sparsity in the learnt intervention targets. In the mixed-state, we verify that the learnt intervention target (dimension 8) corresponds to the ground-truth target (y_1), see Appendix A.4.

mechanisms. The models are evaluated on a validation set of trajectories in the new environment. Fig. 3.8 shows the rollout error of the models trained on datasets of varying size. Across all models, performance improves as the dataset grows. However, in contrast to RSSM and MultiRSSM, which overfit the dataset when the number of trajectories is small, VCD is able to predict significantly more accurately. This is because VCD estimates the intervention targets and reuses trained modules that remain invariant. Remarkably, in the mixed-state experiment, VCD is able to converge to a single latent dimension that encodes the y position of particle 1 with just one training trajectory. In App. A.5, we include more experiments with different interventions, showing consistent results.

3.4 Conclusion and Discussions

In this paper, we propose VCD, a predictive world model with a causal structure that is able to consume high-dimensional observations. This is achieved by jointly training a representation and a causally structured transition model using a modified

causal discovery objective. In doing so, VCD is able to identify causally meaningful representations of the observations and discover sparse relationships in the dynamics of the system. By leveraging the invariance of causal mechanisms, VCD is able to adapt to new environments efficiently by identifying relevant mechanism changes and updating in a modular way, resulting in significantly improved data efficiency.

Limitations.

Our experiments highlight VCD's ability to discover causal variables and mechanisms from image observations. This serves as a promising first step towards learning causal world models from high-dimensional data. However, there are still several limitations that motivate our next steps. In particular, the current experiments are limited to simple environments with a small number of objects and simple interactions. In Chapter 4, we investigate more complex environments that exhibit more diverse interactions and focus on how to disentangle interaction mechanisms from observations. Moreover, another limitation of VCD is that it assumes the causal graph is static across all timesteps and environments. This is, in general, not true in more complex environments, where interactions can be temporally sparse and the number of objects can vary. In Chapter 5, we explore how the sparsity regularisation idea developed in this chapter can be extended to discover local causal influences.

4

Learning Independent Mechanisms via Competition

To reason about environments as rich and complex as our physical world requires the ability to learn efficiently and to flexibly adapt prior knowledge to unseen settings. Whilst humans seem able to generalise knowledge across myriad tasks and situations effortlessly, building artificial agents that can do so with minimal training data remains a significant challenge. At the heart of this challenge are crucial distinctions between what and how humans and machines learn. It has been conjectured that humans represent knowledge internally in a structured and modular way, i.e., by distilling past experience into general principles (or core knowledge) about the world, which can be applied or selectively updated in novel settings [18, 83, 117, 152]. By contrast, current learning-based world models are mostly based on monolithic architectures, and the resulting entangled representations of the world limit the selective reusing of prior knowledge in new environments. Therefore, learning methods that afford modularity are key to world models that can adapt efficiently in diverse settings. In this paper, we address this challenge by developing a model capable of discovering a toolbox of recognisable, generalisable concepts that can be reused across different contexts.

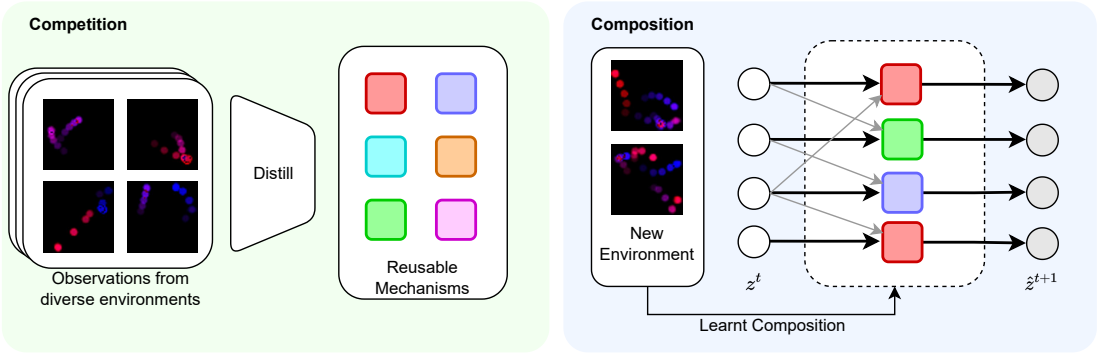


Figure 4.1: Illustration of the *competition* and *composition* training phases. In the first phase, the model learns a set of reusable mechanisms that captures interaction primitives. In the second phase, the model learns to apply these mechanisms in a new environment.

Recent works on object-centric world models [96, 97] serve as an initial step towards a structured and compositional understanding of the world. By decomposing the observed scene into discrete *object slots*, these methods model the interaction between entities in the scene and achieve state-of-the-art results. We argue that, just as the state representation of the scene can be factorised into object slots, the dynamics of the environment, too, can be factorised into discrete and independent *mechanisms*. Our work is motivated by the observation that, while the overall dynamics can change across environments, we can often explain the behaviour of objects by a small set of *interaction primitives* such as "A rests on top of B" and "C collides with D". Our understanding of novel environments is mediated by these primitives: given some observations in a previously unseen setting, one can quickly recognise the relationship between objects in terms of these interaction primitives. As such, we posit that the ability to structurally represent different modes of interactions is crucial to flexible world models. Acquiring such a set of versatile mechanisms from observations without supervision presents challenges both in terms of model architecture and learning algorithm. While current learning methods excel at learning to predict dynamics from i.i.d. environments, learning shared modules that compositionally capture diverse environments cannot be achieved by directly applying gradient updates to the entire model. To this end, we argue that the ability to *selectively* update the model during learning, i.e., to recognise parts of

the model that are relevant to the observed data and perform modular updates, is instrumental to the emergence of discrete independent mechanisms. In this work, we instantiate this capability by utilising a competition of experts training scheme [130], where, during training, we keep a set of independently parameterised mechanisms and only update the module that best explains the observed data. This serves as a natural inductive bias which encourages modules to specialise in specific interaction primitives, and is consequently conducive to the emergence of reusable mechanisms.

We introduce *COmpetitive Mechanisms for Efficient Transfer* (COMET), a modular world model with the capability to explicitly disentangle modes of interactions between objects. In contrast with prior art, COMET is able to learn discrete, abstract concepts from diverse observations and reuse such concepts to predict the evolution of unseen environments. We conjecture that COMET's ability to perceive the world through the lens of structured high-level abstract concepts serves as an important step towards efficient generalisation and transfer across different task settings. Concretely, COMET achieves this via a two-phase training procedure: (i) *competition*, in which COMET learns a set of independently parameterised modules which encode interaction primitives from diverse environments; and (ii) *composition*, wherein COMET is trained to apply these learnt interaction primitives in novel environments (Fig. 4.1).

In our experiments, we evaluate COMET on novel datasets designed to test whether models can systematically learn from multiple environments and transfer to environments with unseen dynamics. We demonstrate both quantitatively and qualitatively that COMET successfully learns reusable mechanisms, i.e. interaction primitives, from image-based observations, and is able to generalise to unseen environments by composing these mechanisms. Moreover, compared to more conventional finetuning baselines, COMET is able to explicitly make use of past knowledge and achieves improved sample efficiency during adaptation.

4.1 COMET: COmpetitive Mechanisms for Efficient Transfer

In this section, we present the training procedure and the architecture of COMET. The main idea of the method is to learn a set of generalisable and composable modules, which encode the different modes of interaction between objects. The intuition behind the approach is that, while dynamics can vary across environments, the ways in which objects or entities interact with each other can often be explained by a small number of independent rules. For example, in a road traffic setting, whilst the behaviour of cars can differ across different locations, the act of stopping at a red light, i.e. the interaction between cars and traffic lights, can be used to explain road behaviours in a wide range of environments. On a conceptual level, this can be considered as a manifestation of the ICM principle [20]. By having a model of *how* objects could interact with each other, the task of adapting to a novel environment reduces to the learning of *when* each rule should be applied.

The goal of COMET is to discover potential interactions between objects through observed data from a diverse set of environments, and to adapt to novel environments by learning when to utilise learnt mechanisms. This is reflected in the training strategy of COMET, which is split into two phases: *competition* and *composition*. In the *competition* phase, COMET learns a set of independent mechanisms from observed sequences using a competition of experts training scheme. In the *composition* phase, COMET adapts to novel environments by learning a classifier that activates the correct mechanism with the correct object pair based on the state. Our main hypothesis is that learning to apply pre-trained, reusable mechanisms in a novel environment facilitates explicit transfer of past knowledge and is hence more efficient than finetuning on new data.

4.1.1 Problem Setup and Model Architecture

COMET learns from a dataset of observed sequences $\{\mathbf{x}^{1:T}\}_N$. In this work, we focus on learning from observations without actions, which is commensurate with the settings in similar works [94, 96], although the framework presented

here can be readily extended to action-conditioned world models. Importantly, these sequences are sampled from environments with varying dynamics where objects can exhibit different behaviours. In order to model interactions between objects, we assume that each observation, $\mathbf{x}^t$, can be factorised into latent object-slots, $\{\mathbf{z}_0^t, \mathbf{z}_1^t, \ldots, \mathbf{z}_K^t\}$, where the subscript denotes the object-id. These object representations can be based on ground-truth state information or obtained from object-centric encoders [87, 88, 90].

COMET consists of two main components, the *mechanisms* and the *composition module*. *Mechanisms* contains M independently parameterised feedforward networks, $f_{mech}^m : \mathbb{R}^{2d} \to \mathbb{R}^d$, with parameters θ_m, where d is the dimension of the object representations. Each mechanism predicts updates to all objects at every timestep, given the state of the object itself and another context object:

$$\Delta \mathbf{z}_i^t(m, j) = f_{mech}^m([\mathbf{z}_i^t \oplus \mathbf{z}_j^t]), \tag{4.1}$$

where $\oplus$ denotes concatenation, i is the index of the object to be predicted and j is the index of the *context* object, i.e. the object with which object i interacts. The mechanisms are trained during the *competition* phase, where each mechanism learns to specialise to cover a particular mode of interaction between objects.

In order to predict transitions for objects using the trained mechanisms, the composition module picks the relevant context object and the active mechanism. The composition module predicts the mechanism-context pair based on the observed state:

$$(\hat{m}_i, \hat{j}_i) = f_{comp}(\mathbf{z}_i^t, \mathbf{z}_{1:K}^t), \tag{4.2}$$

where $(\hat{m}_i, \hat{j}_i)$ is the predicted mechanism-context pair for object i and $\mathbf{z}_{1:K}$ is the representation of all objects. f_{comp} can be considered as a classifier with $M \times K$ classes, where M is the number of mechanisms and K is the number of objects in the scene. In the *composition* phase, COMET adapts to a new environment given a small number of observed sequences $\{\tilde{\mathbf{x}}^{1:T}\}_{\tilde{N}}$ by updating the composition module.

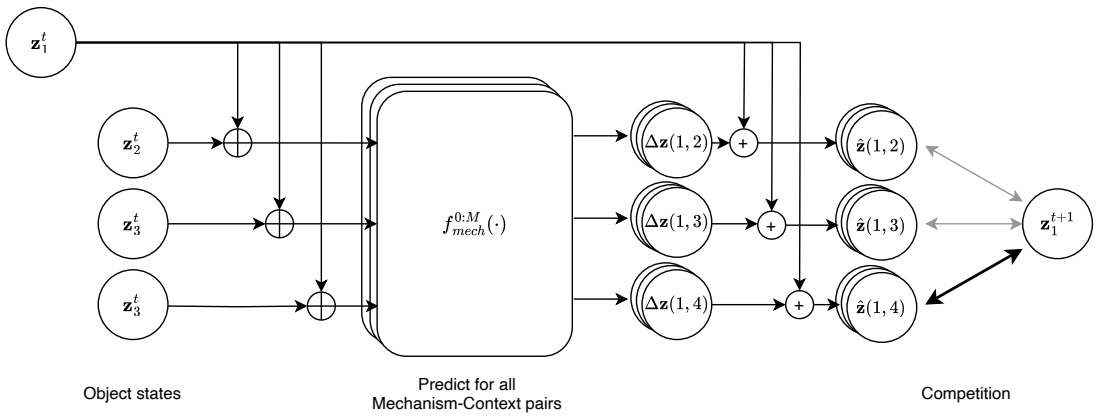


Figure 4.2: In the *competition* phase, predictions are made using all possible mechanism-context pairs for each object. Gradients are only allocated to the mechanism-context pair that produces the most accurate prediction. This encourages specialisation within the mechanisms and enables learning from environments with varying dynamics. The figure describes the prediction step for a single object.

4.1.2 Phase 1: Learning Reusable Mechanisms via Competition

The aim of the *competition* phase is to train the mechanisms in a way such that each mechanism specialises in a particular mode of interaction between objects. Since the training dataset is sampled from multiple environments with varying dynamics, the training procedure needs to recognise shared mechanisms regardless of context. Taking inspiration from [130], we apply a competitive training scheme in the setting of dynamics learning. Concretely, for each object, the model makes predictions using all possible mechanism-context pairs in parallel. Comparing the predictions, we update the mechanism-context pair with the most accurate prediction for a given object. This is illustrated in Fig. 4.2. Given a state transition pair, $(\mathbf{z}_{1:K}^t, \mathbf{z}_{1:K}^{t+1})$, the loss function can be written as:

$$\mathcal{L}(\theta_{1:M}) = \sum_{i=0}^{K} \min_{m,j} \left[d\left(\mathbf{z}_i^t + \Delta\mathbf{z}_i^t(m,j), \mathbf{z}_i^{t+1} \right) \right], \tag{4.3}$$

where d is a function that measures the prediction error. There are several sensible choices for the distance function d. In this work, we find that using Euclidean distance is sufficient. Other potential choices include a contrastive loss [97] or

ELBO loss [54] for training representation and dynamics model jointly, and sparsity-regularised loss, such as discussed in Chapter 3 for causal structure discovery. Importantly, when performing back-propagation on this loss function, only the parameters of the competition winner are updated. This independent update of parameters encourages *specialisation* among the mechanisms and facilitates environment-agnostic learning. Intuitively, insofar as a particular mode of interaction is concerned, a specialised mechanism will outperform other mechanisms and hence win more data as training progresses. This triggers a positive feedback loop, which allows the specialised mechanism to further improve its accuracy. Moreover, the competition procedure ensures the selection of the correct context objects during training since accurate predictions are only possible with the relevant context information.

In practice, one of the main failure modes in this training phase is a case where a single mechanism wins the competition most of the time, which hinders the proper disentanglement of dynamics. Empirically, we find that the robustness of the training dynamics can be significantly improved by adding a warm-starting phase and by increasing the time horizon for mechanism selection.

Warm-starting phase. One failure mode of COMET is that, due to the competitive nature of the training process, one or two mechanisms can dominate the competition and win all of the data. This can happen if the mechanisms are initialised randomly. In this case, one mechanism can, by chance, be the closest approximation to the environment dynamics and keep winning the competition. In order to alleviate this, we add a warm-start phase before the actual competitive training, where the gradient is passed through all mechanism-context pairs regardless of accuracy. The intuition is that this allows all mechanisms to be roughly correct and, as such, have a fair chance of specialising in a particular mode of interactions.

Increased time horizon for mechanism selection. Another failure mode that we encountered is that the model can learn non-interpretable mechanisms, which leads to flickering mechanism selection during rollout. To this end, we

found that increasing the effective time-horizon of the chosen mechanism-context pair is conducive to the emergence of well-disentangled mechanisms. This can be implemented as a modification to the loss function, such that:

$$\mathcal{L}(\theta_{0:M}) = \sum_{i=0}^{K} \min_{m,j} \left[\sum_{\tau=0}^{T} d\left(\mathbf{z}_i^{t+\tau} + \Delta \mathbf{z}_i^{t+\tau}(m,j), \mathbf{z}_i^{t+\tau+1} \right) \right], \tag{4.4}$$

where T is the time horizon, which is set as a hyperparameter. Note that the sum over the time horizon is inside the minimisation, meaning that a mechanism-context pair is picked only if it has the best performance over multiple consecutive time steps. This serves as an inductive bias that encourages the sparsity of mechanism changes across time steps, which empirically results in better disentangled mechanisms.

In App. B.3, we present ablation studies that demonstrate the effectiveness of these two strategies.

4.1.3 Phase 2: Learning to Compose Mechanisms

With a trained set of mechanisms, COMET adapts to new environments by training the composition module on new data.[1] The composition module acts as a classifier that picks the optimal mechanism-context pair for each object in the scene. Given the state of object i, for each mechanism-context pair, we compute the confidence score:

$$c_i^t(m,j) = f_{conf}^m([\mathbf{z}_i^t \oplus \mathbf{z}_j^t]), \tag{4.5}$$

where f_{conf}^m is an independently parameterised MLP for mechanism m, i.e. each mechanism has a corresponding f_{conf}^m. We predict a categorical distribution over all mechanism-context pairs by taking the softmax over the confidence scores for object i at time step t.

Given a small number of observation sequences in a new environment, we obtain the best performing mechanism-context pair $(m^*, j^*)_i^t$ for each object at each time step by investigating which pair minimises the loss function of the competition scheme in Eq. 4.3. These then serve as the target labels for the classifier. The composition module is then trained using the negative log-likelihood loss. At test

[1] Here, we focus on directly reusing mechanisms in new environments. We leave the exploration of systems that can instantiate new mechanisms to future work, see discussion in Sec. 4.3.

time, together with the trained mechanisms, COMET predicts the next state of each object by first picking the mechanism-context pair with the highest score and then feeding the chosen context object through the chosen mechanism. In the following section, we show empirically that this two-step training procedure is able to learn meaningful mechanisms from image observations and can compose mechanisms in novel environments efficiently.

4.2 Experiments

In this section, we demonstrate that COMET is able to disentangle different modes of interaction between objects and can efficiently reuse learnt mechanisms during adaptation. We evaluate, both quantitatively and qualitatively, the performance of COMET on image-based environments. Concretely, the experiments focus on whether COMET can learn reusable and recognisable mechanisms (Sec. 4.2.2) and whether re-composing learnt mechanisms is more sample-efficient compared to finetuning approaches (Sec. 4.2.3).

4.2.1 Experiment Setup

Datasets

We evaluate COMET on three problem domains: *Particle Interactions*, *Traffic*, and *Team Sports*. For each of these domains, we define a set of environments where objects can exhibit different behaviours. These environments are designed to test whether COMET can extract meaningful mechanisms and adapt to unseen environments via composition. The *Particle Interactions* dataset consists of coloured particles that can interact with each other in different ways, such as attraction and repulsion. Environments are defined by a combination of rules, such as "red particles repel each other". The *Traffic* dataset contains observation sequences of traffic scenarios generated with the CARLA simulator [153]. Here, the environments are defined by traffic rules that apply to different vehicles, such as "blue cars do not need to stop at red lights". The *Team Sport* domain consists of a simulated generic hockey game where players can perform different actions such as moving

towards the puck or dribbling the puck towards the opponent's goal, generated by the STS2 simulator [154]. In this setting, differences in the environment amount to the different behaviours of the players, e.g. players in some environments might tend to take more aggressive actions. This is a particularly challenging dataset for COMET as it violates the assumption that all interactions are binary. Observations are masked RGB images for each object, except for the *Team Sports* domain, where state-based observations are more naturally suited. Details of the datasets are provided in App. B.2.

Baselines

We evaluate COMET against two competitive baselines, C-SWM [97] and Neural Production Systems (NPS) [122]. C-SWM learns a world model from observation via contrastive learning with a GNN-based dynamics model. Similar to COMET, C-SWM operates on an object-factorised representation and achieves state-of-the-art prediction accuracy. COMET further disentangles the interactions between objects as independent mechanisms rather than learning a monolithic model that captures all interactions. In Sec. 4.2.3, we show that this disentanglement is conducive to sample-efficient adaptation and facilitates interpretable transfer.

We also compare against NPS, which learns to capture object interactions as independent mechanisms. Architecturally, NPS is similar to COMET, except that the mechanism-context pair is selected using dot-product attention [8], which is trained jointly with the mechanisms. In contrast, COMET deploys a competitive training scheme that allows the model to recognise shared mechanisms across environments. Moreover, in Sec. 4.2.2, we show that competition serves as a strong inductive bias that enables the emergence of generalisable mechanisms. We provide more model details and further discussions around the baselines in App. B.1.

4.2.2 Disentanglement of Mechanisms

We investigate the emergence of recognisable mechanisms from competition. Here, both COMET and NPS are trained on a mixture of environments in each domain.

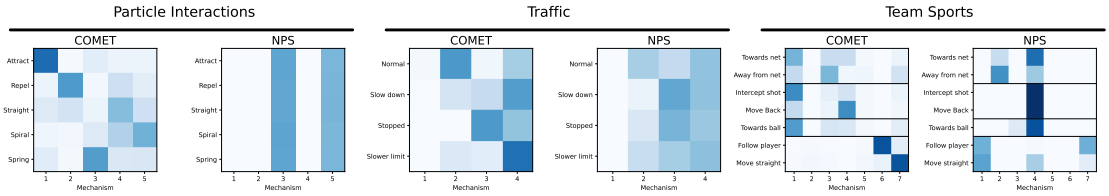


Figure 4.3: Disentanglement plots showing the correlation between mechanisms chosen by the models and ground-truth interaction modes. In the ideal case, the matrices should look like permutation matrices. Here, COMET is able to learn disentangled mechanisms that correspond to ground-truth behaviours in all three domains, as indicated by the fact that each interaction mode has one main corresponding learnt mechanism. In contrast, NPS does not exhibit the same structure.

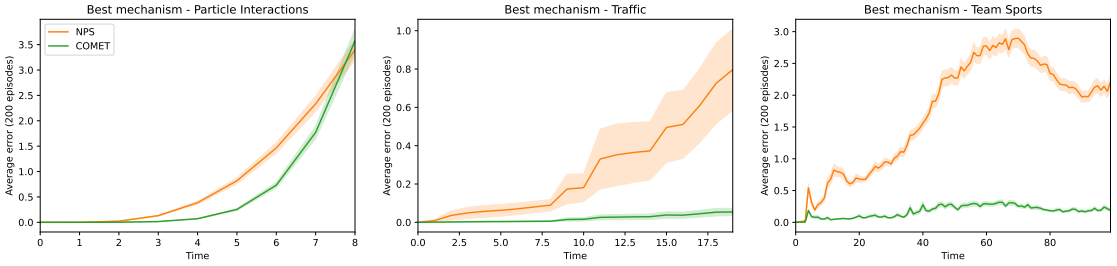


Figure 4.4: Rollout errors (lower is better) in unseen environments with optimal mechanism selection. Shaded areas indicate the standard error of the mean. The lower errors indicate that COMET mechanisms can be readily reused across environments *without* finetuning.

We obtain the ground-truth labels for the object interactions and use these to directly investigate whether the learnt mechanisms correspond to actual interactions without any supervision. These labels are not accessible to the models during training. Fig. 4.3 shows the correlation between the active ground-truth interactions and the winning mechanisms in the competition process in the different domains. COMET achieves successful disentanglement and learns mechanisms that correspond to the ground-truth interactions. In the *particle interaction* dataset, COMET recovers the ground-truth mode of interactions between particles. Similarly, in the *Traffic* and *Team Sports* datasets, COMET learns mechanisms that model interaction primitives such as stopping before a red light and running towards the opponent's goal.[2] Fig. 4.5 qualitatively illustrates that the 'winning' mechanisms switch as the

[2]In the team sports dataset, we observe that mechanism 1 is 'overloaded' in the sense that it represents different behaviours for different objects, i.e. for the attacking team, it models 'move towards net' and for the defending team, it models 'intercept shot'. Crucially, for each object

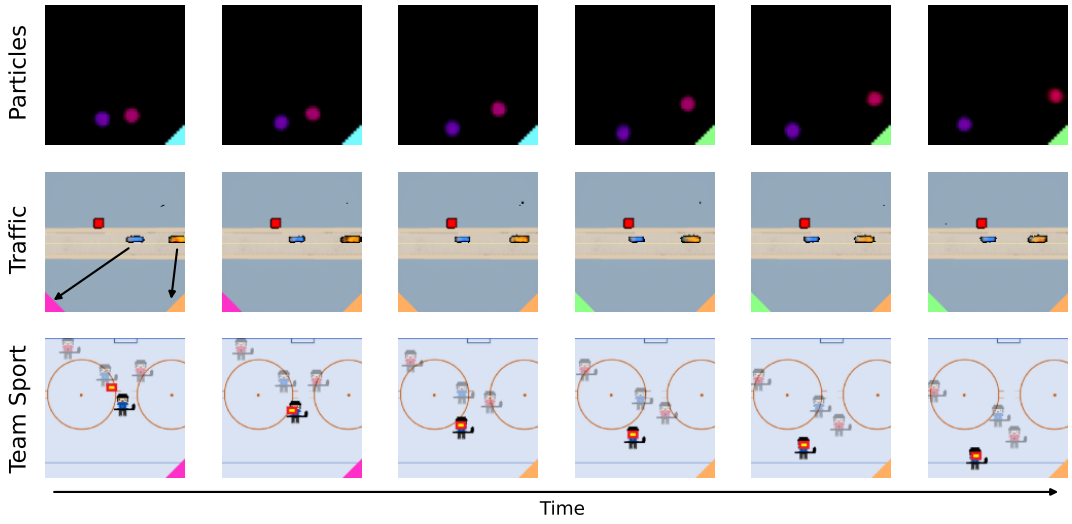


Figure 4.5: Qualitative rollouts. The colour of the tabs on the bottom of each frame indicates the 'winning' mechanism at each time step. Across all environments, the competition winner changes as the underlying interaction mode changes. **Top**: The particles repel each other when they are close (blue) and move independently when they are apart (green). **Middle**: In this traffic environment, the orange car obeys a slower speed limit and always picks the slow mechanism (orange). The blue car approaches the red light with normal driving (pink) → slow down (orange) → stop (green). Note that the orange mechanism is used for slow driving for both cars. **Bottom**: The player first waits to receive the ball (pink) and then moves towards the opponent's goal when in possession of the ball (orange).

underlying interaction type changes. In contrast, the mechanisms learnt by NPS show no correspondence with the ground-truth interactions. This is likely because NPS cannot learn from a mixture of environments with varying dynamics, as it employs a simple dot-product attention for picking mechanisms during training. To this end, COMET's ability to learn from diverse environments is uniquely afforded by the competition scheme, which assigns relevant data to update each mechanism.

Central to our hypothesis is the notion that good mechanisms are *reusable* across environments. To this end, we perform rollout in *unseen* environments by picking the optimal mechanism-object pair, (m^*, j^*), effectively bypassing the confidence module. This offers insights into whether the learnt mechanisms can generalise to new settings. We perform this to the mechanisms learnt with COMET and NPS. Fig. 4.4 shows the rollout errors with optimal mechanism selection. COMET

type (indicated by the horizontal divider lines in Fig. 4.3), different behaviours are captured with different mechanisms. This is not the case for NPS.

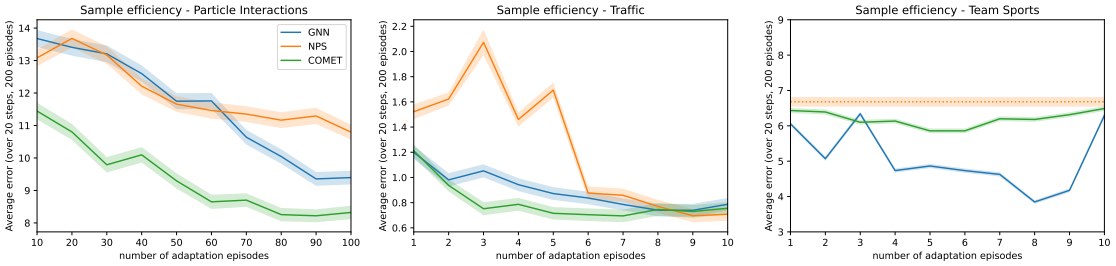


Figure 4.6: The average rollout error in an unseen environment with different amounts of observed data in the new environment (lower is better). In all environments, all models eventually converge to similar errors given enough data. We show this explicitly in App B.4 and offer further discussion on the behaviours of the models at the large-data limit. In terms of sample efficiency, in the *Particle Interactions* and *Traffic* domains, COMET is able to achieve lower errors with few adaptation episodes. This means that COMET can learn to use the correct mechanisms with a small amount of data, thus corroborating our hypothesis that composing learnt mechanisms enables sample-efficient transfer. In the *Team Sports* domain, NPS is not able to generate stable rollouts with the amounts of adaptation episodes shown in the plots. The dotted line indicates the performance of NPS when trained with a large amount of data. Shaded areas represent the standard errors of the mean.

mechanisms perform significantly better than NPS mechanisms, meaning that, when picked correctly, COMET mechanisms can be directly used in new environments without any finetuning. This corroborates our claim that competition training is conducive to the emergence of generalisable and reusable mechanisms.

4.2.3 Adaptation Efficiency

One of our main hypotheses is that learning to compose learnt mechanisms leads to data-efficient adaptation. For each domain, we train all of the models on a mixture of environments and adapt the models to unseen environments. COMET adapts by training the composition module, whereas the baselines adapt by finetuning the entire model on new data. Fig. 4.6 shows the prediction performance of the models when trained on different amounts of observations in the new environment. In general, all models improve as the amount of available data increases. In the *Particle Interactions* and *Traffic* domains, while all models eventually perform comparably given sufficient data, here we see that COMET outperforms the baselines in the low-data regime, illustrating that explicitly reusing learnt mechanisms results in improved sample efficiency compared to gradient-based finetuning. In the *Team*

Sport domain, C-SWM performs better than COMET and NPS. We hypothesise that this is because both COMET and NPS only model binary interactions between objects, whereas C-SWM uses a GNN-based dynamics model that can take into account the entire scene. We further discuss this limitation in Sec. 4.3. Nonetheless, under the same binary interactions restriction, COMET significantly outperforms NPS, which overfits to the adaptation data quickly and is not able to generate stable rollouts.

4.3 Conclusion and Discussions

In this paper, we introduce COMET, a structured world model that encodes discrete abstract mechanisms explicitly from observations. This is achieved through training the model on a set of environments with diverse dynamics using a winner-takes-all competition scheme. This enables the model to perform *selective* updates during the training phase, a central capability which facilitates the emergence of recognisable and reusable mechanisms. We show experimentally that the proposed method is indeed able to disentangle shared mechanisms across different environments from image observations, and thus enables sample-efficient and interpretable adaptation to novel situations. While the presented results are competitive on the tasks considered, we see our main contribution as a conceptual one, serving as a step towards world models exhibiting a structured understanding of the world. Looking forward, we believe that the method developed here opens up several promising avenues of research, such as designing agents that learn a growing repertoire of reusable interaction behaviours and agents that explore the world through the lens of mechanism discovery.

Limitations

There are several limitations of the current work that motivate our next steps:

- **Object Encoders.** In our experiments, we focus our analysis on the dynamics models and opt for simple CNN encoders with ground-truth object segmentation masks as the representation backbone. The application of

the model on more visually complex environments without segmentation masks would require the use of object-centric encoders or other segmentation techniques [88, 90].

- **Adapting Mechanisms.** As it stands, COMET lacks the ability to update the mechanisms during *composition*, which limits the model's ability to adapt to environments with completely new interactions. Designing mechanism-based models that can instantiate new mechanisms as they encounter new data offers an exciting avenue of research, with important implications for life-long and continual learning, where an agent can improve its understanding of the world through a growing set of reusable mechanisms. This is beyond the scope of the present investigation. Nonetheless, we believe the method presented here offers a tantalising glimpse into this potential.

- **Binary Interactions.** The model architecture for the mechanisms implies that the model can only capture *binary* interactions. While this is a good approximation for many interacting systems, this hinders the model's ability to capture more complex dynamics. As seen in the Team Sports experiments, while COMET can disentangle meaningful mechanisms, picking the correct mechanisms may require information from the entire scene, leading to suboptimal prediction accuracy. To this end, in Chapter 5, we explore more flexible architectures that can capture higher-order interactions, with the explicit desiderata of developing methods that can discover causal structures while matching state-of-the-art prediction performance.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

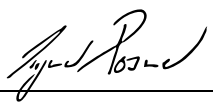
Title of Paper	Compete and Compose: Learning Independent Mechanisms for Modular World Models
Publication Status	Published (Workshop paper)
Publication Details	A workshop version is published as: Anson Lei, Frederik Nolte, Bernhard Schölkopf, Ingmar Posner. Modular World Models with Competitive Independent Mechanisms. In Causal and Object-Centric Representations for Robotics Workshop at CVPR (2024) A Long version available on: https://arxiv.org/abs/2404.15109

Student Confirmation

Student Name:	Anson Lei		
Contribution to the Paper	I conceptualised the project and implemented the main code base. My coauthor Frederik implemented one (out of three) experiment environments and helped with the analysis of the results.		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:			
Supervisor comments			
Signature		Date	

5

Learning Local Dependency Structures with Sparse Transformers

In Chapters 3 and 4, we have focused on learning causal representations and disentangled mechanisms for dynamics, demonstrating the importance of structured models. In particular, these works elucidate how key features of structured models, namely the ability to identify causal relationships and modularly update mechanisms, can lead to improved adaptation efficiency and interpretability. However, the shortcomings of these approaches compared to the current state-of-the-art are equally instructive: whereas standard world models are able to achieve high prediction accuracy in complex, real-life settings, the scope of our investigations thus far has been primarily limited to comparatively simple environments. COMET, for example, is restricted to binary interactions, which limits its applicability to more complex scenarios where multiple entities can interact simultaneously. In this chapter, we aim to bridge this gap between causally structured models and high-performance world models, and set the key desiderata that drive our approach: the model developed should *match* the prediction performance of state-of-the-art world models on complex tasks, while retaining the ability to discover causal structures that facilitate interpretability and efficient adaptation.

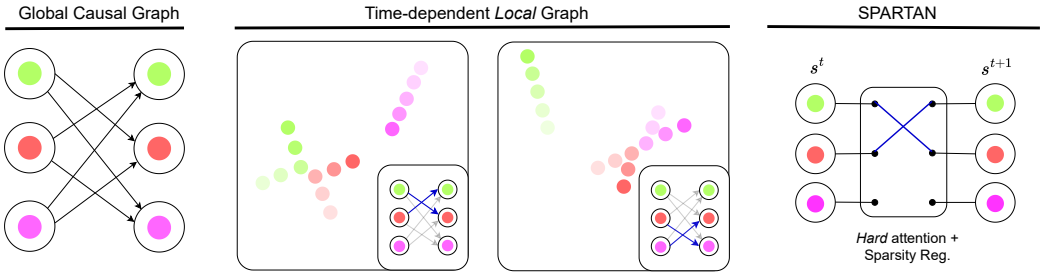


Figure 5.1: In the context of modelling physical interactions, a global causal graph is often uninformative and close to fully connected. A time-dependent *local* causal graph better captures the sparse nature of interactions between entities. We present SPARTAN, a Transformer-based world model that discovers local causal structure using hard attention with sparsity regularisation.

The crux of the challenge lies in identifying *why* a conventional causal model falls short in complex world modelling tasks. We argue that the key limitation stems from a mismatch between the causal structure assumed by causal models and the true underlying interaction structure in common physical environments. In particular, methods inspired by classical causal discovery, such as the VCD architecture introduced in Chapter 3, aim to learn a fixed graph to fit an entire dataset, which leads to two key issues. First, learning a fixed causal graph that explains the data requires the graph to capture all *possible* interactions between entities in the scene. In most realistic settings, where all objects can potentially interact with each other, this often leads to uninformative, fully-connected graphs. Second, in many application domains, such as traffic motion prediction [155], the number of entities can change from scene to scene, which is not compatible with the standard theoretical framework of structural causal models. To remedy this, more recent approaches [24, 141] formulate the problem as learning *state-dependent* local causal graphs (Fig. 5.1). Despite their success in simple state-based settings, scaling prior local causal discovery methods to image observations with complex dynamics remains a significant bottleneck. This hinders the application of these methods to realistic scenarios where, for example, objects are represented via learned embeddings and the number of objects can vary from scene to scene.

Taking inspiration from the conceptual framework of local causal models, the goal of this chapter is to develop a transformer-based world model that faithfully reflects the interaction structure between objects while maintaining the flexibility and accuracy afforded by the transformer architecture. The key insight is that attention mechanisms serve as a natural starting point for constructing state-dependent graphs. By applying sparsity regularisation, a key inductive bias in causal discovery explored in Chapter 3, on a modified transformer architecture with *hard* attention, we develop SPARTAN, a Transformer-based world model with a learnable local causal graph. Specifically, SPARTAN minimises the *expected* number of *local* causal edges between tokens, allowing the discovery of accurate and interpretable causal edges between objects. We evaluate our model on observation-based environments with physical interactions and a real-life traffic motion prediction dataset [155]. To the best of our knowledge, SPARTAN is the first method that learns local causal structures beyond simple state-based settings. Compared to prior transformer world models, our results show that SPARTAN is able to accurately identify causal edges, resulting in significantly improved robustness against distractors and interpretability. We further extend our model to represent sparse causal interventions by introducing *intervention tokens* that only affect a subset of entities in the scene. In so doing, SPARTAN is able to efficiently adapt to changes in the environment in a few-shot manner. We demonstrate that due to the causal structure discovered, our model achieves better adaptation efficiency without compromising on overall prediction accuracy.

5.1 Background

We begin this section by introducing our problem setting and providing a brief recap of CGMs and how sparsity regularisation can be applied to discover causal structures. We then provide a discussion of the role of *local* causal models in relation to the standard notion of CGMs. The goal of our work is to learn a causal model that predicts future states from observation. In order to model interventions, we train our model on a set of intervened environments with *unknown* intervention targets. Specifically, the training data of our model consists of observation sequences

sampled from different environments $\{\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^T, I\}$, where x^t is the observation at timestep t and I is the environment index. Here, an intervention amounts to altering the dynamics, such as changing the friction of an object.[1] We assume access to a representation model that maps observations to factored latent states. $\mathcal{S} = \mathcal{S}_i \times \cdots \times \mathcal{S}_N$, such that each $\mathcal{S}_i$ is an embedding that represents the state of an entity, such as an object in the scene. Examples of such representation models include object-centric models [88, 90] or disentangled causal representations such as VCD, where each $\mathcal{S}_i$ is an object slot or a causal variable, respectively.

Our aim is to learn a transition function $p(s^{t+1}|s^t)$ that captures the local causal structure. Recall the structure of a CGM (see Chapter 2):

$$p'(x_1, \ldots, x_d) = \prod_{i \notin I} p(x_i|PA_i) \prod_{i \in I} p'(x_i|PA_i), \tag{5.1}$$

where PA_i are the parents of the variable x_i in the causal graph. The task of causal discovery amounts to learning the graph structure from data. As discussed in Sec. 3.1, given a set of interventions with unknown targets $\mathcal{I}^{1:K}$ and their corresponding distributions $X \sim p_X^{(k)}$, causal discovery can be formulated as the following optimisation problem [22],

$$\max_{\theta} \sum_{k=1}^{K} \mathbb{E}_{X \sim p_X^{(k)}} \log f^{(k)}(X; \theta) - \lambda_G |\mathcal{G}| - \lambda_I |\mathcal{I}|, \tag{5.2}$$

where θ are the learnable parameters for the conditional distributions, $f^{(k)}(\cdot)$ is the likelihood of the data under the model for the k-th intervention, $|\mathcal{G}|$ is the number of edges in the learnt causal graph, $|\mathcal{I}|$ is the number of intervention targets, and λ_G and λ_I are hyperparameters for sparsity regularisation. Intuitively, solving this optimisation amounts to pruning out spurious edges and interventions, and finding a model that can explain the data with the *sparsest* causal graph and intervention targets.

[1] Note that the case where there is only one single environment with no interventions is included as a special case.

Local Causal Models

While learning a global causal graph is sufficient in some settings, in the context of dynamics models, such a graph is often close to fully connected since any entities that interact with each other at *any point* are connected on the graph regardless of how unlikely the interaction is. This leads to a mismatch between the meaning of the causal edge $A \to B$ and the colloquial meaning of "A causes B": when referring to physical systems, what we mean when we say "A causes B" is that A has a causal influence on B *in the current state*. Consider a traffic modelling scenario where we aim to predict how vehicles behave. A global causal graph would need to connect every vehicle in the scene, as they can all influence each other when close together. However, events such as "vehicle A causes vehicle B to stop" can be more appropriately captured by *local causal graphs* with the edge $A \to B$ but not edges to other vehicles that are irrelevant at the time. As such, we argue that local causal graphs are more suitable for models to fully exploit the sparse structure of the problem at hand.

We define *local causal models* [24, 25, 141]. Suppose we observe the values of a subset of variables, $\mathcal{X}^{obs} \subset \mathcal{X}$, $\mathcal{X}^{obs} = x^{obs}$. For some neighbourhood $\mathcal{L}$ containing x^{obs}, the induced local causal graph $\mathcal{G}^{\mathcal{L}}$ is constructed by removing all edges (i, j) from the global causal graph $\mathcal{G}$ where $x_j \perp\!\!\!\perp x_i | PA_i(x_j) \backslash x_i$ within the neighbourhood, i.e. the local graph is a subgraph of the global graph obtained by removing any "inactive" edges based on the observed state. Note that the structure of the local graph is now dependent on the observed variables. Concretely, in the context of transition functions, the set of observed variables is the states of the current timestep s^t. As a simple example, consider two billiard balls. If it is observed that they are far apart, then there is no local edge since they cannot influence each other. However, when they are about to collide, their next states will depend on each other, and there is a *local* edge between them.

Analogously to standard causal discovery, our goal is to infer the sparsest graph at each timestep, such that the next observation can be explained. In the local case, the optimisation objective for *local* causal discovery is to minimise

the *expected* number of edges and targets with respect to the data distribution, where the causal edges and intervention targets are state-dependent and are built dynamically at each timestep. [24] provides a theoretical analysis of the conditions under which local causal structures can be identified using sparse regularisation. In the following section, we present SPARTAN, a transformer-based instantiation of sparsity-regularised local causal discovery.

5.2 Sparse Transformer World Models

Our goal is to develop a world model that learns *local* causal models as transition functions. We build on the Transformer architecture [8] as it achieves state-of-the-art prediction performance in object-factored world models [98]. Moreover, its attention mechanism provides a natural way to control the flow of information between object tokens. While prior work [25] argues that soft attention is a strong enough inductive bias for learning local causal graphs and proposes a thresholding heuristic to the attention patterns, we posit that this is insufficient beyond simple state-based scenarios and that appropriate sparsity regularisation and hard attention play a crucial role in scaling this up to more complex, observation-based environments. To demonstrate with a simple example, consider a system with only one causal edge, $s_i^t \rightarrow s_j^{t+1}$ and many other irrelevant nodes. A soft-attention model can predict s_j^{t+1} as long as the attention value for s_i^t is non-zero, i.e., information can flow from the i-th token to the j-th token. However, the model is free to have non-zero attention values over other tokens. In this case, there is no guarantee that the attention value on s_i^t is the highest amongst all tokens, and therefore applying a threshold on the attention values may 'catch' spurious edges or miss necessary edges. Instead, we need a model that *masks* the information flow between the tokens with a penalty on connections between tokens. Any model that contains spurious connections would have a high loss due to the sparsity penalty, while models that do not contain the $s_i^t \rightarrow s_j^{t+1}$ would suffer from low prediction accuracy. Under this scheme, the model that achieves the best training loss would be the one that only contains the correct causal edge. In the following section, we present *SPARrse TrANsformer*

World models (SPARTAN), an instantiation of a masked transition model using a Transformer-based architecture with *hard* attention.

5.2.1 Learnable Sparse Connections

We start with the architecture of a single layer before extending to the multi-layer case. Given object representations $s^t_{1:N}$ as input, we obtain the keys, queries, and values, $\{k_i, q_i, v_i\}$ via linear projection as in the standard Transformer. Using these, we sample an adjacency matrix,

$$A_{ij} \sim \text{Bern}(\sigma(q_i^T k_j)), \tag{5.3}$$

where $\text{Bern}(\cdot)$ is the Bernoulli distribution and σ is the sigmoid function. The hidden features are then computed using the standard scaled dot-product attention with the adjacency matrix as a mask.

$$h_i = \sum_j \frac{A_{ij} \exp(q_i^T k_j / \sqrt{d_k}) v_j}{\sum_i \exp(q_i^T k_j / \sqrt{d_k})}, \quad \hat{s}_i^{t+1} = MLP(h_i + s_i) \tag{5.4}$$

where $\hat{s}_i^{t+1}$ is the prediction for the i-th object in the next timestep. Here, the adjacency matrix acts as a mask that disallows information flow from tokens that are not in the parent set of the querying token. Note that the adjacency matrix is dependent on the current state via the key-query pairs. The learnt adjacency matrix can therefore be interpreted as the local causal graph, i.e., $A_{ij} = 1$ means $s_j^t \in Pa^{\mathcal{L}}(s_i^{t+1})$. The sampling step is made differentiable via the Gumbel softmax trick [151]. Typically, multiple Transformer layers are stacked sequentially for expressivity. This presents a problem for the learnt adjacency matrix as information can flow between tokens in a multi-hop manner across different layers, e.g., token i can influence token j via $i \to k \to j$ without the edge (i, j) being in any of the adjacency matrices. SPARTAN mitigates this problem by tracking the number of *paths* from token i to j. Concretely, for L layers, we compute the path matrix,

$$\bar{A} = (A^{L+1} + \mathbb{I}) \dots (A^2 + \mathbb{I})(A^1 + \mathbb{I}), \tag{5.5}$$

where A^l is the adjacency matrix at layer l and $\mathbb{I}$ is the identity matrix. The identity matrix is added due to residual connections. The path matrix $\bar{A}$ has the

property that $\bar{A}_{ij}$ is the number of paths that lead from j to i. In this case, s_j^t is a local causal parent of s_i^{t+1} iff $\bar{A}_{ij} >= 1$.

5.2.2 Interventions

We can extend our model to represent interventions in cases where the model is trained on a dataset with a set of environments with intervened dynamics. During training, the model has access to the environment index I for each transition but does not know which object is affected by the intervention. Conditioned on the current state $s_{1:N}^t$ and the environment index I, the model needs to identify a subset of $s_{1:N}^{t+1}$ as intervention targets and change the predictions (cf. Eq. 5.1). To do this, we keep a learnable codebook of *intervention tokens*, $\mathcal{T}_{1:K} \in \mathbb{R}^{d \times k}$, where d is the dimension of the state tokens and K is the number of interventions. We append the corresponding intervention token $\mathcal{T}_I$ to the object tokens and proceed as described in the previous subsection. Here, $\bar{A}_{i(N+1)} = 0$ indicates that there is no path from the intervention token to s_i^{t+1} and therefore s_i^{t+1} is not an intervention target, i.e. the prediction for s_i^{t+1} is not affected to the changes in this environment.

At test time, the model adapts to unknown interventions by observing sequences without the environment index. We perform adaptation by finding an intervention token $T_{adapt} \in \mathbb{R}^d$ that best fits the observed data via gradient descent in the token space. Note that the adaptation intervention token need not be in the discrete set of intervention tokens from training and can therefore model unseen environments. In Sec. 5.3, we show that this approach can generalise to the composition of previously seen interventions and leads to efficient adaptation.

5.2.3 Training

Our aim is to fit the data distribution with the sparsest possible model in terms of causal edges and interventions. The training objective is to minimise the expected loss with sparsity regularisation,

$$\min_{\theta} \mathbb{E}_{s^t, s^{t+1}, I} \Big[\mathcal{L}(\hat{s}^{t+1}) + \lambda_s |\bar{A}| \Big], \tag{5.6}$$

where θ are the model parameters, including the parameters for the Transformer and the intervention tokens, $\mathcal{L}$ is a loss function such as MSE, and λ_s denotes the regularisation hyperparameter. Note that due to the intervention token being part of the input during training, $|\bar{A}|$ is the sum of the number of causal edges *and* the intervention targets. Hence this objective is analogous to Eq. 5.2 except we penalise the *expected* $|\bar{A}|$.

Lagrangian Relaxation for Stable Training

In practice, the model is sensitive to the choice of λ since a high λ can lead to mode collapse. We alleviate this via Lagrangian relaxation, which schedules the regularisation weight. Following our design criteria of matching the prediction accuracy of a fully connected model, we formulate a constrained optimisation problem where we minimise $|\bar{A}|$ under the constraint that $\mathcal{L} \leq \tau$ where τ is the target loss,

$$\min_{\theta} \quad \mathbb{E}\Big[|\bar{A}|\Big] \quad s.t. \quad \mathbb{E}\Big[\mathcal{L}(\hat{s}^{t+1})\Big] \leq \tau. \tag{5.7}$$

We set the target as the loss achieved by a fully connected model. Applying Lagrangian relaxation [156], we obtain the following min-max objective,

$$\max_{\lambda>0} \min_{\theta} \quad \mathbb{E}\Big[|\bar{A}|\Big] + \lambda\bigg(\mathbb{E}\Big[\mathcal{L}(\hat{s}^{t+1})\Big] - \tau\bigg). \tag{5.8}$$

We solve this by taking gradient steps on θ and λ in an alternating manner. To ensure λ is positive, we perform updates in the form of $\lambda \leftarrow \alpha * \exp(\mathcal{L}(\hat{s}^{t+1}) - \tau) * \lambda$, where α is the step size. Intuitively, this has the effect of increasing λ when the error is higher than the target, i.e., weighting the error term when the error is high. This is analogous to the GECO [157] framework for tuning the KL loss in a VAE. In practice, we initialise λ to be high and set τ to be the error achieved by the fully connected model. We use a moving-average estimator for the $\mathcal{L}(\hat{s}^{t+1}) - \tau$ term for stability. We also rewrite the training objective to $(\mathcal{L}(\hat{s}^{t+1}) - \tau) + |\bar{A}|/\lambda$ so that the loss value is more stable. This acts as a scheduling scheme where the optimisation focuses on learning the dynamics and switches to pruning redundant

edges once the error is below the target. At the start of the training, $\log(\lambda)$ stays low as the model prediction error improves. This allows the increase of the number of active edges. As the prediction error becomes low enough, the sparsity penalty automatically increases and the number of edges gradually decreases while maintaining the prediction accuracy.

5.3 Experiments

Our experiments are designed to investigate the following guiding questions:

1. Does sparsity enable local causal discovery from observations?
2. Can the sparse model match the prediction accuracy of fully-connected models?
3. Does learning interventions lead to improved robustness and adaptation sample efficiency?

5.3.1 Experiment Setup

Datasets

We evaluate our model on three domains: *Interventional Pong*, *CREATE*, and *Traffic*. The *Interventional Pong* dataset [110], a standard benchmark for causal representation learning, is based on the Pong game with interventions. The *CREATE* [158] environment is a 2D physics simulation that consists of interacting objects such as ladders, cannons, and balls. This is similar to the PHYRE dataset [159] on which SlotFormer [98] is evaluated, but has more interaction types, which allows for more interventions to better showcase the capabilities of SPARTAN. For these two domains, we obtain object slot representations by providing ground-truth masks for objects and encoding each object separately using a VAE [44]. In order to evaluate our model on more realistic tasks, the *Traffic* domain uses the Waymo Open Dataset [155], which was collected in real life. The task is to predict the motion of the ego vehicle conditioned on the observed vehicle trajectories (i.e. past positions) and map layout lines. For evaluating the accuracy of the learnt causal graphs, we compare against ground-truth causal graphs for

the two simulated domains. For the *Traffic* domain, we compare against human-labeled causal graphs [17]. In the simulated domains, we also train the models using interventional data as described in Sec. 5.2.2. Interventions are defined as changes to the simulated dynamics, such as changing the strength of gravity. We provide further details of the datasets in App. C.2.

Baselines

In order to investigate whether SPARTAN can maintain the state-of-the-art prediction accuracy achieved by Transformer-based models, we compare against a *Transformer* baseline using a Transformer for learning dynamics. This can be seen as an instantiation of the SlotFormer [98] architecture with one-step context length, except the slots are learnt using ground-truth object masks. For the *Traffic* domain, we use MTR [160], a state-of-the-art transformer-based model designed for traffic data, as the base architecture. In terms of causal discovery, existing methods [23, 24] only operate on state observations where each state is represented as a scalar, and are therefore not applicable in our setup, which observes object embeddings. We provide a more detailed discussion on the applicability of these prior approaches and provide extra experiment comparisons in Sec. 5.3.4. Closest to our work is CoDA [25], which argues that soft attention in Transformers alone is sufficient for learning local graphs and proposes a thresholding heuristic. We compare against this approach and show that sparsity regularisation in SPARTAN is crucial in local causal discovery. To corroborate that local causal graphs are more suitable for modeling physical interactions, we further compare against the *Global Graph* baseline, which is based on the VCD transition function proposed in Chapter 3, and learns a fixed global causal graph using a masked MLP. The details of the baseline models are provided in App. C.1.

5.3.2 Graph Learning

The prediction accuracy of the models are shown in table 5.1. Overall, SPARTAN is able to make predictions at a similar level of accuracy to, and often outperform,

Table 5.1: Rollout prediction error and average SHD between the learned graphs and the ground-truth causal graphs. Lower is better.

Method	Intv. Pong		CREATE		Traffic	
	Pred. Err.	SHD	Pred. Err.	SHD	Pred. Err.	SHD
SPARTAN (Ours)	**8.60**	**1.51**	**246.6**	**1.17**	0.52	**6.84**
Transformer	8.83	6.37	265.5	6.29	**0.51**	10.6
Global Graph	11.36	5.42	277.4	8.77	–	–

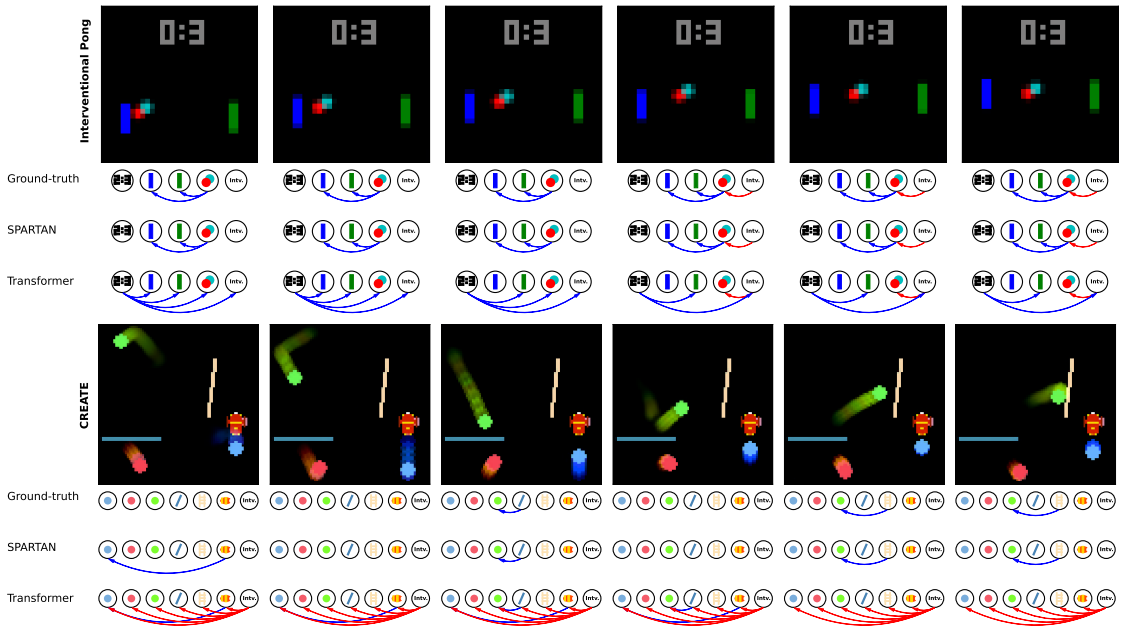


Figure 5.2: Example rollouts in the two simulated environments with the learnt local causal graph visualised. Blue and red arrows between icons indicate the learnt causal edges and intervention targets, respectively. In the *Interventional Pong* example, the intervention is that the ball slows down in the middle. SPARTAN correctly identifies the same causal dependencies as the ground-truth (e.g., ball causes the paddles to follow). The Transformer baseline learns edges that do not correspond to the ground-truth. Similarly, in *CREATE*, SPARTAN learns the correct causal edges (e.g. green ball bounces off the blue plank) while Transformer learns many spurious interventions.

the Transformer baseline. We investigate whether casual discovery via sparsity regularisation is required to identify correct local causal graphs in the setting of dynamics modelling. While [25] suggests that extracting causal graphs by thresholding attention patterns is sufficient in simple environments, here we demonstrate that in more complex environments, learning *hard* attention patterns via sparsity regularisation is *necessary* for accurate local causal discovery. Fig. 5.2 visualises

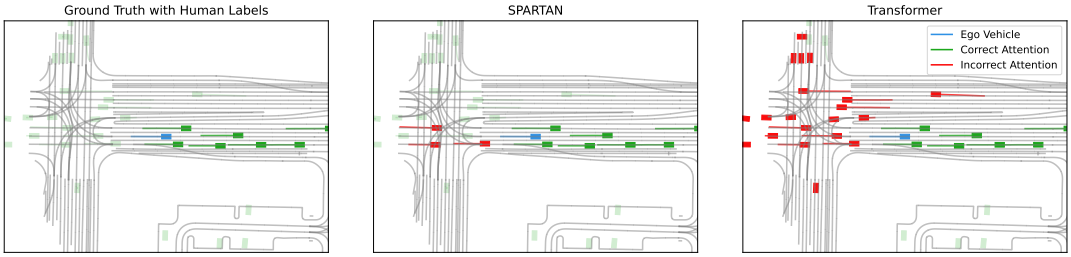


Figure 5.3: Visualisation of the causal relationships learned by the models compared to human-labeled data. The ego vehicle is blue. Transparent rectangles mean that the vehicle is not a causal parent of the ego vehicle. SPARTAN learns a similar attention pattern to human data (e.g. focus on vehicles that are moving in the same direction), whereas Transformer learns many spurious edges.

example rollouts with the local causal graphs.[2] Here, we observe that SPARTAN can reliably recover the relevant causal edges and intervention targets at each timestep. In the Interventional Pong environment, our method identifies that the paddles are constantly influenced by the ball as they follow it. In contrast, the Transformer baseline gives erroneous edges such as "Score → Ball" (frame 1). Similarly, in the CREATE environment, SPARTAN accurately discovers sparse interactions between objects. In the *Traffic* domain, Fig. 5.3 shows that SPARTAN learns to attend to vehicles in neighbouring lanes, which is largely consistent with the human labels, whereas the Transformer baseline attends to many irrelevant vehicles. In this case, heuristics based on spatial proximity would also fail, since there are vehicles that are far ahead that are relevant, while vehicles that are close by but are in the opposite lane, for example, should be ignored. This further highlights the importance of learning local causal graphs. Quantitatively, we evaluate the Structural Hamming Distance, a commonly used metric in graph structure learning, between the learned graphs and the ground-truth. Table 5.1 shows that SPARTAN achieves significantly lower distance compared to the baselines across all domains. We do not apply the Global Graph baseline to the *Traffic* domain since the fixed graph approach cannot be extended to include different numbers of objects.

[2]When evaluating the Transformer baseline, we pick a threshold that achieves the best structural hamming distance. Note that this is not possible in practice without access to ground-truth information.

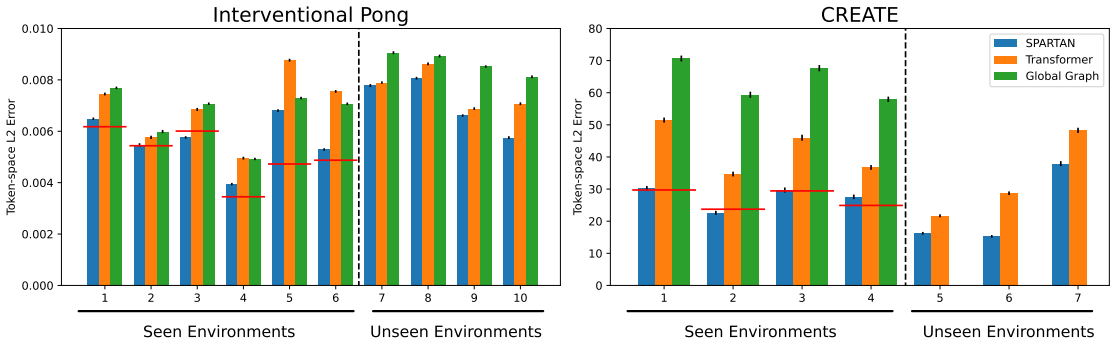


Figure 5.4: Adaptation errors on the two datasets. Each model has access to 5 trajectories with an unknown environment index. "Unseen environments" refers to interventions that are not in the training set. Red lines indicate the prediction error if the environment index is provided, which serves as a lower bound. SPARTAN (blue) consistently achieves the lowest errors across environments. Remarkably, in seen environments, SPARTAN can often reach the lower bound with only 5 observed trajectories.

5.3.3 Robustness and Adaptation Efficiency

Roelofs et al. [17] proposed to evaluate the robustness of models by considering the change of the prediction error when non-causal entities are removed from the scene. This can be seen as a special case of interventions that only target variables that are not the cause of the predicted variable. Concretely, for each object in the scene, we remove any object that is not in the ground-truth parent set and compute the mean absolute value of the percentage change in the prediction error. Here, a model that correctly captures the causal structure would be robust, whilst a model that learns spurious correlations would have large fluctuations. Table 5.2 shows that compared to a non-regularised Transformer, SPARTAN is significantly more robust, suggesting that the Transformer overfits to spurious correlations between entities. In the *Interventional Pong* domain, the ground-truth causal graph is largely consistent across time steps. As such, the Global Graph baseline is sufficient in capturing the causal structure and hence remains robust. However, this degrades in the more complex CREATE domain, where the local causal graph is more state-dependent, while SPARTAN achieves the best robustness. In the Traffic domain, we also implement a local attention variant of the Transformer baseline, proposed in MTR [160], which applies a spatial K-nearest neighbour mask on the attention pattern between the tokens. Whilst this heuristic improves the model, SPARTAN

Table 5.2: Percentage change of error when non-causal entities are removed. Lower is better.

	SPARTAN	Transformer	Local Attn.	Global Graph
Intervention Pong	24.5 ± 4.4	1140.2 ± 15	–	**22.2** ± 1.3
CREATE	**28.6** ± 1.7	138.2 ± 5.6	–	73.6 ± 1.3
Traffic	**13.9** ± 0.2	32.8 ± 0.4	25.5 ± 0.3	–

remains significantly more robust to non-causal changes. A wide range of model architectures is evaluated in [17], with reported changes ranging from 25% to 38%, which is consistent with our findings. Remarkably, the authors in [17] propose a data augmentation scheme that requires ground-truth information, resulting in a change of 22%, which is still significantly worse than what SPARTAN achieves. This further highlights the efficacy of our method. To investigate adaptation efficiency, we adapt the models to a sample of five trajectories from an intervened environment. For SPARTAN, we perform adaptation as described in Sec. 5.2.2. We follow a similar procedure of optimising over intervention tokens for the Global Graph baseline. For the Transformer baseline, adaptation is performed via gradient finetuning on the provided trajectories. We also adapt the models to previously *unseen* environments where the interventions performed are not included in the training set. In the Interventional Pong dataset, the new environments are obtained by combining interventions from the training set, e.g. intervening on the ball and the motion of the paddle at the same time. In CREATE, we change the scene composition as well as the dynamics, i.e. we change the number of objects in the scene. In this scenario, the Global Graph baseline cannot be applied. We note that this is a fundamental challenge for fixed, global graph based methods. We present the results in Fig. 5.4, which shows that SPARTAN consistently outperforms the baselines in this few-shot adaptation setting.

5.3.4 Relations to Related Approaches

The aim of this work is to integrate the inductive bias of sparse interactions common in causal discovery into a transformer-based architecture. As such, we focus our

evaluations on environments that are commensurate in complexity and scale to settings on which prior transformer world models are evaluated. While there exist prior approaches for local causal learning, these methods cannot be readily scaled to the environments in the experiments presented in this chapter. This is due to two principal reasons: first, prior methods operate directly on state-based observations such as x, y position of objects whereas in our environments objects are represented as learnt embeddings (typically 64 dimensions) that are mapped from images; second, and more fundamentally, our environments, in particular CREATE and traffic domains, require the world model to operate on different numbers of objects and different compositions of objects across samples, e.g. different number of vehicles across scenes. A transformer-based world model can accommodate these requirements by treating each object embedding as a separate token. Here, we provide a discussion of the particular architecture choices of prior works that hinder their application in these environments, and highlight the flexibility of our approach.

- ELDEN [23] explores the role of local causal models in the context of intrinsic motivations for RL. Here, the forward dynamics model is trained with a sparsity regularisation, proxied by using the L1 norm, on the partial derivatives between the states. In principle, this technique can be extended beyond scalar states to the case where each object is represented by high-dimensional embedding vectors. However, we implemented ELDEN for our experiments and found that, since the loss is a function of many partial derivatives, the optimisation process becomes prohibitively memory-intensive and highly unstable.

- FCDL [24] investigates the robustness improvement afforded by local causal models. This approach learns a *codebook* of possible local causal graphs and infers the correct structure conditioned on the state. Sparse regularisation is then applied to the set of local causal graphs represented by the codebook. The drawback of this approach is that the local graphs must have the same

set of nodes across samples, meaning that all scenes must have exactly the same composition of objects.

- CAI [141] considers a related but different setting to ours, where the goal is to estimate the causal influence of an agent on the states of the objects in the scene. This approach relies on testing the conditional independence using conditional mutual information estimation between the agent's action and the objects based on a learned dynamics model. However, estimating conditional mutual information is challenging in the multivariate case.

- CODA [25] uses threshold heuristics to determine local structures. We compare against this in our main experiments and show that attention alone is not able to reliably learn the correct dependency structure.

5.4 Conclusion

We tackle the problem of adaptation in world models through the lens of local causal models. To this end, we propose SPARTAN, a structured world model that jointly performs dynamics model learning and causal discovery. We show on image-based datasets that attention mechanisms alone are not sufficient for discovering causal relationships and develop a novel sparsity regularisation scheme that learns accurate causal graphs, resulting in significantly improved interpretability and few-shot adaptation capabilities without compromising on prediction accuracy.

Limitations

There are several limitations to our approach which serve as pointers for future investigations.

- **Identifiability.** We empirically show that local causal graphs can be learnt from data. While the use of sparsity regularisation is grounded in prior works that have theoretical guarantees [24], we do not make theoretical statements about the identifiability of local causal graphs in the setting where the scene

composition can vary between each data sample. Future work should explore the conditions under which local graphs can be identified.

- **New Mechanisms.** During adaptation, SPARTAN adapts by optimising over the learnt "intervention space". We have shown that this is sufficient to generalise to combinations of seen interventions or different numbers of objects. However, in more extreme cases where the test time environment contains completely new behaviours (e.g. ball teleporting), there might not be a corresponding intervention token in the space, while a finetuning approach would be able to converge to the right behaviour (given enough data). One avenue of exploration is to consider procedurally generating intervened environments, akin to domain randomisation [161, 162], to cover the space of all meaningful interventions.

- **Object Encoders.** Our approach relies on pre-disentangled object representations. This is primarily to facilitate the analysis of the learnt causal graphs, which requires a one-to-one mapping between ground-truth objects and embeddings. We have thus focused our analysis on the case where object representations are available. In App. C.3, we present complementary results using a slot-attention based object encoder [90], showing commensurate performance to our main results.

6

Sparse Transformers Beyond World Models

One of the overarching motivations for explicitly learning structures is to improve the *interpretability* of machine learning models. The ability to inspect and understand the internal workings of a model is a crucial aspect of modern machine learning systems, offering a means to build trust, perform model diagnostics, as well as provide insights into the underlying data and tasks. Yet, despite their myriad capabilities, most machine learning models remain opaque black boxes that are difficult to interpret. To this end, the sparse attention mechanisms in SPARTAN provide a natural avenue for interpretability, as the sparse attention patterns offer fine-grained insights into how information flows through the model. In Chapter 5, we verified that sparse attention patterns faithfully reflect the underlying dependency structure in the context of world models. One particular design decision was that the model architecture was to mimic the structure of a standard Transformer, using the same attention mechanism and feedforward layers, but replacing the dense attention with a sparsely sampled *hard attention*. As a result, SPARTAN can be used as a drop-in replacement for standard Transformers, opening up a wide range of potential applications beyond world models.

The goal of this chapter is to explore the benefits of sparse attention in other domains with a focus on interpretability. We present two case studies: first, we explore using a sparse vision Transformer for function approximators in reinforcement

learning, showing that sparse attention successfully captures decision-relevant features; second, we leverage the framework of *mechanistic interpretability* to show that sparse attention offers a post-training strategy that significantly enhances the interpretability of large language models.

6.1 Direct Advantage Estimation with Sparse Transformers

As reinforcement learning (RL) agents are deployed in increasingly complex, real-world environments, developing agents that can act in a robust and generalisable manner in the face of changes in the environment, such as visual distractors, presents a critical challenge. Motivated by this, the notion of causality has garnered significant interest within the context of RL [163] as it offers a principled framework for understanding generalisation in learning systems. In this light, recent works [26, 61] elicit the role of the *advantage function* in causal reinforcement learning, highlighting two principal connections: first, the advantage function can be viewed as modelling the causal effect of actions on the return, and second, under the ExoMDP framework [164], the advantage function inherently only depends on endogenous variables, i.e., variables that are within the agent's control. In other words, the advantage function, and by extension any policy that derives from it, is invariant to distractor variables. Empirically, Pan et al. [61] investigated the properties of learned representations in simple linear environments, showing that learning advantage functions *induces* endogenous representations that isolate the decision-relevant part of the state space. This provides a promising first step towards learning robust and generalisable representations in more complex, high-dimensional environments.

In this case study, we scale this approach of learning advantage functions to image-based environments, and aim to address two questions: (1) Can learning advantage functions lead to better generalisation in image-based environments? (2) If so, is the generalisation capability due to the model attending to only the decision-relevant parts of the image? This latter question is fundamentally one of interpretability, for we seek to understand *why* a model behaves in a certain

way. In contrast to linear environments, where this can be directly verified by inspecting the learned weights, image-based environments pose a unique challenge as the decision-relevant part of the state can be rendered in different parts of the image over time. One possible approach is to use the attention mechanisms in the Transformer architecture as a proxy for the learned dependency structure. However, as empirically demonstrated in Chapter 5, Transformers often attend to spurious features and do not faithfully reflect the causal structure of the problem at hand. To this end, we leverage the sparse attention mechanism developed in Chapter 5 to learn sparse attention patterns that better reflect the underlying dependencies.

In the presented experiments, we train a Vision Transformer [48], augmented by sparsity regularisation, via Direct Advantage Estimation [26] in the CoinRun environment from the Procgen suite [165], an image-based environment with procedurally generated levels. We show that explicit advantage learning outperforms a baseline that learns the Q function via SARSA [166] in terms of generalisation, and that compared to a standard Transformer, a SPARTAN-based function approximator can better extract interpretable dependency structures that attend only to task-relevant parts of the observation.

6.1.1 Background and Approach

We consider the Exogenous MDP (ExoMDP) setting, which provides a more fine-grained view of the decision-making process by separating variables into those that are causally related to the agent's actions (endogenous variables) and those that are not (exogenous variables). Fig. 6.1 shows a comparison between the causal graphs of a normal MDP and an ExoMDP. Specifically, in an ExoMDP, the state space is factorised into $\mathcal{S} = \mathcal{S}^e \times \mathcal{S}^x$, where $s^e \in \mathcal{S}^e$ are the endogenous states and $s^x \in \mathcal{S}^x$ are the exogenous states. The transition probability is factorised as $p(s_{t+1}|s_t, a_t) = p(s^e_{t+1}|s^e_t, a_t)p(s^x_{t+1}|s^x_t)$, where a_t is the action taken by the agent at timestep t, and the reward function is factorised as $r(s, a) = r_e(s^e, a) + r_x(s^x)$, where r_e and r_x are the reward functions for the endogenous and exogenous states, respectively.

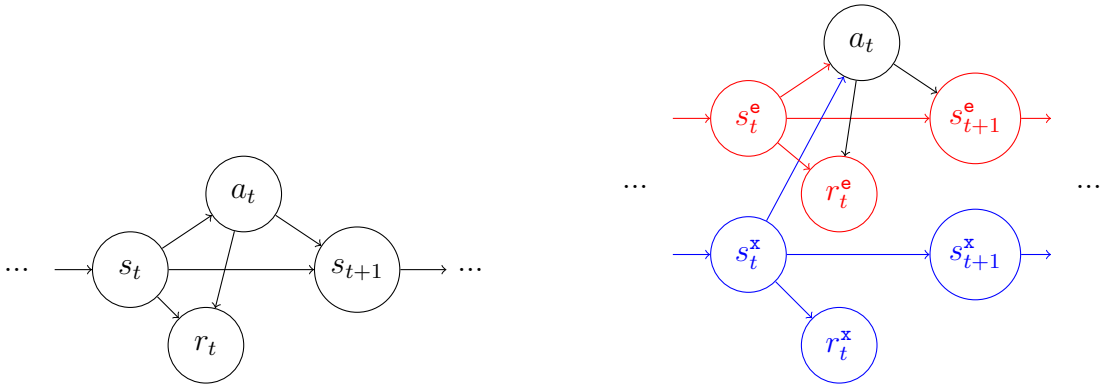


Figure 6.1: Graphical models of MDPs (left) and ExoMDPs (right). Colors and superscripts indicate whether the variable or relationship is **e**ndogenous (red) or e**x**ogenous (blue). In the ExoMDP setting, the state and reward are factorised into endogenous parts, which are dependent on the agent's actions, and exogenous parts, on which the agent has no influence.

Direct Advantage Estimation

Pan et al. [26] showed that the advantage function can be seen as quantifying causal effects of actions on the return and proposed Direct Advantage Estimation (DAE), an on-policy method that can estimate the advantage function directly. Specifically, the learning process is formulated as minimising the following constrained least-squares objective which is similar to standard Q-learning objectives:

$$\mathcal{L}(\hat{A}, \hat{V}) = \mathbb{E}_\pi \left[\left(\sum_{t=0}^{n-1} \gamma^t \left(r_t - \hat{A}_t \right) + \gamma^n V_{\text{target}}(s_n) - \hat{V}(s_0) \right)^2 \right] \tag{6.1}$$

$$\text{s.t.} \quad \sum_a \hat{A}(s,a)\pi(a|s) = 0 \tag{6.2}$$

where V_{target} is the bootstrapping target, and $(\hat{A}, \hat{V})$ are the variables to be minimised. Minimising this objective can be seen as a multi-step estimate of (A^π, V^π). In [61], it is shown that, unlike other value functions such as V or Q, the advantage function is inherently endogenous (i.e., it only depends on the endogenous part of an ExoMDP), which motivates the hypothesis that learning the advantage function using DAE can lead to representations that are also endogenous. The same work empirically verified this in a toy linear ExoMDP by inspecting the linear weights learned by DAE are less dependent on exogenous variables compared to representations induced via learning the Q function. However, scaling this approach up to more complex image-based environments remains an open challenge.

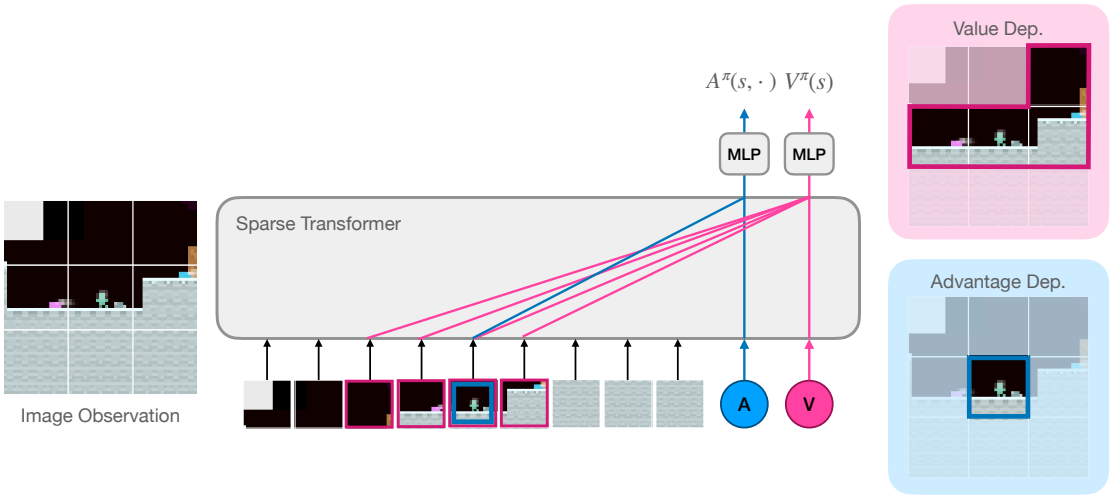


Figure 6.2: The overall architecture of the proposed method. A sparse transformer operates over patched tokens and query tokens, which outputs the learned advantage and value functions. The learned dependencies of the value functions are extracted via the sparse attention patterns. For illustrative purposes, we only show 9 patches.

Sparse Vision Transformer

Our goal is to develop a flexible function approximator that can be used to learn the value and advantage functions. Specifically, we build on the Vision Transformer (ViT) [48] architecture. Image observations are first segmented into patches, which are subsequently encoded into patch embeddings via early convolutions [167]. These patch tokens serve as the input to the Transformer encoder. To extract the advantage and value functions, we append two learnable query tokens (one for advantage and one for value) to the patch tokens. This is akin to the `[cls]` token, which extracts image information for classification in a standard ViT. The transformer outputs for the query tokens are mapped to the correct dimensions, i.e., one for the value function and the number of actions for the advantage function, via separate MLPs. Fig. 6.2 illustrates the overall pipeline of our approach.

The ViT, convolution embedding, and query tokens are jointly trained to minimise the DAE objective (eq. 6.1) plus a sparsity regularisation term via a PPO style [168] on-policy value-based algorithm. Following our approach in Chapter 5, we aggregate sparse attention patterns across layers using the path matrix,

$$\bar{A} = \prod_{l} (A^l + \mathbb{I}), \tag{6.3}$$

where A^l is the adjacency matrix at layer l, and $\mathbb{I}$ is the identity matrix, to extract and visualise the learned dependencies of the advantage and value functions. In the following section, we empirically show that the advantage estimation approach achieves improved generalisation, and verify that sparsity induces attention patterns that only depend on a small number of decision-relevant image patches.

6.1.2 Experiments

We compare the performance of the Direct Advantage Estimation method against a baseline that estimates the Q-function. The baseline uses the same architecture as our method (replace the A/V token by a Q token), but replaces DAE with multi-step SARSA to learn the Q-function. To evaluate the generalisation performance of the learned policies, the policies are trained on a fixed number of procedurally generated CoinRun levels, and test on *unseen* levels. Each policy is trained for 250 iterations, where each iteration corresponds to 8192 (parallel actors) $\times$ 64 (multi-step) = 524288 frames. Details of the training procedure and the experimental setup can be found in Appendix D.1. We evaluate the agent at the end of each iteration by averaging the scores over 1000 episodes from randomly sampled test levels. Each configuration is repeated for 3 random seeds. In the following, we first provide empirical evidence for the improved generalisation and robustness of DAE compared to SARSA, and subsequently demonstrate how the interpretability afforded by sparse attention offers explanations for the improved generalisation performance.

Generalisation and Learning Efficiency

We present the results on generalisation performance in Fig. 6.3 (top row), showing that DAE is not only more data efficient in terms of training frames, but also demonstrates better generalisation performance across different training levels. Here, we find that while the generalisation performance of both methods is similar when the number of training levels is very low (10^2 levels) or very high (10^4 levels), DAE is able to generalise to new levels better when trained on a moderate number

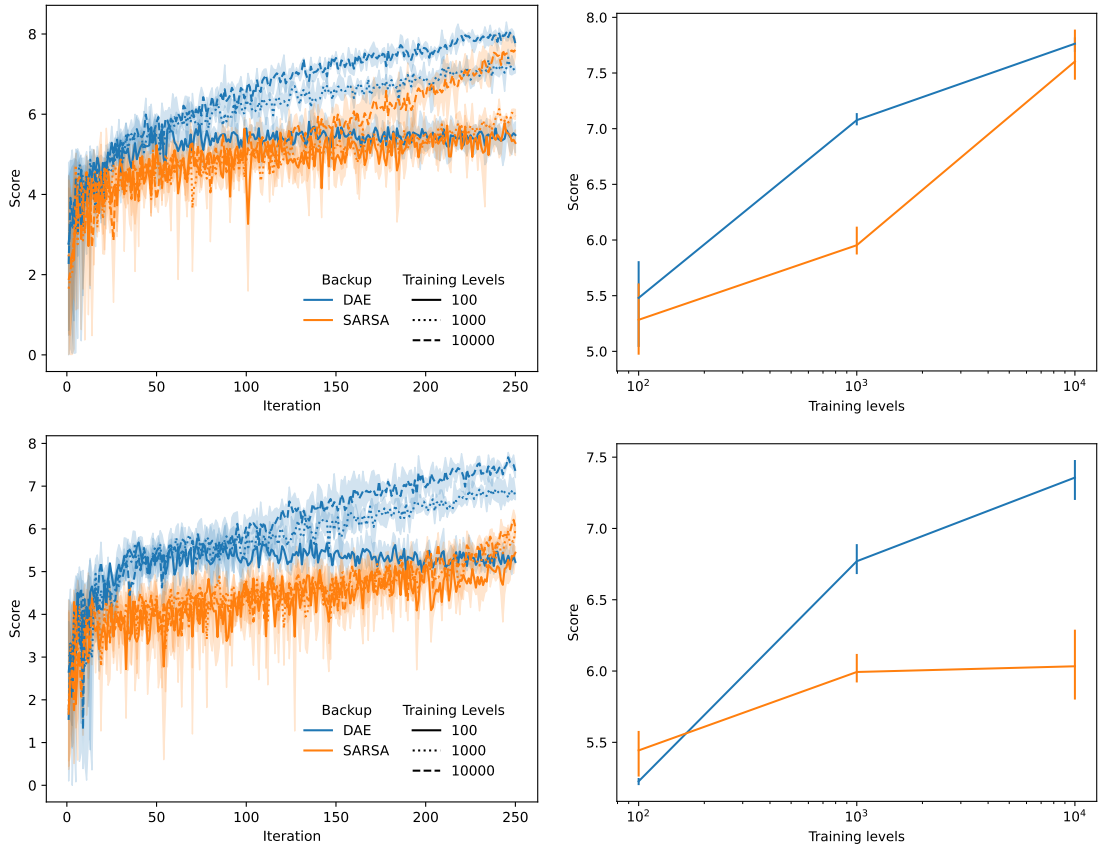


Figure 6.3: Test performance without exogenous rewards (top) and with exogenous rewards (bottom). Left: Learning efficiency. Right: Final evaluation score. Lines and shadings (error bars) represent the mean and min/max score over 3 seeds.

of levels (10^3 levels). This corroborates our hypothesis that learning the advantage function leads to more efficient generalisation.

The CoinRun environment, while capable of generating diverse levels through procedural generation, does not include distractive rewards (e.g., r_x in an ExoMDP). To further test the robustness of both DAE and SARSA, we consider a more challenging version of CoinRun, where we add an additional exogenous reward that depends on the randomly sampled background image of a level. Note that, since the background image does not depend on the agent's actions, this reward function is exogenous by the definition of an ExoMDP. In Fig. 6.3 (bottom row), we find that SARSA struggles to generalise in this case, even with 10^4 training levels, while DAE shows similar trends of generalisation with respect to the number of training levels. This suggests that estimating the advantage directly via DAE

can improve the robustness of the policy when there are distractor rewards that are not causally related to the agent's actions.

Interpreting Dependencies

One potential explanation for the improved generalisation performance of DAE is that the advantage function only depends on the endogenous part of the state, and thus can ignore distractor variables. To qualitatively investigate this hypothesis, the sparse attention patterns learned by the sparse transformer enable us to directly visualise the learned dependency structures. Fig. 6.4 presents two examples of how A, V, and Q depend on the image observation. Here, we see that the advantage function's dependency map is much more sparse compared to the V or the Q function: the advantage function only depends on the velocity of the character (rendered in the upper-left corner as grey blocks) and the immediate obstacles (e.g. the lava floor on the top example, and the enemy sprites on the bottom example). In contrast, both the value and the Q functions depend on various details of the observation, including various entities and the background image (due to the exogenous rewards). Since the policies were constructed either with Q (for SARSA) or A (for DAE), the advantage function's ability to ignore irrelevant features offers a possible explanation for the better generalisation performance. Note that, by construction, these dependency patterns are not just post-hoc explanations of model behaviour, but they capture the actual information flow in the model and mechanistically reflect how the functions are computed. To highlight the importance of our sparse attention approach, we also train a standard Transformer and visualise its attention patterns in the top row of Fig. 6.4. Here, we see that, consistent with the results in Chapter 5, the attention patterns are insufficient to isolate decision-relevant parts of the image and attend to various spurious features.

6.1.3 Discussion

In this case study, we investigated the generalisation improvements of learning advantage functions via Direct Advantage Estimation in image-based environments.

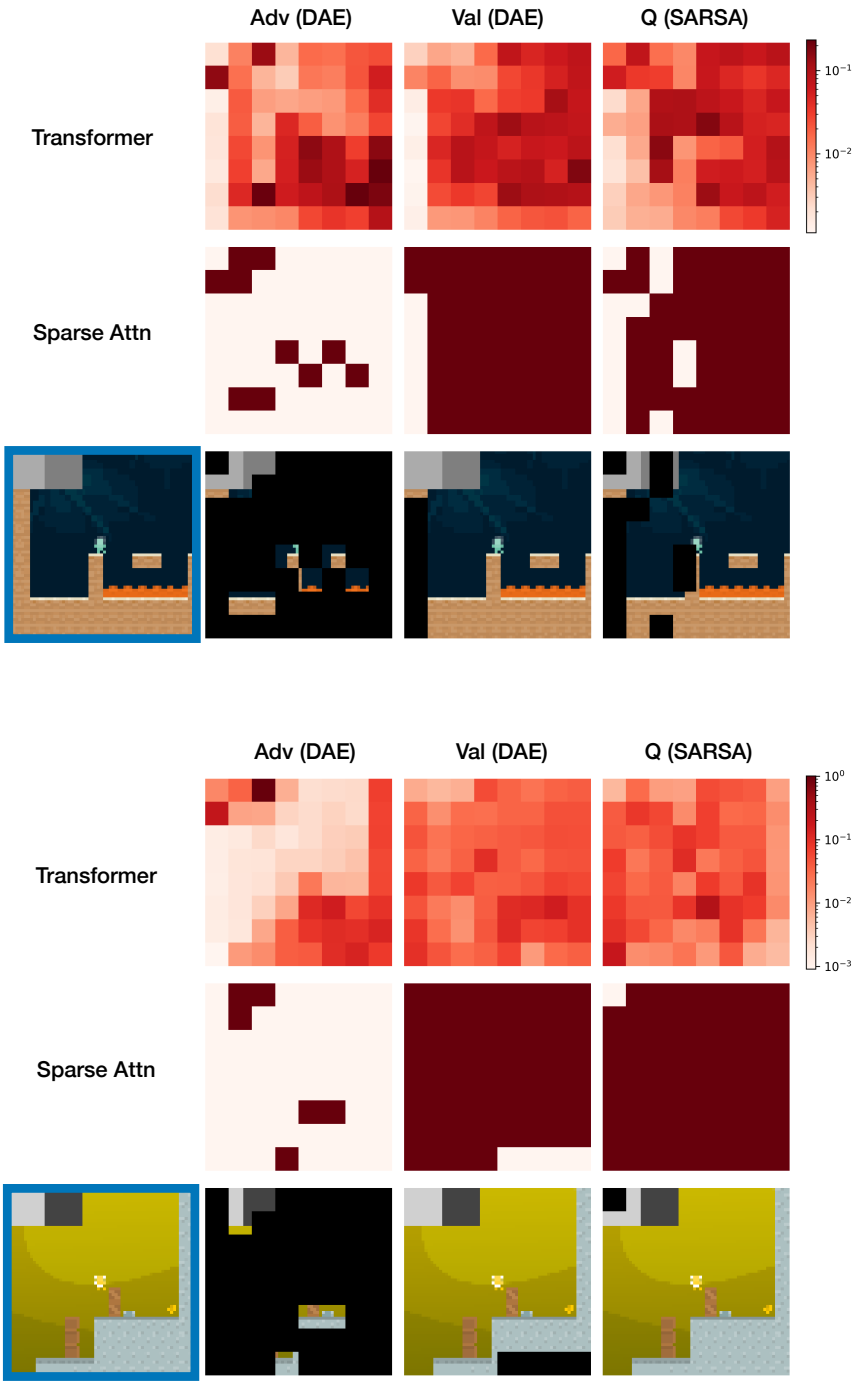


Figure 6.4: Two example attention patterns. Each column shows the attention map for the Advantage, Value, and Q function, respectively. The top row shows the attention patterns obtained from a standard Transformer. The middle row shows the masks obtained from the path matrix from sparse attention. The bottom row shows the image observation and visualises the patches that are not masked out by sparse attention. Only sparse attention is able to isolate the decision-relevant parts of the image (e.g., lava or enemy sprites).

In this setting, we showcased how sparse attention can act as an interpretability tool that offers insights into the learned dependency structures of the value functions, providing qualitative evidence that explains the model's behaviour.

6.2 Sparse Language Models for Mechanistic Interpretability

Scaling has underpinned recent advances in AI: training ever-larger models on internet-scale datasets has yielded striking capabilities across tasks. In natural language processing, large language models (LLMs) now power text generation, translation, and question answering, which further foster their application in many real-world use cases. Yet, this trend of scaling up model complexity exacerbates the growing challenge of interpretability, as it becomes increasingly difficult to understand the inner workings of learned models. In light of this, there has been a growing interest in developing methods that can help interpret and explain the behavior of complex models, in order to ensure their alignment with human values and enable their deployment in safety-critical and ethically sensitive applications [169].

Mechanistic interpretability [170–172] is an emerging paradigm that aims to understand trained models by reverse-engineering their internal mechanisms. The core idea is to analyse the operations of individual model components, such as neuron activations or attention heads, with the goal of identifying the underlying algorithms implemented by the model. Approaches in this field have developed a variety of techniques for post-hoc analysis of trained models, such as sparse dictionary learning and circuit tracing, which has shown promise in revealing the inner workings of complex models. However, the fact that base models are not designed to be interpretable presents a complementary problem: even with sophisticated analysis tools, the revealed algorithms by the models can remain highly complex and opaque. This has led to the question of how we can design models that are *intrinsically interpretable* while still maintaining strong performance. Here, SPARTAN's approach of inducing sparse computation patterns with the constraint of maintaining prediction performance offers a promising avenue towards this goal.

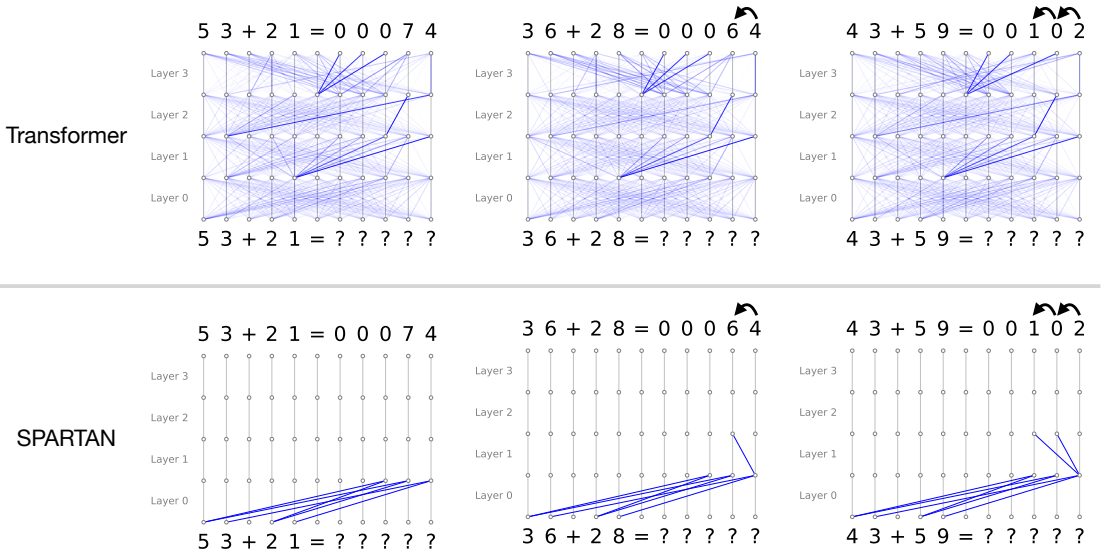


Figure 6.5: Simple example showing the attention patterns (shown in blue) of SPARTAN and Transformer trained on a two digit addition task. In the Transformer plot, the weights of the lines indicate the corresponding attention weights. The input is shown on the bottom and the prediction on the top. Both models are able to correctly predict the sum, but the attention patterns are very different: Transformer solves the task with a dispersed information flow, while SPARTAN uses a highly interpretable attention pattern. In particular, we can see that in Layer 0, the model first attends to the corresponding digits to be added, then in Layer 1, it attends to the carry bit *only if* it is needed (see middle and right columns, where the model has to carry once and twice respectively).

In Fig. 6.5, we present a simple example that illustrates the potential of SPARTAN to induce intrinsically interpretable models. Here, we train a small, 4-layer SPARTAN model on a two-digit addition task, where the input is a sequence of digits and the model is trained to predict the sum. In this example, we can clearly see that the pressure of sparsity leads to the emergence of human-recognisable algorithmic patterns: in the first layer, each digit in the answer attends to the corresponding digits in the input, while the second layer computes the carry bit when necessary. By enforcing selective information flow through sparse message-passing, SPARTAN is able to learn crisp and localised mechanisms that are immediately amenable to interpretation.

In this case study, we explore the potential of sparsity-induced interpretability *at scale*: we investigate the use of sparsity regularisation as a *post-training* technique to improve the interpretability of pre-trained LLMs. Specifically, we take pre-trained

language model weights (in particular GPT-2 [173] and LLaMA3 1B [174]), replace the attention layers with the sparse hard attention mechanism, and fine-tune the models on the original language modeling task with a sparsity penalty. Subsequently, we leverage mechanistic interpretability techniques, namely activation patching and ablation, to identify circuits responsible for carrying specific information in the models. Empirically, we focus on the Indirect Object Identification (IOI) task [170], and show that sparse finetuned models leads to significantly smaller circuits that explains model behaviour compared to their dense counterparts.

6.2.1 Background and Approach

Post-Training Base Models

While it is possible to train SPARTAN as language models from scratch, we focus our experiments on using sparsity as a post-training technique for two main reasons: (1) to avoid the high computational overhead of training large models from scratch, and (2) to showcase how sparse attention can be readily incorporated into existing models, highlighting the versatility of the proposed approach. Note that this two-staged approach broadly aligns with the Lagrangian scheduling strategy used in Chapter 5, where the model first learns to predict well before the sparsity constraint is gradually introduced.

The post-training procedure is straightforward: since SPARTAN uses the same architecture as Transformer, with the only difference being how attention is computed (see Eq. 5.3), we can directly take the weights of a pre-trained Transformer model and replace the attention layers with SPARTAN attention layers. The model is then fine-tuned on the original language modeling task (i.e., next token prediction) with the addition of a sparsity penalty that is scheduled using the Lagrangian relaxation method described in Sec. 5.2.3. To ensure that the model do not lose its original capabilities, we set the target loss in the Lagrangian to be the original loss of the pre-trained model, this is equivalent to minimising the number of edges with the constraint of maintaining the original loss.

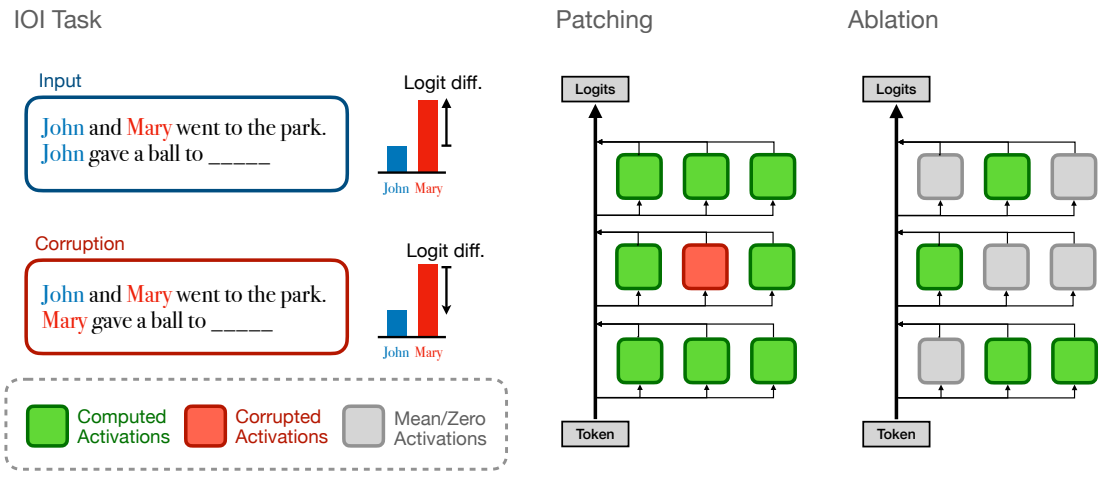


Figure 6.6: We focus on the Indirect-Object Identification (IOI) task. The performance is measured by the logit difference between the correct and incorrect answers. We use activation patching (swapping in activations obtained from corrupted inputs) and ablation (removing specific components) to identify the components responsible for carrying information about the indirect object (IO). Note that in both patching and ablation, we only modify the activations of specific attention heads, while the other components are recomputed as normal.

Activation Patching

There exists a variety of techniques for mechanistic interpretability of language models focusing on different tasks and model components. For this case study, we focus on the Indirect Object Identification (IOI) task [170], which has been widely used as a benchmark for mechanistic interpretability in LLMs. The IOI task involves sentences of the form "[**A**] and [**B**] went to a place. [**A**] gave an object to [**?**]". The model is expected to predict the indirect object (IO) in the sentence (in this case, [**B**]). The goal here is to identify the components in the model that are responsible for moving the information from the IO token to the final prediction.

In our experiments, we use *activation patching* [175] to identify the subcircuits that implement the IOI task. The central idea is to compute the model activations on two different inputs: a 'clean' input that contains the correct input sentence, and a 'corrupted' input where the roles of [**A**] and [**B**] are swapped (see Fig. 6.6, left). By selectively replacing the activations of specific attention heads with the corrupted activations, we can measure the impact of each head on the model's decision (see Fig. 6.6, middle). Concretely, we measure the *logit difference* between the correct

and incorrect answers, [**B**] and [**A**] respectively, as our primary metric. By patching each individual head and measuring the change in logit difference, we can score each attention head based on the size of its effect on the model's prediction.

We use the score obtained from activation patching as a heuristic to identify circuits responsible for carrying the IO information. To quantify whether the identified circuits captures the underlying computation, we choose the top-k heads based on the patching scores, and perform ablation [176] on the rest of the model by 'turning off' the other heads, i.e., setting their activations to zero or a mean value computed from a datset. In the following section, we empirically show that sparse models lead to significantly smaller circuits that explain the IOI behaviour compared to dense models.

6.2.2 Experiments

To evaluate the effectiveness of sparse attention spost-training, we fine-tune pre-trained LLMs and compare their performance and interpretability before and after applying sparsity regularisation. In particular, we perform post-training on two base models: GPT-2 [173] and LLaMA3-1B [174], both of which are Transformer-based language models pre-trained on large text corpora, with 124M and 1B parameters respectively.[1] The models are fine-tuned on the OpenWebText dataset [177], which is an open-source recreation of the WebText corpus used to train GPT-2. Both models are trained on sequences of length 64 with the Lagrangian target set to the original loss of the pre-trained models, as described in the previous section. In the following evaluations, we aim to answer two key questions: (1) Can sparse models maintain comparable performance to their dense counterparts with sparse message-passing? (2) Do sparse models lead to more modularised and interpretable circuits?

Performance vs Sparsity

We first evaluate the performance of the sparse models compared to the dense baselines. Fig. 6.7 shows that both sparse GPT-2 and LLaMA1B models are

[1] We use the open-sourced model weights provided by HuggingFace: `https://huggingface.co/openai-community/gpt2` and `https://huggingface.co/meta-llama/Llama-3.2-1B`.

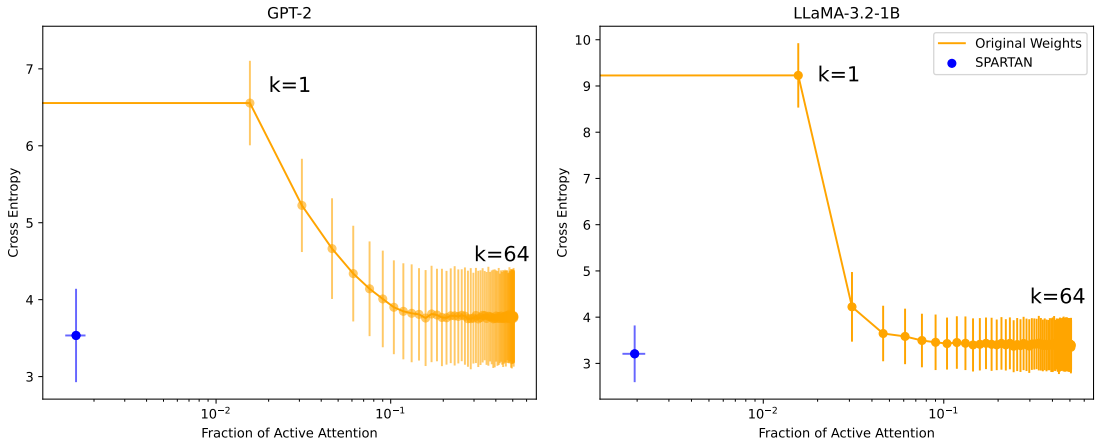


Figure 6.7: Performance, as measured by cross entropy loss (lower is better), for the sparse models and dense baselines. Error bars shows standard deviation over the dataset. The performance of the baselines under top-k attention is also shown for comparison. Sparse models achieve comparable performance to dense models with much fewer active edges, whereas dense models significantly degrades when edges are turned off.

able to maintain comparable performance to their dense counterparts, with only a very small number of active attention edges (roughly 0.2% of the total edges). To further characterise the benefits of sparsity regularisation, we evaluate the dense models using top-k attention, where, for each attention head, only the top-k attention weights are kept with the rest set to zero. In Fig. 6.7, we observe that the performance of dense models degrades significantly under top-k attention, suggesting that the sparse models indeed have learned to implement circuits that move information more efficiently. Remarkably, the average number of active edges in the sparse models is even lower than the top-1 attention setting, indicating that, in general, some attention heads in the sparse models are not used at all.

Circuit Discovery with Activation Patching

The sparsity results above suggest that the sparse models implement efficient algorithms that only make use of a small number of model components. To further investigate the nature of the circuits implemented by the sparse models, we apply activation patching and ablation to identify model components responsible for the IOI task, as described in the previous section. We rank each attention head based on their patching scores, and measure the explained effect when only the top-k heads

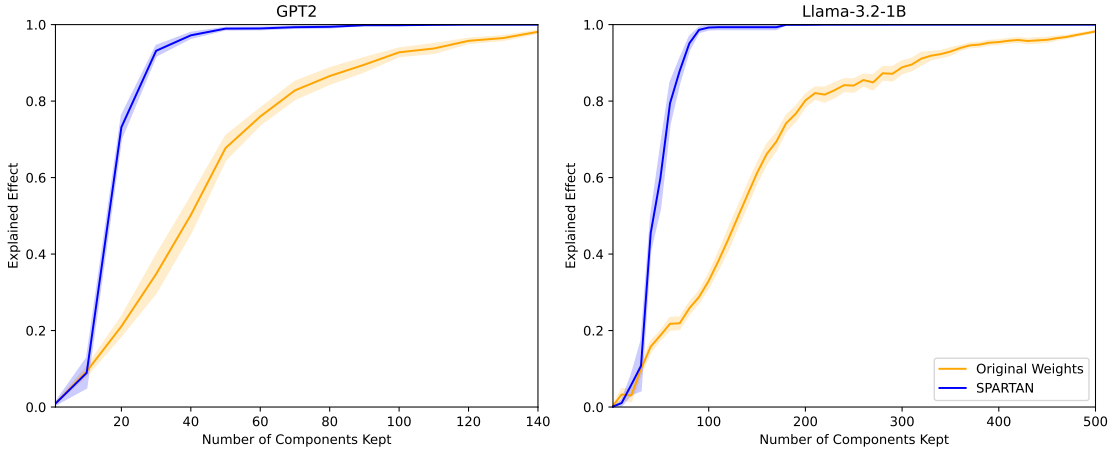


Figure 6.8: Comparison of explained effects between sparse and dense models. In both cases, sparse models are able to achieve high explained effect with fewer attention heads.



Figure 6.9: Attention patterns for GPT2 models for a specific prompt. Red boxes highlights the set of heads that explains 90% of the model behaviour.

are kept. A detailed description of the prompts and task setup, and the ablation procedure can be found in Appendix D.2. Here, the explained effect is defined as:

$$\mathrm{EE}(k) = \frac{|\Delta_k - \Delta_\emptyset|}{|\Delta_{\text{clean}} - \Delta_\emptyset|}, \tag{6.4}$$

where Δ_k is the logit difference when only the top-k heads are kept, $\Delta_\emptyset$ is the logit difference when all heads are ablated, and Δ_{clean} is the logit difference on the clean input. A value of 1 indicates that the top-k heads fully explain the models' behaviour, whereas a value of 0 indicates that the top-k heads have no contribution.



Figure 6.10: Attention pattern for LlaMa models for a specific prompt. Red boxes highlights the set of heads that explains 90% of the model behaviour.

Fig. 6.8 shows the explained effect at different values of k for both sparse and dense models. We observe that sparse models are able to achieve high explained effect with significantly fewer heads compared to dense models, corroborating our hypothesis that sparsity regularisation leads to more easily discoverable circuits. We also visualise the attention patterns of each attention head in Fig. 6.9 and Fig. 6.10 for GPT-2 and LLaMA1B respectively. These visualisations demonstrate the qualitative differences in attention patterns between sparse and dense models: in the sparse models, many attention heads remain inactive while active heads exhibit clear and interpretable attention patterns.

6.2.3 Discussion

In this case study, we have explored the potential of using sparsity regularisation as a post-training technique for large language models, with the goal of improving their interpretability. We show, at scales of up to 1 billion parameters, that the sparsity regularisation scheme developed in SPARTAN is effective in inducing sparse computation circuits without sacrificing performance. By leveraging standard mechanistic interpretability techniques, namely activation patching and ablation, we demonstrate that sparse models lead to significantly smaller circuits that can explain model behaviour. These results highlight the potential of sparsity as a scalable approach for improving the intrinsic interpretability of large language models.

6.3 Conclusion

In this chapter, we investigate applications of the SPARTAN architecture beyond the world model setting. In particular, we leverage the sparse attention mechanism as a means to understand how information flows through the model, enabling us to reveal the underlying computation structure of learned models. We present two case studies that demonstrate the versatility of SPARTAN and highlight its potential for interpretability: first, we explore using a sparse vision Transformer as a function approximator in reinforcement learning, revealing sparse dependency structures that explain the advantage function's generaisation performance; second, we apply SPARTAN as a post-training technique to improve the interpretability of large language models. In both cases, we find that the sparse attention patterns learned by SPARTAN provide meaningful insights into the underlying computation structure of the models, enabling us to extract human-recognisable algorithms and decision-relevant features.

However, we note that the results presented in these case studies are exploratory and are limited in scope. In the RL setting, we only consider one simple visual task. Further work is needed to assess the generality and robustness of our findings across a wider range of tasks and environments. Similarly, in the LLM setting,

we demonstrate the potential of SPARTAN by using a simple activation patching heuristic to select important attention heads at the final token position. More sophisticated methods for discovering circuits, such as path patching [170], and more in-depth analysis of the role of each circuit component could be explored in future work. Nonetheless, these results underscore the potential of sparse attention as a general-purpose tool for interpretability and open many avenues for future research.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

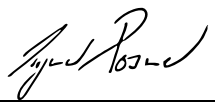
Title of Paper	Learning What Matters: Endogenous Representation through Advantage Estimation and Sparse Attention
Publication Status	Published (Workshop paper)
Publication Details	Hsiao-Ru Pan, Anson Lei, Ingmar Posner, Bernhard Schölkopf. Learning What Matters: Endogenous Representation through Advantage Estimation and Sparse Attention. *In The Causal Reinforcement Learning Workshop at Reinforcement Learning Conference* (2025)

Student Confirmation

Student Name:	Anson Lei		
Contribution to the Paper	The work is carried out during a research exchange at the Max Planck Institute for Intelligent Systems, Tübingen, Germany, as part of the ELLIS programme. My coauthor and I co-conceptualised the project. I implemented the sparse vision transformer code. My coauthor managed the reinforcement learning experiments and analysed the results. I performed the interpretability analysis.		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:		
Supervisor comments		
Signature 	Date	

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

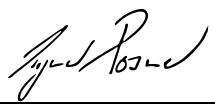
Title of Paper	Sparse Transformers for Intrinsically Interpretable Language Models. (TBC)
Publication Status	To be submitted.
Publication Details	Authors: Anson Lei, Florent Draye, Ingmar Posner, Bernhard Schölkopf

Student Confirmation

Student Name:	Anson Lei		
Contribution to the Paper	The work was initiated during a research exchange at the Max Planck Institute for Intelligent Systems, Tübingen, Germany, as part of the ELLIS programme. My coauthor and I co-conceptualised the project. I implemented and trained the sparse transformer language models. My coauthor implemented the activation patching code. I have produced the results presented in the thesis using my coauthor's implementation.		
Signature 		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:		
Supervisor comments		
Signature 	Date	

7 Conclusion

In this thesis, we explore the role of explicit structures in building flexible, robust, and interpretable world models. The key insight is that while current world model approaches have made significant strides towards capturing the dynamics of complex environments, they do so in a *monolithic* manner. In contrast, our intuitive understanding of how the world behaves is highly *structured*: we recognise objects and entities, their properties, and how they interact with each other. To this end, we take inspiration from the emergent field of *causal machine learning*, and argue that causality provides a natural framework for formalising and reasoning about these structures. Throughout the thesis, we develop a suite of algorithms that leverage causal inductive biases, namely sparsity and modularity, to learn and exploit structures in world models. In this chapter, we summarise the key contributions of this thesis, reflect on the limitations of our current approaches, and outline promising directions for future research.

7.1 Contributions

We organise the review of our contributions around two key questions: (1) *How can we learn structured world models from observations?*, and (2) *In what ways do structures help?* In the following, we address each question in turn, highlighting

the key insights and methodological contributions and summarising the empirical findings that support our claims.

Learning Structured World Models

In Chapter 1, we identified three principal challenges in learning structured world models, namely *Representation Learning*, *Mechanism Learning*, and *Structure Discovery*. In Chapters 3, 4, and 5, we present three approaches, VCD, COMET, and SPARTAN, which address each of these challenges respectively.

Representation Learning. The ability to map high-dimensional observations to well-disentangled and meaningful latent representations, e.g. decomposing images into object representations, is a prerequisite for learning structured world models. The main insight in Chapter 3 is that since sparse causal structures can only be discovered when the underlying variables are well-represented, we can leverage a downstream causal discovery objective as an *inductive bias* to guide representation learning. To this end, we develop VCD, a learning algorithm that uses variational inference to jointly learn a latent representation and perform sparsity-based causal discovery. We provide empirical evidence on an image-based particles environment, demonstrating that, compared to the latent representations learned by monolithic world models, VCD learns significantly more disentangled representations that better capture the underlying factors of variation.

Mechanism Learning. In Chapter 4, we turn our attention to the problem of decomposing the dynamics of the environment into modular, reusable mechanisms. Specifically, we focus on the setting where the environment dynamics can be decomposed into a set of *interaction primitives*, each describing a distinct type of interaction between entities. We propose COMET, a competition-of-experts approach that learns a set of independently parameterised modules, with a competition mechanism that encourages each module to specialise on a specific interaction primitive. We evaluate our approach on a suite of environments with compositional dynamics, ranging from simulated particles to team sports. The results show that, compared to

prior methods which share our aim of learning modular dynamics, the competition mechanism of COMET leads to significantly better decomposition of the dynamics.

Structure Discovery. Another key aspect of structured world models is the ability to explicitly represent how entities interact with each other, i.e. the *dependency structure* between entities. In Chapter 5, we argue that modelling complex physical systems requires the ability to capture *local*, *state-dependent* causal graphs. We present SPARTAN, a sparse attention-based world model that uses sparsity-regularised *hard attention* mechanisms to discover the interaction graph between entities. In the experiments, we show that SPARTAN is able to accurately predict the dynamics of complex systems, including a large-scale real-world autonomous driving dataset. In addition to achieving state-of-the-art prediction performance, we verify that the sparse structure learned by SPARTAN accurately reflects the underlying dependencies in the environment, e.g. influences between vehicles.

Benefits of Structures in World Models

Having developed a suite of algorithms that focus on different aspects of learning structured world models, for each of the three approaches, we investigate empirically how the learned structures can be leveraged to alleviate some of the key shortcommings of current world models. We focus our evaluations on three main benefits: *Adaptation Efficiency*, *Robustness to Distribution Shift*, and *Interpretability*.

Adaptation Efficiency. A key promise of the causal perspective is that, according to the ICM principle and the SMS hypothesis, distribution shifts often only affect a small subset of the underlying causal mechanisms. As a result, one of our main hypotheses is that world models that reflect the underlying causal structure should be able to adapt more efficiently when the environment changes. In Chapter 3 and 5, we demonstrate that both VCD and SPARTAN are able to perform selective adaptation on a small subset of the learned model, and hence adapt significantly faster than monolithic world models when the environment undergoes localised changes. In VCD, this is achieved by leveraging the disentangled representations to

identify and update only the affected latent variables, while in SPARTAN, this is implemented via a learned intervention token that signals which entities are affected by the change. COMET takes an orthogonal approach by explicitly learning reusable mechanisms. Here, COMET gains adaptation efficiency by learning to directly compose existing modules to model novel environments. Across all three approaches, we find strong empirical evidence corroborating our hypothesis that structured world models can adapt more efficiently to localised changes in the environment.

Robustness to Distribution Shift. Another corollary of the ICM principle and the SMS hypothesis is that natural distribution shifts leave much of the underlying causal mechanisms invariant. This suggests that learning the correct causal structure should lead to better prediction robustness, at least for the part of the prediction that is not affected by the distribution shift. In the VCD experiments, this manifests in the fact that the model is able to reuse most of the learned forward model when adapting to a new environment, which still matches baselines that are trained with a large amount of data in the new environment. This suggests that much of the learned prediction models remain valid under distribution shifts. In COMET, our experiments show a similar trend: when composed correctly, learned mechanisms can perform robust predictions in novel environments without the need to update the parameters. We show this robustness property more directly on a larger scale in the SPARTAN experiments, where we evaluate whether the model performance degrades when irrelevant entities are removed from the scene. Remarkably, we find that SPARTAN is more robust even when compared to strong baselines that are trained with privileged, human-annotated data.

Interpretability. Finally, we investigate how explicitly representing structures leads to improved interpretability. Throughout the thesis, we have presented qualitative results that demonstrate how the learned structures can be inspected and interpreted as a means to evaluate the model's behaviour. This, in itself, is a testimony to the improved interpretability of structured world models: in VCD, we can visualise what each latent variable represents; in COMET, we can

inspect the activated mechanisms to understand what type of interactions are occurring; in SPARTAN, we can investigate the sparse attention patterns to probe how information flows through the model. In Chapter 6, we take this a step further and present two case studies that explore the use of the SPARTAN architecture beyond the world model setting. First, we show that by using SPARTAN as a function approximator in the context of reinforcement learning, we can provide a mechanistic explanation of why learning advantage functions leads to improved generalisation. Second, we showcase how SPARTAN can be used as a post-training technique to improve the interpretability of large language models, enabling us to extract human-recognisable algorithms from the learned attention patterns.

Taken together, the methodological contributions and empirical evidence presented in this thesis provide a strong foundation for practical algorithms that can learn and exploit structures in complex environments and paint a compelling picture of the benefits of learning structured world models. In the following section, we reflect on the limitations of our current investigations and outline promising directions for future research.

7.2 Limitations and Future Work

Towards a Unified Framework for Structured World Models

The three approaches presented in this thesis individually address a key aspect of learning structured world models. In reality, however, these facets are intricately intertwined. For instance, in Chapter 3, the main driver for learning disentangled representations is the structure discovery objective; whereas in Chapters 4 and 5, structures and mechanisms are discovered under the assumption that the underlying objects and entities are already well-represented. While studying each aspect in isolation has allowed us to gain insights into the specific challenges and design considerations, this also means that the overall vision of learning fully structured world models directly from observations remains elusive. Nonetheless, the approaches presented in this thesis provide the building blocks to develop a holistic framework for learning structured world models. The techniques developed in VCD, COMET,

and SPARTAN are largely complementary: the sparse attention regularisation scheme in SPARTAN can be used in conjunction with the joint representation learning objective in VCD, while the competition-of-experts mechanism in COMET could be integrated into the transformer-like architecture of SPARTAN as a form of mixture-of-experts. As such, a promising avenue for future research is to investigate how these components can be combined into a unified framework that jointly addresses representation learning, mechanism learning, and structure discovery, and leverage the synergies between them in the context of world model learning.

Towards Agents that Reason about Structures

The scope of this thesis is primarily limited to *learning* world models. However, how a structured world model can be *used* by an agent to plan and make decisions is an equally important, and largely unexplored, question. Here, we highlight two promising directions of future research that build on the findings of this thesis. First, an open question is how decision-making agents can directly benefit from the improved adaptability and robustness of structured world models. The modular nature of the presented world models suggests that they could be well-suited for settings where agents need to adapt to new tasks by composing existing skills and knowledge, or adapt to novel environments with changed dynamics, such as in the context of continual learning. Future research could investigate this hypothesis in the context of embodied agents that act in complex, dynamic environments, such as in robotics or autonomous driving. Second, and perhaps more fundamentally, a promising avenue for future research is to explore how the decision-making process itself can be structured. The reinforcement learning case study in Chapter 6 provides a preliminary step in this direction, demonstrating how sparse attention can be used directly in the policy. Moreover, the explicit representation of structures in the world model also opens up the possibility of agents that reason about the underlying causal structure of the environment, and use this knowledge to guide their actions. For example, existing works on using causal structures as *intrinsic motivation* for

active exploration [e.g. 23, 75] provide a promising starting point that can be readily integrated with the structured world models presented in this thesis.

7.3 Final Remarks

In closing, this thesis puts forth a simple message: in the context of world models, when computation respects causal structure, *adaptability*, *robustness*, and *interpretability* follow. The three approaches proposed in the thesis, VCD, COMET, and SPARTAN, shed light on how sparse, modular mechanisms can be learned from observations, and how these structures can lead to practical and tangible benefits. We hope this thesis serves as a stepping stone towards the broader vision of building intelligent agents that can learn and reason about the world in a structured manner, and that the ideas presented here inspire future research in this nascent and exciting field.

Appendices

A

Appendix for VCD

A.1 Derivations

Using the approximate posterior $q(z^{0:T}|o^{0:T}, a^{0:T}) = \prod_t q(z^t|o^t)$, the variational log lower bound for the latent state-space model (Eq. 3.3) can be derived from importance weighting and Jensen's inequality:

$$\log p(o^{0:T}, a^{0:T}) = \log \int \prod_{t=0}^{T} p(o^t|z^t)p(a^t|z^t)p(z^t|z^{t-1}, a^{t-1})dz^{0:T} \tag{A.1}$$

$$= \log \int \prod_{0}^{T} p(o^t|z^t)p(a^t|z^t)\frac{p(z^t|z^{t-1}, a^{t-1})}{q(z^t|o^t)}q(z^t|o^t)dz^{0:T} \tag{A.2}$$

$$\geq \mathbb{E}_{\prod q(z^t|o^t)}\left[\log\left(\prod_{o}^{T} p(o^t|z^t)p(a^t|z^t)\frac{p(z^t|z^{t-1}, a^{t-1})}{q(z^t|o^t)}\right)\right] \tag{A.3}$$

$$= \sum_{0}^{T} \mathbb{E}_{q(z^t|o^t)}\left[\log p(o^t|z^t) + \log p(a^t|z^t)\right] + \mathbb{E}_{q(z^{t-1}|o^{t-1})}\left[\mathbb{E}_{q(z^t|o^t)}\left[\log p(z^t|z^{t-1}, a^{t-1}) - \log q(z^t|o^t)\right]\right] \tag{A.4}$$

$$= \sum_{0}^{T} \mathbb{E}_{q(z^t|o^t)}\left[\log p(o^t|z^t) + \log p(a^t|z^t)\right] - \mathbb{E}_{q(z^{t-1}|o^{t-1})}\left[\mathbb{KL}\left[q(z^t|o^t)||p(z^t|z^{t-1}, a^{t-1})\right]\right]. \tag{A.5}$$

Since the policy $p(a^t|z^t)$ is constant with respect to the model parameters, we omit this term and write the ELBO objective as

$$\mathbb{ELBO}(\theta,\phi) = \sum_{t=0}^{T} \mathbb{E}_{q_\phi(z^t|o^t)}\Big[\log(p_\theta(o^t|z^t))\Big] \tag{A.6}$$

$$- \mathbb{E}_{q_\phi(z^{t-1}|o^{t-1})}\Big[\mathbb{KL}[q_\phi(z^t|o^t)||p_\theta(z^t|z^{t-1},a^{t-1})]\Big], \tag{A.7}$$

where θ is the model parameter and ϕ is the parameter for the approximate posterior. θ and ϕ are omitted henceforth to simplify notation. In VCD, the KL divergence term can be further decomposed by exploiting the structure of the dynamics model (Eq. 3.8), and the assumption that variables within each timestep are independent:

$$\mathbb{KL}\left[q(z^t|o^t)||p^{(k)}(z^t|z^{t-1},a^{t-1})\right] \tag{A.8}$$

$$= -\int \log\left(\frac{p^{(k)}(z^t|z^{t-1},a^{t-1})}{q(z^t|o^t)}\right) q(z^t|o^t)dz^t \tag{A.9}$$

$$= -\sum_{i=0}^{d}\int \log\left(\frac{p_i^{(0)}(z_i^t|M_i^{\mathcal{G}}\odot[z^{t-1},a^{t-1}])^{1-R_{ki}^{\mathcal{I}}}p_i^{(k)}(z_i^t|M_i^{\mathcal{G}}\odot[z^{t-1},a^{t-1}])^{R_{ki}^{\mathcal{I}}}}{q(z_i^t|o^t)}\right) q(z_i^t|o^t)dz_i^t \tag{A.10}$$

$$= -\sum_{0}^{d}\Bigg((1-R_{ki}^{\mathcal{I}})\int \log\left(\frac{p_i^{(0)}(z_i^t|M_i^{\mathcal{G}}\odot[z^{t-1},a^{t-1}])}{q(z_i^t|o^t)}\right) q(z_i^t|o^t)dz_i^t$$

$$+ R_{ki}^{\mathcal{I}}\int \log\left(\frac{p_i^{(k)}(z_i^t|M_i^{\mathcal{G}}\odot[z^{t-1},a^{t-1}])}{q(z_i^t|o^t)}\right) q(z_i^t|o^t)dz_i^t\Bigg) \tag{A.11}$$

$$= -\sum_{0}^{d}\Bigg((1-R_{ki}^{\mathcal{I}})\mathbb{KL}\left[q(z_i^t|o^t)||p_i^{(0)}(z_i^t|M_i^{\mathcal{G}}\odot[z^{t-1},a^{t-1}])\right]$$

$$+ R_{ki}^{\mathcal{I}}\mathbb{KL}\left[q(z_i^t|o^t)||p_i^{(k)}(z_i^t|M_i^{\mathcal{G}}\odot[z^{t-1},a^{t-1}])\right]\Bigg). \tag{A.12}$$

The KL terms can be computed analytically since the conditional distributions in the last expression are univariate Gaussian distributions. In training time, the gradients through the expectation terms in the ELBO is estimated by drawing a sample from the posterior distribution using the reparameterisation trick [44].

A.2 Implementation Details

Mixed-state

In the mixed-state experiment, all conditional distributions (including encoders, decoders and dynamics models) are parameterised by feedforward MLPs with two

hidden layers of 64 hidden units each. The recurrent modules are implemented as GRUs [149] with 64 hidden units. Distributions in the latent space are 16-dimensional diagonal Gaussian distributions with predicted mean and log variance.

Image

In the image experiment, the encoders and decoders are parameterised as convolutional and deconvolutional networks from [1]. In the RSSM models, the dynamics models are parameterised as feedforward MLPs with two hidden layers of 300 hidden units. The recurrent module is a GRU with 300 hidden units. In VCD, to compensate for the fact each dimension in the latent space has a separate model, the number of hidden units in the GRU and MLP are reduced to 32 to avoid over-parameterisation. We found that initialising the encoders and decoders by pretraining them as a variational autoencoder helped with training stability for both RSSM and VCD.

In both experiments, the training objective is maximised using the ADAM optimiser [178] with learning rate 10^{-3} for mixed-state, and 10^{-4} for images. In both environments, we clip the log variance to -3, with a batch size of two trajectories from each of six environments with $T = 50$. In VCD, the hyperparameters $\lambda_{\mathcal{G}}, \lambda_{\mathcal{I}}$ are both set to 0.01. All models are trained on a single Nvidia Tesla V100 GPU.

A.3 Experiment Details

Interventions

The ground truth states of the multi-body dynamics environment is the x and y coordinates of each particle. The full list of possible interventions on the environment is provided in Table A.1. Note that the forces between particle 1, 3 and 4 are proportional to their masses. Hence intervening on the mass of particle 1 and 3 also affect the dynamics of particle 4. In the experiments, all models are trained in the undisturbed environment and intervened environments 1, 5, 11, 14, 17.

In the mixed-state experiment, the observation function is a mixing matrix where each entry is drawn from a unit Gaussian distribution. In the image experiment, the observation is given by rendering the system to a $128 \times 128 \times 3$ image. In both

Table A.1: List of interventions

ID	Intervention	Intervention targets
1	Remove spring between 1 and 2	x_1, y_1, x_2, y_2
2	Remove spring between 2 and 3	x_2, y_2, x_3, y_3
3	Increase mass 1	x_1, y_1, x_4, y_4
4	Increase mass 2	x_2, y_2
5	Increase mass 3	x_3, y_3, x_4, y_4
6	Decrease mass 1	x_1, y_1, x_4, y_4
7	Decrease mass 2	x_2, y_2
8	Decrease mass 3	x_3, y_3, x_4, y_4
9	Increase spring constant between 1 and 2	x_1, y_1, x_2, y_2
10	Increase spring constant between 2 and 3	x_2, y_2, x_3, y_3
11	Constrain movement of 1 to vertical only	x_1
12	Constrain movement of 1 to horizontal only	y_1
13	Constrain movement of 2 to vertical only	x_2
14	Constrain movement of 2 to horizontal only	y_2
15	Constrain movement of 3 to vertical only	x_3
16	Constrain movement of 3 to horizontal only	y_3
17	Constrain movement of 4 to vertical only	x_4
18	Constrain movement of 4 to horizontal only	y_4

experiments, the models are trained on a training set of 2000 trajectories from each of the six environments and evaluated on a validation set of 400 unseen trajectories.

A.4 Qualitative Exploration

In this section, we qualitatively explore the learnt latent representaitons in both the mixed-state and image experiments.

Mixed-state

Since the observation mixing function is linear and invertable in the mixed-state environment, we can directly access the level of disentanglement of the latent space with respect to the ground-truth state of the environment, i.e. the x and y coordinate of the particles. Fig. A.1 shows the average magnitude of the entries of the Jacobian matrix between the ground-truth state and the learnt latent state. Each entry measures the changes in each ground-truth state variable when the latent representation is perturbed along each dimension. By using the

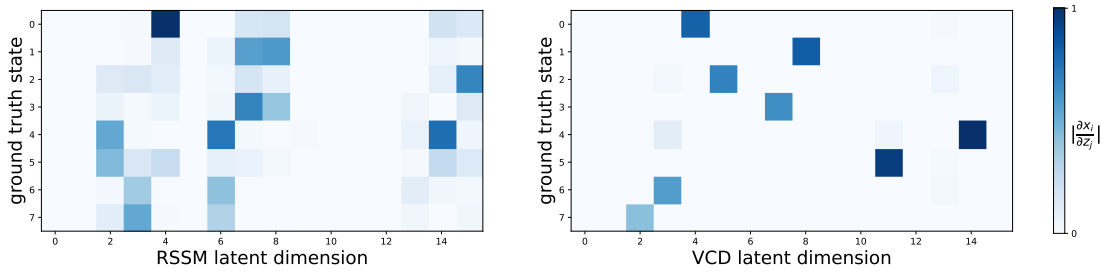


Figure A.1: The average magnitude of the derivative of the ground-truth states x_i w.r.t. the learnt latent states z_j. VCD is able to learn a latent representation where each dimension captures information about one state only, i.e. only one state variable changes when we perturb the latent representation along one dimension. On the other hand, the RSSM latent space encodes state information in a more entangled manner, i.e. multiple ground-truth states are affected when the latent representation is perturbed in one dimension.

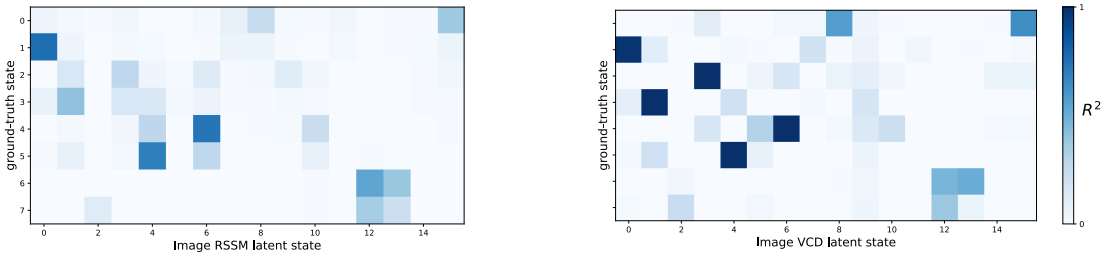


Figure A.2: The R^2 of a linear regression between the ground-truth states and the learnt latent variables. This shows a similar pattern as the mixed-state experiments where the VCD representation in general only capture information about one ground-truth state.

discovery of causal dynamics models as an inductive bias, VCD is able to learn a disentangled representation where each ground-truth state is captured by only one latent dimension. In contrast, RSSM learns a representation that is not sparse.

Image

To inspect the quality of the learnt representation in the image experiment, here we show reconstruction samples along all dimensions of the latent space in the RSSM and VCD representation. Note that since both encoders are initialised from the same pretrained VAE, the difference in the latent space arise because of the causal discovery objective in VCD. Fig. A.4 shows the changes in the reconstruction when the RSSM representation is perturbed in each dimension of the latent space. Fig. A.3 shows the same for VCD. As discussed in the main text, VCD learns a *axis-aligned* representation that affords a sparse causal graph. For example, dimension 1 and

3 captures the y and x coordinates of the green particle respectively. In contrast, RSSM learns a representation that is not axis-aligned. Fig. A.2 plots the R^2 linear regression scores between the learnt latent variables and the ground-truth variables, showing that the VCD representation can in general disentangle individual states, as indicated by the more salient squares.

A.5 Further Experimental Results

In the main text, we have shown that VCD is able to identify sparse interventions and adapt to new environments with significantly improved data efficiency on one particular novel environment (corresponding to intervention 12). Here, we present analogous results on other intervened environments (interventions 4, 9, and 13, which covers all intervention types) in Fig. A.5, Fig. A.6, and Fig. A.7. The results exhibit a consistent pattern where VCD significantly outperforms the baselines in the low data regime by identifying sparse mechanism changes.

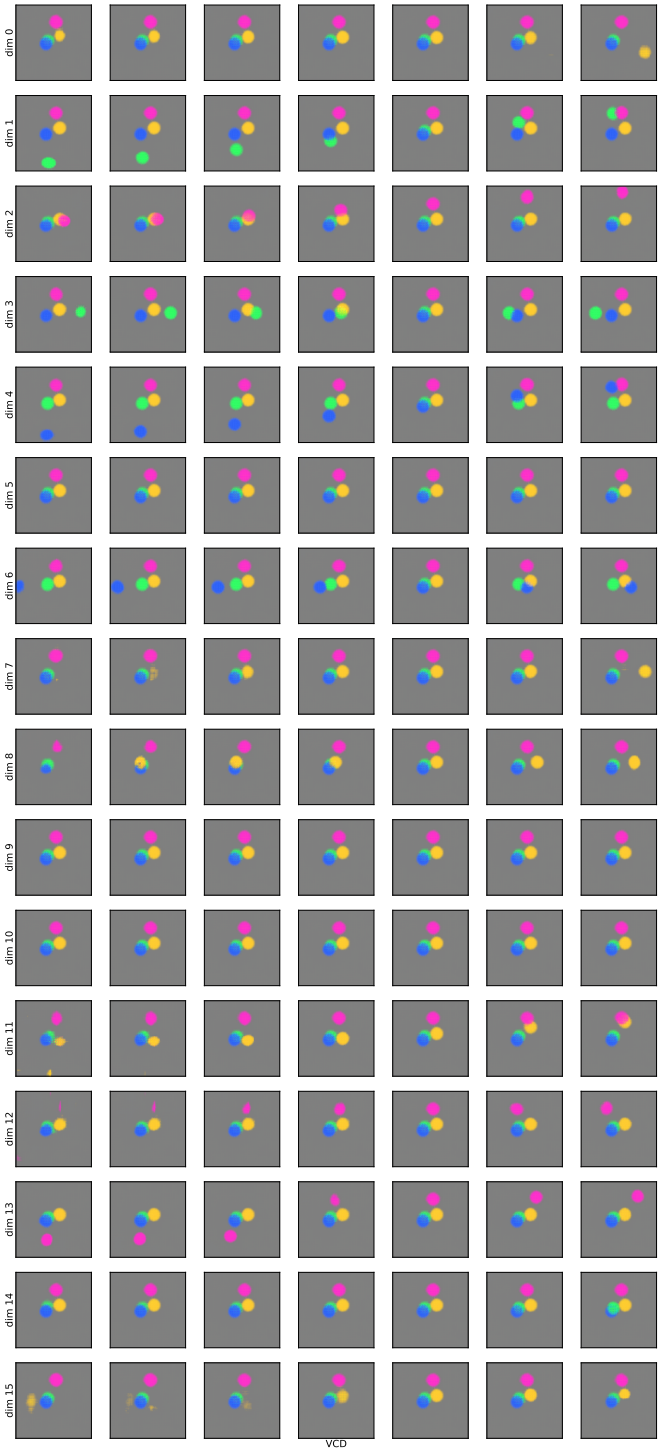


Figure A.3: Visualisation of the learnt latent space in VCD.

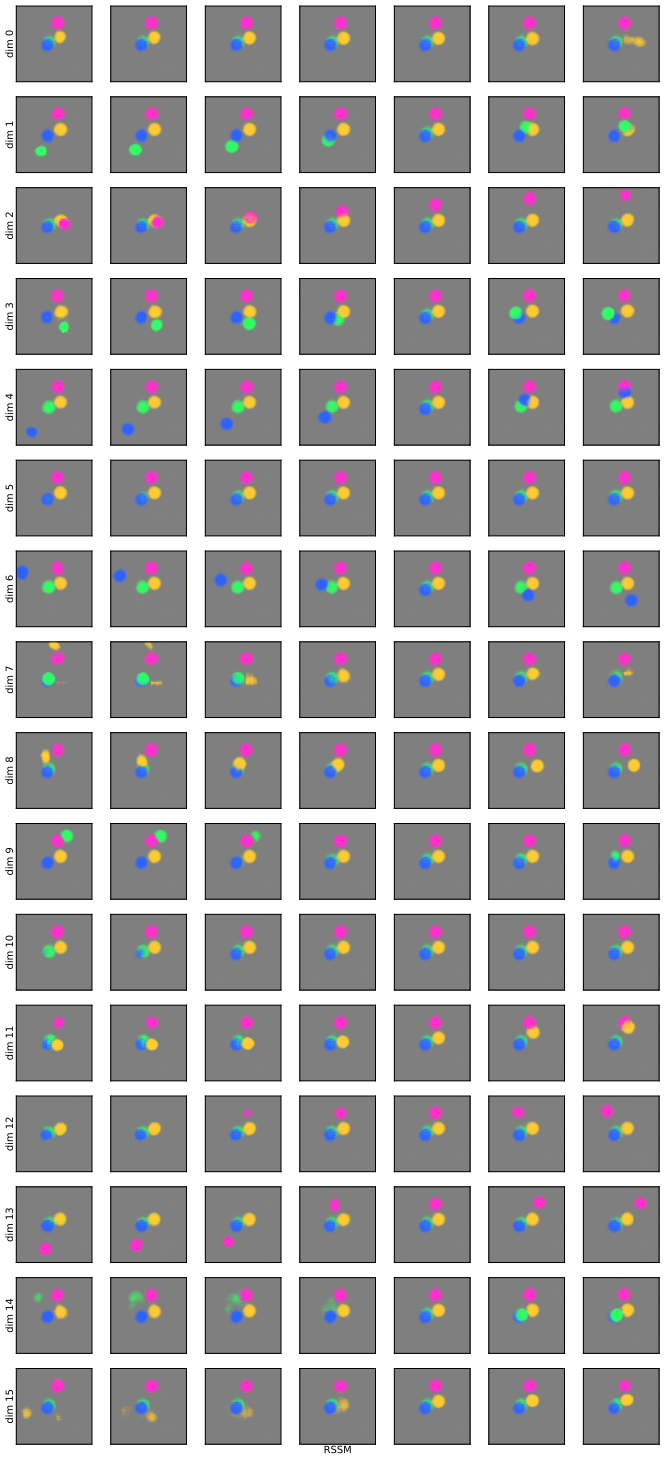


Figure A.4: Visualisation of the learnt latent space in RSSM.

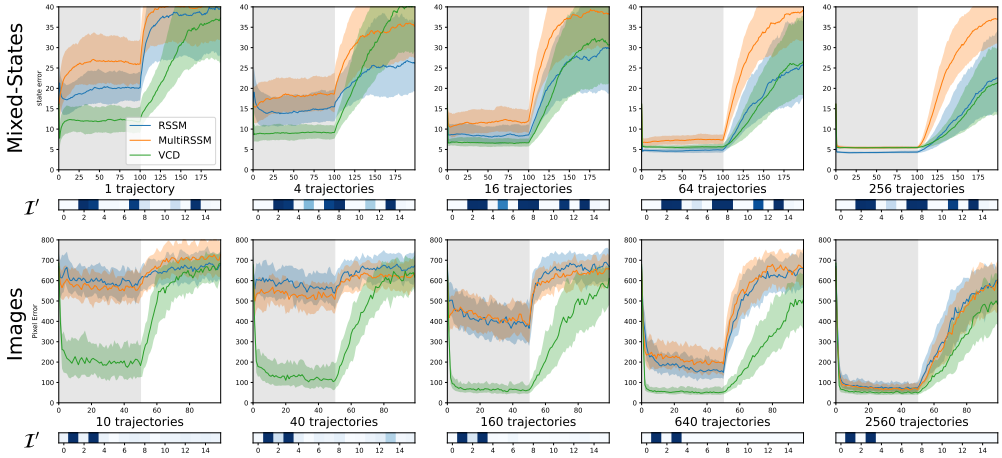


Figure A.5: Adaptation results for intervention 4.

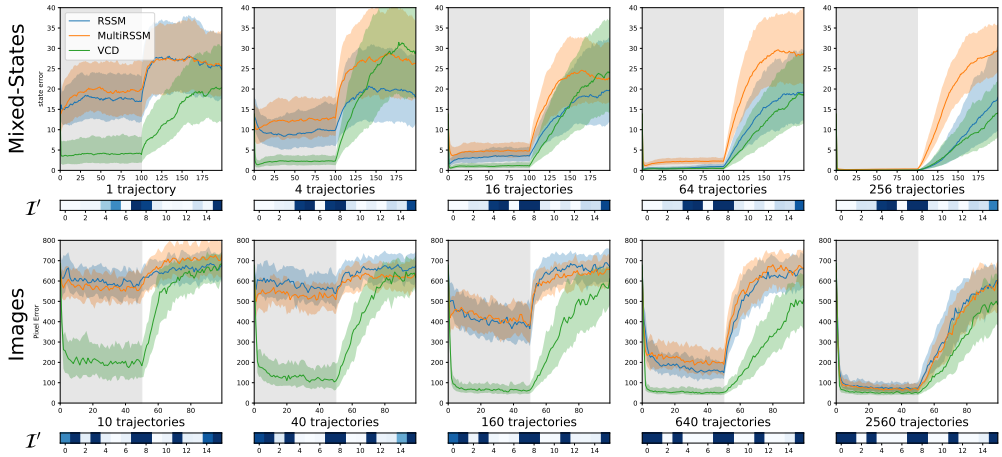


Figure A.6: Adaptation results for intervention 9.

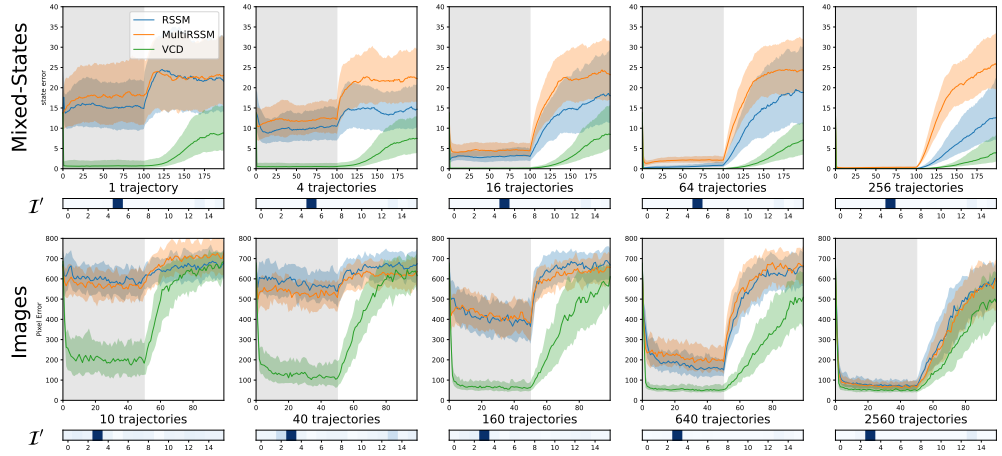


Figure A.7: Adaptation results for intervention 13.

B
Appendix for COMET

B.1 Implementation Details

Representation Model

In our experiments, all models are trained with image input in the *Particle Interactions* and *Traffic* domains. At each timestep, the observation is given as a set of image segmentations — one segmentation frame for each object in the scene. These are then transformed into object embeddings using a pre-trained CNN encoder. Each segmented frames are encoded separately, giving latent representations of each object. To allow for the capture of dynamic visual information, we concatenate two segmentation frames from consecutive time steps along the channel dimension. We pre-train the representation model using the C-SWM [97] framework, which uses uses a contrastive loss signal in latent space. The hyperparameters used during training of the encoder are equivalent to those reported in [97].

For the *Team Sports* domain, the ground-truth states of the players and the ball is used directly.

COMET Architecture

The architecture of COMET consists of two parts, the *mechanisms* and the *composition module*, as described in the main text. The first part, the set of

	Layer Type	in Features	out Features
1	Linear	2 * 8	300
2	ELU		
3	Linear	300	300
4	ELU		
5	Linear	300	8

Table B.1: Architecture of a single mechanism. The mechanism operates on the concatenation of two object embeddings and predicts state updates for the first input object.

	Layer Type	in Features	out Features
1	Linear	2 * 8	300
2	ELU		
3	Linear	300	300
4	ELU		
5	Linear	300	1

Table B.2: Architecture of a single confidence network. There is one confidence network per mechanism and it operates on the concatenation of two object embeddings to predict a score for the given context object and associated mechanism.

mechanisms $f_{mech}^{0:M}(\cdot)$, operates on the concatenated representations of two objects, the primary object and the context object. The mechanisms are implemented as feed-forward networks. The architecture of a single mechanism is shown in Table B.1. The composition module itself consists of a set of independent feed-forward confidence networks, one for each mechanism. Each such independent network receives the concatenated representations of two objects as input, similar to the mechanisms, and outputs a scalar confidence score. The architecture of a single confidence network is shown in Table B.2.

Contrastive Learning of Strutured World Models

Contrastive Learning of Strutured World Models (C-SWM) [97] is a state-of-the-art object-centric world model. At its core is a GNN dynamics model that takes object representations as its nodes and is assumed to be fully-connected. We use the official implementation of the model[1] but replace the default encoder with the encoder detailed above to ensure fairness of comparison. The default encoder of C-SWM

[1] `https://github.com/tkipf/c-swm`

decomposes a single RGB-frame into a set of objects whereas our encoder operates on a set of object segmentation frames. Other than that, we use the original training procedure and model architecture as presented in [97]. This means that the only difference between the C-SWM implementation and the COMET implementation in our experiments is the use of a GNN dynamics model.

Neural Production Systems

Neural Production Systems (NPS) [122] is an object-factorised world model that, like COMET, operates with a set of mechanisms instead of a monolithic dynamics model architecture. Furthermore, NPS also considers sparse contexts, i.e. an object's transition is conditioned on the state of at most one other object in the observation.

The most prominent difference between NPS and our approach is that NPS jointly learns the mechanism selection, as predicted by a dot-product attention layer, and the mechanisms themselves, whereas COMET decouples the mechanism learning and the selection of mechanisms. This limits NPS's capability of being trained on multiple environments, as different mechanisms may be active in different environments given the same state. In contrast, the competition of experts setup of COMET produces predictions for all mechanisms and updates the one with the lowest prediction error. Thus, the only requirement for COMET is that during the competition stage, enough mechanisms exist to capture all environment dynamics. The selection of mechanisms is separately trained during the environment-specific composition phase.

In our experiments, the NPS dynamics model replaces the GNN dynamics model of the C-SWM architecture, and uses the same representation model detailed above. Since no code is provided with the paper, we implemented our own version based on the detailed algorithms provided in the appendix of the original paper [122]. We verified the implementation by confirming that model performance is on par with the results presented in the original paper when evaluated on similar environments.

Hyperparameters and Compute Resources

During training for all of the models, the batch size is set to 1024 and we use the Adam [178] optimiser with a learning rate of 1e-4. Object embeddings are set to be 8-dimensional. All experiments were run on a cluster containing a mixture of Nvidia Quadro RTX 6000 and Nvidia Tesla V100 accelerators. Each training run involved a single GPU only. All training runs are finished within 24 hours.

B.2 Experiment Details

We evaluate COMET and the baselines on three problem domains, *Particle Interactions*, *Traffic* and *Team Sport*. In each of the domains, we define a set of environments with varying dynamics. Below we provide the details of the domains.

Particle Interactions

In the *Particle Interactions* domain, each environment is defined by a set of rules. Each of these rules consists of an interaction and a condition (i.e. when the interaction should happen). The particles interact in one of 5 modes of interactions: *straight line*, *repulsion*, *attraction*, *spring* and *spiral towards centre*. Moreover, each particle is coloured. We pick 6 randomly generated environments as the training set, summarised in Table B.3. The training set consists of 2000 sampled episodes from each of these environments, each with 50 frames.

For the adaptation experiments, COMET adapts to an unseen environment with 4 particles, with the *spring* interaction if the particles are of the same colour and *repel* otherwise.

Traffic

The *traffic* domain is simulated using the CARLA simulator [153].[2] Here, the objects are two cars, one orange and one blue, and a traffic light. In different environments, they follow different traffic rules. We train the models on three environments, 1. cars drive at normal speeds and stop at red traffic light, 2. cars drive under a much

[2]License information can be found on `https://github.com/carla-simulator/carla`.

	Condition	Interaction
Environment 1	close together	repulsion
	otherwise	straight line
Environment 2	same colour	spring
	close together	attraction
	otherwise	straight line
Environment 3	same colour	repulsion
	opposite colour	spring
	otherwise	straight line
Environment 4	same colour	attraction
	is blue	spiral towards centre
	is red	spiral towards centre
	otherwise	straight line
Environment 5	same colour	repulsion
	otherwise	spring
Environment 6	always	spiral towards centre

Table B.3: Environment details for particle interactions domain.

slower speed limit and stop at red traffic light, and 3. cars ignore traffic lights. In the adaptation environment, only the orange car drives at a slow speed limit. As such, the model needs to explain new behaviours such as the blue car slowing down to avoid collision with the orange car in front. The training set has 2000 episodes from each environment. The length of each episode depends on the exact scenario, which ends if both cars are out of frame or have stopped moving over 7 timesteps.

Team Sport

The *Team Sport* domain is simulated using a modified version of STS2 [154][3] As the domain comes with only a single player type, we define additional player types according to heuristics:

Simple Player. If the simple player is in control of the ball it moves towards the opponent goal. If other team players are closer to the opponent goal it tries to pass the ball to them. If a simple player is in the vicinity of the opponent goal and in control of the ball, it tries to score a goal. If a different player of the own team is in control of the ball, the simple player moves towards the opponent goal. If the ball is in control of the opponent team, the simple player moves towards the midway point

[3] `https://github.com/electronicarts/SimpleTeamSportsSimulator`.

between the own goal and the control player to block any attempted shot. If the ball is in the air due to a pass in progress, the simple player will move towards the ball

Defensive Player. The defensive player inherits all offensive behaviour from the simple player. If the defensive player is not in control of the ball, it moves towards the own goal to diffuse any counter attacks.

Shy Player. The shy player inherits all behaviour from the defensive player but instead of moving towards the opponent goal if in control of the ball, it moves away from the opponent goal.

For all player types, we define a specific course of action as being a single mechanism. Player types may share some mechanisms (e.g. moving towards the opponent goal) but differ in others. After executing certain actions such as shooting, passing, receiving, or tackling, players are unable to perform another action within a set number of time steps. This accounts for human players also requiring a certain reaction and orientation time after executing such actions.

We defined separate mechanisms for the ball which is modeled as a separate object. If in possession, the ball traces the control player and if in the air as a result of a pass or shot, the ball follows a straight line at constant speed until it is either caught by a player or bounces off a wall to change its direction. We collect a total of 5,957 episodes over four different team configurations for training and evaluate our model and baselines on a configuration of two simple players in each team, a setting which was not part of the training configurations. The state space consists of three one-hot flags indicating home team, away team, or ball. Further, we include two one-hot variables indicating which team is currently in possession of the ball (if any), a one-hot variable indicating whether the player itself is in control of the ball, a countdown timer denoting for how many time steps the player is unable to execute an action after having executed an action shortly before, as well as position and velocity information. Dimensions that apply only to players but not to the ball (e.g. the countdown timer) are present in the ball state but set to zero.

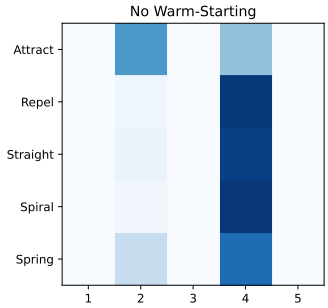


Figure B.1: Disentanglement plot for training without warm-starting, showing that two mechanisms dominate the competition phase.

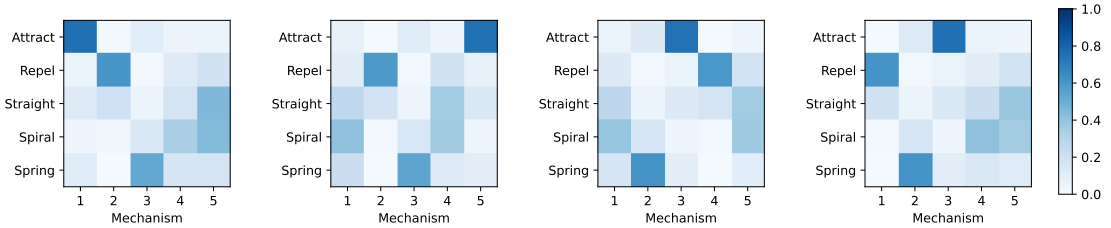


Figure B.2: Disentanglement plots for different random seeds when training with warm-starting. We observe that similar structures emerge across different seeds (up to permutation).

B.3 Ablation Studies

In Sec. 4.1.2, we describe two strategies that stablise the competition phase of COMET: 1. including a warm-starting phase and 2. increasing the effective time-horizon for mechanisms. Here, we present ablation results that demonstrate the effectiveness of these strategies.

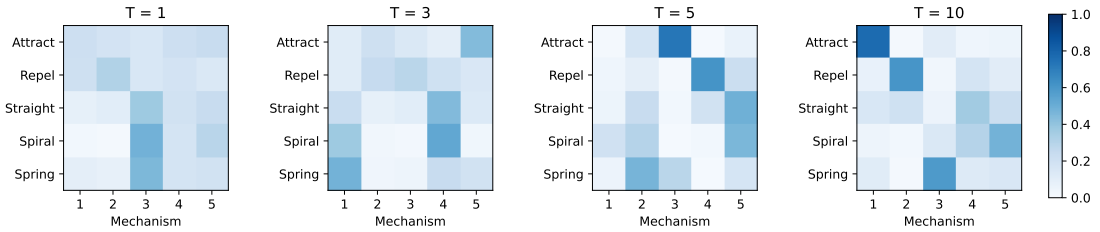


Figure B.3: Effect of increasing the effective time horizon in the competition phase. As the time horizon increases, the quality of disentanglement improves.

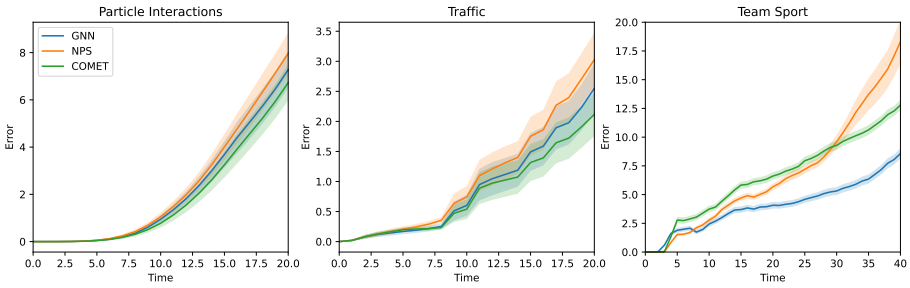


Figure B.4: Given enough data, all model achieves commensurate rollout errors in the Particle Interactions and Traffic dataset. In the Team Sports dataset, COMET performs similarly to NPS, which also considers binary interactions, whereas GNN performs better. This suggests that only considering binary interactions is not sufficient in this environment. Despite this, COMET is still able to disentangle some interaction mechanisms.

Warm-starting. One main failure mode is that due to random initialisation, some mechanisms may be significantly better than others at the start of training. This leads to a situation where the strong mechanisms dominate the competition and the weaker ones are never updated. Fig. B.1 shows an example of this failure mode where two mechanisms dominate the competition. In Fig. B.2, we show that with warm-starting, the model is able to use all mechanisms across multiple random seed.

Increasing Effective Time-Horizon. Another failure mode is that the model fails to learn meaningful mechanisms which leads to high-frequency swithing between mechanisms. We found that increasing the effective time-horizon for mechanisms alleviates this issue. Fig. B.3 shows that as the effective time-horizon increases, the quality of disentanglement improves, demonstrating the efficacy of this strategy.

B.4 Large Data Limit Behaviour

Fig. B.4 shows rollout plots for all of the models in a new environment in the large data limit. This shows that all model eventually converge to the same level of performance given enough data. Note that in the Team Sport environment is designed to be challenging for COMET as it requires reasoning about the entire scene, i.e. high order interactions between multiple objects. Here, GNN outperforms both NPS and COMET, which perform at a similar level. This supports our

hypothesis that considering binary interactions can hinder model expressiveness. Nonetheless, we show that COMET is still able to disentangle mechanisms and adapt much faster than NPS, which also only considers binary interactions.

C

Appendix for SPARTAN

C.1 Implementation Details

SPARTAN and Transformer baseline are implemented as stacked transformer encoder layers with residual connections. The Global Graph baseline is implemented as an ensemble of MLPs with learnable adjacency matrix masks. This is similar to [179] and VCD, with the modification that our baseline work on object embeddings rather than scalar states. Each MLP predicts a separate object embedding. For each MLP, the inputs object tokens are masked according to a learnable adjacency matrix before passing through the model. A sparse regularisation is applied to the adjacency matrix. The hyperparameters for SPARTAN and the baselines are shown in table C.1 and C.2. For the experiments on the simulated datasets, all models are trained on single GPUs (mixture of Nvidia V100 and RTX 6000) and converge within 3 days.

For the Traffic domain, predictions need to be multimodal since each vehicle has multiple possible trajectories. This is common for the dataset we consider. To overcome this, we use MTR [160] as the base architecture. MTR has two stages: first, it uses self-attention between tokenised map lines and vehicle trajectories; second, it applies cross attention using multiple queries, each representing one possible mode of motion. The output for each query is then used as a Gaussian mixture to predict multimodal motion patterns. We adapt SPARTAN to this

Table C.1: Hyperparameters for SPARTAN and the Transformer Baseline.

Hyperparameter	Interventional Pong	CREATE
Token Dimension	32	64
Embedding Dimension	512	512
n. transformer layers	3	3
MLP hidden dimension	512	1024
n. MLP layers per transformer layer	3	3
lr	5e-5	5e-5
Optimiser	Adam[178]	Adam

Table C.2: Hyperparameters for the Global Graph Baseline.

Hyperparameter	Interventional Pong	CREATE
Token Dimension	32	64
MLP hidden dimension	1024	1024
n. MLP layers	5	5
lr	5e-5	5e-5
Optimiser	Adam	Adam

architecture by swapping all attention layers with the sparsified version proposed in this paper. The original work also proposes a local attention masking scheme for each layer which only allow each token to attend to the k-nearest neighbour based on the position of each object. We refer to this as Local Attn. In the traffic domain, the models are trained in parallel on 4 GPUs due to the size of each scene (roughly 1000 tokens). Training takes less than one week for the baseline MTR model and under two weeks for SPARTAN.

C.2 Datasets

Interventional Pong

The *Interventional Pong* environment is based on the Pong game and is originally developed in [110] for investigating causal representation learning. In the original work, the set of interventions included randomly perturbing the positions of the balls and the paddles. We modify the interventions to model dynamics changes,

Table C.3: Environment configurations for Intervention Pong.

	ID	Description
Seen	0	No intervention
	1	Ball follows curved trajectories
	2	Left paddle moves away from ball instead of towards
	3	Right paddle mirrored along y-axis
	4	Strong friction on the ball in the middle section
	5	Changed bouncing dynamics between paddles and ball
	6	Add gravity on the left side
Unseen	7	Env 5 + Env 6
	8	Env 4 + Env 5
	9	Env 1 + Env 2
	10	Env 3 + Env 6

Table C.4: Environment configurations for CREATE

	ID	Objects	Description
Seen	0	3 balls, wall, ladder, cannon	No intervention
	1		Disabled gravity
	2		Balls go *down* ladders
	3		Increased elasticity of walls
	4		Disabled cannon
Unseen	5	3 balls, wall, 3*Ladders	No interventions
	6	3 balls, wall, 3*Ladders	Balls go *down* ladders
	7	3 balls, 2*Walls, 2*Cannons	Increased elasticity

such as adding gravity or changing the way the ball bounces off the paddles, as these are more appropriate in the context of world model learning. A similar setup is also used in [179] in the context of few-shot adaptation. The original Interventional Pong dataset is under the BSD 3-Clause Clear License.

The environment contains 4 objects, the left paddle, the right paddle, the ball and the score. Here, local causal edges include ball $\rightarrow$ left, right paddles as the paddles follow the ball; paddles $\rightarrow$ ball when the ball collides with the paddles; ball $\rightarrow$ score when the ball crosses the left or right boundary lines. Observations are provided as $32 \times 32 \times 3$ RGB images with masks for each object. To capture velocity information for ball, the ball image shows the position of the ball over 2

consecutive timesteps in different colours. Each object is separately encoded into 32-dimensional latent object-slots using a VAE [44]. The configurations of different intervened environments are shown in table C.3.

CREATE

The *CREATE* environment is built using the CREATE simulator [158] which consists of physical interactions between various objects. The simulator is under the MIT License. In our experiments, the environment objects include {*blue*, *red*, *green*} balls, wall, ladder and cannon. The balls collides elastically. All objects are initialised at random positions with random orientations. The wall remains static for each sequence. The ladder object allows the balls to climb up. The cannon object projects the balls at a fixed velocity in the direction it points at. Here, local edges captures when the objects collide or interact with each other. Observations are given as $84 \times 84 \times 3$ images with masks for each object. To capture the velocities of the balls, for each ball, a trail of the past positions (up to two timesteps) is shown in the images. The intervened environments are shown in table C.4. In the unseen environments, we test the models using different scene composition, i.e. different numbers of objects.

Traffic

We use the Waymo Open Dataset [155] for our Traffic domain. The dataset consists of real world scenes which include map lines (such as lane markings) and vehicle trajectories. The map lines are provided as polylines (i.e. each line is represented as a set of positions). The history of the state of each vehicle is provided as a list of positions, headings and velocities. Up to one second of state is given for each vehicle and the task is to predict 9 seconds into the future for the centre vehicle. The number of map lines and vehicle trajectories are different for each scene, with roughly 800 map lines and 50 vehicles per scene. This presents a major challenge for prior approaches in causal world model learning. Since the dataset is not generated by simulation, we do not apply interventions to the data.

Table C.5: Results using SlotAttention object representations on Interventional Pong.

	SPARTAN	Transformer
Prediction Error	7.21 ± 0.69	6.90 ± 0.83
SHD	2.69 ± 0.02	8.52 ± 0.02

C.3 Further Experimental Results

Many existing works, such as SlotFormer [98], use pre-trained object-centric representations such as SlotAttention [90] which handles the segmentation and tracking problems. Our approach can be readily applied to any such pretrained representations. In our experiments, we opted for using ground-truth masked object representations for the principal reason that it enables evaluation of the learned graphs: quantitatively evaluating the SHD of learned graphs requires mapping each latent embedding to the corresponding ground-truth object. Learned object embeddings can make this mapping noisy and ambiguous due to imperfect segmentations. Since learning high-quality representation is not the main focus of our investigation, we use ground-truth masked object representations to ensure the fairness and validity of our evaluation. To demonstrate the generality of our results, we have trained our model on an unsupervised learning slot representation for the interventional pong environment. The results (table C.5) are consistent with the main results in the paper: our model can achieve the same level of prediction while capturing the underlying graph structure more accurately.

D

Appendix for Interpretability Experiments

D.1 Details for Direct Advantage Estimation

Environment Setting and Training Procedure

We use a simplified version of the CoinRun environment by enabling the velocity information via `paint_vel_info`, furthermore, we constrain the action space to actions that are relevant to CoinRun (i.e., the 9 actions that control the character instead of the full 15 actions). These changes were made to reduce partial observability and ease exploration. To further reduce partial observability, we also stack the last 4 observations.

For the exogenous reward experiments, the exogenous reward is defined as a function of the background index (ranging from 0 to 61). More specifically, we define $r_x = 0.1 \times (\frac{\texttt{bkg_idx}}{32} - 1)$. To simulate the effect of infinite horizons, the value of the terminal state is defined by $\frac{r_x}{1-\gamma}$ instead of 0.

We use softmax policies for exploration and smoothing the policy changes, which were found to be crucial for the DAE objective [180]. The parameter τ is the log of the temperature for the softmax function, which is minimised along with a KL penalty to ensure the policy does not change too much between iterations.

Hyperparameter & Network Architecture

We follow the original ViT block design, which consists of a multi-head self-attention layer followed by an MLP block. The patchify the image using a three layer 2D convolutional network as suggested by [167]. We use LayerScale [181] in the ViT blocks, which we found to stabilise training at initialisation. The learning rate follows a one cycle schedule with linear warmup for the first 10% gradient steps followed by linear annealing to 0. Value heads are all MLPs with a single hidden layer.

ViT	
Conv. Stem (channels, kernel size, stride)	[256, 4, 4] [512, 2, 2] [512, 1, 1]
ViT blocks	2
Attention heads	4
MLP dimension	2048
Other Hyperparameters	
Parallel actors	8192
Backup length	64
Epochs	2
Batch size	32
Learning rate	1.25×10^{-4}
Learning rate (τ)	10^{-2}
Adam ($\beta_1, \beta_2, \epsilon$)	(0.9, 0.95, 10^{-5})
β_{KL}	5
β_{S}	Linearly increased from 0 to 0.1

Table D.1: Network Architecture & Hyperparameters

D.2 Details for Mechanistic Interpretability

Indirect Object Identification Task

In the experiment, we curate a dataset of prompts for the Indirect Object Identification (IOI) tasks. These prompts are generated by filling in templates with a set of name pairs. Table D.2 shows the list of templates used to generate the Indirect Object Identification (IOI) tasks. The set of name pairs used are [(Mary, John),

(Alice, Bob), (Michael, David), (Daniel, Sarah), (Emily, Jacob), (Laura, Kevin)]. In total, there are 6 name pairs and 15 templates, resulting in 90 unique IOI prompts.

Ablation Procedure

To perform ablation on the attention heads, we can either set the activation to zero or to a mean value computed from a control dataset. It is found that using the mean value leads to more stable results [170]. Following this, we create a control dataset by generating prompts using the same templates as in Table D.2, but where the subject name [S] is sampled from a random name instead of from [A] or [B]. The idea is that in this control dataset, the subject name [S] does not have a direct relationship with the objects [A] and [B], and therefore serves as a neutral baseline of preference between the two. The set of random names used are [Amy, Ryan, Nina, Carl, Beth, Zoe, Christopher, Jessica, Matthew, Ashley, Jennifer, Joshua, Amanda, James, Joseph, Andrew, Ryan].

Compute Resources

For post-training GPT2, the model is fine-tuned on a single NVIDIA H100 GPU, taking approximately 4 days to fully converge. For LlaMa, the model is fine-tuned on 4 NVIDIA H100 GPUs, taking approximately 7 days to fully converge.

Templates for IOI tasks
Then, [B] and [A] went to the park. [B] gave a ball to
Then, [B] and [A] had a lot of fun at the park. [S] gave a toy to
Then, [B] and [A] were working at the office. [S] decided to give a pen to
Then, [B] and [A] were thinking about going to the beach. [S] gave a shell to
Then, [B] and [A] had a long argument, and afterwards [S] said to
After [B] and [A] went to the mall, [S] gave a gift to
When [B] and [A] got a snack at the cafe, [S] decided to give it to
When [B] and [A] got a toy at the park, [S] decided to give the toy to
While [B] and [A] were working at the office, [S] gave a pen to
While [B] and [A] were at the beach, [S] gave a shell to
After the lunch, [B] and [A] went to the mall. [S] gave a gift to
Afterwards, [B] and [A] went to the park. [S] gave a ball to
Then, [B] and [A] had a long argument. Afterwards [S] said to
The park [B] and [A] went to had a ball. [S] gave it to
Friends [B] and [A] found a toy at the park. [S] gave it to

Table D.2: A list of templates used to generate the Indirect Object Identification tasks. [A] and [B] are randomly sampled name pairs. [S] is the subject name, which is sampled from either [A] or [B] for the IOI task, and from a random name for the control task.

References

[1] D. Ha and J. Schmidhuber. "Recurrent World Models Facilitate Policy Evolution". In: *Advances in Neural Information Processing Systems 31*. https://worldmodels.github.io. Curran Associates, Inc., 2018, pp. 2451–2463. URL: https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution.

[2] Nicklas Hansen, Hao Su, and Xiaolong Wang. "TD-MPC2: Scalable, Robust World Models for Continuous Control". In: *International Conference on Learning Representations (ICLR)*. 2024.

[3] Danijar Hafner et al. "Mastering diverse control tasks through world models". In: *Nature* (2025), pp. 1–7.

[4] Vincent Micheli, Eloi Alonso, and François Fleuret. "Transformers are Sample-Efficient World Models". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=vhFu1Acb0xb.

[5] Philipp Wu et al. "DayDreamer: World Models for Physical Robot Learning". In: *Proceedings of The 6th Conference on Robot Learning*. Ed. by Karen Liu, Dana Kulic, and Jeff Ichnowski. Vol. 205. Proceedings of Machine Learning Research. PMLR, 2023, pp. 2226–2240. URL: https://proceedings.mlr.press/v205/wu23c.html.

[6] Aaron Van Den Oord, Oriol Vinyals, et al. "Neural discrete representation learning". In: *Advances in neural information processing systems* 30 (2017).

[7] Junyu Chen et al. "Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models". In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=wH8XXUOUZU.

[8] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[9] Albert Gu and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *First Conference on Language Modeling*. 2024. URL: https://openreview.net/forum?id=tEYskw1VY2.

[10] Lloyd Russell et al. *GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving*. 2025. arXiv: 2503.20523 [cs.CV]. URL: https://arxiv.org/abs/2503.20523.

[11] Jack Parker-Holder et al. "Genie 2: A Large-Scale Foundation World Model". In: (2024). URL: https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

[12] Philip J. Ball et al. "Genie 3: A New Frontier for World Models". In: (2025).

[13] NVIDIA et al. *Cosmos World Foundation Model Platform for Physical AI*. 2025. arXiv: 2501.03575 [cs.CV]. URL: https://arxiv.org/abs/2501.03575.

[14] Andy Clark. "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and brain sciences* 36.3 (2013), pp. 181–204.

[15] Karl Friston et al. "The free energy principle made simpler but not too simple". In: *Physics Reports* 1024 (2023). The free energy principle made simpler but not too simple, pp. 1–29. URL: https://www.sciencedirect.com/science/article/pii/S037015732300203X.

[16] T. Behrens et al. "What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior". In: *Neuron* 100.2 (2018), pp. 490–509. URL: https://www.sciencedirect.com/science/article/pii/S0896627318308560.

[17] Rebecca Roelofs et al. *CausalAgents: A Robustness Benchmark for Motion Forecasting using Causal Relationships*. 2022. arXiv: 2207.03586 [cs.LG]. URL: https://arxiv.org/abs/2207.03586.

[18] Bernhard Schölkopf et al. "On Causal and Anticausal Learning". In: *29th International Conference on Machine Learning (ICML 2012)*. International Conference on Machine Learning. 2012.

[19] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

[20] Bernhard Schölkopf et al. "Toward Causal Representation Learning". In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[21] Yoshua Bengio et al. "A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=ryxWIgBFPS.

[22] Philippe Brouillard et al. "Differentiable causal discovery from interventional data". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21865–21877.

[23] Zizhao Wang et al. "ELDEN: Exploration via Local Dependencies". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 15456–15474. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/31ed129feae64a7e44a15b148c15558d-Paper-Conference.pdf.

[24] Inwoo Hwang et al. "Fine-Grained Causal Dynamics Learning with Quantization for Improving Robustness in Reinforcement Learning". In: *Proceedings of the 41th International Conference on Machine Learning*. 2024.

[25] Silviu Pitis, Elliot Creager, and Animesh Garg. "Counterfactual Data Augmentation using Locally Factored Dynamics". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 3976–3990. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/294e09f267683c7ddc6cc5134a7e68a8-Paper.pdf.

[26] Hsiao-Ru Pan et al. "Direct Advantage Estimation". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=oNnv9XjClGK.

[27] Anson Lei, Bernhard Schölkopf, and Ingmar Posner. "Variational Causal Dynamics: Discovering Modular World Models from Interventions". In: *Transactions on Machine Learning Research* (2023). URL: https://openreview.net/forum?id=V9tQKYYNK1.

[28] Anson Lei, Bernhard Schölkopf, and Ingmar Posner. "Causal Discovery for Modular World Models". In: *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*. 2022. URL: https://openreview.net/forum?id=VfkjQzdGCH.

[29] Anson Lei et al. "Modular World Models with Competitive Independent Mechanisms". In: *Causal and Object-Centric Representations for Robotics Workshop at CVPR 2024*. 2024. URL: https://openreview.net/forum?id=8exP12DDJo.

[30] Anson Lei et al. *Compete and Compose: Learning Independent Mechanisms for Modular World Models*. 2024. arXiv: 2404.15109 [cs.LG]. URL: https://arxiv.org/abs/2404.15109.

[31] Anson Lei, Bernhard Schölkopf, and Ingmar Posner. "SPARTAN: A Sparse Transformer World Model Attending to What Matters". In: *Advances in Neural Information Processing Systems*. 2025.

[32] Hsiao-Ru Pan et al. "Learning What Matters: Endogenous Representation through Advantage Estimation and Sparse Attention". In: *Submitted to The Causal Reinforcement Learning Workshop, RLC 2025*. under review. 2025. URL: https://openreview.net/forum?id=lhoMGxMJcu.

[33] Basil Kouvaritakis and Mark Cannon. "Model predictive control". In: (2016).

[34] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[35] Christopher G Atkeson and Juan Carlos Santamaria. "A comparison of direct and model-based reinforcement learning". In: *Proceedings of international conference on robotics and automation*. Vol. 4. IEEE. 1997, pp. 3557–3564.

[36] Marc Peter Deisenroth and Carl Edward Rasmussen. "PILCO: a model-based and data-efficient approach to policy search". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 2011, pp. 465–472.

[37] Jakob Hohwy. *The Predictive Mind*. Oxford, GB: Oxford University Press, 2013.

[38] David C Knill and Alexandre Pouget. "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *TRENDS in Neurosciences* 27.12 (2004), pp. 712–719.

[39] Danijar Hafner et al. "Dream to Control: Learning Behaviors by Latent Imagination". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=S1lOTC4tDS.

[40] Danijar Hafner et al. "Mastering Atari with Discrete World Models". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=0oabwyZbOu.

[41] Jiazhi Yang et al. "ReSim: Reliable World Simulation for Autonomous Driving". In: *arXiv preprint arXiv:2506.09981* (2025).

[42] Jake Bruce et al. *Genie: Generative Interactive Environments*. 2024. arXiv: 2402.15391 [cs.LG]. URL: https://arxiv.org/abs/2402.15391.

[43] Sherry Yang et al. "Learning Interactive Real-World Simulators". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=sFyTZEqmUY.

[44] D. Kingma and M. Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR*. 2014. URL: http://arxiv.org/abs/1312.6114.

[45] Vincent Micheli, Eloi Alonso, and François Fleuret. "Efficient World Models with Context-Aware Tokenization". In: *Forty-first International Conference on Machine Learning*. 2024. URL: https://openreview.net/forum?id=BiWIERWBFX.

[46] Younggyo Seo et al. "Masked World Models for Visual Control". In: *6th Annual Conference on Robot Learning*. 2022. URL: https://openreview.net/forum?id=Bf6on28H0Jv.

[47] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[48] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[49] Gaoyue Zhou et al. "DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning". In: *Forty-second International Conference on Machine Learning*. 2025. URL: https://openreview.net/forum?id=D5RNACOZEI.

[50] Yuhang Huang et al. "LaDi-WM: A Latent Diffusion-based World Model for Predictive Manipulation". In: *arXiv preprint arXiv:2505.11528* (2025).

[51] Maxime Oquab et al. "DINOv2: Learning Robust Visual Features without Supervision". In: *Transactions on Machine Learning Research* (2024). Featured Certification. URL: https://openreview.net/forum?id=a68SUt6zFt.

[52] M. Watter et al. "Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper/2015/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

[53] Ershad Banijamali et al. "Robust locally-linear controllable embedding". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1751–1759.

[54] Danijar Hafner et al. "Learning Latent Dynamics for Planning from Pixels". In: *International Conference on Machine Learning*. 2019, pp. 2555–2565.

[55] Maxime Burchi and Radu Timofte. "Learning Transformer-based World Models with Contrastive Predictive Coding". In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=YK9G4Htdew.

[56] Tung D Nguyen et al. "Temporal predictive coding for model-based planning in latent space". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8130–8139.

[57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[58] Amy Zhang et al. "Learning Invariant Representations for Reinforcement Learning without Reconstruction". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=-2FCwDKRREu.

[59] Ruixiang Sun et al. "Learning latent dynamic robust representations for world models". In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 47234–47260.

[60] Fei Deng, Ingook Jang, and Sungjin Ahn. "DreamerPro: Reconstruction-Free Model-Based Reinforcement Learning with Prototypical Representations". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 4956–4975. URL: https://proceedings.mlr.press/v162/deng22a.html.

[61] Hsiao-Ru Pan and Bernhard Schölkopf. "Learning endogenous representation in reinforcement learning via advantage estimation". In: *Causal Representation Learning Workshop at NeurIPS 2023*. 2023.

[62] M. Zhang et al. "Solar: Deep structured representations for model-based reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 7444–7453.

[63] Jan Robine et al. "Transformer-based World Models Are Happy With 100k Interactions". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=TdBaDGCpjly.

[64] Weipu Zhang et al. "STORM: Efficient Stochastic Transformer based World Models for Reinforcement Learning". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=WxnrX42rnS.

[65] Marc Rigter, Jun Yamada, and Ingmar Posner. "World Models via Policy-Guided Trajectory Diffusion". In: *Transactions on Machine Learning Research* (2024). URL: https://openreview.net/forum?id=9CcgOOLhKG.

[66] Fangqi Zhu et al. "Irasim: Learning interactive real-robot action simulators". In: *arXiv preprint arXiv:2406.14540* (2024).

[67] Reuven Rubinstein. "The cross-entropy method for combinatorial and continuous optimization". In: *Methodology and computing in applied probability* 1.2 (1999), pp. 127–190.

[68] Kurtland Chua et al. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models". In: *Advances in neural information processing systems* 31 (2018).

[69] Jyothir SV et al. "Gradient-based Planning with World Models". In: *arXiv preprint arXiv:2312.17227* (2023).

[70] Jun Yamada et al. "D-Cubed: Latent DiffusioN Trajectory Optimisation for Dexterous Deformable Manipulation". In: *arXiv preprint arXiv:2403.12861* (2024).

[71] Sergey Levine. "Reinforcement learning and control as probabilistic inference: Tutorial and review". In: *arXiv preprint arXiv:1805.00909* (2018).

[72] Tuomas Haarnoja et al. "Soft actor-critic algorithms and applications". In: *arXiv preprint arXiv:1812.05905* (2018).

[73] Nicklas A Hansen, Hao Su, and Xiaolong Wang. "Temporal Difference Learning for Model Predictive Control". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 8387–8406.

[74] Ramanan Sekar et al. "Planning to explore via self-supervised world models". In: *International conference on machine learning*. PMLR. 2020, pp. 8583–8592.

[75] Cansu Sancaktar, Sebastian Blaes, and Georg Martius. "Curious Exploration via Structured World Models Yields Zero-Shot Object Manipulation". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=NnuYZ1el24C.

[76] Jonathan Richens and Tom Everitt. "Robust agents learn causal world models". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=pOoKI3ouv1.

[77] J. Pearl. *Causality: Models, Reasoning, and Inference*. 2nd. New York, NY: Cambridge University Press, 2009.

[78] J. Peters, P. Bühlmann, and N. Meinshausen. "Causal inference using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society, Series B: Statistical Methodology (with discussion)* 78.5 (2016), pp. 947–1012.

[79] K. Zhang et al. "Causal Discovery from Nonstationary/Heterogeneous Data: Skeleton Estimation and Orientation Determination". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*. 2017, pp. 1347–1353.

[80] Chaochao Lu et al. "Sample-efficient reinforcement learning via counterfactual-based data augmentation". In: *arXiv preprint arXiv:2012.09092* (2020).

[81] Thomas Mesnard et al. "Counterfactual Credit Assignment in Model-Free Reinforcement Learning". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 7654–7664.

[82] Yoshua Bengio. "Deep learning of representations: Looking forward". In: *International conference on statistical language and speech processing*. Springer. 2013, pp. 1–37.

[83] Brenden M. Lake et al. "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40 (2017), e253.

[84] Hanlin Zhang et al. "Towards Principled Disentanglement for Domain Generalization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 8024–8034.

[85] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy2fzU9gl.

[86] Hyunjik Kim and Andriy Mnih. "Disentangling by factorising". In: *International conference on machine learning*. PMLR. 2018, pp. 2649–2658.

[87] Christopher P. Burgess et al. "MONet: Unsupervised Scene Decomposition and Representation". In: *arXiv preprint arXiv:1901.11390* (2019). (Visited on 11/10/2022).

[88] M. Engelcke, O. Parker Jones, and I. Posner. "GENESIS-V2: Inferring Unordered Object Representations without Iterative Refinement". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 8085–8094. URL: https://proceedings.neurips.cc/paper/2021/file/43ec517d68b6edd3015b3edc9a11367b-Paper.pdf.

[89] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. "Multi-Object Representation Learning with Iterative Variational Inference". In: *International Conference on Machine Learning (ICML)*. 2019. arXiv: 1903.00450.

[90] Francesco Locatello et al. "Object-Centric Learning with Slot Attention". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 11525–11538. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/8511df98c02ab60aea1b2356c013bc0f-Paper.pdf.

[91] Thomas Kipf et al. "Conditional Object-Centric Learning from Video". In: *International Conference on Learning Representations*. 2022.

[92] Mehdi SM Sajjadi et al. "Object scene representation transformer". In: *Advances in neural information processing systems* 35 (2022), pp. 9512–9524.

[93] Sindy Löwe et al. "Learning object-centric video models by contrasting sets". In: *arXiv preprint arXiv:2011.10287* (2020).

[94] Jindong Jiang et al. "SCALOR: Generative World Models with Scalable Object Representations". In: *International Conference on Learning Representations*. 2020. (Visited on 11/10/2022).

[95] Y. Wu et al. "APEX: Unsupervised, object-centric scene segmentation and tracking for robot manipulation". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2021.

[96] Zhixuan Lin et al. "Improving Generative Imagination in Object-Centric World Models". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 6140–6149. URL: https://proceedings.mlr.press/v119/lin20f.html.

[97] Thomas Kipf, Elise van der Pol, and Max Welling. "Contrastive Learning of Structured World Models". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1gax6VtDB.

[98] Ziyi Wu et al. "SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=TFbwV6I0VLg.

[99] Malte Mosbach et al. "SOLD: Slot Object-Centric Latent Dynamics Models for Relational Manipulation Learning from Pixels". In: *Forty-second International Conference on Machine Learning*. 2025. URL: https://openreview.net/forum?id=XOUpHJPYRX.

[100] Rishi Veerapaneni et al. "Entity Abstraction in Visual Model-Based Reinforcement Learning". In: *Proceedings of the Conference on Robot Learning*. Ed. by Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura. Vol. 100. Proceedings of Machine Learning Research. PMLR, Nov. 2020, pp. 1439–1456. URL: https://proceedings.mlr.press/v100/veerapaneni20a.html.

[101] Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. "Identifiability of latent-variable and structural-equation models: from linear to nonlinear". In: *Annals of the Institute of Statistical Mathematics* 76.1 (2024), pp. 1–33.

[102] Francesco Locatello et al. "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[103] Sebastien Lachapelle et al. "Additive Decoders for Latent Variables Identification and Cartesian-Product Extrapolation". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=R6KJN1AUAR.

[104] Jack Brady et al. "Provably learning object-centric representations". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 3038–3062.

[105] Jack Brady et al. "Interaction asymmetry: A general principle for learning composable abstractions". In: *arXiv preprint arXiv:2411.07784* (2024).

[106] William Peebles et al. "The hessian penalty: A weak prior for unsupervised disentanglement". In: *European conference on computer vision*. Springer. 2020, pp. 581–597.

[107] Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. "Weakly supervised representation learning with sparse perturbations". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15516–15528.

[108] Johann Brehmer et al. "Weakly supervised causal representation learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 38319–38331.

[109] Francesco Locatello et al. "Weakly-supervised disentanglement without compromises". In: *International conference on machine learning*. PMLR. 2020, pp. 6348–6359.

[110] Phillip Lippe et al. "CITRIS: Causal Identifiability from Temporal Intervened Sequences". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 13557–13603. URL: https://proceedings.mlr.press/v162/lippe22a.html.

[111] Phillip Lippe et al. "Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=itZ6ggvMnzS.

[112] Phillip Lippe et al. "BISCUIT: Causal Representation Learning from Binary Interactions". In: *The 39th Conference on Uncertainty in Artificial Intelligence*. 2023. URL: https://openreview.net/forum?id=VS7Dn31xuB.

[113] Sébastien Lachapelle et al. "Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 18171–18206.

[114] Dingling Yao et al. "Multi-View Causal Representation Learning with Partial Observability". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=OGtnhKQJms.

[115] Sébastien Lachapelle et al. "Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA". In: *Conference on Causal Learning and Reasoning*. PMLR. 2022, pp. 428–484.

[116] Kartik Ahuja, Jason Hartford, and Yoshua Bengio. "Properties from mechanisms: an equivariance perspective on identifiable representation learning". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=g5ynW-jMq4M.

[117] Anirudh Goyal and Yoshua Bengio. "Inductive biases for deep learning of higher-level cognition". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 478.2266 (2022), p. 20210068.

[118] Philip Becker-Ehmck, Jan Peters, and Patrick Van Der Smagt. "Switching linear dynamics for variational bayes filtering". In: *International conference on machine learning*. PMLR. 2019, pp. 553–562.

[119] A. Goyal et al. "Recurrent Independent Mechanisms". In: *9th International Conference on Learning Representations (ICLR)*. May 2021. URL: https://openreview.net/pdf?id=mLcmdlEUxy-.

[120] Anirudh Goyal et al. "Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems". In: *arXiv preprint arXiv:2006.16225* (2020).

[121] Rim Assouel et al. "VIM: Variational Independent Modules for Video Prediction". In: *Proceedings of the First Conference on Causal Learning and Reasoning*. Ed. by Bernhard Schölkopf, Caroline Uhler, and Kun Zhang. Vol. 177. Proceedings of Machine Learning Research. PMLR, Apr. 2022, pp. 70–89. URL: https://proceedings.mlr.press/v177/assouel22a.html.

[122] Anirudh Goyal et al. "Neural production systems". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25673–25687.

[123] Michael I Jordan and Robert A Jacobs. "Hierarchical mixtures of experts and the EM algorithm". In: *Neural computation* 6.2 (1994), pp. 181–214.

[124] Noam Shazeer et al. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=B1ckMDqlg.

[125] Albert Q. Jiang et al. *Mixtral of Experts*. 2024. arXiv: 2401.04088 [cs.LG]. URL: https://arxiv.org/abs/2401.04088.

[126] Damai Dai et al. *DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models*. 2024. arXiv: 2401.06066 [cs.CL]. URL: https://arxiv.org/abs/2401.06066.

[127] William Fedus, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39.

[128] James Poetzscher, Saumya, et al. *Do Experts Specialize? A Mechanistic Exploration of Mixture of Experts Models*. Tech. rep. Mechanistic interpretability study of MoE specialization. Stanford University (CS231n), 2024. URL: https://cs231n.stanford.edu/2024/papers/do-experts-specialize-a-mechanistic-exploration-of-mixture-of-ex.pdf.

[129] Xingyi Yang et al. "Mixture of Experts Made Intrinsically Interpretable". In: *Forty-second International Conference on Machine Learning*. 2025. URL: https://openreview.net/forum?id=6QERrXMLP2.

[130] Giambattista Parascandolo et al. "Learning independent causal mechanisms". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4036–4044.

[131] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. "Expert gate: Lifelong learning with a network of experts". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3366–3375.

[132] Oleksiy Ostapenko et al. "Continual Learning via Local Module Composition". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=LJjC6DmSkgT.

[133] Francesco Locatello et al. "Competitive training of mixtures of independent deep generative models". In: *arXiv preprint arXiv:1804.11130* (2018).

[134] J. von Kügelgen et al. "Towards causal generative scene models via competition of experts". In: *arXiv preprint arXiv:2004.12906* (2020).

[135] Andrea Tacchetti et al. "Relational Forward Models for Multi-Agent Learning". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=rJlEojAqFm.

[136] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. 2nd. Cambridge, MA: MIT Press, 2000.

[137] A. Hauser and P. Bühlmann. "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs". In: *Journal of Machine Learning Research* 13 (2012), pp. 2409–2464.

[138] Nan Rosemary Ke et al. "Learning neural causal models from unknown interventions". In: *arXiv preprint arXiv:1910.01075* (2019).

[139] Phillip Lippe, Taco Cohen, and Efstratios Gavves. "Efficient Neural Causal Discovery without Acyclicity Constraints". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=eYciPrLuUhG.

[140] Silviu Pitis et al. "MoCoDA: Model-based Counterfactual Data Augmentation". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=w6tBOjPCrIO.

[141] Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. "Causal influence detection for improving efficiency in reinforcement learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22905–22918.

[142] Peter D Grünwald. *The minimum description length principle*. 2007.

[143] Liang Wendong, Simon Buchholz, and Bernhard Schölkopf. "Algorithmic causal structure emerging through compression". In: *Causal Learning and Reasoning*. PMLR. 2025, pp. 201–242.

[144] Dominik Janzing and Bernhard Schölkopf. "Causal inference using the algorithmic Markov condition". In: *IEEE Transactions on Information Theory* 56.10 (2010), pp. 5168–5194.

[145] Andrew Gordon Wilson. "Deep Learning is Not So Mysterious or Different". In: *Forty-second International Conference on Machine Learning Position Paper Track*. 2025. URL: https://openreview.net/forum?id=42Au7FoD8F.

[146] Yedi Zhang, Andrew Saxe, and Peter E Latham. "Saddle-to-Saddle Dynamics Explains A Simplicity Bias Across Neural Network Architectures". In: *arXiv preprint arXiv:2512.20607* (2025).

[147] Andrew Saxe, Shagun Sodhani, and Sam Jay Lewallen. "The neural race reduction: Dynamics of abstraction in gated networks". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 19287–19309.

[148] Lars Buesing et al. "Learning and querying fast generative models for reinforcement learning". In: *arXiv preprint arXiv:1802.03006* (2018).

[149] K. Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014.

[150] Phillip Lippe, Taco Cohen, and Efstratios Gavves. "Efficient Neural Causal Discovery without Acyclicity Constraints". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=eYciPrLuUhG.

[151] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical Reparameterization with Gumbel-Softmax". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=rkE3y85ee.

[152] Elizabeth S Spelke and Katherine D Kinzler. "Core knowledge". In: *Developmental science* 10.1 (2007), pp. 89–96.

[153] Alexey Dosovitskiy et al. "CARLA: An Open Urban Driving Simulator". In: *Proceedings of the 1st Annual Conference on Robot Learning*. 2017, pp. 1–16.

[154] Caedmon Somers et al. *Simple Team Sports Simulator (STS2)*. Version 1.0.0. Feb. 2020.

[155] Pei Sun et al. "Scalability in Perception for Autonomous Driving: Waymo Open Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[156] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[157] Danilo Jimenez Rezende and Fabio Viola. *Taming VAEs*. 2018. arXiv: 1810.00597 [stat.ML].

[158] Ayush Jain, Andrew Szot, and Joseph Lim. "Generalization to New Actions in Reinforcement Learning". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4661–4672. URL: https://proceedings.mlr.press/v119/jain20b.html.

[159] Anton Bakhtin et al. "PHYRE: A New Benchmark for Physical Reasoning". In: *arXiv:1908.05656* (2019).

[160] Shaoshuai Shi et al. "Motion transformer with global intention localization and local movement refinement". In: *Advances in Neural Information Processing Systems* (2022).

[161] Xue Bin Peng et al. "Sim-to-real transfer of robotic control with dynamics randomization". In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2018, pp. 3803–3810.

[162] Josh Tobin et al. "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017, pp. 23–30.

[163] Yan Zeng et al. "A survey on causal reinforcement learning". In: *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[164] George Trimponias and Thomas G Dietterich. "Reinforcement Learning with Exogenous States and Rewards". In: *arXiv preprint arXiv:2303.12957* (2023).

[165] Karl Cobbe et al. "Leveraging procedural generation to benchmark reinforcement learning". In: *International conference on machine learning*. PMLR. 2020, pp. 2048–2056.

[166] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. Vol. 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[167] Tete Xiao et al. "Early convolutions help transformers see better". In: *Advances in neural information processing systems* 34 (2021), pp. 30392–30400.

[168] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[169] Jiaming Ji et al. "AI Alignment: A Comprehensive Survey". In: *CoRR* abs/2310.19852 (2023). URL: https://doi.org/10.48550/arXiv.2310.19852.

[170] Kevin Ro Wang et al. "Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=NpsVSN6o4ul.

[171] Nelson Elhage et al. "A Mathematical Framework for Transformer Circuits". In: *Transformer Circuits Thread* (2021). https://transformer-circuits.pub/2021/framework/index.html.

[172] Chris Olah et al. "Zoom In: An Introduction to Circuits". In: *Distill* (2020). https://distill.pub/2020/circuits/zoom-in.

[173] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[174] Abhimanyu Dubey et al. “The Llama 3 Herd of Models”. In: *CoRR* abs/2407.21783 (2024). URL: https://doi.org/10.48550/arXiv.2407.21783.

[175] Stefan Heimersheim and Neel Nanda. *How to use and interpret activation patching*. 2024. arXiv: 2404.15255 [cs.LG]. URL: https://arxiv.org/abs/2404.15255.

[176] Maximilian Li and Lucas Janson. “Optimal ablation for interpretability”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: https://openreview.net/forum?id=opt72TYzwZ.

[177] Aaron Gokaslan et al. *OpenWebText Corpus*. http://Skylion007.github.io/OpenWebTextCorpus. 2019.

[178] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[179] Biwei Huang et al. “AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning”. In: *International Conference on Learning Representations (ICLR)*. 2022.

[180] Hsiao-Ru Pan and Bernhard Schölkopf. “Skill or luck? return decomposition via advantage functions”. In: *arXiv preprint arXiv:2402.12874* (2024).

[181] Hugo Touvron et al. “Going deeper with image transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 32–42.