# Addressing Data Scarcity in Autonomous Systems through Trustworthy Counterfactual Generation

Zhaoan Ye

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

SCHOOL OF ENGINEERING

JAMES WATT SCHOOL OF ENGINEERING


September 2025

# Abstract

Autonomous systems often operate in environments where collecting large, diverse, and safety-critical datasets is difficult. This data scarcity limits their reliability, particularly in rare or hazardous scenarios that are hard to capture in the real world. This thesis addresses data scarcity by integrating structural causal models with diffusion-based generative models to produce trustworthy, high-fidelity counterfactual images for "what-if'' reasoning. Thus, two frameworks are proposed: Causal DiffuseVAE and Causal DiffuseLLM. Both generate images that follow a directed acyclic graph of semantic factors while preserving visual realism. The thesis first outlines key concepts in causal generative modeling and modern deep generative methods, highlighting that existing approaches either provide interpretable causal control with limited fidelity or achieve photorealism without reliable intervention behavior.

Causal DiffuseVAE structures the latent space using a causal graph and applies a diffusion decoder for detail reconstruction. Experiments show a 40% reduction in generation time and a 30% improvement in counterfactual accuracy compared with state-of-the-art causal diffusion models. Causal DiffuseLLM, which maps language instructions to causal interventions, improves generation accuracy by 15% over its non-LLM baseline and localizes edits to causally affected regions.

Overall, this thesis shows that embedding causal reasoning into diffusion pipelines provides a practical path to generating reliable data for autonomous systems operating under limited data conditions.

# Contents

**Abstract** **iii**

**Acknowledgements** **xi**

**Declaration** **xii**

**Associated Publications** **xiii**

**1 Introduction** **1**

1.1 Background . . . . . 1

1.2 Motivation . . . . . 6

1.3 Tasks for Causal Image Generative Model in Engineering . . . . . 10

1.4 Dissertation Structure . . . . . 13

**2 Related Literature** **15**

2.1 Principles of Causality . . . . . 15

2.1.1 Causal Reasoning . . . . . 16

2.1.2 Causal Inference . . . . . 27

2.2 Image Generation . . . . . 37

2.2.1 Image Generation Models . . . . . 37

2.2.2 Interpretability in Image Generation . . . . . 41

2.3 Large Language Models . . . . . 48

2.3.1 General Large Language Models . . . . . 48

2.3.2 Generation Models Based on Large Language Models . . . . . 49

2.3.3 Generation Models Based on Multimodal Large Language Models 52

2.4 Summary . . . . . 55

**3 Preliminaries and Problem Formulation** **57**

3.1 Intervention and Counterfactual in Structural Causal Models . . . . . 57

3.2 Causal Inference in Machine Learning . . . . . 61

3.3 Variational Autoencoders . . . . . 64

3.3.1 Structure of the Autoencoders . . . . . 64

3.3.2 Structure of the Variational Autoencoders . . . . . 66

3.3.3 Derivation of the Variational Autoencoders . . . . . 67

3.4 Diffusion Models . . . . . 68

3.5 Transformer . . . . . 72

3.5.1 Structure of the Transformer . . . . . 73

3.5.2 Attention Block in the Transformer . . . . . 75

3.5.3 Mechanism of the Transformer . . . . . 76

3.6 Large Language Model . . . . . 79

3.6.1 Mechanism of the LLM . . . . . 80

3.6.2 Finetuning in the LLM . . . . . 84

3.7 Evaluation Metrics . . . . . 86

3.7.1 Pixel-Level and Perceptual Metrics . . . . . 87

3.7.2 Visual and Representation-Level Evaluation . . . . . 87

3.7.3 Training Loss Functions . . . . . 88

3.8 Problem Formulation . . . . . 89

3.8.1 Limitations of Existing Components . . . . . 89

3.8.2 Unified Problem Statement . . . . . 90

3.8.3 Strategy Adopted in This Thesis . . . . . 91

3.9 Chapter Summary . . . . . 92

**4 Trustworthy Counterfactual Generative Model Based on Causal Inference** **93**

4.1 Causal Mechanism . . . . . 94

4.2 Model Learning . . . . . 98

4.2.1 Learning Strategy with No Confounders . . . . . 98

4.2.2 Learning Strategy with Confounders . . . . . 101

4.3 Experiment Setting . . . . . 105
4.3.1 Experimental Dataset . . . . . 107
4.3.2 Experimental Setting . . . . . 109
4.3.3 Experimental Results . . . . . 110
4.4 Conclusion . . . . . 128

**5 Causal Diffusion Model Based on the Large Language Model 130**
5.1 Model Architecture . . . . . 131
5.1.1 Transformer with the Large Language Model . . . . . 132
5.1.2 Large Language Model Integration and Fine-tuning . . . . . 133
5.1.3 Query Transformer . . . . . 134
5.1.4 Cross-Attention Fusion . . . . . 135
5.1.5 Convolutional Latent Projection & Causal Masking . . . . . 136
5.1.6 Diffusion Model in the Causal DiffuseLLM . . . . . 137
5.2 Experiment and Discussion . . . . . 138
5.2.1 Experimental Setting . . . . . 138
5.2.2 Experimental Results . . . . . 141
5.3 Conclusion . . . . . 147

**6 Conclusion 148**
6.1 Research Contribution . . . . . 148
6.2 Limitation . . . . . 151
6.3 Future Work . . . . . 152

# List of Tables


4.1 Network Design of the Diffusion Model (DM) in the Causal DiffuseVAE for MNIST and Flow Datasets . . . 105

4.2 Network Design of CausalVAE Encoders and Decoders for Smile, Age, and Pendulum Datasets . . . 106

4.3 Details of the Diffusion Model . . . 106

4.4 MAE Comparison for the Smile Dataset . . . 118

4.5 Comparison on capabilities of the Causal DiffuseVAE and other baseline methods . . . 120

4.6 LPIPS comparison of ablation results . . . 121

4.7 Comparison of the counterfactual images in evaluation using the MAE criterion . . . 124

4.8 LPIPS comparison of counterfactual image quality . . . 124

4.9 LPIPS comparison of image quality under different training data ratios . . . 125

4.10 LPIPS scores of Causal DiffuseVAE, CDAE, and CDM at different sampling steps . . . 125

4.11 Training and inference time of different models on the Shadow Dataset . . . 126

4.12 MAE Comparison on MNIST and Flow Datasets . . . 127

5.1 Comparison of the counterfactual images using the MAE . . . 143

5.2 Comparison of the counterfactual images using the LPIPS . . . 144

# List of Figures

1.1 The factors in an image of a desktop computer. . . . . 3
1.2 Causal inference helps to find the causation relationship rather than the correlation relationship between factors. . . . . 4
1.3 Pearl's Causal Leverage [20]. . . . . 6
1.4 The example of the counterfactual [24]. A causal model learned from simulated stacking data identifies minimal changes in task variables that would convert a failure into a success, enabling interpretable, transferable explanations of robot actions. . . . . 7
1.5 A robotic dog is investigating a stick hidden in the shadow. (a) The stick in the shadow of a nearby object, (b) the object with shadow and light source, and (c) the object without shadow by counterfactual imagination. . . . . 8
1.6 Process of training robots with counterfactual images. . . . . 9

2.1 The Framework to explain why the action of the robot failed [41]. A causal model is learned to capture relationships between spatial variables and task success. When stacking fails, the model evaluates alternative variable configurations to identify the closest counterfactual conditions that would have led to success. By comparing the current state with these successful counterfactuals, the system generates a contrastive causal explanation. . . . . 16
2.2 Causal analysis when FCI was applied alongside PC, GES and GRaSP [50]. The figure highlights differences in edge orientation, handling of latent confounders, and graph sparsity across score-based and constraint-based approaches. . . . . 20

2.3 Causal graph obtained from the DirectLiNGAM-based Structural Equation Model (SEM) model [55]. Nodes denote socioeconomic, demographic, and behavioral variables, while directed edges represent estimated causal effects, with color and thickness indicating their sign and strength. . . . . 25
2.4 An example of causal inference used to answer the question [60]. Visual and textual inputs affect predictions via an attention-based latent variable, illustrating how confounding can arise when attention encodes non-causal correlations. . . . . 28
2.5 Structural Causal Model and the causal matrix . . . . . 30
2.6 An example of what images the causal generative model will generate [78]. The model learns causal affordances through observation and intervention, generalizes them to novel objects, and supports task-level planning. . . . . 35
2.7 An example of what the VAE trained with disentanglement learning could generate [100]. . . . . 42
2.8 An example of what images the causal generative model will generate [111]. 46
2.9 An overview of DALL · E 2 to generate images based on the human language [127]. . . . . 49

3.1 Causal relationship with confounders. . . . . 60
3.2 Structure of the conventional Autoencoder . . . . . 65
3.3 Structure of the conventional Variational Autoencoder . . . . . 66
3.4 Structure of the conventional diffusion model . . . . . 69
3.5 Structure of the conventional Transformer [95] . . . . . 73
3.6 Attention mechanism in the Transformer [95] . . . . . 74

4.1 Overall architecture of Causal DiffuseVAE. The labels and the images are the inputs of the model. Through the causal layer, the causal relationship among the latent can be learned in the training process. Furthermore, the intervention process can be achieved by changing the values of the labels of the causes to influence the labels of the effects. . . . . 95
4.2 The intervention process in the shadow situation. When the latent of the object size is changed, the shadow area in the causal layer will also change. 97

4.3 The causal relationships in the shadow scenario. . . . . 97
4.4 Architecture of Causal DiffuseVAE with confounders. . . . . 102
4.5 Causal matrix A at the $100^{\text{th}}$ epoch . . . . . 111
4.6 Results on the MNIST Dataset using two different methods . . . . . 112
4.7 Intervention results of shadow datasets using Causal DiffuseVAE. The object in (a) is the cube and in (b) is the polyhedron. . . . . 113
4.8 Counterfactual images generated by Causal DiffuseVAE, CausalVAE and CDRM. . . . . 114
4.9 Intervention results of flow datasets using Causal DiffuseVAE, CausalVAE and CDRM. . . . . 115
4.10 Intervention results of pendulum datasets using Causal DiffuseVAE, CausalVAE and CDRM. . . . . 116
4.11 Intervention results of CelebA datasets (Gender and Age) using Causal DiffuseVAE. . . . . 117
4.12 Results on the Smile Dataset . . . . . 117
4.13 Intervention results of Circuit datasets using Causal DiffuseVAE. . . . . 119
4.14 Ablation results on the shadow dataset. . . . . 121
4.15 Scatter plot of generated data and true data after PCA-based processing. (a) Causal DiffuseVAE. (b) CausalVAE. (c) CDRM. (d) CDAE. . . . . 123
4.16 Expanding intervention results of shadow using Causal DiffuseVAE. . . . . 126
4.17 MAE on the Factor Thickness . . . . . 128

5.1 Overview of the Causal DiffuseLLM architecture. . . . . 131
5.2 Intervention results when using different prompts. . . . . 142
5.3 Intervention results when changing the object size to 0.6 using Causal DiffuseLLM and baseline methods. . . . . 143
5.4 Intervention outcomes on Smile: Causal DiffuseLLM (pre-trained diffusion backbone) vs. InstructPix2Pix . . . . . 145
5.5 Scatter plot of generated data and true data after PCA-based processing. (a) Causal DiffuseLLM. (b) MagicBrush. (c) InstructPix2Pix. (d) CDAE. . 146

# Acknowledgements

I wish to express my sincere gratitude to everyone who supported the completion of this thesis. First and foremost, I thank my parents for their unwavering love and encouragement, which sustained me throughout my PhD journey in both life and study.

I extend my deepest gratitude to my supervisor, Prof. Dezong Zhao, whose rigorous guidance and patient support shaped this thesis from the first ideas to the final manuscript. His clear advice, thoughtful critique, and countless discussions helped me navigate difficulties, refine my methods, and present the results with clarity. He balanced high expectations with generous mentorship, giving me the freedom to explore while steering me in the right direction. Without his steady encouragement, especially during setbacks and final revisions, this work would not have been possible.

I also wish to acknowledge the crucial role of my second supervisor, Prof. David Flynn, whose strategic perspective and practical insight consistently strengthened this work. He provided detailed, constructive comments on multiple manuscript drafts, which improved the clarity and rigor of the finished manuscripts. His mentorship complemented Prof. Dezong Zhao's guidance, and together they provided the supportive framework that enabled this thesis to reach its present form.

Finally, I am deeply grateful to my friends and colleagues for steady encouragement, thoughtful discussions, and practical help throughout this journey. Whether offering feedback on drafts, their support lightened the load and strengthened this thesis. Thanks for being there.

# Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution

**Zhaoan Ye**

# Associated Publications

1. Zhaoan Ye, Dezong Zhao, Li Zhang, Wenjing Zhao, David Flynn, and Chongfeng Wei, "Causal VAE-DM: Trustworthy Generative Modeling for Image Modification Based on Causal Graphs," in Proc. IEEE Int. Conf. on Distributed Computing Systems (ICDCS), Glasgow, U.K., Jul. 2025
2. Zhaoan Ye, Dezong Zhao, Li Zhang, Xidong Yan, Qinglin Bi, Wenjing Zhao, and David Flynn, "Enhancing Data Efficiency with a Trustworthy Counterfactual Generative Model," in IEEE Transactions on Industrial Informatics (Minor Revision)
3. Zhaoan Ye, Dezong Zhao, Wenjing Zhao, Shibei Xue, and David Flynn, "Causal DiffuseLLM: Text-Driven Causal Representation Learning for Counterfactual Image Generation", in International Federation of Automatic Control (IFAC), (Under Review)
4. Xidong Yan, Dezong Zhao, Zhaoan Ye, Qinglin Bi, Xiao Yi, and David Flynn, "Adaptive Eco-Driving and Safe Traffic Control with Knowledge-guided Deep Reinforcement Learning", IEEE Transactions in Intelligent Transportation Systems, (Under Review)

# Chapter 1

# Introduction

Autonomous systems increasingly rely on visual perception to interact safely and effectively with complex real-world environments. However, these systems often operate under conditions that are difficult to capture exhaustively in training data, such as varying illumination, shadows, occlusions, and rare events. Models trained on limited or biased datasets may exhibit brittle behaviour when deployed, leading to perception failures that directly affect downstream decision-making and safety. This thesis addresses this challenge by exploring how causal inference and counterfactual image generation can be integrated into modern generative models to create reliable, interpretable, and data-efficient visual representations. By enabling explicit reasoning about "what-if" scenarios, the proposed approaches aim to improve robustness and trustworthiness in autonomous systems operating under real-world uncertainty.

## 1.1 Background

Engineering systems, from industrial inspection to autonomous robots, must operate reliably despite unpredictable lighting, occlusion, sensor noise, and other scene variations that are difficult to capture exhaustively in real-world datasets [1]. In practice, the scarcity of labelled data and the presence of visually confounding factors such as shadows often lead to brittle perception, causing failures in tasks like object detection, grasping, and scene understanding, especially in dynamic or safety-critical environ-

ments [2]. These challenges are clearly illustrated in robotic applications where a robot cannot recognise or manipulate an object simply because it is partially hidden by a shadow, or where changing environmental conditions cause the vision system to misinterpret key features [3]. Such limitations highlight a broader gap: current systems lack the ability to reason about how a scene would change under different conditions and therefore cannot perform reliable "what-if" analysis[4]. To address these issues, the research in this thesis explores image generative models based on causal inference to solve these application-driven problems, specifically for synthesising missing visual conditions, improving the robustness of robotic perception, and enabling robots to imagine counterfactual scenarios when real data are insufficient.

Although modern deep learning has improved visual perception in engineering applications, these methods still fundamentally rely on correlations in the data and therefore struggle under the challenging conditions outlined above. Convolutional neural networks can extract spatial features from microstructure images or industrial inspection data, and autoencoders compress large datasets into low-dimensional representations useful for simulation or design exploration [5]. Likewise, advances in transfer learning and self-supervised pretraining help reduce the need for extensive labelling [6].

Despite these advances, conventional deep learning models remain correlation-driven and struggle when visual conditions deviate from those seen during training [7]. This limitation extends to reinforcement learning (RL) systems as well [8]. Although RL enables autonomous agents to learn behaviours through trial-and-error interaction, the quality of the learned policy ultimately depends on the visual experiences available during training. If critical edge-case scenarios, such as objects hidden in deep shadow, are missing from the dataset, the agent inherits the same blind spots as its perception model and fails to respond appropriately in the real environment [9]. These challenges highlight the need for methods that can create the missing visual conditions rather than rely solely on those observed.

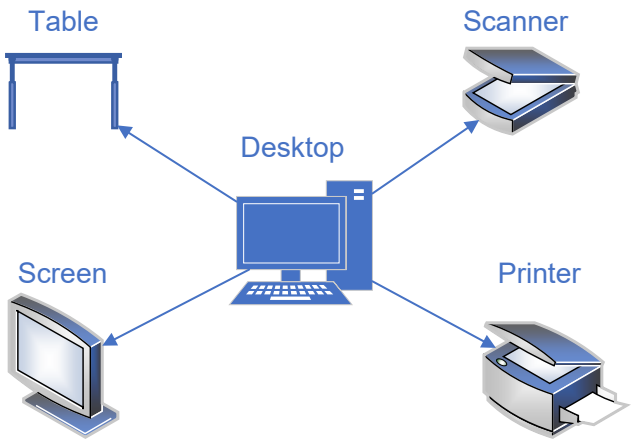



Figure 1.1: The factors in an image of a desktop computer.

Furthermore, the performance of the RL learning in the autonomous systems depends on the reliability of the data. Black-box ML and deep models may struggle to generalize under changing operating conditions, and RL's intensive data demands can limit deployment on physical systems [10]. Unreliable data may cause autonomous systems, specifically for autonomous vehicles, to make unsafe decisions, which will influence the safety of people. Methods that rely only on data usually can't guarantee how a system will behave when something new happens [11]. This becomes a major problem when we want to explore 'what-if' scenarios or optimize the system for conditions it has never seen before. For instance, in Figure 1.1, recognition of a desktop computer in an image is achieved by instructing a machine to search for characteristic elements, such as the table, screen, printer and scanner, which are considered key indicators of a desktop setup in the scene.

However, from a human perspective, these indicators are unreliable because the presence of such peripheral objects does not constitute proof of a desktop computer itself. To address this problem, causal inference is proposed to help a system understand the relationships between different factors. For instance, in autonomous driving, a vehicle must determine whether a pedestrian is about to cross the road based on their motion and orientation, not on correlated but irrelevant factors such as shadows cast by

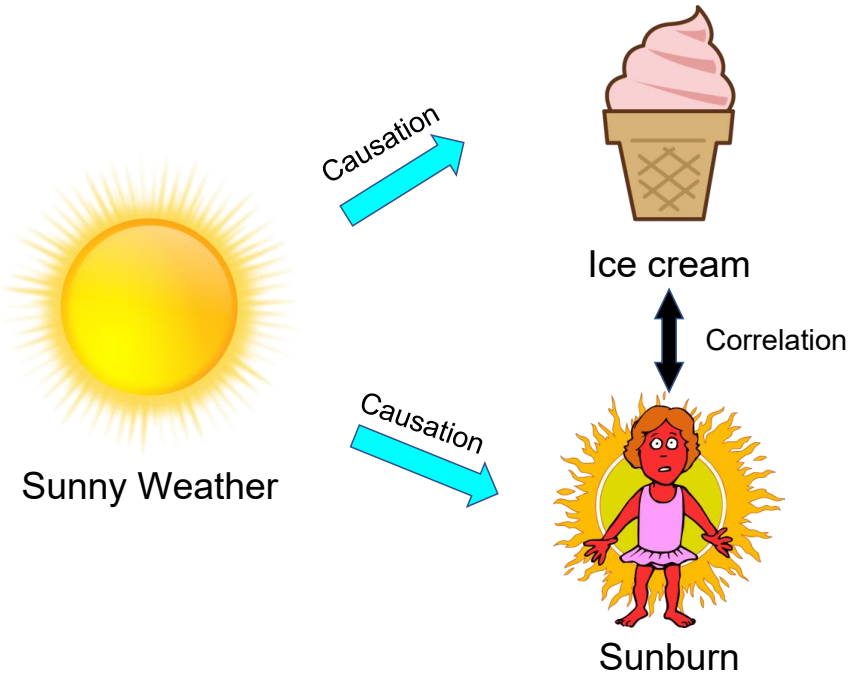



Figure 1.2: Causal inference helps to find the causation relationship rather than the correlation relationship between factors.

buildings or the presence of nearby signs. Figure 1.2 shows that causal inference finds cause-and-effect, rather than correlation. In this Figure, sunny weather causes both ice-cream purchases and sunburn. However, ice cream purchases do not cause sunburn and sunburn does not cause ice cream purchases.

What's more, modern advances in Artificial Intelligence (AI) have been driven not only by ever deeper neural architectures for vision and control, but also by the advent of large-scale Transformer-based language models [12]. These models, often billions of parameters trained in massive corpora, learn to represent, generate, and reason over text via self-supervised objectives [13], which allows the model to learn the chain of thought like a human. This has given rise to systems capable of tasks as diverse as code synthesis, question answering, and even rudimentary planning [14]. However, the learned knowledge is still based on the correlation. By pretraining on general web data and fine-tuning on domain-specific examples, Large Language Models (LLMs) can be adapted to specialized engineering workflows such as automating report generation, drafting design documentation, or translating high-level requirements into mathemat-

ical constraints for simulation code [15]. Their success underscores the broader theme of this thesis, namely that causal structure and interpretability, once embedded in neural models, can dramatically enhance both predictive power and trustworthiness. This holds across vision, control, and language modalities.

Importantly, interpretability remains critical, such as understanding model reasoning enables engineers to validate predictions, diagnose failure modes, and build trust in AI-augmented design tools [16]. To address these gaps, emerging research is integrating causal reasoning and physical priors into ML pipelines, delivering frameworks that provide robust, transparent, and intervention-capable predictions. This trend paves the way for the causal-aware generative methodologies developed in this thesis. The importance of interpretable machine learning is:

• **Enhanced Predictive Accuracy**: Machine learning models can uncover complex, nonlinear relationships in data, leading to more accurate predictions of system behavior, failures, and performance than traditional empirical or physics-only approaches.

• **Automated Decision-Making**: ML enables real-time, data-driven control and optimization, automating tasks such as adaptive process regulation, fault detection, and quality assurance. These capabilities often exceed human responsiveness while reducing the need for continuous manual oversight.

• **Efficient Design Exploration**: Surrogate ML models drastically reduce computational costs, allowing engineers to explore vast design spaces and identify optimal configurations quickly.

## 1.2 Motivation

Interpretability is essential in autonomous systems because perception failures can immediately translate into unsafe actions [17]. Tools such as intrinsic interpretability and post-hoc explanations help engineers in autonomous systems understand why a model behaves incorrectly [18, 19], but they cannot prevent the underlying issue: deep models often rely on spurious correlations instead of the true causal structure of a scene. Embedding domain knowledge, such as illumination physics or object–shadow relationships, into model design allows the learned representations to reflect meaningful causal factors rather than superficial visual patterns. This motivates the causal lens of this thesis, illustrated in Figure 1.3, which shows how autonomous systems must move beyond simple observation to intervention and counterfactual reasoning to operate robustly in real-world settings.

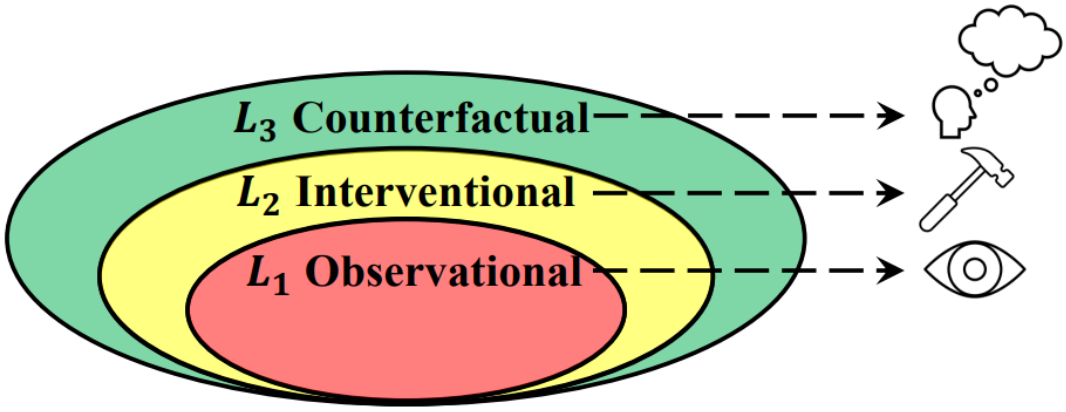



Figure 1.3: Pearl's Causal Leverage [20].

To achieve the intervention and counterfactual reasoning, causal inference provides a formal language for reasoning about cause–and–effect relationships [21], moving beyond mere statistical associations to allow statements of the form "if this variable were intervened upon, then that outcome would change by so much." At its heart lies the structural causal model (SCM), in which variables are connected by directed edges that encode mechanistic dependencies and exogenous noise terms [22]. When a scene or engineering process is represented as a graph of equations, an intervention is modeled

by replacing only the equation that changes [23]. In this way, precise "what-if" questions can be answered. By using the do-operator, passive observation is distinguished from active manipulation. Interventional distributions are then derived to predict how the system would behave under new design changes or operating conditions.

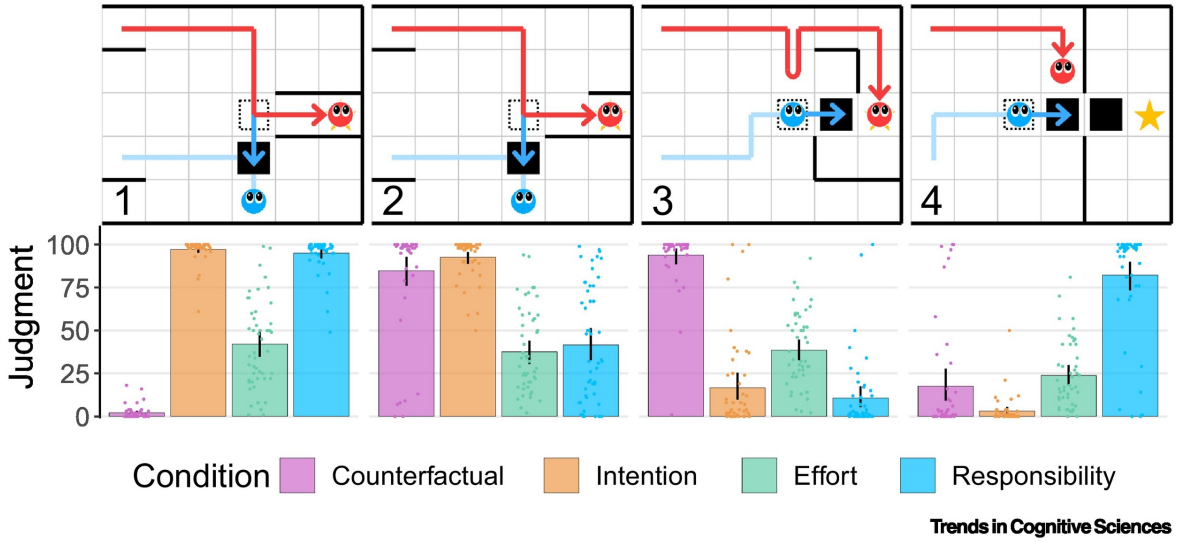



Figure 1.4: The example of the counterfactual [24]. A causal model learned from simulated stacking data identifies minimal changes in task variables that would convert a failure into a success, enabling interpretable, transferable explanations of robot actions.

At the same time, causal reasoning provides a way to formalise these requirements. SCMs represent how factors such as lighting, geometry, and material interact to form an image [25]. Figure 1.4 demonstrates how counterfactual reasoning asks not only "what if we change X?" but also "what would have happened under different conditions?" [24]. Figure 1.4 illustrates how humans use counterfactual reasoning to evaluate actions and assign responsibility in sequential decision-making tasks. The top row shows four grid-world scenarios in which two agents follow different paths toward a goal or obstacle, allowing certain events to occur or preventing them. The bottom row presents human judgment scores across four dimensions, condition, counterfactual, intention, effort, and responsibility, for each scenario [26]. For autonomous perception, this enables the system to test whether a detection failure is caused by a shadow, an occlusion, or a genuine absence of the object [27]. Existing deep generative mod-

els, however, cannot reliably perform such targeted interventions: VAE-based causal models lack photorealism, while diffusion models achieve fidelity but entangle causal factors, making controlled counterfactual edits unreliable. This gap motivates the need for new causal-aware generative architectures.

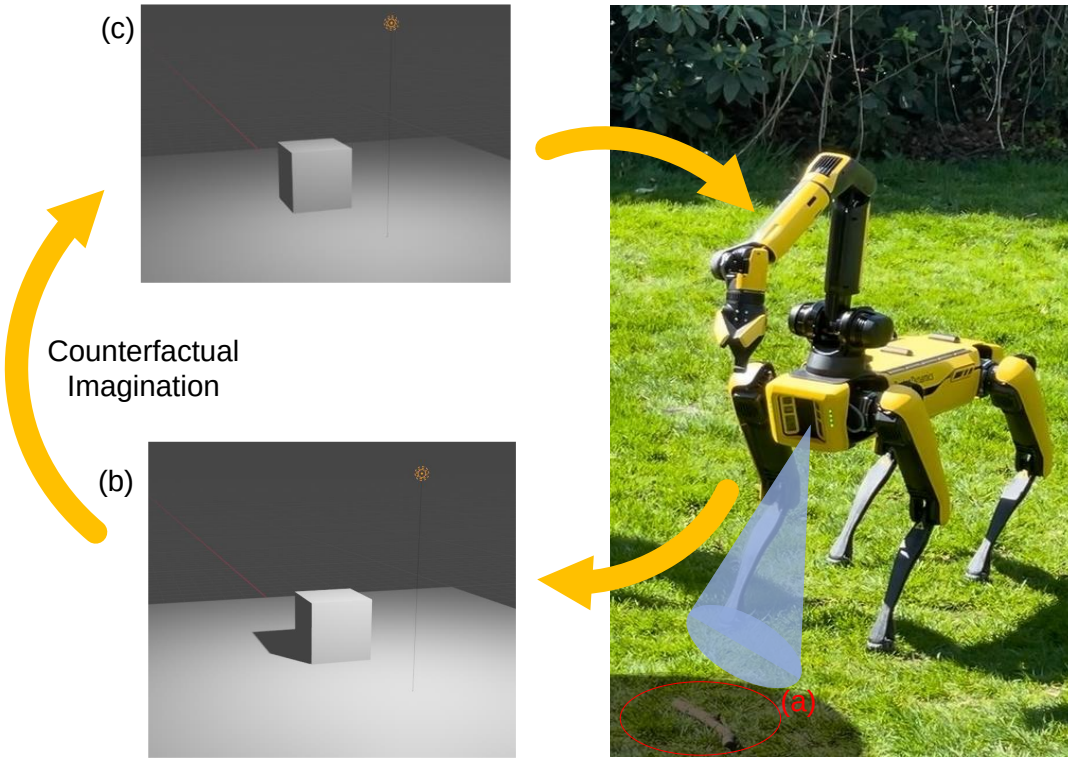



Figure 1.5: A robotic dog is investigating a stick hidden in the shadow. (a) The stick in the shadow of a nearby object, (b) the object with shadow and light source, and (c) the object without shadow by counterfactual imagination.

In the real world, a concrete manifestation of this brittleness was encountered during the development of a quadrupedal robotic platform for autonomous object retrieval. Figure. 1.5 presents an example of a robotic dog trying to pick up the stick hidden in the shadow. In the causal inference, the problem could be regarded as "What if the shadow is removed?" Under uniform illumination, the vision-and-grasp pipeline functioned reliably; however, when objects resided in deep shadow, detection failed and grasp proposals became invalid. This failure mode highlighted that existing generative models, lacking any semantics of lighting interventions, are unable to imagine how an object appears under altered illumination.

Furthermore, the opacity of black-box models undermines trust and hampers diagnostic insight in safety-critical systems [28]. Without interpretability, engineers cannot verify that predictions conform to physical laws, cannot diagnose the reasons for sudden performance degradation, and cannot trust "what-if" analyses essential for informed design modifications. Model-agnostic attribution methods, surrogate models and counterfactual explanations can help explain how decisions are made [29]. However, they fall short when hidden factors do not match known causal relationships.

To address these gaps, this thesis proposes a causal-aware generative framework that embeds structural knowledge of scene formation directly into deep architectures. By defining latent spaces in accordance with causal graphs of illumination, object geometry, and material properties, it is expected that shadow-robust training data can be synthesized through explicit lighting interventions, that interventional grasp planning can be enabled via counterfactual scene generation, and that critical data gaps can be filled without exhaustive field collection. In this way, the models are expected to perform well and explain their decisions clearly.

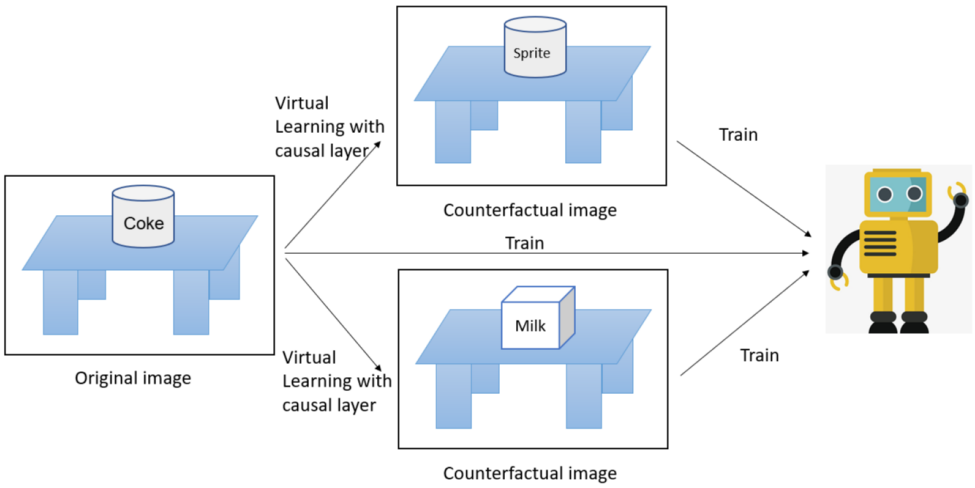



Figure 1.6: Process of training robots with counterfactual images.

Causal image generative models promise the ability to produce and manipulate visual data through explicit interventions on underlying factors such as illumination, geometry, and material properties. For instance, Figure. 1.6 indicates how the counterfactual images could be used to train robots. A causal layer swaps the object while keeping the scene the same, creating versions like Sprite and Milk, so the robot learns to gen-

eralize beyond the original Coke can. However, there is a fundamental obstacle: how to accurately recover latent representations that reflect the distinct causal variables [30]. Images contain many physical and semantic factors that appear and interact together. Without well-designed biases or constraints, learning algorithms often mix these factors into tangled and hard-to-separate representations [31]. Due to highly nonlinear decoders and data issues such as occlusion, noise, and distribution shift, an inference network must invert the generative mapping to recover the latent variables. This makes it hard to tell, for example, whether a change is due to lighting or to the texture of the object. Unmodeled confounders and sensor artifacts further obscure the true causal signals, leading to representations that neither support reliable counterfactual synthesis nor obtain interpretable semantics [32]. Overcoming these challenges requires novel architectures, regularization strategies, and learning paradigms that can disentangle, ground, and validate causal factors within image data. Designing a method with novel architectures, regularization strategies, and learning paradigms is an essential step toward robust, transparent, and intervention-capable generative vision systems.

## 1.3 Tasks for Causal Image Generative Model in Engineering

Efficiently and transparently processing data and learning from data become more essential for perception and decision-making in industry, particularly for robotics. It is natural to raise the following questions:

**Q1: How to efficiently and transparently obtain data for industry tasks?**

By using interpretability tools, such as decision trees, linear models, feature-attribution methods like SHapley Additive exPlanations (SHAP) [33] or Local Interpretable Model-agnostic Explanations (LIME) [34] and partial-dependence plots, engineers can find exactly which visual features cause a model's uncertainty or errors. They can then

focus data collection or synthetic image generation on these problem areas, tagging each new sample with the explanation that motivated it. Counterfactual variations, such as small changes in angle or lighting, can be created to produce paired examples that expose the model's decision boundaries. This makes data collection much more efficient and fully transparent, since the reason for each sample is directly linked to an interpretable insight from the model.

In this thesis, the proposed Causal DiffuseVAE and Causal DiffuseLLM frameworks go beyond traditional interpretability tools by allowing engineers to directly manipulate causal factors in the latent space (e.g., lighting, shadow, object presence). This enables the targeted generation of counterfactual training samples without manually designing perturbations, providing a principled and automated mechanism for transparent data augmentation.

**Q2: How to use the causal inference to generate data for industry tasks?**

First, the industrial imaging process is described as a structural causal model, with nodes for factors like component state, lighting angle, surface material, and sensor noise. A conditional generative model can then be trained on a small set of real, labeled images. During training, each image is matched with the values of its causal variables. When generating new images, specific causal inputs are changed using the do-operator, and the model creates clear, realistic images that match the chosen "what-if" scenarios. Every generated sample is tagged with its intervention provenance, enabling transparent tracing of its origin and a back-door calibration step against held-out real images ensures fidelity. This process creates synthetic images to cover important gaps, like shadowed parts or rare object positions. It greatly reduces the need for extensive real-world image collection and provides traceable, causally based data for later industrial vision tasks.

This thesis develops causal diffusion-based generative models that explicitly encode the causal graph within the latent space. By integrating SCM structure with diffusion decoding, Causal DiffuseVAE and Causal DiffuseLLM produce high-fidelity images that faithfully reflect the specified interventions. This enables controllable generation of industrial scenarios, such as removing shadows or modifying illumination, directly through causal manipulation.

**Q3: How to guarantee the quality of the data for industry tasks?**

To ensure high-quality synthetic images from the model for industrial vision tasks, use a closed-loop assurance process that begins with calibrating generated images against a held-out set of real, labeled images to match appearance and feature statistics. Next, perform counterfactual consistency checks by inferring latent causes from real images, applying targeted changes, and confirming that the synthetic results match those seen in paired real captures. Continuously monitor coverage across all causal conditions to avoid missing important cases or over-representing certain ones, and enforce lightweight physics-based constraints so that shadows, reflections, and occlusions remain realistic. Assign confidence scores to flag uncertain samples for human review. Finally, validate downstream performance by training vision models on the augmented dataset and evaluating them on a real-world benchmark, using any performance gaps to guide further targeted interventions and refine the generator.

The models proposed in this thesis incorporate causal consistency losses, disentangled latent structures, and diffusion refinement modules that enforce visual and semantic realism. Through ablation experiments and quantitative evaluations, the thesis demonstrates that the generated counterfactual images maintain high fidelity while preserving all non-intervened factors, an essential requirement for trustworthy industrial deployment.

**Q4: How to validate the generated data for industry tasks?**

Data validation begins by calculating Mean Absolute Error (MAE) on a per-pixel basis and measuring Learned Perceptual Image Patch Similarity (LPIPS) [35] distances between synthetic and held-out real images to ensure minimal reconstruction error and perceptual drift. These checks are complemented by the Structural Similarity Index Measure (SSIM) [36] and distributional scores such as the LPIPS to confirm overall fidelity. Downstream task performance is then evaluated on a real-world benchmark to verify that the synthetic data improves the intended models. At the same time, coverage statistics are gathered across all causal factors, including illumination levels, object poses, and material types, to identify under- or over-represented cases. Counterfactual consistency tests, where latent causes are inferred from real images, modified through targeted interventions, and regenerated, are used to ensure that synthetic changes reflect real-world behavior.

This thesis introduces evaluation protocols tailored to causal generative models, including tests for causal faithfulness, intervention precision, and stability under distribution shift. Experiments show that models trained with the generated counterfactual data achieve improved robustness in recognition and manipulation tasks, confirming the practical value of the proposed frameworks for industrial applications.

## 1.4 Dissertation Structure

The dissertation consists of five chapters, with the outline as follows:

Chapter 1 outlines the motivation of the thesis, presenting the challenges faced by autonomous vision systems under data limitations and distribution shifts. It introduces the need for causal and counterfactual generative modelling and defines the key tasks that the proposed frameworks aim to address.

Chapter 2 reviews prior research on causal inference, causal representation learning, image generative models, and Large Language Models. It identifies the limitations of existing approaches in achieving controllable, interpretable, and high-fidelity counterfactual image generation.

Chapter 3 presents the theoretical foundations required for the thesis, including structural causal models, variational inference, diffusion processes, and multimodal encoders. These fundamentals provide the basis for understanding the proposed causal generative architectures.

Chapter 4 introduces the Causal Diffuse Variational Autoencoder, detailing its causal latent structure, masked causal layer, and diffusion-based decoder. It demonstrates how the model achieves identifiable causal factors while producing photorealistic counterfactual images.

Chapter 5 presents the LLM-guided causal diffusion framework, which integrates language-conditioned causal reasoning with high-quality generation. It describes how free-form instructions are mapped to structured interventions and how the model performs localized, semantically consistent edits.

Chapter 6 summarises the contributions of the thesis and discusses the implications of causal generative models for autonomous systems. It also outlines limitations and presents directions for future research in scalable causal modelling and domain-adaptive counterfactual generation.

# Chapter 2

# Related Literature

Building on the challenges of data scarcity, visual confounding factors (such as shadows and illumination changes), and the brittleness of correlation-driven perception models in autonomous systems, this chapter reviews existing research relevant to causal and counterfactual image generation for autonomous systems. It surveys prior work in causal inference and causal reasoning, as well as recent advances in image generative models and large language models. Particular attention is given to how current approaches address, or fail to address, issues of interpretability, controllability, and robustness under distribution shift. By analysing the strengths and limitations of these methods, this chapter clarifies the research gaps that motivate the causal diffusion frameworks proposed in the subsequent chapters.

## 2.1 Principles of Causality

To support trustworthy counterfactual image generation in autonomous systems, it is necessary to draw from two complementary strands of causal methodology [37]. Causal reasoning focuses on discovering and exploiting causal structure, often in the form of graphs, to explain failures and to support structured decision-making [38]. Causal inference, in contrast, focuses on estimating the effects of interventions and counterfactual queries under explicit assumptions about confounding and identifiability [39].

Together, these tools enable engineers to predict system responses to design changes, simulate "what-if" queries in digital twins, and augment datasets with interventional or counterfactual examples, all under transparent and auditable assumptions about underlying causal mechanisms [40].

### 2.1.1 Causal Reasoning

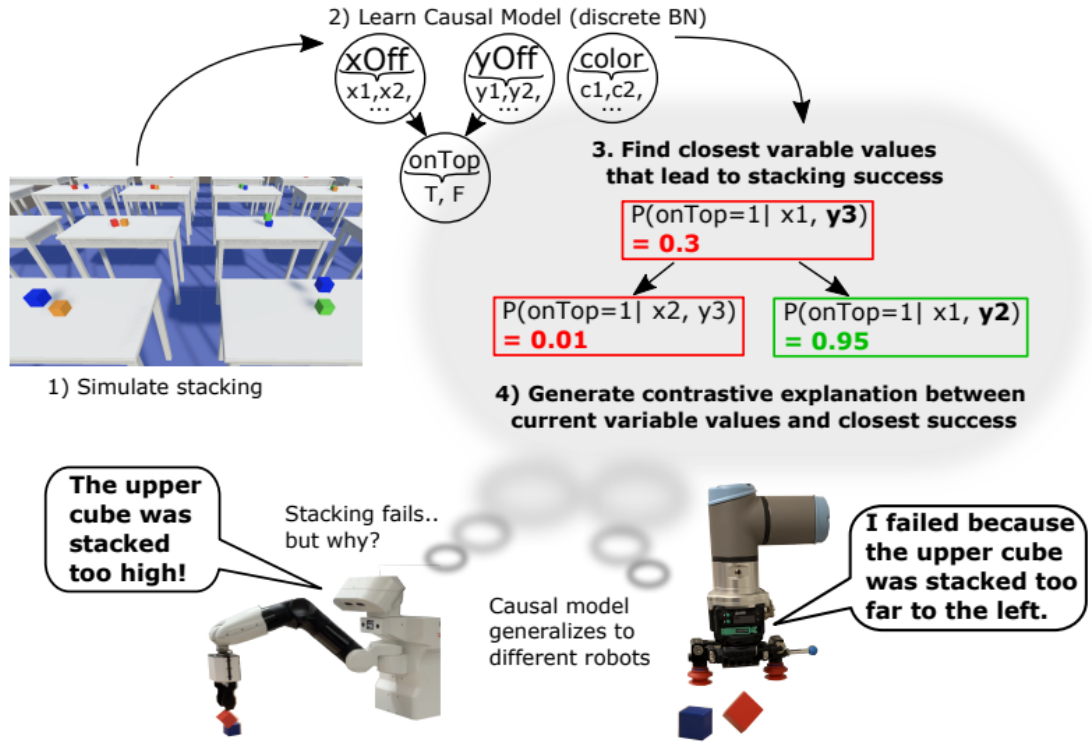



Figure 2.1: The Framework to explain why the action of the robot failed [41]. A causal model is learned to capture relationships between spatial variables and task success. When stacking fails, the model evaluates alternative variable configurations to identify the closest counterfactual conditions that would have led to success. By comparing the current state with these successful counterfactuals, the system generates a contrastive causal explanation.

Causal reasoning seeks to uncover the underlying causal graph that describes mechanistic dependencies among variables directly from observational data. Constraint-based algorithms, such as the Peter–Clark (PC) algorithm [42] and Fast Causal Inference (FCI) [43], use conditional independence tests to gradually add and orient edges while limiting false positives. They rely on strong assumptions such as causal sufficiency and faithfulness. However, these assumptions are often violated in real-world systems,

which means they cannot be used in the industry due to their unreliability. Causal sufficiency requires that all common causes of the observed variables are measured, which is rarely true in practice due to unobserved environmental factors, sensor noise, or latent system states. Faithfulness assumes that all statistical independencies in the data correspond exactly to separations in the true causal graph. In complex or finely tuned systems, causal effects may cancel out, leading to misleading independencies. When these assumptions fail, constraint-based methods can remove true causal edges, misorient directions, or produce partially identified graphs, reducing the reliability of the inferred causal structure, especially under distribution shift or when interventions are later applied.

In the past decade, causal reasoning has been widely used in statistics, economics and sociology. With the structured data in these fields, Causal-learn in [44] provides a tool for up-to-date causal discovery. Causal-learn fills a critical gap in the causal discovery ecosystem by providing the first fully Python-native library that unifies classical, score-based, functional, permutation-based, and Granger-causality methods under a coherent Application Programming Interface (API). In robotics, causal discovery is used to investigate why the motion of a robot failed [41]. Figure. 2.1 indicates the process of explaining why the action of the robot failed. First, a causal model is learned from simulations. When a task fails, a contrastive explanation is generated. Finally, the models are evaluated on two tasks, cube stacking and sphere dropping, and are transferred to two robots. Explanations are then provided whenever errors are made. This framework enables robots to generate contrastive causal explanations for execution failures based on learned causal Bayesian networks from simulated task data. The authors addressed the challenge of acquiring sufficient causal data by transferring models trained in simulation to reality, achieving sim-to-real accuracies of 70 % and 72 % in cube stacking and sphere dropping tasks. This framework demonstrates how explicit causal structure enables interpretable counterfactual reasoning about failure modes,

rather than relying on correlation-based diagnostics. The same principle motivates the causal generative models in this thesis, which aim to embed causal graphs into image generation pipelines to produce trustworthy counterfactual visual data for autonomous systems.

Furthermore, recent works involve causal reasoning with the LLMs. CLADDER, the first large-scale benchmark for formal causal inference in natural language, is proposed for comprising over 10000 questions spanning associational, interventional, and counterfactual queries derived from diverse causal graphs [45]. CLADDER is designed to test whether large language models can genuinely reason over causal structures, distinguishing association from intervention and counterfactual reasoning, rather than relying on superficial patterns in language. This work highlights that highly expressive generative models benefit from explicit causal representations when reliable "what-if" reasoning is required, which parallels the motivation for introducing causal structure into image generation models.

Over the past two years, the PC algorithm has been integrated into a variety of engineering domains to uncover causal structures from complex observational data, enabling more precise diagnosis, monitoring, and optimization. The PC algorithm assumes all common causes are measured and that every statistical independence reflects a true causal separation. It proceeds in two main stages. First, the PC algorithm initializes a fully connected, undirected graph. It then repeatedly tests each pair of variables for conditional independence given progressively larger subsets of their neighbors. When independence is found, the corresponding edge is removed. The conditioning (separating) sets are recorded for later use.

Second, it orients edges by turning any unshielded triple whose middle node was never in its corresponding separating set into a collider. Then it applies Meek's rules to direct as many of the remaining edges as possible without creating new colliders or cycles. PC can become slow when variables have many connections; in the worst case, its runtime grows quickly with node degree. A "stable" version removes order-dependence, so res-

ults don't change with the order you test edges. For the independence test, it uses Fisher's Z for Gaussian data, likelihood-ratio for discrete data and kernel-based tests for mixed or nonlinear data. In [46], a hybrid root-cause diagnosis framework is introduced that couples the PC algorithm with a discrete random genetic optimization routine. Starting from high-dimensional sensor data across the blast furnace and rolling mills, the PC algorithm learns a skeleton graph of conditional independencies, which the genetic component then refines to pinpoint minimal intervention sets. This approach reduced the mean time to fault isolation by over 30 %, translating directly into reduced downtime and scrap rates. However, its reliance on heuristic and randomized optimization to learn the Bayesian network structure brings substantial computational overhead and potential instability.

Moreover, a chemistry study applies PC-based causal discovery to steady-state concentration and flow measurements to automatically reconstruct the process topology [47]. By iteratively testing for conditional independencies among sensor pairs, the algorithm recovers which units are upstream or downstream with over 90 % edge-recovery accuracy under realistic noise levels. While without embedding time lags or dynamic context, the algorithm often produces ambiguous or incorrect edge orientations in recycle networks and fails to capture true causal directionality [48].

Besides, a PC-based method, RUN, is proposed to identify which services directly influence others [49]. However, temporal ordering in time-series data often violates the PC algorithm's simultaneous-test assumption, leading to missed or reversed edges. RUN, a neural Granger causal discovery model, embeds temporal context into each Conditional Independence (CI) test, significantly improving root-cause localization accuracy over the conventional PC baseline. Although RUN improves the capability of the PC-based method, it only captures predictive dependencies rather than guaranteed causal effects. Hidden confounders, non-stationary or multi-periodic signals, and com-

plex higher-order interactions can produce spurious causal relationships that won't hold under intervention. The heavy neural-network machinery also demands large volumes of high-quality time-series data and significant compute. These requirements limit the framework's real-time robustness and interoperability.

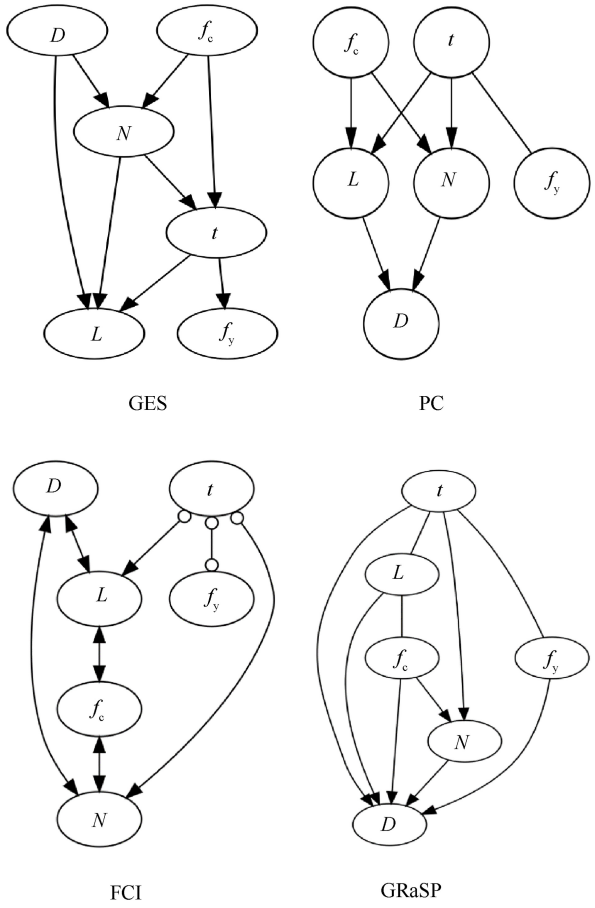



Figure 2.2: Causal analysis when FCI was applied alongside PC, GES and GRaSP [50]. The figure highlights differences in edge orientation, handling of latent confounders, and graph sparsity across score-based and constraint-based approaches.

The FCI algorithm begins by iteratively removing edges from a complete undirected graph through conditional-independence tests, just as in PC. It then performs additional tests on the remaining adjacencies using specially constructed d-separation (D-SEP) sets to detect dependencies induced by hidden confounders. Whenever an unshielded triple's middle node was never in its separating set, that triple is oriented as a collider. Next, an extended suite of orientation rules, including those that place "circle" marks on arrowheads to denote ambiguity from latent variables or selection bias, is applied repeatedly. This process continues until no further orientations or endpoint refinements are possible. The final output is a Partial Ancestral Graph that

compactly encodes all causal relations among the observed variables consistent with the data, despite unmeasured confounding. In [50], FCI was applied alongside PC, GES, and GRaSP to learn causal graphs for four structural-engineering problems. The analysis of the causal relationships was shown in Figure. 2.2. In this figure, GES (top left), a score-based method, produces a fully directed graph by optimizing a global score, but may introduce spurious edges when the scoring criterion favors complex models. PC (top right), a constraint-based method, yields a partially directed graph that reflects conditional independencies in the data, but its orientation is limited by assumptions of causal sufficiency and faithfulness. FCI (bottom left) extends PC by accounting for latent confounders, representing uncertainty using bidirected or circle-marked edges; this improves robustness to hidden variables but increases structural ambiguity. GRaSP (bottom right), a recent greedy randomized search procedure, balances scalability and flexibility by exploring multiple candidate structures, but may still produce dense graphs in the presence of correlated variables. Only the GES discovered a complete Directed Acyclic Graph (DAG). The problems addressed were the prediction of axial load-bearing capacity of columns, fire resistance of members, shear strength of beams, and blast resistance of walls. It was demonstrated that interpretable cause–effect relationships could be uncovered directly from experimental and simulated data using FCI. This approach obviated the need for manually derived formulas.

However, the exhaustive testing of conditional independencies over both neighbor sets and D-SEP sets, together with the iterative orientation rules, imposes a heavy computational and sample-size burden. This amount of work becomes especially prohibitive as the number of variables grows in structural-engineering datasets. While algorithms such as FCI aim to recover causal structure from observational data, they rely on large sample sizes, steady-state assumptions, and significant expert intervention. These requirements make them unsuitable for robotics and autonomous driving systems that demand real-time perception and decision-making, where high-dimensional visual inputs, rapidly changing environments, and strict latency constraints prevent extensive data collection and offline causal graph refinement. What's more, FCI was used to infer the causal structure among sensors in a wastewater-treatment plant [51]. The process

topology, including connections between bioreactors, clarifiers and recycle loops, was automatically reconstructed from steady-state concentration and flow measurements. The results demonstrated both the strengths and the limitations of applying FCI to systems with hidden dynamics. In particular, edges were sometimes misoriented when temporal dependencies were not taken into account. A primary drawback is that temporal dependencies were not modeled, resulting in occasional misorientation of edges. Steady-state data were used, which can fail to capture transient behaviors and lead to spurious independencies. A large sample size was required to detect conditional independencies reliably in noisy process measurements. Significant expert intervention was needed to resolve ambiguities in the reconstructed topology. These limitations highlight a key gap between causal discovery methods and practical deployment in perception-driven autonomous systems. While algorithms such as FCI aim to recover causal structure from observational data, they rely on large sample sizes, steady-state assumptions, and significant expert intervention. These requirements make such methods unsuitable for high-dimensional visual domains, where latent factors, temporal variation, and data scarcity are the norm. This motivates the approach taken in this thesis: rather than discovering causal graphs purely from data, causal structure is incorporated directly into generative models, enabling controllable and interpretable counterfactual image generation even when observations are limited or confounded.

On the other hand, score-based methods define a scoring criterion over graph structures. Algorithms like Greedy Equivalence Search (GES) and its continuous-data variants search for the graph that maximizes this score, rewarding fit and penalizing complexity. In practice, these methods trade higher computational cost for greater flexibility.

The GES algorithm operates in two phases over the space of equivalence classes, such as a Completed Partially Directed Acyclic Graph (CPDAG), using a chosen score function. In the forward phase, it starts with an empty graph and iteratively adds the single edge whose inclusion obtains the greatest improvement in score, updating the CPDAG at each step, until no addition can improve the score. In the backward phase, it starts with the graph produced by the forward phase. It then repeatedly removes the single edge

whose deletion gives the largest score improvement. After each removal, the CPDAG is updated. The process stops when no further deletion improves the score. By searching greedily over equivalence classes rather than individual DAGs, GES achieves score equivalence and consistency under causal sufficiency and score-equivalence assumptions, at the expense of potentially high computational cost when evaluating many candidate moves. Although GES provides a principled, score-consistent approach to causal structure learning, its greedy search over equivalence classes requires evaluating a large number of candidate edge additions and deletions, leading to substantial computational overhead. This makes GES impractical for robotics and autonomous systems that require rapid adaptation and real-time responses to changing visual environments. These limitations further motivate the approach adopted in this thesis, which embeds causal structure directly into generative models rather than relying on expensive offline causal discovery during deployment.

In the GES algorithm, the Bayesian Information Criterion (BIC) balances model fit and complexity. BIC scores each candidate structure using the maximized log-likelihood of the data. It then subtracts a penalty that grows with the number of free parameters and with the logarithm of the sample size. Specifically, it rewards models that explain the data well while penalizing those with more parameters, thereby favoring simpler structures when the sample is large. Under regularity conditions and when the true model is included among the candidates, BIC is known to select the correct model with probability approaching 1 as the sample size increases. In [52], candidate Bayesian network structures were scored using the BIC. The score balanced data fit against model complexity. This approach guided the selection of causal relations most predictive of equipment faults. However, the combination of reinforcement learning and Bayesian network structure learning introduces heavy computational complexity, since learning Bayesian networks from data is NP-hard.

Furthermore, the Conditional Probability Table (CPT) was discretized by imposing upper and lower probability bounds [53]. Values were grouped into a limited set of representative levels. A two-stage optimization process was then used in which these quantization limits were tuned to minimize mean-squared reconstruction error on held-out data while simultaneously penalizing model complexity via the BIC-based structure score. The two-stage optimization for setting quantization bounds adds computational overhead and may require careful tuning to avoid suboptimal trade-offs between fidelity and complexity.

Moreover, Fast Greedy Equivalence Search (FGES) is a scalable adaptation of GES designed to handle high-dimensional datasets more efficiently. It accelerates both the forward and backward phases by parallelizing score computations and caching intermediate results to avoid redundant evaluations. This parallelization produces significant speedups on large continuous or discrete data, making FGES well-suited for modern industrial monitoring tasks. However, the heuristic pruning and distributed coordination it relies on can sometimes lead to slight deviations from the true optimal equivalence class when score differences between candidate moves are small.

Additionally, the overhead of managing parallel processes may limit its effectiveness in environments with constrained computational resources. In [54], FGES was identified as the optimal causal discovery method. Candidate graphs were evaluated using a score-based search to model dependencies among operational, calendar, and weather factors. Interpretable causal structures were produced. These structures were then used to inform transit system management. However, in this work, FGES was applied to cross-sectional snapshots of delay data, so it could not explicitly model the sequential and time-lagged propagation of delays. As a result, some causal directions among delay factors may be misoriented when temporal dependencies are strong.

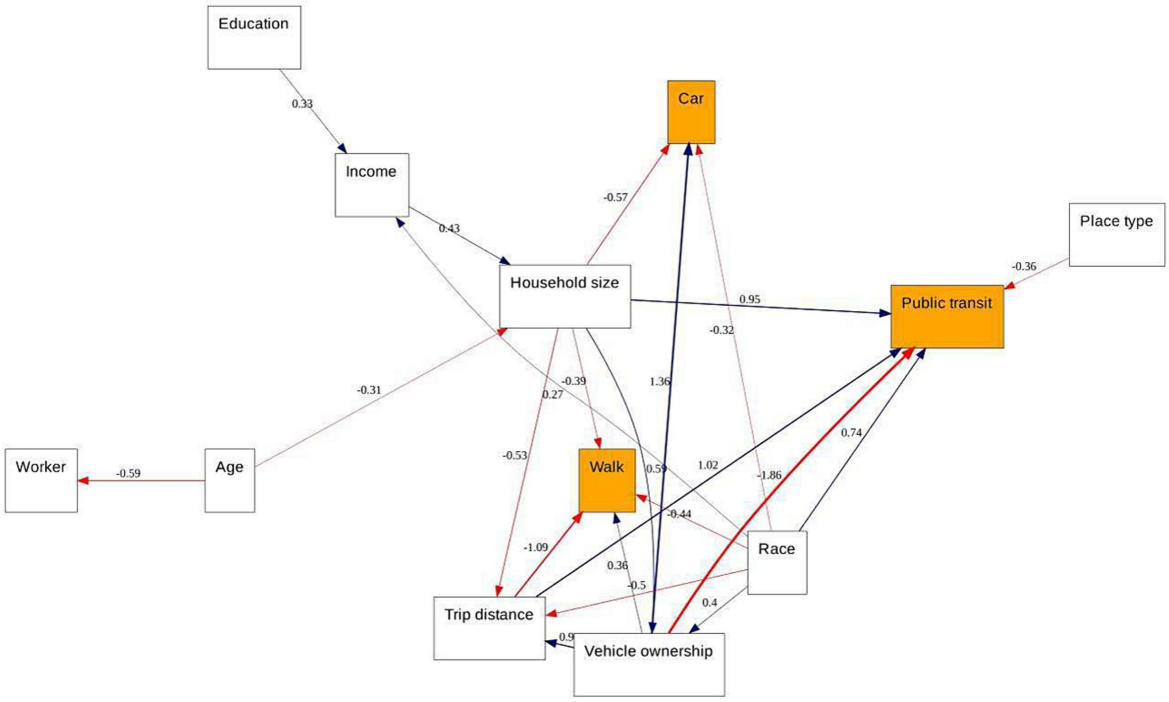



Figure 2.3: Causal graph obtained from the DirectLiNGAM-based Structural Equation Model (SEM) model [55]. Nodes denote socioeconomic, demographic, and behavioral variables, while directed edges represent estimated causal effects, with color and thickness indicating their sign and strength.

In [55], FGES was tested alongside PC, FCI, and Linear Non-Gaussian Acyclic Model (LiNGAM) to determine causal influences among socioeconomic, infrastructural, and behavioral variables. Figure. 2.3 presents the causal graph estimated with the DirectLiNGAM structural equation model. Blue edges denote positive path coefficients, while red edges denote negative ones. The thickness of each edge is scaled to the absolute size of its coefficient. The figure illustrates how interpretable causal pathways can be recovered from observational data, enabling reasoning about how changes in demographic or infrastructural factors propagate to travel behavior. At the same time, the complexity of the graph and the presence of many interacting variables highlight the challenges of identifying stable causal structures in real-world systems with latent factors and heterogeneous populations. The DirectLiNGAM structural equation model outperformed the other algorithms in recovering known causal relationships, guiding more accurate transportation planning and policy analysis. However, the method was applied to cross-sectional survey data, which could not capture temporal dynamics in travel behavior. Unobserved factors such as individual preferences and changes in infrastructure may have violated the causal sufficiency assumption.

Hybrid approaches combine these paradigms, first pruning the search space via independence tests and then scoring candidate graphs. For instance, functional causal models extend discovery by leveraging specific data-generating functions or distributional asymmetries to identify edge directions in linear or nonlinear settings.

More recently, continuous-optimization frameworks and gradient-based techniques have reframed graph search as a constrained optimization problem, enabling scalable discovery in high dimensions. Interventional and active learning extensions further improve identifiability by allowing targeted experiments that break confounding, while stability selection and bootstrap aggregation quantify uncertainty over learned structures. Together, these methods form the foundation for automatically extracting causal diagrams from data. A soft-sensing framework for mass customization in discrete manufacturing integrates an additive noise model with a tailored scoring mechanism to infer reliable causal graphs among process variables [56]. This framework enables robust sensor selection and accurate predictions even under high noise levels. Nevertheless, this framework struggles with the modular production, customized assembly, and multi-scale nature of mass customization processes. These characteristics can overwhelm the static causal graph structure. As a result, prediction accuracy is degraded.

In [57], a Fast Causal Discovery Algorithm based on the Additive Noise Model (FANM) was proposed to scale the Functional Causal Model (FCM) to large sensor networks, which employs coreset-based subsampling to accelerate Additive Noise Model (ANM)-based causal discovery on big industrial datasets. It achieved comparable causal-graph accuracy with up to 100 times faster runtimes. However, FANM relies on fitting a separate regression model for each candidate edge to estimate residuals and assign coreset weights, which can become computationally demanding and fragile to model errors in high-dimensional settings.

Moreover, except FCMs, the Causality-Driven Sequence Segmentation (CDSS) method was introduced that automatically breaks long process traces into causal phases. Then CDSS trains specialized regressors on each phase [58]. The CDSS method relies on detecting shifts in causal mechanisms to segment process phases. Consequently, noise or abrupt changes can cause these shifts to be misdetected, leading to inaccurate phase segmentation. These segmentation errors can then propagate to the soft-sensing models and degrade their predictive performance.

For the continuous-optimization frameworks, Non-combinatorial Optimization via Trace Exponential and Augmented Lagrangian for Structure learning (NOTEARS) was proposed to reformulate DAG structure learning as a smooth and constrained optimization problem [59]. It encodes the acyclicity constraint using a differentiable trace-exponential function. The resulting problem is then solved with augmented Lagrangian methods. This approach enables efficient gradient-based discovery of graph structures without resorting to combinatorial search. However, every augmented-Lagrangian iteration in the NOTEARS requires computing a matrix exponential, which scales cubically in the number of variables and thus becomes infeasible for very high-dimensional problems. Meanwhile, the resulting optimization is nonconvex. As a result, standard solvers can converge to suboptimal local minima. The continuous edge-weight estimates must then be thresholded to create a discrete DAG. This thresholding introduces sensitivity to the penalty and to the chosen threshold hyperparameters.

### 2.1.2 Causal Inference

Causal inference provides a unified framework for distinguishing correlation from causation by encoding variables and their mechanistic dependencies within structural causal models. These models consist of directed acyclic graphs augmented by the do operator. They are used to simulate interventions and answer what-if and counterfactual questions. Figure. 2.4 shows an example of how causal inference is used to answer the questions [60]. The upper part shows a vision–language model that combines visual

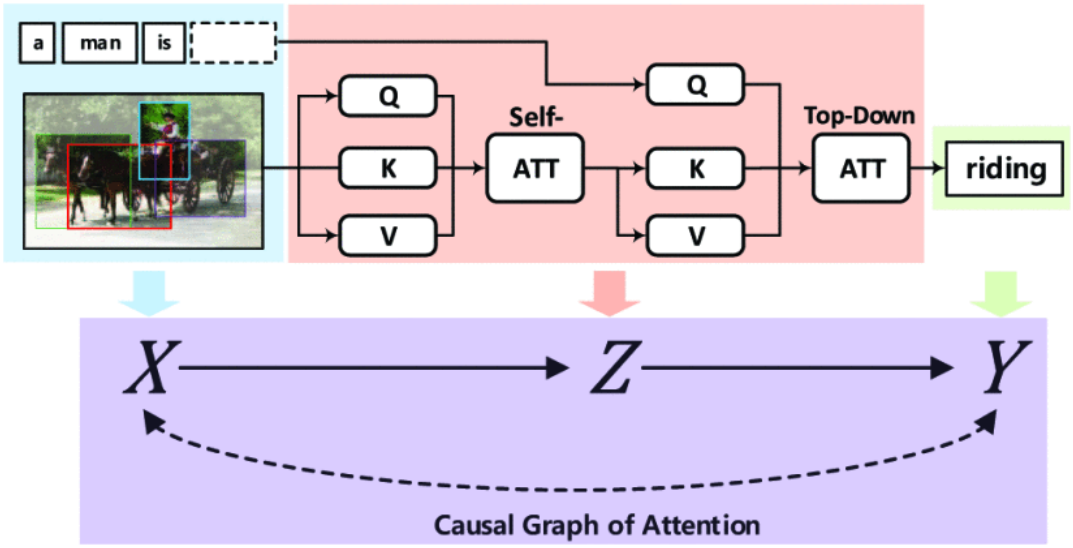



Figure 2.4: An example of causal inference used to answer the question [60]. Visual and textual inputs affect predictions via an attention-based latent variable, illustrating how confounding can arise when attention encodes non-causal correlations.

features and textual context through self-attention and top-down attention to predict an action label. Queries, keys, and values are computed from both image regions and language tokens, allowing the model to focus on semantically relevant visual evidence. The lower part abstracts this process as a causal graph, where the input observations X (image and text) influence an intermediate latent representation Z (attention), which in turn determines the output prediction Y. The dashed arrow highlights potential confounding effects when attention captures spurious correlations rather than true causal relationships. Through the causal inference, the word in the sentence is finally ”riding” rather than ”driving”.

Moreover, estimability depends on identifiable criteria such as back-door and front-door adjustments. These criteria guide techniques ranging from regression adjustment to propensity score weighting and instrumental variables. Modern methods include doubly robust estimators and causal forests. Causal discovery algorithms exploit conditional independencies, score-based searches and functional asymmetries to learn causal graphs directly from data. Extensions for time series and interventional data sets sup-

port dynamic modeling and active experimentation. In engineering domains, causal inference underlies digital twins for scenario analysis, supports targeted fault diagnosis and informs efficient data acquisition. These applications enable transparent models capable of predicting the effects of interventions in complex systems.

In practice, causal inference has been used in different fields, which motivated the beginning of this thesis. In medicine, causal inference is deployed for enabling more accurate, data-driven estimates of treatment and exposure effects while preserving the interpretability and robustness required for public-health decision making [61]. This emphasis on transparent intervention analysis directly parallels the goal of this thesis: generating counterfactual visual data that makes the effect of specific interventions, such as illumination or occlusion changes, explicit and auditable for autonomous systems. In robotic, causal inference is introduced into Robot Operating System (ROS) to enable seamless integration of robot and human state monitoring, asynchronous data batching, and real-time causal model induction onboard the robot [62]. This aligns closely with the motivation of this work, where causal generative models are designed to support real-time perception and decision-making by enabling robots to reason about how visual scenes would change under alternative conditions. In economics, by merging the potential-outcomes tradition with causal inference, causal inference bridge classical econometrics and modern causal inference, offering both interpretability and robustness in the face of unobserved confounding, selection bias, and structural heterogeneity [63]. These challenges mirror those in visual perception, where latent factors and dataset bias can mislead correlation-based models, motivating the causal latent representations proposed in this thesis. In sociology, causal inference is used to investigate human behaviour grapples with the intricate interplay of external stimuli and internal states, requiring methods that can disentangle ambiguous causal language, control for confounding, and address effect heterogeneity, interference, and varied timescales of influence [64]. Similarly, this thesis seeks to disentangle intertwined visual factors, such as lighting, geometry, and material properties, so that autonomous systems can generate and reason over meaningful counterfactual visual scenarios rather than relying on spurious correlations.

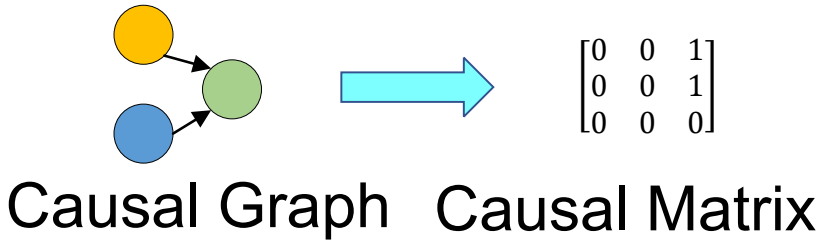



Figure 2.5: Structural Causal Model and the causal matrix

SCMs provide a formal framework for representing and reasoning about causal mechanisms in complex systems. Figure 2.5 shows two separate root nodes, each directing an arrow into a single downstream node, indicating that each root node has a direct causal influence on that downstream node. There are no arrows between the two root nodes, nor any arrows returning from the downstream node, so no feedback or interaction is implied among the roots or from the effect back to its causes. It is commonly used in causal learning, and is defined with a triple $\langle U, V, E \rangle$ [65]. Here, $U$ represents a set of exogenous variables determined by factors outside the model; $V$ indicates the system's internal variables; $E$ represents the unexplained variation in an endogenous variable not captured by the model's deterministic part. Interventions are implemented by modifying the structural equations for one or more variables, which corresponds to Pearl's *do*-operator, and allows it to compute interventional distributions that answer "what-if" queries. Counterfactual reasoning leverages the structural equations and assigned noise variables to imagine alternative scenarios that are consistent with observed data. Under the causal Markov and faithfulness assumptions, the support algorithmic effect identification via *do*-calculus, guiding engineers in designing valid adjustment sets and simulating system behavior under hypothetical changes. Model intervention is to convert an original image to a counterfactual image based on a causal graph. Normally, causal inference is performed through interventions, typically represented by setting a variable to a fixed value regardless of its natural state. This is commonly expressed using the *do*-operator, denoted as $do(X = x)$, which implies actively setting

$X$ to $x$ and observing the resultant changes in other variables [66]. The intervention aims to identify how changes in one variable (the cause) directly affect another variable (the effect). This process helps in answering counterfactual questions, such as "What would happen if we had changed $X$?"

In engineering, SCMs give clear, modular models of how a system behaves over time. They allow customers to safely run "what-if" tests in a digital twin, like turning off a component or changing a controller, and use counterfactual reasoning to diagnose faults, check safety, and pick better designs. By combining the graph-based semantics of SCMs with modern data-driven parameter estimation, engineers can build intervention-capable models that both explain observed data and predict the consequences of novel actions in complex systems.

A range of causal inference methods is employed to estimate the effects of interventions while confounding is accounted for. Back-door adjustment [67] identifies covariates that block spurious paths and estimates intervention effects by conditioning on those variables. Front-door adjustment [68] uses a mediator to recover unbiased estimates even when confounders remain unobserved. Propensity score methods [69] estimate treatment probabilities and apply matching or weighting to balance comparison groups. Instrumental variables techniques [70] rely on external instruments that influence the treatment but not the outcome directly to address hidden confounding. textcolorblue$g$-computation [71] fits outcome models and simulates counterfactual outcomes under different intervention levels. Doubly robust estimators combine outcome and propensity score models so that estimates remain consistent if one model is incorrect. Structural nested models and $g$-estimation handle time-varying treatments and confounders through sequential regression. Causal machine learning methods, such as causal forests, Bayesian additive regression trees, and meta learners, flexibly estimate heterogeneous treatment effects. Bayesian causal modeling specifies full probabilistic models and uses Monte Carlo or variational inference to derive posterior distributions over causal effects.

While these causal inference methods provide rigorous tools for estimating intervention effects under well-defined assumptions, they typically operate on low- to moderate-dimensional tabular data and require explicit specification of treatments, covariates, mediators, or instruments. In high-dimensional visual domains relevant to robotics and autonomous systems, many causal factors, such as illumination, occlusion, and material properties, are latent, intertwined, and difficult to measure or intervene on directly. This limits the practical applicability of traditional adjustment-based methods for generating visual counterfactuals in real time. The models proposed in this thesis address this gap by embedding causal structure into generative image models, enabling counterfactual synthesis and intervention analysis directly in the image space without requiring explicit enumeration of all confounding variables.

The Back-door adjustment involves identifying a set of covariates that blocks all non-causal pathways from the intervention to the outcome. The back door criterion guarantees that conditioning on this set removes bias from confounding variables. In practice, engineers use causal diagrams or domain expertise to select covariates that satisfy the blocking condition. They then estimate the causal effect by comparing outcomes between units that share the same covariate values but differ in their treatment status. This approach produces unbiased estimates of intervention effects provided all relevant confounders have been measured and appropriately adjusted for.

In a recent approach to adaptive, AI-based causal control for structural engineering, the back-door criterion is used to block misleading cause–and–effect paths [72]. By conditioning on disturbance variables, the method can estimate the effect of a control policy without bias, both in offline simulations and in real-time updates. However, identifying and measuring every relevant disturbance variable in real time can be impractical, so any unobserved or mismeasured confounders will bias the estimated intervention effects. Conditioning on a large set of covariates also increases estimation variance and computational burden, which can undermine real-time applicability. Finally, if some covariates act as colliders rather than confounders, adjusting for them can itself introduce spurious associations and distort causal estimates.

Furthermore, in a graph neural network (GNN) framework for fault diagnosis of complex industrial processes, back-door adjustment was recognized as a key causal-intervention technique, a causal intervention graph neural network (CIGNN) framework, to mitigate confounding from irrelevant sensor signals [73]. Although the inability to observe all confounders directly led the authors to adopt instrumental-variable methods instead, this framework analyzes the causality in the GNN-based fault diagnosis process based on causal theory and provides a method that automatically constructs sensor signals into graph data using an attention mechanism.

The front door criterion identifies a mediator variable that fully transmits the effect of the treatment to the outcome. It requires that there be no unmeasured confounders between the treatment and the mediator, and none between the mediator and the outcome. In practice, engineers first model how the treatment influences the mediator while accounting for any confounders of that link. They then model how the mediator affects the outcome, conditioning on the treatment. By combining these two models, they recover an unbiased estimate of the total causal effect even when direct confounders between treatment and outcome remain unobserved.

A Front-door Regulator was also introduced in a few-shot object-detection framework to block spurious generative factors, achieving more robust detection performance under domain shifts [74]. However, the Front-door Regulator depends on correctly specified semantic mediators that fully capture the causal generative factors and assumes there are no unmeasured confounders between those factors and the mediators or between the mediators and detection outcomes. This assumption is hard to verify in few-shot vision tasks. Furthermore, the two plug-and-play regularization terms introduce additional hyperparameters whose misconfiguration can cause over-regularization. Over- regularization leads to degraded detection performance when faced with strong domain shifts.

Moreover, a Hidden Confounder Removal (HCR) framework is presented that leverages front-door adjustment to decompose the total causal effect of user and item features into two partial effects via an observed mediator [75]. Through this decomposition, unobserved confounders are corrected for. As a result, recommendation accuracy under biased observational data is improved. However, the conditions of this framework is hard to satisfy as front-door adjustment demands a valid mediator that fully transmits the treatment effect and has no unmeasured confounders on either side.

In other causal inference methods, instrumental variable methods use a third variable that influences the treatment but has no direct effect on the outcome except through that treatment. In practice, the treatment is first predicted from the instrument and covariates.

Next, the outcome is modeled as a function of this predicted treatment. This two-stage procedure obtains consistent causal-effect estimates even when the original treatment is endogenous due to hidden confounding. A networked instrumental-variable approach was applied to sensor data from an industrial wastewater-treatment plant [76]. Instrumental variables were constructed from the process network topology to adjust for unobserved confounders among correlated sensors. The method was used to estimate the causal effect of aeration-rate changes on downstream biochemical-oxygen-demand measurements. Effect estimates were shown to be unbiased. These results facilitated improved process control and more efficient resource allocation. Nevertheless, external instruments derived from process network topology may inadvertently affect the outcome through unaccounted pathways, violating the exclusion restriction and biasing the estimated effects.

Furthermore, *g*-computation proceeds by first fitting a model for the outcome as a function of treatment and measured confounders. Predictions are then generated for each unit by setting the treatment to each level of interest while keeping confounders at their observed values. These predicted outcomes are averaged across the sample to estimate the marginal effect of the treatment. When confounders are continuous

or multidimensional, the required integration is approximated via Monte Carlo simulation. The G-Transformer framework integrated the *g*-computation algorithm into a deep sequential model for counterfactual outcome prediction in sepsis management [77]. The method first learned a sequence-to-sequence model of patient trajectories under observed treatments, and then simulated hypothetical treatment regimens through this model to generate counterfactual outcome estimates. This method effectively implements the *g*-computation formula. This approach produced more accurate predictions of organ-failure progression and mortality under alternative intervention strategies than standard time-series models.

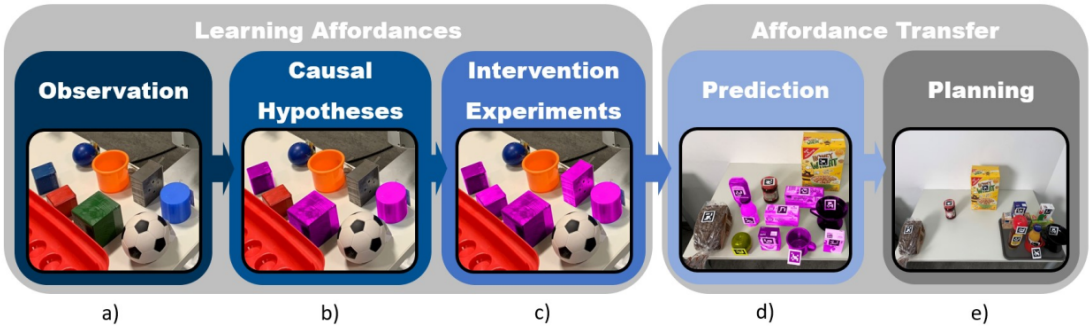



Figure 2.6: An example of what images the causal generative model will generate [78]. The model learns causal affordances through observation and intervention, generalizes them to novel objects, and supports task-level planning.

Causal machine learning is the most frequently used method in recent years. Causal machine learning uses predictive models to estimate how intervention effects vary across different settings. In [78], an early method was proposed to introduce causal machine learning into engineering. Figure. 2.6 shows this causal method using the affordance "supports stacking other objects on top". Objects the system classifies correctly are pink; mistakes are yellow. From left to right: (a) in a new environment, it watches human demonstrations and notes which object properties correlate with the stacking affordance; (b) it turns these observations into causal guesses about which properties actually cause stacking; (c) it runs the most informative small experiments by itself to test and refine those guesses; (d) it carries this property-based knowledge into a new environment with unfamiliar objects and predicts their affordances; and (e) using the learned causal models, it plans purposeful actions with previously unseen objects to achieve a goal.

In causal machine learning, causal forests extend random forests by partitioning on covariates that maximize treatment effect heterogeneity while using sample splitting to avoid bias [79]. Bayesian additive regression trees fit an ensemble of regularised regression trees and draw samples from their posterior to quantify uncertainty in individual treatment effect estimates [80]. Meta learner approaches, such as the T learner, S learner and X learner, decompose the problem into separate models for treated and control groups before combining predictions to improve accuracy and robustness [81]. Cross-fitting and sample splitting techniques are often employed to mitigate overfitting and achieve valid inference. These methods have been applied in engineering for tasks like predictive maintenance, where they estimate how different operational policies affect failure risk, and in energy systems to tailor control strategies to individual units based on their unique response profiles. In recent applications, a Double machine learning (DML) model was used to identify how adjustments to climate-control setpoints and external weather conditions propagate through a building's thermal network to affect indoor temperature and energy consumption [82]. The approach combined counterfactual simulation with ensemble-based causal effect estimators to estimate heterogeneous treatment effects for proposed control strategies. Case studies in commercial office zones showed that these causal estimates produced more reliable energy-saving measures than traditional black-box forecasting. However, the DML method relies on access to debiased building-operation data to ensure valid causal-effect estimates, while obtaining and validating truly debiased observational datasets from real facilities can be challenging. Moreover, causal forests were used to estimate how variations in supplier portfolio consistency affect procurement savings [83]. By treating supplier stability as a treatment and employing a forest of causal trees, individual and aggregate treatment effects were recovered, revealing that more stable supplier groups experienced markedly different savings patterns. The analysis found subtle signs of collusion, which guided targeted anti-corruption actions in the e-procurement platform. However, if there are unobserved factors that affect both supplier stability and procurement savings, and they aren't captured by the measured covariates, the estimated effects can be biased.

## 2.2 Image Generation

### 2.2.1 Image Generation Models

Image generation models are a class of generative algorithms designed to synthesize realistic images from abstract representations. Early approaches relied on variational autoencoders (VAEs) to learn compact latent spaces from which images could be decoded. Generative adversarial networks (GANs) then advanced the field by training paired generator and discriminator networks in a minimax game, producing sharper and more detailed outputs. More recently, diffusion models have achieved state-of-the-art fidelity by iteratively denoising random noise into coherent images through learned reverse-process transitions. Transformer-based architectures have further expanded capabilities by modeling image pixels or latent codes autoregressively or via attention mechanisms, enabling controllable generation and high scalability. Together, these methods have driven rapid progress in tasks ranging from art creation and data augmentation to conditional synthesis and interactive editing.

A GAN is composed of two neural networks, a generator and a discriminator [84]. They are trained simultaneously in opposition. The generator learns to map random input vectors to synthetic images. The discriminator learns to distinguish those synthetic images from real samples. Training proceeds as a minimax game. The generator improves by trying to fool the discriminator. The discriminator improves by spotting ever subtler artifacts in generated images. Early convolutional variants such as Deep Convolutional GAN (DCGAN) introduced architectural guidelines to stabilize training, including replacing pooling with strided convolutions and using batch normalization. Despite these advances, GANs remain sensitive to hyperparameter choices and can suffer from mode collapse, where the generator produces limited image diversity. Subsequent improvements, such as Wasserstein GANs and hinge-loss formulations, address convergence issues by changing the loss landscape.

In [85], synthetic data for energy systems was generated using both conditional GAN (cGAN) and Wasserstein GAN (wGAN) to model critical heat flux phenomena and power-grid demand forecast scenarios. It was found that wGAN offered greater robustness to missing input features and better generalization to unseen experimental conditions. wGANs improve training stability by using the Earth Mover distance. However, enforcing the required Lipschitz constraint originally via weight clipping, can severely limit model capacity and demands careful hyperparameter tuning. Gradient-penalty variants alleviate this limitation. However, they introduce substantial computational overhead at each training step. Furthermore, an infinite high-fidelity cloud generated method (IHFCGen) was proposed to construct a massive "cloudy & cloud-free" pair dataset [86]. This method demonstrated how coupling physical scattering laws with a GAN framework can generate high-resolution synthetic thin-cloud imagery for Earth-observation applications. Nevertheless, IHFCGen relies on an assumed scattering law whose accuracy can vary with actual atmospheric properties, and mismatches between the model and real scattering behavior can lead to visible synthesis artifacts.

Moreover, a deep convolutional conditional GAN was applied to rotating-machinery fault diagnosis with severely imbalanced datasets [87]. By augmenting minority fault samples, the method boosted both training efficiency and generalization performance in real-world imbalanced fault scenarios. While the synthetic minority samples may fail to capture the full diversity of real fault signatures, leading to mode collapse and biased diagnosis.

A VAE consists of two paired neural networks: an encoder that maps each input into the parameters of a latent probability distribution, and a decoder that reconstructs the input by sampling from that distribution [88]. Training optimizes a variational lower bound on the data likelihood, which combines a reconstruction term measuring how well the decoder reproduces the inputs and a regularization term that encourages the latent distribution to match a chosen prior. The reparameterization trick is used to enable backpropagation through the stochastic sampling step by expressing samples as deterministic functions of the encoder outputs plus noise. As a result, VAEs learn a

smooth, continuous latent space that supports interpolation between points and allows for straightforward generative sampling. Model capacity and generative quality are influenced by architectural choices such as the dimensionality of the latent space, the depth and width of the encoder/decoder networks, and the relative weighting of the reconstruction versus regularization terms.

Recently, a knowledge-sharing and correlation-weighting VAE (KSCW-VAE) was developed for concurrent fault detection in manufacturing processes [89]. This model embeds process-quality variables alongside operating data to learn an interpretable latent space that highlights deviations indicative of faults. However, the dual embedding of process-quality variables and operating data increases model complexity, requiring larger training sets to avoid overfitting. Besides, a VAE-parameterized estimator, combining a generative model and classical estimation theory, was proposed to model unknown signal distributions as conditionally Gaussian and parameterize a linear minimum mean squared error estimator [90]. This approach obtained accurate conditional first and second moment estimates from noisy observations, improving estimation performance over conventional methods. While this framework assumes that the true data distribution is conditionally Gaussian given the latent variables, which may be violated in real-world signals and thus limits estimator accuracy.

In addition, the Deep Variational AutoEncoder-Based Support Vector Data Description with Adversarial Learning (DVAA-SVDD) was applied to power-battery systems for anomaly detection [91]. By guiding the SVDD boundary with VAE-extracted features and Gaussian reconstruction losses, the method achieved robust detection of rare fault modes under varying charge–discharge conditions. However, the Gaussian reconstruction loss may not capture the full complexity of battery signal distributions, causing subtle anomalies to be missed.

A diffusion model is a generative framework that learns to produce complex data by reversing a gradual noising process [92]. The forward process corrupts each example by adding small amounts of Gaussian noise over many steps until only noise remains. A neural network is then trained to perform the reverse process, removing noise step by step to recover the original data. This network is typically based on a U-shaped architecture with skip connections and includes embeddings that indicate the current noise level. Training optimises a simplified objective that encourages accurate noise prediction at each timestep. At generation time, sampling begins from pure noise and repeatedly applies the learned denoiser to produce a clean sample. Class conditional outputs can be obtained by guiding the denoising steps with extra signals from a classifier or by using a classifier-free approach.

In recent engineering works, a data-driven structural generative design method employed a diffusion model to propose component geometries that satisfy mechanical loading constraints [93]. By learning to sample from the distribution of valid designs, the approach sped up the exploration of design alternatives in computer-aided engineering workflows. However, the iterative denoising used for sampling is computationally intensive, which limits real-time or interactive design exploration. Besides, the inverse problem of analog integrated-circuit sizing was tackled using denoising diffusion probabilistic models [94]. Trained to gradually denoise random inputs into valid circuit parameter sets, the models achieved high-accuracy sizing under scarce training data, outperforming conventional optimization techniques. However, the denoising diffusion approach to analog circuit sizing relies on an iterative reverse-noise process that requires tens to hundreds of network evaluations per sample, making the generation of each candidate parameter set computationally expensive and slow.

Transformer-based image generation architectures replace convolutional layers with layers of multi-head self-attention and feed-forward networks to model long-range dependencies across an image [95]. They begin by splitting an image into a sequence of patches or tokens, each augmented with positional embeddings to retain spatial information. Stacked Transformer blocks then use multi-head attention to allow every token to attend dynamically to all others, capturing global context that is difficult for convolutional filters to learn.

Recently, Defect Transformer (DefT), an efficient hybrid Convolutional Neural Network (CNN), was proposed to capture both local patterns and global context for detecting surface defects on steel and composite parts [96]. DefT outperformed pure-CNN baselines on multiple industrial datasets while reducing inference time by over 30 %. However, Defect Transformer introduces additional architectural modules that increase model size and memory usage. Deploying this model on resource-constrained devices can therefore be challenging. Moreover, Diagnosisformer, a lightweight Vision Transformer–based architecture, was applied to vibration-signal representations of rolling bearings operating under harsh noise and load variations [87]. It automatically fused multi-scale frequency–domain features and achieved up to 98 % classification accuracy on public and proprietary bearing datasets. Nevertheless, Diagnosisformer relies on Fast Fourier Transform (FFT)-based time–frequency preprocessing and a full Transformer backbone, which incurs high computational and memory costs that can hinder real-time or edge-device deployment.

### 2.2.2 Interpretability in Image Generation

Interpretability in image-generation models aims to explain the hidden steps by which architectures like GANs, VAEs, and diffusion models transform learned representations into coherent visuals. It involves tracing how specific latent dimensions, activations, or training examples influence textures, shapes, and semantic content [97]. Techniques such as activation maximization, concept attribution, and latent-space probing enable

practitioners to diagnose biases and uncover failure modes [98]. Ultimately, understanding the "why" behind a model's creative choices deepens our theoretical grasp of generative mechanisms and fosters safer, more transparent deployment in real-world applications [99].

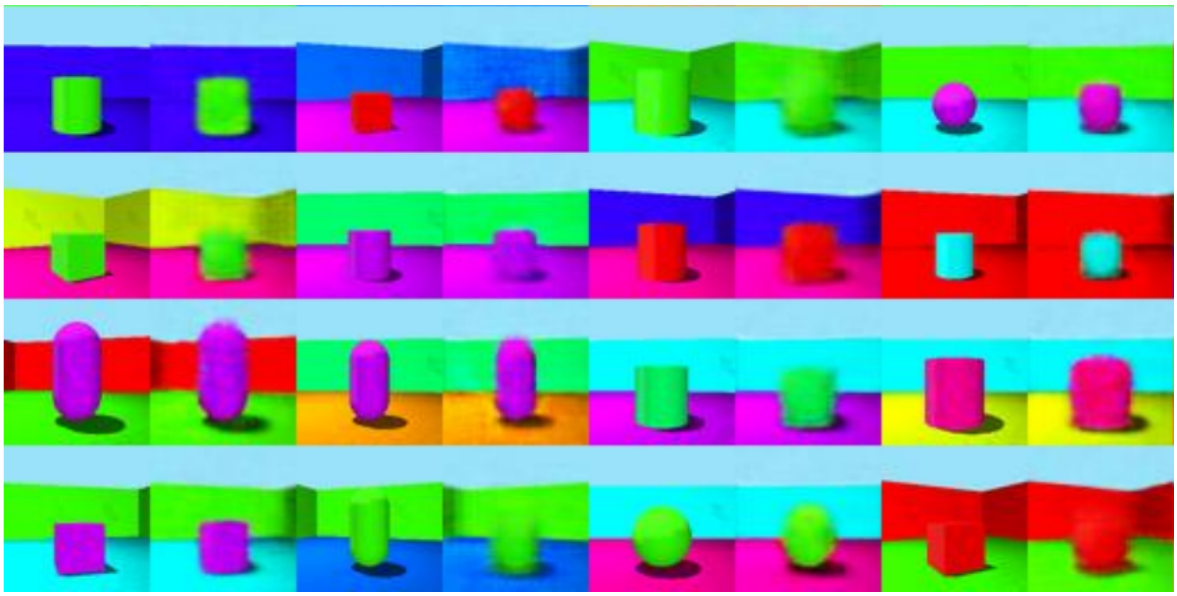


Figure 2.7: An example of what the VAE trained with disentanglement learning could generate [100].

To learn representations from real-world data, disentanglement learning is one of the most popular methods [100]. Figure. 2.7 shows the images generated by the VAE trained with disentanglement learning. Each row corresponds to a fixed object identity and geometry, while variations across columns reflect controlled changes in underlying generative factors such as color, lighting, background, and viewpoint. Because each factor is manipulated independently, the figure illustrates how the model disentangles semantic attributes in its latent space rather than entangling them into a single appearance code. This figure indicates that the shape of the object and the different colors are learned by the VAE, which proves that interpretability is achieved by the disentanglement VAE. Moreover, multimodal inputs are decoupled into a mode-specific appearance code and a mode-invariant content code [101]. Then the codes are fused into a shared representation to achieve disentanglement. The superior performance of disentanglement is demonstrated when processing medical images with different modes by dividing image features into domain-invariant features (DIFs) and domain-specific features (DSFs) [102]. An unsupervised learning method is proposed to obtain disentangled represent-

ations [103]. The obtained disentangled representations are used to handle the conversion between cross-modal medical images, such as computed tomography (CT) and magnetic resonance imaging (MRI). This unsupervised learning method enables efficient adaptation between different medical imaging modes by splitting images into a shared, domain-independent content space and a domain-specific style space. A disentangling controllable generation method, disentangled controllable dialogue generation model (DCG), is proposed in [104]. The DCG method learns to attribute concepts from observable values to unobservable combinations through shared mappings of attribute-oriented prompts. The DCG method also uses disentangling losses to separate different attribute combinations.

Among these interpretability strategies, disentanglement learning restructures the latent space. Each dimension captures one meaningful factor of variation on its own, making the model more transparent and easier to control. Disentanglement learning aims to uncover latent representations where each dimension corresponds to a single, interpretable factor of variation in the data. In most deep generative models, latent codes mix together attributes like object shape, lighting, and pose, which makes it hard to manipulate one aspect without affecting others [105]. Disentanglement techniques resolve this by enforcing statistical independence or by minimizing the total correlation among latent dimensions [106]. As a result, adjusting one latent coordinate produces a predictable change in the generated output while leaving all other features untouched, greatly improving clarity and controllability.

Early approaches, such as the $\beta$-Variational Autoencoder ($\beta$-VAE), achieved disentanglement by up-weighting the Kullback–Leibler term in the variational objective, penalizing deviations from a factorized prior. Subsequent methods, such as FactorVAE, DIP-VAE and InfoGAN introduced regularizers or adversarial critics to directly minimize dependencies among codes. In each case, the learned representations afford interpretability and engineers can visualize and manipulate individual factors. This capability facilitates efficient "what-if" reasoning, since interventions on a single latent coordinate correspond to real-world manipulations of the associated attribute.

$\beta$-VAE extends the standard VAE by introducing a hyperparameter $\beta$ to the training objective that scales the Kullback–Leibler divergence term [107]. The encoder and decoder architectures remain unchanged. By setting $\beta > 1$, the latent posterior is pushed closer to the prior. This forces each dimension of the latent space to capture a distinct factor of variation. Training still uses the reparameterization trick for backpropagation through stochastic sampling. As $\beta$ increases, the model discovers interpretable features such as object shape, position, or color without supervision. Overly large values of $\beta$ can sacrifice reconstruction fidelity and even collapse the latent code. Balancing $\beta$ therefore requires careful tuning to achieve the desired trade-off between disentanglement and image quality. However, high values of $\beta$ degrade reconstruction fidelity, often producing overly smooth or blurry outputs. Moreover, extensive tuning of the $\beta$ hyperparameter is required because the optimal trade-off between disentanglement and reconstruction quality is dataset-specific.

Later approaches, like FactorVAE, DIP-VAE and InfoGAN, increased the disentanglement capability compared with $\beta$-VAE. FactorVAE extends $\beta$-VAE by explicitly penalizing the total correlation of the latent code [108]. It introduces an auxiliary discriminator network that learns to distinguish samples drawn jointly from the encoder's aggregated posterior from samples drawn independently from each marginal. The discriminator's classification error provides an estimate of total correlation, which is added as a penalty to the standard VAE objective alongside reconstruction and KL-divergence terms. Training alternates between updating the VAE to both minimize reconstruction loss and fool the discriminator, and updating the discriminator to better distinguish joint from marginal samples. This adversarial procedure encourages the latent dimensions to become statistically independent, resulting in more robust disentanglement and an improved balance between reconstruction fidelity and factorization compared to $\beta$-VAE. However, the adversarial total-correlation penalty requires training an auxiliary discriminator network, which adds complexity and can introduce the same instability and convergence issues seen in GAN training.

Furthermore, DIP-VAE extends the standard VAE by adding a regularization term that matches moments of the model's aggregated posterior to those of the factorized prior [109]. Two variants are defined: DIP-VAE-I penalizes only the off-diagonal entries of the aggregated posterior's covariance, while DIP-VAE-II penalizes both off-diagonal and variance deviations from the prior. The added moment-matching penalty is incorporated into the evidence lower bound alongside reconstruction and KL-divergence terms. Training uses the reparameterization trick to enable gradient backpropagation through stochastic latent sampling. This approach produces more disentangled and statistically independent latent factors than $\beta$-VAE without requiring adversarial networks, all while preserving data likelihood quality. Nevertheless, DIP-VAE only makes sure that, overall, each latent dimension has the right average value and spread. It doesn't force those dimensions to be fully independent, so more subtle links between them can stay hidden.

Moreover, InfoGAN extends the standard GAN framework by decomposing the latent input into an incompressible noise vector and a structured code, then maximizing the mutual information between that code and the generated outputs [110]. An auxiliary network Q is introduced to approximate the posterior of the structured code given a generated sample, and a variational lower bound on mutual information is optimized alongside the usual adversarial loss. Training alternates between updating the generator and Q to both fool the discriminator and maximize mutual information, and updating the discriminator to better distinguish real from fake samples. This encourages the generator to produce outputs whose salient features, such as digit style, pose, or facial attributes, correspond to interpretable dimensions of the structured code. Because Q can share most of its weights with the discriminator, the additional computational overhead is modest. Still, training InfoGAN can be unstable due to its adversarial setup, which may lead to mode collapse, where the generator produces limited variations of outputs.

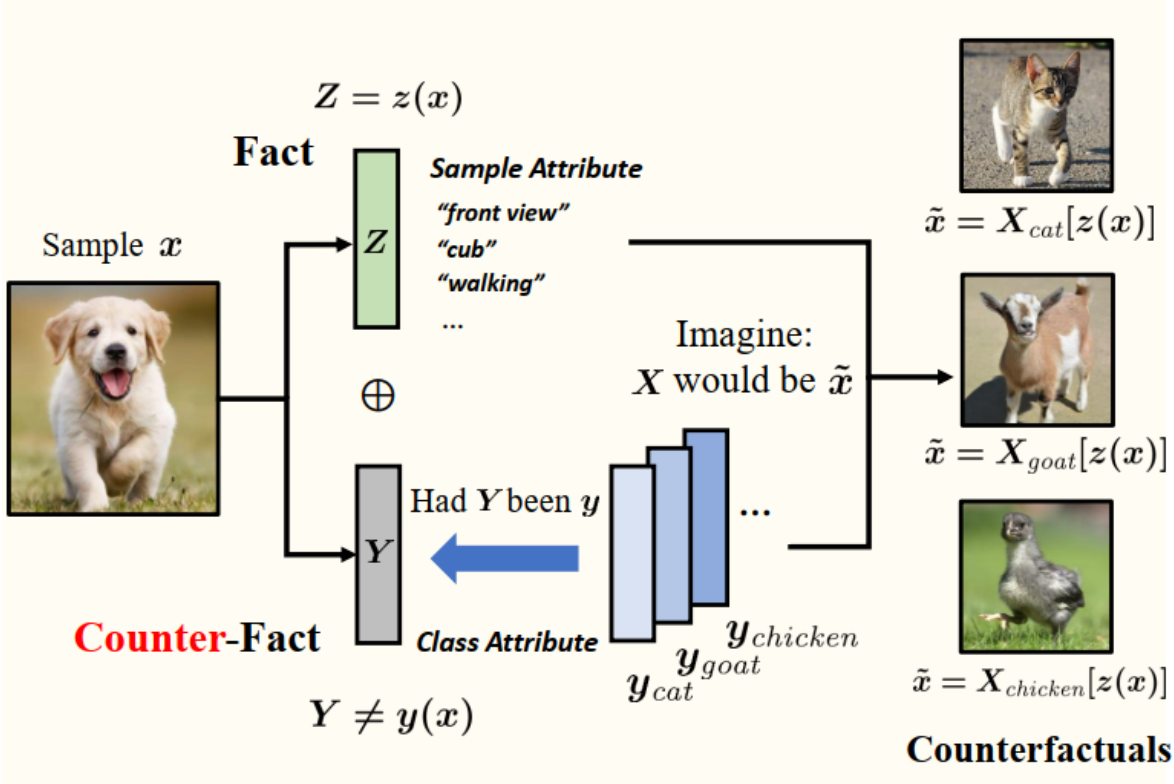



Figure 2.8: An example of what images the causal generative model will generate [111].

In the causal generative model, disentanglement learning is adopted as a key mechanism for embedding causal semantics into generative architectures. Figure. 2.8 shows an example that what images the causal generative models produce [111]. In this case, the dog in the original image is replaced by a cat, a cow and a bird by doing an intervention. By enforcing that latent dimensions correspond to independent causal factors, such as lighting direction, object geometry, and material reflectance, the causal generative model can be trained to produce both interpretable codes and robust, intervention-capable synthesis. This alignment between disentangled representations and causal mechanisms lays the groundwork for systematic generation of shadow training data and interventional grasp planning in robotic systems. Causal inference in generative modeling has emerged as a promising solution to address interpretability and counterfactual reasoning. SCMs and related DAG-based frameworks enable the explicit modeling of causal relationships among latent factors [112]. In recent works, VAEs are combined with masked causal layers that learn a latent adjacency matrix for disentangling exogenous factors, resulting in more interpretable embeddings, though at the cost of sample quality inherent to standard VAE decoders [113].

Diffusion-based Models, on the other hand, focus on photorealistic image synthesis by iteratively reversing a noise corruption process [114, 115]. Although they have achieved high fidelity, general diffusion architectures lack explicit low-dimensional latents. They also typically require hundreds to thousands of denoising steps, making them computationally expensive at inference time and less amenable to direct causal manipulation [116]. Initial attempts to bridge causal inference and diffusion, such as Diff-SCM [117] and the Causal Diffusion Autoencoder (CDAE) [118], integrate causal constraints into the diffusion framework. Moreover, Diff-SCM operates directly in the high-dimensional diffusion space and uses gradient-based updates for both reasoning and intervention. However, Diff-SCM does not learn an explicit low-dimensional latent, making direct manipulation impossible. CDAE uses the Autoencoder (AE) to address the challenge that diffusion architectures lack explicit low-dimensional latents. AEs, which lack the Kullback–Leibler (KL) divergence regularization term found in VAEs, can "memorize" the training set and produce blurry or implausible reconstructions on novel inputs. In contrast, VAEs trade some reconstruction fidelity for a smoother, more robust latent embedding [119]. A more recent method, CausalFusion [120], integrates causal inference into Diffusion Transformers. However, the model only applies causal structure when generating samples and ignores any causal relationships between its internal representations. Furthermore, CCDiff introduces a compositional diffusion framework for closed-loop traffic scenario generation by embedding causal constraints into each denoising step [121]. Nevertheless, CCDiff assumes a fixed causal graph, which limits its ability to adapt to new or evolving causal structures.

## 2.3 Large Language Models

### 2.3.1 General Large Language Models

LLMs are a type of artificial intelligence model designed to understand and generate human language [122]. They are trained on vast amounts of text data and can perform a wide range of tasks, including text summarization, translation, sentiment analysis, and more. LLMs are characterized by their enormous scale, often containing tens of billions of parameters, enabling them to learn the complex patterns present in linguistic data. These models are typically based on deep learning architectures such as Transformers, allowing them to achieve impressive performance across various Natural Language Processing (NLP) tasks. LLMs, typically built on the Transformer architecture, have revolutionized natural language processing by learning to predict the next token in massive text corpora through self-supervised pretraining. Models such as GPT, BERT, and the following models encode rich, contextualized representations of semantics and world knowledge in their latent vectors, enabling tasks from free-form text completion to code synthesis and dialogue. LLMs can be adapted to a specific domain by fine-tuning on task data, guiding them with prompts, or using reinforcement learning from human feedback. This lets them produce high-quality text and structured outputs with very little extra supervision.

In engineering, LLMs have been applied to automate the extraction of synthesis protocols, propose novel reaction pathways, and predict material properties, which significantly accelerates the materials discovery cycle. GPT-4 and similar language models can read scientific literature to pull out detailed reaction pathways. They can suggest step-by-step plans for how to synthesize new materials. They can also scan and summarize vast amounts of papers in a fraction of the time it would take a human. Together, these capabilities have dramatically accelerated materials research and development in a way that feels like an industrial revolution for the field [123]. Moreover, LP-COMDA was proposed as a physics-informed LLM agent that chats with users to gather system

specifications. It then iteratively coordinates with optimization tools to automatically generate and refine power-converter modulation schemes. This approach reduced design errors by over 60 % and accelerated expert workflows by a factor of 30 [124]. In the circuit design, LayoutCopilot uses a multi-agent LLM framework to translate high-level natural-language layout intents into executable analog-design scripts, greatly simplifying the user interface for circuit designers [125].

### 2.3.2 Generation Models Based on Large Language Models

Recent advances have shown how large language models can be used for image synthesis. These approaches couple powerful text representations with state-of-the-art generative image engines [126]. First, the language model processes a natural-language prompt and generates a rich, context-sensitive embedding. Next, that embedding is used to condition a decoder network. The decoder may be a diffusion model or a transformer-based image generator. Finally, the conditioned network produces high-fidelity visuals that closely align with the user's instructions.

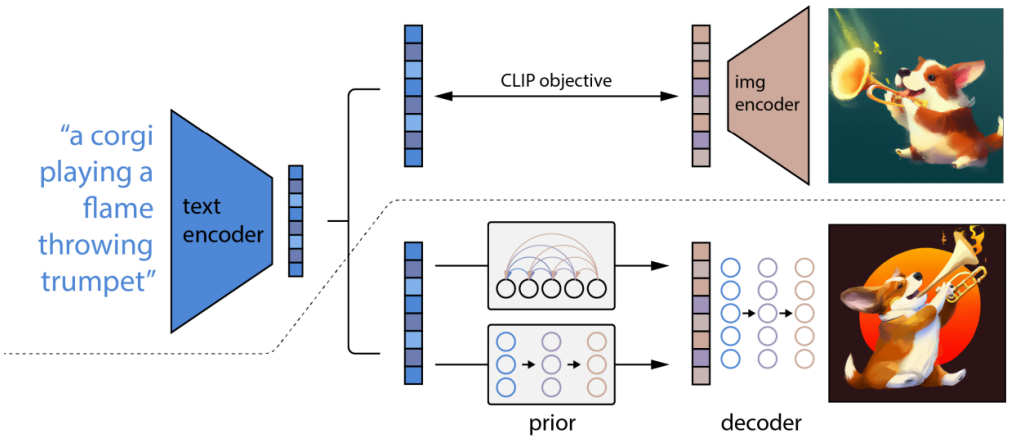



Figure 2.9: An overview of DALL · E 2 to generate images based on the human language [127].

Early examples include DALL · E [128] and CogView [129]. Figure. 2.9 shows the architecture of the DALL · E 2 to generate images based on the human language [127]. A natural-language prompt (e.g., "a corgi playing a flame-throwing trumpet") is first encoded by a text encoder into a semantic embedding. This embedding is aligned with visual representations produced by an image encoder through a CLIP objective, ensuring semantic consistency between text and images. These systems use transformer decoders to map discrete text tokens directly to image tokens. Coherent scenes are produced from textual descriptions. Rather than directly decoding images from text, the aligned representation is mapped into a latent space via a learned generative prior, which models the distribution of visual concepts consistent with the text. A decoder then samples from this latent space to synthesize high-fidelity images. The figure highlights how separating semantic alignment (CLIP) from image synthesis (prior + decoder) enables flexible, expressive image generation while maintaining strong correspondence between linguistic concepts and visual content. More recent systems, such as Imagen and Stable Diffusion, employ a two-stage approach. First, a language model or text encoder transforms the prompt into a semantic latent vector. Then, a diffusion model iteratively refines a noise map into a photorealistic image using that vector as guidance. Classifier-free guidance or cross-attention mechanisms ensure that the generated pixels reflect both the object content and the stylistic cues specified by the prompt.

These integrated LLM-to-Diffusion workflows are good at capturing abstract concepts and fine details, while enabling flexible control through prompt engineering or learned control nets. Using the understanding of linguistic structure and world knowledge, these systems can produce images with coherent composition, accurate object relationships, and nuanced styles, significantly advancing the state of text-to-image generation without relying on external causal frameworks. Text-to-image generation pipelines typically fall into two main categories. Autoregressive models such as DALL · E treat text and image tokens as a single sequence and use a large transformer to predict image tokens one by one from a prompt, after which a discrete VAE decoder reconstructs pixels from those tokens [128]. In diffusion-based systems like Imagen, a pretrained

language encoder transforms the text prompt into a rich semantic embedding, which then conditions a U-Net denoising network to iteratively refine random noise into a photorealistic image [130]. Both approaches often employ cross-attention or classifier-free guidance mechanisms to ensure that generated visuals faithfully reflect both the content and style specified in the prompt. Continuous innovations in noise schedules, sampling acceleration, and hybrid latent-code architectures have further improved fidelity, diversity, and computational efficiency in engineering applications.

LLMs have impressive generative capacity. However, they share many challenges with deep generative vision models. Latent representations often remain entangled. These models can produce hallucinated or semantically inconsistent outputs when given out-of-distribution prompts. Moreover, the implicit "knowledge" in an LLM's weights is difficult to inspect or intervene upon directly. This limitation restricts interpretability and causal reasoning. Integrating causal structure into LLM-driven text generation holds promise for improving robustness under distributional shift. For example, the causal LLM could decompose outputs into content, style, and factuality latents linked by a causal graph to enable counterfactual queries.

In recent research, the generation models based on LLM were used in engineering. In conceptual architecture, Stable Diffusion was used to generate building massing and material renderings from textual briefs, demonstrating faster ideation cycles and more diverse form exploration compared with traditional shape-grammar workflows [131]. However, Stable Diffusion's initial $512 \times 512$ training resolution means that image quality often degrades noticeably when outputs are requested at higher or non-native sizes. Furthermore, a conditional latent diffusion model (CLDM) was applied to the structural design of a UAV reflector support, producing novel support-structure geometries that met specified load-bearing constraints with minimal manual iteration [93]. CLDM can generate structures that satisfy load-bearing constraints but may also propose overly complex shapes that are difficult or costly to manufacture. Vector-Quantized Computer-Aided Design CAD (VQ-CAD), a vector-quantized diffusion model, converted text-based parametric design intents directly into editable Computer-Aided Design

(CAD) geometries, streamlining the transition from concept to CAD model in standard engineering processing [132]. Although the design processing is simplified by VQ-CAD, the use of hierarchical vector-quantized codebooks can introduce quantization errors, leading to blocky or imprecise geometry that lacks fine-scale detail.

### 2.3.3 Generation Models Based on Multimodal Large Language Models

Recent advances have begun to leverage Multimodal Large Language Models (MLLMs) for end-to-end image synthesis by unifying language understanding and pixel generation within a single transformer framework. A key avenue for enhancing interpretability in MLLM-based image generators is to incorporate chain-of-thought reasoning directly into the synthesis pipeline [133]. Given a user prompt, the MLLM first produces an explicit sequence of intermediate "thoughts" or semantic steps, such as identifying scene entities, laying out spatial relationships, and selecting lighting and stylistic attributes, which are then grounded in successive decoding stages of a diffusion or autoregressive image model. This reasoning trace can be inspected and edited by practitioners to verify that each sub-task aligns with the intended prompt semantics and to pinpoint the origin of any anomalies [130]. Cross-attention heatmaps visualized alongside each thought further reveal how language tokens map to image regions at different stages, making it possible to diagnose whether errors arise from ambiguous prompt interpretation or latent embedding misalignments [95]. Chain-of-thought methods explain their steps in plain language. For image generation, chain-of-thought methods turn a black box into a readable, step-by-step process that builds trust and lets users control fine details of multimodal outputs.

Recent works combine image generators with MLLMs. MLLMs bring LLM-style reasoning together with the ability to receive, interpret, and generate outputs across multiple modalities, such as images, audio, and video.

MLLMs–based image generation models leverage a unified transformer backbone to jointly process text and visual signals, enabling the model to both understand enough prompts and synthesize corresponding images in a single end-to-end framework. During inference, an MLLM first encodes the user's text prompt [134]. It also encodes any optional visual context, such as a sketch or reference image. These inputs are fused into a single latent representation that captures semantics, spatial relationships, and style cues. This latent vector then conditions a generative decoder, often implemented via a diffusion or autoregressive image network, that iteratively refines pixel-level outputs to match the described scene [135]. By training on large-scale paired text–image corpora and employing cross-modal attention mechanisms, these models can generate coherent, high-fidelity images that faithfully reflect complex instructions all within a single transformer architecture without separate language and vision modules [136].

In Visual ChatGPT [137], a powerful LLM orchestrates multiple visual foundation models such as Stable Diffusion and inpainting modules. This setup allows users to generate, edit, and iteratively refine images via conversational prompts. Visual ChatGPT relies on GPT-4 to orchestrate separate visual foundation models rather than using a unified multimodal architecture. This prompt-manager approach makes the pipeline brittle and prevents end-to-end fine-tuning, so failures in individual modules can break the overall system.

Large Language and Vision Assistant (LLaVA) [138] extends a Vicuna-based LLM with a vision encoder to jointly perform multimodal comprehension and to condition a diffusion decoder, enabling both image understanding and generation from natural-language instructions. However, Multiple-image understanding is not supported, limiting its applicability to single-image tasks [139]. Moreover, training remains prolonged when fine-tuning on high-resolution images.

InstructPix2Pix [140] demonstrates how a frozen text–image diffusion backbone can be steered by an LLM into semantically accurate image edits. Free-form editing descriptions are translated into targeted pixel-space transformations. These MLLM-based pipelines excel at capturing complex scene semantics like “a misty mountain lake at sunrise” and detailed style cues such as “oil painting, impasto brushwork,” all while maintaining a flexible and dialogue-driven interface. However, they still treat generation as a largely black-box process. Furthermore, the training of the InstructPix2Pix relies on a large synthetic dataset generated from LAION captions, which is quite noisy and contains nonsensical or undescriptive instructions.

SmartEdit [141] demonstrates how MLLMs can be guided to perform highly structured, multi-step image edits from natural-language instructions. Complex revision requests, such as “replace the vase with a ceramic bowl, then adjust the lighting to simulate late afternoon sun”, are decomposed into a sequence of sub-operations. An LLM controller invokes specialized image-editing modules such as segmentation, inpainting, and tone mapping in a coherent pipeline, achieving robust adherence to long-horizon editing goals. Nevertheless, SmartEdit depends on accurate bidirectional interactions between the image and the language model, so any errors in the segmentation or not fully printed modules can influence and degrade overall edit quality.

In parallel, Image Content Generation with Causal Reasoning integrates explicit causal graphs of scene factors and material properties into a diffusion-based MLLM framework [142]. Each denoising step is conditioned on intervened causal latents via a do-operator module. The model generates photorealistic images from prompts such as “a red car parked under a streetlamp at dusk” and also provides counterfactual variants with guaranteed consistency to the underlying causal structure. Although the proposed model showed its potential capability in generating counterfactual images of an event, this model was only evaluated on the dataset ”Tom and Jerry”, its potential

capability in the real world hasn't been proven. Together, these advances illustrate how MLLM-driven image generation and editing can be enriched by both modular instruction decomposition and principled causal interventions, paving the way for more controllable and interpretable multimodal synthesis.

Multimodal LLM–based image generators integrate language understanding and image synthesis within a single architecture to turn natural-language prompts into high-fidelity visuals. Exposing intermediate reasoning steps and attention maps makes these pipelines more transparent and helps diagnose where errors occur. Many implementations still stitch together separate specialized modules under LLM control, which can be brittle and resist end-to-end optimization. More unified models promise deeper multimodal comprehension and smoother workflows but often trade off scalability and robustness. Embedding explicit causal structure into the generation process offers a route to truly controllable, intervention-capable synthesis, although broad real-world validation remains a challenge.

## 2.4 Summary

The review of causal reasoning and causal inference highlights both their conceptual strength and their practical limitations for autonomous systems. Classical causal discovery and inference methods provide principled tools for explaining failures, reasoning about interventions, and answering counterfactual queries, but they typically rely on strong assumptions, low-dimensional variables, and offline analysis. In high-dimensional visual domains, such as robotics and autonomous driving, many causal factors (e.g., illumination, occlusion, material properties) are latent, intertwined, and difficult to measure explicitly, making traditional causal pipelines unsuitable for real-time deployment. This gap motivates the approach adopted in this thesis: instead of discovering causal structure solely from observational data, causal mechanisms are embedded directly into generative models, enabling controllable, interpretable, and intervention-consistent counterfactual image synthesis.

The literature on image generative models demonstrates remarkable progress in visual fidelity and diversity through VAEs, GANs, diffusion models, and transformer-based generators. However, most existing image generation methods remain correlation-driven, with entangled latent representations that hinder reliable intervention and "what-if" reasoning. Disentanglement techniques improve interpretability but often sacrifice realism, while diffusion models achieve high fidelity at the cost of computational efficiency and explicit causal control. These limitations directly relate to the challenges of data scarcity and robustness under distribution shift identified in this thesis. To address them, the proposed models combine causal structure with diffusion-based generation, leveraging structured latent spaces that align with causal factors while retaining the photorealism and generalization capability required for autonomous perception tasks.

Finally, the review of large language models and multimodal generation systems shows how language-guided image synthesis enables flexible, high-level control over visual content. Nevertheless, LLM-based generators often operate as black boxes, with limited transparency regarding why a particular image is produced or how changes in prompts map to concrete visual interventions. Moreover, their reliance on implicit correlations can lead to hallucinations and brittle behavior in rare or safety-critical scenarios. This motivates the integration of causal representations with LLM-guided generation in this thesis, where language is used not only to specify desired outputs but also to structure causal interventions explicitly. By aligning linguistic instructions with causal variables in generative models, the proposed approach aims to support transparent data generation, systematic coverage of edge cases, and robust training of autonomous systems under realistic deployment conditions.

# Chapter 3

# Preliminaries and Problem Formulation

In this chapter, the relevant methods used in this thesis are introduced. A review is provided of the key theoretical tools that underpin the algorithmic approach. The central problem is then formalized by specifying its objective function, constraints and standing assumptions. Finally, the manner in which this formulation both generalizes and unifies existing approaches in the literature is discussed. Building on these foundations, the chapter concludes by formulating the counterfactual image generation problem addressed in this thesis, clarifying how the reviewed causal, generative, and multimodal models are integrated to overcome the limitations of existing methods.

## 3.1 Intervention and Counterfactual in Structural Causal Models

A Bayesian network is a fundamental tool for constructing Structural Causal Models (SCMs) [143]. It is a probabilistic graphical model defined over a Directed Acyclic Graph (DAG), wherein each node denotes a random variable and each directed edge encodes a conditional dependence. Although a Bayesian network itself does not prescribe causal semantics, its directed edges merely encode statistical dependencies rather than

causal relationships. This DAG structure, however, is later combined with causal meaning in SCMs and serves as their foundational graph. Moreover, the factorization of the joint probability distribution induced by a Bayesian network underlies the inferential procedures employed in causal analysis.

A DAG consists of a set of nodes connected by directed edges. Along each edge, the upstream node is called the *parent*, and the downstream node is the *child*. Under the usual Markov assumption for DAGs, each node is conditionally independent of its non-descendants given its parents [144]. By applying the chain rule of probability together with these conditional independencies and denoting $X_i$ as the $i_{th}$ random variable, the joint distribution $P$ over all nodes in the graph factorizes as

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\big(X_i \mid \mathrm{Pa}(X_i)\big) \tag{3.1}$$

where $\mathrm{Pa}(X_i)$ denotes the set of parents of $X_i$. This factorization forms the basis of probabilistic inference in Bayesian networks and is closely related to the structural decomposition employed in structural causal models.

Each DAG produces a unique joint distribution, but a given distribution may be compatible with multiple DAGs. For example, the bi-variate distribution $P(X_1, X_2)$ could be generated by either

$$X_1 \longrightarrow X_2 \quad \text{or} \quad X_2 \longrightarrow X_1,$$

These two graphs imply opposite causal directions. Such observational equivalence prevents a Bayesian network alone from identifying causal structure.

To introduce a DAG with causal semantics, the *do*–operator is introduced to represent an intervention. The notation

$$do(X_i = \text{constant})$$

is used to indicate that all incoming edges into $X_i$ are removed and $X_i$ is set deterministically to the constant. In the context of this thesis, each variable $X_i$ represents a semantically meaningful factor of variation underlying the image generation process rather than an individual pixel value. Concretely, these variables correspond to latent causal factors such as illumination conditions, object attributes, or shadow-related properties, which are not directly observed but give rise to the visual appearance of the image.

An intervention $do(X_i = \text{constant})$ models a counterfactual manipulation of a specific generative factor while holding all other factors fixed according to the causal graph. This formulation directly reflects the research objective of this thesis, to answer counterfactual questions of the form "how would the image change if a particular causal factor were altered, while all other factors remained unchanged?" By operating on such latent variables, the framework enables controlled and interpretable counterfactual image generation rather than unconstrained pixel-level editing.

Under this intervention, the joint distribution [145] is modified to

$$P\big(X_1,\ldots,X_n \mid do(X_i = \text{constant})\big) = \delta_{X_i,\text{constant}} \prod_{j\neq i} P\big(X_j \mid \text{Pa}(X_j)\big), \tag{3.2}$$

where $\delta_{X_i,\text{constant}}$ denotes the point-mass enforcing $X_i = \text{constant}$ and $\text{Pa}(X_j)$ are the original parents of $X_j$. By comparing these interventional distributions across different settings of the constant, observationally equivalent DAGs can be distinguished and the true causal structure recovered.

SCM is commonly used in causal learning, and is defined with a triple $\langle U, V, E\rangle$ [65]. Here, $U$ represents a set of exogenous variables determined by factors outside the model; $V$ indicates the system's internal variables; $E$ represents the unexplained variation in an endogenous variable not captured by the model's deterministic part. Model intervention is to convert an original image to a counterfactual image based on a causal graph. Normally, causal inference is performed through interventions, typically represented by

setting a variable to a fixed value regardless of its natural state. This is commonly expressed using the *do*-operator, which implies actively setting $X$ to a constant and observing the resultant changes in other variables [66]. The intervention aims to identify how changes in one variable (the cause) directly affect another variable (the effect). This process helps in answering counterfactual questions, such as "What would happen if we had changed $X$?"

In structural causal models, the structural equation is treated as a foundational concept. An unobserved exogenous variable $u_i$ (often termed an *omitted factor*) is associated with each endogenous variable $X_i$. The value of $X_i$ is uniquely specified by a function $f_i(\cdot)$ of its parents $\mathrm{Pa}(X_i)$ and the corresponding exogenous variable $u_i$ [146]:

$$X_i = f_i\big(\mathrm{Pa}(X_i), u_i\big) \tag{3.3}$$

Disturbances are introduced into the system by these exogenous variables, shown in Figure 3.1. The exogenous variables are unobserved and may be stochastic. No assumptions are imposed on the mechanism that generates their variation.

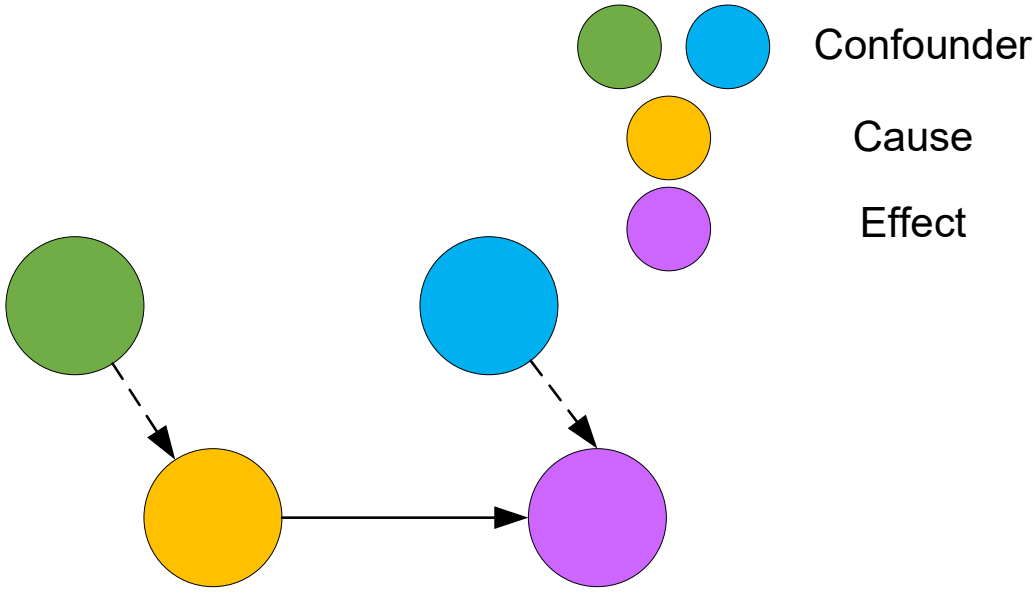



Figure 3.1: Causal relationship with confounders.

Unlike algebraic equations, structural equations encode the generative mechanisms of the variables, which assign values strictly from the right-hand side to the left-hand side and cannot be algebraically rearranged. Each exogenous variable captures all stochastic influences on its corresponding endogenous variable and is endowed with a fixed probability distribution. These exogenous factors are typically neither observed nor manipulable, and, within a structural causal model, they are assumed to be mutually independent.

## 3.2 Causal Inference in Machine Learning

In recent years, causal inference has been applied extensively to observational studies in areas such as epidemiology and economics. It has also been integrated into machine learning frameworks to enhance model robustness and to facilitate invariant representation learning across changing environments. Causal inference further enables counterfactual reasoning within predictive pipelines.

A mask causal layer is proposed to achieve causal disentanglement in [147]. Starting from the structural equation for the latent variables $\mathbf{z}$:

$$\mathbf{z} = \mathbf{A}^{\mathbb{T}}\mathbf{z} + \boldsymbol{\varepsilon}, \tag{3.4}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$ is an independent Gaussian exogenous factor and $\mathbf{A}$ is the adjacency matrix of a DAG, where $I$ is the identity matrix and $\mathbb{T}$ denotes the transpose of a matrix. Collect the $\mathbf{z}$ -terms on the left:

$$\mathbf{z} - \mathbf{A}^{\mathbb{T}}\mathbf{z} = \boldsymbol{\varepsilon} \quad \Longleftrightarrow \quad \left(I - \mathbf{A}^{\mathbb{T}}\right)\mathbf{z} = \boldsymbol{\varepsilon} \tag{3.5}$$

This formulation relies on the assumption that the adjacency matrix $\mathbf{A}$ corresponds to a directed acyclic graph. The acyclicity assumption implies that the latent causal variables admit a recursive ordering, such that no variable is an ancestor of itself through a directed path. Conceptually, this reflects the causal modeling principle that each latent factor is generated by its direct causes and an independent exogenous disturbance, rather than through instantaneous feedback loops.

From a mathematical perspective, acyclicity ensures that the matrix $I - \mathbf{A}^{\mathbb{T}}$ is invertible. This property is required to obtain a closed-form solution for the latent variables and to express them as a function of independent noise sources. Without this assumption, the system of structural equations may be ill-posed or admit multiple solutions, making causal interpretation and intervention unclear.

Under the acyclicity assumption on $\mathbf{A}$, the matrix $I - \mathbf{A}^{\mathbb{T}}$ is invertible, obtaining

$$\mathbf{z} = \left(I - \mathbf{A}^{\mathbb{T}}\right)^{-1} \boldsymbol{\varepsilon} \tag{3.6}$$

Equivalently, the inverse could be expanded as a Neumann series,

$$(I - \mathbf{A}^{\mathbb{T}})^{-1} = \sum_{k=0}^{\infty} (\mathbf{A}^{\mathbb{T}})^k, \tag{3.7}$$

so that

$$\mathbf{z} = \sum_{k=0}^{\infty} (\mathbf{A}^{\mathbb{T}})^k \boldsymbol{\varepsilon}, \tag{3.8}$$

which makes explicit how each latent component aggregates noise from its $k$-step ancestors in the causal graph.

The causal layer is to learn causal latent variables $\mathbf{z}$ with $\mathbf{A}$, which is obtained from a causal graph:

$$\mathbf{z} = \mathbf{A}^{\mathbb{T}}\mathbf{z} + \boldsymbol{\varepsilon} = (I - \mathbf{A}^{\mathbb{T}})^{-1}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I) \tag{3.9}$$

This masked causal layer enables the integration of causal inference with generative models.

An invariant and robust learning method is proposed to exploit invariances across multiple environments [148]. It assumes that, although the marginal distributions of inputs and outputs may vary, the underlying causal mechanisms that generate the target from its direct causes remain stable. By leveraging data from multiple training environments, this approach seeks to isolate features whose predictive power persists under arbitrary shifts in covariate distributions.

A prototypical approach is Invariant Risk Minimization (IRM) [149]. In IRM, it seeks a data representation $\Phi$ mapping from $\mathscr{X}$ to $\mathscr{Z}$ and a classifier $w$ mapping from $\mathscr{Z}$ to $\mathscr{Y}$ such that, for every environment $e \in \mathscr{E}$, $w \circ \Phi$ minimizes the empirical risk

$$R^e(w, \Phi) = \mathbb{E}_{(x,y)\sim P^e}\big[\ell\big(w(\Phi(x)), y\big)\big], \tag{3.10}$$

where $x \in \mathscr{X}$ is the image input, $R^e(w, \Phi)$ denotes the expected loss in environment $e$, and $\ell$ is the point-wise loss. Concretely, the IRM objective is

$$\min_{\Phi, w} \sum_{e\in\mathscr{E}} R^e(w, \Phi) \quad \text{subject to} \quad w \in \arg\min_{\tilde{w}} R^e(\tilde{w}, \Phi) \;\; \forall e \in \mathscr{E}, \tag{3.11}$$

which enforces that the same classifier $w$ is optimal across environments. In practice, this bi-level problem is relaxed by penalizing the stationarity of $w$ at a fixed $w_0$, obtaining the "IRM-v1" variant via a gradient-norm penalty.

Beyond IRM, Distributionally Robust Optimization (DRO) casts robustness as worst-case risk minimization over an uncertainty set [150]:

$$\min_{\theta} \max_{P\in\mathscr{U}} \mathbb{E}_{(x,y)\sim P}\big[\ell(\theta; x, y)\big], \tag{3.12}$$

where $\theta$ denotes model parameters and $\mathscr{U}$ is a family of distributions capturing potential shifts away from the empirical training distribution.

In this thesis, DRO is not adopted as a primary optimization objective. Instead, it is introduced to situate the proposed causal generative framework within the broader literature on robustness under distributional shift. DRO formalizes robustness through worst-case risk minimization, whereas the approach developed in this thesis addresses robustness by explicitly modeling causal mechanisms and performing interventions on latent causal factors. By reasoning about how images change under controlled interventions, the proposed method aims to achieve robustness to spurious correlations, such as those induced by illumination or shadow variations, without relying on adversarial distributional uncertainty sets.

## 3.3 Variational Autoencoders

VAEs are deep latent–variable models that learn a generative mapping from a simple prior over a low-dimensional code to the data space, using an approximate inference network and the evidence lower bound (ELBO) objective [151].

### 3.3.1 Structure of the Autoencoders

An autoencoder, shown in Figure. 3.2, is an unsupervised neural network architecture that seeks to learn a compact representation of input data and then reconstruct the original inputs from this representation via a decoder. An autoencoder consists of two components: an encoder network, which maps the input to its low-dimensional code, and a decoder network, which maps the code back to the input space.

A standard autoencoder consists of an encoder that maps the input image $\mathbf{x} \in \mathbb{R}^{d_x}$ to a low-dimensional code $\mathbf{c} \in \mathbb{R}^{d_c}$ and a decoder that reconstructs $\mathbf{x}$ from $\mathbf{c}$. Hidden widths $h_1, h_2 \in \mathbb{N}$ specify the sizes of intermediate layers. The following formulations represent the networks in the autoencoder [152].

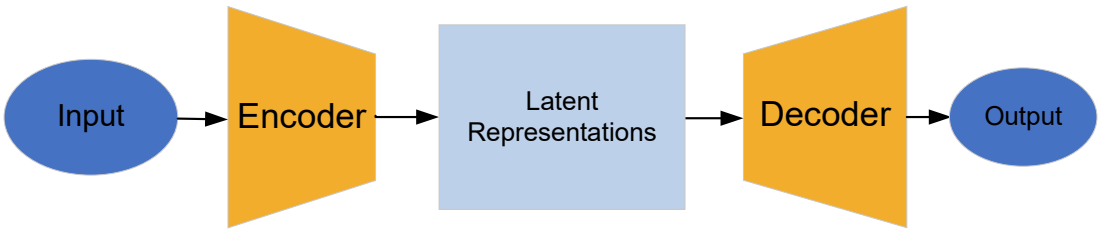



Figure 3.2: Structure of the conventional Autoencoder

$$\mathbf{h}^{(1)} = \mathrm{ReLU}\big(W^{(1)}\mathbf{x} + b^{(1)}\big), \qquad W^{(1)} \in \mathbb{R}^{h_1 \times d_x},\ b^{(1)} \in \mathbb{R}^{h_1},\ \mathbf{h}^{(1)} \in \mathbb{R}^{h_1} \tag{3.13}$$

where $\mathbf{h}$ is the output of the hidden layer $W$ is the weight of the input image and $b$ is the bias.

The Rectified Linear Unit (ReLU) is defined pointwise by [153]

$$\mathrm{ReLU}(s) = \max\big(0, s\big), \tag{3.14}$$

where negative inputs are rectified to zero and positive inputs pass through unchanged.

$$\mathbf{h}^{(2)} = \mathrm{ReLU}\big(W^{(2)}\mathbf{h}^{(1)} + b^{(2)}\big), \qquad W^{(2)} \in \mathbb{R}^{h_2 \times h_1},\ b^{(2)} \in \mathbb{R}^{h_2},\ \mathbf{h}^{(2)} \in \mathbb{R}^{h_2} \tag{3.15}$$

$$\mathbf{c} = W^{(3)}\mathbf{h}^{(2)} + b^{(3)}, \qquad W^{(3)} \in \mathbb{R}^{d_c \times h_2},\ b^{(3)} \in \mathbb{R}^{d_c},\ \mathbf{c} \in \mathbb{R}^{d_c} \tag{3.16}$$

$$\mathbf{d}^{(1)} = \mathrm{ReLU}\big(W^{(4)}\mathbf{c} + b^{(4)}\big), \qquad W^{(4)} \in \mathbb{R}^{h_2 \times d_c},\ b^{(4)} \in \mathbb{R}^{h_2},\ \mathbf{d}^{(1)} \in \mathbb{R}^{h_2} \tag{3.17}$$

$$\mathbf{d}^{(2)} = \mathrm{ReLU}\big(W^{(5)}\mathbf{d}^{(1)} + b^{(5)}\big), \qquad W^{(5)} \in \mathbb{R}^{h_1 \times h_2},\ b^{(5)} \in \mathbb{R}^{h_1},\ \mathbf{d}^{(2)} \in \mathbb{R}^{h_1} \tag{3.18}$$

$$\widehat{\mathbf{x}} = \mathrm{sigmoid}\big(W^{(6)}\mathbf{d}^{(2)} + b^{(6)}\big), \qquad W^{(6)} \in \mathbb{R}^{d_x \times h_1},\ b^{(6)} \in \mathbb{R}^{d_x},\ \widehat{\mathbf{x}} \in \mathbb{R}^{d_x} \tag{3.19}$$

Here, $\mathbf{d}$ is the output of the decoder, $\widehat{\mathbf{x}}$ is the reconstructed image, $\mathrm{sigmoid}(\cdot)$ denotes the element-wise output activation used to produce the reconstruction $\widehat{\mathbf{x}}$.

### 3.3.2 Structure of the Variational Autoencoders

In an autoencoder, the latent representation is encoded as a deterministic vector. In contrast, in a variational autoencoder in Figure. 3.3, the latent variables are treated as stochastic and described by a conditional probability distribution. The encoder network is therefore employed to infer this distribution over the latent space given each input. In the original VAE formulation, a Gaussian distribution was assumed.

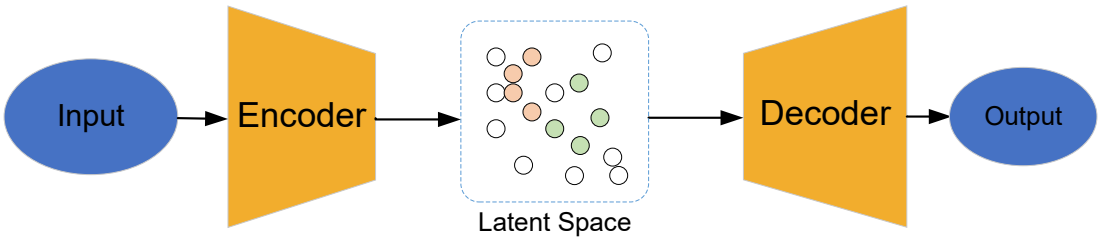



Figure 3.3: Structure of the conventional Variational Autoencoder

Meanwhile, the intermediate quantity is no longer referred to as a "latent variable" but is instead termed the "latent space". The variational autoencoder is composed of three main components: the encoder network, the stochastic latent layer, and the decoder network.

The encoder network is tasked with mapping each input example into the parameters of a latent distribution. In practice, this is implemented as a series of hidden layers, either fully connected or convolutional, depending on the data modality, with nonlinear activation functions. At its output, two parallel projection heads produce the mean and log-variance vectors that characterize a Gaussian distribution in the latent space.

The stochastic latent layer draws a sample from this Gaussian distribution using the reparameterization trick. This construction ensures that sampling remains a differentiable operation, allowing gradients to flow through the random draw back into the encoder parameters.

The decoder network mirrors the encoder's structure in reverse. It accepts a latent sample and passes it through successive hidden layers, again with nonlinear activations, to reconstruct the original input. The final layer outputs the parameters of the chosen likelihood.

Architectural enhancements, such as batch normalization, dropout, or residual connections, may be employed within both the encoder and decoder to stabilize training and improve generalization. The dimensionality of the latent space is typically set much lower than that of the input, creating a bottleneck that encourages the model to capture the most salient factors of variation.

During training, the entire network is optimized end-to-end by maximizing the evidence lower bound. This objective balances accurate reconstruction of the input against adherence of the inferred latent distribution to the prior.

### 3.3.3 Derivation of the Variational Autoencoders

The VAE specifies a prior, a likelihood, and an approximate posterior [154]:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I) \tag{3.20}$$

where $p(\mathbf{z})$ is a standard normal prior over the latent variables. The likelihood of the image given the latent is defined as

$$p_\theta(\mathbf{x} \mid \mathbf{z}) = f_\theta(\mathbf{z}) \tag{3.21}$$

where $\theta$ are parameters of the decoder network. $f_\theta$ is the decoder network that outputs the parameters of the likelihood for $\mathbf{x}$. The approximate posterior is defined as

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\big(\mu_\phi(\mathbf{x}), \operatorname{diag}(\sigma^2_\phi(\mathbf{x}))\big) \tag{3.22}$$

where $\phi$ are parameters of the encoder network. $\mu_\phi(\mathbf{x})$ and $\sigma_\phi(\mathbf{x})$ are the encoder network's outputs for the posterior mean and standard deviation. The marginal log-likelihood of the image is defined as

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x} \mid \mathbf{z})\, p(\mathbf{z})\, \mathrm{d}\mathbf{z} \tag{3.23}$$

where the marginal log-likelihood is intractable due to the integral over the latent space. By introducing the variational distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$ and applying Jensen's inequality, the lower bound is obtained as

$$\log p_\theta(\mathbf{x}) \;\geq\; \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{x} \mid \mathbf{z})\, p(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\, \mathrm{d}\mathbf{z} \tag{3.24}$$

where the inequality follows from Jensen's inequality after introducing the variational distribution. Overall, the loss function of the VAE is defined as

$$\mathscr{L}_{\mathrm{VAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\big] - \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big) \tag{3.25}$$

where the first term is the expected reconstruction log-likelihood and the second term is the Kullback–Leibler (KL) divergence regularizing the approximate posterior toward the prior.

## 3.4 Diffusion Models

The original Diffusion Model, Denoising Diffusion Probabilistic Model (DDPM), was designed to model the process of transforming structured data into noisy data by adding Gaussian noise [135]. Diffusion models, shown in Figure. 3.4, operate in two phases: a forward process incrementally adds Gaussian noise to an image until only noise remains, and a trainable reverse denoising process employs a neural network to iteratively remove noise from a pure-noise input, ultimately reconstructing a realistic image. In the figure, the black arrow indicates the forward process and the blue arrow represents the reverse denoising process.

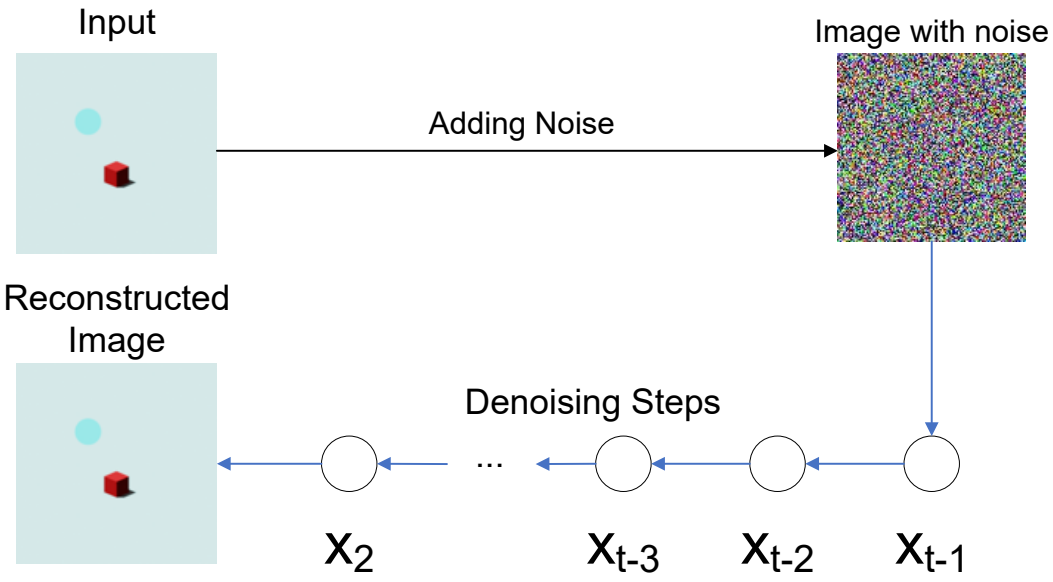



Figure 3.4: Structure of the conventional diffusion model

In the forward process, noise is added to the input image over a fixed number of discrete time steps according to a predefined noise schedule. At each step, a small amount of Gaussian noise is injected, ensuring that the data distribution is corrupted in a controlled and incremental manner. Early steps preserve most of the image structure, while later steps progressively destroy fine details, eventually transforming the image into an approximately isotropic Gaussian noise distribution. This gradual corruption avoids abrupt information loss and enables the reverse process to be learned effectively.

The reverse process is a trainable denoising procedure that starts from pure noise and iteratively removes noise step by step. At each denoising step, a neural network predicts how the current noisy image should be adjusted to recover a slightly less noisy version, conditioned on the diffusion timestep. By repeating this procedure across all timesteps, the model gradually reconstructs a clean and realistic image. Unlike the forward process, which is fixed and non-learnable, the reverse process is learned from data and captures the underlying data distribution through denoising.

In Figure. 3.4, the black arrows illustrate the forward noising trajectory from a clean image to noise, while the blue arrows depict the learned reverse trajectory that maps noise back to a realistic image. Together, these two processes enable diffusion models to combine stable training with high-quality image synthesis.

The forward diffusion process is defined as a Markov chain from $\mathbf{x}_0$ to $\mathbf{x}_T$:

$$q(\mathbf{x}_{t+1} \mid \mathbf{x}_t) = \mathcal{N}\big(\sqrt{1-\beta_{t+1}}\,\mathbf{x}_t,\ \beta_{t+1} I\big), \tag{3.26}$$

where each noisy step depends only on the previous state and $\beta_t$ is a fixed noise schedule. Here $t \in \{0, \ldots, T-1\}$ indicates diffusion step, $\mathbf{x}_t \in \mathbb{R}^{d_\mathbf{x}}$ is the image at time $t$, and $I = I_{d_\mathbf{x}}$ denotes the $d_\mathbf{x} \times d_\mathbf{x}$ identity.

The sequence $\{\beta_t\}_{t=1}^{T}$ is referred to as the noise schedule and controls the rate at which information is destroyed during the forward diffusion process. Each $\beta_t$ specifies the variance of the Gaussian noise injected at diffusion step $t$, thereby determining how much of the original image structure is preserved or corrupted at that step. Small values of $\beta_t$ result in gentle perturbations that retain most semantic content, while larger values accelerate the loss of fine details and drive the distribution toward pure noise.

The noise schedule is fixed in advance and is not learned during training. Its design plays a crucial role in stabilizing training and ensuring that the reverse denoising process is well-conditioned. By distributing the total noise injection across many small steps rather than a single large perturbation, the diffusion process avoids abrupt information loss and allows the neural network to learn a sequence of simpler denoising transformations. Common choices include linear or cosine schedules, which balance training stability with sample quality.

In this thesis, $\beta_t$ serves as a mechanism for defining a smooth interpolation between structured image data and an isotropic Gaussian noise distribution, providing a principled foundation for iterative generative modeling.

By marginalizing out intermediary steps, the distribution of $\mathbf{x}_t$ given the initial sample $\mathbf{x}_0$ admits a closed-form:

$$q\big(\mathbf{x}_t \mid \mathbf{x}_0\big) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1-\bar{\alpha}_t)\,I\big), \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s) \tag{3.27}$$

The quantity $\alpha_t = 1-\beta_t$ represents the fraction of signal retained at diffusion step $t$, while $\bar{\alpha}_t$ denotes the cumulative signal preservation after $t$ noising steps. Specifically, $\bar{\alpha}_t$ measures how much of the original image $\mathbf{x}_0$ remains in expectation within the noisy sample $\mathbf{x}_t$. As $t$ increases, the product structure of $\bar{\alpha}_t$ causes it to decay monotonically toward zero, indicating the progressive loss of information about the original image.

Introducing $\bar{\alpha}_t$ enables a closed-form expression for the distribution of $\mathbf{x}_t$ conditioned on $\mathbf{x}_0$, which greatly simplifies both training and sampling. This formulation allows the model to directly sample noisy images at arbitrary diffusion steps without explicitly simulating all intermediate transitions. Conceptually, $\bar{\alpha}_t$ defines a smooth interpolation between the data distribution at early timesteps and an approximately isotropic Gaussian distribution at later timesteps, providing a principled bridge between structured images and pure noise.

The reverse process seeks to recover $\mathbf{x}_0$ from $\mathbf{x}_T$. Although the true backward kernel

$$q\big(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\big) = \mathcal{N}\big(\mu_{\text{post}}(\mathbf{x}_t, \mathbf{x}_0), \Sigma_{\text{post}}(t)\big) \tag{3.28}$$

is analytically tractable, it cannot be used directly for sampling. $\mu_{\text{post}}(\mathbf{x}_t, \mathbf{x}_0)$ is the posterior mean of the exact one-step backward distribution and $\Sigma_{\text{post}}(t)$ is the posterior covariance of $q\big(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\big)$. Instead, a neural network

$$p_\theta\big(\mathbf{x}_{t-1} \mid \mathbf{x}_t\big) \tag{3.29}$$

is trained to approximate this posterior by minimizing

$$\begin{aligned}\mathrm{KL}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big) = \frac{1}{2}\Big[\log \frac{|\Sigma_\theta(t)|}{|\Sigma_{\text{post}}(t)|} - d + \mathrm{tr}\big(\Sigma_\theta(t)^{-1}\,\Sigma_{\text{post}}(t)\big) \\ + \big(\mu_{\text{post}}(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\big)^T\,\Sigma_\theta(t)^{-1}\,\big(\mu_{\text{post}}(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\big)\Big]\end{aligned} \tag{3.30}$$

where $\mu_\theta(\mathbf{x}_t, t)$ is the model's predicted mean of the reverse step and $\Sigma_\theta(t)$ is the model's covariance for that reverse step. Furthermore, $\mathrm{KL}(\cdot\|\cdot)$ denotes Kullback–Leibler divergence, $\mathrm{tr}(\cdot)$ is the trace, $|\cdot|$ is the determinant, $d$ is the data dimensionality in the Gaussian, and $\|\cdot\|_2$ is the Euclidean norm.

In the usual DDPM parameterization, it sets $\Sigma_\theta(t) = \beta_t I$, $\Sigma_{\text{post}}(t) = \tilde{\beta}_t I$, and the constants drop out, leaving [135]

$$\mathrm{KL}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big) \;\propto\; \frac{1}{2\beta_t}\left\|\mu_{\text{post}}(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\right\|^2 \;\approx\; \frac{1}{2\beta_t}\left\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t, t)\right\|^2 \tag{3.31}$$

where $\varepsilon$ is the Gaussian noise and $\varepsilon_\theta(\mathbf{x}_t, t)$ is neural network's predicted noise at time step $t$.

## 3.5 Transformer

The Transformer is a neural network architecture that replaces recurrence and convolution with a pure attention mechanism [95]. Each layer uses multi-head self-attention to compute weighted sums of all input positions, enabling direct, parallel modeling of long-range dependencies. This is followed by a position-wise feed-forward network, with residual connections and layer normalization around each sub-layer. Positional encodings are added to the input embeddings to retain sequence order. In a sequence-to-sequence setting, the decoder further applies masked self-attention and cross-attention over the encoder outputs. Together, these design choices produce a highly parallelizable model that has become the basis for virtually all state-of-the-art language and many vision models.

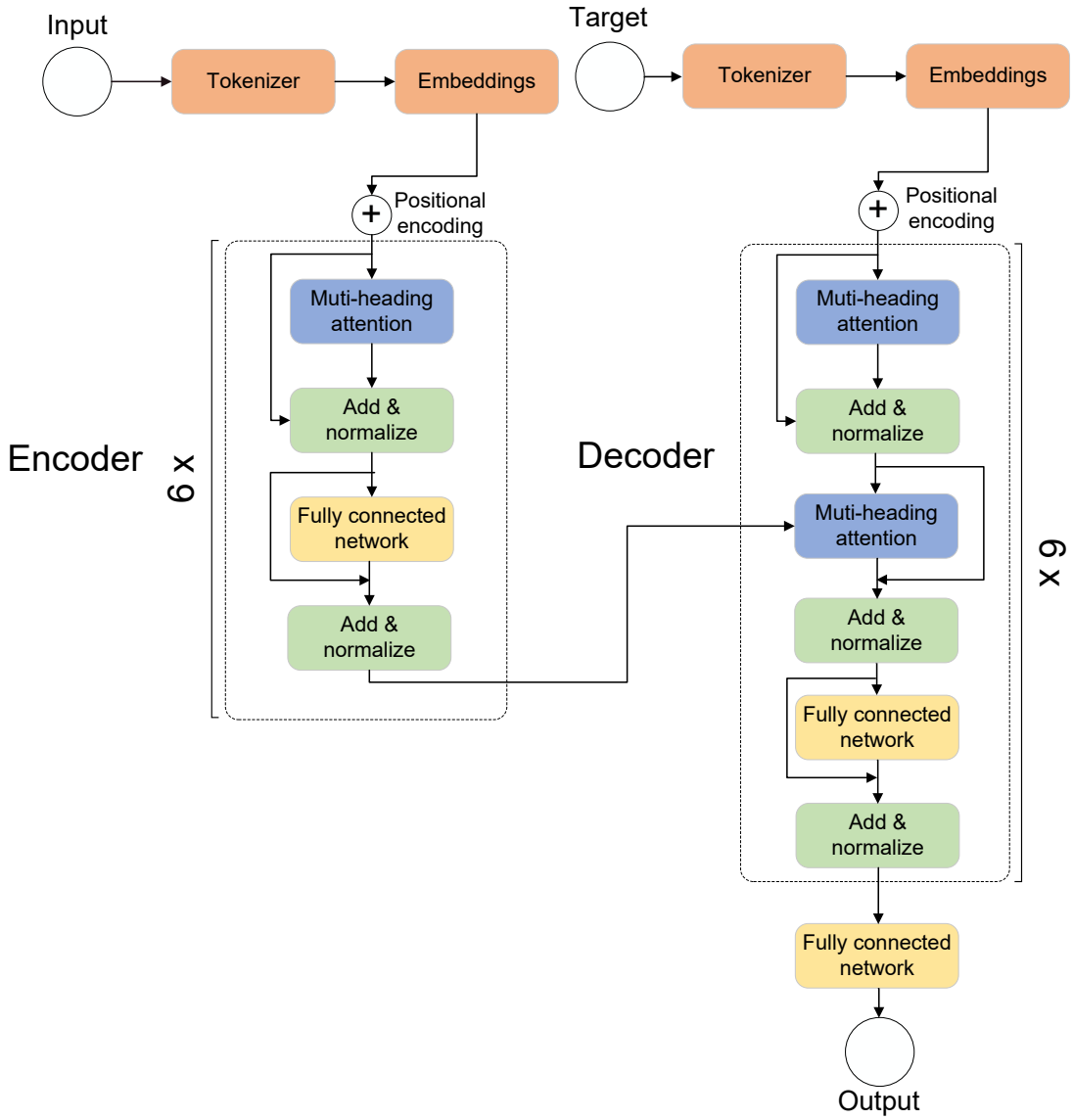



Figure 3.5: Structure of the conventional Transformer [95]

### 3.5.1 Structure of the Transformer

Figure 3.5 shows the Transformer architecture. Each input token is embedded into a dense vector. A fixed or learned positional encoding adds order information. The encoder contains several identical layers. In each layer multi-head self-attention lets every position attend to the rest. A residual connection and layer normalization follow. A position-wise feed-forward network adds nonlinearity. It uses two linear layers with a ReLU or Gaussian Error Linear Unit (GELU). Another residual connection and layer norm wrap this block. For sequence-to-sequence tasks the decoder mirrors the encoder.

It begins with masked self-attention to maintain autoregressive causality. It then performs cross-attention over the encoder outputs. The same feed-forward block follows with residual connections and layer norms. The top decoder states are projected to the vocabulary or target space by a linear layer. A softmax converts logits to probabilities. This design replaces recurrence and convolution with attention. It produces high parallelism and strong modeling of long-range dependencies.

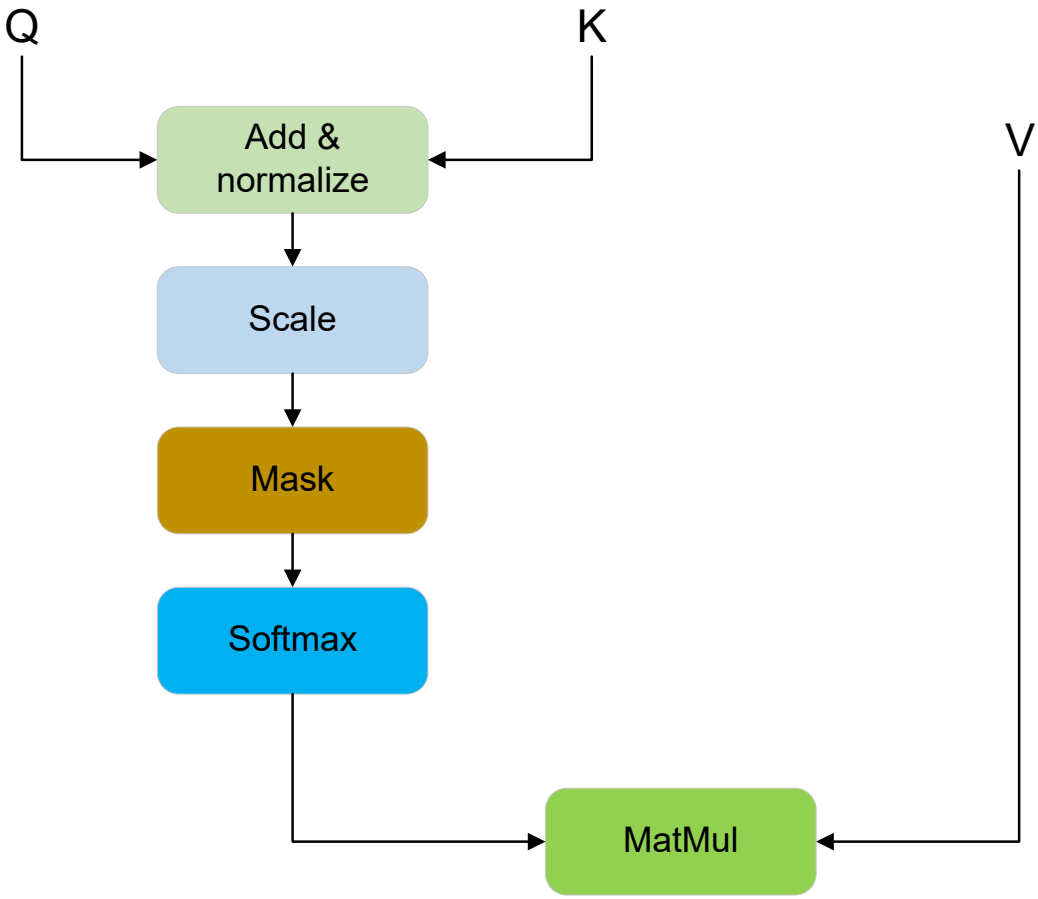



Figure 3.6: Attention mechanism in the Transformer [95]

Unlike a convolutional VAE, which relies on stacks of convolutional and deconvolutional layers for its encoder and decoder, the Transformer's encoder and decoder are composed exclusively of attention mechanisms and small feed-forward networks. Each layer applies multi-head self-attention, while the decoder additionally performs cross-attention over the encoder's outputs, followed by a position-wise two-layer Multi-Layer Perceptron (MLP). Every sub-layer is wrapped in a residual connection and layer normalization. Positional encodings are added to the inputs to convey sequence order.

The attention mechanism is regarded as the most critical component within the Transformer. When a scene is processed by the human visual system, an exhaustive scan of every detail is typically not performed. Rather, focus is directed toward particular regions based on interest or task requirements. In the example image, shown in Figure 2.9, attention is initially drawn to the animal's face.

### 3.5.2 Attention Block in the Transformer

In the Transformer, the attention mechanism computes attention scores by matching queries (Q) with keys (K) via scaled dot products. These scores are then converted into weights and applied to the value (V) matrix to produce the final attention vector. The whole process is represented in Figure. 3.6. Scaled dot-product attention is applied to a single sequence by first computing pairwise attention scores between all positions in that sequence via the dot product of queries and keys, scaled by the square root of the key dimension. These scores are then normalized and used to weight the corresponding value vectors, producing a weighted sum at each position that serves as the attention-enhanced representation for that position. The output is calculated by [95]:

$$Z = \text{softmax}\Big(\frac{QK^T}{\sqrt{d_k}}\Big)V, \tag{3.32}$$

where $Q \in \mathbb{R}^{n\times d_k}$ is the query matrix, $K \in \mathbb{R}^{m\times d_k}$ is the key matrix, $V \in \mathbb{R}^{m\times d_v}$ is the value matrix, and $d_k$ is the dimensionality of the queries and keys. Here $\text{softmax}(\cdot)$ is applied row-wise to $\frac{QK^T}{\sqrt{d_k}}$ to produce non-negative attention weights that sum to one in each row. The product $QK^T/\sqrt{d_k}$ computes pairwise similarities, the softmax normalizes these into attention weights, and multiplication by $V$ to get the final attention-weighted output $Z$. It quantifies the extent to which a particular piece of information is attended to by the attention mechanism, thereby reflecting its relative importance within that mechanism.

In the past decades, recurrent neural networks (RNNs) have become central to sequence modeling and transformation [155]. In particular, long short-term memory networks (LSTMs) and gated recurrent units (GRUs) have achieved state-of-the-art performance [156, 157]. Their applications include language modeling and machine translation. Compared with RNNs, the Transformer benefits from self-attention, which attends to all positions in a sequence at once, capturing long-range dependencies and enabling a more accurate understanding of extended contexts. It also processes tokens in parallel rather than sequentially, getting significantly greater computational efficiency and scalability.

### 3.5.3 Mechanism of the Transformer

In the conventional Transformer, the encoder is composed of six identical layers. Each layer comprises two sublayers: a multi-head self-attention mechanism and a position-wise feed-forward network. After each sublayer, a residual connection and layer normalization, together called Add & Norm in the Figure. 3.5, are applied. This design allows the encoder to capture dependencies across all positions in the input sequence. The Transformer's decoder is also composed of six identical layers. Each layer contains three sub-layers: a masked self-attention mechanism, an encoder–decoder attention mechanism, and a position-wise feed-forward network. After each sub-layer, a residual connection and layer normalization, Add & Norm, are applied. This design ensures that, during sequence generation, the decoder attends to all previous outputs while preventing information from future tokens from leaking into its predictions. The essential distinction between the encoder and the decoder lies in the masking applied to their self-attention layers. In the encoder, self-attention is unmasked, allowing each token to attend to every other token in the input sequence. In the decoder, however, a causal mask is imposed so that each position can only attend to previous tokens, thereby enforcing autoregressive generation.

The Transformer uses three distinct attention mechanisms. In the encoder, multi-head self-attention allows each input token to attend to all others, producing context-rich representations. Multi-Head Attention allows the model to jointly attend to information from different representation subspaces at different positions. Given $X \in \mathbb{R}^{n \times d_{\text{model}}}$, three learned projections produce the query, key, and value matrices [95]:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V, \tag{3.33}$$

where $W^Q, W^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $W^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$. For each of the $H$ heads, scaled dot-product attention is computed [95]:

$$\text{head}_i = \text{softmax}\Big(\frac{Q_i K_i^{\mathbb{T}}}{\sqrt{d_k}}\Big) V_i, \quad Q_i = XW_i^Q,\ K_i = XW_i^K,\ V_i = XW_i^V. \tag{3.34}$$

where $\text{head}_i$ is the $i_{th}$ attention operation with its own learned projections $W_i^Q, W_i^K, W_i^V$.

The outputs of all heads are concatenated and linearly projected:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H) W^O, \quad W^O \in \mathbb{R}^{H d_v \times d_{\text{model}}}. \tag{3.35}$$

where $\text{Concat}(\text{head}_1, \ldots, \text{head}_H)$ denotes concatenation along the feature dimension, getting a matrix in $\mathbb{R}^{n \times (H d_v)}$. $W^O$ is the learned output projection matrix of the multi-head attention block.

Furthermore, in the decoder's first sub-layer, multi-head causal self-attention applies a look-ahead mask so that each position can only attend to previous tokens, enabling autoregressive generation. Let $X^{\text{dec}} = [X_1^{\text{dec}}, \ldots, X_n^{\text{dec}}] \in \mathbb{R}^{n \times d_{\text{model}}}$ be the input sequence. Three learned projections produce queries, keys, and values:

$$Q = X^{\text{dec}} W^Q, \quad K = X^{\text{dec}} W^K, \quad V = X^{\text{dec}} W^V, \quad W^Q, W^K \in \mathbb{R}^{d_{\text{model}} \times d_k},\ W^V \in \mathbb{R}^{d_{\text{model}} \times d_v}. \tag{3.36}$$

A causal mask $M \in \mathbb{R}^{n\times n}$ is defined by

$$M_{ij} = \begin{cases} 0, & j \le i, \\ -\infty, & j > i, \end{cases} \tag{3.37}$$

so that each position can attend only to itself and previous positions. For each head $i = 1,\dots,H$, denote

$$Q_i = X^{\text{dec}} W_i^Q, \quad K_i = X^{\text{dec}} W_i^K, \quad V_i = X^{\text{dec}} W_i^V. \tag{3.38}$$

The scaled, masked attention scores and output are computed as

$$\text{head}_i = \text{softmax}\Big(\frac{Q_i K_i^{\mathbb{T}}}{\sqrt{d_k}} + M\Big)\, V_i, \tag{3.39}$$

where the softmax is applied row-wise. Finally, the heads are concatenated and projected back to $d_{\text{model}}$:

$$\text{MultiHead}(X^{\text{dec}}) = \text{Concat}(\text{head}_1,\dots,\text{head}_H)\, W^O, \quad W^O \in \mathbb{R}^{H\,d_v\times d_{\text{model}}}. \tag{3.40}$$

Let $X^{\text{dec}} \in \mathbb{R}^{n\times d_{\text{model}}}$ be the decoder input at this layer and $H^{\text{enc}} \in \mathbb{R}^{m\times d_{\text{model}}}$ be the encoder outputs. For each head $i = 1,\dots,H$,

$$Q_i = X^{\text{dec}} W_i^Q, \quad K_i = H^{\text{enc}} W_i^K, \quad V_i = H^{\text{enc}} W_i^V, \tag{3.41}$$

with $W_i^Q \in \mathbb{R}^{d_{\text{model}}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}}\times d_v}$. The scores and row-wise attention are

$$S_i = \frac{Q_i K_i^{\mathbb{T}}}{\sqrt{d_k}} \in \mathbb{R}^{n\times m}, \qquad R_i = \text{softmax}(S_i) \in \mathbb{R}^{n\times m}. \tag{3.42}$$

The head output is

$$\text{head}_i = R_i V_i \in \mathbb{R}^{n\times d_v}. \tag{3.43}$$

Concatenating heads and projecting produces

$$\text{MultiHeadCross}(X^{\text{dec}}, H^{\text{enc}}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\, W^O, \qquad W^O \in \mathbb{R}^{Hd_v \times d_{\text{model}}}. \tag{3.44}$$

In Figure 3.5, positional encoding (PE) is used to inject information about the order of tokens into the model, since the attention mechanism itself is permutation-invariant. In the original Transformer, it was defined as a fixed, deterministic function of position *pos* and embedding dimension index $i$ [95]:

$$\text{PE}(pos, 2i) = \sin\Big(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\Big), \quad \text{PE}(pos, 2i+1) = \cos\Big(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\Big), \tag{3.45}$$

where $d_{\text{model}}$ is the model dimension, *pos* is the integer token position in the sequence.

Using the PE allows the model to learn to attend by relative as well as absolute positions, and it generalizes to sequence lengths not seen during training. Alternatively, the PE can replace these fixed encodings with learned positional embeddings of the same shape, which are learned jointly with the rest of the parameters.

## 3.6 Large Language Model

Within the scope of this thesis, the LLM is introduced to address the limitation of manual or low-level control over generative factors. While causal generative models enable interpretable and disentangled manipulation of underlying factors, they typically require explicit specification of these factors in latent space, which is neither intuitive nor scalable for complex scenes. By incorporating an LLM, high-level textual descriptions can be translated into structured conditioning signals that align semantic intent with the underlying generative process.

This translation is achieved through multimodal fusion mechanisms, most commonly implemented via cross-attention networks, which allow textual representations to modulate visual feature generation dynamically. Through this mechanism, language embeddings guide the image generation process without directly operating on pixels, enabling coherent alignment between textual descriptions and visual content. As a result, natural-language-driven control becomes compatible with structured generative models, supporting end-to-end image generation that is both semantically meaningful and visually consistent.

From a research perspective, this integration enables the investigation of whether natural language can serve as an effective interface for controlling complex generative processes under causal constraints. It directly supports the thesis objective of achieving controllable, interpretable, and trustworthy image generation by allowing users to specify counterfactual or hypothetical scenarios in natural language, while ensuring that the resulting images remain consistent with learned structural dependencies.

### 3.6.1 Mechanism of the LLM

An LLM refers to a machine-learning model with a very large number of parameters and a correspondingly complex computational structure. Such models are typically built as deep neural networks containing tens or even hundreds of billions of parameters. They are designed to enhance representational capacity and predictive performance, enabling them to handle more complex tasks and data modalities. LLMs have been applied broadly across domains, including natural language processing, computer vision, speech recognition, and recommendation systems. By training on massive datasets, these models learn intricate patterns and features, obtaining strong generalization capabilities and the ability to make accurate predictions on previously unseen inputs. During pre-training, the model parameters are optimized to minimize the negative log-likelihood of held-out text, often with variants such as causal masking or span corruption to encourage fluency and coherence.

A large model is fundamentally a deep neural network trained on vast amounts of data. Its enormous scale in both parameters and training examples gives rise to emergent intelligence, exhibiting behaviors that resemble human-like cognition [158]. After pre-training, LLMs can be adapted to downstream tasks via fine-tuning on labeled examples or prompted directly in a zero- or few-shot fashion [159]. In zero-shot prompting, natural-language instructions suffice to elicit desired behaviors. In few-shot prompting, a small number of input–output exemplars are provided within the prompt [160]. As model size and training data grow, performance on diverse language benchmarks, ranging from question answering to code generation, improves markedly.

Inference in LLMs is performed autoregressively. At each step, the model computes [161]

$$P(w_t \mid w_{<t}) = \text{softmax}\big(W_O \,\text{Attention}(Q,K,V)\big) \tag{3.46}$$

where $w_t$ is the random variable for the token at position $t$ and $w_{<t}$ are the previously generated tokens. Sampling strategies such as greedy decoding, top-k sampling, or nucleus (top-p) sampling are used to balance quality and diversity in the output.

The deployment of LLMs raises important considerations in terms of computational cost, latency, and ethical safeguards. Techniques such as model quantization, distillation, and retrieval-augmented generation have been developed to reduce resource usage, while alignment methods, such as reinforcement learning from human feedback (RLHF), are applied to steer outputs toward safe and helpful behavior.

Based on the type of input data, LLMs can be classified into the three main categories: Language Models, Vision Models, and Multimodal Models. For the Language Models, Language Models in the NLP domain are almost exclusively implemented as deep Transformer–based architectures trained on massive text corpora. Two principal pretraining paradigms are employed. For the decoder-only autoregressive modeling, in which the network maximizes the likelihood [162]

$$\prod_{t=1}^{L} p(w_t \mid w_{<t})$$

over a sequence of tokens $w_1, \ldots, w_L$, and masked denoising modeling, in which the model reconstructs corrupted or masked tokens by minimizing the negative log-likelihood of the true tokens given the unmasked context. Model sizes range from the hundreds of millions of parameters to hundreds of billions or more. After pretraining, the networks in the LLM are adapted to downstream tasks either by fine-tuning on labeled data with task-specific heads or by prompting via natural-language instructions and exemplars. Language models have demonstrated state-of-the-art performance across tasks such as text generation, summarization, translation, question answering, code synthesis, and dialogue, with capabilities that scale and often emerge as model size and data diversity increase. Language models employ decoder-only Transformer architectures. Input tokens are first embedded into a high-dimensional space and enriched with positional encodings. These embeddings are then processed by a series of identical Transformer layers, each comprising masked multi-head self-attention followed by a position-wise feed-forward network. Residual connections and layer normalization are applied around every sub-layer. Finally, the resulting hidden states are projected into the vocabulary space via a linear transformation and normalized with softmax to produce next-token probabilities.

Vision models are tailored to process two-dimensional image data and extract hierarchical visual features. Traditional architectures employ convolutional layers with local receptive fields and weight sharing, arranged in successive blocks of convolution, nonlinearity, and pooling to progressively increase abstraction and receptive field. Residual connections and batch normalization are commonly used to enable very deep networks. More recently, Vision Transformers (ViT) have been introduced. An image is first divided into a sequence of fixed-size patches, each of which is flattened and linearly projected into a high-dimensional embedding. Then positional encodings are added to retain spatial information, and the resulting patch embeddings are processed by a stack of Transformer encoder layers. Finally, a classification head, typically a linear layer applied to the embedding of a special “class” token, is used to produce the model’s output. This attention-based paradigm enables global context modeling across the entire image and has been shown to match or exceed convolutional back-

bones when trained on sufficiently large datasets. Vision models typically follow either a convolutional or an attention-based paradigm. Convolutional backbones use successive layers of small-kernel convolutions to extract local features, interleaved with nonlinear activations and pooling to increase spatial abstraction. Residual connections and normalization layers enable the construction of very deep networks.

Multimodal models are designed to process and integrate multiple data modalities within a single architecture. Typically, separate modulality-specific encoders, such as a convolutional or Vision-Transformer encoder for images and a Transformer encoder for text, first extract independent embeddings. These embeddings are then fused through cross-attention blocks or by interleaving modality-specific self-attention with cross-attention layers in a joint Transformer backbone. Pretraining objectives commonly include contrastive alignment losses to bring paired image–text representations into a shared space. Generative reconstruction losses are applied to predict one modality from another. Task-specific heads are then attached to support downstream classification, retrieval, or generation tasks. Pretraining on large-scale paired datasets enables the model to learn rich cross-modal correspondences, supporting downstream applications such as image captioning, visual question answering, text-to-image synthesis, and multimodal retrieval.

Beyond these types, LLMs also guide image generation. Recent methods use them to drive text-to-image pipelines. They first expand a brief instruction into a detailed prompt, specifying objects, composition, style, lighting, and mood. The enriched prompt is then fed to a generative image model. During sampling, a vision–language scorer evaluates each intermediate output against the LLM-produced prompt, guiding the denoising trajectory toward semantically faithful images. In advanced systems, the generated image is re-captioned by a vision–language model and the resulting description is passed back to the LLM for further refinement, forming a closed-loop that iteratively enhances coherence between text and visual content.

In a typical LLM-driven text-to-image workflow, the user's initial instruction is first converted by the language model into a structured prompt embedding that captures not only object descriptions but also spatial relationships, stylistic attributes, and lighting cues. This embedding is then concatenated with learned positional encodings and fed into the image generator's cross-modal conditioning layer, where it modulates the denoising U-Net via cross-attention at multiple resolutions. During each reverse diffusion step, the model computes a guided update by combining the unconditional noise prediction with the prompt-conditioned prediction, weighted by a guidance scale that balances fidelity against diversity. Optionally, a separate vision–language matcher, such as Contrastive Language–Image Pre-training (CLIP), scores intermediate samples and back-propagates a small adjustment to the prompt embedding, sharpening semantic alignment. Once the final image is produced, it can be captioned by the vision–language model and the resulting text re-injected into the LLM for iterative refinement, closing the loop between description and visual output.

Large language models are trained on very large, mixed datasets. Typical sources include web pages, digitized books and encyclopedias, code repositories, conversation logs, and domain-specific collections. Using this mix teaches the model many writing styles, facts, and ways of reasoning. Web text adds breadth and current usage. Books and articles provide long, edited prose. Code teaches precise syntax and structured thinking. Dialogue data teaches turn-taking and practical conversational cues.

### 3.6.2 Finetuning in the LLM

Furthermore, the pretrained LLM models cannot deal with the details of all tasks well. For a specific task, finetuning an LLM model has become a popular choice for most researchers. Fine-tuning an LLM typically begins with a pretrained base model, such as Llama and GPT, and a task-specific dataset of input–output pairs. In the simplest supervised setup, the model's parameters $\boldsymbol{\theta}$ are updated to minimize the cross-entropy

loss [162]

$$\mathscr{L}_{\mathrm{CE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{L_i}\log p_\theta\left(y_t^{(i)} \mid y_{<t}^{(i)}, x^{(i)}\right) \tag{3.47}$$

using a small learning rate, gradient accumulation, and the Adam with decoupled weight decay (AdamW) optimizer with weight decay. To improve sample efficiency and reduce storage, parameter-efficient methods such as Adapters (small bottleneck layers inserted between Transformer blocks), Low-Rank Adaptation (LoRA) and Prefix-Tuning are often employed.

To reduce both memory footprint and latency at serving time, quantization techniques convert full-precision weights and activations to lower bit-width representations. Post-training quantization (PTQ) maps 32-bit floats to 8-bit integers with minimal accuracy loss by calibrating scale and zero-point parameters on a small calibration set, while quantization-aware training (QAT) simulates quantization effects during fine-tuning to recover performance. Recent advances such as GPTQ or QLoRA extend these methods to extreme 3 or 4-bit quantization with clever grouping and Hessian-based rounding, enabling multi-billion-parameter models to run on a single graphics processing unit (GPU).

In these finetuning methods, LoRA is a popular choice for most developers. LoRA has become a popular standard for efficient LLM fine-tuning because it combines strong task performance with minimal extra compute and storage overhead. By freezing the majority of the pretrained weights and only learning a small low-rank update $\Delta W = BA$ (with $r \ll d$), LoRA typically adds fewer than 1 %–5 % more parameters per layer, dramatically reducing GPU memory usage and enabling much larger batches or longer contexts during fine-tuning. The low-rank adapters can be merged into the original weights at inference with no extra runtime cost, preserving the model's original throughput. Moreover, LoRA is architecture-agnostic. It can be applied to any transformer block, and has been shown empirically to match or even surpass full fine-tuning on many downstream tasks, making it an attractive, plug-and-play solution for both research and production.

In detail, LoRA freezes the original pretrained weight matrices $W_0 \in \mathbb{R}^{d\times k}$ and injects a trainable low-rank decomposition $\Delta W = BA$, where

$$A \in \mathbb{R}^{r\times k}, \quad B \in \mathbb{R}^{d\times r}, \quad r \ll \min(d,k).$$

Here, $A$ and $B$ are the down-projection (dimension-reducing) matrix and up-projection (dimension-expanding) matrix, respectively. In practice, $A$ is typically initialized from a Gaussian distribution while $B$ is initialized to the zero matrix.

The adapted weight is [163]

$$W = W_0 + \alpha_{lora}\frac{BA}{r} \tag{3.48}$$

with a scalar scaling factor $\alpha_{lora}$ ,which is often set to $r$, to stabilize training. During fine-tuning, only the parameters of $A$ and $B$ are updated, while $W_0$ remains fixed, dramatically reducing GPU memory and storage requirements. At inference time, the low-rank update $BA$ can be merged into $W_0$ with no extra compute overhead, preserving the original architecture's speed. LoRA is most commonly applied to the query, key, value, and output projection matrices within each multi-head attention block, as well as the two weight matrices in the feed-forward network, obtaining efficient, high-performance adaptation with minimal additional parameters.

## 3.7 Evaluation Metrics

To comprehensively evaluate the performance of the proposed models, both quantitative and qualitative evaluation metrics are employed. These metrics are designed to assess reconstruction fidelity, perceptual similarity, and the structure of learned latent representations. In addition, standard loss functions are used during training to guide optimization. Together, these criteria provide a multi–level evaluation of model performance.

### 3.7.1 Pixel-Level and Perceptual Metrics

Mean Absolute Error is used to measure pixel-wise reconstruction accuracy between a generated image $\widehat{\mathbf{x}}$ and the corresponding ground-truth image $\mathbf{x}$. It is defined as

$$\mathrm{MAE}(\mathbf{x}, \widehat{\mathbf{x}}) = \frac{1}{d_{\mathbf{x}}} \sum_{i=1}^{d_{\mathbf{x}}} \left| x_i - \widehat{x}_i \right|, \tag{3.49}$$

where $d_{\mathbf{x}}$ denotes the dimensionality of the image. MAE is robust to outliers and provides an interpretable measure of average absolute deviation at the pixel level.

While pixel-wise metrics capture low-level differences, they often fail to reflect perceptual similarity as judged by humans. To address this limitation, the Learned Perceptual Image Patch Similarity (LPIPS) metric is adopted [164]. LPIPS measures the distance between deep feature representations extracted from a pretrained convolutional network:

$$\mathrm{LPIPS}(\mathbf{x}, \widehat{\mathbf{x}}) = \sum_{l} w_l \left\| \phi_l(\mathbf{x}) - \phi_l(\widehat{\mathbf{x}}) \right\|_2^2, \tag{3.50}$$

where $\phi_l(\cdot)$ denotes the activation at layer $l$ of a fixed feature extractor and $w_l$ are learned layer weights. Lower LPIPS values indicate higher perceptual similarity. This metric is particularly suitable for evaluating generative models, as it correlates well with human visual perception.

### 3.7.2 Visual and Representation-Level Evaluation

To qualitatively assess the structure of learned latent representations, Principal Component Analysis (PCA) is used as a visualization tool [165]. Given a set of latent vectors $\{\mathbf{z}_i\}_{i=1}^{N}$, PCA projects them onto a low-dimensional subspace spanned by the leading eigenvectors of the covariance matrix. This enables visual inspection of clus-

tering behavior, disentanglement properties, and the separation of samples across different conditions or interventions. PCA does not provide a scalar performance score but serves as an important diagnostic for understanding representation geometry and causal structure in the latent space.

### 3.7.3 Training Loss Functions

Mean Squared Error is used as a reconstruction loss during training, penalizing large deviations more strongly than MAE:

$$\mathrm{MSE}(\mathbf{x},\widehat{\mathbf{x}}) = \frac{1}{d_{\mathbf{x}}}\sum_{i=1}^{d_{\mathbf{x}}}\left(x_i-\widehat{x}_i\right)^2. \tag{3.51}$$

MSE provides smooth gradients and is commonly employed in autoencoders and diffusion models for stable optimization.

For models with Bernoulli likelihood assumptions or sigmoid output activations, Binary Cross-Entropy is adopted:

$$\mathrm{BCE}(\mathbf{x},\widehat{\mathbf{x}}) = -\frac{1}{d_{\mathbf{x}}}\sum_{i=1}^{d_{\mathbf{x}}}\Big[x_i\log\widehat{x}_i+(1-x_i)\log(1-\widehat{x}_i)\Big]. \tag{3.52}$$

BCE is particularly suitable when pixel intensities are normalized to $[0,1]$ and interpreted probabilistically.

In summary, MAE and LPIPS are used as quantitative evaluation metrics to assess reconstruction accuracy and perceptual quality, respectively. PCA provides a qualitative, visual assessment of latent-space structure and disentanglement. During training, MSE and BCE serve as loss functions that guide optimization under different likelihood assumptions. This combination of metrics ensures a balanced evaluation across pixel-level accuracy, perceptual fidelity, and representation quality.

## 3.8 Problem Formulation

This thesis addresses the problem of trustworthy and controllable counterfactual image generation under limited data availability and visually confounding conditions, such as illumination changes, shadows, and occlusions. The objective is to generate counterfactual images that are not only visually realistic, but also causally consistent, interpretable, and controllable, either through explicit factor manipulation or through natural-language instructions.

In this context, a counterfactual image is defined as an image that answers a causal "what-if" question: what would the scene look like if a specific generative factor were changed, while all other factors remained consistent with the underlying data-generating process. Achieving this goal requires models that go beyond correlation-based image synthesis and instead reason about the causal structure of visual scenes.

### 3.8.1 Limitations of Existing Components

Recent research on generating counterfactual images has shown some drawbacks. Causal representation learning methods based on VAEs and related latent-variable models provide an important foundation for counterfactual reasoning. By learning structured and interpretable latent spaces aligned with structural causal models, these approaches enable explicit interventions on semantic factors. However, due to the combination of smooth reconstruction objectives and distributional regularization, such models typically produce over-smoothed or blurry images, which limits their applicability in realistic vision scenarios where fine-grained visual details are essential.

Diffusion models, in contrast, have demonstrated remarkable success in generating high-fidelity and photorealistic images. Nevertheless, standard diffusion models lack an explicit low-dimensional latent space that can be directly intervened upon. As a result, edits performed by diffusion-based models are primarily driven by statistical correlations learned from data, rather than by causal mechanisms. Even when conditioned on labels or prompts, these models cannot guarantee that an edit corresponds to a valid causal intervention or that non-target factors remain invariant.

Recent LLM-guided image editing methods enable flexible and intuitive control through natural-language instructions. While these approaches improve usability and semantic alignment, they typically operate at a heuristic or symbolic level and do not explicitly model causal dependencies between scene factors. Consequently, language-guided edits may introduce unintended changes, violate physical relationships, or suffer from hallucination effects, especially in visually ambiguous environments. This highlights the need for causal constraints beneath language-driven control.

### 3.8.2 Unified Problem Statement

To overcome these limitations, this thesis formulates counterfactual image generation as a problem of causal latent intervention followed by high-fidelity conditional synthesis. The central challenge is to design a generative framework that simultaneously supports:

- interpretable and structured latent representations that reflect underlying causal factors,
- principled intervention mechanisms that modify only the intended factors,
- and high-quality image synthesis that preserves visual realism and fine details.

Rather than operating directly in pixel space, interventions are performed in a learned latent space that captures semantic and causal structure. The modified latent representation is then translated into a realistic image through a powerful generative model. This separation between causal reasoning and image synthesis allows the framework to combine interpretability with perceptual quality.

### 3.8.3 Strategy Adopted in This Thesis

Building on the insights from Causal DiffuseVAE and Causal DiffuseLLM, this thesis adopts a unified strategy that integrates multiple complementary components. Variational inference is used to learn compact and stable latent representations that support uncertainty modeling and data efficiency. Masked causal layers are introduced to encode structural dependencies among latent factors and to ensure that interventions are causally valid. Diffusion-based generators are employed to decode intervened latents into high-fidelity images, addressing the visual limitations of conventional causal generative models. When natural-language control is required, Transformer- and LLM-based modules are incorporated to translate user instructions into structured interventions that respect the learned causal structure.

This integration resolves a fundamental trade-off in existing methods: causal models provide interpretability but lack realism, while diffusion models provide realism but lack causal controllability. By combining these components within a single framework, the proposed approach enables trustworthy counterfactual reasoning together with practical, high-quality image generation in both data-efficient and text-driven settings.

In summary, the problem addressed in this thesis is the design of a unified causal–generative framework that reconciles interpretability, controllability, and visual fidelity. The following chapters instantiate this formulation in concrete model architectures and validate it empirically across multiple datasets and counterfactual intervention scenarios.

## 3.9 Chapter Summary

This chapter has established the theoretical and methodological foundations required for the study of counterfactual image generation. Core concepts from structural causal models were introduced to formalize interventions and counterfactual reasoning, clarifying how causal structure enables principled manipulation of generative factors. Building on this foundation, causal inference techniques in machine learning were reviewed, highlighting how invariant and robust learning objectives relate to the challenges posed by distribution shifts and spurious correlations in visual data.

The chapter then surveyed the key generative and representational models employed in this thesis, including variational autoencoders, diffusion models, and Transformer-based architectures. Variational autoencoders were presented as a mechanism for learning structured latent representations, diffusion models as a powerful framework for high-fidelity image synthesis, and Transformers and large language models as flexible interfaces for semantic and multimodal control. Together, these components provide complementary strengths but also exhibit individual limitations when applied in isolation.

Finally, these ideas were unified into a formal problem formulation for trustworthy and controllable counterfactual image generation. The formulation emphasizes causal latent interventions followed by realistic conditional synthesis, addressing the core trade-offs between interpretability, controllability, and visual fidelity. This chapter thus provides a coherent conceptual framework that motivates the proposed models and algorithms.

# Chapter 4

# Trustworthy Counterfactual Generative Model Based on Causal Inference

In this chapter, the design and implementation of Causal DiffuseVAE are presented in detail. First, the joint VAE–diffusion architecture, motivated by the need to disentangle and ground latent factors according to an explicit structural causal model of scene formation, is introduced, in which images are encoded into causally structured latents and subsequently denoised via a conditioned diffusion process. Next, interventions in latent space are described, wherein the do-operator is applied to individual causal variables and the modified latents are decoded back into pixel space to generate counterfactual images. The training objectives, comprising reconstruction loss, diffusion-based denoising loss, and causal regularization terms that enforce factor independence and interventional consistency, are then detailed. Finally, the inference pipeline, the synthetic and real-world evaluation datasets, and a suite of experiments and ablation studies are outlined to demonstrate the model's ability to produce physically plausible, semantically controlled images under varied lighting and occlusion scenarios.

## 4.1 Causal Mechanism

Current causal generative models, like CausalVAE [166] and causal disentangled representation learning for missing data (CDRM) [167], cannot generate 3-Dimensional images with both high accuracy and high quality due to the VAE's smooth loss function. The combination of a pixel-wise reconstruction term and the continuous KL divergence encourages averaged, smooth outputs and heavily penalizes sharp transitions. This loss function results in overly blurry reconstructions. The limitation leads to inconsistencies between generated images and the underlying engineering principles they aim to represent. The architecture of the Causal DiffuseVAE is illustrated in Figure. 4.1, which combines causal inference with the VAE and the Diffusion Model. Figure. 4.1 shows how data flow through an encoder, transforming inputs into causal latent representations. Causal mechanism specifies how one causal variable influences an effect variable, which is achieved by a causal layer. The causal layer organizes these latent variables based on an underlying causal graph, where labeled concepts, such as Concept 1, Concept 2, etc., are assigned numerical values. These values change when causal interventions are applied. Then the causal layer incorporates a decoder to generate conditions for the Denoising Diffusion Probabilistic Model (DDPM), which ensures the generative process of the DDPM respects causal dependencies. Furthermore, model learning refers to the joint optimization of the encoder, causal layer, and diffusion network under combined VAE and diffusion objectives. The causal mechanism and model learning are introduced below to show the details of the Causal DiffuseVAE.

Causal Mechanism is to transform unstructured data into a structured latent space that explicitly captures and leverages the causal relationships among the underlying factors. In the proposed Causal DiffuseVAE, the causal latent variables $\mathbf{z}$ serve as the structured data, capturing the underlying causal factors that influence the observed features. A structured encoding process extracts $\mathbf{z}$, which maps labeled concepts to their respective causal influences. These structured latent variables then guide the con-

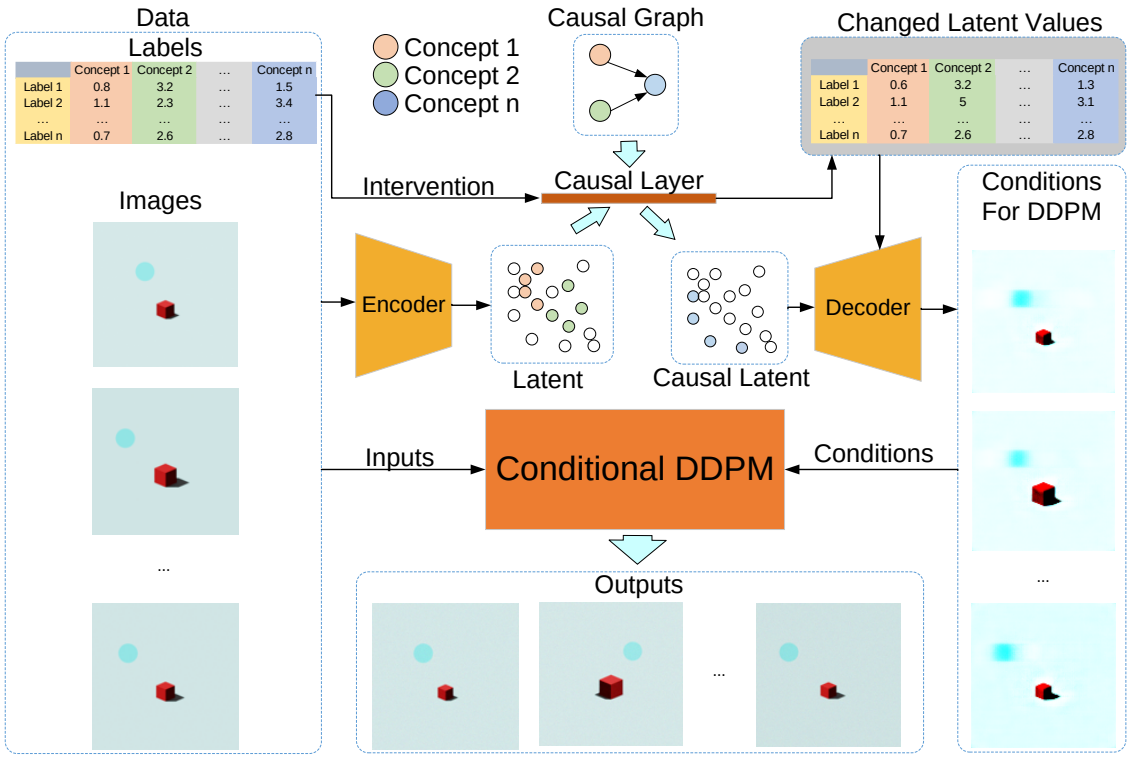



Figure 4.1: Overall architecture of Causal DiffuseVAE. The labels and the images are the inputs of the model. Through the causal layer, the causal relationship among the latent can be learned in the training process. Furthermore, the intervention process can be achieved by changing the values of the labels of the causes to influence the labels of the effects.

ditional DDPM, ensuring that generated images reflect the true causal dependencies. This approach enhances interpretability and reliability, allowing for meaningful interventions and counterfactual generation. In the Causal DiffuseVAE, the causal latent variables $\mathbf{z}$ correspond to the causal features in the real world.

Figure 4.2 illustrates the mechanism of causal inference. During an intervention, when a specific latent in $\mathbf{z}$ is altered, the causal layer modifies the corresponding effect based on the encoded causal dependencies. The causal layer serves as a transformation matrix derived from a predefined causal graph. This causal layer encodes the relationships among $\mathbf{z}$, ensuring that changes in cause factors propagate their effects accordingly. While the causal graph remains fixed, the causal matrix is designed to be trainable, allowing the model to adaptively refine its learned causal relationships throughout the training process.

The mechanism of causal inference is represented as

$$z_i = g_i(A_i \circ \mathbf{z}; \eta_i) + \varepsilon_i \tag{4.1}$$

where $z_i$ is the $i$th element of $\mathbf{z}$, $g_i$ is the $i$th element of the set of mild nonlinear and invertible functions $\mathbf{g}$, $A_i$ is the $i$th column vector in the adjacency matrix $\mathbf{A}$, $\circ$ is the elementwise product, and $\eta_i$ is the learnable parameter of $\mathbf{g}$.

By enforcing $A_i \circ \mathbf{z}$, we ensure that only the true parent nodes influence each $z_i$. This results in a disentangled, interpretable causal code as the true parent nodes capture the relationships between causes and effects. Because each $z_i$ is built only from its masked parent nodes, any change to a "cause" latent will automatically propagate through all downstream nodes. This enforces that interventions on one latent spread through the network exactly following the causal arrows in the SCM.

The causal layer in (4.1) is designed to mirror a structural causal model in latent space: each latent factor $z_i$ is generated from its (masked) parents and an independent disturbance.

Throughout this chapter, it is assumed that (i) the underlying causal graph is a directed acyclic graph (DAG), (ii) the masking operation $A_i \circ \mathbf{z}$ restricts each mechanism to use only its parent variables, and (iii) the functions $g_i(\cdot)$ are mild nonlinear and invertible, so that different causal factors remain distinguishable in latent space.

Under these assumptions, interventions applied to a parent latent are expected to propagate to downstream latents through the learned mechanisms, which encourages disentangled and causally consistent counterfactual changes rather than arbitrary correlated edits.

Importantly, this provides a modeling justification and an inductive bias, not an absolute guarantee for every dataset.

Once $\mathbf{z}$ has been recomputed under these causal rules, it conditions the VAE decoder and downstream diffusion sampler. Because the decoder is explicitly trained to treat each dimension of $\mathbf{z}$ as a semantically meaningful factor, changes in a cause latent produce coherent and localized changes in the output image. In this way, the causal layer guarantees that tuning a cause dimension leads to the correct effect in the generated image and reflects the learned causal graph.

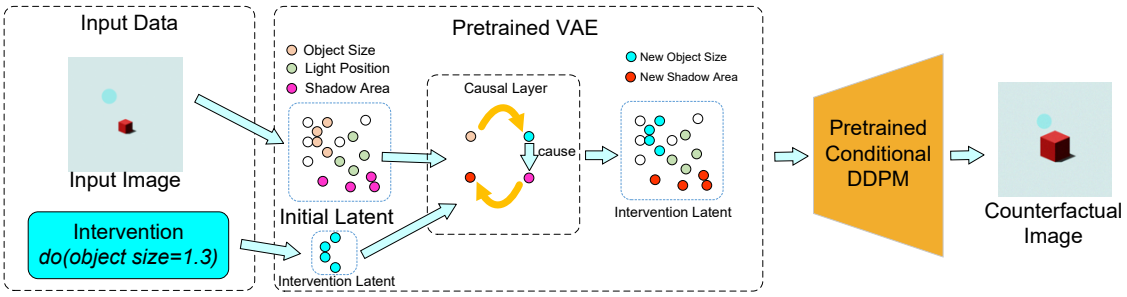



Figure 4.2: The intervention process in the shadow situation. When the latent of the object size is changed, the shadow area in the causal layer will also change.

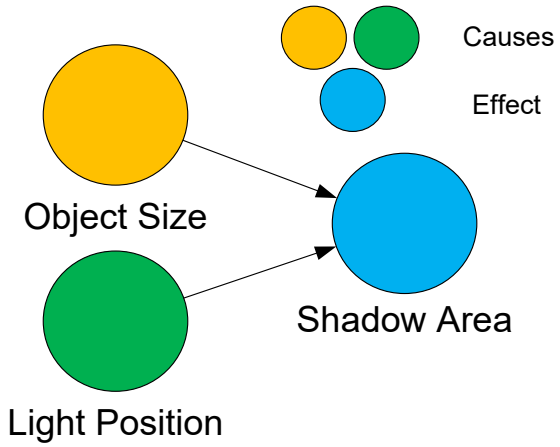



Figure 4.3: The causal relationships in the shadow scenario.

Considering the case that a light source casts a shadow under an object, in Figure. 4.2, whose causal relationships are shown in Figure. 4.3. Let $\mathbf{z} = (z_{\text{light}}, z_{\text{object\_size}}, z_{\text{shadow}})$, where $z_{\text{light}}$ encodes the light position, $z_{\text{object\_size}}$ indicates scale of the object, and $z_{\text{shadow}}$ denotes the shadow area. If we increase $z_{\text{light}}$ (simulating a right-side light position) as the intervention, the causal layer recomputes the shadow area:

$$z_{\text{shadow}} = g_{\text{shadow}}\left(A_{\text{shadow}} \circ \mathbf{z}\right) \tag{4.2}$$

When the light angle increases, the shadow shrinks as expected. The diffusion decoder then generates an image showing the shadow shifted and reduced, proving that adjusting the causal latent produces the correct effect. These causal latents guide the diffusion decoder at every denoising step. By fusing this structured embedding with guidance information, the model follows the learned causal graph.

## 4.2 Model Learning

Model learning is to capture the complex causal dependencies within the observed data by refining both the structured latent space and the transformation functions. To model the evolution of latent variables across different contexts in the diffusion process, the loss function is defined as

$$L = \mathscr{L}_{\text{VAE}} + \mathscr{L}_{\text{DDPM}} \tag{4.3}$$

where $\mathscr{L}_{\text{DDPM}}$ is the loss function of the Diffusion Models.

### 4.2.1 Learning Strategy with No Confounders

#### 4.2.1.1 Learning Strategy of the VAE

In $\mathscr{L}_{\text{VAE}}$, the generative process is defined as

$$p_\theta(\mathbf{x}, \mathbf{z}, \varepsilon \mid \mathbf{u}) = p_\theta(\mathbf{x} \mid \mathbf{z}, \varepsilon, \mathbf{u}) p_\theta(\varepsilon, \mathbf{z} \mid \mathbf{u}) \tag{4.4}$$

where $\mathbf{u}$ is the learning guidance. The inference process is defined as

$$q_\phi(\mathbf{z}, \varepsilon \mid \mathbf{x}, \mathbf{u}) \equiv q(\mathbf{z} \mid \varepsilon) q_\zeta(\varepsilon - h(\mathbf{x}, \mathbf{u})) \tag{4.5}$$

where $\zeta$ is the vector of independent noise with probability densities $q_\zeta$, and $h(\mathbf{x}, \mathbf{u})$ represents the mechanisms of the encoder.

The general ELBO is defined for uncausal latent. For the causal layer, the ELBO in the Causal DiffuseVAE is redefined to learn the parameters $\theta$ and $\phi$ as

$$\begin{aligned} \text{ELBO} = \mathbb{E}_X \Bigg\{ & \underbrace{\mathbb{E}_{\varepsilon, \mathbf{z} \sim q_\phi} \left[ \log p_\theta(\mathbf{x} \mid \mathbf{z}, \varepsilon, \mathbf{u}) \right]}_{\text{Reconstruction Loss}} \\ & - \underbrace{D_{\text{KL}} \left[ q_\phi(\varepsilon, \mathbf{z} \mid \mathbf{x}, \mathbf{u}) \,\|\, p_\theta(\varepsilon, \mathbf{z} \mid \mathbf{u}) \right]}_{\text{KL Divergence Regularization}} \Bigg\} \end{aligned} \tag{4.6}$$

The KL divergence between $q_\phi$ and $p_\theta$ is factorized as

$$\begin{aligned} D_{\text{KL}} \Big[ q_\phi(\varepsilon, \mathbf{z} \mid \mathbf{x}, \mathbf{u}) \,\|\, p_\theta(\varepsilon, \mathbf{z} \mid \mathbf{u}) \Big] = \\ \mathbb{E}_{q_\phi} \left[ \log \frac{q_\phi(\varepsilon, \mathbf{z} \mid \mathbf{x}, \mathbf{u})}{p_\theta(\varepsilon, \mathbf{z} \mid \mathbf{u})} \right] \end{aligned} \tag{4.7}$$

The reconstruction loss is normally computed using Mean Squared Error (MSE) or Binary Cross-Entropy (BCE). MSE produces high-quality images but smooths out causal latent structures by focusing on pixel-wise differences. BCE ensures high accuracy but reduces image quality as it treats pixel values independently. This mechanism makes it better suited for binary data and potentially introducing artifacts in continuous image reconstruction. To generate images with both high quality and accuracy, we define the reconstruction loss $\mathscr{L}_{\text{recon}}$ in the Causal DiffuseVAE as

$$\mathscr{L}_{\text{recon}} = \alpha \cdot \text{BCE}(\mathbf{x}, \hat{\mathbf{x}}) + \nu \cdot \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) \tag{4.8}$$

where $\hat{\mathbf{x}}$ is the reconstructed image; $\alpha$ and $\nu$ are the coefficients of the BCE and MSE, respectively, whose sum equals one.

Combined with (4.7) and (4.8), the reconstruction loss in the ELBO is replaced by using $\mathscr{L}_{\text{recon}}$ so the ELBO is derived as

$$\begin{aligned}\text{ELBO} &= \mathscr{L}_{\text{recon}} - D_{\text{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta(\varepsilon, \mathbf{z} \mid \mathbf{u})\big) \\ &= \alpha\,\text{BCE}(\mathbf{x}, \hat{\mathbf{x}}) + \nu\,\text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) - \mathbb{E}_{q_\phi}\Big[\log \frac{q_\phi(\varepsilon, \mathbf{z} \mid \mathbf{x}, \mathbf{u})}{p_\theta(\varepsilon, \mathbf{z} \mid \mathbf{u})}\Big] \quad (4.9)\end{aligned}$$

Besides, to ensure the identifiability of the adjacency matrix $\mathbf{A}$, the learning guidance $\mathbf{u}$ is used to guide the identifiability. $\mathbf{u}$ can be regulated by minimizing the loss function $l_u$:

$$l_u = \mathbb{E}_X \|\mathbf{u} - \sigma(\mathbf{A}^{\mathbb{T}}\mathbf{u})\|_2^2 \quad (4.10)$$

where the notation $\|\cdot\|_2^2$ represents the squared $L_2$-norm of a vector.

Learning with a similar method, $\mathbf{z}$ is regulated in the loss function by minimizing the loss function $l_z$:

$$l_z = \mathbb{E}_{\mathbf{z}\sim q_\phi}\left[\sum_{i=1}^{n} \|z_i - g_i(A_i \circ \mathbf{z}; \eta_i)\|_2^2\right] \quad (4.11)$$

To follow the training strategy of the $l_u$ and $l_z$, Causal DiffuseVAE minimizes the negating ELBO, which is unlike the standard VAE training objective of maximizing the ELBO. Overall, we have the new loss function

$$\mathscr{L}_{\text{VAE}} = -\text{ELBO} + \gamma l_u + \lambda\, l_z \quad (4.12)$$

where $\gamma$ and $\lambda$ are the regularization coefficients.

Therefore, we have

$$\begin{aligned}\mathscr{L}_{\text{VAE}} &= -\,\text{ELBO} + \gamma l_u + \lambda\, l_z \\ &= \underbrace{-\alpha\,\text{BCE}(\mathbf{x},\hat{\mathbf{x}}) \;-\; \nu\,\text{MSE}(\mathbf{x},\hat{\mathbf{x}})}_{\text{Reconstruction Loss}} \\ &\quad + \underbrace{\mathbb{E}_{q_\phi}\left[\log \frac{q_\phi(\boldsymbol{\varepsilon},\mathbf{z}\mid\mathbf{x},\mathbf{u})}{p_\theta(\boldsymbol{\varepsilon},\mathbf{z}\mid\mathbf{u})}\right]}_{\text{KL Divergence Regularization}} \\ &\quad + \gamma l_u + \lambda\, l_z\end{aligned} \tag{4.13}$$

#### 4.2.1.2 Learning Strategy of the Diffusion Model

In the Diffusion Model, $\mathbf{u}$ needs to be included. Specifically, the loss function will measure the discrepancy between the forward process and the reverse process. The forward process generates $\mathbf{x}_{1:T}$ conditioned on the additional information $\mathbf{y}$; $\mathbf{z}$; and $\mathbf{u}$, modeling how data evolve over time. The reverse process attempts to recover $\mathbf{x}_0$ from $\mathbf{x}_{1:T}$ to reconstruct the original data. This discrepancy can be factorized as

$$\mathscr{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{z}\sim q_\psi(\mathbf{z}|\mathbf{y},\mathbf{x}_0)}\left[\mathbb{E}_{q_\phi(\mathbf{x}_{1:T}|\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{x}_0)}\log\frac{q_\phi(\mathbf{x}_{1:T}\mid\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{x}_0)}{p_\psi(\mathbf{x}_{0:T}\mid\mathbf{y},\mathbf{z},\mathbf{u})}\right] \tag{4.14}$$

where $\psi$ is the learnable parameter of the Diffusion Model. Therefore, the loss function in (4.3) is obtained by combining (4.13) and (4.14).

### 4.2.2 Learning Strategy with Confounders

Confounding factors are variables that influence both the cause and the effect within a study, leading to potential biases in estimating causal relationships. The structure of the model with confounders is shown in Figure 4.4. These factors create a spurious association that can obscure the true causal effect or suggest a relationship where

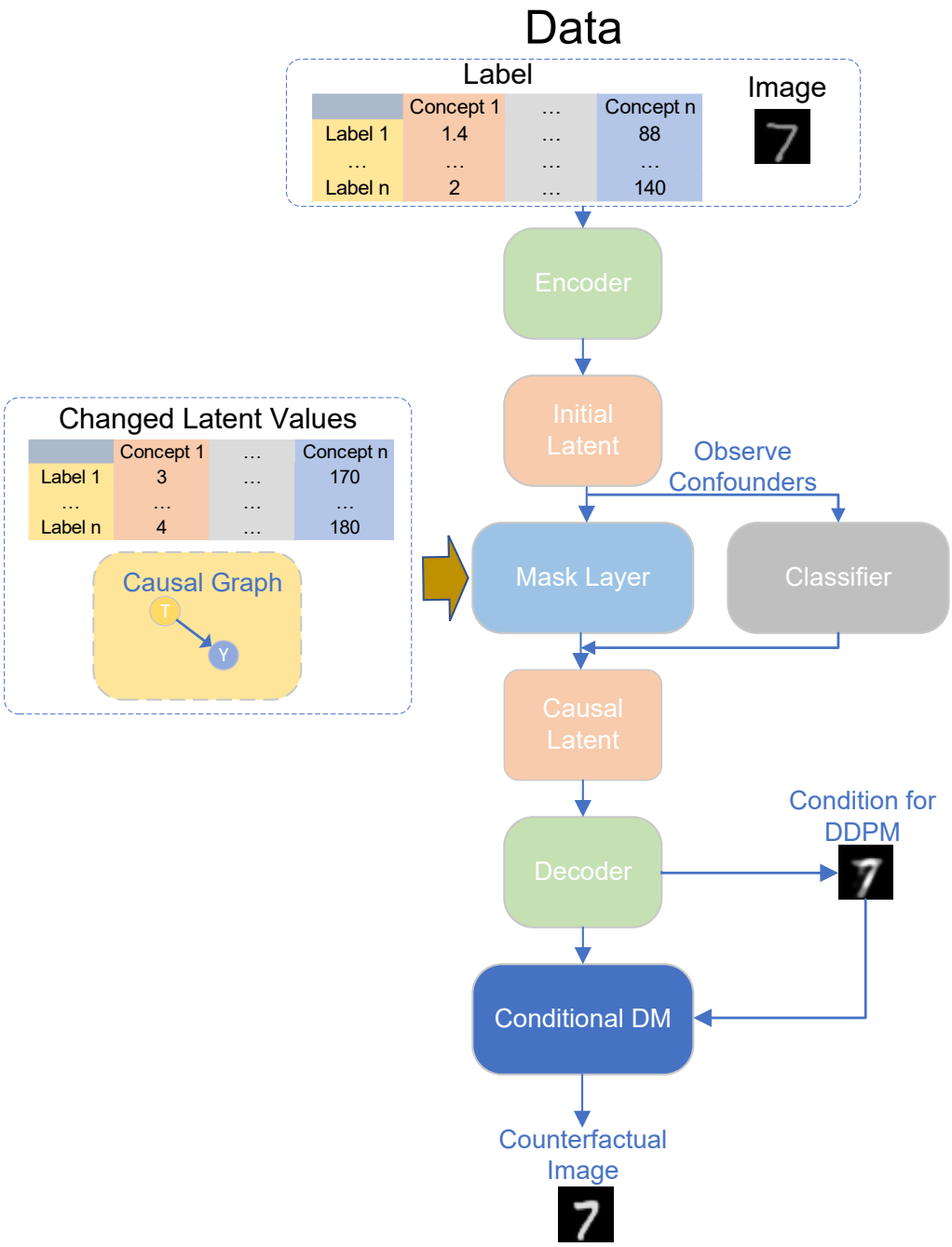



Figure 4.4: Architecture of Causal DiffuseVAE with confounders.

none exists. For example, in image generation tasks, confounders such as background patterns or lighting conditions can simultaneously affect both the input features (cause) and the target outputs (effect). This dual influence makes it challenging to disentangle genuine causal relationships.

A classifier is introduced after obtaining the initial latent representation to address the presence of confounding factors, particularly if these confounding factors are easily identifiable. The mechanism of the classifier is represented as

$$\mathbf{z} = \mathbf{W} \cdot \mathbf{x}_0 + \boldsymbol{\varepsilon}_{\text{classifier}}, \tag{4.15}$$

$$\mathbf{k}_{\text{classifier}} = \mathbf{M} \cdot \mathbf{z} + \eta_{\text{classifier}}, \tag{4.16}$$

where $\mathbf{W}$ is the weight matrix, $\mathbf{x}_0$ is the input image, $\boldsymbol{\varepsilon}_{\text{classifier}}$ denotes independent Gaussian noise in the classifier, $\mathbf{k}_{\text{classifier}}$ is the output of the classifier, and $\mathbf{M}$ and $\eta_{\text{classifier}}$ are additional parameters. The classifier utilizes a fully connected network and applies cross-entropy as the loss function, effectively distinguishing between relevant features and confounding factors within the latent space that encodes abstract data features. The loss function $\mathscr{L}_{\text{CE}}$ is defined as

$$\mathscr{L}_{\text{CE}} = -\sum_{i=1}^{N_{\text{sample}}} \sum_{c=1}^{C} l_{i,c} \log \hat{l}_{i,c}, \tag{4.17}$$

where $N_{\text{sample}}$ is the number of samples, $C$ is the number of classes, $l_{i,c}$ is the true label, and $\hat{l}_{i,c}$ is the predicted probability that sample $i$ belongs to class $c$.

By doing so, the classifier helps isolate the true causal variables, ensuring that the model focuses on meaningful causal relationships rather than spurious associations introduced by confounders. This step enhances the robustness and interpretability of the model's outputs, ultimately supporting more accurate and trustworthy image generation. The overall loss function could be represented as

$$\mathscr{L} = \mathscr{L}_{\text{VAE}} + \mathscr{L}_{\text{DDPM}} + \mathscr{L}_{\text{CE}} \tag{4.18}$$

where the $\mathscr{L}_{\text{VAE}}$ is same as (4.13), $\mathscr{L}_{\text{DDPM}}$ is same as (4.14) and $\mathscr{L}_{\text{CE}}$ is same as (4.17).

**Algorithm 1** Causal DiffuseVAE Inference

**Input:** (image, label) pairs $(\mathbf{x}_0, \mathbf{u})$, number of concepts $n$
**Output:** Generated counterfactual $\mathbf{x}_0^{\mathrm{DM}}$

Sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
**for** $i = 1$ to $n$ **do**
  **if** $i =$ Intervention variable index **then**
    $\mathbf{z} \leftarrow$ Desired value
  **else**
    $\mathbf{z} = g_i(A_i \circ \mathbf{z}, \eta_i) + \varepsilon_i$
  **end if**
**end for**
$\hat{\mathbf{x}}_0 \leftarrow p_\theta(\mathbf{x}_0 \mid \mathbf{z}, \varepsilon, \mathbf{u})$
$\mathbf{x}_T^{\mathrm{DM}} \leftarrow \mathbf{x}_T^{\mathrm{DM}} + \hat{\mathbf{x}}_0$
**for** $t = T$ to 1 **do**
  Sample noise $\varepsilon_t \sim \mathcal{N}(0, I)$
  $\mathbf{x}_{t-1}^{\mathrm{DM}} \leftarrow \frac{\mathbf{x}_t^{\mathrm{DM}} - \sqrt{\beta_t} \cdot \varepsilon_t}{\sqrt{1-\beta_t}}$
**end for**
**Return** $\mathbf{x}_0^{\mathrm{DM}}$

The Causal DiffuseVAE is a weakly supervised learning method. Although the learning guidance $\mathbf{u}$ is utilized during the training, the provided labels lack full precision, as they do not specify the exact regions of the labeled features. This uncertainty in the labeling process contributes to the weakly supervised nature of the model, where the supervision is incomplete or inexact, making the learning process more challenging. With the pre-trained Causal DiffuseVAE model, the counterfactual images are generated by changing the value of one specific causal variable in $n$ concepts. Then, the generating process is continued, which is elaborated in Algorithm 1 with the following steps. First, sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. For each concept index $i$, set $z_i$ to the desired value if $i$ is the intervention variable; otherwise, compute (4.1). Next, combine $\mathbf{z}$, $\varepsilon$, and learning guidance $\mathbf{u}$ in the decoder to produce $\hat{\mathbf{x}}_0$. Afterwards, add $\hat{\mathbf{x}}_0$ to the noisy diffusion state $\mathbf{x}_T^{\mathrm{DM}}$, and perform the reverse process for $T$ steps by sampling $\varepsilon_t$ and updating

$$\mathbf{x}_{t-1}^{\mathrm{DM}} = \frac{\mathbf{x}_t^{\mathrm{DM}} - \sqrt{\beta_t}\,\varepsilon_t}{\sqrt{1-\beta_t}} \tag{4.19}$$

The final output $\mathbf{x}_0^{\mathrm{DM}}$ is the generated counterfactual image reflecting the applied causal intervention.

Table 4.1: Network Design of the Diffusion Model (DM) in the Causal DiffuseVAE for MNIST and Flow Datasets

| **Detail** | **MNIST** | **Flow** |
|---|---|---|
| Base channels | 128 | 128 |
| Channel multipliers | [1, 2, 2] | [1, 2, 4, 8] |
| Training set | 60k | 8k |
| Test set | 10k | 2k |
| Image resolution | $28 \times 28 \times 1$ | $96 \times 96 \times 4$ |
| $\beta_1$ | 0.0001 | 0.0001 |
| $\beta_2$ | 0.02 | 0.02 |
| Diffusion loss | MSE | MSE |
| Optimizer | Adam | Adam |
| Epochs | 1000 | 800 |
| Learning rate | $10^{-4}$ | $10^{-4}$ |

Furthermore, with the loss functions, the Causal DiffuseVAE could obtain the desired causal latent. The intervention is deployed during the testing to evaluate the performance of the Causal DiffuseVAE. The intervention process refers to actively changing or setting the value of a variable (or variables) to observe how this manipulation affects other variables in the system. This process is foundational for distinguishing causation from correlation, as it allows researchers to observe the direct effects of changes in one variable on others. The intervention process is commonly denoted by *do*-calculus, where an intervention is represented as $do(X = x)$. This operation means that the variable $X$ is set to the value $x$, irrespective of its natural causes.

## 4.3 Experiment Setting

The experiments are deployed on a server with an Ubuntu 20.04 operating system and two NVIDIA RTX A6000 graphics cards. $\alpha$ and $\nu$ in (4.8) are set to 0.7 and 0.3, respectively. At the same time, $\gamma$ and $\lambda$ in (4.12) are set to 0.01 and 0.1, respectively. The parameters used in the MNIST and Flow Datasets are shown in Table 4.1. The detailed architecture of the models for the CelebA Dataset and Pendulum Dataset can be found in Table 4.2 and 4.3.

Table 4.2: Network Design of CausalVAE Encoders and Decoders for Smile, Age, and Pendulum Datasets

| Dataset | Encoder | Decoder |
|---|---|---|
| Smile and Age | $3 \times 128^2 \to 32 \times 64^2$ | $512 \to 512 \times 8^2$ |
| | Conv2d + ReLU | ConvT2d + ReLU |
| | $32 \times 64^2 \to 64 \times 32^2$ | $512 \times 8^2 \to 256 \times 16^2$ |
| | Conv2d + ReLU | ConvT2d + ReLU |
| | $64 \times 32^2 \to 128 \times 16^2$ | $256 \times 16^2 \to 128 \times 32^2$ |
| | Conv2d + ReLU | ConvT2d + ReLU |
| | $128 \times 16^2 \to 256 \times 8^2$ | $128 \times 32^2 \to 64 \times 64^2$ |
| | Conv2d + ReLU | ConvT2d + ReLU |
| | $256 \times 8^2 \to 512 \times 4^2$ | $64 \times 64^2 \to 32 \times 128^2$ |
| | Conv2d + ReLU | ConvT2d + ReLU |
| | $512 \times 4^2 \to 512 \times 1^2$ | $32 \times 128^2 \to 3 \times 128^2$ |
| | Conv2d + ReLU | ConvT2d + Sigmoid |
| Pendulum | $4 \times 96^2 \to 900$ | $4 \times (512 \to 300)$ |
| | Linear + ELU | Linear + ELU |
| | $900 \to 300$ | $4 \times (300 \to 300)$ |
| | Linear + ELU | Linear + ELU |
| | $300 \to 2 \times 512$ | $4 \times (300 \to 1024)$ |
| | Linear + ELU | Linear + ELU |
| | – | $4 \times (1024 \to 4 \times 96^2)$ |
| | | Linear |

Table 4.3: Details of the Diffusion Model

| Parameter | Smile | Age | Pendulum |
|---|---|---|---|
| Batch size | 16 | 16 | 64 |
| Base channels | 128 | 128 | 128 |
| Channel multipliers | [1, 2, 2, 2, 4] | [1, 2, 2, 2, 4] | [1, 2, 4, 8] |
| Training set | 17k | 17k | 8k |
| Test set | 3k | 3k | 2k |
| Image resolution | $128 \times 128 \times 3$ | $128 \times 128 \times 3$ | $96 \times 96 \times 4$ |
| Size of causal variables | 512 | 512 | 512 |
| $\beta_1$ | 0.0001 | 0.0001 | 0.0001 |
| $\beta_2$ | 0.02 | 0.02 | 0.02 |
| Learning rate | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ |
| Optimizer | Adam | Adam | Adam |
| Diffusion steps | 1000 | 1000 | 1000 |
| Epoch | 500 | 500 | 1000 |
| Diffusion loss | MSE | MSE | MSE |

### 4.3.1 Experimental Dataset

Eight distinct datasets are selected for evaluation. In [168], a dataset generation method is proposed for shadow analysis. Two synthetic shadow datasets were created using this method with Blender, each containing 10,000 images (8,500 for training, 1,000 for validation, and 500 for testing). Each image includes a light source, an object, and its shadow. Variations in light source size and object properties systematically influence shadow formation, making the dataset ideal for causal analysis.

The MNIST dataset [169] comprises 70,000 grayscale images of handwritten digits (0–9) at 28×28 pixels, divided into 51,000 for training, 9,000 for validation and 10,000 for test images, serving as a standard benchmark for image recognition and deep learning.

For general dataset evaluation, Causal DiffuseVAE was tested on the Pendulum and Flow datasets [166], each containing 7,000 images in Red, Green, Blue, and Alpha (RGBA) format. The Pendulum Dataset (5950 for training, 700 for validation, 350 for testing) captures causal interactions between pendulum angle, light angle, and shadow characteristics. The Flow Dataset (5,100 for training, 900 for validation, 1,000 for testing) simulates fluid dynamics as a ball interacts with liquid in a broken vessel, with transparency effects aiding visualization.

For real-world validation, the CelebA, a large scalable dataset, is used for facial attribute analysis [170]. CelebA is a large-scale facial attributes dataset widely used in computer vision and machine learning for face recognition, attribute prediction, and causal representation learning. The dataset includes a variety of facial attributes, with annotations for over 200,000 celebrity images, which are in RGB format. Two notable sub-datasets within CelebA are CelebA(SMILE) and CelebA(BEARD). Each dataset consists of 20,000 images, with 70% allocated for training, 15% for validation, and the remaining 15% reserved for testing. Then CelebA(SMILE) focuses on the attributes of Gender, Smile, Eyes Open, and Mouth Open, allowing for the exploration of causal relationships and interactions between these facial expressions. On the other

hand, CelebA(BEARD) centers around Age, Gender, Baldness, and Beard, providing a framework to study how these attributes influence one another, particularly in the context of age and gender-related features. Both sub-datasets are essential for understanding and disentangling the complex relationships between facial attributes in various applications. For these two datasets, only the facial parts in the images are focused on, so the images are cropped and resized to be square.

For validation of the industry situation, the Causal Circuit dataset was used [171]. The CausalCircuit dataset comprises $512 \times 512$ RGB images of a robot arm interacting with a causally connected circuit of buttons and lights, capturing four underlying causal variables. The dataset is divided into an 80% of training set and a 20% of test set, with no separate validation subset employed.

The datasets are selected to progressively evaluate the proposed model under increasing levels of causal complexity, visual realism, and practical relevance. Synthetic shadow datasets provide a fully controlled environment with known ground-truth causal relationships between lighting, object properties, and shadows, allowing direct validation of causal interventions. MNIST serves as a low-dimensional benchmark with inherent confounding between digit identity and visual attributes, enabling assessment of disentanglement under spurious correlations. The Pendulum and Flow datasets introduce physically grounded, multi-factor causal systems with richer visual structure, testing scalability to more complex interactions. CelebA further extends evaluation to real-world images with weak supervision and highly correlated semantic attributes, reflecting realistic causal representation challenges. Finally, the Causal Circuit dataset represents an industry-inspired control scenario, validating the model's ability to generate trustworthy counterfactuals in environments resembling real-world decision-making systems. Together, this progression ensures that improvements are not limited to idealized settings but generalize across diverse causal and visual regimes.

### 4.3.2 Experimental Setting

The most popular methods in the causal generative model are employed as baselines in evaluation, including CausalVAE [166], Causal disentangled representation learning for missing data (CDRM) [167], Causal Diffusion Autoencoder (CDAE) [118] and Conditional Diffusion Models (CDM) [172].

CausalVAE extends the variational autoencoder framework by explicitly imposing a structural causal model over the latent variables. A directed acyclic graph constrains the dependencies among latent factors, enabling causal interventions to be performed directly in latent space. While CausalVAE provides interpretable and identifiable latent representations, its reliance on a pixel-level reconstruction objective limits image fidelity, often producing blurred outputs in visually complex scenarios.

CDRM (Causal Disentangled Representation Learning for Missing Data) is designed to learn causal latent representations under incomplete or partially observed data. It integrates causal graphs with a VAE-based generative model to support counterfactual reasoning and data imputation. Although CDRM achieves robust causal disentanglement and stable training, it inherits the visual limitations of VAEs and therefore struggles to generate high-quality or photorealistic images.

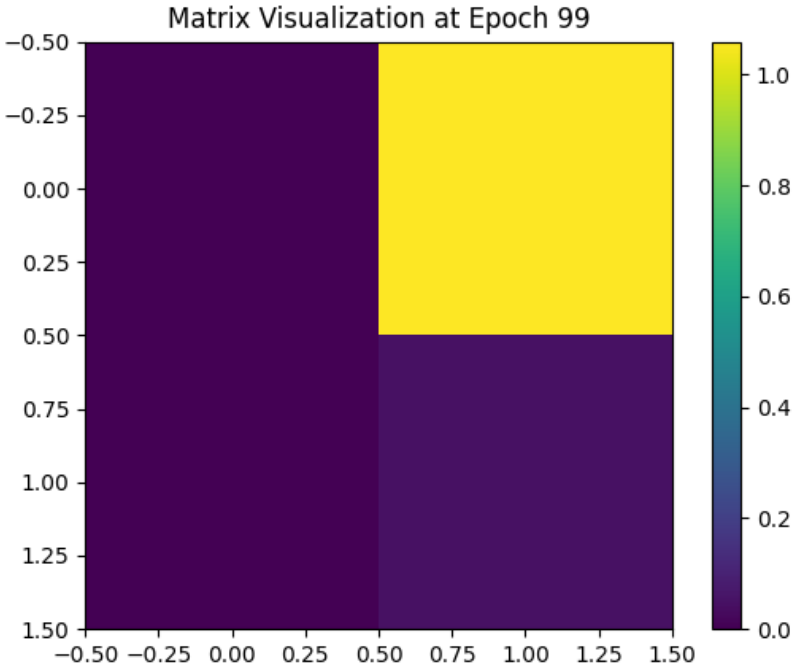
CDAE (Causal Diffusion Autoencoder) combines causal representation learning with diffusion-based generation. It first encodes images into a causal latent space using an autoencoder and then applies a diffusion model conditioned on these latents to generate counterfactual images. CDAE improves visual fidelity compared to VAE-only methods and supports causal interventions, but it lacks a fully probabilistic encoder with reparameterization-based inference, which can reduce stability and uncertainty modeling in the learned representations.

Conditional Diffusion Models (CDM) represent a class of diffusion-based generators that condition the denoising process on auxiliary information such as labels or attributes. These models excel at producing high-resolution and photorealistic images, but they do not incorporate an explicit causal latent structure. As a result, conditional edits are driven by statistical correlations rather than causal mechanisms, limiting their suitability for trustworthy counterfactual reasoning.

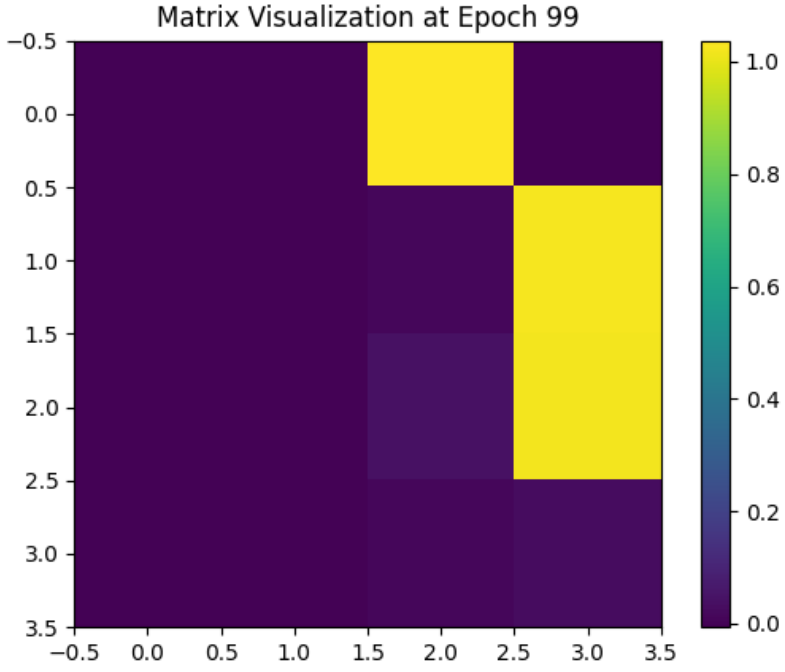
The evaluation is based on two primary metrics: MAE and LPIPS. The details of the MAE and LPIPS could be found in Section 3.7.1. In the MNIST dataset, it is found that the digit itself is the confounder in the causal relationship. It implies that when the cause is changed, the digit is also changed. To address the confounder, the classifier is introduced into the Causal DiffuseVAE. Furthermore, no observed confounder is found in the Flow dataset so there is no classifier in the model.

### 4.3.3 Experimental Results

In Figure 4.5, the images appear to be a heatmap visualization of a matrix at the $100^{\text{th}}$ epoch. In Figure 4.5(a), on both axes, the section from -0.50 to 0.50 indicates the cause. At the same time, the section from 0.50 to 1.50 indicates the effect. The yellow part illustrates that the cause could influence the effect, which means the factor thickness could influence the intensity. It is similar to Figure 4.5(b). On both axes, the section from -0.50 to 0.50 is the factor ball size. The section from 0.50 to 1.50 is the factor hole. Moreover, the section from 1.50 to 2.50 is the factor of water height. The section from 2.50 to 3.50 is the factor of water flow. The results in Figure 4.5(a) and Figure 4.5(b) prove the model learns correct causal relationships.




(a) The causal matrix A of the MNIST Dataset



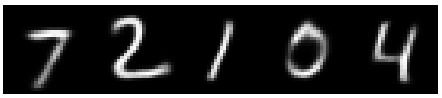

(b) The causal matrix A of the Flow Dataset

Figure 4.5: Causal matrix A at the $100^{\text{th}}$ epoch

(a) Original Images from the MNIST Dataset

(b) do(Thickness = 2) from Causal DiffuseVAE

(c) do(Thickness = 4) from Causal DiffuseVAE



(d) do(Thickness = 2) from Conventional CausalVAE

(e) do(Thickness = 4) from Conventional CausalVAE

Figure 4.6: Results on the MNIST Dataset using two different methods

#### 4.3.3.1 Results of the MNIST Dataset

To evaluate the results of the intervention, the experiments are deployed on two datasets. For the MNIST, the range of the label thickness is from 1 to 5.8 and the range of the intensity is from 67 to 255. The results of the intervention are shown in Figure 4.6. The images in Figure 4.6a are the original images from the dataset. The images in Figure 4.6b and Figure 4.6c are the results produced by the Causal DiffuseVAE when $do(\text{Thickness} = 2)$ and $do(\text{Thickness} = 4)$, respectively. To evaluate the performance of the Causal DiffuseVAE. The images in Figure 4.6d and Figure 4.6e are the results from the conventional CausalVAE with the same intervention. The ideal intervention occurs when a decrease in thickness leads to a corresponding decrease in intensity and an increase in thickness results in a proportional increase in intensity. The results of

the Causal DiffuseVAE correspond to the ideal intervention. However, for the conventional CausalVAE, the change of intensity is not clear. Moreover, when the thickness is higher, the images show some errors in the background, which do not appear in the Causal DiffuseVAE.

#### 4.3.3.2 Shadow Dataset

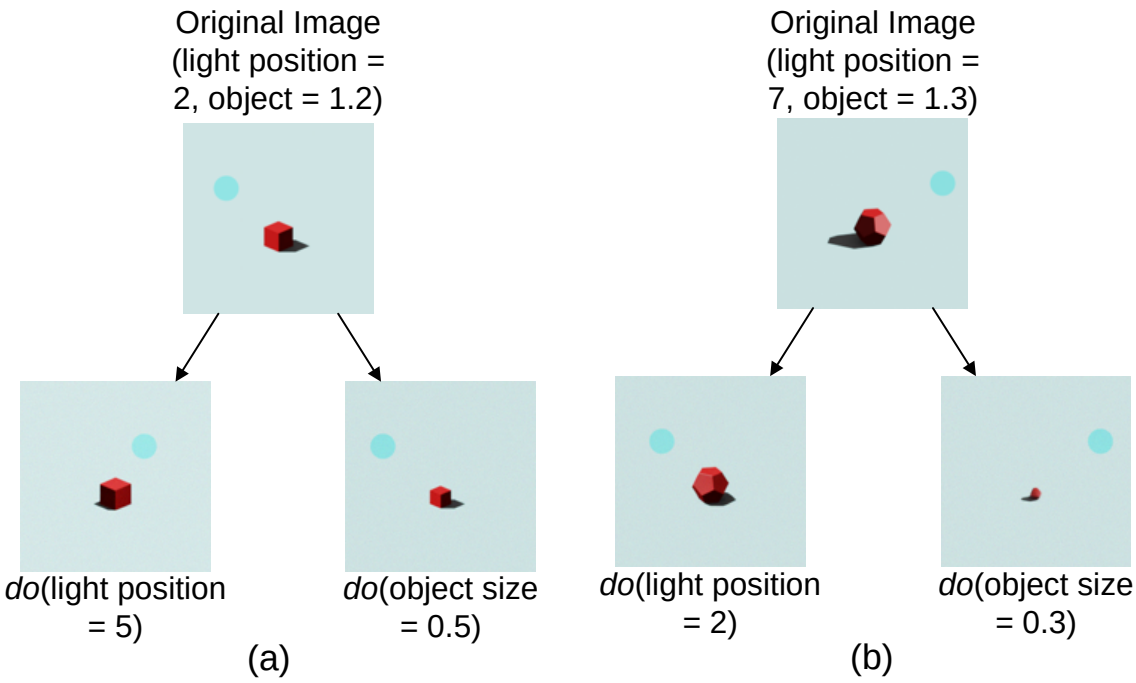



Figure 4.7: Intervention results of shadow datasets using Causal DiffuseVAE. The object in (a) is the cube and in (b) is the polyhedron.

Intervention results on the Shadow Dataset are presented in Figure. 4.7, where shadows cast by a cube and a polyhedron under light positions from 1 to 8 and object sizes from 0.3 to 1.7 are displayed. As the light position and object size are varied, the shadow area is adjusted accordingly. The shape and the direction of the shadow areas demonstrate that Causal DiffuseVAE accurately models both shadow formation and object geometry.

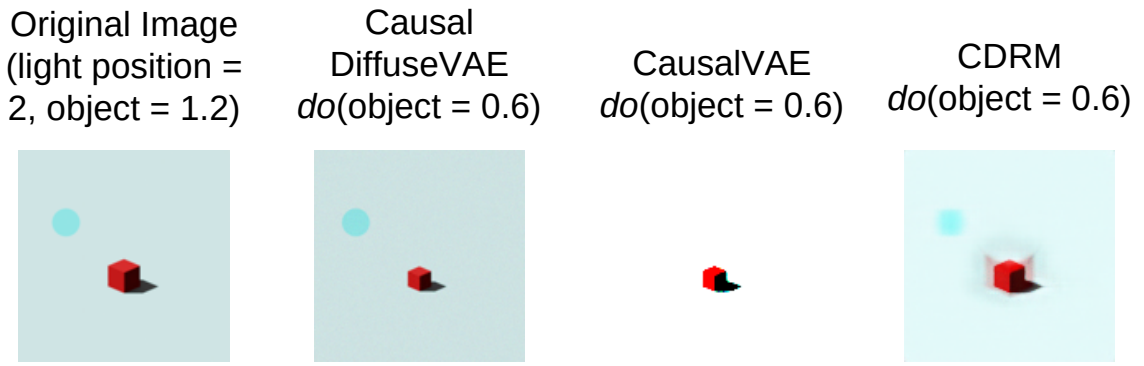



Figure 4.8: Counterfactual images generated by Causal DiffuseVAE, CausalVAE and CDRM.

The intervention performance of the Causal DiffuseVAE, CausalVAE, and CDRM on the 3-dimensional dataset is presented in Figure. 4.8. When the object size changes, CausalVAE fails to generate the counterfactual 3-dimensional image due to the absence of the light source. CDRM successfully reconstructs all factors but fails to capture the underlying causal relationships. In contrast, Causal DiffuseVAE outperforms both methods by achieving high-quality reconstructions while accurately preserving causal dependencies.

#### 4.3.3.3 Flow

Intervention results on the Flow Dataset are presented in Figure. 4.9, where ball sizes vary from 5 to 35, hole sizes from 0 to 4, water heights from 33 to 85, and water flow rates from 6 to 15. As the ball size is tuned, water height is adjusted accordingly, which in turn influences the water flow. These variations follow physical laws, confirming that causal relationships have been accurately captured by Causal DiffuseVAE. Although CausalVAE and CDRM can also learn causal relationships, they fail to reconstruct the details, like the color and the hole of the vessel.

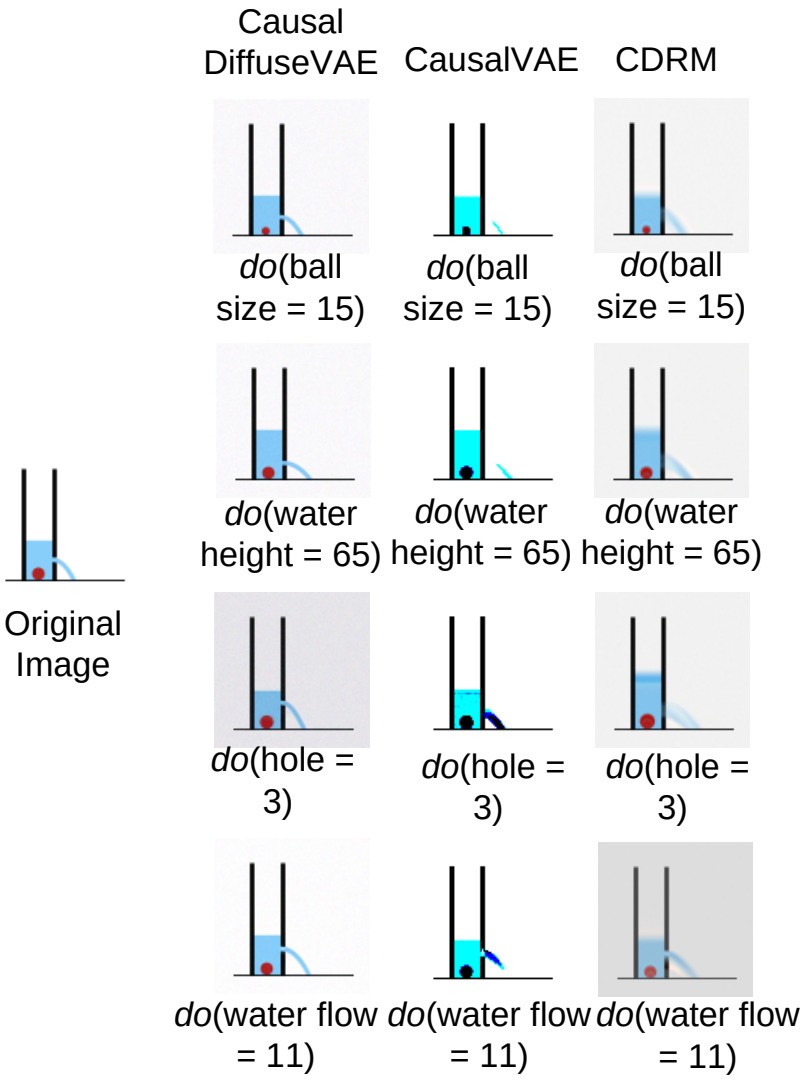



Figure 4.9: Intervention results of flow datasets using Causal DiffuseVAE, CausalVAE and CDRM.

#### 4.3.3.4 Pendulum

In the Pendulum Dataset, for each causal variable, the value of the label is in a fixed range. For the pendulum angle, the label values are set from -44 to 44, which is the angle between the pendulum and the vertical line. For the light position, shadow length and shadow position, the label values are set from 60 to 140, 3 to 10, and 5 to 15, respectively. Figure. 4.10 illustrates the result of the intervention using Causal DiffuseVAE,

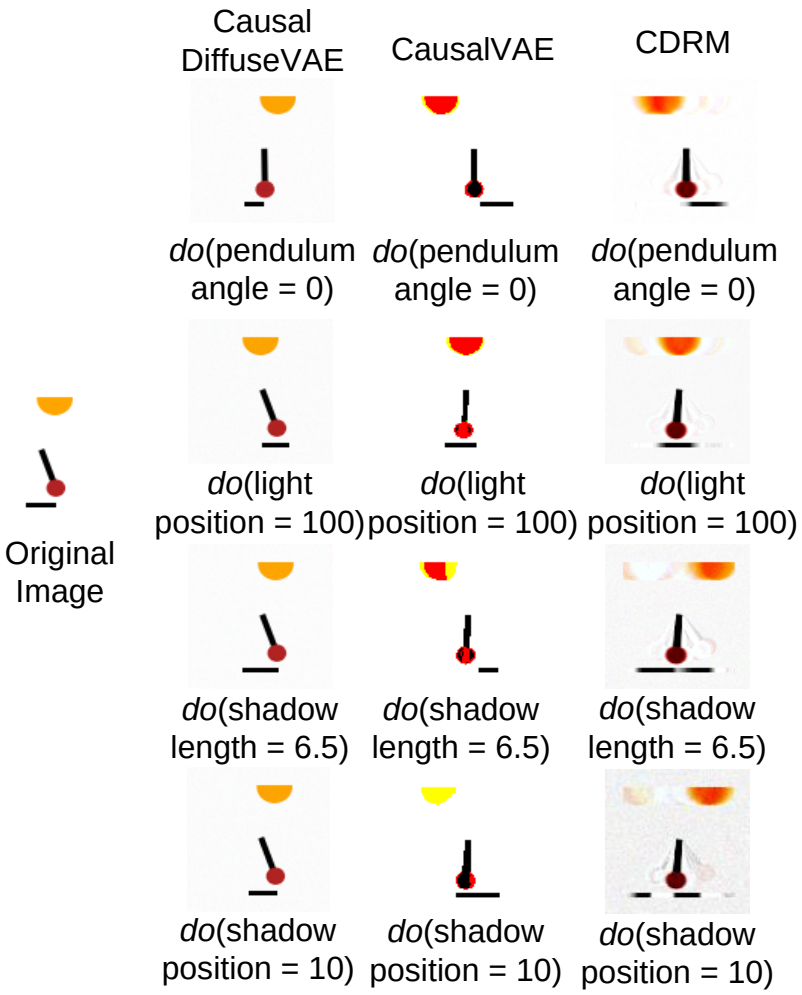



Figure 4.10: Intervention results of pendulum datasets using Causal DiffuseVAE, CausalVAE and CDRM.

CausalVAE and CDRM. Interventions on the pendulum angle and the light position lead to changes in the shadow length and shadow position. The results are similar to the results of the flow dataset. The CausalVAE and the CDRM fail to learn the details of the color and the shadow well.

#### 4.3.3.5 CelebA

Figure. 4.11 shows the generated counterfactual images with the CelebA (Gender and Age) Dataset. Although the value of the labels is limited between -1 and 1, the "DO" operation can use the values from -1 to 1. When intervening the gender, from -1 to 1, the faces in the generated images changed from female to male, which corresponds to

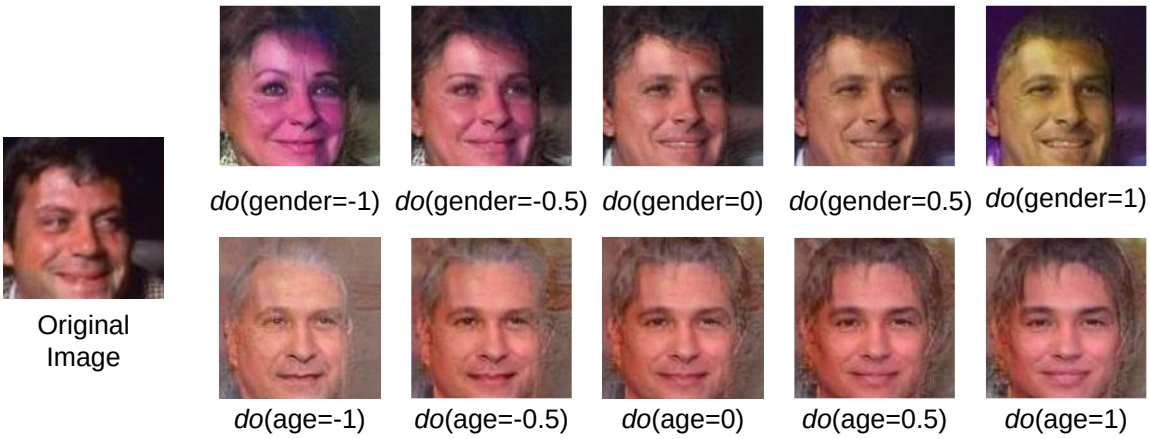



Figure 4.11: Intervention results of CelebA datasets (Gender and Age) using Causal DiffuseVAE.

the labels. With the change of the gender, the hair and the beard are also changed. The female's hair is more than the male's. The female has no beard while the male has. When intervening the age, from -1 to 1, the faces changed from an old person to a young person. The old person's hair is less than the young's. The beard on the old person's face is clearer than the young person's.

(a) Real image


(b) Results of the Causal DiffuseVAE on the Smile Dataset


(c) Results of the conventional CausalVAE on the Smile Dataset

Figure 4.12: Results on the Smile Dataset

Table 4.4: MAE Comparison for the Smile Dataset

| **"DO" operation** | **Causal DiffuseVAE** | **Conventional CausalVAE** |
|---|---|---|
| *do*(gender = -1) | **0.203** | 0.670 |
| *do*(gender = 1) | **0.142** | 0.673 |
| *do*(smile = -1) | **0.068** | 0.648 |
| *do*(smile = 1) | **0.102** | 0.618 |
| *do*(eyes open = -1) | **0.075** | 0.608 |
| *do*(eyes open = 1) | **0.063** | 0.650 |
| *do*(mouth open = -1) | **0.075** | 0.607 |
| *do*(mouth open = 1) | **0.060** | 0.651 |

Similar to the Age Dataset, the relationships among four causal variables, gender, smile, eyes open and mouth open are represented in Figure 4.12(b). The labels used in the Age Dataset are also limited to two values -1 and 1. The value -1 means female, no smile, eyes open and mouth closed. The value 1 means male, smile, narrow eyes and mouth open. The comparison between the Causal DiffuseVAE and conventional CausalVAE is presented in Table 4.4, which proves that the performance of the Causal DiffuseVAE on the Smile Dataset is also better than the baselines. Figure 4.12 shows the comparison among the real image, the generated counterfactual images of the Causal DiffuseVAE and the generated counterfactual images of the conventional CausalVAE. By comparing the results in Figure 4.12(b) and Figure 4.12(c), the resolution of images in Figure 4.12(b) is higher than Figure 4.12(c), which corresponds to the results of the Fréchet Inception Distance (FID) scores. Furthermore, when *do*(gender) is applied, the eyes change while the mouth stays the same. When *do*(smile) is applied, both the eyes and the mouth change. This result corresponds to the causal graph. Moreover, when *do*(mouth open) and *do*(eyes open) are applied, only the mouth open and eyes open are changed, rather than the smile and the gender are not changed. This result proves the unidirectional character as the causal graph should be a DAG.

#### 4.3.3.6 Causal Circuit

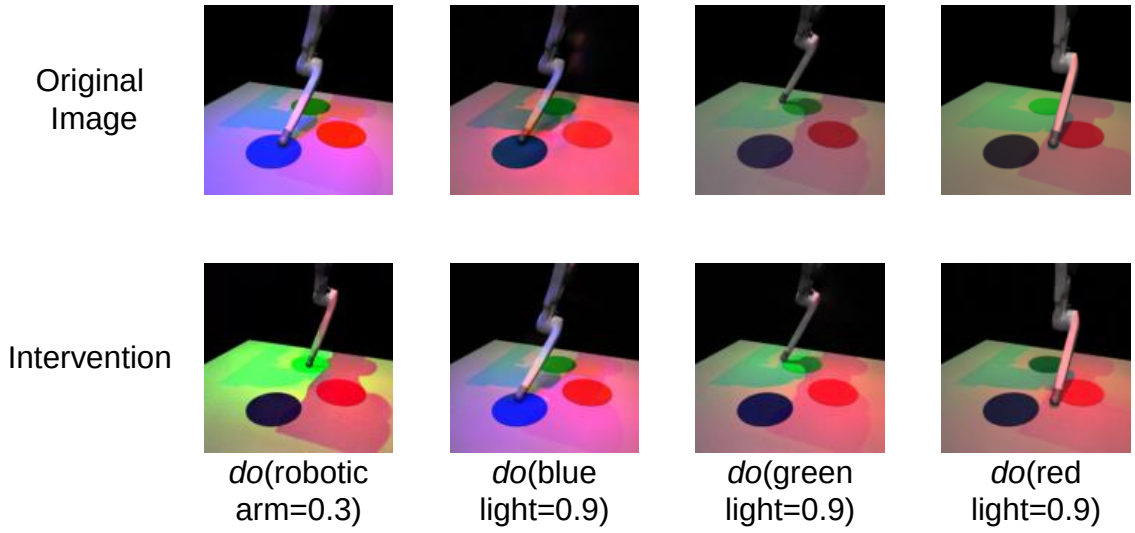



Figure 4.13: Intervention results of Circuit datasets using Causal DiffuseVAE.

To evaluate the generation capability of the Causal DiffuseVAE in the industry situation, the Causal DiffuseVAE is trained using the Causal Circuit Dataset. Figure. 4.13 shows the generated counterfactual images. Based on the causal relationships, when the robotic arm moves to a location, the corresponding light will be on or off. When the intensity of blue light or green light is changed, the intensity of the red light must be changed. When the intensity of the red light changed, the other lights would remain unchanged. As the intensity of the light increases, the light becomes brighter. In the results, when the robotic arm moves to the green light, the blue light is off, and the green light is on. When the intensity of blue light is tuned to 0.9, the red light and the blue light are on. The intervention of the green light is the same as the blue light. However, when the intensity of the red light is modified to 0.9, only the red light is on. These results follow the causal relationships among these factors.

Table 4.5: Comparison on capabilities of the Causal DiffuseVAE and other baseline methods

| Capabilities | Causal DiffuseVAE | CausalVAE | CDRM | CDAE | conditional DDPM |
|---|---|---|---|---|---|
| Explicit low-dimensional causal latent | ✓ | ✓ | ✓ | ✓ | × |
| High-fidelity image synthesis | ✓ | × | × | ✓ | ✓ |
| Stable training via reparameterization | ✓ | ✓ | ✓ | × | × |
| Trustworthy causal counterfactuals | ✓ | ✓ | ✓ | ✓ | × |
| Diffusion-based architecture | ✓ | × | × | ✓ | × |

#### 4.3.3.7 Result Analysis

Table 4.5 shows that only Causal DiffuseVAE offers all of the following: an explicit low-dimensional causal latent; photorealistic outputs; stable end-to-end training via the reparameterization trick; reliable counterfactual generation even with incomplete observations; and a diffusion-based pipeline. No other baseline model combines this full set of capabilities.

By contrast, CausalVAE also provides an interpretable causal latent and benefits from stable, probabilistic training, but lacks the high-fidelity synthesis afforded by diffusion decoders. CDRM retains a structured latent space and supports stable training. It is designed to impute missing entries and generate consistent counterfactuals from incomplete data. However, because it relies solely on VAE architecture, it cannot produce photorealistic images. CDAE recovers both a causal latent and high-quality diffusion synthesis, supporting reliable counterfactuals. It lacks the VAE's reparameterization-based encoder and full Bayesian uncertainty modeling. Finally conditional DDPMs excel at raw image fidelity. However, they lack an explicit causal latent, cannot handle missing data in a principled way for counterfactual inference, and do not support stable, one-shot latent encoding.

At a high level, all of these models optimize composite loss functions that balance pixel-level fidelity with latent-space regularization. In the meantime, they also incorporate diffusion or imputation objectives to achieve both high-quality synthesis and principled causal reasoning.

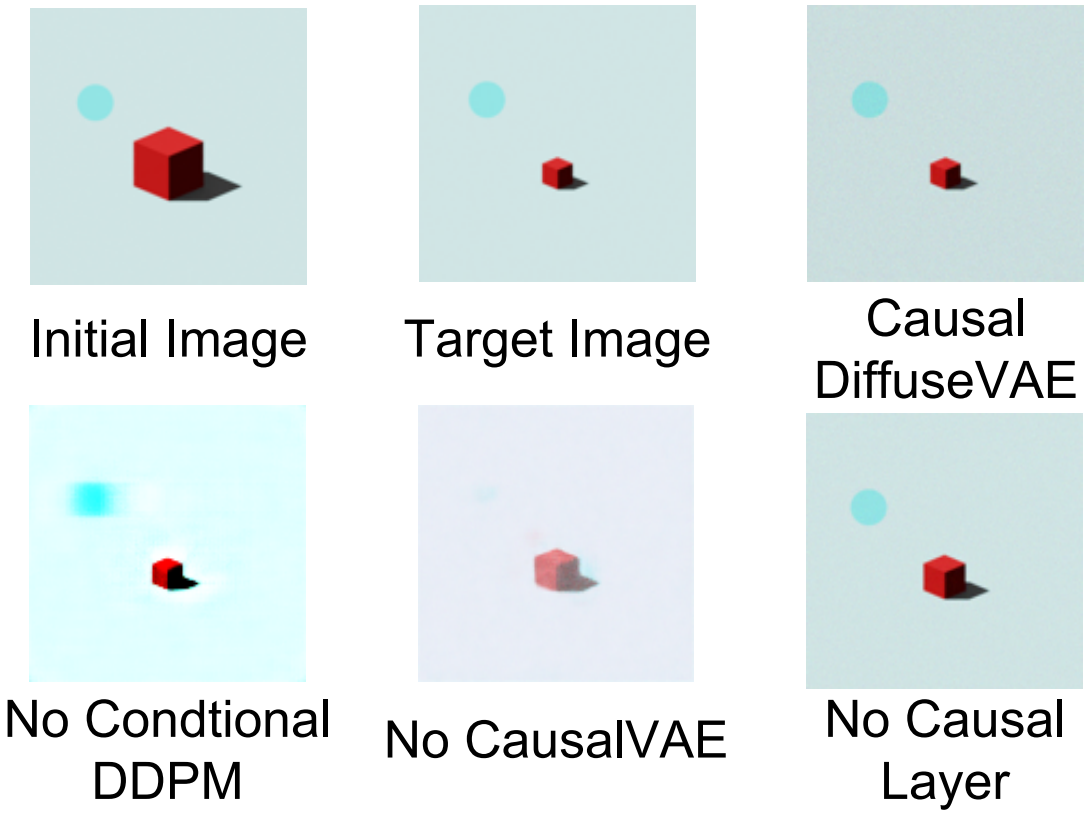



Figure 4.14. Ablation results on the shadow dataset.

Table 4.6. LPIPS comparison of ablation results

| **Experiment settings** | **LPIPS** |
|---|---|
| Causal DiffuseVAE | $\mathbf{0.0185 \pm 0.0140}$ |
| Without Diffusion Decoder | $0.0483 \pm 0.0230$ |
| Without Causal Layer | $0.0503 \pm 0.0240$ |
| Without CausalVAE Module | $0.1410 \pm 0.1070$ |

The results of the ablation experiments are summarized in Figure. 4.14. When the diffusion decoder is removed, blurring is observed in the images, indicating decreased quality. When both the VAE and causal layer are removed, color and lighting details are lost and object sizes are not adjusted correctly. When only the causal layer is removed, visual details are retained but adjustments of object size are unsuccessful. These findings confirm that each component of Causal DiffuseVAE is indispensable.

The quantitative comparison in Table 4.6 shows that the full model achieves the best perceptual similarity, while each ablated setting leads to a degradation in image quality. These findings confirm that every component of Causal DiffuseVAE is essential for achieving high-fidelity and causally consistent generation.

Moreover, Causal DiffuseVAE jointly minimizes the revised **ELBO** of $\mathscr{L}_{\text{VAE}}$, in (21), and a latent-conditioned diffusion variational bound. An additional acyclicity penalty is comprised on its causal mask, producing reliable reconstructions and a well-structured DAG in the latent. CausalVAE and CDRM likewise optimizes a modified ELBO augmented by the same acyclicity constraint on its adjacency matrix, enforcing a valid causal graph over its latents but cannot generate high-quality images without diffusion-based models. CDAE adds a supervised alignment loss on labeled semantic factors to its autoencoder, then conditions a Denoising Diffusion Implicit Model (DDIM)-style denoising objective on those factors to enable "DO" intervention counterfactual sampling. With the autoencoder, the model may generate overfitting results. Finally, conditional DDPMs are trained by minimizing the weighted diffusion variational bound across all timesteps, fitting a conditional denoiser for image synthesis but without any explicit causal or DAG regularization.

The principal component analysis (PCA)-based scatter plot analysis of comparison in Figure 4.15 presents a comparison of the distributions formed by real samples and generated counterfactual samples across different models, enabling a direct assessment of how well each method preserves the underlying data structure after generation. For CausalVAE, the generated samples cluster tightly around a limited region of the PCA space, indicating that although the model learns a compact latent representation, it fails to capture the full variability of the data. This collapse reflects the well-known smoothing effect of VAE reconstruction loss, which limits its ability to represent fine-grained causal variations. In the case of CDRM, the generated samples exhibit noticeable shifts relative to the real data clusters, with partial overlap but clear displacement along principal components. This suggests that while CDRM captures some causal dependencies, inaccuracies accumulate when reconstructing complex visual attributes,

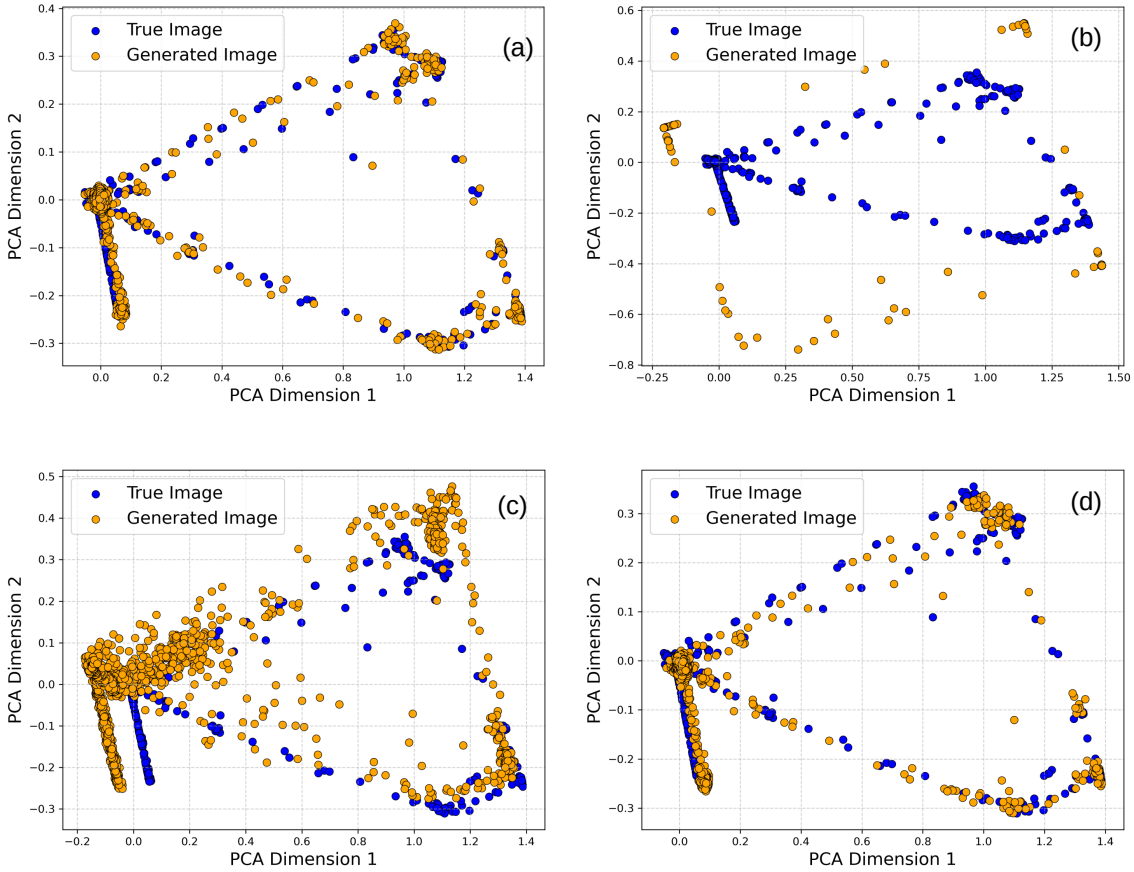



Figure 4.15. Scatter plot of generated data and true data after PCA-based processing. (a) Causal DiffuseVAE. (b) CausalVAE. (c) CDRM. (d) CDAE.

leading to deviations in global data geometry. CDAE shows improved coverage of the real data manifold due to the diffusion-based decoder. However, the generated samples are fragmented into multiple sub-clusters, indicating that different causal factors are not consistently aligned across samples. By contrast, Causal DiffuseVAE produces generated samples that closely overlap with the real data distribution along the principal components, preserving both the spread and orientation of the true data manifold. This indicates that the model not only reconstructs visually realistic samples but also maintains the structural relationships induced by causal factors. For the target application of counterfactual image generation, this alignment is critical, as it implies that interventions modify samples within the valid data distribution rather than pushing them into unrealistic or causally inconsistent regions of the space.

Table 4.7. Comparison of the counterfactual images in evaluation using the MAE criterion

| Dataset | Causal DiffuseVAE | CausalVAE | CDRM | CDAE | CDM |
|---|---|---|---|---|---|
| Cube Shadow | **0.0098 ± 0.0050** | 0.0930 ± 0.0300 | 0.5400 ± 0.1300 | 0.2930 ± 0.0150 | 1.2100 ± 0.4590 |
| Polyhedron Shadow | **0.0103 ± 0.0040** | 0.1450 ± 0.0350 | 0.6600 ± 0.1320 | 0.4360 ± 0.0460 | 2.4500 ± 0.6770 |
| Pendulum | **0.0190 ± 0.0100** | 21.3020 ± 3.4400 | 17.9590 ± 2.5430 | 0.2980 ± 0.0100 | 0.8570 ± 0.3050 |
| Flow | **0.0230 ± 0.0100** | 16.6500 ± 3.3700 | 14.5900 ± 2.3220 | 0.3150 ± 0.0100 | 1.0100 ± 0.4680 |
| CelebA (Gender) | **0.0850 ± 0.0140** | 0.6500 ± 0.2400 | 0.7800 ± 0.4110 | 0.1340 ± 0.0369 | 0.9460 ± 0.3880 |

Table 4.8. LPIPS comparison of counterfactual image quality

| Dataset | Causal DiffuseVAE | CausalVAE | CDRM | CDAE | CDM |
|---|---|---|---|---|---|
| Cube Shadow | **0.0185 ± 0.0140** | 0.1059 ± 0.0980 | 0.0590 ± 0.0400 | 0.0482 ± 0.0210 | 0.2490 ± 0.1500 |
| Polyhedron Shadow | **0.0292 ± 0.0160** | 0.1982 ± 0.1200 | 0.0680 ± 0.0400 | 0.0295 ± 0.0200 | 0.3670 ± 0.1700 |

As the result shown in Table 4.7, Causal DiffuseVAE consistently outperforms other methods across all datasets, achieving the lowest error rates. This indicates that incorporating diffusion-based causal modeling enhances the model's ability to capture structural dependencies and improve reconstruction accuracy. In contrast, CausalVAE and CDRM exhibit significantly higher errors, particularly in complex datasets, highlighting their limitations in handling intricate variations. While CDAE and CDM show moderate performance, they still lag behind Causal DiffuseVAE, reinforcing the advantage of diffusion-based approaches in causal representation learning. These results demonstrate the potential of Causal DiffuseVAE in tasks requiring precise and robust generative modeling.

The results of the Causal DiffuseVAE and the baseline methods are listed in Table 4.8. Causal DiffuseVAE achieves the lowest perceptual similarity error in one dataset and remains highly competitive in the other. This indicates that the method preserves detailed structural information and produces perceptually similar reconstructions. CDAE also demonstrates impressive performance, particularly in one dataset, indicating its capability in certain scenarios. In contrast, CausalVAE, CDRM, and CDM exhibit higher perceptual errors, suggesting greater discrepancies between generated and original images. These findings reinforce the effectiveness of diffusion-based approaches in improving perceptual quality while maintaining causal consistency in image generation.

Table 4.9. LPIPS comparison of image quality under different training data ratios

| Dataset | Causal DiffuseVAE | CausalVAE | CDRM | CDAE | CDM |
|---|---|---|---|---|---|
| Full dataset | $\mathbf{0.0185 \pm 0.0108}$ | $0.1059 \pm 0.0980$ | $0.0590 \pm 0.0400$ | $0.0482 \pm 0.0217$ | $0.2490 \pm 0.1544$ |
| 50% dataset | $\mathbf{0.0193 \pm 0.0113}$ | – | $0.0800 \pm 0.0347$ | $0.0735 \pm 0.0483$ | $0.3140 \pm 0.1613$ |
| 30% dataset | $\mathbf{0.0235 \pm 0.0150}$ | – | $0.1200 \pm 0.1153$ | $0.0917 \pm 0.0657$ | $0.3470 \pm 0.1650$ |

Table 4.10. LPIPS scores of Causal DiffuseVAE, CDAE, and CDM at different sampling steps

| Sampling Steps | Causal DiffuseVAE | CDAE | CDM |
|---|---|---|---|
| 50 | $\mathbf{0.2156 \pm 0.0580}$ | $0.3018 \pm 0.0890$ | $0.3566 \pm 0.1930$ |
| 100 | $\mathbf{0.0440 \pm 0.0295}$ | $0.1486 \pm 0.0470$ | $0.2044 \pm 0.1544$ |
| 1000 | $\mathbf{0.0185 \pm 0.0108}$ | $0.0482 \pm 0.0217$ | $0.1410 \pm 0.1070$ |

To evaluate the efficiency of the Causal DiffuseVAE and the baseline methods, the models are trained on 50% and 30% of the dataset and use LPIPS to assess the quality of the generated images. Table 4.9 shows the LPIPS scores under different training-data ratios. Causal DiffuseVAE achieves the lowest LPIPS, which demonstrates robust and data-efficient performance. In contrast, CausalVAE fails to learn complete representations when trained on only 50% or 30% of the data. Moreover, other baselines exhibit much larger LPIPS increases as data decreases. These results demonstrate that Causal DiffuseVAE achieves data efficiency by maintaining low LPIPS scores even when trained on just 50% or 30% of the data, whereas competing LPIPS of methods degrade sharply under the same data constraints.

In conventional diffusion models, high-resolution image generation is costly. Full pixel-space denoising is applied over hundreds or thousands of steps, causing significant computational and memory demands. By contrast, diffusion is performed within a compact latent space in Causal DiffuseVAE, reducing per-step complexity and memory footprint. As shown in Table 4.10, equivalent LPIPS scores are achieved in 50 steps for CDM instead of 1,000, and in 100 steps for CDAE instead of 1,000, corresponding to $20\times$ and $10\times$ fewer iterations. These findings demonstrate that high-resolution images can be generated with fewer iterations and higher data efficiency.

Table 4.11. Training and inference time of different models on the Shadow Dataset

| Model | Training Time (h) | Inference Time (s/per image) |
|---|---|---|
| Causal DiffuseVAE | 17.3 | 0.68 |
| CausalVAE | 2.0 | 0.66 |
| CDRM | 2.5 | 0.75 |
| CDAE | 55.2 | 1.8 |
| CDM | 17.4 | 6.6 |

Table 4.11 presents training and inference time measured on two RTX A6000 GPUs. Training and inference are performed quickly for CausalVAE and CDRM, but high-resolution images are not generated. Among models capable of high-resolution generation, training time is shorter for Causal DiffuseVAE and CDM than for CDAE, but inference time is longer for CDM and causal control cannot be performed. Inference time for diffusion-based models is defined as the time required to generate images that achieve the same LPIPS score. Under these matched-quality conditions, the inference of using Causal DiffuseVAE is faster, demonstrating higher data efficiency in realizing equivalent image quality.

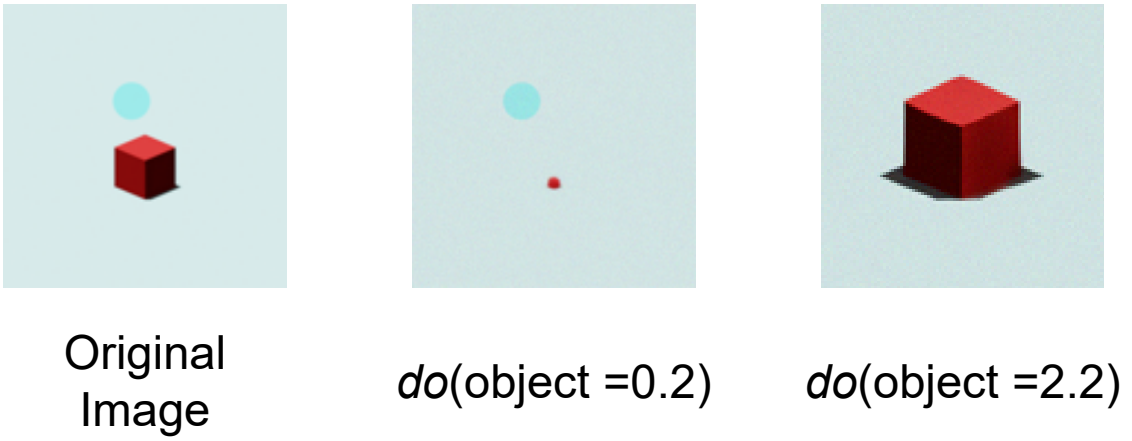



Figure 4.16. Expanding intervention results of shadow using Causal DiffuseVAE.

Furthermore, the object size in the training data for the shadow dataset is limited from 0.3 to 1.7. To evaluate more on the efficiency of the Causal DiffuseVAE, the images are generated using the data out of the range, which is shown in Figure. 4.16. When the object fully hides the light source, Causal DiffuseVAE still predicts the correct shadow, demonstrating its generation capability outside the training range.

Across datasets, Causal DiffuseVAE achieves the lowest MAE and highest LPIPS, outperforming CDAE, CausalVAE, CDRM, and CDM by precisely controlling causal factors while preserving visual detail. These findings underscore the promise of diffusion-based causal modeling for realistic and interpretable image generation. Hence, the generated counterfactual images can provide more reliable scenes for vision systems. The various scenes improve scene understanding in dynamic environments, such as avoiding the disruptions of the shadow.

Table 4.12. MAE Comparison on MNIST and Flow Datasets

| **Dataset** | **Intervention** | **Causal DiffuseVAE** | **Conventional CausalVAE** |
|---|---|---|---|
| MNIST | $do(\text{thickness} = 2)$ | 0.060 | 1.010 |
| | $do(\text{thickness} = 4)$ | 0.081 | 0.845 |
| Flow | $do(\text{ball size} = 15)$ | 0.023 | 16.655 |
| | $do(\text{water height} = 65)$ | 0.021 | 19.104 |
| | $do(\text{hole} = 3)$ | 0.091 | 8.790 |
| | $do(\text{water flow} = 11)$ | 0.011 | 9.724 |

To evaluate the accuracy of the model's control over latent factors, the MAE is used. TABLE 4.12 illustrates the comparison of the MAE between the Causal DiffuseVAE and Conventional CausalVAE. The blue part is the MAE of the Causal DiffuseVAE while the orange part is the MAE of the Conventional CausalVAE. As indicated in TABLE I, the MAE of the Causal DiffuseVAE is much lower than that of the Conventional CausalVAE, ascertaining the competence of the Causal DiffuseVAE against the Conventional CausalVAE for photo-realistic image generation.

Moreover, the models, like the Conventional Conditional DMs, are not able to change the effects by changing the causes. Only the performance of these models on the factor thickness is evaluated by MAE in Figure 4.17. while the Causal DiffuseVAE shows a preferable performance in controlling the latent factors.

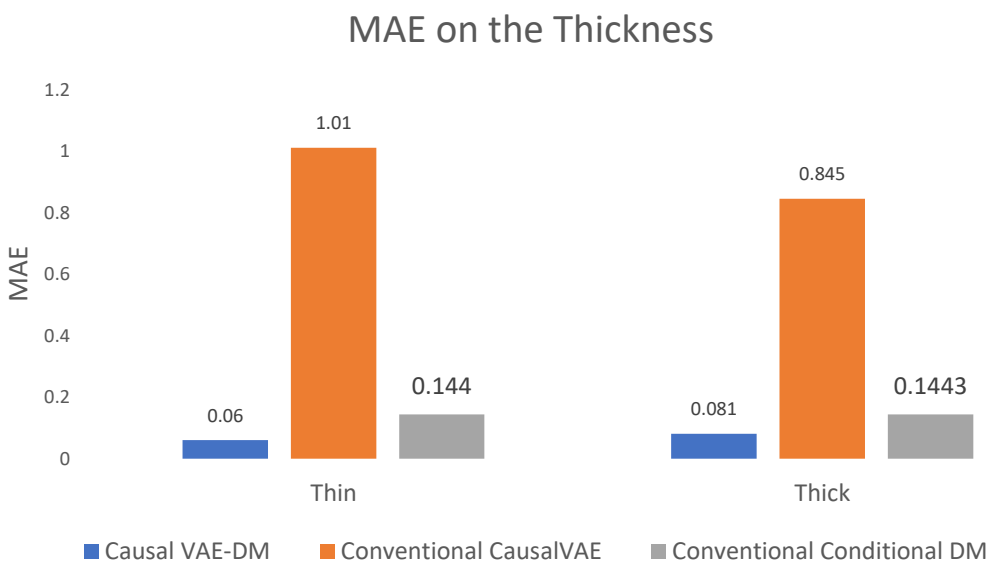



Figure 4.17. MAE on the Factor Thickness

## 4.4 Conclusion

In this chapter, Causal DiffuseVAE, a novel generative method, is proposed that integrate causal reasoning into the VAE and Diffusion Model framework. Its effectiveness has been demonstrated through quantitative evaluations on the Cube Shadow and Polyhedron Shadow datasets. Experimental results indicate that Causal DiffuseVAE consistently outperforms conventional methods in both reconstruction accuracy and perceptual similarity. The method effectively preserves structural details while ensuring causal consistency. Moreover, the causal-layer formulation in Section 4.1 provides a theoretical rationale for why interventions can be implemented in a structured latent space: under the DAG assumption and masked parent-dependence in (4.1), each latent factor is generated from its direct causes plus an independent disturbance, which encourages downstream effects to change consistently when a parent variable is intervened upon. This should be understood as an inductive bias rather than a strict guarantee; therefore, the experimental results and ablation studies in Section 4.3 are used to verify that the learned latent structure yields stable and semantically meaningful counterfactuals in practice. The capability of the Causal DiffuseVAE to generate

high-quality, causally controlled images shows its superiority in disentangling causal factors and reducing errors compared to CausalVAE, CDRM, CDAE, and CDM. The proven capability in generalization across different datasets suggests its potential for applications in shadow-aware vision systems and autonomous navigation, where accurate reconstruction and causal reasoning are essential. Future work could explore its deployment in real-world vision systems, enabling shadow removal while preserving object integrity, ultimately improving the robustness of perception-based decision-making in complex environments.

# Chapter 5

# Causal Diffusion Model Based on the Large Language Model

Chapter 4 showed that integrating a structural causal model with a diffusion-based generator enables precise and causally consistent latent interventions, but these interventions still require manually selecting and tuning causal variables. This limitation motivated Chapter 5, which investigates how a large language model can translate natural-language instructions into structured do-operations over causal factors while preserving the high-fidelity diffusion-based generation validated previously.

In this chapter, an end-to-end Causal Diffusion framework guided by a large language model is presented. The input image is first tokenized by a vision Transformer. Causal representations are then extracted from the LLM's output. These representations are integrated with the image tokens via a Query Transformer (Q-Former) cross-attention module. The fused features are processed by a masked causal layer that enforces the learned structural equations and enables explicit interventions, and an intermediate image is reconstructed by a lightweight decoder. Finally, this intermediate reconstruction is refined into a high-quality output by a conditional diffusion model trained with combined likelihood, causal regularization, and diffusion objectives. The architecture, training strategy, and evaluation are detailed throughout.

## 5.1 Model Architecture

In this section, the architecture of the Causal DiffuseLLM is introduced. The architecture is divided into two parts, which are Causal Transfomer with LLM and Conditional DDPM. The overall architecture is divided into the following components: Vision–Language Transformer, LLM Integration, Query Transformer, Cross-Attention Fusion, Convolutional Latent Projection & Causal Masking, and Diffusion Model.

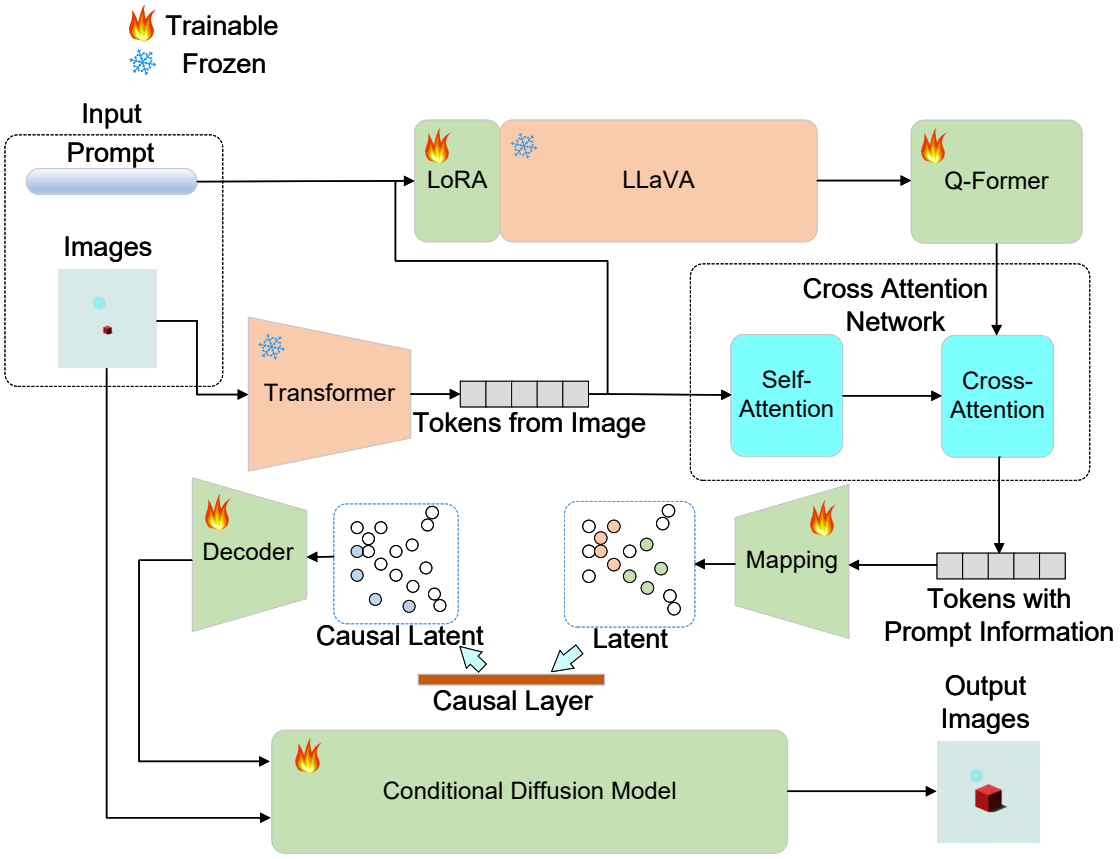



Figure 5.1. Overview of the Causal DiffuseLLM architecture.

Figure 5.1 illustrates the end-to-end Causal DiffuseLLM pipeline. An input image is patch-tokenized and, together with a text prompt, is fed to a frozen LLaVA backbone (lightly adapted via LoRA) to produce multimodal hidden states. A Q-Former distills prompt-aware query tokens, which are fused with visual tokens through a cross-attention module. The fused representation is projected into causal latents and passed

through a masked causal layer that enforces a learned structural graph and enables interventions. A lightweight decoder maps these latents to an intermediate reconstruction, which subsequently conditions a trainable diffusion model that iteratively denoises from noise to generate the final high-fidelity image.

### 5.1.1 Transformer with the Large Language Model

An input image $\mathbf{x}$ and its associated binary-attribute prompt sequence $c = (s_1, \ldots, s_T)$ are first embedded by separate encoders. A convolutional encoder plus linear projection is used to obtain

$$\mathscr{E}_v : \mathbf{x} \mapsto v_\mu(\mathbf{x}), \tag{5.1}$$

and a token embedding layer produces

$$\mathscr{E}_t : (s_1, \ldots, s_T) \mapsto \big(e(s_1), \ldots, e(s_T)\big). \tag{5.2}$$

where $e$ is the token embeddings.

The joint image–text representations are obtained by a sequence of transformer modules. First, the visual feature $v_\mu(\mathbf{x})$ and token embeddings $\{e(s_i)\}_{i=1}^{T}$ are merged by a frozen multimodal transformer:

$$H = \mathscr{M}\big(v_\mu(\mathbf{x}), e(s_1), \ldots, e(s_T)\big), \tag{5.3}$$

where $\mathscr{M}$ denotes the multi-layer transformer that processes both modalities in parallel. In Causal DiffuseLLM, the Vision–Language Transformer is instantiated from the Large Language and Vision Assistant (Llava) [173]. The image $\mathbf{x}$ is first preprocessed by the processor of the Llava, which divides $\mathbf{x}$ into a sequence of overlapping patches and projects them into patch embeddings of dimension $d_v$. Simultaneously, the binary-attribute prompt $c = (s_1, \ldots, s_T)$ is tokenized by the tokenizer in the Llava and mapped to embeddings of dimension $d_\ell$. These two streams are concatenated (with learned positional encodings) and passed into multimodal encoder–decoder, which con-

sists of 12 Vision Transformer layers in its vision tower and 24 causal Transformer decoder layers in its language branch, each with 16 attention heads. Cross-attention is performed in every decoder layer, allowing the language branch to attend to visual tokens and the vision tower to incorporate textual context. The fused hidden states are produced as

$$H = \mathscr{M}\big([v_\mu(\mathbf{x}) \,;\, e(s_1), \ldots, e(s_T)]\big) \in \mathbb{R}^{B\times N\times d}, \tag{5.4}$$

where $N$ is the total sequence length, $d$ is the embedding dimension, same as $d_\ell$ and $d_v$ and $B$ is the batch size used in the training and testing process. These representations $H$ encode aligned semantic and visual information and are forwarded to the subsequent Query Transformer.

### 5.1.2 Large Language Model Integration and Fine-tuning

The textual reasoning component is based on the Llava model. After loading the pre-trained weights, a LoRA configuration is applied, only the query, key and value projection matrices in each self-attention and cross-attention block are augmented with trainable adapters of rank $r = 8$, the scaling factor is defined as 32, and dropout is set as 0.05. All other parameters of the Llava model remain frozen, and gradient checkpointing is enabled to minimize memory consumption.

Let the IMG tokens occupy positions $t_1 = T+1, \ldots, t_r = T+r$ in the decoder output. Their hidden states are collected as

$$h \;=\; \big(\tilde{H}_{b,t_i,:}\big)_{\substack{b=1,\ldots,B\\ i=1,\ldots,r}} \;\in\; \mathbb{R}^{B\times r\times d} \tag{5.5}$$

where $\tilde{H} \in \mathbb{R}^{B\times(T+r)\times d}$ are the hidden states in the decoder, $t_i = T+i$ indicates the $i$-th appended IMG token, $B$ is the batch size used in the training and testing process, $T$ is the number of text tokens in the prompt, $r$ is the number of appended IMG tokens, and $d$ is the shared embedding dimension.

The full output token sequence produced by the chat template is denoted $y_{1:L}$, and a mask $m_t \in \{0,1\}$ is defined with $m_t = 1$ when $t \in \{t_1, \ldots, t_r\}$. With the base parameters $\theta$ frozen and the LoRA parameters $E_{lora}$ trainable, the following objective is minimized:

$$\mathscr{L}_{\text{LLM}}(E_{lora}) \;=\; \mathbb{E}_{(\mathbf{x},c)\sim\mathscr{D}}\left[-\frac{1}{\sum_{t=1}^{L} m_t}\sum_{t=1}^{L} m_t \log p_{\theta,E_{lora}}\big(y_t \mid y_{<t}, \mathbf{x}, c\big)\right], \tag{5.6}$$

where $t_i = T + i$ denote the IMG positions, $\sum_{t=1}^{L} m_t = r$ by construction, $\mathbf{x}$ is the input image, $c = (s_1, \ldots, s_T)$ is the tokenized prompt, $\mathscr{D}$ is the dataset of $(\mathbf{x}, c)$ pairs, $y_{<t}$ denotes the tokens strictly preceding position $t$, $p_{\theta,E}$ is the conditional token distribution with frozen base weights $\theta$ and trainable LoRA parameters $E$, and $L = T + r$ plus additional special tokens.

An equivalent form, obtained by summing only over the IMG positions, is given by

$$\mathscr{L}_{\text{LLM}}(E_{lora}) \;=\; \mathbb{E}_{(\mathbf{x},c)\sim\mathscr{D}}\left[-\frac{1}{r}\sum_{i=1}^{r} \log p_{\theta,E_{lora}}\big([\texttt{IMG}]_i \mid y_{<t_i}, \mathbf{x}, c\big)\right], \tag{5.7}$$

where $[\texttt{IMG}]_i$ denotes the $i$-th appended IMG token at position $t_i = T + i$, and all remaining symbols are as defined above.

### 5.1.3 Query Transformer

The Query Transformer $\mathscr{Q}$ is introduced to distill and compress the multimodal hidden states $H$ into a compact, fixed–size embedding suitable for downstream fusion. It is configured as a 6-layer Transformer decoder with dimension $d$, 16 self-attention heads, and an intermediate feed-forward dimension of $4d$. A set of $n_q = 32$ learnable query tokens $Q \in \mathbb{R}^{1\times n_q \times d}$ is prepended to each batch and expanded to $\mathbb{R}^{B\times n_q \times d}$.

During the forward pass, a small set of learnable query tokens $Q$ read from the entire hidden-state sequence $H$ using cross-attention. At the same time, the queries attend to self-attention to share information. Each attention or MLP block is wrapped with layer normalization and a residual connection. The module outputs a compact set of

features:

$$f = \mathcal{Q}(Q, H), \tag{5.8}$$

where $H \in \mathbb{R}^{B\times N\times d}$ is the input sequence of hidden states, $Q \in \mathbb{R}^{B\times n_q\times d}$ are the expanded query tokens, and $f \in \mathbb{R}^{B\times n_q\times d}$ are the distilled vision–language features. Gradient checkpointing is used inside $\mathcal{Q}$ to reduce memory, and no extra parameters are added to the frozen backbone beyond $\mathcal{Q}$ and the $n_q$ queries.

### 5.1.4 Cross-Attention Fusion

The Cross-Attention Fusion module $\mathcal{C}$ combines the distilled vision–language queries $f$ with the pure visual features $v$. The visual features are obtained by passing the image through the frozen vision tower:

$$v = \mathcal{V}(\mathbf{x}) \in \mathbb{R}^{B\times N_v\times d}, \tag{5.9}$$

where $\mathcal{V}(\cdot)$ is the network of the vision tower, $B$ is the batch size, $N_v$ is the number of visual tokens, and $d$ is the shared embedding dimension.

Inside $\mathcal{C}$, cross-attention is applied once, using $f$ as queries and $v$ as values:

$$f' = \mathrm{softmax}\Big(\frac{(fW^Q)(vW^K)^\top}{\sqrt{d_k}}\Big)(vW^V), \tag{5.10}$$

followed by a projection and a residual add:

$$\tilde{f} = f + f'W^O. \tag{5.11}$$

Layer normalization is applied around the attention and MLP blocks for stability. The fused embedding $\tilde{f} \in \mathbb{R}^{B\times n_q\times d}$ is then sent to the convolutional latent projection stage.

where $f \in \mathbb{R}^{B\times n_q\times d}$ are the query tokens, $d_k$ is the key/query head dimension, and $W^Q \in \mathbb{R}^{d\times d_k}$, $W^K \in \mathbb{R}^{d\times d_k}$, $W^V \in \mathbb{R}^{d\times d}$, $W^O \in \mathbb{R}^{d\times d}$ are learned projections. Intuitively, $f$ "looks up" relevant details in $v$ via cross-attention, and the result is merged back into $f$ with a residual connection.

### 5.1.5 Convolutional Latent Projection & Causal Masking

After cross-attention, the feature tensor $f' \in \mathbb{R}^{B\times n_q\times d}$ is reshaped into a spatial grid and passed through a small convolutional mapping network $\mathscr{G}$. The network outputs per-factor Gaussian parameters for $n$ causal subvectors:

$$\{(\mu^{(i)}, \sigma^{2\,(i)})\}_{i=1}^{n} = \mathscr{G}(f'), \qquad \mu^{(i)}, \sigma^{2\,(i)} \in \mathbb{R}^{d_c}. \tag{5.12}$$

An adjacency matrix $A \in \mathbb{R}^{n\times n}$ is learned. From $A$, a differentiable binary mask $M^{(i)}(A) \in \{0,1\}^{d_c}$ is derived for each subvector to enforce the causal relationship. Each latent subvector is then sampled with element-wise masking:

$$\mathbf{z}^{(i)} \sim \mathscr{N}\Big(\mu^{(i)} \odot M^{(i)}(A),\ \sigma^{2\,(i)} \odot M^{(i)}(A)\Big), \qquad i = 1,\ldots,n. \tag{5.13}$$

where $B$ is the batch size; $n_q$ is the number of query tokens; $d$ is the feature dimension; $n$ is the number of causal factors; $d_c$ is the dimensionality of each causal subvector; $\odot$ denotes element-wise multiplication; and $A$ provides the parent–child structure used to build the masks $M^{(i)}(A)$, so non-parent channels are deactivated for factor $i$. Intuitively, this implements the masked parent influence used in the causal layer of Chapter 4: only dimensions selected by $M^{(i)}(A)$ are allowed to affect $\mathbf{z}^{(i)}$.

### 5.1.6 Diffusion Model in the Causal DiffuseLLM

The intermediate image from the causal decoder is used to condition a denoising diffusion model. A UNet denoiser receives a noisy RGBA image and the causal latent code. Timestep information is encoded with sinusoidal embeddings and injected via Sigmoid Linear Unit (SiLU)-activated linear layers. The UNet consists of downsampling residual blocks, an attention bottleneck, and matching upsampling blocks, all with group normalization and skip connections. Training minimizes a noise-prediction loss; at inference, noise is removed step by step to produce a high-fidelity image consistent with the learned causal structure.

The forward process is

$$\mathbf{x}_t \;=\; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 \;+\; \sqrt{1-\bar{\alpha}_t}\,\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,I), \tag{5.14}$$

where $\mathbf{x}_0$ is the clean image, $\mathbf{x}_t$ is the noisy image at step $t$, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative noise schedule, and $\varepsilon$ is standard Gaussian noise.

The denoiser $\varepsilon_\theta$ is trained to predict the injected noise:

$$\mathscr{L}_{\text{DDPM}}(\theta) \;=\; \mathbb{E}_{\mathbf{x}_0,\mathbf{z},t,\varepsilon}\big\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t,t,\mathbf{z})\big\|^2, \tag{5.15}$$

where $\mathbf{z}$ is the causal latent code used to condition the UNet, t is a random timestep ranged from 0 to $T$, $\mathbf{x}_t$ is formed from $\mathbf{x}_0$ and $\varepsilon$ as above, and the expectation is taken over the data distribution and the sampling of $t$ and $\varepsilon$.

An auxiliary reconstruction term may be added:

$$\mathscr{L}(\theta) \;=\; \mathscr{L}_{\text{DDPM}}(\theta) \;+\; \lambda_{\text{recon}}\,\mathbb{E}\big[\|\mathbf{x}_{\text{recon}} - \mathbf{x}_0\|_2^2\big], \tag{5.16}$$

where $\mathbf{x}_{\text{recon}}$ is the decoder's intermediate image and $\lambda_{\text{recon}} > 0$ balances fidelity to the causal reconstruction.

## 5.2 Experiment and Discussion

The experiments are deployed on a server with an Ubuntu 20.04 operating system and two NVIDIA RTX A6000 graphics cards. For the first training step, which trains the model of the CausalLLM, two NVIDIA RTX A6000 graphics cards are necessary. For the second training step, which trains the model of the Diffusion Model, the cache file could be generated first to reduce memory consumption and accelerate data loading, allowing the training to be performed efficiently on a single NVIDIA RTX A6000 graphics card.

### 5.2.1 Experimental Setting

The shadow dataset in Chapter 4 is used in this experiment. The dataset provides a controlled environment where variations in light source size and object properties serve as causal factors, enabling the model to learn the underlying mechanisms of shadow formation rather than relying solely on correlations. In this dataset, 7,500 images are used for training, 1,000 images are used for validation, and 1500 images are used for testing. In the second experiment, the Smile dataset in Chapter 4 is also used. The Smile dataset consists of 20,000 images, with 70% allocated for training, 15% for validation, and the remaining 15% reserved for testing. It focuses on the attributes of Gender, Smile, Eyes Open, and Mouth Open, allowing for the exploration of causal relationships and interactions between these facial expressions.

Based on the findings in Chapter 4, where controlled synthetic and real-world datasets (e.g., shadow and Smile datasets) proved effective for validating causal consistency and intervention accuracy, Chapter 5 deliberately reuses these datasets to isolate the added value of language-driven control rather than changing the data domain. The model design in Chapter 5 is therefore guided by the need to preserve the same causal factors and evaluation settings while extending the intervention mechanism to an end-to-end, instruction-driven framework enabled by a large language model.

Furthermore, building on Chapter 4, where the shadow and Smile datasets were shown to provide clear causal structures and reliable evaluation of intervention accuracy, Chapter 5 retains these datasets to ensure that any performance gains arise from the language-guided causal design rather than changes in data distribution. In addition, while Chapter 4 relied on a fully trainable DDPM to demonstrate the effectiveness of causal conditioning, Chapter 5 explores replacing it with a pretrained Stable Diffusion backbone to reduce training cost and time, and to test whether the proposed causal–LLM front end can generalize to powerful off-the-shelf diffusion models without sacrificing causal controllability or image quality.

In the evaluation, two LLM-based methods, MagicBrush [174] and InstructPix2Pix [140], and a diffusion-based causal model, Causal Diffusion Autoencoder (CDAE) [118], are used as the baseline methods. The details of the CDAE could be found in Section 4.3.2.

MagicBrush introduces an instruction-guided image editing framework that leverages a tool-augmented multimodal large language model to decompose natural-language instructions into a sequence of localized editing operations. By explicitly reasoning about what to edit, where to edit, and how to perform the modification, MagicBrush enables complex, multi-step image edits driven by free-form textual input. The method benefits from a manually annotated dataset that provides strong supervision for instruction decomposition and region localization. However, MagicBrush does not impose any explicit causal structure over the underlying visual factors. Editing operations are executed sequentially in image space rather than through interventions on disentangled latent variables, which prevents the model from reasoning about causal dependencies among attributes. As a result, changes to one attribute may inadvertently affect others, and the framework cannot support counterfactual queries or guarantee consistency under physically grounded interventions.

InstructPix2Pix formulates instruction-following image editing as a single-stage conditional diffusion problem, learning a direct mapping from an input image and a textual instruction to an edited output image. The model is trained on large-scale synthetic triplets generated by pairing language model–produced editing instructions with diffusion-based image synthesis, enabling efficient and flexible instruction-conditioned editing without explicit region supervision or intermediate reasoning steps. While this approach achieves strong performance on general semantic and stylistic edits, it operates purely through correlational learning in pixel and feature space. InstructPix2Pix lacks an explicit latent representation in which individual generative factors are disentangled or causally related, and thus cannot perform controlled interventions or preserve invariant attributes under modification. Consequently, the model often entangles multiple visual properties when responding to an instruction, limiting its suitability for counterfactual image generation and causally consistent editing.

To achieve better evaluations, MagicBrush and InstructPix2Pix are finetuned using the shadow dataset. Furthermore, CDAE performs counterfactual editing by intervening on disentangled causal latents learned by a diffusion autoencoder.

Meanwhile, counterfactual images are generated and compared with Blender-rendered references using LPIPS, which assesses perceptual similarity via deep feature embeddings from pretrained networks rather than raw pixels. The metric has been reported to correlate more strongly with human judgments. Moreover, MAE is defined as the mean of the absolute differences between the predictions and the ground truth, computed over all elements. Lower MAE indicates higher fidelity. The details of the LPIPS and MAE could be found in Section 3.7.1.

### 5.2.2 Experimental Results

In this application, the large language model serves as a high-level reasoning and planning module that translates free-form, human-readable instructions into structured and consistent causal interventions. By interpreting semantic constraints, numeric targets, and relational queries, the LLM enables users to specify complex counterfactual goals without manually manipulating latent variables. The resulting images are therefore not merely visually realistic, but causally grounded: each generated counterfactual reflects a valid "what-if" scenario consistent with the underlying scene mechanics. Such images are particularly useful for objectives such as robustness evaluation and data augmentation in vision systems, where controlled variations of lighting, object properties, or shadows are required to probe model behavior under distribution shifts while preserving physical plausibility.

Causal DiffuseLLM provides an end-to-end intervention mechanism that maps natural-language instructions to precise do-operations over interpretable factors. Figure 5.2 represents the intervention results using different prompts. When "Moving the light source to position 7 (right side), how does this affect the shadow direction?" is sent to the LLM, the light sources in the original images are edited to position on the right side, which indicates location 7. Same to the intervention of the light position, when "The object size directly affects shadow formation. Setting size to 0.6 should produce specific shadow characteristics." is sent to LLM, Causal DiffuseLLM changes the object size smaller, which indicates size 0.6. At the same time, the light positions keep unchanged. Unlike the Causal DiffuseVAE in Chapter 4, which follows the causal graph strictly, Causal DiffuseLLM provides the capability to the model to consider what would happen if the effects in the causal graph are changed. When "If the shadow area becomes 5.0, what lighting conditions would cause this?" is sent to the LLM, the object sizes in the original images remain unchanged, and the light positions are tuned to get the shadow whose area is 5.0. The intervention produced an ordered set of do-operations that preserved previously set values and avoided conflicts. Guard conditions were applied to keep non-target attributes invariant unless explicitly specified.

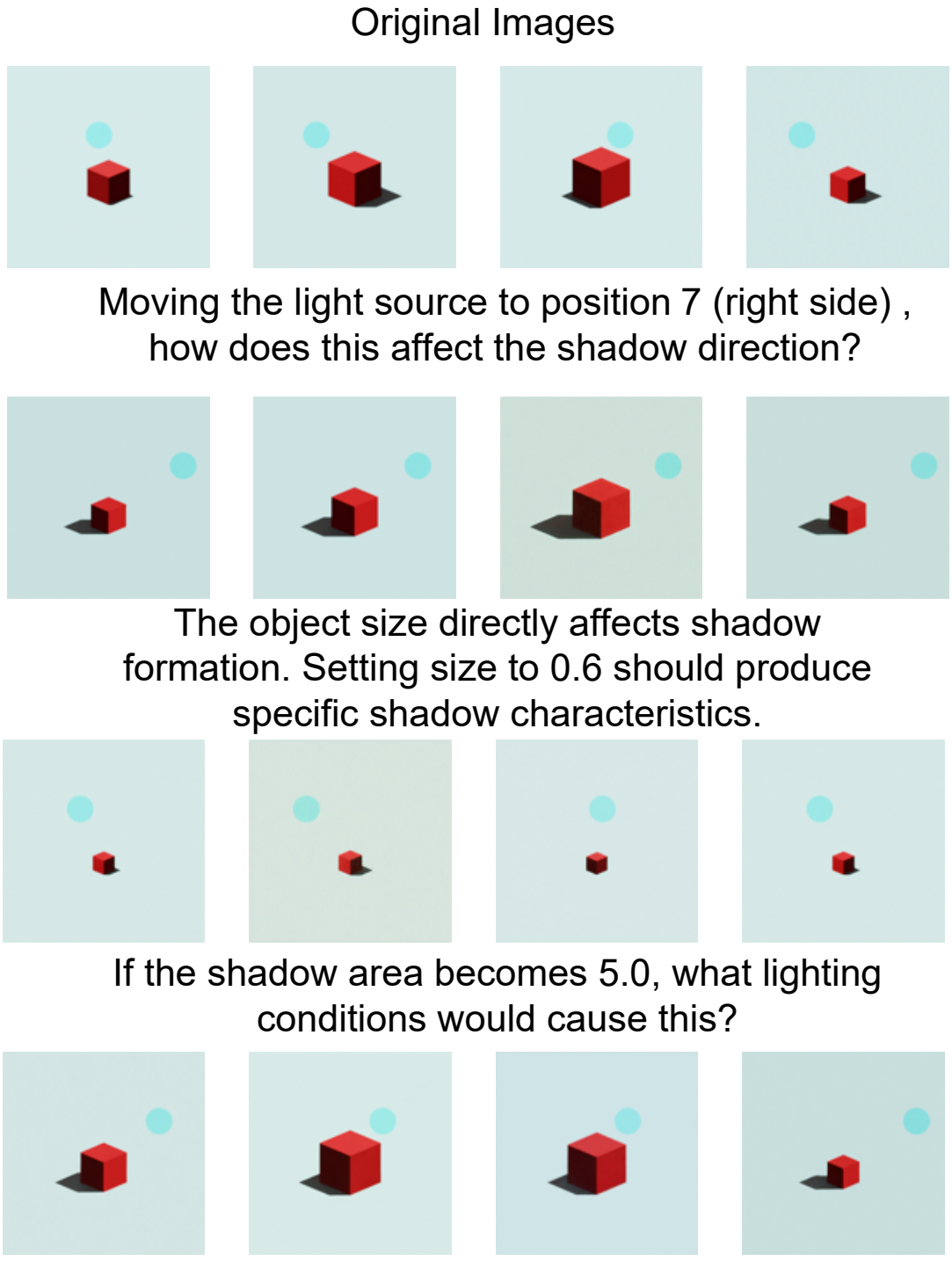



Figure 5.2. Intervention results when using different prompts.

Figure 5.3 shows the results when doing the intervention on the object size using Causal DiffuseLLM, CDAE, InstructPix2Pix and MagicBrush. CDAE almost achieves all the details of the target image, while the object size is not edited well. InstructPix2Pix and MagicBrush generate the image based on the prompt ”The object size directly affects

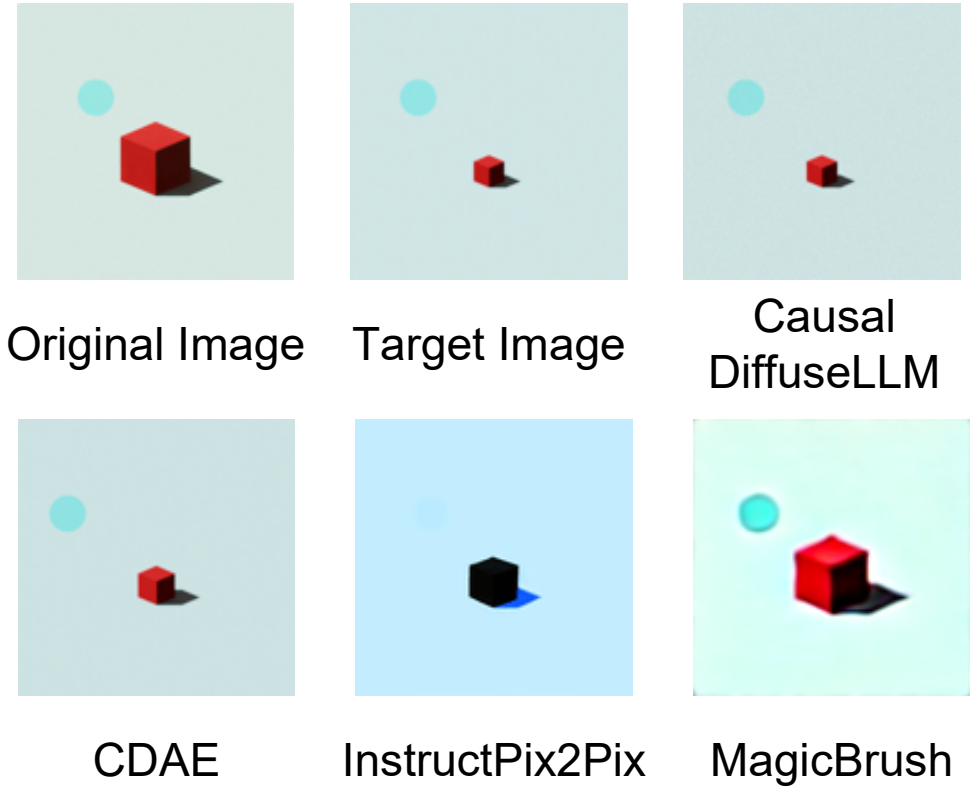



Figure 5.3. Intervention results when changing the object size to 0.6 using Causal DiffuseLLM and baseline methods.

Table 5.1. Comparison of the counterfactual images using the MAE

| Intervention | Causal DiffuseLLM | CDAE | InstructPix2Pix | MagicBrush |
|---|---|---|---|---|
| *do*(light position=7) | **0.0133** ± **0.0060** | 0.1877 ± 0.040 | 0.7720 ± 0.1800 | 0.2312 ± 0.090 |
| *do*(object size=0.6) | **0.0107** ± **0.0050** | 0.1653 ± 0.035 | 0.6970 ± 0.1620 | 0.2674 ± 0.080 |
| *do*(shadow area=5.0) | **0.0142** ± **0.0090** | 0.2043 ± 0.047 | 0.8451 ± 0.1770 | 0.2963 ± 0.070 |

shadow formation. Setting size to 0.6 should produce specific shadow characteristics.". However, although finetuning is deployed on these two models, they failed to understand what is the object size. InstructPix2Pix reduces the object size but fails to keep other features unchanged. MagicBrush fails to change the object size and also fails to reconstruct the other features.

Table 5.2. Comparison of the counterfactual images using the LPIPS

| Intervention | Causal DiffuseLLM | CDAE | InstructPix2Pix | MagicBrush |
|---|---|---|---|---|
| *do*(light position=7) | $\mathbf{0.0051 \pm 0.0008}$ | $0.0413 \pm 0.020$ | $0.2119 \pm 0.200$ | $0.2708 \pm 0.330$ |
| *do*(object size=0.6) | $\mathbf{0.0067 \pm 0.007}$ | $0.0562 \pm 0.037$ | $0.2735 \pm 0.240$ | $0.2587 \pm 0.290$ |
| *do*(shadow area=5.0) | $\mathbf{0.0103 \pm 0.0010}$ | $0.0861 \pm 0.040$ | $0.2514 \pm 0.215$ | $0.2465 \pm 0.266$ |

Table 5.1 shows the results evaluated by MAE. The lowest reconstruction error is consistently achieved by Causal DiffuseLLM across all three intervention types, with the smallest dispersion, indicating both higher fidelity and greater stability. CDAE ranks second, while the two instruction-driven baselines underperform, with MagicBrush generally outperforming InstructPix2Pix. The MAE results prove that Causal DiffuseLLM outperforms other baseline methods in controlling the factors.

Based on Table 5.2, Causal DiffuseLLM achieves the lowest perceptual distance across all intervention types and exhibits the smallest dispersion, indicating stable, faithful edits. CDAE forms a consistent second tier, whereas the instruction-driven baselines, InstructPix2Pix and MagicBrush, get higher LPIPS and substantially larger variability, with no consistent winner between them. These results highlight the capability of the Causal DiffuseLLM in generating precise counterfactual images with high quality.

Furthermore, a cost-efficient configuration was realized by adopting Stable Diffusion [175] and validating it on the Smile Dataset. Stable Diffusion is a latent diffusion text-to-image model that performs denoising in a compressed latent space via a VAE, enabling high-resolution synthesis with modest compute. Language prompts are injected through a Contrastive Language–Image Pretraining (CLIP) text encoder and cross-attention in a U-Net denoiser, supporting controllable generation and efficient fine-tuning for tasks such as image editing and inpainting. In the Smile Dataset, the smile is considered causing narrow eyes and a mouth opening while the gender is considered causing the eyes changed. Figure. 5.4 shows the results of the Causal DiffuseLLM using Stable Diffusion and the results obtained using InstructPix2Pix with

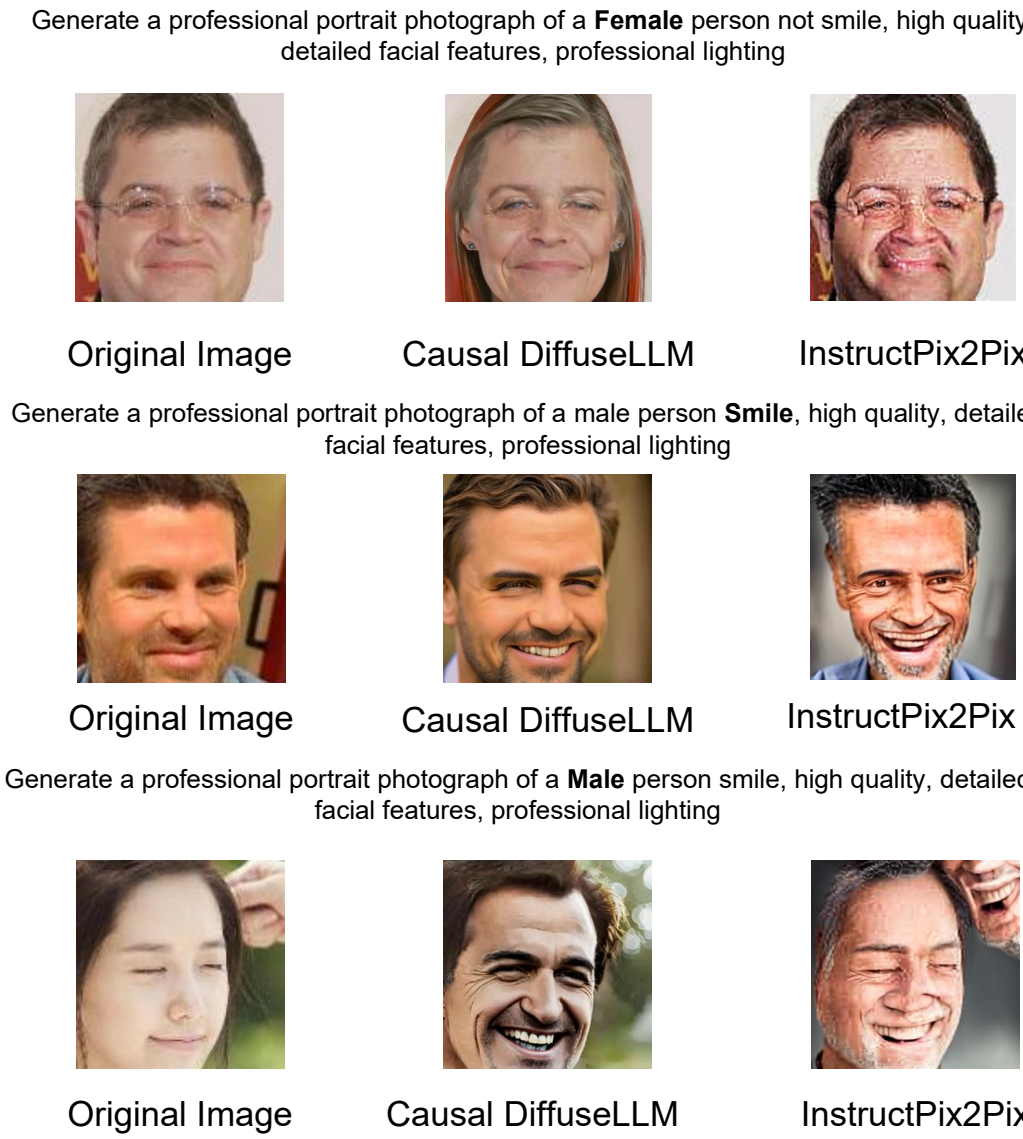



Figure 5.4. Intervention outcomes on Smile: Causal DiffuseLLM (pre-trained diffusion backbone) vs. InstructPix2Pix

the same prompts. This figure indicates that the framework of the Causal DiffuseLLM is compatible with the pretrained diffusion model to generate counterfactual results. Compared with the InstructPix2Pix, Causal DiffuseLLM is able to edit the image following the given prompt on the causal features and obtain better results.

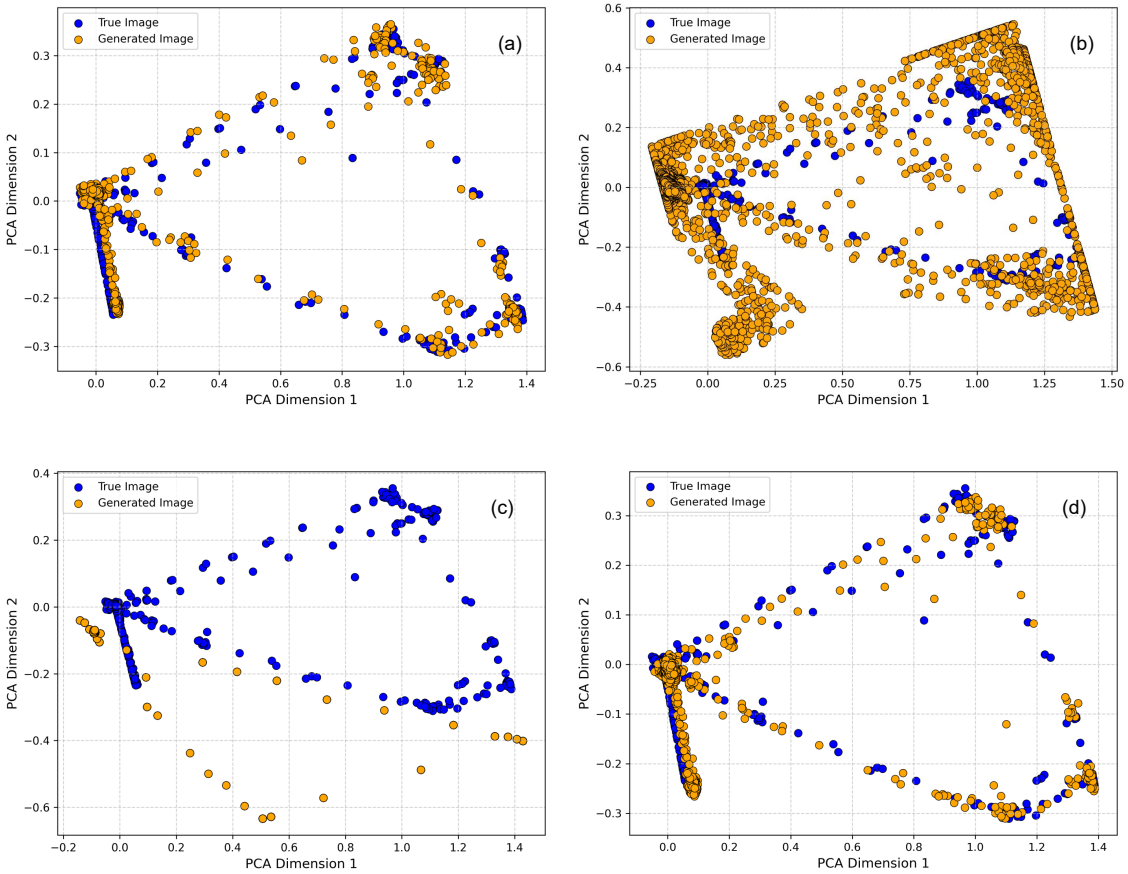



Figure 5.5. Scatter plot of generated data and true data after PCA-based processing. (a) Causal DiffuseLLM. (b) MagicBrush. (c) InstructPix2Pix. (d) CDAE.

In Figure 5.5, PCA projects high-dimensional latent representations of generated and true samples into a two-dimensional subspace spanned by the first two principal components. Each point denotes one sample in this reduced latent space. Overlapping distributions indicate that the generated data manifold aligns closely with the real data distribution. Among all compared models, Causal DiffuseLLM shows the most consistent overlap with the ground truth, demonstrating superior semantic and causal alignment in the latent space.

Overall, the results show that Causal DiffuseLLM enables precise, causally grounded interventions that translate free-form instructions into reliable counterfactual edits. By mapping language directly to do-operations over interpretable factors, edits are localized to causally affected regions and non-target attributes are preserved. This approach achieves consistent improvements over diffusion- and LLM-based baselines, reducing error, perceptual distance, and variance. These results establish Causal DiffuseLLM

as an effective end-to-end framework for controllable counterfactual generation in the shadow domain. They also lay a strong foundation for extending causal, language-driven control to more complex scenes and factors. Moreover, to reduce training time and cost, the conditional DDPM can be replaced with a general pretrained diffusion model, such as Stable Diffusion. Unlike non-causal LLM-guided generators, Causal DiffuseLLM leverages its causal structure to produce precise, prompt-based counterfactual images.

## 5.3 Conclusion

In this chapter, Causal DiffuseLLM, an end-to-end, language-guided causal diffusion framework, was presented. By mapping free-form instructions to explicit do-operations over interpretable factors and coupling a causal Transformer front end with a conditional diffusion back end, the model delivers precise, localized counterfactual edits. On the shadow dataset, it consistently outperforms CDAE, InstructPix2Pix, and MagicBrush, achieving lower MAE and LPIPS with smaller variance while preserving non-target attributes and satisfying numerical constraints on light position, object size, and shadow area. The language-to-intervention interface also produces transparent, reproducible edit logs and robust behavior under paraphrase, underscoring its practicality for instruction-driven causal control. The learned causal layer and intervention planner promote factor disentanglement and stable counterfactuals, enabling faithful what-if analysis within the domain. Future work will move beyond discretized lighting toward richer scene physics and continuous controls, and explore adaptation to real-world tasks with online refinement of the causal graph.

# Chapter 6

# Conclusion

This chapter provides a summary of the key contributions of the present work, directly addressing the research objectives outlined in Section 1.4. It also discusses the limitations of the study and proposes potential directions for future research to address these limitations.

## 6.1 Research Contribution

This thesis advances causal generative modeling by introducing two novel frameworks, Causal DiffuseVAE and Causal DiffuseLLM, that integrate explicit causal reasoning with high-fidelity diffusion-based synthesis. Causal DiffuseVAE combines a variational autoencoder with a causal latent layer and a conditional diffusion model, enabling interpretable, controllable, and photorealistic counterfactual image generation.

The Causal DiffuseVAE framework integrates explicit structural causal modeling with a latent-conditioned diffusion process, enabling high-fidelity image synthesis alongside interpretable and controllable counterfactual generation. A causal layer is embedded within the VAE latent space to enforce directional dependencies consistent with a predefined DAG while supporting flexible "do-calculus" interventions. By combining the

efficient inference of a VAE with the generative fidelity of diffusion models, Causal DiffuseVAE achieves state-of-the-art performance in reconstruction accuracy, perceptual similarity, and latent controllability across diverse synthetic and real-world datasets, demonstrating its effectiveness for both visual quality and causal interpretability.

On the other hand, the Causal DiffuseLLM is the framework to leverage an LLM for extracting semantically rich causal representations from image–text pairs and integrating them into a causal diffusion pipeline. It employs a Q-Former cross-attention fusion mechanism to combine LLM-derived semantic latents with vision encoder features, enabling precise control over causal factors conditioned on both visual content and textual prompts. A causal masking mechanism in the latent projection stage further enforces structural dependencies defined by a causal graph while supporting text-driven interventions, allowing for flexible and interpretable multimodal image editing.

The works in this thesis present a unified architectural framework for embedding causal graphs into latent representations while preserving high image fidelity through diffusion-based refinement. The approach generalizes effectively across diverse domains, including shadows, physical systems, human faces, and industrial circuits, as well as different types of causal variables, such as continuous, discrete, and multimodal factors. By combining explicit causal structure with diffusion refinement, the framework achieves superior trade-offs between interpretability, controllability, and visual quality compared to existing baselines, including CausalVAE, CDRM, CDAE, and CDM.

Furthermore, the proposed frameworks are supported by a theoretical identifiability analysis showing that, under mild assumptions, they can recover disentangled causal latents consistent with the true causal graph. Extensive quantitative evaluations using MAE, LPIPS, and efficiency metrics, along with qualitative case studies, confirm their superior performance in generalization and robustness. Notably, the models maintain causal consistency and visual fidelity even in out-of-distribution intervention scenarios, highlighting their reliability for both controlled experimentation and real-world applications.

Moreover, the proposed methods demonstrate strong potential for applications such as shadow-aware vision systems and autonomous navigation, where integrating causal reasoning enhances the robustness of perception and decision-making. Furthermore, by introducing LLM-augmented causal generation, this work opens a new pathway for interactive, instruction-driven scene manipulation with explicit causal guarantees, expanding the scope of controllable and interpretable generative modeling in real-world and safety-critical domains. Causal DiffuseLLM extends causal interventions from purely visual domains to multimodal scenarios, where interventions can be triggered by natural language descriptions as well as direct manipulation of latent variables. The expanding results of Causal DiffuseVAE demonstrate that the proposed methods maintain causal interpretability and synthesis quality when performing interventions outside the training distribution, highlighting their robustness in realistic, unconstrained scenarios.

Last but not least, learning dynamic or time-varying causal graphs is a natural extension of the proposed frameworks when considering video data. In videos, causal relationships between factors such as object motion, lighting, occlusion, and interaction are not static but evolve over time. While this thesis focuses on fixed DAGs for single-image counterfactual generation, the latent causal formulation and masked structural updates introduced in Chapters 4 and 5 provide a foundation for modeling temporal causal mechanisms. By extending the causal latent variables to sequences and allowing the adjacency matrix to change across time steps, future work could capture evolving cause–and–effect relations in dynamic scenes, such as moving shadows, object interactions, or action-driven state transitions. This would enable causally consistent video generation and counterfactual reasoning, where interventions not only affect individual frames but propagate coherently across time, aligning with real-world physical and semantic dynamics.

## 6.2 Limitation

Despite the promising results, the proposed frameworks have several limitations that point to opportunities for improvement. Causal relationships are obtained from physical laws, so the causal graphs are assumed to be fixed in the Causal DiffuseVAE. While the causal graph remains fixed, the causal matrix is designed to be trainable, allowing the model to adaptively refine its learned causal relationships throughout the training process. Both Causal DiffuseVAE and Causal DiffuseLLM assume a fixed, acyclic causal graph, which limits their applicability to domains with evolving structures, feedback loops or unknown causal relations.

In Causal DiffuseLLM, the LLM backbone is largely frozen, which constrains its adaptability to highly specialized or low-resource domains where pretrained multimodal knowledge may be insufficient. Meanwhile, although Causal DiffuseVAE is more efficient than full diffusion pipelines, Causal DiffuseLLM still requires substantial computational resources for training and inference, especially when scaling to high-resolution or multi-object scenes. While LLM provides strong capability in semantic reasoning and multimodal inference, its large model size and frozen backbone limit adaptability to highly specialized domains, and the reliance on extensive pretrained parameters increases memory usage and slows down fine-tuning in resource-constrained environments.

Furthermore, performance is sensitive to the quality and diversity of training data. Both models can experience degradation under severe domain shift, especially in cases where causal factors present in the test domain are absent in training. The Causal Circuit benchmark uses synthetic data but ignores real-world factors such as sensor noise, calibration drift and lighting changes. These factors can reduce image quality and causal accuracy in practice. To address this, the encoder–decoder can be fine-tuned on a

small set of real sensor data. Additionally, synthetic training samples can be augmented with realistic noise, and noise-robust encoder architectures can be employed. Hardware-aware optimizations such as weight quantization and pruning can further help meet latency and memory constraints.

## 6.3 Future Work

Based on the findings of this thesis, several directions can be explored to overcome the identified limitations and extend the applicability of the proposed frameworks.

First, future work could investigate dynamic or learnable causal graphs that adapt to evolving relationships in the data. This extension is particularly relevant for video and sequential settings, where causal relationships between factors such as motion, lighting, and interaction change over time. Incorporating causal discovery methods or Bayesian structure learning into the training process would allow the model to handle partially unknown, time-varying, or even cyclic causal structures.

Second, the adaptability of the large language model in Causal DiffuseLLM could be further improved. While the current framework relies on a largely frozen LLM backbone, future research may explore joint or selective fine-tuning using parameter-efficient methods such as LoRA, adapters, or prompt tuning. In addition, domain-specific multimodal pretraining could enhance performance in low-resource or highly specialized application domains.

Third, reducing the computational cost of Causal DiffuseLLM remains an important direction for practical deployment. This could be achieved through model compression techniques, including knowledge distillation, weight quantization, and pruning, as well as more efficient diffusion sampling strategies to lower inference latency without sacrificing image quality.

Fourth, enhancing robustness to domain shift is essential for real-world applications. Future work could incorporate mixed real–synthetic training, domain adaptation techniques, and data augmentation strategies that simulate realistic environmental factors such as sensor noise, lighting variation, and calibration drift. Hardware-aware optimizations should also be explored to enable reliable deployment on embedded and resource-constrained platforms.

Finally, beyond discrete and predefined interventions, future research could extend the intervention space to support continuous, compositional, and higher-order manipulations. Such extensions would enable richer causal reasoning and more flexible control over generated content in complex, multi-object, and multimodal scenes.

# References


[1] S. Lin, G. Zheng, Z. Wang *et al.*, 'Embodied neuromorphic synergy for lighting-robust machine vision to see in extreme bright', *Nature Communications*, vol. 15, no. 1, p. 10 781, 2024. DOI: 10.1038/s41467-024-54789-8. [Online]. Available: https://doi.org/10.1038/s41467-024-54789-8.

[2] S. Neupane *et al.*, *Security considerations in ai-robotics: A survey of current methods, challenges, and opportunities*, 2024. arXiv: 2310.08565 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2310.08565.

[3] Y.-T. Chen, C.-Y. Hsu, C.-M. Yu, M. Barhamgi and C. Perera, 'On the private data synthesis through deep generative models for data scarcity of industrial internet of things', *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 551–560, 2023. DOI: 10.1109/TII.2021.3133625.

[4] L. Jiao *et al.*, 'Causal inference meets deep learning: A comprehensive survey', *Research*, vol. 7, p. 0467, 2024. DOI: 10.34133/research.0467. eprint: https://spj.science.org/doi/pdf/10.34133/research.0467. [Online]. Available: https://spj.science.org/doi/abs/10.34133/research.0467.

[5] C. Chen, N. A. Mat Isa and X. Liu, 'A review of convolutional neural network based methods for medical image classification', *Computers in Biology and Medicine*, vol. 185, p. 109 507, 2025, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2024.109507. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482524015920.

[6] M. Gholizade, H. Soltanizadeh, M. Rahmanimanesh and S. S. Sana, 'A review of recent advances and strategies in transfer learning', *International Journal of System Assurance Engineering and Management*, vol. 16, no. 3, pp. 1123–1162, Mar. 2025. DOI: 10.1007/s13198-024-02684-2.

[7] H. A. Mehrtens, A. Kurz, T.-C. Bucher and T. J. Brinker, 'Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise', *Medical Image Analysis*, vol. 89, p. 102 914, 2023, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2023.102914. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001743.

[8] J. Wang *et al.*, *Large language models for robotics: Opportunities, challenges, and perspectives*, 2024. arXiv: 2401.04334 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2401.04334.

[9] Z. Hau, S. Demetriou and E. C. Lupu, *Using 3d shadows to detect object hiding attacks on autonomous vehicle perception*, 2022. arXiv: 2204.13973 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.13973.

[10] E. Şahin, N. N. Arslan and D. Özdemir, 'Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning', *Neural Computing and Applications*, vol. 37, pp. 859–965, 2025. DOI: 10.1007/s00521-024-10437-2.

[11] D. Medina-Ortiz, A. Khalifeh, H. Anvari-Kazemabad and M. D. Davari, 'Interpretable and explainable predictive machine learning models for data-driven protein engineering', *Biotechnology Advances*, vol. 79, p. 108 495, 2025, ISSN: 0734-9750. DOI: https://doi.org/10.1016/j.biotechadv.2024.108495. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0734975024001897.

[12] Y. Wang *et al.*, 'Segt-go: A graph transformer method based on ppi serialization and explanatory artificial intelligence for protein function prediction', *BMC Bioinformatics*, vol. 26, no. 1, p. 46, 2025. DOI: 10.1186/s12859-025-06059-7.

[13] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang and W. Zhang, 'History, development, and principles of large language models—an introductory survey', *arXiv preprint arXiv:2402.06853*, 2024. [Online]. Available: https://arxiv.org/abs/2402.06853.

[14] Y. Annepaka and P. Pakray, 'Large language models: A survey of their development, capabilities, and applications', *Knowledge and Information Systems*, vol. 67, pp. 2967–3022, 2025. DOI: 10.1007/s10115-024-02310-4. [Online]. Available: https://link.springer.com/article/10.1007/s10115-024-02310-4.

[15] D. Anisuzzaman, J. G. Malins, P. A. Friedman and Z. I. Attia, 'Fine-tuning large language models for specialized use cases', *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 1, p. 100 184, 2025, ISSN: 2949-7612. DOI: https://doi.org/10.1016/j.mcpdig.2024.11.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949761224001147.

[16] S. R. Gantla, 'Exploring mechanistic interpretability in large language models: Challenges, approaches, and insights', in *2025 International Conference on Data Science, Agents Artificial Intelligence (ICDSAAI)*, 2025, pp. 1–8. DOI: 10.1109/ICDSAAI65575.2025.11011640.

[17] D. Liang and F. Xue, 'Integrating automated machine learning and interpretability analysis in architecture, engineering and construction industry: A case of identifying failure modes of reinforced concrete shear walls', *Computers in Industry*, vol. 147, p. 103 883, 2023, ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2023.103883. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166361523000337.

[18] C. Zhang, P.-J. Hoes, S. Wang and Y. Zhao, 'Intrinsically interpretable machine learning-based building energy load prediction method with high accuracy and strong interpretability', *Energy and Built Environment*, 2024, ISSN: 2666-1233. DOI: https://doi.org/10.1016/j.enbenv.2024.08.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666123324000825.

[19] D. Bhati, M. Amiruzzaman, Y. Zhao, A. Guercio and T. Le, 'A survey of post-hoc xai methods from a visualization perspective: Challenges and opportunities', *IEEE Access*, vol. 13, pp. 120 785–120 806, 2025. DOI: 10.1109/ACCESS.2025.3581136.

[20] A. Komanduri, X. Wu, Y. Wu and F. Chen, 'From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling', *Transactions on Machine Learning Research*, 2024, ISSN: 2835-8856. [Online]. Available: https://openreview.net/forum?id=PUpZXvNqmb.

[21] L. Yang, J. Cheng, Y. Luo, T. Zhou and X. Zhang, 'Detecting and rationalizing concept drift: A feature-level approach for understanding cause–effect relationships in dynamic environments', *Expert Systems with Applications*, vol. 260, p. 125 365, 2025, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2024.125365. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424022322.

[22] R. González-Sendino, E. Serrano and J. Bajo, 'Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making', *Future Generation Computer Systems*, vol. 155, pp. 384–401, 2024, ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2024.02.023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X24000694.

[23] B. I. Igoche, O. Matthew and D. Olabanji, 'A structural causal model ontology approach for knowledge discovery in educational admission databases', *Knowledge*, vol. 5, no. 3, 2025, ISSN: 2673-9585. [Online]. Available: https://www.mdpi.com/2673-9585/5/3/15.

[24] T. Gerstenberg, 'Counterfactual simulation in causal cognition', *Trends in Cognitive Sciences*, vol. 28, no. 10, pp. 924–936, 2024, ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2024.04.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661324001074.

[25] J. Zhang and S. Chen, 'Expand horizon: Graph out-of-distribution generalization via multi-level environment inference', in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI-25 Technical Tracks 12, vol. 39, 2025.

[26] S. Liu, J. Chen, Z. Shi, L. Song and S. He, 'Representations aligned counterfactual domain learning for open-set fault diagnosis under speed transient conditions', *Knowledge-Based Systems*, vol. 310, p. 112 932, 2025, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2024.112932. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705124015661.

[27] E. Medcalf, F. Stanaway, R. M. Turner, D. Espinoza and K. J. Bell, 'Using the counterfactual framework to estimate non-intention-to-treat estimands in randomised controlled trials: A methodological scoping review', *Contemporary Clinical Trials*, vol. 153, p. 107 912, 2025, ISSN: 1551-7144. DOI: https://doi.org/10.1016/j.cct.2025.107912. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1551714425001065.

[28] L. E. J. Bynum *et al.*, *Black box causal inference: Effect estimation via meta prediction*, 2025. arXiv: 2503.05985 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2503.05985.

[29] E. S. Ortigossa, F. F. Dias, B. Barr, C. T. Silva and L. G. Nonato, 'T-explainer: A model-agnostic explainability framework based on gradients', *IEEE Intelligent Systems*, pp. 1–10, 2025. DOI: 10.1109/MIS.2025.3564330.

[30] Y. Wang *et al.*, 'Causal invariant geographic network representations with feature and structural distribution shifts', *Future Generation Computer Systems*, vol. 169, p. 107 814, 2025, ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2025.107814. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X25001098.

[31] A. Tejada-Lapuerta, P. Bertin, S. Bauer, H. Aliee, Y. Bengio and F. J. Theis, 'Causal machine learning for single-cell genomics', *Nature Genetics*, vol. 57, no. 4, pp. 797–808, Apr. 2025. DOI: 10.1038/s41588-025-02124-2.

[32] Z. Lu, B. Lu and F. Wang, 'Causalsr: Structural causal model-driven super-resolution with counterfactual inference', *Neurocomputing*, vol. 646, p. 130 375, 2025, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2025.130375. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231225010471.

[33] S. M. Lundberg and S. Lee, 'A Unified Approach to Interpreting Model Predictions', in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4768–4777.

[34] M. T. Ribeiro, S. Singh and C. Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, 'The unreasonable effectiveness of deep features as a perceptual metric', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, 'Image quality assessment: From error visibility to structural similarity', *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: 10.1109/TIP.2003.819861.

[37] A. Zanga, E. Ozkirimli and F. Stella, 'A survey on causal discovery: Theory and practice', *International Journal of Approximate Reasoning*, vol. 151, pp. 101–129, Dec. 2022, ISSN: 0888-613X. DOI: 10.1016/j.ijar.2022.09.004. [Online]. Available: http://dx.doi.org/10.1016/j.ijar.2022.09.004.

[38] A. Holzinger, G. Langs, H. Denk, K. Zatloukal and H. Müller, 'Causability and explainability of artificial intelligence in medicine', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019. DOI: 10.1002/widm.1312.

[39] Y. Liu, Y. Wei, H. Yan, G. Li and L. Lin, *Causal reasoning meets visual representation learning: A prospective study*, 2023. arXiv: 2204.12037 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.12037.

[40] E. Hosseinkhani, M. Leucker, M. Sachenbacher, H. Streichhahn and L. B. Vosteen, 'A model-based approach for monitoring and diagnosing digital twin discrepancies', in *35th International Conference on Principles of Diagnosis and Resilient Systems (DX 2024)*, ser. Open Access Series in Informatics (OASIcs), vol. 125, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024, 2:1–2:15. DOI: 10.4230/OASIcs.DX.2024.2. [Online]. Available: https://doi.org/10.4230/OASIcs.DX.2024.2.

[41] M. Diehl and K. Ramirez-Amaro, 'Why did i fail? a causal-based method to find explanations for robot failures', *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8925–8932, Oct. 2022. DOI: 10.1109/LRA.2022.3188889. [Online]. Available: https://doi.org/10.1109/LRA.2022.3188889.

[42] P. Spirtes, C. Glymour and R. Scheines, 'Causation, prediction, and search', 2000.

[43] P. Spirtes, C. Meek and T. Richardson, 'An algorithm for causal inference in the presence of latent variables and selection bias', *Computational Statistics & Data Analysis*, vol. 39, no. 1, pp. 1–26, 2001.

[44] Y. Zheng *et al.*, 'Causal-learn: Causal discovery in python', in *arXiv preprint*, vol. arXiv:2307.16405, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.16405.

[45] Z. Jin *et al.*, 'Cladder: Assessing causal reasoning in language models', in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, Eds., vol. 36, Curran Associates, Inc., 2023, pp. 31 038–31 065. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/631bb9434d718ea309af82566347d607-Paper-Conference.pdf.

[46] C. Zhang, Y. Wang, S. Li and X. Chen, 'Root cause diagnosis in process industry via bayesian network enhanced by prior knowledge and randomized optimization', *Chemical Engineering Science*, vol. 312, p. 121 683, 2025, ISSN: 0009-2509. DOI: https://doi.org/10.1016/j.ces.2025.121683. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0009250925005068.

[47] H. Dewantoro, A. Smith and P. Daoutidis, 'Causal discovery for topology reconstruction in industrial chemical processes', *Industrial & Engineering Chemistry Research*, vol. 63, no. 26, pp. 11 530–11 543, 2024. DOI: 10.1021/acs.iecr.4c01155. eprint: https://doi.org/10.1021/acs.iecr.4c01155. [Online]. Available: https://doi.org/10.1021/acs.iecr.4c01155.

[48] C. Gong, C. Zhang, D. Yao, J. Bi, W. Li and Y. Xu, 'Causal discovery from temporal data: An overview and new perspectives', *ACM Comput. Surv.*, vol. 57, no. 4, Dec. 2024, ISSN: 0360-0300. DOI: 10.1145/3705297. [Online]. Available: https://doi.org/10.1145/3705297.

[49] C.-M. Lin, C. Chang, W.-Y. Wang, K.-D. Wang and W.-C. Peng, *Root cause analysis in microservice using neural granger causal discovery*, 2024. arXiv: 2402.01140 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2402.01140.

[50] M. Naser, 'Discovering causal models for structural, construction and defense-related engineering phenomena', *Defence Technology*, vol. 43, pp. 60–79, 2025, ISSN: 2214-9147. DOI: https://doi.org/10.1016/j.dt.2024.04.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214914724000849.

[51] L. Allen and J. Cordiner, 'Knowledge-enhanced data-driven modeling of wastewater treatment processes for energy consumption prediction', *Computers Chemical Engineering*, vol. 194, p. 108 982, 2025, ISSN: 0098-1354. DOI: https://doi.org/10.1016/j.compchemeng.2024.108982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135424004009.

[52] G. Valverde, D. Quesada, P. Larrañaga and C. Bielza, 'Causal reinforcement learning based on bayesian networks applied to industrial settings', *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106 657, 2023, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2023.106657. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623008412.

[53] R. R. M. Ribeiro, J. Natal, C. P. de Campos and C. Dias Maciel, 'Conditional probability table limit-based quantization for bayesian networks: Model quality, data fidelity and structure score', *Applied Intelligence*, vol. 54, no. 6, pp. 4668–4688, 2024, ISSN: 0924-669X. DOI: 10.1007/s10489-023-05153-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135424004009.

[54] Q. Zhang, Z. Ma and Z. Cui, 'A causality-based explainable ai method for bus delay propagation analysis', *Communications in Transportation Research*, vol. 5, p. 100 178, 2025, ISSN: 2772-4247. DOI: https://doi.org/10.1016/j.commtr.2025.100178. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772424725000186.

[55] R. S. Chauhan *et al.*, 'Determining causality in travel mode choice', *Travel Behaviour and Society*, vol. 36, p. 100 789, 2024, ISSN: 2214-367X. DOI: https://doi.org/10.1016/j.tbs.2024.100789. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214367X24000528.

[56] W. Hu, Y. Wang, Y. He, L. Qian, D. Zhang and Y. Meng, 'Soft sensing technique for mass customization based on heterogeneous causal graph attention networks', *Advanced Engineering Informatics*, vol. 65, p. 103 139, 2025, ISSN: 1474-0346. DOI: https://doi.org/10.1016/j.aei.2025.103139. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034625000321.

[57] B. Zhao *et al.*, 'Coresets for fast causal discovery with the additive noise model', *Pattern Recognition*, vol. 148, p. 110 149, 2024, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2023.110149. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323008464.

[58] Y. He, X. Zhang, X. Kong, L. Yao and Z. Song, 'Causality-driven sequence segmentation assisted soft sensing for multiphase industrial processes', *Neurocomputing*, vol. 631, p. 129 612, 2025, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2025.129612. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122500284X.

[59] X. Zheng, B. Aragam, P. Ravikumar and E. P. Xing, 'Dags with no tears: Continuous optimization for structure learning', in *arXiv preprint*, 2018. arXiv: 1803.01422 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1803.01422.

[60] X. Yang, H. Zhang, G. Qi and J. Cai, 'Causal attention for vision-language tasks', in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9842–9852. DOI: 10.1109/CVPR46437.2021.00972.

[61] C. Moccia, G. Moirano, M. Popovic and E. Al., 'Machine learning in causal inference for epidemiology', *European Journal of Epidemiology*, vol. 39, pp. 1097–1108, 2024. DOI: 10.1007/s10654-024-01173-x. [Online]. Available: https://doi.org/10.1007/s10654-024-01173-x.

[62] L. Castri, G. Beraldo, S. Mghames, M. Hanheide and N. Bellotto, 'Ros-causal: A ros-based causal analysis framework for human-robot interaction applications', *arXiv preprint*, vol. arXiv:2402.16068, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.16068.

[63] P. Hünermund and E. Bareinboim, 'Causal inference and data fusion in econometrics', *The Econometrics Journal*, vol. 28, no. 1, pp. 41–82, Mar. 2023. DOI: 10.1093/ectj/utad008. [Online]. Available: https://doi.org/10.1093/ectj/utad008.

[64] D. H. Bailey, A. J. Jung, A. M. Beltz *et al.*, 'Causal inference on human behaviour', *Nature Human Behaviour*, vol. 8, pp. 1448–1459, 2024. DOI: 10.1038/s41562-024-01939-z. [Online]. Available: https://doi.org/10.1038/s41562-024-01939-z.

[65] J. Dörfler, B. van der Zander, M. Bläser and M. Liskiewicz, *On the complexity of identification in linear structural causal models*, 2024. arXiv: 2407.12528 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2407.12528.

[66] G. V. Goffrier, L. Maystre and C. Gilligan-Lee, *Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding*, 2023. arXiv: 2302.10625 [stat.ML]. [Online]. Available: https://arxiv.org/abs/2302.10625.

[67] M. A. Hernán and J. M. Robins, 'Causal inference: What if', *Chapman & Hall/CRC*, 2020.

[68] I. Shpitser and J. Pearl, 'Complete identification methods for the causal hierarchy', *Journal of Machine Learning Research*, vol. 9, pp. 1941–1979, 2008.

[69] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, UK: Cambridge University Press, 2015.

[70] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

[71] H. Bang and J. M. Robins, 'Doubly robust estimation in missing data and causal inference models', *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[72] M. Herchenbach, S. Weinzierl, S. Zilker, E. Schwulera and M. Matzner, 'A methodology for adaptive ai-based causal control: Toward an autonomous factory in solder paste printing', *Computers in Industry*, vol. 167, p. 104 256, 2025, ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2025.104256. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166361525000211.

[73] R. Liu, Q. Zhang, D. Lin, W. Zhang and S. X. Ding, 'Causal intervention graph neural network for fault diagnosis of complex industrial processes', *Reliability Engineering System Safety*, vol. 251, p. 110 328, 2024, ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2024.110328. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832024004009.

[74] Y. Zhang *et al.*, 'Intervening on few-shot object detection based on the front-door criterion', *Neural Networks*, vol. 185, p. 107 251, 2025, ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2025.107251. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608025001303.

[75] X. Zhu, Y. Zhang, F. Feng, X. Yang, D. Wang and X. He, ' Mitigating Hidden Confounding Effects for Causal Recommendation ', *IEEE Transactions on Knowledge & Data Engineering*, vol. 36, no. 09, pp. 4794–4805, Sep. 2024, ISSN: 1558-2191. DOI: 10.1109/TKDE.2024.3378482. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TKDE.2024.3378482.

[76] Z. Zhao *et al.*, 'Networked instrumental variable for treatment effect estimation with unobserved confounders', *IEEE Trans. on Knowl. and Data Eng.*, vol. 37, no. 2, pp. 823–836, Feb. 2025, ISSN: 1041-4347. DOI: 10.1109/TKDE.2024.3491776. [Online]. Available: https://doi.org/10.1109/TKDE.2024.3491776.

[77] H. Xiong, F. Wu, L. Deng, M. Su, Z. Shahn and L. H. Lehman, 'G-Transformer: Counterfactual outcome prediction under dynamic and time-varying treatment regimes', in *Proceedings of Machine Learning Research*, PMID:40433313, PMCID:PMC12113242, vol. 252, PMLR, Aug. 2024. [Online]. Available: https://proceedings.mlr.press/v252/xiong24a.html.

[78] C. Uhde, N. Berberich, H. Ma, R. Guadarrama and G. Cheng, 'Learning causal relationships of object properties and affordances through human demonstrations and self-supervised intervention for purposeful action in transfer environments', *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 015–11 022, 2022. DOI: 10.1109/LRA.2022.3196125.

[79] S. Wager and S. A. and, 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018. DOI: 10.1080/01621459.2017.1319839. eprint: https://doi.org/10.1080/01621459.2017.1319839. [Online]. Available: https://doi.org/10.1080/01621459.2017.1319839.

[80] H. A. Chipman, E. I. George and R. E. McCulloch, 'Bart: Bayesian additive regression trees', *The Annals of Applied Statistics*, vol. 4, no. 1, Mar. 2010, ISSN: 1932-6157. DOI: 10.1214/09-aoas285. [Online]. Available: http://dx.doi.org/10.1214/09-AOAS285.

[81] H. Gharoun, F. Momenifar, F. Chen and A. H. Gandomi, *Meta-learning approaches for few-shot learning: A survey of recent advances*, 2023. arXiv: 2303.07502 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2303.07502.

[82] J. Mun and C. S. Park, 'Beyond correlation: A causality-driven model for indoor temperature control', *Energy and Buildings*, vol. 338, p. 115 739, 2025, ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2025.115739. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778825004694.

[83] G. Demeter, R. Delina, J. Urminsky and M. Zajarosova, 'Assessing the impact of supplier group stability on public procurement performance in slovakia: A causal forest analysis of collusion signals', *IEEE Access*, vol. 13, pp. 41 607–41 624, 2025. DOI: 10.1109/ACCESS.2025.3547688.

[84] I. J. Goodfellow *et al.*, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1406.2661.

[85] U. M. Nabila, L. Lin, X. Zhao, W. L. Gurecky, P. Ramuhalli and M. I. Radaideh, 'Data efficiency assessment of generative adversarial networks in energy applications', *Energy and AI*, vol. 20, p. 100 501, 2025, ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2025.100501. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666546825000333.

[86] L. Xu, H. Li, C. Shao, M. Gao and H. Shen, 'Infinite high fidelity thin cloud synthesis by coupling scattering law and generative adversarial network', in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 2528–2531. DOI: 10.1109/IGARSS53475.2024.10642090.

[87] Z. Wang, H. Xia, J. Zhang, B. Yang and W. Yin, 'Imbalanced sample fault diagnosis method for rotating machinery in nuclear power plants based on deep convolutional conditional generative adversarial network', *Nuclear Engineering and Technology*, vol. 55, no. 6, pp. 2096–2106, 2023, ISSN: 1738-5733. DOI: https://doi.org/10.1016/j.net.2023.02.036. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1738573323001092.

[88] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2022. arXiv: 1312.6114 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1312.6114.

[89] Z. Wang, C. Wang and Y. Li, 'Variational autoencoder based on knowledge sharing and correlation weighting for process-quality concurrent fault detection', *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108 051, 2024, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2024.108051. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197624002094.

[90] M. Baur, B. Fesl and W. Utschick, 'Leveraging variational autoencoders for parameterized mmse estimation', *IEEE Transactions on Signal Processing*, vol. 72, pp. 3731–3744, 2024. DOI: 10.1109/TSP.2024.3439097.

[91] J. Chan, T. Han and E. Pan, 'Variational autoencoder-driven adversarial svdd for power battery anomaly detection on real industrial data', *Journal of Energy Storage*, vol. 103, p. 114 267, 2024, ISSN: 2352-152X. DOI: https://doi.org/10.1016/j.est.2024.114267. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352152X24038532.

[92] J. Ho, A. Jain and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2006.11239.

[93] C. Zeng, J. Wang, W. Wang, K. Hai, S. Ma and L. Wei, 'Data-driven structural generative design based on diffusion model for flexible support of optical mirror', *Engineering Structures*, vol. 338, p. 120 578, 2025, ISSN: 0141-0296. DOI: https://doi.org/10.1016/j.engstruct.2025.120578. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141029625009691.

[94] F. P. de Azevedo, N. C. Correia Lourenço and R. M. Ferreira Martins, 'Comprehensive application of denoising diffusion probabilistic models towards the automation of analog integrated circuit sizing', *Expert Systems with Applications*, vol. 290, p. 128 414, 2025, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2025.128414. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425020330.

[95] A. Vaswani *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[96] J. Wang, G. Xu, F. Yan, J. Wang and Z. Wang, 'Defect transformer: An efficient hybrid transformer architecture for surface defect detection', *Measurement*, vol. 211, p. 112 614, 2023, ISSN: 0263-2241. DOI: https://doi.org/10.1016/j.measurement.2023.112614. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224123001781.

[97] D. Bau *et al.*, *Gan dissection: Visualizing and understanding generative adversarial networks*, 2018. arXiv: 1811.10597 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1811.10597.

[98] E. Härkönen, A. Hertzmann, J. Lehtinen and S. Paris, *Ganspace: Discovering interpretable gan controls*, 2020. arXiv: 2004.02546 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2004.02546.

[99] J. Schneider, 'Explainable generative ai (genxai): A survey, conceptualization, and research agenda', *arXiv preprint arXiv:2404.09554*, 2024. DOI: 10.48550/arXiv.2404.09554. [Online]. Available: https://doi.org/10.48550/arXiv.2404.09554.

[100] F. Locatello *et al.*, 'Challenging common assumptions in the unsupervised learning of disentangled representations', in *arXiv preprint*, 2019. arXiv: 1811.12359 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1811.12359.

[101] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin and P.-A. Heng, 'Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion', in *arXiv preprint*, 2020. arXiv: 2002.09708 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2002.09708.

[102] C. Pei, F. Wu, L. Huang and X. Zhuang, 'Disentangle domain features for cross-modality cardiac image segmentation', in *Medical Image Analysis*, vol. 71, Elsevier, 2021, p. 102 078. DOI: 10.1016/j.media.2021.102078. [Online]. Available: https://doi.org/10.1016/j.media.2021.102078.

[103] J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin and J. S. Duncan, 'Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation', in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, Epub 2019 Oct 10, vol. 11765, Cham, Switzerland: Springer, Oct. 2019, pp. 255–263. DOI: 10.1007/978-3-030-32245-8_29.

[104] W. Zeng *et al.*, 'Seen to unseen: Exploring compositional generalization of multi-attribute controllable dialogue generation', in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 179–14 196. DOI: 10.18653/v1/2023.acl-long.793. [Online]. Available: https://aclanthology.org/2023.acl-long.793.

[105] X. Wang, H. Chen, S. Tang, Z. Wu and W. Zhu, 'Disentangled representation learning', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 46, no. 12, pp. 9677–9696, Dec. 2024. DOI: 10.1109/TPAMI.2024.3420937.

[106] Z. Han, T. Luo, H. Fu, Q. Hu, J. T. Zhou and C. Zhang, 'A principled framework for explainable multimodal disentanglement', *Information Sciences*, vol. 675, p. 120 768, 2024, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2024.120768. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025524006820.

[107] I. Higgins *et al.*, 'Beta-VAE: Learning basic visual concepts with a constrained variational framework', in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl.

[108] H. Kim and A. Mnih, *Disentangling by factorising*, 2019. arXiv: 1802.05983 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1802.05983.

[109] A. Kumar, P. Sattigeri and A. Balakrishnan, *Variational inference of disentangled latent concepts from unlabeled observations*, 2018. arXiv: 1711.00848 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1711.00848.

[110] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, *Infogan: Interpretable representation learning by information maximizing generative adversarial nets*, 2016. arXiv: 1606.03657 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1606.03657.

[111] Z. Yue, T. Wang, H. Zhang, Q. Sun and X.-S. Hua, *Counterfactual zero-shot and open-set visual recognition*, 2021. arXiv: 2103.00887 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.00887.

[112] A. Dhir, M. Ashman, J. Requeima and M. van der Wilk, *A meta-learning approach to bayesian causal discovery*, 2025. arXiv: 2412.16577 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2412.16577.

[113] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao and J. Wang, *Causalvae: Structured causal disentanglement in variational autoencoder*, 2023. arXiv: 2004.08697 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2004.08697.

[114] Y. Shi *et al.*, 'Diffusion models for medical image computing: A survey', *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 357–383, 2025. DOI: 10.26599/TST.2024.9010047. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2024.9010047.

[115] K. Sueyoshi and T. Matsubara, *Predicated diffusion: Predicate logic-based attention guidance for text-to-image diffusion models*, 2024. arXiv: 2311.16117 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2311.16117.

[116] K. Preechakul, N. Chatthee, S. Wizadwongsa and S. Suwajanakorn, *Diffusion autoencoders: Toward a meaningful and decodable representation*, 2022. arXiv: 2111.15640 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2111.15640.

[117] P. Sanchez and S. A. Tsaftaris, *Diffusion causal models for counterfactual estimation*, 2022. arXiv: 2202.10166 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2202.10166.

[118] A. Komanduri, C. Zhao, F. Chen and X. Wu, *Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models*, 2024. arXiv: 2404.17735 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2404.17735.

[119] H. Aetesam and S. K. Maji, 'Deep variational magnetic resonance image denoising via network conditioning', *Biomedical Signal Processing and Control*, vol. 95, p. 106 452, 2024, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2024.106452. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S174680942400510X.

[120] C. Deng, D. Zhu, K. Li, S. Guang and H. Fan, *Causal diffusion transformers for generative modeling*, 2024. arXiv: 2412.12095 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2412.12095.

[121] H. Lin *et al.*, *Causal composition diffusion model for closed-loop traffic generation*, 2025. arXiv: 2412.17920 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2412.17920.

[122] S. Minaee *et al.*, *Large language models: A survey*, 2025. arXiv: 2402.06196 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2402.06196.

[123] X. Jiang, W. Wang, S. Tian, H. Wang, T. Lookman and Y. Su, 'Applications of natural language processing and large language models in materials discovery', *npj Computational Materials*, vol. 11, p. 79, 2025. DOI: 10.1038/s41524-025-01554-0. [Online]. Available: https://www.nature.com/articles/s41524-025-01554-0.

[124] J. Liu, F. Lin, X. Li, K. H. Lim and S. Zhao, *Physics-informed llm-agent for automated modulation design in power electronics systems*, 2024. arXiv: 2411.14214 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2411.14214.

[125] B. Liu *et al.*, *Layoutcopilot: An llm-powered multi-agent collaborative framework for interactive analog layout design*, 2025. arXiv: 2406.18873 [cs.AR]. [Online]. Available: https://arxiv.org/abs/2406.18873.

[126] Y. He *et al.*, *Llms meet multimodal generation and editing: A survey*, 2024. arXiv: 2405.19334 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2405.19334.

[127] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, *Hierarchical text-conditional image generation with clip latents*, 2022. arXiv: 2204.06125 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.06125.

[128] A. Ramesh *et al.*, *Zero-shot text-to-image generation*, 2021. arXiv: 2102.12092 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2102.12092.

[129] M. Ding *et al.*, *Cogview: Mastering text-to-image generation via transformers*, 2021. arXiv: 2105.13290 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2105.13290.

[130] C. Saharia *et al.*, *Photorealistic text-to-image diffusion models with deep language understanding*, 2022. arXiv: 2205.11487 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2205.11487.

[131] J. Alcaide-Marzal and J. A. Diego-Mas, 'Computers as co-creative assistants. a comparative study on the use of text-to-image ai models for computer aided conceptual design', *Computers in Industry*, vol. 164, p. 104 168, 2025, ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2024.104168. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166361524000964.

[132] H. Wang, M. Zhao, Y. Wang, W. Quan and D.-M. Yan, 'Vq-cad: Computer-aided design model generation with vector quantized diffusion', *Computer Aided Geometric Design*, vol. 111, p. 102 327, 2024, ISSN: 0167-8396. DOI: https://doi.org/10.1016/j.cagd.2024.102327. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016783962400061X.

[133] J. Wei *et al.*, *Chain-of-thought prompting elicits reasoning in large language models*, 2023. arXiv: 2201.11903 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2201.11903.

[134] C. Wang *et al.*, *Mllm-tool: A multimodal large language model for tool agent learning*, 2025. arXiv: 2401.10727 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2401.10727.

[135] J. Ho, A. Jain and P. Abbeel, 'Denoising diffusion probabilistic models', in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 2020-December, Jun. 2020. [Online]. Available: https://arxiv.org/abs/2006.11239v2.

[136] J. Wang *et al.*, *A comprehensive review of multimodal large language models: Performance and challenges across different tasks*, 2024. arXiv: 2408.01319 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2408.01319.

[137] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang and N. Duan, 'Visual chatgpt: Talking, drawing and editing with visual foundation models', *arXiv preprint*, vol. arXiv:2303.04671, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.04671.

[138] H. Liu, C. Li, Q. Wu and Y. J. Lee, 'Visual instruction tuning', *arXiv preprint*, vol. arXiv:2304.08485, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.08485.

[139] H. Liu, C. Li, Y. Li and Y. J. Lee, *Improved baselines with visual instruction tuning*, 2024. arXiv: 2310.03744 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2310.03744.

[140] T. Brooks, A. Holynski and A. A. Efros, 'Instructpix2pix: Learning to follow image editing instructions', *arXiv preprint*, vol. arXiv:2211.09800, Nov. 2023. [Online]. Available: https://arxiv.org/abs/2211.09800.

[141] Y. Huang *et al.*, 'Smartedit: Exploring complex instruction-based image editing with multimodal large language models', *arXiv preprint*, vol. arXiv:2312.06739, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2312.06739.

[142] X. Li *et al.*, 'Image content generation with causal reasoning', *arXiv preprint*, vol. arXiv:2312.07132, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2312.07132.

[143] D. Heckerman, *A tutorial on learning with bayesian networks*, 2022. arXiv: 2002.00269 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2002.00269.

[144] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.

[145] J. Peters, D. Janzing and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[146] B. Schölkopf *et al.*, *Towards causal representation learning*, 2021. arXiv: 2102.11107 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2102.11107.

[147] J. Chung, Y. Ahn, D. Shin and G. Park, 'Learning distribution-free anchored linear structural equation models in the presence of measurement error', *Journal of the Korean Statistical Society*, vol. 54, pp. 361–385, 2025. DOI: 10.1007/s42952-024-00298-9.

[148] M. Arjovsky, L. Bottou, I. Gulrajani and D. Lopez-Paz, *Invariant risk minimization*, 2020. arXiv: 1907.02893 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1907.02893.

[149] M. Arjovsky, L. Bottou, I. Gulrajani and D. Lopez-Paz, 'Invariant risk minimization', in *International Conference on Learning Representations (ICLR)*, 2020.

[150] H. Rahimian and S. Mehrotra, 'Distributionally robust optimization: A review', *SIAM Review*, vol. 61, no. 3, pp. 464–501, 2019.

[151] S. Zhao *et al.*, *Cv-vae: A compatible video vae for latent generative video models*, 2024. arXiv: 2405.20279 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2405.20279.

[152] G. E. Hinton and R. R. Salakhutdinov, 'Reducing the dimensionality of data with neural networks', in *Science*, 5786, vol. 313, 2006, pp. 504–507.

[153] V. Nair and G. E. Hinton, 'Rectified linear units improve restricted boltzmann machines', in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[154] D. P. Kingma and M. Welling, 'Auto-encoding variational bayes', *International Conference on Learning Representations (ICLR)*, 2014.

[155] R. M. Schmidt, *Recurrent neural networks (rnns): A gentle introduction and overview*, 2019. arXiv: 1912.05911 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1912.05911.

[156] R. C. Staudemeyer and E. R. Morris, *Understanding lstm – a tutorial into long short-term memory recurrent neural networks*, 2019. arXiv: 1909.09586 [cs.NE]. [Online]. Available: https://arxiv.org/abs/1909.09586.

[157] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, *Empirical evaluation of gated recurrent neural networks on sequence modeling*, 2014. arXiv: 1412.3555 [cs.NE]. [Online]. Available: https://arxiv.org/abs/1412.3555.

[158] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud *et al.*, 'Emergent abilities of large language models', *Transactions on Machine Learning Research*, 2022.

[159] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal *et al.*, 'Language models are few-shot learners', *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[160] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora *et al.*, 'On the opportunities and risks of foundation models', *arXiv preprint arXiv:2108.07258*, 2021.

[161] A. Fan, M. Lewis and Y. Dauphin, 'Hierarchical neural story generation', *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[162] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, 'Language models are unsupervised multitask learners', *OpenAI Technical Report*, 2019.

[163] E. J. Hu *et al.*, 'Lora: Low-rank adaptation of large language models', in *International Conference on Learning Representations (ICLR)*, 2022.

[164] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, 'The unreasonable effectiveness of deep features as a perceptual metric', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[165] I. T. Jolliffe and J. Cadima, *Principal Component Analysis*, 2nd ed. Springer, 2016.

[166] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao and J. Wang, 'Causalvae: Structured causal disentanglement in variational autoencoder', in *arXiv preprint*, Apr. 2020. DOI: 10.48550/arxiv.2004.08697. [Online]. Available: https://arxiv.org/abs/2004.08697.

[167] M. Chen, H. Wang, R. Wang, Y. Peng and H. Zhang, 'Cdrm: Causal disentangled representation learning for missing data', *Knowledge-Based Systems*, vol. 299, p. 112 079, 2024, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2024.112079. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705124007135.

[168] J. Zhu *et al.*, *Shadow datasets, new challenging datasets for causal representation learning*, 2023. arXiv: 2308.05707 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2308.05707.

[169] D. C. Castro, J. Tan, B. Kainz, E. Konukoglu and B. Glocker, *Morpho-mnist: Quantitative assessment and diagnostics for representation learning*, 2019. arXiv: 1809.10780 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1809.10780.

[170] Z. Liu, P. Luo, X. Wang and X. Tang, *Deep learning face attributes in the wild*, 2015. arXiv: 1411.7766 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1411.7766.

[171] J. Brehmer, P. de Haan, P. Lippe and T. Cohen, *Weakly supervised causal representation learning*, 2022. arXiv: 2203.16437 [stat.ML]. [Online]. Available: https://arxiv.org/abs/2203.16437.

[172] P. Dhariwal and A. Nichol, *Diffusion models beat gans on image synthesis*, 2021. arXiv: 2105.05233 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2105.05233.

[173] F. Li *et al.*, *Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models*, 2024. arXiv: 2407.07895 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2407.07895.

[174] K. Zhang, L. Mo, W. Chen, H. Sun and Y. Su, *Magicbrush: A manually annotated dataset for instruction-guided image editing*, 2024. arXiv: 2306.10012 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2306.10012.

[175] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2021. arXiv: 2112.10752 [cs.CV].