# Toward Causal Generative Modeling: From Representation to Generation

**Aneesh Komanduri**

Department of Electrical Engineering and Computer Science
University of Arkansas, Fayetteville, AR, USA, 72701
akomandu@uark.edu

## Abstract

Deep learning has given rise to the field of representation learning, which aims to automatically extract rich semantics from data. However, there have been several challenges in the generalization capabilities of deep learning models. Recent works have highlighted beneficial properties of causal models that are desirable for learning robust models under distribution shifts. Thus, there has been a growing interest in causal representation learning for achieving generalizability in tasks involving reasoning and planning. The goal of my dissertation is to develop theoretical intuitions and practical algorithms that uncover the nature of causal representations and their applications. In my work, I focus on causal generative modeling with an emphasis on either representation or generation. For representation learning, I investigate the disentanglement of causal representations through the lens of independent causal mechanisms. For generation tasks, I develop algorithms for counterfactual generation under weak supervision settings by leveraging recent advances in generative modeling. The proposed approaches have been empirically shown to be effective in achieving disentanglement and generating counterfactuals.

## Introduction

*One major contribution of AI to the study of cognition has been the paradigm "Representation first, acquisition second." — (Pearl and Mackenzie 2018)*

The hallmark of human intelligence is causal reasoning, the ability to infer causes and effects from observation and experimentation. Many have conjectured that the knowledge represented in our brain is *causal* in nature and takes the form of reusable and modular mechanisms (Schölkopf et al. 2021). Infants, in their early development, learn about the world through manipulation, a concept equivalent to interventions in causal inference through which we can reason about cause and effect. Thus, artificial intelligence agents should be equipped with representations that enable learning and inferring causality from data to achieve generalizability.

Representation learning aims to extract interpretable and semantically meaningful concepts from data that can be used for robust learning in downstream tasks. Recently, there has been a growing effort in bridging structural causal models (Pearl and Mackenzie 2018) and representation learning. The goal of *causal representation learning* is to extract causal variables from high-dimensional data (Schölkopf et al. 2021). The key feature of causal variables is that we can intervene on them to reason about causal effects on other factors and generate counterfactuals, which are desired in several critical domains.

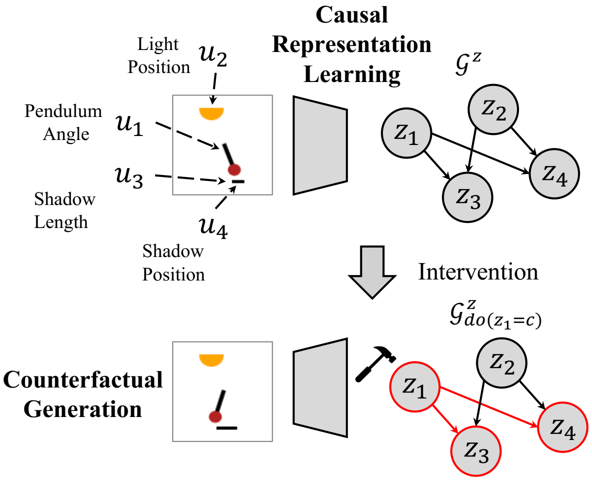

Figure 1: Illustration of Causal Representation Learning and Controllable Counterfactual Generation

Motivated by the current shift toward causal generative modeling, I have written an extensive survey that delves into **causal representation learning** and **counterfactual generation** (Komanduri et al. 2024a), which are the two main themes of my dissertation. The former is useful for a variety of downstream learning tasks and offers distribution shift robustness. The latter has significant applications in domains such as healthcare, where counterfactuals are crucial for reducing data collection costs and reasoning about treatment plans. My dissertation research aims to study three main questions: (1) How can we incorporate causality into generative models? (2) Under what conditions can we provably recover and disentangle causal factors? (3) Can we utilize causal representation guidance in generative models for high-fidelity counterfactual generation?

## Proposed Work

Causal representation learning fundamentally stems from the well-established area of independent component analysis (ICA) (Comon 1994), where we are interested in separating independent latent sources that explain the observational data. Recent work has shown that the unsupervised disentanglement of latent factors, using generative models such as VAEs, is impossible without weak supervision (Locatello et al. 2018). My work builds on ICA and disentangled representation learning to develop theory and models for recovering and disentangling *causally related* factors and enabling controllable counterfactual generation. The following is my current research progress in causal generative modeling.

### Exploring Causality in Generative Models

Komanduri et al. (2022) explore to what extent we can incorporate causality in variational autoencoders for causal representation learning. Specifically, we propose SCM-VAE, a VAE-based framework to learn causal representations, and conclude that providing more useful causal information in the form of priors is critical for causal representation learning. We improve upon previous methods by introducing nonlinearities in the structural causal model and providing better guidance for causal representation learning via a causal prior. Our empirical results show that the learned representation aligns closer to the ground-truth causal variables and is compatible with counterfactual generation.

### Learning Causally Disentangled Representations

Based on the intuition that priors are crucial in generative modeling, we study how to achieve disentanglement of the learned causal factors through the lens of causal mechanisms in Komanduri et al. (2024b). Traditionally, the notion of *disentanglement* implies that the learned factors are variable along orthogonal axes and are non-overlapping. Intuitively, this means that each factor is represented by a unique set of latent codes and can be recovered up to some permutation and scaling of the true factors. However, in causal models, the notion of disentanglement must also satisfy the recoverability of the causal mechanisms. To incorporate this notion, we propose a new definition of disentanglement from the perspective of independent causal mechanisms for causal models. Then, we develop ICM-VAE, a framework for causal representation learning that (1) parameterizes causal mechanism via autoregressive normalizing flows to learn flexible and general bijective mappings in the latent space, and (2) utilizes a causally factorized prior distribution to learn disentangled causal mechanisms. Empirical results show high causal disentanglement, robustness of interventions, and counterfactual generation capability.

### Toward High-Fidelity Counterfactual Generation

Besides representation learning, controllable generation is also a significant task of great interest in many domains. Recently, diffusion probabilistic models have surged in popularity due to their impressive capabilities in high-quality image generation. The question we aim to answer is: How can we bridge the gap between causal learning and diffusion models to enable high-quality *counterfactual* image generation? In Komanduri et al. (2024c), we propose CausalDiffAE, a diffusion-based causal representation learning framework for high-fidelity counterfactual generation. We aim to learn a joint representation capturing both causal and stochastic information from images and condition the diffusion model on the learned representation. We model causal mechanisms among factors by utilizing developments in neural causal models. To facilitate counterfactual generation during inference, we develop a deterministic sampling algorithm subject to interventions. Furthermore, we propose a weak-supervision training paradigm inspired by classifier-free guidance for scenarios with limited labeled data, which enables granular control over generated counterfactuals.

## Future Work

One particularly interesting direction that I am currently exploring is integrating causality with large-scale generative models. With the rise of pre-trained generative models, precisely controlling data generation has become a popular research direction. For instance, in text-to-image diffusion models, the text prompt can be interpreted as inducing a distribution over concepts. However, such concepts often embedded in prompts may be causally related and not necessarily independent. My future work aims to investigate the concept space of large language models and text-to-image generative models through a causal lens. In other words, how can we model the causal mechanisms among prompt concepts in the latent space? I am interested in guiding large-scale generative models via causal knowledge to enable counterfactual inference and model interpretability.

## References

Comon, P. 1994. Independent component analysis, A new concept? *Signal Processing*.

Komanduri, A.; Wu, X.; Wu, Y.; and Chen, F. 2024a. From Identifiable Causal Representations to Controllable Counterfactual Generation: A Survey on Causal Generative Modeling. *Transactions on Machine Learning Research (TMLR)*.

Komanduri, A.; Wu, Y.; Chen, F.; and Wu, X. 2024b. Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms. In *IJCAI*.

Komanduri, A.; Wu, Y.; Huang, W.; Chen, F.; and Wu, X. 2022. SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge. In *IEEE Big Data*.

Komanduri, A.; Zhao, C.; Chen, F.; and Wu, X. 2024c. Causal Diffusion Autoencoders: Toward Counterfactual Generation via Diffusion Probabilistic Models. In *ECAI*.

Locatello, F.; Bauer, S.; Lucic, M.; Gelly, S.; Scholkopf, B.; and Bachem, O. 2018. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *ICML*.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why*. New York: Basic Books. ISBN 978-0-465-09760-9.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5).