Causal Diffusion Autoencoders: Toward Representation-Enabled Counterfactual Generation via Diffusion Probabilistic Models

Aneesh Komanduri[1] Chen Zhao[2] Feng Chen[3] Xintao Wu[1]
[1]University of Arkansas [2]Baylor University
[3]University of Texas at Dallas
{akomandu, xintaowu}@uark.edu chen_zhao@baylor.edu feng.chen@utdallas.edu

Abstract

Diffusion probabilistic models (DPMs) have become the state-of-the-art in high-quality image generation. However, DPMs have an arbitrary noisy latent space with no interpretable or controllable semantics. Although there has been significant research effort to improve image sample quality, there is little work on representation-enabled controllable generation using diffusion models. Specifically, controllable counterfactual generation using DPMs has been an underexplored area. In this work, we propose CausalDiffAE, a diffusion-based causal representation learning framework to enable counterfactual generation according to a specified causal model. We encode the high-dimensional image into a low-dimensional representation corresponding to causally related semantic factors. We model causal dependencies among latent variables using neural structural causal models and ensure their disentanglement via an alignment prior. Given a pre-trained CausalDiffAE, we propose a DDIM-based counterfactual generation procedure subject to do-interventions. We empirically show that CausalDiffAE learns a disentangled latent space and is capable of generating high-quality counterfactual images.

1. Introduction

Diffusion probabilistic models (DPMs) [7, 11, 22–24] are a class of likelihood-based generative models that have achieved remarkable successes in the generation of high-resolution images with many large-scale implementations such as DALLE-2 [17], Stable Diffusion [18], and Imagen [19]. Thus, there has been great interest in evaluating the capabilities of diffusion models. Two of the most promising approaches are formulated as discrete-time [7] and continuous-time [24] step-wise perturbations of the data distribution. A model is then trained to estimate the reverse process which transforms noisy samples to samples from the underlying data distribution. Representation learning has been an integral component of generative models such as GANs [5] and VAEs [9] for extracting robust and interpretable features from complex data [1, 16, 21]. Recently, a thrust of research has focused on whether DPMs can be used to extract a semantically meaningful and decodable representation that increases the quality of and control over generated images [12, 15]. However, there has been little work in modeling causal relations among the semantic latent codes to learn causal representations and enable *counterfactual generation* capabilities at inference time in DPMs. Generating high-quality counterfactual images is critical for domains such as healthcare and medicine [10, 20]. The ability to generate accurate counterfactual data from a causal graph obtained from domain knowledge can significantly cut the cost of data collection. Furthermore, reasoning about hypothetical scenarios unseen in the training distribution can be quite insightful for gauging the interactions among causal variables in complex systems. Given a causal graph of a system, we study the capability of DPMs as causal representation learners and evaluate their ability to generate counterfactuals upon interventions on causal variables.

In this paper, we focus on learning disentangled causal representations, where the high-level semantic factors are causally related. We propose **CausalDiffAE**, a learning framework for causal representation learning and controllable counterfactual generation in DPMs. Our key idea is to learn a causal representation via a learnable stochastic encoder and model the relations among latents via causal mechanisms parameterized by neural networks. We formulate a variational objective with a label alignment prior to enforce disentanglement of the learned causal factors. We then utilize a conditional DDIM [23] for decoding and modeling the stochastic variations. Intuitively, the causal representation encodes compact information that is *causally relevant* for image decoding in reverse diffusion. Furthermore, the modeling of causal relations in the latent space enables the generation of counterfactuals upon interventions on learned causal variables. Finally, we propose a DDIM variant for counterfactual generation subject to $\mathbf{do}(\cdot)$ interventions.

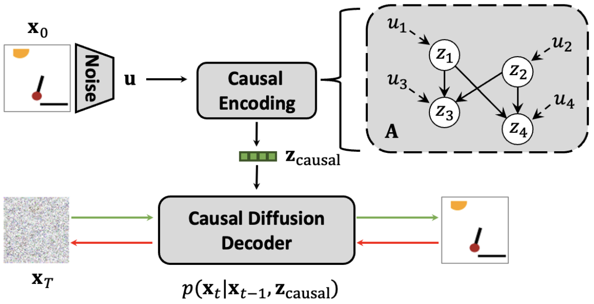


Figure 1. CausalDiffAE architecture

2. Causal Diffusion Autoencoders

The proposed framework, CausalDiffAE, consists of two main components. Firstly, we introduce a causal encoding scheme in the latent space as a structuring of the latent space and a training objective of a causal representation conditioned reverse diffusion process. In our formulation, we learn a causal representation $\mathbf{z}_{\text{causal}}$, which captures *causally relevant* information, in addition to $\mathbf{x}_T$, which captures low-level stochastic information. Together, the two latent variables $(\mathbf{z}_{\text{causal}}, \mathbf{x}_T)$ capture all the detailed causal semantics and stochasticity in the image. Secondly, given a pretrained diffusion model from the aforementioned method, we propose a counterfactual generation algorithm that utilizes do-interventions and the DDIM [23] sampling algorithm. The overall framework of CausalDiffAE is shown in Figure 1.

2.1. Causal Encoding

Let $\mathbf{x}_0 \in \mathbb{R}^d$ be the observed input image. We carry out the forward diffusion process until we have a set of T perturbed samples $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$, each at a different noise scale. Suppose there are n abstract causal variables that describe the high-level semantics of the observed image. To learn a meaningful representation, we propose to encode the input image $\mathbf{x}_0$ to a low-dimensional noise encoding $\mathbf{u} \in \mathbb{R}^n$. We then map the noise encoding to latent causal factors $\mathbf{z}_{\text{causal}} \in \mathbb{R}^n$ corresponding the the abstract causal variables. In this formulation, each noise term u_i is the exogenous noise term for causal variable z_i in the SCM. Let A be the adjacency matrix encoding the causal graph among the underlying factors. Then, we parameterize the mechanisms between causal variables as follows

$$z_i = f_i(z_{\mathbf{pa}_i}, u_i) \tag{1}$$

where f_i is the causal mechanism generating causal variable z_i as a function of its parents and exogenous noise term. In practice, we can implement f_i as a post-nonlinear additive noise model such that

$$z_i = f_i(A_i \circ z; \nu_i) + u_i \tag{2}$$

This module captures the causal relations between latent variables using neural structural causal models.

2.2. Generative Model

The forward diffusion process defines the perturbation of the image as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I) \tag{3}$$

where $\beta_t \in (0,1)$ is a variance parameter that controls the step size of noise. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Given a high-dimensional input image $\mathbf{x}_0$, an auxiliary weak supervision signal $\mathbf{y}$, a latent noise encoding $\mathbf{u}$, latent representation $\mathbf{z}_{\text{causal}}$, and a sequence of T latent representations $\mathbf{x}_{1:T}$ via forward diffusion, the CausalDiffAE generative process can be factorized as follows:

$$p(\mathbf{x}_{0:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y}) = p_\theta(\mathbf{x}_{0:T}|\mathbf{u}, \mathbf{z}_{\text{causal}})p(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y}) \tag{4}$$

where θ are the parameters of the reverse process of the conditional diffusion model, $p(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y}) = p(\mathbf{u})p(\mathbf{z}_{\text{causal}}|\mathbf{y})$, and $p(\mathbf{u}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The log-likelihood of the input data distribution can be obtained as follows:

$$\log p(\mathbf{x}_0, \mathbf{y}) = \log \int p(\mathbf{x}_{0:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{y}) \, d\mathbf{x}_{1:T} \, d\mathbf{u} \, d\mathbf{z}_{\text{causal}} \tag{5}$$

The joint posterior distribution $p(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})$ is intractable, so we approximate it using a variational distribution $q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})$ which can be factorized into the following conditional distributions

$$\begin{aligned} q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y}) = q_\phi(\mathbf{z}_{\text{causal}}, \mathbf{u}|\mathbf{x}_0, \mathbf{y}) \\ q(\mathbf{x}_{1:T}|\mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{x}_0) \end{aligned} \tag{6}$$

where ϕ are the parameters of the variational encoder network.

2.3. Causal Diffusion Decoder

We use a conditional DDIM decoder that takes as input the pair of latent variables $(\mathbf{z}_{\text{causal}}, \mathbf{x}_T)$ to generate the output image. We approximate the inference distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ by parameterizing the probabilistic decoder via a conditional DDIM $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z}_{\text{causal}})$. The joint distribution of the reverse generative process is defined as follows:

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{z}_{\text{causal}}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z}_{\text{causal}}) \tag{7}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z}_{\text{causal}}) = \mathcal{N}(\mathbf{x}_{t-1}|\mathbf{x}_t, \epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{causal}})) \tag{8}$$

where ϵ_θ is a noise prediction UNet [3]. By leveraging the reparameterization trick, we can optimize the following mean squared error between noise terms

$$\mathcal{L}_{\text{simple}} = \sum_{t=1}^{T} \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{causal}}) - \epsilon_t\|_2^2\right] \tag{9}$$

Algorithm 1 CausalDiffAE Training

```
Input: (image, label) pairs $(\mathbf{x}_0, y)$
Output: trained causal diffusion autoencoder $\epsilon_\theta$
repeat
    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
    $\mathbf{u} \sim q_\phi(\mathbf{u}|\mathbf{x}_0)$                          ▷ Noise encoding
    $\mathbf{z}_{\text{causal}} = \{f_i(u_i, z_{\mathbf{pa}_i}; \nu_i)\}_{i=1}^n$     ▷ Causal encoding
    $t \sim \mathcal{U}(\{1, \dots, T\})$                      ▷ Sample timestep
    $\epsilon \sim \mathcal{N}(0, I)$
    $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1 - \alpha_t}\epsilon_t$    ▷ Corrupt data to sampled time
    Take gradient step on $\nabla_\theta \mathcal{L}_{\text{CausalDiffAE}}$
until convergence
```

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1 - \alpha_t}\epsilon_t$, and T is number of diffusion steps.

2.4. Learning Objective

To ensure the causal representation is disentangled, we incorporate label information $\mathbf{y} \in \mathbb{R}^n$ as a prior in the variational objective to aid in learning semantic factors and for identifiability guarantees [8]. We define the following joint loss objective:

$$
\begin{aligned}
\mathcal{L}_{\text{CausalDiffAE}} = \mathcal{L}_{\text{simple}} \\
&+ \gamma \Big\{ \mathcal{D}_{KL}(q_\phi(\mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y}) \| p(\mathbf{z}_{\text{causal}}|\mathbf{y})) \\
&+ \mathcal{D}_{KL}(q_\phi(\mathbf{u}|\mathbf{x}_0) \| \mathcal{N}(\mathbf{0}, \mathbf{I})) \Big\}
\end{aligned}
\tag{10}
$$

where γ is a regularization hyperparameter similar to the bottleneck parameter in β-VAEs [6], and the alignment prior over latent variables is defined as the following exponential family distribution

$$
p(\mathbf{z}_{\text{causal}}|\mathbf{y}) = \prod_{i=1}^{n} p(z_i|y_i) = \prod_{i=1}^{n} \mathcal{N}(z_i; \mu_\nu(y_i), \sigma_\nu^2(y_i)\mathbf{I})
\tag{11}
$$

where μ_ν and σ_ν^2 are functions that estimate the mean and variance of the Gaussian, respectively. Intuitively, this prior ensures that the learned factors are one-to-one mapped to an indicator of the underlying ground truth factors. DiffAE requires training a latent DDIM in the latent space of the pretrained autoencoder to enable sampling of latent semantic representation. However, CausalDiffAE is formulated as a variational objective with a stochastic encoder. Thus, we can sample the representation from the defined prior directly without having to train a separate diffusion model in the latent space. The training procedure for CausalDiffAE is outlined in Algorithm 1.

Algorithm 2 CausalDiffAE Counterfactual Generation

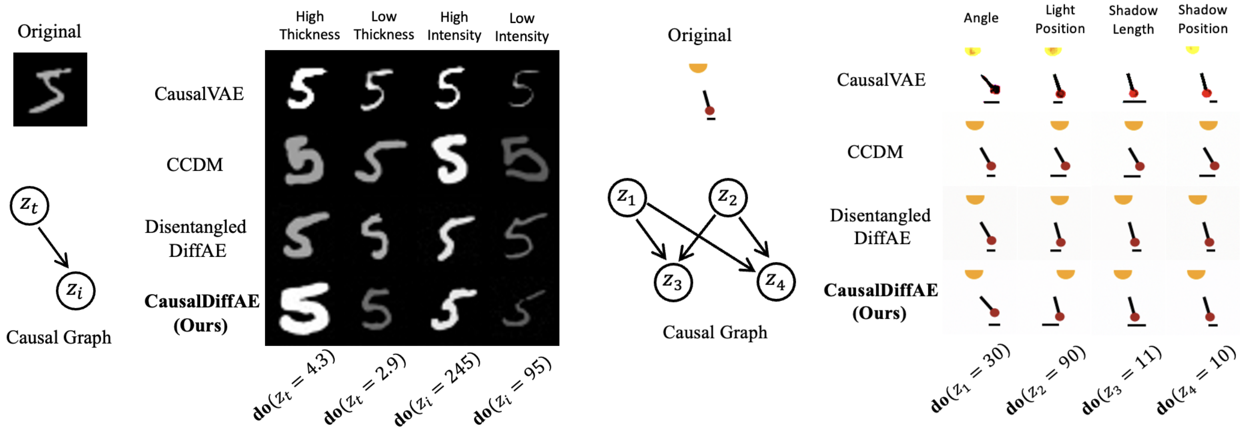
```
Input: Factual sample $\mathbf{x}_0$, intervention target set $\mathcal{I}$ with intervention values $c$, pre-trained causal diffusion autoencoder $\epsilon_\theta$
Output: Counterfactual sample $\mathbf{x}_0^{CF}$
 1: $\mathbf{u} \sim q_\phi(\mathbf{u}|\mathbf{x}_0)$                          ▷ Noise encoding
 2: for $i = 1$ to $n$ do                           ▷ in topological order
 3:     if $i \in \mathcal{I}$ then
 4:         $z_i = c_i$
 5:     else
 6:         $z_i = f_i(u_i, z_{\mathbf{pa}_i})$
 7:     end if
 8: end for
 9: $\bar{\mathbf{z}}_{\text{causal}} = \{z_1, \dots, z_n\}$           ▷ Intervened representation
10: $\mathbf{x}_T \sim \mathcal{N}(\sqrt{\alpha_T}\mathbf{x}_0, (1 - \alpha_T)\mathbf{I})$
11: for $t = T, \dots, 1$ do                             ▷ DDIM sampling
12:     $\mathbf{x}_{t-1}^{CF} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\mathbf{x}_t^{CF} - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t^{CF}, t, \bar{\mathbf{z}}_{\text{causal}})}{\sqrt{\bar{\alpha}_t}}\right)$
13:                 $+ \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t^{CF}, t, \bar{\mathbf{z}}_{\text{causal}})$
14: end for
15: return $\mathbf{x}_0^{CF}$
```

2.5. Counterfactual Generation

A fundamental property of causal models is the ability to perform interventions, facilitated by the $\mathbf{do}(\cdot)$ operator, and observe changes to a system. In generative models, this enables the sampling of counterfactual data. Given a pretrained CausalDiffAE, we can controllably manipulate any factor of variation, propagate the causal effects to descendants, and perform reverse diffusion to sample from the counterfactual distribution. Algorithm 2 shows the process of generating counterfactuals from a trained CausalDiffAE, where $\mathbf{x}_0$ refers to the factual observation and $\mathbf{x}_0^{CF}$ refers to the generated counterfactual sample. We utilize the DDIM sampling algorithm to ensure the stochastic noise $\mathbf{x}_T$ is a deterministic encoding to enable semantic manipulations. In lines 12-13, we use the DDIM non-Markovian deterministic generative process to generate counterfactual instances.

3. Experiments

We investigate the generative capability of the proposed CausalDiffAE model. We compare our model with CausalVAE [25], class-conditional diffusion model (CCDM) [11], and diffusion autoencoder (DiffAE) [15]. We evaluate disentanglement of the learned latent space and the quality of generated counterfactuals on visual datasets. We run experiments on MorphoMNIST [2, 13], where thickness causes intensity, and Pendulum [25], where pendulum angle and light position cause shadow length and position. We note that since the latent space of DiffAE does not allow precise control over variables, we modify DiffAE to align the representation to be disentangled (DisentangledDiffAE).



(a) MorphoMNIST results

(b) Pendulum results

Figure 2. Counterfactual trajectories generated for (a) MorphoMNIST and (b) Pendulum datasets with fully supervised model

Table 1. DCI disentanglement

Dataset	Model	DCI ↑
MorphoMNIST	CausalVAE	0.7838
	DiffAE	0.3578
	CausalDiffAE (Ours)	**0.9934**
Pendulum	CausalVAE	0.8850
	DiffAE	0.3525
	CausalDiffAE (Ours)	**0.9995**

3.1. Disentanglement

We are often interested in utilizing learned representations for downstream tasks. Thus, we evaluate to what extent a diffusion-based decoder helps to disentangle the latent factors of variation compared to standard VAE-based methods. We use the DCI disentanglement metric [4], which measures the level of one-to-one correspondence to the ground truth factors. A high DCI score also suggests the effectiveness of controllable generation. In the context of a causal representation, this means that we can intervene on latent codes in an isolated fashion without any entanglements (i.e., two different factors are encoded in separate latent codes). We find that CausalDiffAE can disentangle the factors to a higher degree than CausalVAE and DiffAE, as shown in Table 1. We evaluate representation disentanglement of the original DiffAE model [15] since it learns an arbitrary representation.

3.2. Qualitative Evaluation

We show qualitative counterfactual generation results of CausalDiffAE compared to other baseline models. The distinction between a conditional and causal model lies in the difference between conditioning and intervening. When we condition, we narrow down the scope of possibilities. When we intervene, we fix the value of a variable and compute downstream causal effects. For example, increasing the shadow length of the pendulum system with other factors unchanged produces a counterfactual that does not exist in the training distribution. Sampling from a conditional model will change the angle of the pendulum to generate an image with a longer shadow, as shown in Figure 2b. Semantic manipulations in DiffAE [15] are done as a post-processing step by training a linear classifier and performing linear interpolation in the latent space. However, such a method is not compatible with interventions. For a fair comparison in counterfactual generation, we extend the DiffAE to learn disentangled factors through an alignment prior, similar to CausalDiffAE. We observe that CCDM and DisentangledDiffAE generate images that are not consistent with the causal model upon intervention. For example, in MorphoMNIST, intensity does not change after intervening on thickness. For CausalDiffAE, intervening on a causal factor changes causal variables downstream and intervening on a child node keeps the parents unchanged, as shown in Figures 2a and 2b.

4. Conclusion

In this work, we propose causal diffusion autoencoders, a diffusion-based causal representation learning framework for counterfactual generation. We propose a causal encoding and learn causal mechanisms among variables via neural networks. We formulate a variational learning objective to learn disentangled causal representations, which we use to condition the reverse diffusion process. To enable the generation of counterfactuals, we propose a DDIM-based counterfactual generation algorithm subject to do-interventions. Experiments show that diffusion models are a promising class of generative models for high-quality controllable counterfactual generation. Future work includes exploring reduced supervision scenarios and text-controlled counterfactual generation.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1

[2] Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20(178), 2019. 3

[3] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 2, 1, 5

[4] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. 4

[5] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT Press, 2016. 1

[6] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 3

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 5

[8] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020. 3, 5

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 1

[10] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, 2022. 1

[11] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 1, 3, 5

[12] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. DiffuseVAE: Efficient, controllable and high-fidelity generation from low-dimensional latents. *Transactions on Machine Learning Research*, 2022. 1

[13] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, 2020. 3

[14] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. 1

[15] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 4, 5

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 1

[17] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[20] Pedro Sanchez, Jeremy P. Voisey, Tian Xia, Hannah I. Watson, Alison Q. ONeil, and Sotirios A. Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 2022. 1

[21] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109:612–634, 2021. 1, 3

[22] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015. 1

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 2

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 3

[25] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3, 5

Appendices

A. Background

A.1. Structural Causal Model

A structural causal model (SCM) [14] is formally defined by a triple $\mathcal{M} = \langle Z, U, F \rangle$, where Z is the set of n endogenous variables, U is a set of n exogenous independent noise variables, and F is a collection of n structural equations of the form:

$$z_j := f_j(\mathbf{pa}_j, u_j), j = 1, \dots, n \tag{12}$$

where $\mathbf{pa}_j$ are called parents or direct causes of z_j and the exogenous noise u_j ensures to represent a general conditional distribution $P(z_j|\mathbf{pa}_j)$. An SCM where the exogenous noise variables are jointly independent (no hidden confounders) is known as a Markovian model, which is the setting we assume for the purposes of this work. In this work, we assume the additive noise model, $z_j := f_j(\mathbf{pa}_j) + u_j$ for $j = 1, \dots, n$, where f_j is a deterministic function and u_j's are mutually independent noise variables with strictly positive densities. A (hard) intervention on a causal variable z_j is facilitated by the $\mathbf{do}(\cdot)$ operator, which fixes the value of the variable to some constant c (i.e., $\mathbf{do}(z_j = c)$).

A.2. Diffusion Probabilistic Models

Diffusion Probabilistic Models (DPMs) [7, 11] have shown impressive results in image generation tasks, even beating out GANs in many cases [3]. The idea of the diffusion model is to define a Markov chain of diffusion steps to slowly destroy the structure in a data distribution through a forward diffusion process by adding noise [7] and learn a reverse diffusion process that restores the structure of the data. Some proposed methods, such as DDIM [23], break the Markov assumption to speed up the sampling in the diffusion process by carrying out a deterministic encoding of the noise.

Forward Diffusion. Given some input data sampled from a distribution $\mathbf{x}_0 \sim q(\mathbf{x})$, the forward diffusion process is defined by adding small amounts of Gaussian noise to the sample in T steps thereby producing noisy samples $\mathbf{x}_1, \dots, \mathbf{x}_T$. The distribution of the noisy sample at time step t is defined as a conditional distribution as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I) \tag{13}$$

where $\beta_t \in (0, 1)$ is a variance parameter that controls the step size of noise. As $t \to \infty$, the input sample $\mathbf{x}_0$ loses its distinguishable features. In the end, when $t = T$, $\mathbf{x}_T$ follows an isotropic Gaussian. From Eq (13), we can then define a closed-form tractable posterior over all time steps factorized as follows:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{14}$$

Now, $\mathbf{x}_t$ can be sampled at any arbitrary time step t using the reparameterization trick. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I) \tag{15}$$

Reverse Diffusion. In the reverse process, to sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, the goal is to recreate the true sample $\mathbf{x}_0$ from a Gaussian noise input $\mathbf{x}_T \sim \mathcal{N}(0, I)$. Unlike the forward diffusion, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is not analytically tractable and thus requires learning a model p_θ to approximate the conditional distributions as follows:

$$\begin{gathered} p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \\ p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \end{gathered} \tag{16}$$

where μ and Σ are learned via neural networks. It turns out that conditioning on the input $\mathbf{x}_0$ yields a tractable reverse conditional probability

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t I) \tag{17}$$

The overall learning objective of diffusion probabilistic models is to maximize the following variational lower bound

$$\begin{aligned} \log q(\mathbf{x}_0) \geq \mathbb{E}_{q(\mathbf{x}_0)}[&\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \\ &- \sum_{t=2}^{T} \mathcal{D}_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \\ &- \mathcal{D}_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)||q(\mathbf{x}_t))] \end{aligned} \tag{18}$$

Equivalently, for Gaussian diffusion, we can simplify this objective via reparameterization and minimize the following mean squared error loss

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right] \tag{19}$$

where ϵ_t is the noise that takes an analytical form via a reparameterization from $\mathbf{x}_0$, as shown in [7].

B. Derivation of ELBO

Given a high-dimensional input image $\mathbf{x}_0$, an auxiliary weak supervision signal $\mathbf{y}$, a latent noise encoding $\mathbf{u}$, latent representation $\mathbf{z}_{\text{causal}}$, and a sequence of T latent representations $\mathbf{x}_{1:T}$ learned by the diffusion model, the CausalDiffAE generative process can be factorized as follows:

$$p(\mathbf{x}_{0:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y}) = p_\theta(\mathbf{x}_{0:T}|\mathbf{u}, \mathbf{z}_{\text{causal}})p(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y}) \tag{20}$$

where θ are the parameters of the reverse process of the conditional diffusion model. The log-likelihood of the input data distribution can be obtained as follows:

$$\log p(\mathbf{x}_0, \mathbf{y}) = \log \int p(\mathbf{x}_{0:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{y}) \; d\mathbf{x}_{1:T} \; d\mathbf{u} \; d\mathbf{z}_{\text{causal}} \tag{21}$$

The joint posterior distribution $p(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})$ is intractable, so we approximate it using a variational distribution $q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})$ which can be factorized into the following conditional distributions

$$q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y}) = q_\phi(\mathbf{z}_{\text{causal}}, \mathbf{u}|\mathbf{x}_0, \mathbf{y})q(\mathbf{x}_{1:T}|\mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{x}_0) \tag{22}$$

where ϕ are the parameters of the variational encoder network. Since the likelihood of the data is intractable, we can approximate it by maximizing the following evidence lower bound (ELBO):

$$\begin{align}
\log p(\mathbf{x}_0, \mathbf{y}) &\geq \mathbb{E}_{q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})}\left[\log \frac{p(\mathbf{x}_{0:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{y})}{q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})}\right] \tag{23}\\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})}\left[\log \frac{p(\mathbf{u})p(\mathbf{z}_{\text{causal}}|\mathbf{y})p_\theta(\mathbf{x}_{0:T}|\mathbf{u}, \mathbf{z}_{\text{causal}})}{q_\phi(\mathbf{z}_{\text{causal}}, \mathbf{u}|\mathbf{x}_0, \mathbf{y})q(\mathbf{x}_{1:T}|\mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{x}_0)}\right] \tag{24}\\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})}\left[\log \frac{p(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y})}{q_\phi(\mathbf{z}_{\text{causal}}, \mathbf{u}|\mathbf{x}_0, \mathbf{y})} + \log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{u}, \mathbf{z}_{\text{causal}})}{q(\mathbf{x}_{1:T}|\mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{x}_0)}\right] \tag{25}\\
&= \mathbb{E}_{q(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})}\left[\log \frac{p(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y})}{q_\phi(\mathbf{z}_{\text{causal}}, \mathbf{u}|\mathbf{x}_0, \mathbf{y})}\right] + \mathbb{E}_{q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{u}, \mathbf{z}_{\text{causal}})}{q(\mathbf{x}_{1:T}|\mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{x}_0)}\right] \tag{26}\\
&= \underbrace{\mathbb{E}_{q(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})}\left[\mathbb{E}_{q(\mathbf{x}_{1:T}, \mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0)}\left[\frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{u}, \mathbf{z}_{\text{causal}})}{q(\mathbf{x}_{1:T}|\mathbf{u}, \mathbf{z}_{\text{causal}}, \mathbf{x}_0)}\right]\right]}_{\text{Representation-conditioned DDPM Loss}} - \underbrace{\mathcal{D}_{KL}(q_\phi(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})\|p(\mathbf{u}, \mathbf{z}_{\text{causal}}|\mathbf{y}))}_{\text{Joint Latent Posterior Loss}} \tag{27}\\
& \tag{28}
\end{align}$$

In the learning process, we minimize the negative of the derived ELBO. We simplify this objective by using the ϵ_θ parameterization to optimize the representation-conditioned DDPM loss. Further, since $\mathbf{u}$ and $\mathbf{z}_{\text{causal}}$ are one-to-one mapped, we can split the joint conditional distribution into separate conditional distributions. Thus, we have the following final objective for CausalDiffAE:

$$\mathcal{L}_{\text{CausalDiffAE}} = \sum_{t=1}^{T} \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{causal}}) - \epsilon_t\|_2^2\right] + \gamma\Big\{\mathcal{D}_{KL}(q_\phi(\mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})\|p(\mathbf{z}_{\text{causal}}|\mathbf{y})) + \mathcal{D}_{KL}(q_\phi(\mathbf{u}|\mathbf{x}_0)\|\mathcal{N}(\mathbf{0}, \mathbf{I}))\Big\} \tag{29}$$

C. Connection to Score-based Generative Models

Diffusion models can also be represented as stochastic differential equations (SDEs) [24] to model continuous-time perturbations. Specifically, the forward diffusion process can be modeled as the solution to an SDE on a continuous-time domain $t \in [0, T]$ with stochastic trajectories:

$$d\mathbf{x} = f(\mathbf{x}, t)\; dt + g(t)\; dw \tag{30}$$

where w is the standard Weiner process, f is a vector-valued function known as the drift coefficient of $\mathbf{x}(t)$ and g is a scalar function known as the diffusion coefficient of $\mathbf{x}(t)$. The drift and diffusion coefficients can be considered as the mean and variance of the noise perturbations in the diffusion process, respectively. The reverse diffusion process can be modeled by the solution to the reverse-time SDE of Eq. (30), which can be derived analytically as:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\nabla_x \log p_t(\mathbf{x})]\; dt + g(t)\; d\bar{w} \tag{31}$$

where $\bar{w}$ is the standard Weiner process in reverse time and $\nabla_x \log p_t(\mathbf{x})$ is the score of the data distribution at timestep t. Once we know the score of the marginal distribution for all timesteps t, we can derive the reverse diffusion process from Eq. (31).

Song et al [24] showed that the denoising diffusion probabilistic model (DDPM) is a discretization of the following Variance Preserving SDE (VP-SDE)

$$d\mathbf{x} = \frac{1}{2}\beta(t)\mathbf{x}\; dt + \sqrt{\beta(t)}\; dw \tag{32}$$

Thus, learning a noise prediction network ϵ_θ and minimizing MSE in diffusion probabilistic models is equivalent to approximating the score of the data distribution in the SDE formulation. From a score-based perspective, we aim to minimize the following conditional denoising score-matching form of our objective

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}_{\text{causal}}|\mathbf{x}_0,\mathbf{y})}\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\Big[\log p(\mathbf{u}) + p(\mathbf{z}_{\text{causal}}|\mathbf{y}) \\ &- \log q_\phi(\mathbf{u}|\mathbf{x}_0) - \log q_\phi(\mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y}) \\ &+ \lambda(t)\|s_\theta(\mathbf{x}_t, \mathbf{z}_{\text{causal}}, t) - \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_0)\|\Big] \end{aligned} \tag{33}$$

where s_θ approximates the score of the data distribution conditioned on $\mathbf{x}_0$ and $\lambda(t)$ is a positive weighing function. The ideal for modeling natural phenomena in the world is by using differential equations to model the physical mechanisms [21]. In the SDE formulation, the causal variables are used to denoise the high-dimensional data, which is modeled as a reverse-time stochastic trajectory. We can interpret this idea as modeling the dynamics of high-dimensional systems by incorporating causal information. As opposed to simply learning an arbitrary latent representation, a disentangled causal representation encodes the causal information that the denoising process can use to reconstruct *causally relevant* features in high-dimensional data.

D. Experiment Details

D.1. Dataset Details

MorphoMNIST. The MorphoMNIST dataset [2] is produced by applying morphological transformations on the original MNIST handwritten digit dataset. The digits can be described by measurable shape attributes such as stroke thickness, stroke length, width, height, and slant of digit. Pawlowski et al [13] impose a 3-variable SCM to generate the morphological transformations, where stroke thickness is a cause of the brightness of each digit. That is, thicker digits are often brighter, whereas thinner digits are dimmer. The data-generating process is as follows

$$\begin{aligned} t &= f_T(u_T) = 0.5 + u_T\,, & u_T &\sim \Gamma(10, 5)\,, \\ i &= f_I(u_I; t) = 191 \cdot \sigma(0.5 \cdot u_I + 2 \cdot t - 5) + 64\,, & u_I &\sim \mathcal{N}(0, 1)\,, \\ x &= f_X(u_X; i, t) = \text{SetIntensity}(\text{SetThickness}(u_X; t)\,; i)\,, & u_X &\sim \text{MNIST}\,, \end{aligned} \tag{34}$$

where x is the resulting image, u is the exogenous noise for each variable, and $\sigma(\cdot)$ is the logistic sigmoid.

Pendulum. The Pendulum dataset [25] consists of a set of 7K images with resolution $96 \times 96 \times 4$ describing a physical system of a pendulum and light source that cause the length and position of a shadow. The causal variables of interest are the angle of the pendulum, the position of the light source, the length of the shadow, and the position of the shadow. The data generating process is as follows:

$$
\begin{aligned}
&y_1 \sim U(-45, 45); \qquad \theta = y_1 * \frac{\pi}{200}; \qquad x = 10 + 9.5 \sin\theta \\
&y_2 \sim U(60, 145); \qquad \phi = y_2 * \frac{\pi}{200}; \qquad y = 10 - 9.5 \cos\theta \\
&y_3 = \max(3, \left|9.5 \frac{\cos\theta}{\tan\phi} + 9.5 \sin\theta\right|) \\
&y_4 = \frac{-11 + 4.75 \cos\theta}{\tan\phi} + (10 + 4.75 \sin\theta)
\end{aligned}
$$

D.2. Additional Experiments

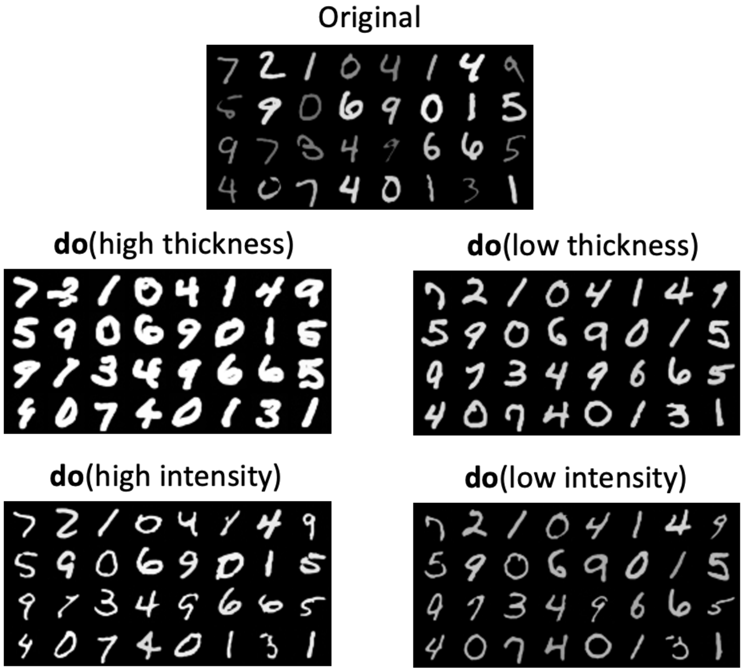


Figure 3. CausalDiffAE generated counterfactuals (MorphoMNIST)

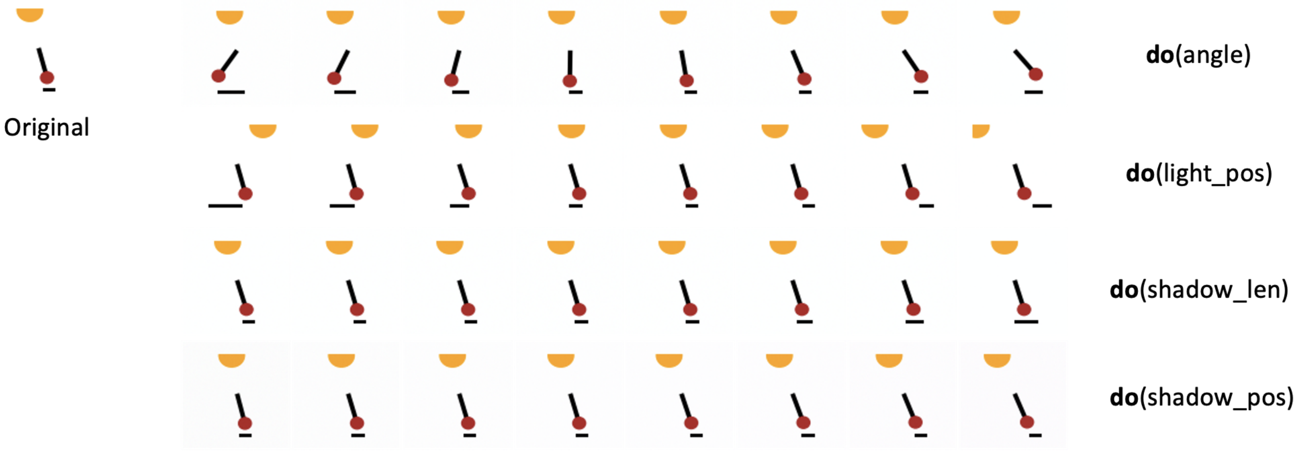


Figure 4. latent traversals in the normalized range $(-1, 1)$ generated counterfactuals (Pendulum)

Table 2. Implementation details of CausalDiffAE

Parameter	MorphoMNIST	Pendulum
Batch size	768	128
Base channels	128	128
Channel multipliers	$[1, 2, 2]$	$[1, 2, 4, 8]$
Training set	60K	5K
Image resolution	$28 \times 28 \times 1$	$96 \times 96 \times 4$
Num causal variables	2	4
z_{causal} size	512	512
β scheduler	Linear	Linear
Learning rate	10^{-4}	10^{-4}
Optimizer	Adam	Adam
Diffusion steps	1000	1000
Iterations	10K	50K
Diffusion loss	MSE	MSE
Sampling	DDIM	DDIM
Bottleneck γ	1.0	0.1

D.3. Implementation Details

We use the same network architectures used in other works based on diffusion models [3, 7, 11]. We set the causal latent variable size to 512 to ensure a large enough capacity to capture causally relevant information. The representation-conditioned noise predictor is parameterized by a UNet with the attention mechanism. Similar to [7], we use a linear noise scheduling for the variance parameter β between $\beta_1 = 10^{-4}$ and $\beta_2 = 0.02$ during training. We also use different bottleneck parameters γ for each dataset. Note that we start at $\gamma = 0$ and linearly increase γ throughout training.

D.4. Baselines

CausalVAE. [25] proposed CausalVAE, a framework for causal representation learning from weak label supervision. CausalVAE assumes the underlying factors of variation are related by a linear structural causal model and utilizes a causal masking layer to transform noise encodings into causal variables. The latent space is structured by a label prior that regularizes the posterior and ensures the identifiability of the representation according to [8].

Class-conditional Diffusion Model (CCDM). The class-conditional diffusion model [3] is a simple conditional generative model that conditions the reverse diffusion on predefined class labels. Thus, one can generate images conditioned on specified discrete or continuous labels.

Diffusion Autoencoder (DiffAE). [15] proposed diffusion autoencoders (DiffAE), a representation learning objective in diffusion models to learn manipulable and semantically meaningful latent codes. DiffAE learns a compact semantic subcode that is then used in the reverse diffusion process for image decoding. However, this approach learns an arbitrary representation in an unsupervised fashion and does not disentangle the latent space. Manipulations are performed using a post-hoc classifier for linear interpolation. Thus, the learned representation would not be ideal to perform causal interventions. For a fair comparison, we modify the objective to disentangle the latent space by incorporating label information in a prior to regularize the posterior. We call this extension **DisentangledDiffAE**.