Diffusion Based Causal Representation Learning

Amir Mohammad Karimi Mamaghan [1] **Andrea Dittadi** [2 3] **Stefan Bauer** [2 4] **Francesco Quinzan** [5]

Abstract

Causal reasoning can be considered a cornerstone of intelligent systems. Having access to an underlying causal graph comes with the promise of cause-effect estimation and the identification of efficient and safe interventions. However, depending on the application and the complexity of the system one causal graph might be insufficient and even the variables of interest and levels of abstractions might change. This is incompatible with currently deployed generative models including popular VAE approaches which provide only representations from a point estimate. In this work, we study recently introduced diffusion-based representations which offer access to infinite dimensional latent codes which encode different levels of information in the latent code. In a first proof of principle, we investigate the use of a single point of these infinite dimensional codes for causal representation learning and demonstrate experimentally that this approach performs comparably well in identifying the causal structure and causal variables.

1. Introduction

Causal representation learning consists of uncovering a system's latent causal factors and their relationships, from observed low-level data. Causal representation learning finds applicability in domains such as autonomous driving (Schölkopf et al., 2021), robotics (Hellström, 2021), healthcare (Anwar et al., 2014), climate studies (Runge et al., 2019), epidemiology (Hernán et al., 2000; Robins et al., 2000), and finance (Hiemstra & Jones, 1994). In these tasks, the underlying causal variables are often unknown, and we only have access to low-level representations.

There has been a growing interest in leveraging the power of generative models in order to learn causal representations with specific properties such as disentanglement. Variational Autoencoders (VAE) are one of the most common frameworks used in the literature (Locatello et al., 2020). Recently, diffusion models have proven to be highly effective in modeling the underlying data distribution (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021), and they have demonstrated remarkable success across several domains (Dhariwal & Nichol, 2021; Ramesh et al., 2022; Saharia et al., 2022; Ho et al., 2022; Höppe et al., 2022; Abstreiter et al., 2022). However, diffusion models have not yet been employed for causal representation learning, indicating that their potential has yet to be explored in this context.

Causal representation learning is a challenging problem, because learning causal mechanisms from observational data is impossible. There has been an ongoing effort to study sets of assumptions that ensure the identifiability of causal variables and their relationships (Yang et al., 2020; Schölkopf et al., 2021; Liu et al., 2022; Subramanian et al., 2022; Brehmer et al., 2022). These approaches consider the availability of additional information or interventional data, but they differ in how they incorporate and utilize such information. Interestingly, Brehmer et al. (2022) consider a weak form of supervision in which we have access to a data pair, corresponding to the state of the system before and after a random, unknown intervention. Brehmer et al. (2022) prove that, in this weakly-supervised setting, the structure and the causal variables are identifiable up to a relabeling and element-wise reparameterization.

Diffusion-based representations have the appealing property of infinite-dimensional latent codes, which allow manual control of the level of detail encoded in the representation (Abstreiter et al., 2022). This is of particular importance for causal representation learning since for different downstream tasks different causal graphs with possibly different causal abstractions (variables) might be needed from the same input. As opposed to other generative models, such as generative adversarial networks or VAEs, diffusion-based representations encode different levels of information in the latent code at different time steps.

[1] Department of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden [2] Helmholtz AI, Munich, Germany [3] Max Planck Institute for Intelligent Systems, Tübingen, Germany [4] TU Munich, Munich, Germany [5] Department of Computer Science, University of Oxford, Oxford, England. Correspondence to: Amir Mohammad Karimi Mamaghan <amkm@kth.se>.

Accepted to *ICML workshop on Structured Probabilistic Inference & Generative Modeling*

In this work, we study the connection between diffusion-based models and causal structure learning. In particular, our contributions are the following:

- We incorporate diffusion models for causal representation learning and study and test the connection between the learned representations of diffusion score matching with causal variables.
- We derive the Evidence Lower Bound (ELBO) for learning both the causal variables and the underlying mechanisms using conditional diffusion models and weak supervision.
- We empirically illustrate that the noise and diffusion-based representations contain equivalent information about the underlying causal variables and causal mechanisms.

2. Related Work

Several previous methods rely on additional knowledge on the data generating process, such as knowledge of the causal graph or labels for the high-level causal variables. CausalGAN (Kocaoglu et al., 2017) requires the structure of the underlying causal graph to be known. Yang et al. (2020); Liu et al. (2022) assume a linear structural equation model, and they require additional information associated with the true causal concepts as supervising signals. Similar to Yang et al. (2020), Komanduri et al. (2022) assume the availability of supplementary supervision labels, but without requiring mutual independence among factors. Von Kügelgen et al. (2021) investigates self-supervised causal representation learning by utilizing a known, but non-trivial, causal graph between content and style factors. Subramanian et al. (2022) applies Bayesian structure learning in the latent space and relies on having interventional samples. Brehmer et al. (2022) proposes ILCM, a VAE-based model that identifies the causal structure in a weakly supervised setting. A unifying framework between independent component analysis and VAEs underlying many approaches is presented in Khemakhem et al. (2020) and for an overview of causal representation learning we refer to Schölkopf et al. (2021).

3. Background

3.1. Structural Causal Model

We describe the causal relationships between latent variables as a Structural Causal Model (SCM). An SCM consists of a set of equations associated with a directed graph $\mathcal{G}$ where the nodes $z_1, ..., z_n \in \mathcal{Z}$ are high-level latent causal variables. Each variable z_i is defined by a structural equation of the form

$$z_i = f_i(z_{pa(i)}, \epsilon_i).$$

Here, $z_{pa(i)} \subseteq \mathcal{Z}$ is a set of causal variables, and ϵ_i is exogenous noise. The variables $z_{pa(i)}$ are commonly referred to as the causal parents of z_i. Directed edges in $\mathcal{G}$ capture cause-effect relationships, i.e., there is an edge $z_j \to z_i$ if $z_j \in z_{pa(i)}$. An intervention set I changes the causal mechanisms so that they no longer depend on the parents of the causal variable being intervened on, i.e. $z_i = f_i'(\epsilon_i),\ i \in I$.

3.2. Diffusion Models

Diffusion models are a class of generative models that comprise two processes: a forward process and a backward process. The forward process is defined by a stochastic differential equation (SDE) across a continuous time domain $t \in [0, 1]$, aiming to transform the data distribution to a known prior distribution, typically a standard Gaussian. Given $x_0 \in \mathbb{R}^d$ sampled from a data distribution $p(x)$, the forward process constructs a trajectory $(x_t)_{t\in[0,1]}$ across the time domain. We utilize the Variance Exploding SDE (Song et al., 2021) for the forward process, defined as:

$$dx = f(x,t) + g(t)dw := \sqrt{\frac{d[\sigma^2(t)]}{dt}}dw,$$

where w is the standard Wiener process and $\sigma^2(t)$ is the noise variance of the diffusion process at time t. The backward process is also formulated as an SDE:

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

where $\bar{w}$ is the standard Wiener process in reverse time.

The backward process requires the score function $\nabla_x \log p_t(x)$, which is not available. Vincent (2011) addresses this issue by proposing Denoising Score Matching. A score function s_θ is trained by minimizing the loss function:

$$\begin{aligned}\mathcal{L}(\theta) =& \mathbb{E}_t\Bigg[\lambda(t)\mathbb{E}_{x_0}\mathbb{E}_{p(x_t|x_0)}\Big[||s_\theta(x_t,t) \\ &- \nabla_{x_t}\log p_t(x_t|x_0)||^2\Big]\Bigg],\end{aligned}$$

where the conditional distribution of x_t given x_0 is $p_t(x_t|x_0) = \mathcal{N}(x_t; x_0, [\sigma^2(t) - \sigma^2(0)]\mathbf{I})$, and $\lambda(t)$ is a positive weighting function.

We can modify Denoising Score Matching, to perform representation learning while training the score function. Following Abstreiter et al. (2022), the objective then becomes:

$$\begin{aligned}\mathcal{L}(\theta,\phi) =& \mathbb{E}_t\Bigg[\lambda(t)\mathbb{E}_{x_0}\mathbb{E}_{p(x_t|x_0)}\Big[||s_\theta(x_t, E_\phi(x_0), t) \\ &- \nabla_{x_t}\log p_t(x_t|x_0)||^2\Big]\Bigg], \qquad (1)\end{aligned}$$

where additional information about the data $E_\phi(x_0)$ is provided to the diffusion model through a learned encoder with parameters ϕ. With this formulation, the encoder learns meaningful representations of the data (Abstreiter et al., 2022; Mittal et al., 2022). Later on, we will develop this objective function in the ELBO, and we will use it in our subsequent analysis.

The key difference between the other generative models and diffusion-based representations is that the first two are only concerned with one final state that generates samples from a desired distribution while in the latter, different levels of information are encoded along an infinite-dimensional code, i.e., the encoder will be conditioned on time t: $E_\phi(x_0, t)$. Different steps along this latent code contain different levels of detail and in this work, we start by investigating a single point of this latent code and study the benefits and implications of this formulation for causal representation learning.

4. Diffusion-based Causal Representation Learning

4.1. Identifiability

It is well-known that it is impossible to learn causal variables from low-level observational data only (Schölkopf et al., 2021). Causal representation learning requires datasets that comprise both *observational* and *interventional* data. We follow the weakly-supervised framework by Brehmer et al. (2022). We consider a weak supervision in which the dataset is in a paired format with each pair representing the system before and after a random, unknown, and atomic intervention. They demonstrate that under this weakly supervised setting, it is possible to identify the causal variables and causal mechanisms up to a permutation and elementwise reparameterization of the variables. For the assumptions we use in the data generation process, we refer to Brehmer et al. (2022).

4.2. Problem Setup

We consider a system that is described by an unknown underlying SCM among latent variables and we have access to data pairs $(x, \tilde{x}) \sim p(x, \tilde{x})$. The objective is to learn an SCM that accurately represents the true underlying SCM associated with the given data, up to a permutation and elementwise reparameterizations of causal variables. We consider causal mechanisms and therefore, solution functions in the SCM to be diffeomorphic. Under this assumption, the noise encodings hold the same information as the causal variables and mechanisms. Therefore, We try to learn noise pairs $(e, \tilde{e})$ -that are identical to the SCM noise variables- from the input $(x, \tilde{x})$ and map each noise pair to the corresponding latent pair $(z, \tilde{z})$ with learned solution functions. We train the framework by minimizing β-VAE loss (Higgins et al., 2017).

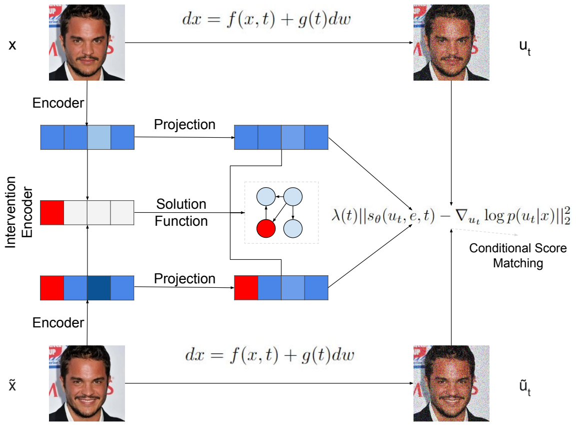

Figure 1. The framework overview. A diffusion model receives the input pair and is conditioned on noise encodings, while intervention targets are obtained through the intervention encoder. Causal variables are derived by applying solution functions to the noise encodings and intervention targets.

4.3. Proposed Method

We incorporate several elements of the Implicit Latent Causal Model (ILCM) introduced by Brehmer et al. (2022) into our framework. We consider latent variables to be noise encodings $(e, \tilde{e})$. The inputs are mapped to the noise encodings through an encoder $q(e|x)$, and we formulate the mapping from noise encodings to the input data by a diffusion model that is conditioned on noise encodings. The causal structure is represented with solution functions $s(e)$. After the training, the framework contains information about the underlying causal structure and causal variables and therefore, can be used in different related downstream tasks.

Figure 1 provides a visual representation of the framework's architecture. In summary, our framework consists of:

- A noise encoder $q(e|x)$ implemented as a VAE;
- An intervention encoder $q(I|x, \tilde{x})$;
- A projection phase on the noise encodings $(e, \tilde{e})$ in which for the noise components e_i that are not intervened upon, $i \notin I$, the pre-intervention and post-intervention noise encodings will be equal, $e_i = \tilde{e}_i$. This prevents solution functions from deviating from the weakly supervised structure. We write the combination of the noise encoder and the projection phase as $(e, \tilde{e}) \sim q(e, \tilde{e}|x, \tilde{x}, I)$, and refer it to as the noise encoding module;
- A conditional diffusion model $q(u|x)$ where u is the collection of the trajectory of latent variables in the diffusion model $(u_t)_{t\in[0,1]}$ corresponding to the input x. We consider noise encodings to be diffusion-

based representations. In other words, we consider the encoder in Eq. 1 to be the noise encoding module $q(e, \tilde{e}|x, \tilde{x}, I)$.

- Solution functions $s_i(e_i; e_i)$ for $i = 1, ..., n$ where n is the number of causal variables. They are defined as invertible affine transformations with parameters learned with neural networks;
- The prior $p(e, \tilde{e}, I)$ which encodes the causal structure and is defined as

  $$p(e, \tilde{e}, I) = p(I)p(e)p(\tilde{e}|e, I),$$

  where $p(I)$ and $p(e)$ are prior distributions we choose as uniform categorical and standard Gaussian, respectively, and

  $$p(\tilde{e}|e, I) = \prod_{i \notin I} \delta(\tilde{e}_i - e_i) \prod_{i \in I} p(\tilde{e}_i|e), \qquad (2)$$

  where $p(\tilde{e}_i|e)$ is parameterized with a conditional normalizing flow containing the solution functions $z_i = s_i(\tilde{e}_i; e_i)$ and a prior over z_i which we choose as standard Gaussian.

For more details, see Appendix B.

Putting everything together, the lower bound on the paired data distribution will be:

$$\begin{aligned}\log p(x, \tilde{x}) \geq\ & \mathbb{E}_{p(x,\tilde{x})}\mathbb{E}_{q(I|x,\tilde{x})}\mathbb{E}_{q(e,\tilde{e}|x,\tilde{x},I)}\mathbb{E}_{t\sim U(0,1)} \\ & \mathbb{E}_{q(u_t|x)}\mathbb{E}_{q(\tilde{u}_t|\tilde{x})}\Bigg[\log p(I) + \log p(e) \\ & + \log p(\tilde{e}|e, I) - \log q(I|x, \tilde{x}) - \log q(e, \tilde{e}|x, \tilde{x}, I) \\ & + \lambda(t)||s_\theta(u_t, e, t) - \nabla_{u_t} \log p(u_t|x)||_2^2 \\ & + \lambda(t)||s_\theta(\tilde{u}_t, \tilde{e}, t) - \nabla_{\tilde{u}_t} \log p(\tilde{u}_t|\tilde{x})||_2^2\Bigg],\end{aligned}$$

where $\lambda(t)$ is a positive weighting function. For more details, see Appendix A.

We train the model by minimizing β-VAE loss:

$$\begin{aligned}\mathcal{L}_{model} =\ & \mathbb{E}_{p(x,\tilde{x})}\mathbb{E}_{q(I|x,\tilde{x})}\mathbb{E}_{q(e,\tilde{e}|x,\tilde{x},I)} \\ & \mathbb{E}_{t\sim U(0,1)}\mathbb{E}_{q(u_t|x)}\mathbb{E}_{q(\tilde{u}_t|\tilde{x})} \\ & \Bigg[\lambda(t)||s_\theta(u_t, e, t) - \nabla_{u_t} \log p(u_t|x)||_2^2 \\ & + \lambda(t)||s_\theta(\tilde{u}_t, \tilde{e}, t) - \nabla_{\tilde{u}_t} \log p(\tilde{u}_t|\tilde{x})||_2^2 \\ & + \beta\Big[\log p(I) + \log p(e) + \log p(\tilde{e}|e, I) \\ & - \log q(I|x, \tilde{x}) - \log q(e, \tilde{e}|x, \tilde{x}, I)\Big]\Bigg],\end{aligned}$$

where $\lambda(t)$ is a positive weighting function. For more details, see Appendix A.

To prevent a collapse of the latent space to a lower-dimensional subspace, we add the negative entropy of the batch-aggregate intervention posterior ($q_I^{batch}(I) = \mathbb{E}_{x,\tilde{x}\in batch}[q(I|x, \tilde{x}])$ as a regularization term:

$$\mathcal{L}_{entropy} = \mathbb{E}_{batches}\Big[-\sum_I q_I^{batch}(I) \log q_I^{batch}(I).\Big]$$

5. Experiments

Here we analyze the performance of the proposed model on synthetic data. We employ our method for the task of causal discovery and apply ENCO (Lippe et al., 2021), a continuous optimization structure learning method that leverages observational and interventional data, on top of the proposed method to infer the underlying causal graph. Furthermore, we evaluate how good the learned latent variables are by employing DCI scores (Eastwood & Williams, 2018).

Data Generation In order to generate latent variables, we adopted random graphs where each edge in a fixed topological order is sampled from a Bernoulli distribution with a probability of 0.5. We utilize an SCM with 5, 10, and 15 variables. For more details, see Appendix B.

Baselines We consider ILCM as our main baseline. To the best of our knowledge, there aren't any other methods that consider the same weakly-supervised assumptions. We also compare the results with a modification of disentanglement VAE (Locatello et al., 2020) for the weakly supervised setting. Similarly, we apply ENCO on top of both to obtain the learned graph.

Metrics We assess the quality of the representations with DCI disentanglement and completeness scores. DCI disentanglement score measures the extent to which the latent factors in a representation are independent, while the completeness score assesses how well the representation captures all the factors of variation in the data. In order to evaluate how well the models recover the true causal graph, we measure the Structural Hamming Distance (SHD) between the inferred and the true graph.

Results Our method demonstrates superior or competitive performance compared to the baselines, as indicated by the metrics shown in Figure 2, and Figure 3 in the appendix. In higher dimensions, our method excels by acquiring more information about the causal variables and underlying causal structure.

6. Conclusion

Identifying the underlying causal variables and mechanisms of a system solely from observational data is considered

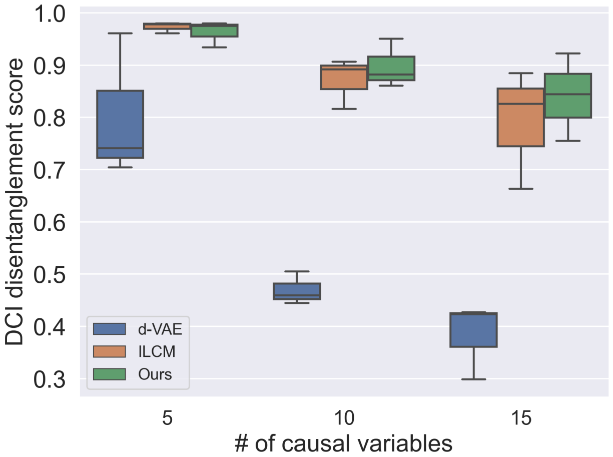

Figure 2. Comparison of models on Causal Disentanglement score. Our method is able to disentangle causal variables comparatively well.

impossible without additional assumptions. In this project, we use weak supervision as an inductive bias and study if the information encoded in the latent code of diffusion-based representations contains useful knowledge of causal variables and the underlying causal graph.

This study serves as an initial exploration of applying diffusion models to Causal Representation Learning, acknowledging that there may exist more effective approaches for integrating diffusion models within this domain. Additionally, the representation learning process relies on an encoder, which acts as an information channel, regulating the amount of input information transmitted to the score function during each step of the diffusion process. It is important to note that in certain scenarios, the encoder may not be essential to the diffusion process and could potentially result in collapsing behavior.

Therefore, significant further research and extensions are needed. This can include the extension to video or multi-view data as well as the application in reinforcement learning or experimental design settings.

References

Abstreiter, K., Mittal, S., Bauer, S., Schölkopf, B., and Mehrjou, A. Diffusion-based representation learning, 2022.

Anwar, A. R., Mideska, K. G., Hellriegel, H., Hoogenboom, N., Krause, H., Schnitzler, A., Deuschl, G., Raethjen, J., Heute, U., and Muthuraman, M. Multi-modal causality analysis of eyes-open and eyes-closed data from simultaneously recorded eeg and meg. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2825–2828. IEEE, 2014.

Brehmer, J., De Haan, P., Lippe, P., and Cohen, T. S. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Hellström, T. The relevance of causation in robotics: A review, categorization, and analysis. *Paladyn, Journal of Behavioral Robotics*, 12(1):238–255, 2021.

Hernán, M. Á., Brumback, B., and Robins, J. M. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pp. 561–570, 2000.

Hiemstra, C. and Jones, J. D. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models, 2022.

Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., and Dittadi, A. Diffusion models for video prediction and infilling, 2022.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. Causalgan: Learning causal implicit generative models with adversarial training, 2017.

Komanduri, A., Wu, Y., Huang, W., Chen, F., and Wu, X. Scm-vae: Learning identifiable causal representations via structural knowledge. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1014–1023. IEEE, 2022.

Lippe, P., Cohen, T., and Gavves, E. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021.

Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., Hengel, A. v. d., Zhang, K., and Shi, J. Q. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2022.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Luo, C. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

Mittal, S., Lajoie, G., Bauer, S., and Mehrjou, A. From points to functions: Infinite-dimensional representations in diffusion models, 2022.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022.

Robins, J. M., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pp. 550–560, 2000.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations, 2021.

Subramanian, J., Annadani, Y., Sheth, I., Ke, N. R., Deleu, T., Bauer, S., Nowrouzezahrai, D., and Kahou, S. E. Learning latent structural causal models. *arXiv preprint arXiv:2210.13583*, 2022.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv e-prints*, pp. arXiv–2004, 2020.

A. Problem Formulation & ELBO

The ELBO for the proposed framework will be:

$$\log p(x,\tilde{x}) \geq \mathbb{E}_{q(e,\tilde{e},u,\tilde{u},I|x,\tilde{x})}\Big[\log \frac{p(x,\tilde{x},u,\tilde{u},e,\tilde{e},I)}{q(e,\tilde{e},I,u,\tilde{u}|x,\tilde{x})}\Big] \tag{3}$$

$$=\mathbb{E}_{q(e,\tilde{e},u,\tilde{u},I|x,\tilde{x})}\Big[\log \frac{p(I)}{q(I|x,\tilde{x})} + \log \frac{p(e)p(\tilde{e}|e,I)}{q(e,\tilde{e}|x,\tilde{x},I)} + \log \frac{p(x,u|e)}{q(u|x)} + \log \frac{p(\tilde{x},\tilde{u}|\tilde{e})}{q(\tilde{u}|\tilde{x})}\Big] \tag{4}$$

$$=\mathbb{E}_{q(I|x,\tilde{x})}\mathbb{E}_{q(e,\tilde{e}|x,\tilde{x},I)}\mathbb{E}_{q(u|x)}\mathbb{E}_{q(\tilde{u}|\tilde{x})}\Bigg[\Big[\log p(I) + \log p(e) + \log p(\tilde{e}|e,I) - \log q(I|x,\tilde{x}) - \log q(e,\tilde{e}|x,\tilde{x},I)\Big] \tag{5}$$

$$+\Big[\log \frac{p(x,u|e)}{q(u|x)} + \log \frac{p(\tilde{x},\tilde{u}|\tilde{e})}{q(\tilde{u}|\tilde{x})}\Big]\Bigg] \tag{6}$$

The terms in Eq. 5 correspond to the intervention encoder and the noise encoding module, respectively, and the terms in Eq. 6 correspond to the diffusion model conditioned on pre- and post-intervention noise encodings.

It's been shown that the discretization of SDE formulations of the diffusion model is equivalent to discrete-time diffusion models (Song et al., 2021). Therefore, for simplicity, we derive the ELBO for discrete-time diffusion models. For a discrete-time diffusion model where $t \in [1,T]$, we have (Luo, 2022):

$$\mathbb{E}_{q(I|x,\tilde{x})}\mathbb{E}_{q(e,\tilde{e}|x,\tilde{x},I)}\mathbb{E}_{q(u|x)}\mathbb{E}_{q(\tilde{u}|\tilde{x})}\left[\log \frac{p(x,u|e)}{q(u|x)}\right] = \mathbb{E}_{q(I|x,\tilde{x})}\mathbb{E}_{q(e,\tilde{e}|x,\tilde{x},I)}\mathbb{E}_{q(u|x)}\mathbb{E}_{q(\tilde{u}|\tilde{x})}\Bigg[\mathbb{E}_{q(u_1|x)}[\log p(x|u_1)]$$

$$-D_{KL}(q(u_T|x)||p(u_T)) - \sum_{t=2}^{T}\mathbb{E}_{q(u_t|x)}[D_{KL}(q(u_{t-1}|u_t,x,e)||p(u_{t-1}|u_t,e)]\Bigg]$$

where

- $\mathbb{E}_{q(u_1|x)}[\log p(x|u_1)]$ is the reconstruction term and it can be defined in a way that it is constant so it can be ignored during training;
- $D_{KL}(q(u_T|x)||p(u_T))$ is the prior matching term and can similarly be defined in a way that it is constant;
- $\mathbb{E}_{u_t|x}[D_{KL}(q(u_{t-1}|u_t,x,e)||p(u_{t-1}|u_t,e)]$ is a denoising matching term. This term is the origin of different interpretations of the score-based diffusion models. For the SDE formulation of the forward process, the denoising matching term becomes (Song et al., 2021):

  $$\lambda(t)||s_\theta(u_t,e,t) - \nabla_{u_t}\log p(u_t|x)||_2^2$$

  It's been shown that the likelihood weighting of denoising matching terms is related to the diffusion coefficient of the forward SDE, i.e. $\lambda(t) = g^2(t)$. Therefore, for a Variance Exploding SDE, we have:

  $$\lambda(t) = 2\sigma^2(t)\log\left(\frac{\sigma_{max}}{\sigma_{min}}\right)$$

  where

  $$\sigma(t) = \sigma_{min} \cdot \left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t$$

Therefore, the ELBO will become:

$$\log p(x,\tilde{x}) \geq \mathbb{E}_{p(x,\tilde{x})}\mathbb{E}_{q(I|x,\tilde{x})}\mathbb{E}_{q(e,\tilde{e}|x,\tilde{x},I)}\mathbb{E}_{t\sim U(0,1)}\mathbb{E}_{q(u_t|x)}\mathbb{E}_{q(\tilde{u}_t|\tilde{x})}$$

$$\Bigg[\log p(I) + \log p(e) + \log p(\tilde{e}|e,I) - \log q(I|x,\tilde{x}) - \log q(e,\tilde{e}|x,\tilde{x},I)$$

$$+\lambda(t)\Big[||s_\theta(u_t,e,t) - \nabla_{u_t}\log p(u_t|x)||_2^2 + ||s_\theta(\tilde{u}_t,\tilde{e},t) - \nabla_{\tilde{u}_t}\log p(\tilde{u}_t|\tilde{x})||_2^2\Big]\Bigg]$$

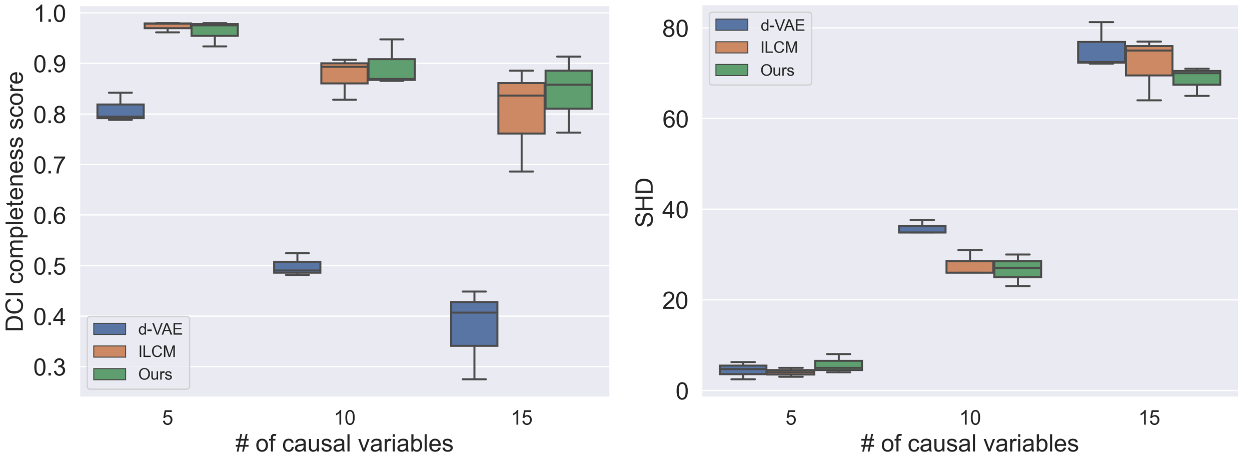

Figure 3. Comparison of models on Causal Completeness score and Structural Hamming Distance (SHD). Our method performs well compared to baselines.

B. Experiments

Data Generation We generate random graphs by sampling edges from a Bernoulli distribution with a probability of 0.5. We consider the SCM to be linear Gaussian and we sample the weights from a Multivariate Normal distribution with a mean of 0 and a variance of 1. We make sure the weights are not close to zero to avoid the violation of the faithfulness assumption. We introduce additive Gaussian noise with equal variances across all nodes, with its variance set to 0.1. Data was then sampled using ancestral sampling, and we generate 100k training samples, 10k train, and 10k validation samples. Finally, to generate input data x, we apply a random linear projection on the obtained latent variables. We keep the dimension of x fixed to 16. We generate data for 4 different seeds and present the results.

Training For the training, we follow the 4-phase training of Brehmer et al. (2022) and consider the same hyperparameters for the training, with the new β-VAE loss as the objective function and consider the coefficient of the regularization term $\mathcal{L}_{entropy}$ to be 1.

Architectures & Hyperparameters The noise encoder is considered Gaussian, with mean and standard deviation parameterized as an MLP with two hidden layers and 64 units each and ReLU activation functions. The noise encodings have a property that only for the elements that are intervened upon, we have $e_i \neq \tilde{e}_i, i \in I$. Considering this property, the intervention encoder is defined as:

$$\log q(i \in I|x, \tilde{x}) = \frac{1}{Z}(\alpha + \beta|\mu_e(x)_i - \mu_e(\tilde{x})_i| + \gamma|\mu_e(x)_i - \mu_e(\tilde{x})_i|^2)$$

Where $\mu_e(x)$ is the mean of the noise encoder, α, β, γ are learnable parameters, and Z is a normalization constant. Finally, in order to define $p(\tilde{e}_i|e)$ in Eq. 2 for $i \in I$, we use a conditional normalizing flow as:

$$p(\tilde{e}_i|e) = \tilde{p}(s_i(\tilde{e}_i|e_i))\left|\frac{\partial s_i(\tilde{e}_i; e_i)}{\tilde{e}_i}\right|$$

The architecture of the score function in the diffusion model is based on NCSN++ architecture (Song et al., 2021). As the input x is 16-dimensional and the score model follows a convolutional architecture, we reshape the input into a 4×4 format and then feed it into the diffusion model. Furthermore, In the forward SDE, σ_{min} and σ_{max} are set to 0.01 and 50, respectively.