Towards Understanding When Causal Structure Improves Robustness: Evidence from Generative Models

Manal Benhamza[1] **Marianne Clausel**[2] **Myriam Tami**[1]
[1]Paris-Saclay University, CentraleSupélec, MICS [2]Lorraine University, CRAN

Abstract

Causal Generative Models aim to incorporate causal knowledge into black box deep generative architectures, thereby improving transparency and interpretability. It is commonly claimed that integrating the causal structure also enhances robustness. However, to the best of our knowledge, no prior work has systematically compared the robustness of standard generative models with that of their causal counterparts. Hence, in this work, our aim is to address this gap by providing a principled comparative study of robustness between standard and causal generative frameworks that is theoretically grounded in the VAE setting. Our extensive experiments conducted on synthetic and real-world datasets across different configurations and scales show that the difference in robustness levels between the standard and causal frameworks is tightly related to the structure of the encoded causal mechanisms. In particular, we provide an intuitive explanation based on the critical properties of the underlying causal graph.

1 Introduction

Recent advances in generative modeling have led to substantial improvements in modeling complex probability distributions (Komanduri et al., 2024a). Despite these advances, standard generative models only provide limited insight into the learned latent representations, thus hindering interpretability. Causal generative models mitigate this limitation by modeling causal relationships among latent variables. Incorporating causal structures into generative models has led to more interpretability, thus fostering their adoption across diverse application domains. They can be used, for example, in biological sciences with sequencing data to learn latent causal variables and their related causal relationships, hence providing some explanation to complex biological systems (Squires et al., 2023). In highstakes domains, such as medecine and industry, these models further facilitate counterfactual reasoning. In particular, they allow the generation of scenarios with unseen specific attributes, through controlled interventions on latent causal factors (Benhamza et al., 2025). More generally, causal representations enable multiple aspects of trustworthy AI by achieving fairness (Schölkopf et al., 2021; Zuo et al., 2023), interpretability (Wu et al., 2023), transparency (Bouchattaoui et al., 2024; Bouchattaoui, 2025), and avoiding spurious correlations (Beery et al., 2018). However, standard generative models remain vulnerable to distribution shifts, in which small perturbations in input can lead to significant changes in generated output (Cui et al., 2024; Kos et al., 2017; Liu et al., 2025). Thereafter, we assume that despite their promising advantages, causal generative models are not inherently robust to such shifts and may be susceptible to the same robustness limitations as their standard counterparts.

Many prior works have addressed the robustness of standard generative models (Cui et al., 2024; Pope et al., 2020; Kos et al., 2017; Liu et al., 2025). Others have provided explanations for the difference in causal generative model's robustness levels across interventions on causal latent variables (Benhamza et al., 2025). The latter have shown that the robustness of counterfactual causal generative models is tightly related to the causal mechanisms. However, to the best of our knowledge, none has ever compared the robustness of standard generative models to their causal counterparts. Hence, in this work, we bridge this gap by conducting a principled comparative analysis of robustness in both standard and causal generative settings, which is theoretically grounded in the Variational Autoencoder (VAE) framework. Our study, more specifically, aims to characterize the conditions under which incorporating causal structure into latent representations leads to improved robust-

ness guarantees.

In particular, we conduct extensive experiments on the Pendulum (Yang et al., 2020) and CelebA (Liu et al., 2015) datasets to compare the robustness of common standard generative models, namely Variational Autoencoders (VAE) (Kingma and Welling, 2019), Generative Adversarial Networks (GAN) (Goodfellow et al., 2020) and Diffusion based models (Sohl-Dickstein et al., 2015), with their respective causal counterparts Causal/SCM-VAE (Yang et al., 2020; Komanduri et al., 2022), DEAR (Shen et al., 2022) and CausalDiffAE (Komanduri et al., 2024b) to distribution shifts, under different model configurations. It is worth noting that our analysis focuses mainly on the VAE framework, as many causal generative models are built upon it, as illustrated in the taxonomy proposed in (Komanduri et al., 2024a). Nevertheless, we also consider other generative frameworks, such as GANs and diffusion models, due to their remarkable generative performance and their growing relevance in large-scale and high-fidelity generation settings. Experiments results are to be presented during the workshop. They indicate that the robustness depends mainly on the structure of the encoded causal mechanisms. More specifically, they suggest that the difference in robustness levels can be further explained by leveraging key properties of the underlying causal graph. These empirical results are interpreted through the lens of the theoretical analysis conducted in (Benhamza et al., 2025). Notably, we state here bounds on the minimal input perturbation required to significantly change an extended CausalVAE's reconstruction, thereby extending existing results on standard VAEs (Camuto et al., 2021) and counterfactual causal VAEs (Benhamza et al., 2025). Unlike VAEs, the resulting bounds in the causal setting depend on key structural properties of the causal graph's adjacency matrix. These theoretical findings are consistent with our empirical results.

2 Related Work

Causal Generative Models. Causal Generative models seek to incorporate causal structure into standard generative frameworks, such as VAE, GAN, or Diffusion models, to enhance their interpretability. Notably, CausalVAE (Yang et al., 2020) is a VAE based framework, which transforms the encoded independent latent factors $\boldsymbol{\eta}$ into latent causal ones $\boldsymbol{z}$, through a linear Structural Causal Model (SCM). To overcome the limitations of the CausalVAE, mainly the linear SCM assumption, SCM-VAE (Komanduri et al., 2022) adopts a post-nonlinear additive noise SCM, thus enabling more general relationships among the causal latent variables. Among the causal counterparts defined in the literature for the VAE framework, we focus on the CausalVAE and SCM-VAE to assess the impact of non-linear SCMs on the robustness of causal generative models. DEAR (Shen et al., 2022) which stands for Disentangled gEnerative cAusal Representation, leverages an SCM as the prior distribution for a bidirectional GAN. As for CausalDiffAE (Komanduri et al., 2024b), it extends the SCM-VAE to diffusion-based models, by using a denoising diffusion implicit model (DDIM) for the reconstruction. To ensure the identifiability of the learned representations, all models use labels as additional supervisory signals.

Generative Models Robustness. (Camuto et al., 2021), by providing global margin bounds for the Variational Autoencoder (VAE), have put in evidence which parameters can be controlled to guarantee more robustness. Building on this result, (Barrett et al., 2022) established that a VAE can be designed with a predefined level of robustness by carefully controlling the Lipschitz constants of both the encoder and decoder. Their theoretical framework, however, is specifically tailored to standard VAE architectures and does not readily generalize to other classes of generative models. While (Benhamza et al., 2025) extend it to counterfactual models derived from causal generative frameworks under interventions on latent causal variables, we express global margin bounds for the extended CausalVAE itself. In a complementary perspective, (Kos et al., 2017) proposed an approach for constructing adversarial perturbations capable of significantly altering the outputs of various generative models. (Liu et al., 2025) introduce a new training method to enhance the robustness of diffusion-based purification. (Benhamza et al., 2025) investigate the robustness of counterfactual models. They demonstrate that differences in counterfactual robustness levels can be attributed to the underlying causal mechanisms. More specifically, they show that the edges in a causal graph have robustness scores, and thus their removal with an intervention can make the counterfactual model either more or less robust. To the best of our knowledge, however, no prior work has systematically compared the robustness of standard generative models with that of their causal counterparts. In this work, we address this gap through an extensive experimental analysis, theoretically grounded in the VAE framework, to assess the impact of incorporating causal structure on the robustness of standard generative models.

3 Theoretical Results

In this section, we present insights about the robustness margin bounds for causal extensions of VAE.

Background. Let $\mathcal{X} = (\mathbf{x}_j)_{1\leq j'\leq N}$ be a set of observed variables where $\mathbf{x}_{j'} \in \mathbb{R}^{N_d}$. We introduce a set of latent variables $\boldsymbol{z} = (z_j)_{1\leq j\leq M}$, interpreted as underlying causal latent variables. The latter are assumed to follow a structural causal model (SCM) $S = (E, P^{\eta})$, where $E = (E_1, \ldots, E_M)$ denotes a collection of structural equations of the form

$$E_j : \quad z_j = f_j(\mathrm{PA}_j, \eta_j),$$

with PA_j denoting the parent set of z_j. The noise variables $(\eta_j)_{j=1}^{M}$ are assumed to be mutually independent. We further assume that the causal relationships among the latent variables are represented by a known directed acyclic graph (DAG), characterized by its adjacency matrix $\boldsymbol{A}$.

3.1 Path Matrices

Let G be a directed acyclic graph, and $\boldsymbol{A}$ its adjacency matrix. We recall that for powers of the adjacency matrix $\boldsymbol{A}^k$, if $\boldsymbol{A}$ is binary, the entry $[\boldsymbol{A}^k]i, j$ gives the total number of paths of length k from node i to node j. In contrast, when $\boldsymbol{A}$ is weighted, $[\boldsymbol{A}^k]i, j$ corresponds to the sum of the weights of all such paths. We define the path matrix $\boldsymbol{P}$ as::

$$\boldsymbol{P} = \sum_{k\geq 0} (\boldsymbol{A}^{\boldsymbol{T}})^k = (\boldsymbol{I} - \boldsymbol{A}^T)^{-1} \tag{1}$$

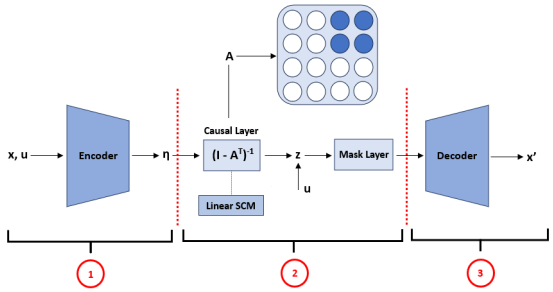
Note that the nilpotence of $\boldsymbol{A}$ implies that this sum is finite. Each entry in the path matrix $[\boldsymbol{P}]_{ij}$, thus represents the aggregate weight of all directed paths from node i to node j across all admissible lengths. In the binary case, where each edge weight is equal to 1, this quantity reduces to the total number of such directed paths.

3.2 Extended CausalVAE

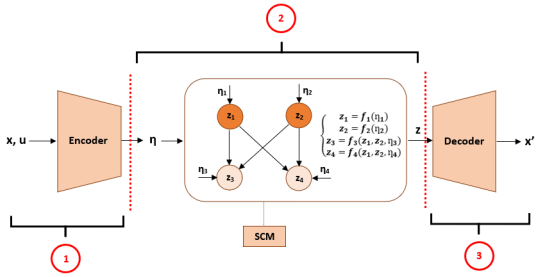
As in (Benhamza et al., 2025), we call extended CausalVAE model, all VAE based causal models where the causal layer, which encodes the causal mechanisms as illustrated in block 2 of Fig. 1a and 1b, implements the general non linear SCM (Yue Yu and Yu, 2019) in the following Eq. 2:

$$\boldsymbol{z} = \pi_1((\boldsymbol{I} - \boldsymbol{A}^T)^{-1}\pi_2(\boldsymbol{\eta})) \tag{2}$$

where π_1 and π_2 are non-linear element-wise transformations. By integrating a parametric SCM within the causal layer, the extended CausalVAE framework encompasses both the linear CausalVAE and the SCM-VAE models, depending on the specific choices of the functions π_1 and π_2. In the CausalVAE setting, Fig. 1a, both π_1 and π_2 are identity mappings, while in the SCM-VAE setting, Fig. 1b, π_2 is the identity function and π_1 is a non-linear transformation learned by a neural network.



(a) CausalVAE



(b) SCM-VAE

Figure 1: Frameworks of the CausalVAE (a) and SCM-VAE (b) model. Blocks 1 refer to the encoding process, blocks 2 enclose the causal mechanisms, and blocks 3 hold the decoding process.

3.3 Robustness Margin Bounds

In Appendix A, building on (Camuto et al., 2021) we express lower robustness margin bounds for an extended CausalVAE leveraging its structure and the Lipschitz continuity of its components. We should first recall the definition of a robustness margin (Camuto et al., 2021).

Definition 1 (Robustness Margin) *For $r \in \mathbb{R}^+$ and $\epsilon \in [0, 1)$, a generative model F has an (r, ϵ)-robustness margin $R_{(r,\epsilon)}$ about input $\mathbf{x}$ if :*

$$\|\boldsymbol{\delta}\|_2 < R_{(r,\epsilon)}(\mathbf{x}) \implies P\Big[\|F(\mathbf{x}+\boldsymbol{\delta}) - F(\mathbf{x})\|_2 \leq r\Big] > \epsilon$$

The Def. 1 stipulates that a generative model F with a robustness margin $R_{(r,\epsilon)}(\mathbf{x})$ can withstand for its input $\mathbf{x}$ all perturbations $\boldsymbol{\delta}$ with a norm $\|\boldsymbol{\delta}\|_2 < R_{(r,\epsilon)}(\mathbf{x})$. Lower global bounds on this margin therefore inform

us about the robustness of the model over the entire input space. In particular, for an extended CausalVAE, they are expressed as a function of $\lambda_1(\boldsymbol{P})$, the largest singular value of the path matrix $\boldsymbol{P} = (\boldsymbol{I} - (\boldsymbol{A})^T)^{-1}$. Unlike standard VAEs, whose robustness margin bounds only depend on the parameters of the model, mainly the Lipschitz constants of both the encoder and decoder, its causal counterparts robustness also include additional terms $\lambda_1(\boldsymbol{P})$ and the Lipschitz constants of π_1 and π_2, as established in Appendix A. *Notably, under careful choice of the Lipschitz constants, we show in Appendix A, that for the same VAE and CausalVAE parameters configurations, lower values of* $\lambda_1(\boldsymbol{P})$ *yield higher robustness margin bounds compared to the VAE.* The latter hence suggests that if there is a direction in the causal graph that contains multiple direct causal paths, the perturbation effect may be amplified, hence undermining robustness. This result will be illustrated at the workshop.

4 Experimental Results

We compare the robustness of multiple standard generative frameworks to their causal counterparts on synthetic and real-world datasets, Pendulum (Yang et al., 2020) and CelebA (Liu et al., 2015), under different configurations and scales. Pendulum is a synthetic dataset that simulates the dynamic behavior of a physical pendulum, whereas CelebA is a large-scale real-world dataset of human faces containing more than $200k$ images. These datasets were selected to reflect distinct causal structures, enabling us to analyze how the underlying causal mechanisms influence robustness properties. The results for VAE and GAN based models are presented in Appendix B. Diffusion-based models results will be presented at the workshop.

5 Conclusion

In this work, we conduct a comparative study between the robustness of standard generative frameworks and their causal counterparts, under distribution shifts. Our experimental results along with a rigorous theoretical analysis in the case of the extended CausalVAE show that the difference in robustness levels is tightly related to the encoded causal mechanisms. While our theoretical study considers a general perturbation framework, some of the distribution shifts used in our experimental setup can, in specific cases, be related to interventions in a structural causal model. In particular, shifts that correspond to targeted modifications of certain generative factors; while leaving the underlying data-generating mechanisms unchanged; are conceptually close to interventions. However, this correspondence remains partial and depends on the nature of the shift under consideration. Not all distribution shifts admit such an interpretation, and the extent to which a causal perspective applies varies across settings. While causal models are, by construction, well-suited to represent and reason about interventions, this does not imply that they are systematically advantaged across all the shifts considered, nor that alternative modeling approaches are inherently inadequate. Further extending the experimental analysis beyond the distribution shifts already considered, to encompass a broader class of perturbations, would enable a more comprehensive evaluation of robustness, including scenarios relevant to the security of machine learning systems. In such settings, structured causal representations may provide advantages by constraining how perturbations propagate through the generative process. Exploring this connection between causal modeling and robustness under more general threat models is an important direction for future work.

6 Acknowledgements

This research was funded in whole or in part by the National Research Agency (ANR) under the project "ANR-23-CE23-0008-01".

References

Barrett, B., Camuto, A., Willetts, M., and Rainforth, T. (2022). Certifiably Robust Variational Autoencoders . In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 3663–3683.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Benhamza, M., Clausel, M., and Tami, M. (2025). Counterfactual robustness: A framework to analyze the robustness of causal generative models across interventions. In *European Conference, ECML PKDD 2025,*, page 391–408.

Bouchattaoui, M. E. (2025). Learning causality for longitudinal data.

Bouchattaoui, M. E., Tami, M., LEPETIT, B., and Cournède, P.-H. (2024). Toward a more transparent causal representation learning. In *9th Causal Inference Workshop at UAI 2024*.

Camuto, A., Willetts, M., Roberts, S., Holmes, C., and Rainforth, T. (2021). Towards a Theoretical Understanding of the Robustness of Variational Autoencoders. pages 3565–3573.

Cui, X., Aparcedo, A., Jang, Y. K., and Lim, S.-N. (2024). On the Robustness of Large Multimodal

Models Against Image Adversarial Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*, pages 24625–24634.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in neural information processing systems*, vol 30.

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392.

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*.

Komanduri, A., Wu, Y., Huang, W., Chen, F., and Wu, X. (2022). SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge. In *IEEE International Conference on Big Data*, pages 1014–1023.

Komanduri, A., Zhao, C., Chen, F., and Wu, X. (2024b). Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *Proceedings of the 27th European Conference on Artificial Intelligence*.

Kos, J., Fischer, I., and Song, D. X. (2017). Adversarial Examples for Generative Models. *IEEE Security and Privacy Workshops SPW*, pages 36–42.

Liu, Y., Liu, K., Xiao, Y., Dong, Z., Xu, X., Wei, P., and Lin, L. (2025). Towards understanding the robustness of diffusion-based purification: A stochastic perspective. In *The Thirteenth International Conference on Learning Representations*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision ICCV*, pages 3730–3738.

Pope, P., Balaji, Y., and Feizi, S. (2020). Adversarial Robustness of Flow-Based Generative Models. In *International Conference on Artificial Intelligence and Statistics*, pages 3795–3805.

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*, vol 109:612–634.

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. (2022). Weakly Supervised Disentangled Generative Causal Representation Learning. *Journal of Machine Learning Research*, pages 1–55.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.

Squires, C., Seigal, A., Bhate, S., and Uhler, C. (2023). Linear causal disentanglement via interventions. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Wu, Z., Geiger, A., Icard, T., Potts, C., and Goodman, N. (2023). Interpretability at scale: Identifying causal mechanisms in alpaca. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2020). CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. *CVPR*, pages 9588–9597.

Yue Yu, Jie Chen, T. G. and Yu, M. (2019). DAG-GNN: DAG Structure Learning with Graph Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*.

Zuo, Z., Khalili, M. M., and Zhang, X. (2023). Counterfactually fair representation. NIPS '23. Curran Associates Inc.

Towards Understanding When Causal Structure Improves Robustness: Evidence from Generative Models Supplementary Materials

A Theoretical Results

Theorem A.1 (Extended CausalVAE Probability Bound) *Let Mask denote the mask layer. Given that* $\boldsymbol{z} = Mask(\pi_1((\boldsymbol{I} - \boldsymbol{A}^T)^{-1}\pi_2(\boldsymbol{\eta})))$*, the variational posterior* $q_\phi(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{u}) = \mathcal{N}(\boldsymbol{\eta}; \mu_\phi(\mathbf{x}, \boldsymbol{u}), \text{diag}(\sigma^2_\phi(\mathbf{x}, \boldsymbol{u})))$ *and that the deterministic component of the extended CausalVAE decoder* $g_\theta(\cdot)$ *is a-Lipschitz, the encoder mean* $\mu_\phi(\cdot)$ *is b-Lipschitz, and the encoder standard deviation* $\sigma_\phi(\cdot)$ *is c-Lipschitz. The element-wise transformations of the general non-linear SCM are also* π_j *respectively* γ_j *Lipschitz. The mid-linear functions* g_j *of the Mask layer are respectively* β_j *Lipschitz. Finally, let* $\boldsymbol{\eta}_{\boldsymbol{\delta}} \sim q_\phi(\boldsymbol{\eta}|\mathbf{x} + \boldsymbol{\delta}, \boldsymbol{u})$ *and* $\boldsymbol{\eta}_{\neg\boldsymbol{\delta}} \sim q_\phi(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{u})$ *since we conserve the same label for the perturbed and non-perturbed inputs. For* $\forall\, r \in \mathbb{R}^+$, $\forall\, \mathbf{x} \in \mathcal{X}$ *and* $\forall\, \boldsymbol{\delta} \in \mathcal{X}$,

$$P\left[\|g_\theta(\boldsymbol{z}_{\boldsymbol{\delta}}) - g_\theta(\boldsymbol{z}_{\neg\boldsymbol{\delta}})\|_2 \leq r\right] \geq 1 - \min\{p_1(\mathbf{x}, \boldsymbol{u}), p_2(\mathbf{x}, \boldsymbol{u})\}$$

where

$$p_1(\mathbf{x}, \boldsymbol{u}) := \min\left(1, \frac{k^2(b^2\|\boldsymbol{\delta}\|_2^2 + (c\|\boldsymbol{\delta}\|_2 + 2\|\sigma_\phi(\mathbf{x}, \boldsymbol{u})\|_2)^2)}{r^2}\right)$$

$$p_2(\mathbf{x}, \boldsymbol{u}) := \begin{cases} C(M)\dfrac{n(\mathbf{x}, \boldsymbol{u})^{\frac{M}{2}} \exp\left(-\frac{n(\mathbf{x}, \boldsymbol{u})}{2}\right)}{(n(\mathbf{x}, \boldsymbol{u}) - M + 2)} & \text{if } \frac{r}{k} - b\|\boldsymbol{\delta}\|_2 \geq 0; M \geq 2; n(\mathbf{x}, \boldsymbol{u}) > M - 2 \\ 1 & o.w. \end{cases}$$

$$n(\mathbf{x}, \boldsymbol{u}) := \left(\frac{\frac{r}{k} - b\|\boldsymbol{\delta}\|_2}{c\|\boldsymbol{\delta}\|_2 + 2\|\sigma_\phi(\mathbf{x}, \boldsymbol{u})\|_2}\right)^2$$

$$C(M) := \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{2}\left(M - (M-1)\log M\right)\right)$$

and

$$k = a\gamma_1\gamma_2 \left(\sqrt{\sum_{i=1}^{M} \beta_i^2 \max(\boldsymbol{A}_i)^2}\right) \|(\boldsymbol{I} - \boldsymbol{A}^T)^{-1}\|_2$$

Proof. Step 1 : Lower bound on $P\left[\|g_\theta(\boldsymbol{z}_{\boldsymbol{\delta}}) - g_\theta(\boldsymbol{z}_{\neg\boldsymbol{\delta}})\|_2 \leq r\right]$
We first have to prove the Lipschitz continuity of the Mask Operator. Let $\boldsymbol{z}$ and $\boldsymbol{z}'$ be two causal representations in the latent space, we have:

$$\begin{aligned}
\|\text{Mask}(\boldsymbol{z}') - \text{Mask}(\boldsymbol{z})\|_2 &= \sqrt{\Sigma_{i=1}^{M}(\text{Mask}(\boldsymbol{z}')_i - \text{Mask}(\boldsymbol{z})_i)^2} \\
&= \sqrt{\Sigma_{i=1}^{M}(g_i(\boldsymbol{A}_i \odot \boldsymbol{z}') - g_i(\boldsymbol{A}_i \odot \boldsymbol{z}))^2} \\
&\leq \sqrt{\Sigma_{i=1}^{M}\beta_i^2\|\boldsymbol{A}_i \odot \boldsymbol{z}' - \boldsymbol{A}_i \odot \boldsymbol{z}\|_2^2} \\
&\leq \sqrt{\Sigma_{i=1}^{M}\beta_i^2 \max(\boldsymbol{A}_i)^2\|\boldsymbol{z}' - \boldsymbol{z}\|_2^2} \\
&\leq \left(\sqrt{\Sigma_{i=1}^{M}\beta_i^2 \max(\boldsymbol{A}_i)^2}\right)\|\boldsymbol{z}' - \boldsymbol{z}\|_2
\end{aligned}$$

Where g_i are MLP with a β_i Lipschitz constant, $\boldsymbol{A}_i$ is the i^{th} column of the adjacency matrix $\boldsymbol{A}$ and $\odot$ is the element wise product. Since the SCM-VAE model does not contain a Mask Layer, we equivalently consider this

component as the identity transformation. By definition of $\boldsymbol{z_\delta}, \boldsymbol{z}_{\neg\boldsymbol{\delta}}$ and Lipschitz continuity of π_1, π_2 one has

$$\begin{aligned}\|\boldsymbol{z_\delta} - \boldsymbol{z}_{\neg\boldsymbol{\delta}}\|_2 &= \left(\sqrt{\Sigma_{i=1}^M \beta_i^2 \max(\boldsymbol{A}_i)^2}\right) \|\pi_1((\boldsymbol{I}-\boldsymbol{A}^T)^{-1}\pi_2(\boldsymbol{\eta_\delta})) - \pi_1((\boldsymbol{I}-\boldsymbol{A}^T)^{-1}\pi_2(\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}))\|_2 \\ &\leq \gamma_1 \left(\sqrt{\Sigma_{i=1}^M \beta_i^2 \max(\boldsymbol{A}_i)^2}\right) \|(\boldsymbol{I}-\boldsymbol{A}^T)^{-1}(\pi_2(\boldsymbol{\eta_\delta}) - \pi_2(\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}))\|_2 \\ &\leq \gamma_1 \left(\sqrt{\Sigma_{i=1}^M \beta_i^2 \max(\boldsymbol{A}_i)^2}\right) \|(\boldsymbol{I}-\boldsymbol{A}^T)^{-1}\|_2 \|\pi_2(\boldsymbol{\eta_\delta}) - \pi_2(\boldsymbol{\eta}_{\neg\boldsymbol{\delta}})\|_2 \\ &\leq \gamma_1 \left(\sqrt{\Sigma_{i=1}^M \beta_i^2 \max(\boldsymbol{A}_i)^2}\right) \gamma_2 \|(\boldsymbol{I}-\boldsymbol{A}^T)^{-1}\|_2 \|\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2\end{aligned} \tag{3}$$

Since $g_\theta(\cdot)$ is a-Lipschitz:

$$\|g_\theta(\boldsymbol{z_\delta}) - g_\theta(\boldsymbol{z}_{\neg\boldsymbol{\delta}})\|_2 \leq k\|\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2 \tag{4}$$

where we set $k := a\gamma_1\gamma_2\|(\boldsymbol{I}-\boldsymbol{A}^T)^{-1}\|_2 \left(\sqrt{\Sigma_{i=1}^M \beta_i^2 \max(\boldsymbol{A}_i)^2}\right)$. It implies that:

$$\{\|g_\theta(\boldsymbol{z_\delta}) - g_\theta(\boldsymbol{z}_{\neg\boldsymbol{\delta}})\|_2 \leq r\} \supseteq \{k\|\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2 \leq r\} \tag{5}$$

which in turn yields:

$$P\left[\|g_\theta(\boldsymbol{z_\delta}) - g_\theta(\boldsymbol{z}_{\neg\boldsymbol{\delta}})\|_2 \leq r\right] \geq P\left[k\|\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2 \leq r\right] = 1 - P\left[\|\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2 \geq \frac{r}{k}\right] \tag{6}$$

Step 2 : We use the assumptions on the distribution of the noise variable

$$\boldsymbol{\eta_\delta} \sim q_\phi(\boldsymbol{\eta}|\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) = \mathcal{N}(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}), \mathrm{diag}(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}))) \tag{7}$$

and

$$\boldsymbol{\eta}_{\neg\boldsymbol{\delta}} \sim q_\phi(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{u}) = \mathcal{N}(\mu_\phi(\mathbf{x}, \boldsymbol{u}), \mathrm{diag}(\sigma_\phi^2(\mathbf{x}, \boldsymbol{u}))) \tag{8}$$

Elements from $q_\phi(\boldsymbol{\eta}|\cdot)$ are sampled independently in every extended CausalVAE forward pass, and $\boldsymbol{\eta_\delta}$,$\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}$ are independent. Since the difference of independent multivariate Gaussian random variables is also multivariate Gaussian, we thus have:

$$\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}} \sim \mathcal{N}(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) - \mu_\phi(\mathbf{x}, \boldsymbol{u}), \mathrm{diag}(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u})) + \mathrm{diag}(\sigma_\phi^2(\mathbf{x}, \boldsymbol{u}))) \tag{9}$$

We will now provide two lower bounds. The proof is hence splitted into two parts, yielding $p_1(\mathbf{x}, \boldsymbol{u})$ and $p_2(\mathbf{x}, \boldsymbol{u})$ respectively.

Step 2a : Bound based on the Markov's inequality.
Markov's Inequality yields:

$$P\left[\|\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2 \geq \frac{r}{k}\right] = P\left[\sum_{i=1}^M (\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2 \geq \left(\frac{r}{k}\right)^2\right] \leq \frac{E\left[\sum_{i=1}^M (\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2\right]}{\left(\frac{r}{k}\right)^2} \tag{10}$$

Leveraging Assumption 9, we infer that the variable

$$\frac{(\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2}{(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) + \sigma_\phi^2(\mathbf{x}, \boldsymbol{u}))_i}$$

for each i, follows a non central χ^2 square distribution with one degree of freedom its centrality parameter is:

$$\frac{(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) - \mu_\phi(\mathbf{x}, \boldsymbol{u}))_i^2}{(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) + \sigma_\phi^2(\mathbf{x}, \boldsymbol{u}))_i}$$

and hence:

$$E\left(\frac{(\boldsymbol{\eta_\delta} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2}{(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) + \sigma_\phi^2(\mathbf{x}, \boldsymbol{u}))_i}\right) = 1 + \frac{(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) - \mu_\phi(\mathbf{x}, \boldsymbol{u}))_i^2}{(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) + \sigma_\phi^2(\mathbf{x}, \boldsymbol{u}))_i}$$

Following on 10, we obtain:

$$
\begin{aligned}
&E\left[\sum_{i=1}^{M}(\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2\right] \\
&=\sum_{i=1}^{M}(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi^2(\mathbf{x},\boldsymbol{u}))_i\left(1+\frac{(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))_i^2}{(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi^2(\mathbf{x},\boldsymbol{u}))_i}\right) \\
&=\sum_{i=1}^{M}(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi^2(\mathbf{x},\boldsymbol{u}))_i+\sum_{i=1}^{M}(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))_i^2
\end{aligned}
$$

Drawing on the definition of the l_2-norm:

$$
\sum_{i=1}^{M}(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))_i^2=\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2^2
$$

and the Lipschitz continuity of the encoder mean $\mu_\phi(.)$

$$
\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2\leq b\|\boldsymbol{\delta}\|_2
$$

We hence obtain:

$$
\begin{aligned}
\sum_{i=1}^{M}(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))_i^2&=\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2^2 \\
&\leq(b\|\boldsymbol{\delta}\|_2)^2
\end{aligned}
\tag{11}
$$

Similarly, following from $\sigma_\phi : X\to\mathbb{R}_M^{\geq 0}$ and the definition of the l_2-norm:

$$
\begin{aligned}
\sum_{i=1}^{M}(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi^2(\mathbf{x},\boldsymbol{u}))_i&\leq\sum_{i=1}^{M}\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})_i+\sigma_\phi^2(\mathbf{x},\boldsymbol{u})_i+2\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})_i\sigma_\phi(\mathbf{x},\boldsymbol{u})_i \\
&=\sum_{i=1}^{M}(\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi(\mathbf{x},\boldsymbol{u}))_i^2 \\
&=\left(\sqrt{\sum_{i=1}^{M}(\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi(\mathbf{x},\boldsymbol{u}))_i^2}\right)^2 \\
&=\|\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2^2
\end{aligned}
\tag{12}
$$

Then leveraging the triangle inequality and the Lipschitz continuity of $\sigma_\phi(\cdot)$:

$$
\begin{aligned}
\|\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2&=\|\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\sigma_\phi(\mathbf{x},\boldsymbol{u})+2\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2 \\
&\leq\|\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2 \\
&\leq c\|\boldsymbol{\delta}\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2
\end{aligned}
\tag{13}
$$

Based on the inequalities in 12 and 13, we get:

$$
\sum_{i=1}^{M}(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi^2(\mathbf{x},\boldsymbol{u}))_i\leq\|\sigma_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2^2\leq(c\|\boldsymbol{\delta}\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2 \tag{14}
$$

Hence, returning to 10, we obtain:

$$\begin{aligned}
&\frac{E\left[\sum_{i=1}^{M}(\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2\right]}{(\frac{r}{k})^2}\\
&=\frac{\sum_{i=1}^{M}\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})+\sigma_\phi^2(\mathbf{x},\boldsymbol{u})\right)_i+\sum_{i=1}^{M}(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))_i^2}{(\frac{r}{k})^2}\\
&\leq\frac{b^2\|\boldsymbol{\delta}\|_2^2+(c\|\boldsymbol{\delta}\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2}{(\frac{r}{k})^2}\\
&\leq\frac{k^2(b^2\|\boldsymbol{\delta}\|_2^2+(c\|\boldsymbol{\delta}\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2)}{r^2}
\end{aligned}\tag{15}$$

such that:

$$\begin{aligned}
P\left(\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2\geq\frac{r}{k}\right)&\leq\frac{E\left[\sum_{i=1}^{M}(\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2\right]}{(\frac{r}{k})^2}\\
&\leq\frac{k^2(b^2\|\boldsymbol{\delta}\|_2^2+(c\|\boldsymbol{\delta}\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2)}{r^2}
\end{aligned}\tag{16}$$

Since the right-most term is non-negative, and to have a well-defined probability, we take

$$p_1(\mathbf{x},\boldsymbol{u}):=\min\left(1,\frac{k^2(b^2\|\boldsymbol{\delta}\|_2^2+(c\|\boldsymbol{\delta}\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2)}{r^2}\right)$$

The latter verifies:

$$P\left(\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2\geq\frac{r}{k}\right)\leq p_1(\mathbf{x},\boldsymbol{u})$$

Step 2b : Obtaining $p_2(\mathbf{x})$: By leveraging the triangular inequality in 6, we have:

$$\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2\leq\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}-(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))\|_2+\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2\tag{17}$$

Thus, it implies

$$\begin{aligned}
&P\left(\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2\geq\frac{r}{k}\right)\\
&\leq P\left(\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}-(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))\|_2+\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2\geq\frac{r}{k}\right)\\
&=P\left(\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}-(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))\|_2\geq\frac{r}{k}-\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta})-\mu_\phi(\mathbf{x})\|_2\right)
\end{aligned}\tag{18}$$

Then, again recalling the latent space $Z=\mathbb{R}^M$, the inequality in 18 and the definition of the l_2 norm:

$$\begin{aligned}
&P\left(\|\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}-(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}))\|_2\geq\frac{r}{k}-\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2\right)\\
&=P\left(\sum_{i=1}^{M}(\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}-(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})))_i^2\geq\left(\frac{r}{k}-\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2\right)^2\right)\\
&\leq P\left(\sum_{i=1}^{M}\frac{(\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}-(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})))_i^2}{\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})_i+\sigma_\phi^2(\mathbf{x},\boldsymbol{u})_i}\geq\frac{\left(\frac{r}{k}-\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})\|_2\right)^2}{(c\|\boldsymbol{\delta}\|_2+2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2}\right)
\end{aligned}\tag{19}$$

Since

$$\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\sim\mathcal{N}\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u}),\operatorname{diag}\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})\right)+\operatorname{diag}\left(\sigma_\phi^2(\mathbf{x},\boldsymbol{u})\right)\right)$$

it follows that

$$\frac{(\boldsymbol{\eta}_{\boldsymbol{\delta}}-\boldsymbol{\eta}_{\neg\boldsymbol{\delta}}-(\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})-\mu_\phi(\mathbf{x},\boldsymbol{u})))_i}{\sqrt{\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})_i+\sigma_\phi^2(\mathbf{x},\boldsymbol{u})_i}}\sim\mathcal{N}(0,1)\tag{20}$$

In particular, we shall note that since $\boldsymbol{\eta}_{\boldsymbol{\delta}} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}$ is a diagonal-covariance multivariate Gaussian, then

$$\frac{(\boldsymbol{\eta}_{\boldsymbol{\delta}} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u}) - \mu_\phi(\mathbf{x},\boldsymbol{u})))_i}{\sqrt{\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})_i + \sigma_\phi^2(\mathbf{x},\boldsymbol{u})_i}}$$

are jointly independent for all $i = 1, \ldots, M$. Because the sum of squares of M independent standard Gaussian random variables has a standard χ^2 distribution with M degrees of freedom, then:

$$\sum_{i=1}^{M} \frac{(\boldsymbol{\eta}_{\boldsymbol{\delta}} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u}) - \mu_\phi(\mathbf{x},\boldsymbol{u})))_i^2}{\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u})_i + \sigma_\phi^2(\mathbf{x},\boldsymbol{u})_i} =: T \sim \chi^2(M) \tag{21}$$

Denoting

$$n'(\mathbf{x}) := \frac{\left(\frac{r}{k} - \|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u}) - \mu_\phi(\mathbf{x},\boldsymbol{u})\|_2\right)^2}{(c\|\boldsymbol{\delta}\|_2 + 2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2}$$

and

$$n(\mathbf{x}) := \frac{\left(\frac{r}{k} - b\|\boldsymbol{\delta}\|_2\right)^2}{(c\|\boldsymbol{\delta}\|_2 + 2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2)^2}$$

It implies that $n'(\mathbf{x}) \geq n(\mathbf{x})$ because $\mu_\phi(\cdot)$ is b-Lipschitz, and therefore

$$\|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u}) - \mu_\phi(\mathbf{x},\boldsymbol{u})\|_2 \leq b\|\boldsymbol{\delta}\|_2$$

and therefore

$$\frac{r}{k} - \|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u}) - \mu_\phi(\mathbf{x},\boldsymbol{u})\|_2 \geq \frac{r}{k} - b\|\boldsymbol{\delta}\|_2 \tag{22}$$

We also have:

$$(c\|\boldsymbol{\delta}\|_2^2 + 2\|\sigma_\phi(\mathbf{x},\boldsymbol{u})\|_2^2)^2 \geq 0$$

Then, based on 19 while requiring that:

$$\frac{r}{k} - \|\mu_\phi(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{u}) - \mu_\phi(\mathbf{x},\boldsymbol{u})\|_2^2 \geq \frac{r}{k} - b\|\boldsymbol{\delta}\|_2^2 \geq 0$$

to ensure the meaningfulness of the inequality in the first line of 19:

$$P[T \geq n'(\mathbf{x})] \leq P[T \geq n(\mathbf{x})]$$

The tail bound for standard χ^2 random variables (Inglot, 2010) infers:

$$P[T \geq n(\mathbf{x})] \leq C(M)\frac{n(\mathbf{x})^{\frac{M}{2}}\exp\left(-\frac{n(\mathbf{x})}{2}\right)}{(n(\mathbf{x}) - M + 2)} \tag{23}$$

with

$$C(M) := \frac{1}{\sqrt{\pi}}\exp\left(\frac{1}{2}(M - (M-1)\log M)\right)$$

Since the expression on the right-hand side is positive under specific conditions, we denote

$$p_2(\mathbf{x},\boldsymbol{u}) := \begin{cases} C(M)\frac{n(\mathbf{x})^{\frac{M}{2}}\exp\left(-\frac{n(\mathbf{x})}{2}\right)}{(n(\mathbf{x}) - M + 2)} & \text{if } \frac{r}{k} - b\|\boldsymbol{\delta}\|_2 \geq 0; M \geq 2; n(\mathbf{x}) > M-2 \\ 1 & \text{o.w.} \end{cases}$$

Thus we will get:

$$P\left[\|\boldsymbol{\eta}_{\boldsymbol{\delta}} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2 \geq \frac{r}{k}\right] \leq p_2(\mathbf{x},\boldsymbol{u})$$

Step 3 : Obtaining the final bound Choosing the least of $p_1(\mathbf{x}, \boldsymbol{u})$ and $p_2(\mathbf{x}, \boldsymbol{u})$ for a tight probability bound.

$$\begin{aligned} & P\left[\|g_\theta(\boldsymbol{z}_{\boldsymbol{\delta}}) - g_\theta(\boldsymbol{z}_{\neg\boldsymbol{\delta}})\|_2 \leq r\right] \\ & \geq 1 - P\left[\|\boldsymbol{\eta}_{\boldsymbol{\delta}} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}}\|_2 \geq \frac{r}{k}\right] \\ & \geq 1 - \min\{p_1(\mathbf{x}, \boldsymbol{u}), p_2(\mathbf{x}, \boldsymbol{u})\} \end{aligned} \tag{24}$$

Lemma A.1 (Local Margin Bound Extended CausalVAE) *Given the assumptions in Th.A.1 and an* $\epsilon \in [0, 1)$*, the* (r, ϵ)*-robustness margin of an extended CausalVAE on input* $\mathbf{x}$*:*

$$R_{(r,\epsilon)}(\mathbf{x}, \boldsymbol{u}) \geq \max\{m_1(\mathbf{x}, \boldsymbol{u}), m_2(\mathbf{x}, \boldsymbol{u})\}$$

for

$$m_1(\mathbf{x}, \boldsymbol{u}) = \frac{-4c||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2 + \sqrt{\Delta}}{2(c^2 + b^2)}$$

where

$$\Delta = (4c\|\sigma_\phi(\mathbf{x}, \boldsymbol{u})\|_2)^2 - 4(c^2 + b^2)\left(4\|\sigma_\phi(\mathbf{x}, \boldsymbol{u})\|_2 - (1 - \epsilon)\left(\frac{r}{k}\right)^2\right)$$

and

$$m_2(\mathbf{x}, \boldsymbol{u}) := \sup\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}, \mathbf{x}, \boldsymbol{u}) \leq (1 - \epsilon)\}$$

Proof. Referring to Th. A.1, for any input perturbation $\boldsymbol{\delta} \in X$ and any input $\mathbf{x} \in X$ associated to the label $\boldsymbol{u}$,

$$P\left[||g_\theta(\boldsymbol{z}_{\boldsymbol{\delta}}) - g_\theta(\boldsymbol{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] \geq 1 - \min\{p_1(\mathbf{x}, \boldsymbol{u}), p_2(\mathbf{x}, \boldsymbol{u})\} \tag{25}$$

Hence, the extended CausalVAE is (r, ϵ)-robust to perturbation $\boldsymbol{\delta}$ on input $\mathbf{x}$ for threshold $\epsilon \in [0, 1)$, if it verifies:

$$1 - \min\{p_1(\mathbf{x}, \boldsymbol{u}), p_2(\mathbf{x}, \boldsymbol{u})\} > \epsilon$$

We define the (r, ϵ) robustness margin $R_{(r,\epsilon)}(\mathbf{x}, \boldsymbol{u})$ for a model F as:

$$||\boldsymbol{\delta}||_2 < R_{(r,\epsilon)}(\mathbf{x}, \boldsymbol{u}) \implies P\left[||F(\mathbf{x} + \boldsymbol{\delta}) - F(\mathbf{x})||_2 \leq r\right] > \epsilon$$

for an extended CausalVAE $R_{(r,\epsilon)}(\mathbf{x}, \boldsymbol{u})$ will hence be the maximum perturbation verifying

$$1 - \min\{p_1(\boldsymbol{\delta}, \mathbf{x}, \boldsymbol{u}), p_2(\boldsymbol{\delta}, \mathbf{x}, \boldsymbol{u})\} \geq \epsilon$$

(We make the dependence on the perturbation $\boldsymbol{\delta}$ explicit in the probability bounds), or equivalently the robustness margin can be described as:

$$\max\left\{\sup\{||\boldsymbol{\delta}||_2 : p_1(\boldsymbol{\delta}, \mathbf{x}, \boldsymbol{u}) \leq (1 - \epsilon)\}, \sup\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}, \mathbf{x}, \boldsymbol{u}) \leq (1 - \epsilon)\}\right\}$$

Let's denote $m_1(\mathbf{x}, \boldsymbol{u}) := \sup\{||\boldsymbol{\delta}||_2 : p_1(\boldsymbol{\delta}, x, \boldsymbol{u}) \leq (1 - \epsilon)\}$. After rearranging the terms, $m_1(\mathbf{x})$ is expressed

$$\sup\left\{||\boldsymbol{\delta}||_2 : (c^2 + b^2)||\boldsymbol{\delta}||_2^2 + 4c||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2||\boldsymbol{\delta}||_2 + 4||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2^2 - (1 - \epsilon)\left(\frac{r}{k}\right)^2 \leq 0\right\}$$

Assuming c is strictly positive, the supremum is attained at the maximum square root of Eq.26 if one exists:

$$(c^2 + b^2)||\boldsymbol{\delta}||_2^2 + 4c||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2||\boldsymbol{\delta}||_2 + 4||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2^2 - (1 - \epsilon)(\frac{r}{k})^2 = 0 \tag{26}$$

Therefore, by applying the quadratic formula:

$$m_1(\mathbf{x}, \boldsymbol{u}) = \frac{-4c||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2 + \sqrt{(4c||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2)^2 - 4(c^2 + b^2)(4||\sigma_\phi(\mathbf{x}, \boldsymbol{u})||_2 - (1 - \epsilon)(\frac{r}{k})^2)}}{(c^2 + b^2)}.$$

The second expression does not admit a closed-form solution, thus:

$$m_2(\mathbf{x}, \boldsymbol{u}) := \sup\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}, \mathbf{x}, \boldsymbol{u}) \leq (1-\epsilon)\}$$

Choosing the maximum of $m_1(\mathbf{x}, \boldsymbol{u})$ and $m_2(\mathbf{x}, \boldsymbol{u})$ then yields

$$R_{(r,\epsilon)}(\mathbf{x}, \boldsymbol{u}) \geq \max\{m_1(\mathbf{x}, \boldsymbol{u}), m_2(\mathbf{x}, \boldsymbol{u})\}$$

Theorem A.2 (Global Margin Bound Extended CausalVAE) *Given the last assumptions, with $\sigma_\phi(\mathbf{x}) = \sigma \in \mathbb{R}^M_{\geq 0}$, the (r, ϵ)-robustness global margin of an extended CausalVAE all inputs considered:*

$$R_{(r,\epsilon)} \geq \max\{m_1, m_2\},$$

where

$$m_1 := \frac{\sqrt{-\left(4\|\sigma\|_2^2 - (1-\epsilon)\left(\frac{r}{k}\right)^2\right)}}{b}$$

and

$$m_2 := \sup\{\|\boldsymbol{\delta}\|_2 : p_2(\boldsymbol{\delta}) \leq (1-\epsilon)\}$$

Proof. When fixing the encoder std deviation, i.e., $\sigma_\phi(\mathbf{x}, \boldsymbol{u}) = \sigma$, the probability bound $p_1(\mathbf{x}, \boldsymbol{u})$ changes accordingly. Mainly we will no more use the Lipschitz continuity of $\sigma_\phi(.)$ and will have instead in inequality 15:

$$\begin{aligned} &\frac{E\left[\sum_{i=1}^M (\boldsymbol{\eta}_{\boldsymbol{\delta}} - \boldsymbol{\eta}_{\neg\boldsymbol{\delta}})_i^2\right]}{(\frac{r}{k})^2} \\ &= \frac{\sum_{i=1}^M \left(\sigma^2 + \sigma^2\right)_i + \sum_{i=1}^M (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}, \boldsymbol{u}) - \mu_\phi(\mathbf{x}, \boldsymbol{u}))_i^2}{(\frac{r}{k})^2} \\ &\leq \frac{b^2\|\boldsymbol{\delta}\|_2^2 + 4\|\sigma\|_2^2}{(\frac{r}{k})^2} \end{aligned} \tag{27}$$

and then by following the same steps in Th.7.3 and Lemma.7.2, we get:

$$m_1 = \frac{\sqrt{-\left(4\|\sigma\|_2^2 - (1-\epsilon)\left(\frac{r}{k}\right)^2\right)}}{b}$$

As for m_2, we can directly plug the new fixed value in $p_2(\boldsymbol{\delta}, \mathbf{x}, \boldsymbol{u})$ hence obtaining:

$$m_2 = \sup\{\|\boldsymbol{\delta}\|_2 : p_2(\boldsymbol{\delta}) \leq (1-\epsilon)\}$$

Similarly to what has been established in (Camuto et al., 2021) for the decoder's Lipschitz constant a, we show, in this work, that the robustness margin bounds are decreasing functions of k. Therefore lower values of k yield higher robustness margin bounds. Mainly for $k < a$, the VAE's corresponding marging bounds are lower than it causal counterparts. More specifically, under careful choice of the Lipschitz constants, for the same VAE and CausalVAE parameters configurations, lower values of $\lambda_1(\boldsymbol{P})$ yield higher robustness margin bounds compared to the VAE.

B Experimental Results

To evaluate the robustness of both traditional and causal generative models, we follow the pipeline illustrated in Fig.2. We consider a set of 16 perturbations, each applied at 5 levels of severity, to simulate a range of realistic distribution shifts. Model performance is assessed using the Fréchet Inception Distance (FID) as a similarity measure between the clean and perturbed reconstructed datasets. The latter measures the similarity between two datasets based on feature representations extracted using a pre-trained InceptionV3 network (Heusel et al., 2017). Lower FID scores indicate a higher similarity between the evaluated datasets and, consequently, higher robustness.

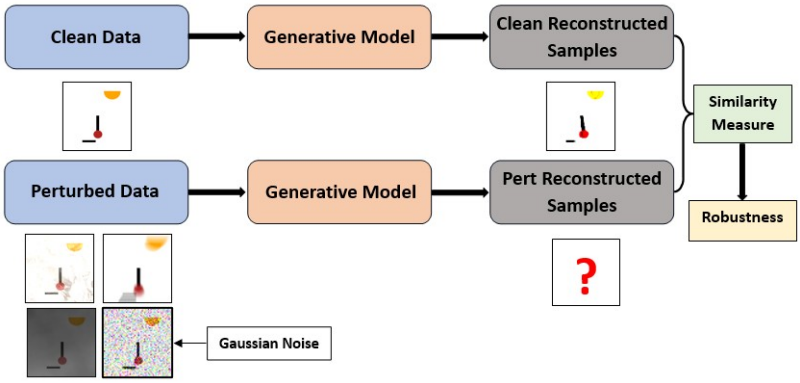
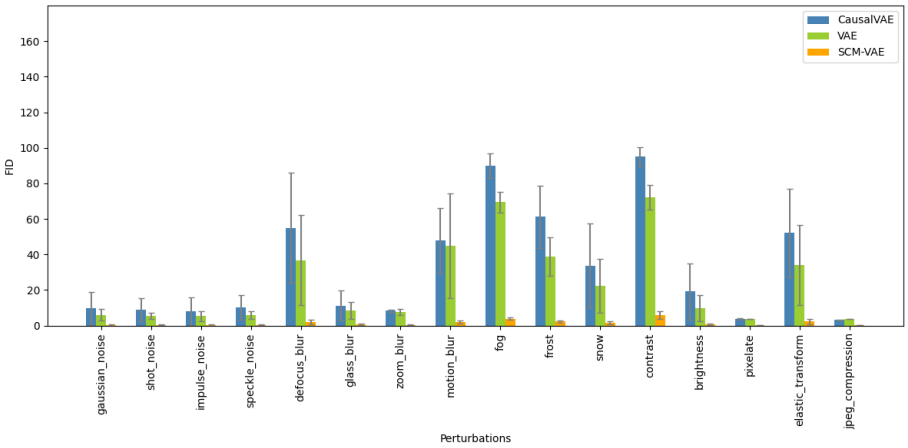
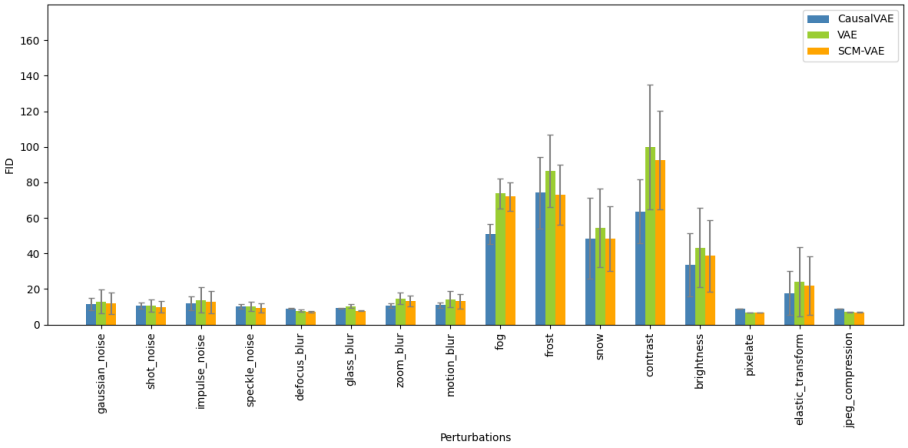


Figure 2: Robustness Evaluation Pipeline

In Fig.3 and 4, we report the mean and standard deviation of FID scores computed across the 5 severity levels for each perturbation. The results indicate that for most perturbations, traditional generative models VAE and BiGAN exhibit greater robustness than their causal counterparts CausalVAE and DEAR. This empirical observation is consistent with our theoretical analysis in the VAE case, which shows that, under careful choices of the Lipschitz constants, large values of $\lambda_1(P)$, here 1.71 can adversely affect the robustness of causal generative models. Interestingly, the SCM-VAE, which incorporates a non-linear causal layer, demonstrates improved robustness compared to both the standard and CausalVAE. This performance can be attributed to its robustness margin bound, which additionally depends on γ_1, leading in this setting to higher theoretical and empirical robustness bounds. Although CelebA(BEARD) and Pendulum share the same value of $\lambda_1(P) = 1.71$, Fig.3 and 4 demonstrate greater robustness for causal models compared to traditional ones on CelebA(BEARD). This improvement is primarily explained by the causal generative models' Lipschitz constants resulting, in this setting, to larger robustness margin bounds.

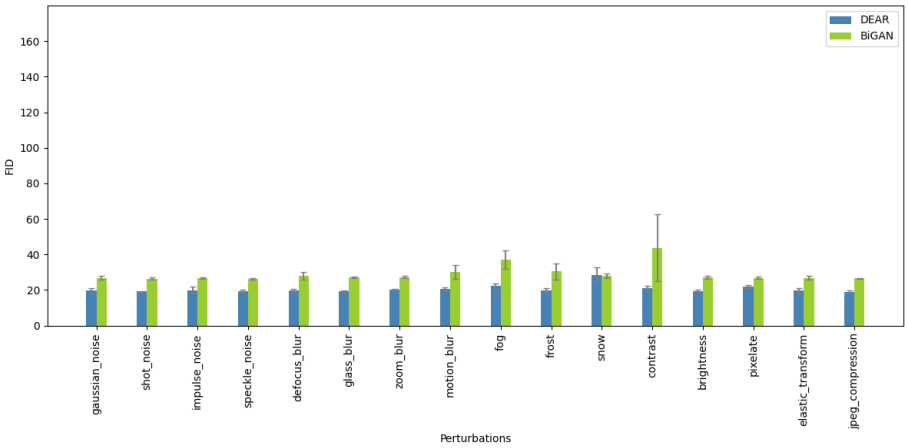


(a) Pendulum

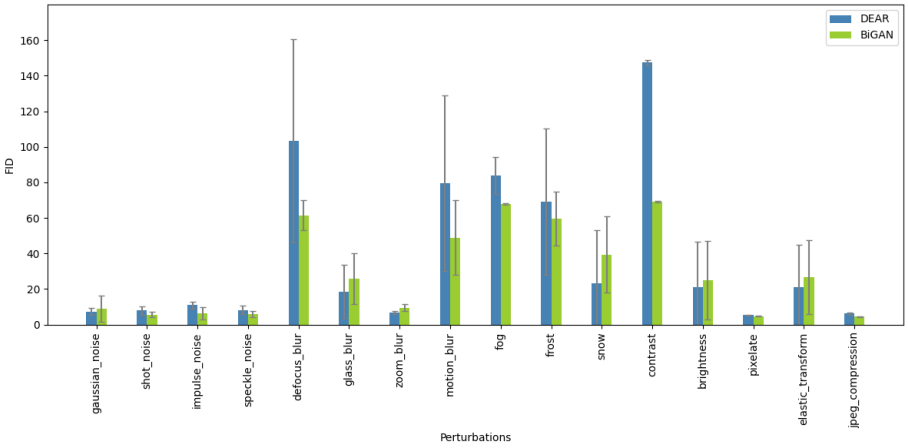


(b) CelebA(BEARD)

Figure 3: VAE based models



(a) Pendulum



(b) CelebA(BEARD)

Figure 4: GAN based models