Counterfactual Robustness: a framework to analyze the robustness of Causal Generative Models across interventions

Manal Benhamza[1] (✉), Marianne Clausel[2], and Myriam Tami[1]

[1] Paris-Saclay University, CentraleSupélec, MICS Lab, France
{manal.benhamza,myriam.tami}@centralesupelec.fr
[2] Lorraine University, Elie Cartan Institute, IECL, France
marianne.clausel@univ-lorraine.fr

Abstract. Data generation using generative models is one of the most impressive growing field of artificial intelligence. However, such models are black boxes trained on huge datasets lacking interpretability properties. Causality is a natural framework to include expert knowledge into deep generative models. Other expected beneficial properties of causal generative models are fairness, transparency and robustness of the generation process. Up to our best knowledge, while many works have analyzed general generative models' robustness, surprisingly none have focused on their causal counterpart even if their robustness is a common claim. In the present paper, we introduce the fundamental concept of counterfactual robustness, which evaluates how sensitive causal generative models are to interventions with respect to distribution shifts. Through a series of experiments on synthetic and real-life datasets, we demonstrate that all the studied causal generative models are not equal with respect to counterfactual robustness. More surprisingly, we show that all causal interventions are also not equally robust. We provide a simple explanation based on the causal mechanisms between the variables, that is theoretically grounded in the case of an extended CausalVAE. Our in-depth analysis also yields an efficient way to identify the most robust intervention based on prior knowledge on the causal graph.

Keywords: Counterfactual Robustness · Causal Representation Learning · Generative Models.

1 Introduction

Generative AI models have gained widespread recognition for their ability to model complex distributions and generate high-quality outputs [5,3]. These models, however, often lack interpretability properties, making it difficult to understand the relationships between the learned representations. Causal generative models [14] address this issue by capturing causal dependencies between the extracted latent features, assumed as latent causal factors, hence offering

enhanced transparency and interpretability [24,23]. Incorporating causal structures in generative models has furthermore enabled the generation of counterfactual data. By intervening on an extracted causal factor, we derive a new counterfactual model that can generate samples from unexplored contexts with specific attributes [2,20], providing answers to counterfactual what-if questions. The latter are commonly expressed in the form: "What will happen to the model's output when setting one of the input variables to a specific value?", and aim to analyze the impact of an intervention on the model's outcome. For example, for CelebA dataset [18], a counterfactual model with respect to the intervention "Gender=Male" is capable of generating the male counterparts of female input images, hence highlighting the facial attributes that change with the gender in the model's reconstruction. Since counterfactual models enhance the interpretability through responding "What-if" questions [12,26,27], they are leveraged in overcoming multiple AI research *challenges*. These models play a crucial role in defining *fairness* [17,28] by verifying whether a model's predictions remain consistent when only sensitive variables are altered. In *mitigating data biases* [11], counterfactuals correct imbalances by generating synthetic samples that counteract spurious correlations. They are also used in *reinforcement learning* [19] to simulate alternative actions, hence improving policy optimization and decision-making. Nonetheless, generative models are known to be vulnerable to distribution shifts [16], i.e., whereby small input perturbations induce unwanted changes in the output. Counterfactual models can therefore also be subject to this unwanted robustness limitation.

Surprisingly, many works exist on generative models' robustness [6,16,21], but to the best of our knowledge, none has ever studied causal generative models robustness across different interventions. Hence, this work provides a new theoretical framework and an experimental study addressing the stated robustness limitation. In Section 3, we introduce a new theoretical concept of counterfactual robustness, which allows to characterize the vulnerability of a causal generative model to distribution shifts across interventions. In Section 4, we conduct a series of experiments on Pendulum and CelebA datasets [29,18] to analyze the counterfactual robustness of popular observational data causal representation learning (CRL) models: CausalVAE [29], DEAR [25] and SCM-VAE [15] on different interventions. The obtained results show that counterfactual models respond differently to the considered perturbations. This observation motivated to develop in Section 5, a rigorous theoretical proof in the extended CausalVAE case. We demonstrate that the substantial difference in the counterfactual models' robustness levels to noise perturbations can be explicitly explained by the causal graph structure, more specifically by the critical properties of the removed edges with each intervention. Thus, in Section 5 we define a new theoretical concept of "Edge Robustness Score" ERS based on the adjacency matrix of the causal graph. Considering the causal graph as a prior knowledge, we propose an algorithm allowing to rank the counterfactual models' robustness to noise perturbations. The contributions of this research work are the following:

- We introduce a new theoretical concept of counterfactual robustness that evaluates the sensitivity of counterfactual causal models to distribution shifts.
- We analyze the counterfactual robustness of counterfactual models derived from popular observational data CRL models CausalVAE, SCM-VAE, and DEAR.
- We show through the conducted experimental study coupled with a theoretical analysis for an extended CausalVAE, that the counterfactual models exhibit different responses to the considered perturbations.
- Based on the causal graph, we propose a novel interpretability perspective for the differences in robustness to noise perturbations.
- We define the new theoretical concept of "Edge Robustness Score" ERS leveraging the adjacency matrix of the causal graph and establish that prior knowledge of the latter, allows the ranking of counterfactual models' robustness to noise perturbations solely based on the computation of the ERS.
- We propose a robustness score ranking algorithm based on the ERS.

2 Related Work

Generative Models Robustness [4], have introduced the so-called r-robustness to evaluate the robustness of Variational Autoencoders (VAE). This notion quantifies the robustness of a VAE reconstruction with respect to a given perturbation. In [4], r-robustness local margin bounds are provided, putting in evidence which parameters can be controlled to guarantee more robustness. Building on this result, [1] proves that it is possible to construct a VAE model with an a priori known level of robustness, based on fine control of the Lipschitz constants of the encoder and decoder. Their proposed theoretical framework focuses solely on the robustness of traditional VAEs and can not be extended to other generative models. Other works, such as [21], have studied the adversarial robustness of flow-based generative models, whereas [16] presented a method to craft adversarial attacks capable of changing different generative models' outputs. These papers thoroughly analyze the robustness of generative models. However, to the best of our knowledge, no prior work has explicitly and comprehensively examined the counterfactual robustness of causal generative models, a gap that we aim to address in this study. Notably, we provide an extensive experimental analysis for several causal generative models and a full theoretical analysis of counterfactual robustness for the extended CausalVAE.

Causal Representation Learning Generative Models Recently, CRL models have advanced quickly due to their ability to learn causal latent factors from high-dimensional data along with their underlying causal structure. These learned factors describe meaningful semantics of data and hence guarantee more interpretability and explainability. State-of-the-art CRL models assume the data-generation process to be either observational, interventional, or counterfactual. Referring to the causal generative models mapping proposed in [14], we chose to study the robustness of 3 popular observational data models CausalVAE, DEAR,

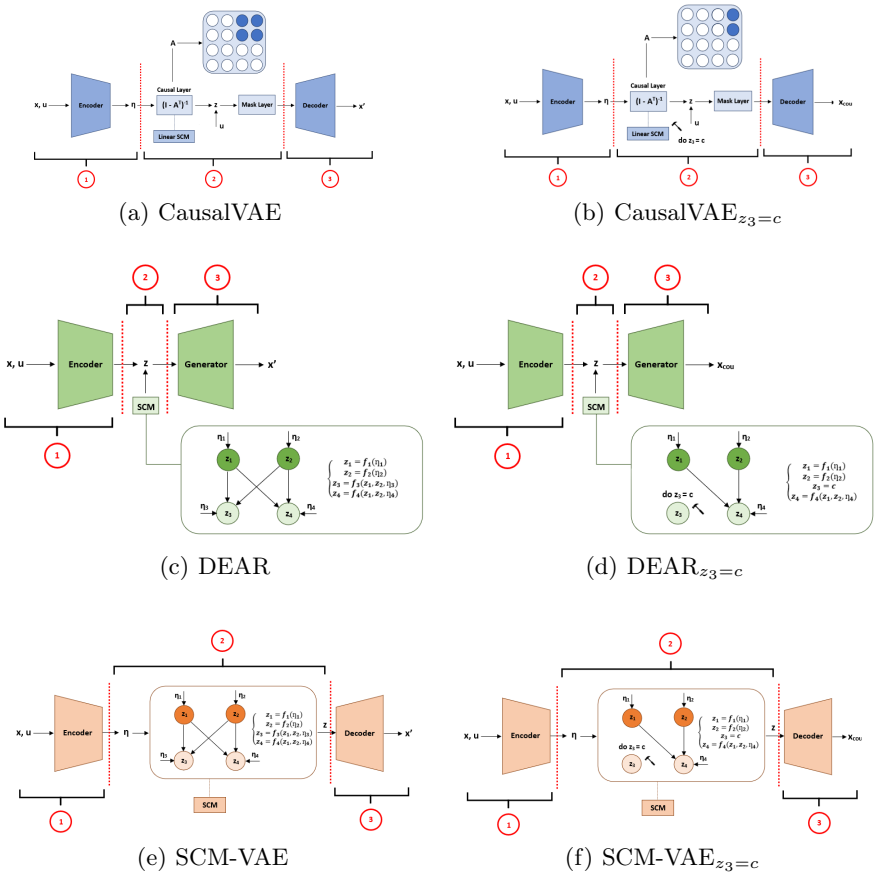
and SCM-VAE, that can both learn a latent causal representation and perform counterfactual generation. Focusing on this models paradigm is justified by its broader applicability compared to its counterparts that focus solely on controlled counterfactual generation, e.g., CausalGAN [13]. Moreover, observational data is always accessible, whereas interventional and counterfactual data are often limited or unavailable in real-world settings. This accessibility makes observational models suitable for studying robustness, as it allows for a comprehensive evaluation across diverse datasets and perturbations. CausalVAE uses a linear Structural Causal Model (SCM) parameterized by the causal adjacency matrix A, as illustrated in Fig.1(a) to transform the encoded independent latent factors η into latent causal ones z. The input labels u are used as additional information to ensure the identifiability of the model. DEAR [25], in Fig.1(c), as in Disentangled gEnerative cAusal Representation, builds a new disentangling method using an SCM as the prior distribution for a bidirectional generative model. DEAR and CausalVAE both utilize labels as weak supervision signals. However, unlike CausalVAE, DEAR does not learn intermediate independent encodings of the inputs. SCM-VAE overcomes the limitations of the CausalVAE, mainly the linear SCM, by learning a post-nonlinear additive noise SCM to describe more general relations between the causal variables, as presented in Fig.1(e). The use of a non-linear SCM in both DEAR and SCM-VAE particularly interests us, as linear SCMs are limited in capturing complex causal relationships. We also seek to investigate the impact of non-linear SCMs on the robustness of CRL models.

3 Counterfactual Robustness of Causal Generative Models to Distribution Shifts

3.1 Counterfactual Models

Given a dataset $\mathcal{X} = (x_j)_{1\leq j\leq N}$, we suppose that each observation x_j is a vector of $\mathbb{R}^{N_d}$, representing the set of observed variables. We define the latent variables as the set $\mathcal{Z} = (z_j)_{1\leq j\leq M}$ considered as causal factors. We assume that the graph of relationships between the latent variables is known. The latter is characterized by its adjacency matrix A. Both expressions, latent variables, and causal factors will hence be used interchangeably in this paper.

Each model is characterized by its so-called latent SCM, describing the functional mechanism between latent variables $S := (E, P^\eta)$, where $E = (E_1, \ldots, E_M)$ is a collection of M equations of the form $E_j : z_j = f_j(\mathrm{PA}_j, \eta_j)$, with $\mathrm{PA}_j \subseteq \{z_1, \ldots, z_M\} \setminus \{z_j\}$. The variables in the subset PA_j are called parents of z_j. We denote $P^\eta = P^{\eta_1,\ldots,\eta_M}$ the joint distribution of the noise variables, that are supposed to be mutually independent. Let G be a CRL model with latent SCM S. We provide a description of the latent SCM of each model in Block 2 of Fig.1(a), 1(c), and 1(e). Let *do* be the operator that performs hard interventions on the causal latent factors by replacing one or more structural equations in E with a constant c. This process results in a new SCM, denoted $\tilde{S}$. For the considered



(a) CausalVAE (b) CausalVAE$_{z_3=c}$

(c) DEAR (d) DEAR$_{z_3=c}$

(e) SCM-VAE (f) SCM-VAE$_{z_3=c}$

Fig. 1: Frameworks of the considered CRL models and their derived counterfactual models. In all Figures, Block 1 refers to the encoding process, Block 2 encloses the causal mechanisms, and Block 3 holds the decoding process. We shall note that both CausalVAE and SCM-VAE models consider the output of the encoder as noise variables for their latent SCM in Block 2.

models, the new latent SCMs $\tilde{S}$ with respect to the intervention $do_{z_3=c}$ are represented in Block 2 of Fig.1(b), 1(d) and 1(f). We call a counterfactual model derived from the CRL model G with respect to the intervention $do_{z_j=c}$, the model $G_{z_j=c}$ that generates counterfactual samples $\mathbf{x}_{\text{cou}}$ in Block 3 of Fig.1(b), 1(d) and 1(f), by intervening on the causal factor z_j in the SCM fixing its value to c. In this work, we only consider counterfactual models intervening on a single causal latent variable.

Each $G_{z_j=c}$ is responsible for generating counterfactual samples with specific attributes that correspond to the intervention $do_{z_j=c}$. The intervening process of $G_{z_j=c}$ depends on the structure of G. For example, to obtain CausalVAE$_{z_j=c}$ in Fig.1(b), the interventions are performed on the linear latent SCM breaking specific causal relations and hence modifying the matrix A. The intervened factors are then passed through the Mask Layer to propagate the effect of the parent variables to the children variables. The Mask Layer yields a final latent causal representation, which is used by the decoder in Block 3 to generate $\mathbf{x}_{\text{cou}}$. As for DEAR$_{z_j=c}$ in Fig.1(d), $\mathbf{x}_{\text{cou}}$ are obtained by passing through the generator, Block 3, the latent factors sampled from the interventional SCM $\tilde{S}$. SCM-VAE$_{z_j=c}$ also yields $\mathbf{x}_{\text{cou}}$ by sampling from $\tilde{S}$ and then passing samples through the decoder Block 3 as explained in Fig.1(f).

3.2 Counterfactual Robustness

We refer to a model's G reconstruction of an observation x_k, as $G(x_k)$ and the counterfactual reconstruction as $G_{z_j=c}(x_k)$. We test the robustness of each model, considering several perturbations of the dataset $\mathcal{X}$ detailed in Section 4. The perturbed dataset is denoted $\mathcal{X}^* := (x_j^*)_{1\leq j\leq N}$. We give a general definition of the counterfactual robustness that we shall instantiate in the experimental section later.

Definition 1. *A causal generative model is said to be counterfactually γ-robust to a perturbation $*$ with respect to the intervention $do_{z_j=c}$ and a similarity measure SIM if:*

$$SIM\left(\{G_{z_j=c}(\mathcal{X}^*)\}, \{G_{z_j=c}(\mathcal{X})\}\right) \geq \gamma \tag{1}$$

where, SIM is a similarity measure that evaluates how similar the distributions of the two considered datasets are in terms of features.

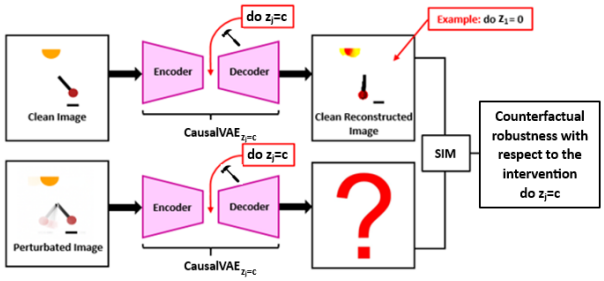


Fig. 2: Counterfactual Robustness Evaluation Pipeline exemplified for CausalVAE$_{z_j=c}$ on the Pendulum Dataset

Fig.2 shows the pipeline we propose for assessing the counterfactual robustness, exemplified in the case of CausalVAE$_{z_j=c}$ on the Pendulum dataset. We

start by fixing an intervention variable z_j and a value c, we perform the intervention on G and then recover $G_{z_j=c}$. In this example, z_1 is the Pendulum Angle and $c = 0$, i.e., we enforce the pendulum angle to be 0. We also choose a distribution shift perturbation $*$, apply it to our dataset to obtain a perturbed dataset, and afterward, pass the datasets of clean and perturbed images through $G_{z_1=0}$. We compute SIM in Eq.1 between the two datasets of reconstructed images with and without perturbation and thereafter provide a measure of the robustness of $G_{z_1=0}$ with respect to the considered perturbation.

4 Experiments

Here, we evaluate the robustness of counterfactual models derived from 3 observational data CRL models CausalVAE, DEAR, and SCM-VAE on two annotated synthetic and real-world datasets, Pendulum and CelebA for 16 common perturbations. We follow the pipeline described in Fig.2. To approximate real-world scenarios where perturbations occur at different intensities [22], we define 5 severity levels for each considered image corruption.

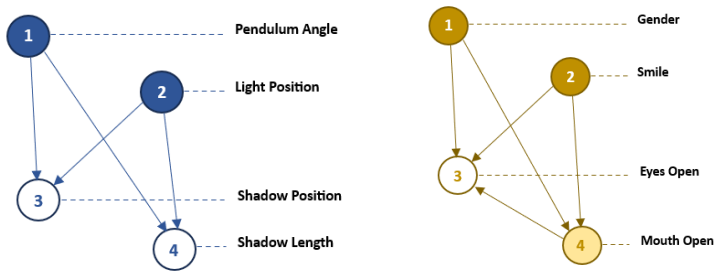
4.1 Datasets and Models

Datasets

- **Pendulum** [29] is a synthetic dataset that contains four causal variables: Pendulum Angle, Light Position, Shadow Position, and Shadow Length. It simulates the dynamic behavior of a pendulum and its interaction with a light source, capturing how the motion of the pendulum affects the position and length of its shadow. Its causal graph is presented in Fig.3(a). The counterfactual models for Pendulum are generated respectively by intervening on the Pendulum Angle, Light Position, Shadow Position, or Shadow Length.

- **CelebA** [18] is a popular resource in the computer vision community. This dataset contains 200k images of human faces, each labeled with various attributes. In the literature, we consider two subsets of causally related attributes for this dataset. The first subset, CelebA(SMILE), includes: Gender, Smile, Eyes Open, and Mouth Open. The second subset CelebA(BEARD) consists of: Age, Gender, Bald, and Beard. We choose to work with CelebA-(SMILE) to have the causal graph in Fig.3(b), different in its structure from the Pendulum graph. In this setting, the counterfactual models are respectively obtained by intervening on the causal variables Gender, Smile, Eyes Open, or Mouth Open.

Datasets Perturbations To simulate data distribution shifts on the datasets Pendulum and CelebA(SMILE), we inject the 16 perturbations proposed by [8] for ImageNet-C. The proposed perturbations belong to five main categories:



(a) Pendulum (b) CelebA(SMILE)

Fig. 3: Causal Graphs of the Pendulum and CelebA(SMILE) datasets, highlighting the causal relationships between the considered causal variables. For example, in CelebA the Smile influences the eyes and mouth openness.

noise, blur, weather, and digital. Each of the latter contains several techniques: (1) Noise: gaussian noise, shot noise, impulse noise, speckle noise; (2) Blur: defocus blur, frosted glass blur, motion blur, zoom blur; (3) Weather: snow, frost, fog, brightness; (4) Digital: contrast, elastic, pixelated, JPEG compression. Since distribution shifts in the real world happen with different intensities, we use 5 severity levels for each perturbation technique following [7]. The number of input perturbations, taking into account their varying intensities sum to 80. The crafted corruptions for Pendulum are illustrated in Fig.4.

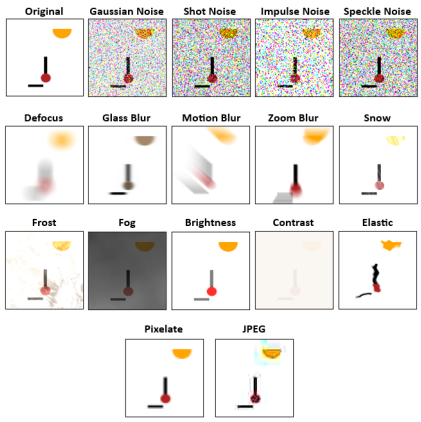


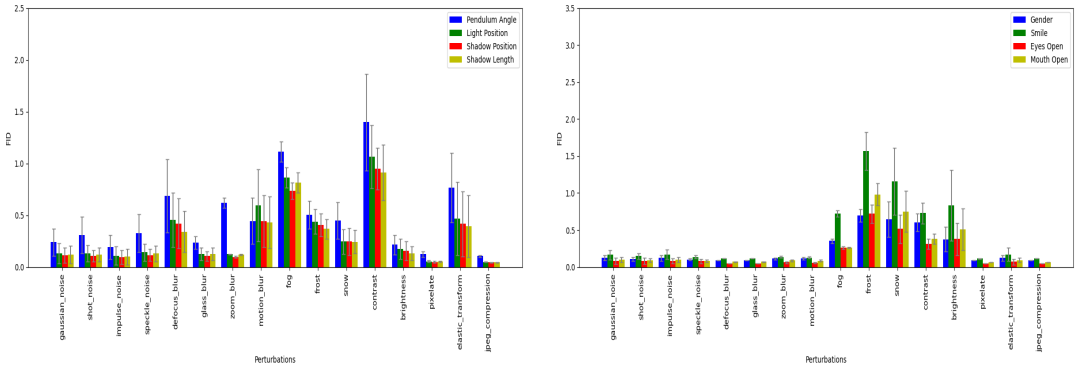
Fig. 4: Examples of 16 image perturbations. The original image (top left) is taken from the Pendulum dataset.

Models We derive counterfactual models from two trained versions of **Causal-VAE**, **DEAR**, and **SCM-VAE**, one for each dataset, using the default baseline parameters and architectures from [29], [25], and [15] respectively.

4.2 Results

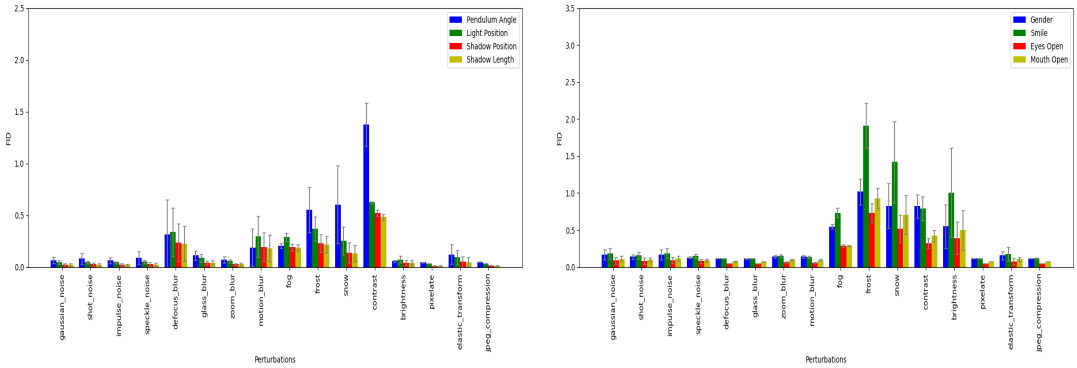
To compare the robustness of counterfactual models derived from CausalVAE, DEAR, and SCM-VAE, we use the Frechet Inception Distance (FID) as a similarity evaluation metric between the clean and perturbed reconstructed datasets for each intervention. FID was first introduced by [9]. The latter evaluates how similar the distributions of two datasets are in terms of features extracted by the pre-trained Inceptionv3 model. Low FID values indicate high similarity between the evaluated datasets. In this work, the Inceptionv3 model was fine-tuned on the considered datasets Pendulum and CelebA to capture their unique patterns.

We report in Fig.5, 6 and 7 the mean and standard deviation of the FID between the clean and perturbed images over all perturbations, on both datasets Pendulum and CelebA. The lower the FID score, the more robust the counterfactual model. The counterfactual models were obtained by performing interventions on individual causal variables, setting their values to 0 and 0.8, respectively, for the Pendulum and CelebA(SMILE) datasets. Note that an intervention value of 0 is nonsense for the CelebA(SMILE) dataset. The figure scales differ between the two datasets for visualization purposes and Appendix C explores additional intervention values.



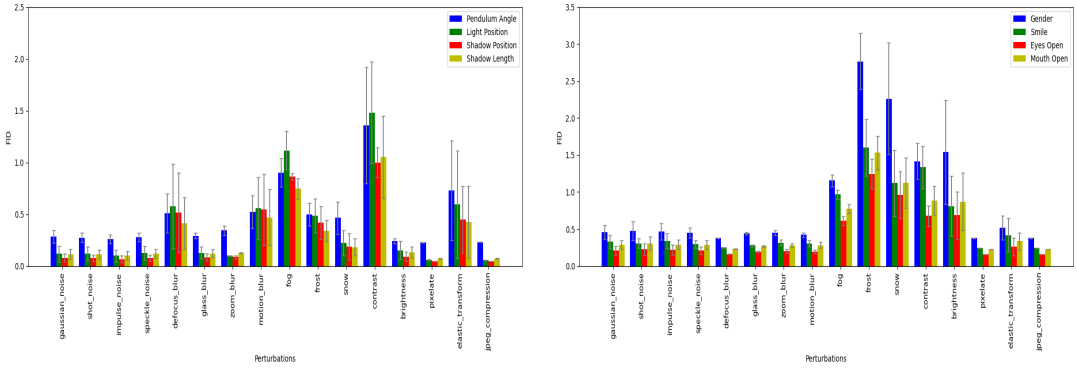
(a) Pendulum (b) CelebA(SMILE)

Fig. 5: CausalVAE



(a) Pendulum (b) CelebA(SMILE)

Fig. 6: DEAR



(a) Pendulum (b) CelebA(SMILE)

Fig. 7: SCM-VAE

We can observe from Fig.5, 6, and 7 that counterfactual models obtained from the same CRL model are not equally impacted by the considered perturbations. For Pendulum, the contrast perturbation is the one affecting the counterfactual models the most, whereas it is frost for CelebA(SMILE). The JPEG compression has the least impact on both datasets. Moreover, it is to be noted that the counterfactual models derived by intervening on the exogenous causal variable Pendulum Angle for the Pendulum dataset, are generally the least robust to image perturbations. Counterfactuals derived from DEAR are more robust than those of the other considered CRL models CausalVAE and SCM-VAE. We shall note that the obtained FID are scaled by the spatial importance of the intervened causal variable to ensure a fair comparison across different interventions,

taking into account the intrinsic significance of each causal variable in terms of picture space. The latter is calculated for complex attributes like Gender and Smile, by a decomposition following the semantic attribute graphs available in Appendix D.

Note also that for all types of **noise perturbations** and all CRL models, the counterfactual models derived by intervening on the endogeneous variables representing the causal effect of an exogeneous causal variable, e.g., Shadow Position and Shadow Length, are more robust than those obtained through interventions on the exogenous causal variables, e.g., Pendulum Angle and Light Position, for the Pendulum dataset. Whereas for CelebA(SMILE), intervening on the endogenous variable, Eyes Open yields the most robust counterfactual models for all analyzed CRL models and interventions on the exogenous causal variables Gender and Smile are associated with the least robust counterfactual models to noise perturbations.

The latter observation provided insight into the possibility of interpreting the difference in counterfactual robustness levels to noise perturbations based on the dataset's causal mechanisms, more specifically, by leveraging the causal structure of the causal graph and its related adjacency matrix. Since each intervention on the causal variables implies the removal of specific edges in the causal graph, then the counterfactual robustness is necessarily tied with the properties of the removed edges, as we will show in Section 5. In this part of the paper, we introduce an ERS metric and demonstrate through experimental results and theoretical proof in the case of an extended CausalVAE, that the most robust counterfactual model is the one for which an intervention removes the edges with the highest cumulative ERSs. Moreover, we propose an algorithm to identify the most robust intervention to noise perturbation based on a prior knowledge of the causal graph.

5 Causal Graph Edge Robustness Score

5.1 Motivation of the definition

An intervention on a causal system $do_{z_j=c}$ implies analytically fixing the causal variable z_j to be equal to the intervention value c. Whereas graphically, it signifies removing all the incoming edges, in the causal graph, to the node z_j since the latter no longer depends on its causal parents and its value is rather determined by the intervention. The removal of edges eliminates all causal pathways that include them. Pathways are sequences of nodes and edges, and therefore removing an edge causes the connection to be cut in the pathways where it belongs. Hence, for a given CRL and intervention value, a derived counterfactual model can be characterized by the sets of edges and paths that are removed from the causal graph by its corresponding intervention.

Having different robustness scores for counterfactual models, as observed in Fig.5, 6, and 7, implies thus that the causal graph edges are not equally robust. The computation of their robustness scores will hence be of great importance in understanding the robustness of counterfactual models solely by leveraging the causal graph. We therefore propose in Subsection 5.2 to define an edge robustness score ERS based on the paths in which an edge is included and the singular vectors of $(I - A^T)^{-1}$. Existing works [10] have analyzed the centrality scores of edges in social networks based on different indicators. The centrality score of an edge is determined by the proportion of walks or paths that traverse it or by the amount of information it conveys. The latter centrality metrics give answers to questions related to the frequency with which information flows through an edge, the duration it takes, and the path multiplicity to reach the target node. No measure, however, was designed to describe the robustness of the edges.

Our proposed ERS enables to interpret the robustness of counterfactual models to noise perturbations based on the structure of the causal graph. We particularly show that the most robust counterfactual model is the one for which the removed edges have the highest cumulative ERS.

5.2 Theoretical Insights

Let $\mathcal{G}$ be a causal graph and A its corresponding adjacency matrix.

Definition 2 (Edge Robustness Score). *Given an edge $e =< s_e, t_e >$ in a causal graph $\mathcal{G}$, represented by its source node and target node respectively, noted s_e and t_e, the edge robustness score is defined as the sum of the products of the eigenvector scores at both ends of the paths containing the edge, weighted by the path intensity:*

$$ERS_e := \sum_{p \in P_e} w_1(s_p) y_1(t_p) Int(p) \tag{2}$$

where A is the adjacency matrix of $\mathcal{G}$, w_1 and y_1 are respectively the right and left singular vectors of $(I - A^T)^{-1}$ corresponding to the largest singular value $\lambda_1((I - A^T)^{-1})$, $Int(p)$ is the scalar representing the path intensity, given by the product of the weights along the path edges, P_e is the set of paths where e is included. s_p and e_p denote respectively the source and target nodes of the path p and $w_1(s_p)$ is the s_p^{th} coordinate of w_1, $y_1(t_p)$ the t_p^{th} coordinate of y_1.

Fig.8 and 9 illustrate the edge robustness scores, respectively, for the Pendulum and CelebA(SMILE) datasets. The considered order of the variables is the same as in the causal graphs Fig.3(a) and 3(b). Non existing edges in the causal graph are affected a robustness score of 0. Fig.8 suggests, in the case of Pendulum, that the edges linking the exogenous causal variables, Pendulum Angle and Light Position, to the endogenous ones, Shadow Position and Shadow Length are equally robust. For CelebA(SMILE), Fig.9 indicates that there are three levels of edge robustness. Edges linking the cause variable Gender to its effects Eyes Open and Mouth Open exhibit low robustness highlighted in light

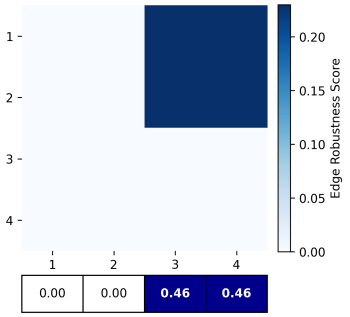


Fig. 8: Pendulum Edge Robustness Scores

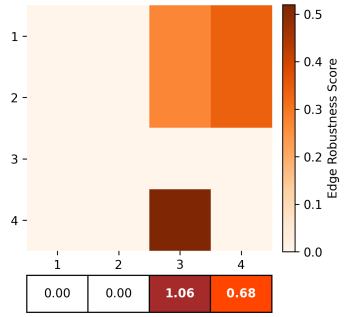


Fig. 9: CelebA Edge Robustness Scores

orange, the ones in dark orange linking Smile to the same effects Eyes Open and Mouth Open display moderate robustness, and only one edge in dark brown from Mouth Open to Eyes Open demonstrates high robustness.

This notion of edge robustness is important, because the ordering of the cumulative ERS for each intervention, aligns with the ranking of the counterfactual robustness experimentally observed in Section 4. Specifically, we display below Fig.8 and 9, the sum over the columns of the edge robustness scores. The sum over column j corresponds to the cumulative ERS of the edges removed when intervening on the variable z_j. This provides a structural justification for the observed differences in counterfactual robustness, as interventions that remove edges with higher cumulative robustness scores yield more robust counterfactual models.

5.3 An extensive theoretical analysis in the case of an extended CausalVAE

We denote extended CausalVAE, a VAE based causal model where the causal layer implements the general non linear SCM [30] in Eq.3:

$$z = \pi_1((I - A^T)^{-1}\pi_2(\eta)) \tag{3}$$

where A is the adjacency matrix of the causal graph corresponding to the causal variables $\mathcal{Z}$, π_1 and π_2 are element-wise transformations generally nonlinear.
By incorporating a parametric SCM in the causal layer, the extended CausalVAE framework can encompass both the linear CausalVAE and the SCM-VAE models, depending on the choice of the functions π_1 and π_2. In the linear CausalVAE case, both π_1 and π_2 are identity functions, while in the SCM-VAE setting, π_2 remains the identity function and π_1 is a non-linear transformation learned

by a neural network.

In Appendix A, building on [4] we express leveraging the structure of an extended CausalVAE and the Lipschitz continuity of its components, the counterfactual robustness probability, and the margin bounds of the counterfactual models against adversarial attacks. These attacks involve generating imperceptible noise that, when applied to the inputs of an extended CausalVAE counterfactual models, result in unintended reconstructions. The obtained counterfactual robustness margin bounds with respect to the intervention $do_{z_j=c}$ are expressed as a function of $\lambda_1((I-(A^j)^T)^{-1})$, the largest singular value of $(I-(A^j)^T)^{-1}$, where A^j corresponds to the adjacency matrix of the new causal graph after the intervention $do_{z_j=c}$. We show in Appendix A that the lower $\lambda_1((I-(A^j)^T)^{-1})$, the larger the counterfactual robustness margin bounds. Th.1 provides, based on the proposed ERS, an approximation of $\lambda_1((I-(A^j)^T)^{-1})$. We show, using first order matrix perturbation theory, that removing edges with the highest cumulative ERS effectively reduces the largest singular value $\lambda_1((I-(A^j)^T)^{-1})$.

Theorem 1 (Edge Removal Impact on Singular Values). *Let $\mathcal{G}$ be a causal graph and A its corresponding adjacency matrix. Let $\widehat{\lambda_1}$ be the first singular value of $(I-\widehat{A}^T)^{-1}$, where $\widehat{A}$ is the adjacency matrix of the perturbed causal graph $\widehat{\mathcal{G}}$, obtained by removing the edges indexed by the set $\mathcal{E}$ from $\mathcal{G}$; such that $(I-\widehat{A}^T)^{-1} = (I-A^T)^{-1}+Q$. Let α be the singular value gap, w_1 and y_1 be respectively the right and left singular vectors of $(I-A^T)^{-1}$. If λ_1, $\widehat{\lambda_1}$ are respectively the first singular values of $(I-A^T)^{-1}$, $(I-\widehat{A}^T)^{-1}$ and $\alpha \geq 2\|Q\|$, then:*

$$\lambda_1 - \widehat{\lambda_1} \simeq \sum_{e\in\mathcal{E}} \sum_{p\in P_e} w_1(s_p) y_1(t_p) Int(p) \tag{4}$$

$$\simeq \sum_{e\in\mathcal{E}} ERS_e \tag{5}$$

where P_e is the set of paths in which each edge e is included, s_p, t_p are respectively the source and target nodes of each path $p \in P_e$ and $Int(p)$ the intensity of the path p.

The quality of the approximation $\lambda_1 - \widehat{\lambda_1}$ in Th. 1 depends on the singular value gap α and the Frobenius norm of Q. To evaluate the quality of the proposed approximator on real causal graphs, we measure the linear correlation between the sets of real and approximated singular values, considering different sparsity levels l ($l = 0.2, 0.4, 0.6, 0.8, 1$) and graph sizes m ($m = 4, 6, 8, 10, 20$). We report in Tab. 1 and 2 the mean and standard deviation of the obtained correlation results on 300 randomely simulated causal graph for each setting. We put $-$ whenever the considered sparsity level leads to isolated nodes in the causal graph.

Table 1: Approximation Quality (Interventions on all the variables)

	$l = 0.2$	$l = 0.4$	$l = 0.6$	$l = 0.8$	$l = 1$
$m = 4$	–	–	–	0.989 ± 0.017	0.979 ± 0.014
$m = 6$	–	0.979 ± 0.038	0.977 ± 0.019	0.969 ± 0.017	0.957 ± 0.018
$m = 8$	–	0.973 ± 0.025	0.971 ± 0.017	0.957 ± 0.018	0.943 ± 0.018
$m = 10$	–	0.973 ± 0.017	0.962 ± 0.018	0.948 ± 0.019	0.931 ± 0.019
$m = 20$	0.972 ± 0.047	0.973 ± 0.018	0.96 ± 0.018	0.947 ± 0.021	0.943 ± 0.018

Table 2: Approximation Quality (Interventions on the $(m-2)$ first variables)

	$l = 0.2$	$l = 0.4$	$l = 0.6$	$l = 0.8$	$l = 1$
$m = 4$	–	–	–	0.989 ± 0.017	0.979 ± 0.014
$m = 6$	–	0.984 ± 0.016	0.988 ± 0.023	0.994 ± 0.012	0.998 ± 0.004
$m = 8$	–	0.98 ± 0.029	0.988 ± 0.015	0.994 ± 0.009	0.995 ± 0.007
$m = 10$	–	0.986 ± 0.016	0.988 ± 0.016	0.992 ± 0.011	0.994 ± 0.006
$m = 20$	0.985 ± 0.016	0.989 ± 0.015	0.99 ± 0.01	0.992 ± 0.012	0.995 ± 0.005

It can be seen that the proposed approximator is good for ranking the first singular values of the perturbed matrices $(I - (A^j)^T)^{-1}$, since correlation values are all greater than 0.94. It is also to be noted that for dense causal graphs of size m, the approximator is better in ranking the first singular values corresponding to the $m-2$ first interventions (correlation often near 1). This is due to $\|Q\|$ being large for interventions on the last variables in dense causal graphs, which degrades the quality of the approximation.

In general, Th.1 provides a good approximation of how the first singular value of $(I - A^T)^{-1}$ changes when removing a set of edges in the causal graph, as a function of the cumulative ERS of the removed edges. As confirmed by Tab.3 and 4, the interventions that remove the edges with the highest cumulative ERS are the ones that effectively reduce the largest singular value of the perturbed matrix $(I - \widehat{A}^T)^{-1}$. The latter interventions are thus the ones associated with the counterfactual models with the highest counterfactual robustness margin bounds.

5.4 Algorithm

Based on Subsections 5.2 and 5.3, we propose an Algorithm 1 to rank the robustness of counterfactual models derived from a CRL using a unique intervention value.

Table 3: Pendulum Intervention Edge Robustness and Largest Singular Values $\widehat{\lambda_1}$

Causal Variable	Edges Robustness	$\widehat{\lambda_1}$
Pendulum Angle	0	1.71
Light Position	0	1.71
Shadow Position	0.46	1.51
Shadow Length	0.46	1.51

Table 4: CelebA(SMILE) Intervention Edge Robustness and Largest Singular Values $\widehat{\lambda_1}$

Causal Variable	Edges Robustness	$\widehat{\lambda_1}$
Gender	0	2.23
Smile	0	2.23
Eyes Open	1.06	1.51
Mouth Open	0.68	1.64

Algorithm 1 Counterfactual Robustness Ranking

Input: Adjacency matrix A of the causal graph $\mathcal{G}$
Output: Ascending order sorting of the counterfactual robustness
Initialize the cumulative ERS vector: $ranking_i \leftarrow 0$
Compute the leading eigenvalue λ_1 of $(I - A^T)^{-1}$, let w_1 and y_1 be respectively the right and left singular vectors
for $j = 1$ **to** M (Considering all possible intervention variables) **do**
  for Edge e in parent edges of z_j, the intervened variable **do**
    Identify the paths P_e including edge e
    for path (s_p, t_p) in P_e **do**
      Compute the intensity of the path $Int(p)$
      $ranking_i \leftarrow ranking_i + w_1(s_p)y_1(t_p)Int(p)$
    end for
  end for
end for
Sort in ascending order the cumulative ERS vector $ranking$
return Return the indexes of the sorting

6 Conclusion

This paper introduces a novel theoretical framework for evaluating the counterfactual robustness of causal generative models under distribution shifts. Through extensive experiments on the Pendulum and CelebA datasets, we demonstrate that the studied counterfactual models exhibit varying degrees of robustness to perturbations, which can be explained by the causal graph structure. To formalize this relationship, we define the edge robustness score, a theoretical measure leveraging the causal adjacency matrix. Our findings reveal that intervening on variables with the highest cumulative robustness scores for their incoming edges yields the most robust counterfactual models. Future research could explore counterfactual models that intervene simultaneously on multiple variables at a time and extend our theoretical work to learning based perturbations for other CRL models like DEAR.

Acknowledgments. This work was performed using HPC resources from the "Mésocentre" computing center of CentraleSupélec and Ecole Normale Supérieure Paris-Saclay supported by CNRS and Région Ile-de-France. This work was partially supported by the ANR CLearDeep ANR-23-CE23-0008-01.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Barrett, B., Camuto, A., Willetts, M., Rainforth, T.: Certifiably Robust Variational Autoencoders . In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. pp. 3663–3683 (2022)
2. Besserve, M., Sun, R., Janzing, D., Schölkopf, B.: A theory of Independent Mechanisms for Extrapolation in Generative Models. Proceedings of the AAAI Conference on Artificial Intelligence (2020)
3. Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G.: Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
4. Camuto, A., Willetts, M., Roberts, S., Holmes, C., Rainforth, T.: Towards a Theoretical Understanding of the Robustness of Variational Autoencoders. pp. 3565–3573 (2021)
5. Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z.: A Survey on Generative Diffusion Models. IEEE Transactions on Knowledge and Data Engineering **vol 36**, 2814–2830 (2024)
6. Cui, X., Aparcedo, A., Jang, Y.K., Lim, S.N.: 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. pp. 24625–24634 (2024)
7. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F., Brendel, W.: ImageNet-trained CNNs are biased Towards Texture. ArXiv (2018)
8. Hendrycks, D., Dietterich, T.: Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. Proceedings of the International Conference on Learning Representations (2019)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Advances in neural information processing systems **vol 30** (2017)
10. Huang, X., Huang, W.: Eigenedge: A Measure of Edge Centrality for Big Graph Exploration. Journal of Computer Languages (2019)
11. Hutchinson, B., Denton, E., Mitchell, M., Gebru, T.: Detecting Bias with Generative Counterfactual Face Attribute Augmentation (2019)
12. Hvilshøj, F., Iosifidis, A., Assent, I.: ECINN: Efficient Counterfactuals from Invertible Neural Networks. In: 32nd British Machine Vision Conference Virtual (2021)
13. Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S.: CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In: International Conference on Learning Representations (2018)

14. Komanduri, A., Wu, X., Wu, Y., Chen, F.: From Identifiable Causal Representations to Controllable Counterfactual Generation: A Survey on Causal Generative Modeling. Transactions on Machine Learning Research (2024)
15. Komanduri, A., Wu, Y., Huang, W., Chen, F., Wu, X.: SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge. In: IEEE International Conference on Big Data. pp. 1014–1023 (2022)
16. Kos, J., Fischer, I., Song, D.X.: Adversarial Examples for Generative Models. IEEE Security and Privacy Workshops SPW pp. 36–42 (2017)
17. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual Fairness. In: Advances in Neural Information Processing Systems. vol. vol 30 (2017)
18. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: 2015 IEEE International Conference on Computer Vision ICCV. pp. 3730–3738 (2015)
19. Lu, C., Huang, B., Wang, K., Hernández-Lobato, J.M., Zhang, K., Schölkopf, B.: Sample-Efficient Reinforcement Learning via Counterfactual-Based Data Augmentation. In: Offline Reinforcement Learning - Workshop at the 34th Conference on Neural Information Processing Systems NeurIPS (2020)
20. Melistas, T., Spyrou, N., Gkouti, N., Sanchez, P., Vlontzos, A., Panagakis, Y., Papanastasiou, G., Tsaftaris, S.A.: Benchmarking Counterfactual Image Generation. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
21. Pope, P., Balaji, Y., Feizi, S.: Adversarial Robustness of Flow-Based Generative Models. In: International Conference on Artificial Intelligence and Statistics. pp. 3795–3805 (2020)
22. Qiu, J., Zhu, Y., Shi, X., Tang, Z., Zhao, D., Li, B., Li, M.: Benchmarking Robustness under Distribution Shift of Multimodal Image-Text Models. In: NeurIPS Workshop on Distribution Shifts: Connecting Methods and Applications (2022)
23. Schölkopf, B., Von Kügelgen, J.: From Statistical to Causal Learning. In: Proceedings of the International Congress of Mathematicians. p. 1 (2022)
24. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward Causal Representation Learning. Proceedings of the IEEE **vol 109**, 612–634 (2021)
25. Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., Zhang, T.: Weakly Supervised Disentangled Generative Causal Representation Learning. Journal of Machine Learning Research pp. 1–55 (2022)
26. Van Looveren, A., Klaise, J.: Interpretable Counterfactual Explanations Guided by Prototypes. In: Machine Learning and Knowledge Discovery in Databases Research Track. pp. 650–665 (2021)
27. Verma, S., Dickerson, J.P., Hines, K.: Counterfactual Explanations for Machine Learning: Challenges Revisited. ArXiv (2021)
28. Wang, Z., Chu, Z., Blanco, R., Chen, Z., Chen, S.C., Zhang, W.: Advancing Graph Counterfactual Fairness Through Fair Representation Learning. In: Machine Learning and Knowledge Discovery in Databases Research Track. pp. 40–58 (2024)
29. Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR pp. 9588–9597 (2020)
30. Yue Yu, Jie Chen, T.G., Yu, M.: DAG-GNN: DAG Structure Learning with Graph Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning (2019)