Generative Models for Counterfactual Explanations

Daniil Kirilenko[1,*], Pietro Barbiero[1], Martin Gjoreski[1], Mitja Luštrek[2] and Marc Langheinrich[1]

[1]Università della Svizzera italiana, Lugano, Switzerland
[2]Jožef Stefan Institute, Ljubljana, Slovenia

Abstract

Counterfactual explanations have emerged as an effective method of explaining machine learning models. These explanations elucidate how to tweak the model input in order to flip its output. Generative approaches serve as a tool for creating meaningful counterfactuals for complex problems, where other methods fail or require too much computation. This work presents an overview of generative approaches and their applications in the generation of counterfactual explanations. We highlight the prevailing challenges, such as diversity and distinction from adversarial examples, and identify open questions with future research directions, such as ensuring the stability of counterfactuals and automatic reasoning with counterfactual explanations.

Keywords

Counterfactual Explanations, Generative Models, Explainable AI

1. Introduction

Counterfactual explanations clarify complex system decisions by answering "what if" scenarios, showing how minimal input changes can lead to different outcomes [1]. This is crucial in Machine Learning (ML), where understanding the rationale of a model is as important as the decision itself [2]. By examining hypothetical alternatives, counterfactual explanations make ML models' decision-making more transparent and comprehensible.

Despite growing interest in counterfactual explanations, there is a gap in the literature on the generative methods used to create them. Variational Autoencoders (VAEs) [3], Generative Adversarial Networks (GANs) [4], and Denoising Diffusion Probabilistic Models (DDPMs) [5] are notable for generating counterfactuals, especially for complex data modalities such as images, where tweaking uninterpretable features fall short. However, existing surveys often overlook the generative aspects or high-dimensional data scenarios [6, 7, 8]. Our work addresses this gap by focusing on generative models for counterfactual explanations in complex data, offering a comprehensive understanding of their capabilities and limitations.

In this paper, we explore the common use cases of generative models for counterfactual explanations and highlight primary challenges. We categorize methods by their generative techniques and examine modifications to standard processes to meet counterfactual requirements. Our discussion aims to stimulate further research by identifying key challenges and potential directions for advancing generative methods in counterfactual explanations. While

HI-AI@KDD, Human-Interpretable AI Workshop at the KDD 2024, 26th of August 2024, Barcelona, Spain
*Corresponding author.
daniil.kirilenko@usi.ch (D. Kirilenko)

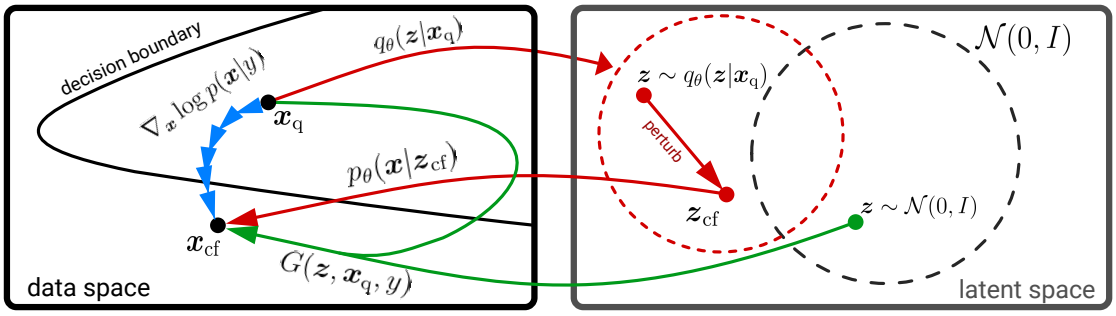

Figure 1: The figure illustrates the prevalent use cases of VAE (red), GAN (green), and DDPM (blue) in generating a counterfactual $\boldsymbol{x}_\mathrm{cf}$ from a given query sample $\boldsymbol{x}_\mathrm{q}$. The VAE approach employs an encoder to approximate $q_\phi(\boldsymbol{z}|\boldsymbol{x}_\mathrm{q})$, from which $\boldsymbol{z}$ is sampled. Subsequent modifications yield $\boldsymbol{z}_\mathrm{cf}$, and the decoder then generates $\boldsymbol{x}_\mathrm{cf}$ from $p_\theta(\boldsymbol{x}|\boldsymbol{z}_\mathrm{cf})$. In contrast, GAN leverages a generator, trained using adversarial loss, that inputs a latent noise vector $\boldsymbol{z}$, the original sample $\boldsymbol{x}_\mathrm{q}$, and a target label y to synthesize $\boldsymbol{x}_\mathrm{cf}$. DDPM integrates a trained score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ with a classifier into a conditional score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}|y)$. This function facilitates the transition from $\boldsymbol{x}_\mathrm{q}$ to $\boldsymbol{x}_\mathrm{cf}$ through iterative noise injections and denoising steps.

counterfactual generation is often seen through the lens of causal generative modeling [9], we focus on noncausal approaches.

2. Counterfactual Explanations

Counterfactual explanations are essential for explainability in machine learning [6, 1]. Formally, given a dataset X with corresponding labels Y, a counterfactual for a query sample $\boldsymbol{x}_\mathrm{q} \in X$ is an alternative sample $\boldsymbol{x}_\mathrm{cf} \in X$ that results in a different outcome $y \in Y$ under a predictive model $f : X \to Y$, providing insight into model behavior. The quality of a counterfactual explanation depends on several conditions [1]. **Proximity** and **validity** require a counterfactual to be as similar as possible to the query sample $\boldsymbol{x}_\mathrm{q}$ but lead to a different desired outcome. Meanwhile, **plausibility** and **actionability** demand that the suggested modifications be meaningful and practical to users. For example, suggesting a reduction in age is impractical, as age cannot be reversed.

Counterfactuals and adversarial examples. To further understand counterfactual explanations, it is useful to compare them with adversarial examples, since both involve modifying the original samples to change the output of a model. Counterfactual explanations and adversarial examples modify the original samples to change the output of a model but differ in their objectives. Counterfactuals introduce semantically reasonable changes to provide meaningful insights, while adversarial examples use subtle, imperceptible perturbations to mislead the model. Distinguishing between them is challenging. Thus, it is crucially important for counterfactual approaches to ensure modifications that are perceptible and semantically significant [10, 11].

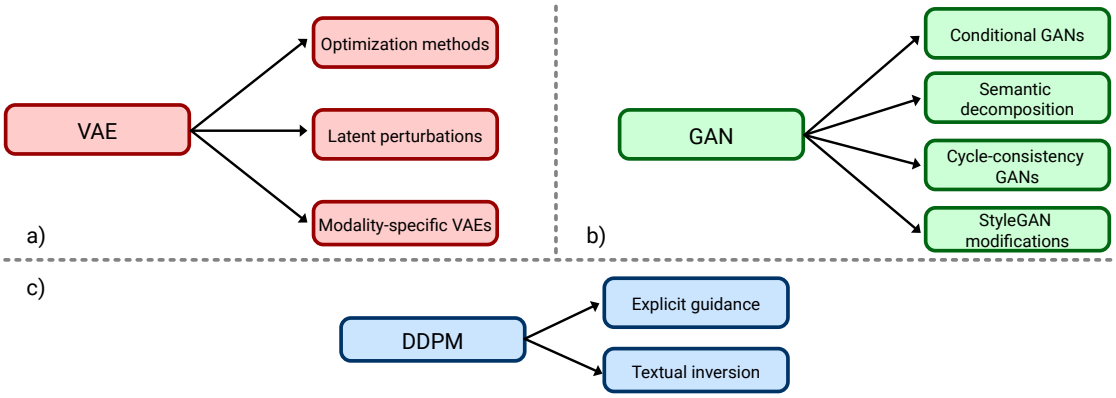

Figure 2: The taxonomy of considered approaches. We consider three of the most popular generative approaches and categorize them based on the types of modifications made to the standard sampling process to ensure counterfactual properties.

Counterfactuals and generative models. While most counterfactual explanation methods are developed for tabular data with interpretable features, extending them to high-dimensional domains like images or time-series is challenging. In tabular data, ensuring properties like validity and feasibility is straightforward. However, high-dimensional data with non-interpretable features poses significant difficulties. Generative models capable of approximating data distributions offer a promising solution. Figure 1 illustrates common uses of VAEs, GANs, and DDPMs for counterfactual generation. These models can satisfy counterfactual conditions through specific modifications and regularizations, such as incorporating distance metrics for validity or controllable generation techniques for actionability.

3. Variational Autoencoders

VAEs are useful for counterfactual generation by approximating data density $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, ensuring modifications remain within the data distribution (see the Appendix A for details). Since VAEs can produce interpretable latent factors, they are useful for counterfactual explanations. Our taxonomy for VAE-based counterfactual generation methods includes **optimization methods**, which employ optimization of an expression defining a good counterfactual while using VAE to stay in the data manifold; **latent perturbations**, which encode a sample into a latent space, modify it, and decode to obtain a counterfactual; we also mention some **modality-specific VAEs** to highlight the versatility of this approach.

Optimization methods. This is the most common approach to use density approximation for counterfactual explanations [1, 6]. They optimize an expression involving a classifier f, desired outcome y, classifier loss ℓ, and a cost function c that enforces desired properties, balanced by λ:

$$\mathcal{L}_{\text{cf}} = \ell\Big(f\big(p_\theta(\boldsymbol{x}|\boldsymbol{z})\big), y\Big) + \lambda c\big(\boldsymbol{x}_{\text{q}}, p_\theta(\boldsymbol{x}|\boldsymbol{z})\big), \quad \boldsymbol{z}_{\text{cf}} = \arg\min_{\boldsymbol{z}} \mathcal{L}_{\text{cf}} \tag{1}$$

The counterfactual explanation is derived by $\boldsymbol{x}_{\text{cf}} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\text{cf}})$. Despite using the learnable posterior $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, stochastic optimization process can result in $\boldsymbol{z}$ being outside the prior $p(\boldsymbol{z})$, making generating actionable counterfactuals a challenge.

Latent perturbations. [12] introduced a conditional VAE with factorized encoder and decoder. It uses mixture priors for clustering in the latent space. Counterfactuals are generated by small perturbations to the latent representation and reconstruction through the decoder, ensuring proximity and high-density data alignment. [13] proposed an approach to generate non-trivial, diverse explanations by varying less influential latent factors.

Modality-specific VAEs. [14] presented a framework for counterfactual explanations for graph ML models using a conditional graph VAE. It handles graph data challenges and generalizes to out-of-distribution graphs. [15] focused on anomaly detection in multivariate time-series, segmenting the latent space into general and salient components with supervised contrastive loss. Counterfactuals replace the salient component with a *healthy latent prototype*, estimated using kernel density estimation.

4. Generative Adversarial Networks

The capability of GANs to generate high-quality samples makes them useful for realistic counterfactuals. However, they face challenges with unstable training and mode collapse, limiting diversity [4, 16, 17, 18]. We categorize GAN-based approaches into four groups: **Conditional GANs**, where the generator G combines latent noise $\boldsymbol{z}$, encoded features from the original sample $\boldsymbol{x}_{\text{q}}$, and the class label y to generate counterfactuals: $\boldsymbol{x}_{\text{cf}} = G(\boldsymbol{z}, \boldsymbol{x}_{\text{q}}, y)$; **Semantic decomposition**, which involves segmenting original images into meaningful regions and treating each region individually during generation, these approaches utilize the individual editing of semantically distinct regions to enhance actionability; **Cycle-consistency GANs**, which produce a counterfactual and its reversal (as a counterfactual with respect to the counterfactual), aligning the reversal with the original sample; and **StyleGAN modifications**, which use StyleGAN modifications for detailed and high-quality counterfactuals.

Conditional GANs. [19] introduced a GAN-based approach for counterfactual generation by finding the latent encoding of a query image and using a class-conditional GAN to produce three instances: a reconstructed original, a modified image, and a change mask. The final counterfactual blends the original and modified images based on the predicted mask. [20] used typical conditional GAN training with additional constraints to ensure validity and counterfactual properties. [21] modified the GAN architecture, training the generator to output residuals instead of complete data points. [22] used an external classifier as a discriminator to improve robustness against adversarial attacks.

Semantic decomposition. [23] decomposed counterfactual generation into three components: background, foreground, and object mask, using a conditional GAN for each. [24] used

semantic maps and embeddings to generate counterfactuals. [25] combined conditional GANs with saliency maps to target specific regions for modification.

Cycle-consistency GANs. [26] used cycle-consistency loss to ensure coherence and reversibility of counterfactual changes. [27] added latent concept vectors for disentangled concept learning. [28] enforced cycle-consistency between original and counterfactual latent embeddings.

StyleGAN modifications. StyleGAN [29] uses latent style vectors for image generation. [30] combined StyleGAN vectors with classifier outputs for counterfactuals. [31] integrated a style vector with a CLIP [32] embedding to allow user-defined modifications in natural language.

5. Denoising Diffusion Probabilistic Models

DDPMs have emerged as the state-of-the-art generative approach, particularly useful for counterfactual generation (see the Appendix C for details). We classify these approaches into two main groups: **explicit guidance** and **textual inversion** methods. **Explicit guidance** methods exploit an external differentiable classifier $p(y|\boldsymbol{y})$ to replace the original score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ with a conditional one $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}|y)$ to generate specific counterfactuals. **Textual inversion** is a technique used in text-conditioned generative models, where unique trainable embeddings are assigned to images that share common concepts. These embeddings are optimized so that, when used as conditioning inputs, the generative model produces images that are similar to the original ones, effectively capturing and reproducing the shared concepts. Counterfactual generation here involves modifications of the original concept embedding.

Explicit guidance. Advances in DDPMs have led to innovative applications in counterfactual generation, such as DiME [33] and DVCE [34]. These methods use classifier-guidance during the diffusion process. Since diffusion models operate with noised samples, DVCE employs a one-step denoising approximation, while DiME uses multiple iterations. [35] emphasized the significance of techniques for handling noised samples, such as gradient cone projections and intermediate denoising steps. [36] introduced a two-stage approach using classifier feedback for initial image modifications, followed by iterative denoising with DDPM. [37] utilized the latent diffusion method from [38], known for computational efficiency by operating in a lower-dimensional latent space. The introduced *consensus guidance mechanism* filters gradients to ensure plausible counterfactual changes.
DDPMs gained wide application in medical image processing. One significant task addressed with counterfactual generation is the medical anomaly detection, which involves identifying disease-specific regions within samples. [39] and [40] used reverse diffusion with classifier guidance to generate healthy counterparts of pathological images. [41] combined DDIM [42], DDPM, and saliency maps for precisely targeted modifications. [43] explores counterfactual generation for fMRI data, using a transformer-only model for long-range dependencies and a modified diffusion sampling process for enhanced efficiency.

Textual inversion. Textual inversion [44] learns distinctive tokens for specific classes or concepts, enabling refined control over the generation process. [45] used this technique for counterfactual generation by combining learned concept tokens with additional *counterfactual shift*. [46] applied textual inversion to visual counterfactuals by learning concept embeddings and prompts for predefined objects or classes. In contrast to the previous approaches, this method implemented counterfactual modifications as transitions between concepts.

6. Challenges, open questions, and future directions

6.1. Challenges

Dealing with adversarial perturbations. All methods for counterfactual explanations vary significantly in their approach to ensuring semantically reasonable modifications. Some strategies use discriminative models that are adversarially resilient [34]. Other approaches apply changes within a structured latent space [37, 28, 13], resulting in more feasible and meaningful modifications. Another promising method involves the use of a structured latent space with representations of concepts as individual entities present in a sample [47, 48], learned with or without supervision [46, 45, 27].

Diversity of generated explanations. In high-dimensional data, generating diverse counterfactual explanations is crucial but rarely addressed [13]. Multiple plausible changes can lead to the same desired outcome, making it a complex task to consider all possible directions. However, too many diverse explanations can overwhelm users and hinder decision-making, requiring a balance between variety and interpretability. The challenge is to combine potential alterations, identify the most reasonable and relevant ones, and allow users to choose specific directions of change without being inundated. This diversity of counterfactuals, vital for comprehensive explainability, remains an open area for research and development. However, providing users with multiple diverse explanations can invoke problems related to the Rashomon Effect [49], where different potentially contradictory explanations may cause potentially contradictory interpretations of the same phenomena.

6.2. Open questions

Are counterfactuals stable? The stability of counterfactuals — ensuring minimal changes in the input sample do not lead to disproportionate changes in the output—remains a critical challenge. Research in tabular data has revealed vulnerabilities, where slight manipulations in the input can drastically alter the counterfactual [50]. This raises the question of whether generative approaches, increasingly used for counterfactual explanations, exhibit similar instability or offer more robust solutions. Understanding and improving the stability of these models is crucial for their reliability and trustworthiness in practical applications.

How to evaluate generated counterfactuals? Evaluating counterfactual explanations in AI is challenging due to a disconnect between theoretical metrics and real-world applicability. Traditional metrics, focusing on conditions such as proximity to factual instances or diversity,

do not necessarily translate to practical utility for end-users [51]. Research [52] has shown that while counterfactual explanations may satisfy theoretical criteria, their impact on user trust and understanding is inconsistent. To address this, recent approaches integrate users and domain experts into the evaluation process [34, 53]. This expert-in-the-loop methodology aligns theoretical constructs with practical realities, especially in critical areas like healthcare, providing a more comprehensive evaluation of counterfactual explanations' utility. However, a universal benchmark for comparing different methods is still missing.

6.3. Future directions

Modalities and multimodal counterfactual explanations. Much of the current research on counterfactual explanations focuses on image data. Extending these methodologies to other modalities, such as graphs, time-series, audio, and video is an important challenge. In addition, there is a gap in methods capable of handling multimodal data, which is increasingly prevalent in practical applications. Developing techniques that can generate counterfactuals across various modalities, or even within multimodal contexts, is a critical area for future exploration.

Grounding and reasoning in counterfactual explanations. The typical process of generating counterfactuals often lacks detailed explanations for why certain changes lead to specific outcomes, leaving users to interpret these changes on their own, which can lead to misunderstandings. Recent advances in Large Language Models (LLMs) with multimodal capabilities offer a promising solution to this issue [54]. Integrating advanced LLMs with counterfactual generation can enhance user comprehension by providing reasoning for suggested changes. This aligns with the concept of Evaluative AI [55]. Efforts like [56] highlight the versatility of LLMs in improving model explainability, especially with tabular data. Combining various forms of explanations, including semi-factuals [57], with multimodal LLMs conditioned on specific scenarios, could lead to more comprehensive and transparent AI systems.

7. Conclusion

This work highlights the crucial role of generative models in producing counterfactual explanations for high-dimensional data. We emphasize the need for semantically rich, intuitive explanations and robust user-centered evaluation describing existing approaches. We discussed future research directions, which include application of counterfactuals across diverse data modalities and integrating them with LLMs and other explanatory methods.

Acknowledgment

This study was funded by the projects TRUST-ME (205121L_214991), BASE (200021_182109), and XAI-PAC (PZ00P2_216405).

References

[1] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[2] R. Hoffman, T. Miller, S. T. Mueller, G. Klein, W. J. Clancey, Explaining explanation, part 4: a deep dive on deep nets, IEEE Intelligent Systems 33 (2018) 87–95.

[3] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[5] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Advances in neural information processing systems 34 (2021) 8780–8794.

[6] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: A review, arXiv preprint arXiv:2010.10596 (2020).

[7] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.

[8] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55.

[9] A. Komanduri, X. Wu, Y. Wu, F. Chen, From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling, arXiv preprint arXiv:2310.11011 (2023).

[10] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, H. Lakkaraju, Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 4574–4594.

[11] T. Freiesleben, The intriguing relation between counterfactual explanations and adversarial examples, Minds and Machines 32 (2022) 77–109.

[12] M. Pawelczyk, K. Broelemann, G. Kasneci, Learning model-agnostic counterfactual explanations for tabular data, in: Proceedings of the web conference 2020, 2020, pp. 3126–3132.

[13] P. Rodríguez, M. Caccia, A. Lacoste, L. Zamparo, I. Laradji, L. Charlin, D. Vazquez, Beyond trivial counterfactual explanations with diverse valuable explanations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1056–1065.

[14] J. Ma, R. Guo, S. Mishra, A. Zhang, J. Li, Clear: Generative counterfactual explanations on graphs, Advances in Neural Information Processing Systems 35 (2022) 25895–25907.

[15] W. Todo, M. Selmani, B. Laurent, J.-M. Loubes, Counterfactual explanation for multivariate times series using a contrastive variational autoencoder, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[16] L. Mescheder, A. Geiger, S. Nowozin, Which training methods for gans do actually converge?, in: International conference on machine learning, PMLR, 2018, pp. 3481–3490.

[17] K. Liu, W. Tang, F. Zhou, G. Qiu, Spectral regularization for combating mode collapse in

gans, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6382–6390.
[18] H. Thanh-Tung, T. Tran, Catastrophic forgetting and mode collapse in gans, in: 2020 international joint conference on neural networks (ijcnn), IEEE, 2020, pp. 1–10.
[19] P. Samangouei, A. Saeedi, L. Nakagawa, N. Silberman, Explaingan: Model explanation via decision boundary crossing transformations, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 666–681.
[20] A. Van Looveren, J. Klaise, G. Vacanti, O. Cobb, Conditional generative models for counterfactual explanations, arXiv preprint arXiv:2101.10123 (2021).
[21] D. Nemirovsky, N. Thiebaut, Y. Xu, A. Gupta, Countergan: Generating realistic counterfactuals with residual generative adversarial nets, arXiv preprint arXiv:2009.05199 (2020).
[22] R. Bischof, F. Scheidegger, M. A. Kraus, A. C. I. Malossi, Counterfactual image generation for adversarially robust and interpretable classifiers, arXiv preprint arXiv:2310.00761 (2023).
[23] A. Sauer, A. Geiger, Counterfactual generative networks, arXiv preprint arXiv:2101.06046 (2021).
[24] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, M. Cord, Steex: steering counterfactual explanations with semantics, in: European Conference on Computer Vision, Springer, 2022, pp. 387–403.
[25] A. Samadi, A. Shirian, K. Koufos, K. Debattista, M. Dianati, Safe: Saliency-aware counterfactual explanations for dnn-based automated driving systems, arXiv preprint arXiv:2307.15786 (2023).
[26] S. Singla, B. Pollack, J. Chen, K. Batmanghelich, Explanation by progressive exaggeration, arXiv preprint arXiv:1911.00483 (2019).
[27] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, R. W. Picard, Dissect: Disentangled simultaneous explanations via concept traversals, arXiv preprint arXiv:2105.15164 (2021).
[28] S. Khorram, L. Fuxin, Cycle-consistent counterfactuals by latent transformations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10203–10212.
[29] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
[30] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, et al., Explaining in style: Training a gan to explain a classifier in stylespace, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 693–702.
[31] J. Luo, Z. Wang, C. H. Wu, D. Huang, F. De la Torre, Zero-shot model diagnosis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11631–11640.
[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
[33] G. Jeanneret, L. Simon, F. Jurie, Diffusion models for counterfactual explanations, in:

Proceedings of the Asian Conference on Computer Vision, 2022, pp. 858–876.
[34] M. Augustin, V. Boreiko, F. Croce, M. Hein, Diffusion visual counterfactual explanations, Advances in Neural Information Processing Systems 35 (2022) 364–377.
[35] P. Vaeth, A. M. Fruehwald, B. Paassen, M. Gregorova, Diffusion-based visual counterfactual explanations–towards systematic quantitative evaluation, arXiv preprint arXiv:2308.06100 (2023).
[36] G. Jeanneret, L. Simon, F. Jurie, Adversarial counterfactual visual explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16425–16435.
[37] K. Farid, S. Schrodi, M. Argus, T. Brox, Latent diffusion counterfactual explanations, arXiv preprint arXiv:2310.06668 (2023).
[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
[39] J. Wolleb, F. Bieder, R. Sandkühler, P. C. Cattin, Diffusion models for medical anomaly detection, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2022, pp. 35–45.
[40] P. Sanchez, A. Kascenas, X. Liu, A. Q. O'Neil, S. A. Tsaftaris, What is healthy? generative counterfactual diffusion for lesion localization, in: MICCAI Workshop on Deep Generative Models, Springer, 2022, pp. 34–44.
[41] A. Fontanella, G. Mair, J. Wardlaw, E. Trucco, A. Storkey, Diffusion models for counterfactual generation and anomaly detection in brain images, arXiv preprint arXiv:2308.02062 (2023).
[42] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502 (2020).
[43] H. A. Bedel, T. Çukur, Dreamr: Diffusion-driven counterfactual explanation for functional mri, arXiv preprint arXiv:2307.09547 (2023).
[44] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, D. Cohen-Or, An image is worth one word: Personalizing text-to-image generation using textual inversion, arXiv preprint arXiv:2208.01618 (2022).
[45] J. Vendrow, S. Jain, L. Engstrom, A. Madry, Dataset interfaces: Diagnosing model failures using controllable counterfactual generation, arXiv preprint arXiv:2302.07865 (2023).
[46] G. Jeanneret, L. Simon, F. Jurie, Text-to-image models for counterfactual explanations: a black-box approach, arXiv preprint arXiv:2309.07944 (2023).
[47] D. Kirilenko, V. Vorobyov, A. Kovalev, A. Panov, Object-centric learning with slot mixture module, in: The Twelfth International Conference on Learning Representations, 2023.
[48] M. R. Arefin, Y. Zhang, A. Baratin, F. Locatello, I. Rish, D. Liu, K. Kawaguchi, Unsupervised concept discovery mitigates spurious correlations, arXiv preprint arXiv:2402.13368 (2024).
[49] R. Anderson, The rashomon effect and communication, Canadian Journal of Communication 41 (2016) 249–270.
[50] D. Slack, A. Hilgard, H. Lakkaraju, S. Singh, Counterfactual explanations can be manipulated, Advances in neural information processing systems 34 (2021) 62–75.
[51] E. Delaney, A. Pakrashi, D. Greene, M. T. Keane, Counterfactual explanations for misclassified images: How human and machine explanations differ, Artificial Intelligence 324

(2023) 103995.
[52] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, E. Kasneci, Towards human-centered explainable ai: A survey of user studies for model explanations, IEEE Transactions on Pattern Analysis & Machine Intelligence (2023) 1–20.
[53] S. Sankaranarayanan, T. Hartvigsen, L. Oakden-Rayner, M. Ghassemi, P. Isola, Real world relevance of generative counterfactual explanations, in: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, 2022.
[54] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, N. Duan, Visual chatgpt: Talking, drawing and editing with visual foundation models, arXiv preprint arXiv:2303.04671 (2023).
[55] T. Miller, Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 333–342.
[56] D. Slack, S. Krishna, H. Lakkaraju, S. Singh, Explaining machine learning models with interactive natural language conversations using talktomodel, Nature Machine Intelligence 5 (2023) 873–883.
[57] S. Aryal, M. T. Keane, Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai, arXiv preprint arXiv:2301.11970 (2023).

A. VAE Background

Introduced by [3], VAE has become a significant tool in deep learning for training latent variable models through variational inference. A VAE comprises an encoder and a decoder, with its primary objective typically being the minimization of the reconstruction error of data samples. The encoder, parameterized by trainable parameters ϕ, approximates a variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, where $\boldsymbol{z}$ is a latent variable with a prior distribution $p_\theta(\boldsymbol{z})$, often chosen as the standard Gaussian distribution $\mathcal{N}(0, I)$. The decoder models $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. VAEs are trained by maximizing the Evidence Lower Bound (ELBO), a variational lower bound of the exact log-likelihood:

$$\log p_\theta(\boldsymbol{x}) \geq \text{ELBO} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\big(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})\big), \tag{2}$$

where D_{KL} is Kullback–Leibler divergence.

B. GAN Background

In contrast to VAEs, GANs [4] operate on a different principle, as they do not explicitly learn the likelihood of data samples. Instead, GANs employ two neural networks: a generator G and a discriminator D. The generator G maps input noise, sampled from a distribution $p(\boldsymbol{z}) = \mathcal{N}(0, I)$, to the data space with the objective of learning the distribution of the generator p_g on the data samples. The discriminator D, on the other hand, aims to estimate the probability that a given data sample $\boldsymbol{x}$ originated from the actual data distribution p_{data}. The training of D involves distinguishing real samples drawn from p_{data} and generated samples from p_g. Concurrently, G is trained to maximize the probability of its generated samples being misclassified by D, effectively minimizing $\log(1 - D(G(\boldsymbol{z})))$. This training process sets up a minimax zero-sum game between G and D, where each network continuously improves its performance in response to the other, leading to the generation of increasingly realistic samples:

$$\min_G \max_D \Big(\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]\Big). \tag{3}$$

Compared to VAEs, GANs sample the latent code $\boldsymbol{z}$ from the same prior distribution during both training and inference, and are capable of generating more complex and high-fidelity data samples [29].

C. DDPM Background

In recent years, Denoising Diffusion Probabilistic Models (DDPMs) have solidified their position as a leading framework in generative modeling [5]. The central component of DDPMs is an iterative process where noise is incrementally added to a data sample $\boldsymbol{x}_0$, transforming it into a complete noise sample $\boldsymbol{x}_T \sim p(\boldsymbol{x}_T) = \mathcal{N}(0, I)$:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I), \tag{4}$$

where $\{\beta_1, \ldots, \beta_T\}$ is a predefined variance schedule. This is coupled with a learned reversal of this process:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\big(\mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)\big), \tag{5}$$

where $\mu_\theta(\boldsymbol{x}_t, t)$ and $\Sigma_\theta(\boldsymbol{x}_t, t)$ are predicted by models parameterized with learnable parameters θ. Since DDPMs, similar to VAEs, represent latent variable models, they are trained by optimizing the following variational lower bound:

$$\log p_\theta(\boldsymbol{x}_0) \geq -\log \frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_{0:T})}, \tag{6}$$

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{7}$$

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t). \tag{8}$$

This procedure results in training a so-called score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$. The resultant sampling procedure is executed by initial sampling of $\boldsymbol{x}_T \sim \mathcal{N}(0, I)$, followed by a sequence of $\boldsymbol{x}_{t-1} \sim p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ until the final $\boldsymbol{x}_0 \sim p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)$. This technique, involving multiple trainable denoising iterations, empowers DDPMs to produce outputs that are both highly detailed and diverse, setting them apart in the generative modeling arena.

A notable feature of DDPMs, with significant potential for counterfactual generation, is the development of a classifier guidance mechanism [5]. This approach diverges from traditional generative models, which typically necessitate explicit conditioning during the training phase. Guided diffusion introduces a paradigm where an unconditional generative model is first trained. Subsequently, this model is adapted for conditional sampling via the integration of an auxiliary classifier model by replacing $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ with

$$p_{\theta,\phi}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y) \propto p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) p_\phi(y|\boldsymbol{x}_t, t), \tag{9}$$

where $p_\phi(y|\boldsymbol{x}_t, t)$ is an external model to be explained. This strategy provides remarkable flexibility in conditional generation but introduces a pivotal challenge: the classifier must be either trained on noise-augmented samples to align with the DDPM's intermediate stages, or a denoising mechanism should be applied prior to classifier usage.