Facial Counterfactual Generation via Causal Mask-Guided Editing

Anonymous authors
Paper under double-blind review

Abstract

Generating counterfactual facial images is an important tool for interpretable machine learning, fairness analysis, and understanding the causal relationships among facial attributes. In this work, we propose a novel neuro-symbolic framework for *causal editing*, which integrates causal graph discovery, mask-guided counterfactual generation, and semantic interpretation to produce facial images that are both realistic and causally consistent. We first employ the Fast Causal Inference (FCI) algorithm to uncover latent causal relationships among facial attributes, enabling the identification of direct and indirect factors for target interventions. Using these causal graphs, we construct spatially informed masks that guide a DDPM-based generative model, ensuring that only regions relevant to the causal factors are modified. Finally, we leverage CLIP-based embeddings to provide logical, human-understandable explanations of the semantic changes in the counterfactuals. Experiments on CelebA and CelebA-HQ demonstrate that our approach produces high-fidelity counterfactuals, achieves superior performance on sparsity and realism metrics, and mitigates bias compared to state-of-the-art methods. This framework offers a principled approach to causally grounded, interpretable facial image editing.

1 Introduction

Counterfactual reasoning has become an essential tool in interpretable and responsible machine learning (Velazquez et al., 2023). By considering how changes in inputs would alter outputs, counterfactuals allow researchers and practitioners to explore “what-if” scenarios in a controlled and understandable manner (Goyal et al., 2019). In the context of facial images, the goal extends beyond explaining a model’s decision: it involves generating realistic, causally consistent counterfactual instances that reflect plausible alternative appearances or attributes. Unlike traditional counterfactual explanation methods, which aim to minimally perturb inputs to reveal a model’s decision boundary, counterfactual generation focuses on creating images that maintain fidelity to the underlying causal relationships among facial features.

Realism and causal consistency are critical in applications such as identity verification, biometric authentication, and human-computer interaction, as shown in Figure 1, an example in the visa application scenario and consequences of non-causal inference counterfactual generation. Existing image-based counterfactual methods often treat all features as independently manipulable, ignoring dependencies between attributes (Jeanneret et al., 2023; 2022). As a result, edits may produce implausible or entangled changes. For example, altering hair colour might unintentionally modify perceived age or gender, and changing gender could introduce unrelated expressions. Such inconsistencies not only reduce trust in the generated images but can also amplify dataset biases, undermining fairness in downstream tasks (Xu et al., 2020).

Several methods have been proposed for generating visual counterfactual explanations. The work in (Goyal et al., 2019) developed a method for generating visual explanations with a counterfactual perspective. Another work that emphasizes causal relation in counterfactual estimation is (Sanchez & Tsaftaris, 2022), a deep structural causal model that leverages recent advancements in diffusion models that involve iteratively sampling gradients from the marginal and conditional distributions that are part of the causal model. Peng et al. (Peng et al., 2022) simplified the explanation of model-detected face swap images by introducing au-

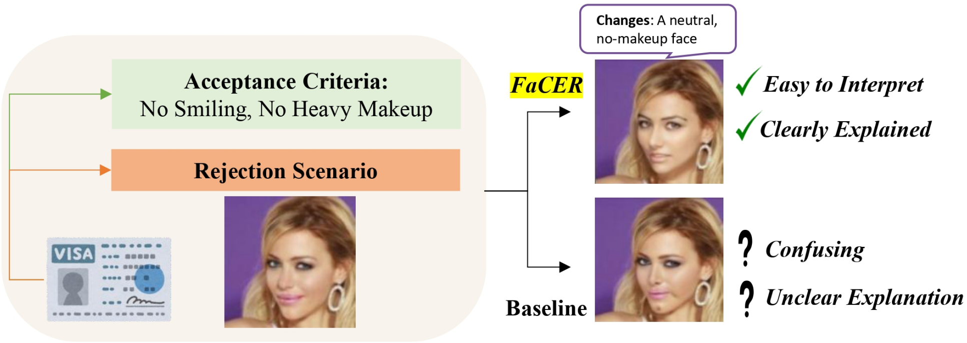

Figure 1: Motivation scenario: a rejected visa photo. Our method highlights the actual causal reasons, enhancing user understanding.

tomatic methods to create more authentic and manipulated counterfactual versions. Similarly, (Hendricks et al., 2018) introduced a phrase-critic model to enhance candidate explanations generated with inverted phrases.

However, methods that do not account for causal dependencies can yield visually unrealistic or non-actionable counterfactuals (Slack et al., 2021). De Sousa Ribeiro et al. (De Sousa Ribeiro et al., 2023) show that incorporating a causal model helps distinguish true causes from spurious associations, enabling accurate, high-fidelity counterfactual image generation. Similarly, Melistas et al. (Melistas et al., 2024) emphasize that realistic image edits must respect causal relationships inherent to the data generation process. These works demonstrate that causal models help isolate genuine causal factors and avoid spurious correlations. Existing methods face limitations that hinder the generation of causally consistent and interpretable facial counterfactuals. These gaps highlight the need for a framework that not only enforces causal consistency during counterfactual generation but also provides human-understandable explanations.

To address these challenges, we propose a framework that explicitly incorporates causal knowledge into the generation process. We introduce *Facial Counterfactual Generation via Causal Mask-Guided Editing*, a method that first learns a causal graph over facial attributes and perceptible features. This graph identifies direct causes and confounders, which then guide mask-based interventions during reverse diffusion. By modifying only causally relevant regions while holding confounders fixed, our approach ensures that generated counterfactuals are both realistic and semantically coherent. To further enhance interpretability, we leverage CLIP-based embeddings to provide textual descriptions of the visual edits, allowing users to understand which features were causally modified without implying an explanation of any specific model.

This work makes four primary contributions: (i) we shift the focus from counterfactual explanation to the causally coherent generation of facial images in a neuro-symbolic generation framework, emphasizing realistic and plausible outputs; (ii) we introduce a framework that learns causal structure over attributes and applies mask-guided reverse diffusion to intervene selectively on direct causes; (iii) we provide visual and textual outputs that narrate the applied edits, improving interpretability and user understanding; and (iv) we demonstrate both quantitative and qualitative improvements in realism, sparsity, and causal consistency compared to state-of-the-art baselines on CelebA and CelebA-HQ.

2 Literature Review

There are two primary research gaps and challenges motivating this work: first, the difficulty of accurately discovering causal relationships among facial attributes in images; and second, the prevalence of stereotypical biases in current generative models.

Causality in Facial Images Analysis The literature on causality in facial attributes has explored methods to improve model interpretability and achieve counterfactual fairness, particularly with respect to sensitive attributes such as race and gender. While prior work has addressed fairness by explicitly modeling the role of protected attributes (Mazumder & Singh, 2022), methods in facial attribute editing still face important limitations. For example, Kang et al. (Kang et al., 2022) propose generating synthetic facial replicas under factual and counterfactual assumptions using causal graph-based attribute translation, producing realistic counterfactual images. Other approaches incorporate wavelet scattering transforms (Liu et al., 2024) and channel attention for efficient face attribute classification, while NCINet (Wang & Boddeti, 2022) aims to identify causal relations from high-dimensional image representations.

Despite these advances, existing methods face several limitations that hinder their practical application for causally consistent facial counterfactual generation. First, while some approaches build causal graphs to guide attribute changes, they rarely integrate these graphs directly into the image generation process, resulting in edits that may violate causal dependencies. Second, previous methods focus solely on visual output, without providing interpretable or human-understandable explanations of the applied interventions. Third, many causal discovery techniques rely on approximations that are sensitive to noise or missing variables, limiting robustness and scalability to large, high-dimensional facial features.

Stereotyping in Generative Models Recent studies have highlighted significant concerns regarding stereotyping in image generation models, particularly in text-to-image systems. For instance, Bianchi et al. (Bianchi et al., 2023) demonstrated that widely accessible text-to-image models can amplify demographic stereotypes at large scale, producing biased images even from neutral prompts. Similarly, Barve et al. Barve et al. (2025) found that while prompt refinement can mitigate stereotypes, it often limits contextual alignment, indicating a trade-off between bias reduction and semantic accuracy. These issues underscore the need for more nuanced approaches to address bias in generative models.

However, existing literature predominantly focuses on text-to-image models, leaving a gap in understanding and mitigating bias in facial image generation. This oversight is particularly concerning given the widespread use of facial images in sensitive applications such as identity verification and biometric assessments. The lack of causal reasoning in current models often leads to unrealistic and biased facial edits, as they fail to account for the interdependencies among facial attributes.

3 Methodology

Our methodology begins with causal graph learning (Section 3.1), which uncovers the relationships among facial attributes. Next, we generate a reference counterfactual image via guided diffusion (Section 3.2) to achieve the target attribute change. The third stage applies mask-guided causal refinement using the learned graph (Section 3.3) to enforce interventions only on direct causes. Finally, we provide logical interpretations of the edits through CLIP-based semantic mapping (Section 3.4).

Table 1: Edge types in a Partial Ancestral Graph (PAG) and their interpretations.

Edge Type	**Interpretation**
(a) $V_m \rightarrow V_n$	Directed causal link from V_m to V_n (i.e., V_n is a child of V_m).
(b) $V_m {\circ\!\!\rightarrow} V_n$	Rules out V_n as an ancestor of V_m; may reflect the presence of unmeasured confounders C.
(c) $V_m \circ\!-\!\circ V_n$	Undetermined orientation; no d-separation (an unblocked path).
(d) $V_m \leftrightarrow V_n$	Suggests an unobserved confounder between V_m and V_n.

3.1 Causal Graph Learning

To generate the causal graph, let $D \in \{0,1\}^{N \times d}$ be the data matrix of binary facial attributes, where each row corresponds to an instance and each column to an attribute. Given D, we use the Fast Causal Inference (FCI) algorithm to estimate a Partial Ancestral Graph (PAG) over the d attributes. Unlike conventional

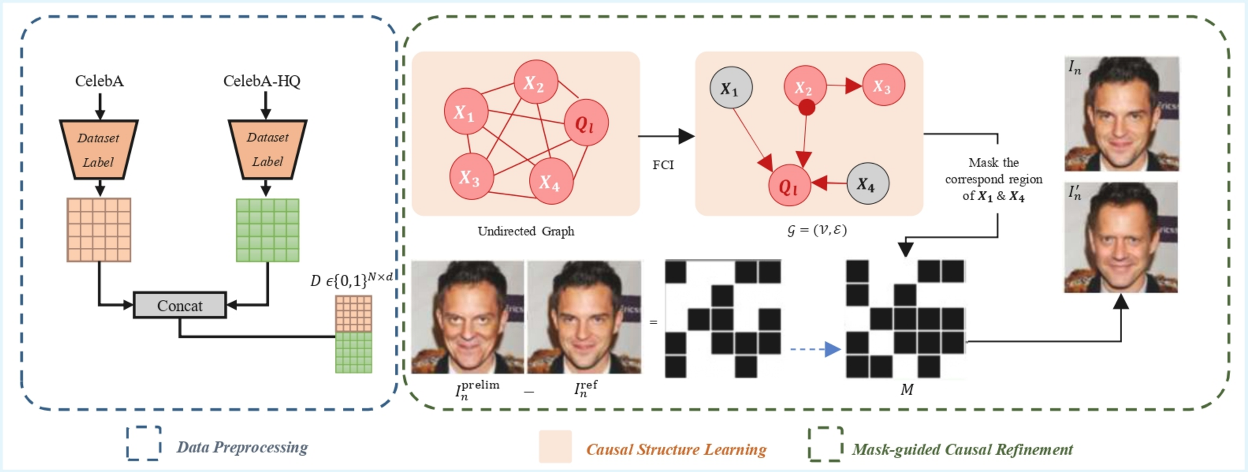

Figure 2: Our proposed causal editing counterfactual generation pipeline. We begin with data preprocessing to construct the facial attribute matrix D, which serves as input to the causal structure learning algorithm FCI. In this figure, we assume $d = 5$ and $N = 10$ to simplify the illustration and aid understanding. Using the resulting partial ancestral graph $\mathcal{G}$, we mask the corresponding region of the direct cause of the target attribute Q_l. The initial mask is obtained as a binary difference between I_n^{ref} and I_n^{prelim}. This mask is then applied in the final generation process to produce the counterfactual image I'_n.

causal graphs, PAGs encode both directed causal edges and edges with uncertain orientation, as summarised in Table 1.

Causal Graph Discovery. To ensure causally consistent counterfactual generation, we first uncover the underlying causal structure among facial attributes. We compared established causal discovery algorithms, focusing on the FCI (Spirtes, 2001). While Peter-Clark (PC) (Spirtes et al., 2000) may suggest simplistic direct links, FCI uncovers richer, more nuanced structures that capture indirect effects and potential confounding relationships, allowing for more precise and valid interventions during counterfactual generation. Ultimately, we adopted FCI for its superior handling of hidden confounders—unobserved variables that affect multiple facial traits.

Mathematically, FCI employs chi-square tests to evaluate conditional dependencies between categorical variables:

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \tag{1}$$

where O is the observed frequency of a category and E is the expected frequency under independence. Higher χ^2 values indicate stronger dependence, with significance determined via comparison to a critical value or a p-value threshold of 0.05.

Our goal is to infer the causal structure among these attributes. Formally, we seek a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{X_1, X_2, \dots, X_d\}$ is the set of nodes and $\mathcal{E}$ is the set of edges encoding causal relations.

Practical Advantages. FCI combines conceptual robustness with computational efficiency. Its use of conditional independence tests avoids the heavy resource demands of deep learning-based causal discovery methods, making it scalable for large facial attribute datasets. For a complete mathematical treatment of FCI, see (Spirtes, 2001).

By integrating FCI-based causal graphs with do-calculus interventions in subsequent counterfactual generation, our approach ensures that edits are grounded in causal mechanisms rather than superficial correlations. This foundation enhances both the interpretability and fairness of the generated counterfactual images.

3.2 Reference Image Generation via Guided Diffusion

The first stage of our counterfactual generation produces a reference image that achieves the desired target attribute change without any causal constraints. Let I_n denote the original image and y' the target attribute. We use a guided DDPM to generate a reference counterfactual I_n^{ref}:

$$x_{t-1} \sim \mathcal{N}\big(\mu_\theta(x_t, t) + \gamma \nabla_{x_t} \log p(y'|x_t), \sigma_t^2 I\big), \tag{2}$$

where $\mu_\theta(x_t, t)$ is the predicted mean from the diffusion model, σ_t^2 is the variance schedule, and γ controls the guidance strength toward the target attribute y'.

This reference image serves as a minimally perturbed counterfactual that successfully flips the target label. At this stage, edits are guided purely by the attribute change, and no causal reasoning is applied. The image may include unrelated or unrealistic changes in other regions, which are addressed in the subsequent mask-guided refinement.

3.3 Masked Guided Counterfactual Generation with Causal Graphs

After obtaining I_n^{ref}, we identify regions relevant for causal intervention by comparing it to a second-round, minimally guided counterfactual or a preliminary refinement. The difference highlights areas that changed due to the target attribute, Q_l and serves as the basis for constructing a binary mask M:

$$M_{i,j} = \begin{cases} 1, & \text{if } |I_n^{\text{ref}}(i,j) - I_n^{\text{prelim}}(i,j)| > \lambda \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Here, λ is a threshold controlling sensitivity to changes. The mask indicates candidate regions for intervention and is subsequently refined using the learned causal graph, ensuring that only direct causes are modified while confounders remain fixed. In our framework, type (a) edges, as stated in Table 1, representing direct causal relations, are prioritised for intervention, whereas edges with circles or bidirectional marks are treated as uncertain or potentially confounded and are not directly manipulated.

The mask is applied as a *do-calculus* operation within the DDPM generative process:

$$I_n' = G\big(do(I_n, M)\big), \tag{4}$$

where G is the DDPM and $do(I_n, M)$ denotes intervening on I_n along causally relevant regions defined by M. This strategy ensures that the final counterfactual I_n' is both realistic and causally consistent.

3.4 Logical Interpretation

To provide a logical and human-understandable interpretation of the generated counterfactuals, we leverage CLIP (Radford et al., 2021) to map visual changes into semantic descriptions, as illustrated in Figure 3. Let I_n denote the original image and I_n' its counterfactual counterpart, and let $\mathcal{T}$ denote the space of textual descriptions. CLIP provides encoders E_I and E_T that embed images and texts into a shared d-dimensional space, enabling cross-modal reasoning.

We first compute the normalised image embeddings:

$$e = E_I(I_n), \quad e' = E_I(I_n'), \tag{5}$$

and define the *semantic transformation vector* as

$$\Delta e = e' - e. \tag{6}$$

This vector captures the overall change induced by the counterfactual transformation in the semantic embedding space.

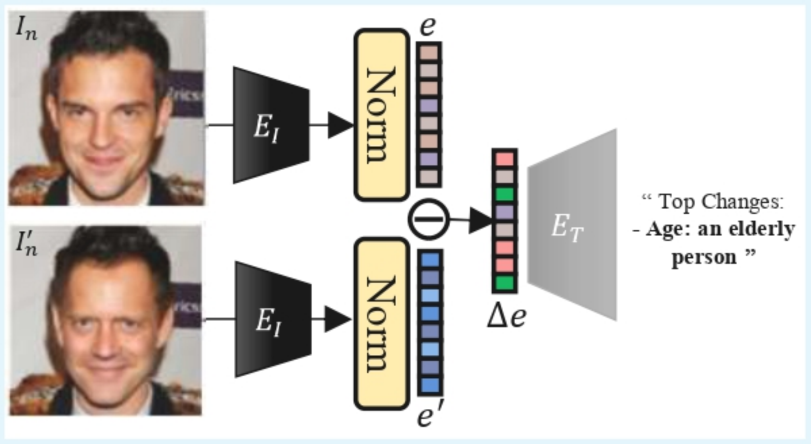

Figure 3: Our proposed counterfactual-to-text interpretation model with CLIP inspired image and text encoder.

To interpret these changes logically, we compare Δe with a curated set of text embeddings $\{E_T(d_i)\}_{i=1}^{n}$, where each $d_i \in \mathcal{D} \subset \mathcal{T}$ represents an interpretable facial attribute or semantic concept. The relevance of each description is quantified using cosine similarity:

$$S(d_i) = \frac{\Delta e \cdot E_T(d_i)}{\|\Delta e\|_2 \cdot \|E_T(d_i)\|_2}. \tag{7}$$

Descriptions with high alignment are selected as the logical explanations for the visual changes:

$$\mathcal{D}^* = \{d_i \in \mathcal{D} \mid S(d_i) > \tau\}, \tag{8}$$

where τ is a similarity threshold.

This logical interpretation framework allows us to translate causal manipulations in the image into human-readable semantic statements, providing transparent insights into which facial attributes were modified and how these changes relate causally to the target attributes. By connecting visual counterfactuals to linguistic concepts, we bridge the gap between low-level image transformations and high-level reasoning about attribute causality.

4 Experiments

We conducted all our experiments on two large facial datasets to evaluate the performance of our model. Comparison is conducted with the current state-of-the-art counterfactual explanations model. Our experimental pipeline was implemented using PyTorch 2.0.1, with computations accelerated by NVIDIA A-100 GPUs on a SLURM-managed computing cluster using CUDA 11.7. The diffusion process is performed with 500 steps, with timestep respacing to 50, a stochastic sampling fraction of 0.1, and an inpainting ratio of 0.15. The causal discovery component of our work leveraged the `causal-learn` library (Zheng et al., 2023), a comprehensive Python package that implements both classical and cutting-edge causal discovery algorithms.

4.1 Datasets

Our work draws strength from the diversity and scale of the **CelebA** dataset (Liu et al., 2015), which is a collection featuring over 200,000 facial images with 40 precisely annotated attributes. To ensure our findings generalise across different image quality conditions, we also conducted experiments on **CelebA-HQ** (Lee et al., 2020), which is a high-resolution counterpart featuring 30,000 meticulously detailed images of 1,000 celebrities. These two datasets together provide a sufficiently diverse and large sample, covering a wide range of facial variations, attribute combinations, and image resolutions, making them well-suited for robust evaluation of our counterfactual image generation method.

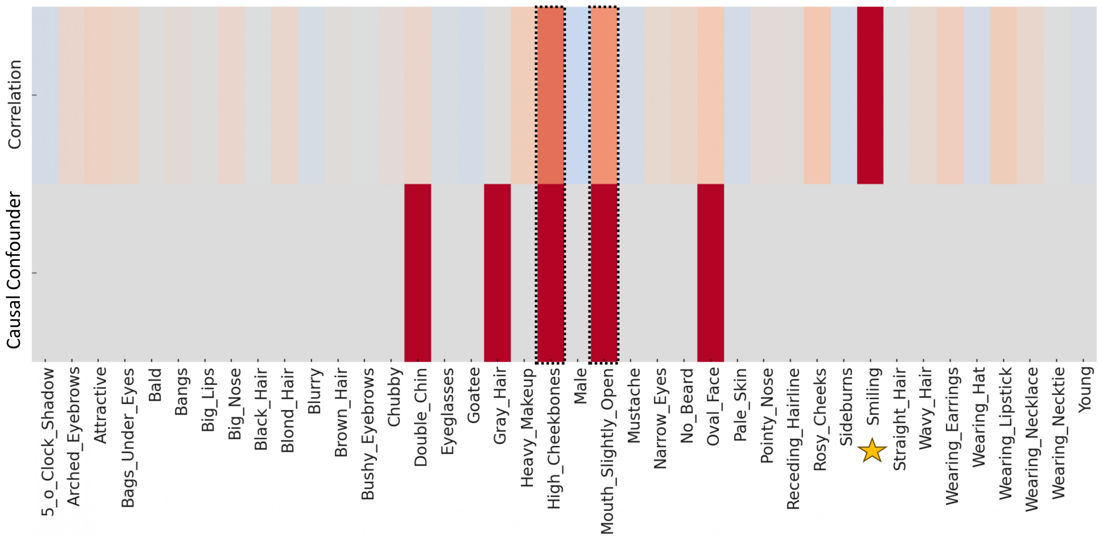

Figure 4: The heatmap visualises correlations vs confounders in causal relationships of attribute “Smile”. It (dashed box) highlights the conflict between highly correlated attributes and the confounders the causal algorithm identifies. The conflict is that while causal algorithms correctly identify confounders as indirect causes, traditional correlation-based models mistakenly learn these confounders as direct predictors simply because they are correlated with the target attribute.

4.2 Evaluation Metrics

Realism of Generated Counterfactuals: Our evaluation methodology employs three complementary metrics, each capturing different aspects of counterfactual quality. **Fréchet Inception Distance (FID)** serves as our primary gauge of generated image quality, with lower scores indicating better quality and closer resemblance to real images. Let μ_r, Σ_r be the mean and covariance of features from real images, and μ_g, Σ_g be the corresponding statistics for generated images. FID is computed as:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big). \tag{9}$$

We supplement this with **Scaled Fréchet Inception Distance (sFID)**, which adjusts the feature distributions to better capture differences across attribute scales:

$$\text{sFID} = \frac{1}{d}\sum_{k=1}^{d} \frac{(\mu_r^k - \mu_g^k)^2 + (\sigma_r^k - \sigma_g^k)^2}{(\sigma_r^k)^2 + \epsilon}, \tag{10}$$

where d is the feature dimension, μ^k and σ^k are the mean and standard deviation of the k-th feature, and ϵ is a small constant for numerical stability.

Sparsity in Counterfactuals: Additionally, we introduce **Mean Number of Attributes Changed (MNAC)** to quantify how many facial attributes are modified in the counterfactual generation process. Let x denote the original image and $\hat{x}$ the generated counterfactual, with corresponding binary attribute vectors $A(x)$ and $A(\hat{x})$. MNAC is defined as:

$$\text{MNAC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d} \mathbf{1}\big(A_j(x_i) \neq A_j(\hat{x}_i)\big), \tag{11}$$

where N is the total number of test images, d is the number of attributes, and $\mathbf{1}(\cdot)$ is the indicator function. Counterfactual explanations should exhibit minimal visual variations from the original image while achieving the desired change in target attributes.

	CelebA						CelebAHQ					
	Smile			Age			Smile			Age		
Method	FID	sFID	MNAC	FID	sFID	MNAC	FID	sFID	MNAC	FID	sFID	MNAC
STEEX (Jacob et al., 2022)	10.2	-	4.11	11.8	-	3.44	21.9	-	5.27	26.8	-	5.63
DiVE (Rodríguez et al., 2021)	29.4	-	-	33.8	-	4.58	107.0	-	7.41	107.5	-	6.76
DiME (Jeanneret et al., 2022)	3.17	4.89	3.72	4.15	5.89	3.13	18.1	27.7	2.63	18.7	27.8	2.10
ACE ℓ_1(Jeanneret et al., 2023)	1.27	3.97	2.94	1.45	4.12	3.20	3.21	20.2	1.56	5.31	21.7	1.53
ACE ℓ_2(Jeanneret et al., 2023)	1.90	4.56	2.77	2.08	4.62	2.94	6.93	22.0	1.87	16.4	28.2	1.92
Ours	**1.08**	**3.79**	**1.63**	1.51	**4.10**	**1.92**	**1.27**	**5.12**	**1.55**	12.28	**18.83**	**1.06**

Table 2: Quantitative Results for the CelebA and CelebAHQ datasets.

Notably, we deliberately avoid using correlation-based metrics such as Correlation Difference (CD) because our method is specifically designed to account for causal relationships rather than mere correlations between facial attributes, as illustrated in Figure 4. While Face Similarity (FS) and Face Verification Accuracy (FVA) are widely used in counterfactual tasks, they are not the primary focus of our work, which aims to generate realistic counterfactual explanations, and these metrics' limitations are also stated in existing work (Jeanneret et al., 2023). Nonetheless, we note that our approach exhibits qualitatively strong performance on both FS and FVA.

4.3 Baseline Methods

We compared our model with several baselines.

1. **Adversarial Counterfactual visual Explanations (ACE)** (Jeanneret et al., 2023) leverages adversarial attacks on a diffusion model to refine counterfactual images, focusing on generating subtle yet effective modifications that change the target attribute while maintaining realism.

2. **Steering Counterfactual Explanations with Semantics (STEEX)** (Jacob et al., 2022) generates counterfactuals by steering the latent representations of images in a pretrained generative model, guided by semantic attribute changes. It allows controlled modifications of specific attributes while preserving other visual details.

3. **Diffusion Model Counterfactual Explanations (DiME)** (Jeanneret et al., 2022) uses diffusion-based generative models to produce realistic counterfactual images through iterative denoising of latent representations conditioned on target attributes. This approach enables smooth transitions between original and modified attribute states.

4. **Counterfactual Explanations with Diverse Valuable Explanations (DiVE)** (Rodríguez et al., 2021) focuses on generating a diverse set of counterfactual explanations that are both visually plausible and informative. It employs a diversity-promoting objective to ensure multiple counterfactuals for a given input capture different plausible modifications, providing richer insights into model behaviour.

These baselines cover a wide spectrum of counterfactual generation paradigms, from adversarial refinement and latent-space steering to diffusion-based synthesis and diversity-driven explanations, making them suitable benchmarks for evaluating the performance of our proposed method.

4.4 Results and Discussions

We presented and discussed both the causal graph results as illustrated in Fig. 5, 6 and the generated counterfactual explanations images quantitative and qualitative evaluation are concluded in Table 2 and 3.

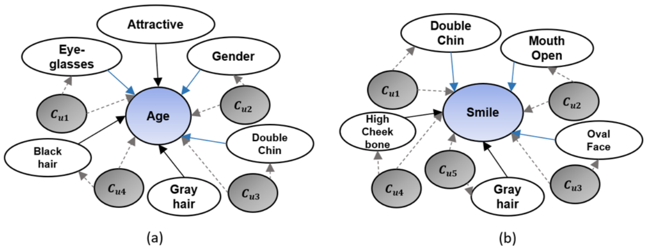

Figure 5: Partial Ancestral Graph (PAG) output for the query label "young" and "smile" as children nodes. The label in this graph represents the dataset attribute, indicating the attribute class type without specifying its actual binary class value in the CelebA dataset. The blue node is the target label Q_t, the grey node indicates the confounding variable C, and the white node is the input variables I related to the Q_t.

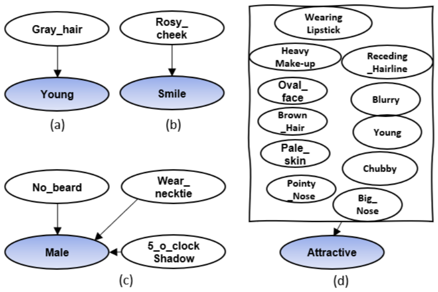

Figure 6: CelebA dataset causal relation of the target variable (in blue) with the PC algorithm. Note that the label in this graph indicates the attribute class type without specifying its actual binary class value in CelebA.

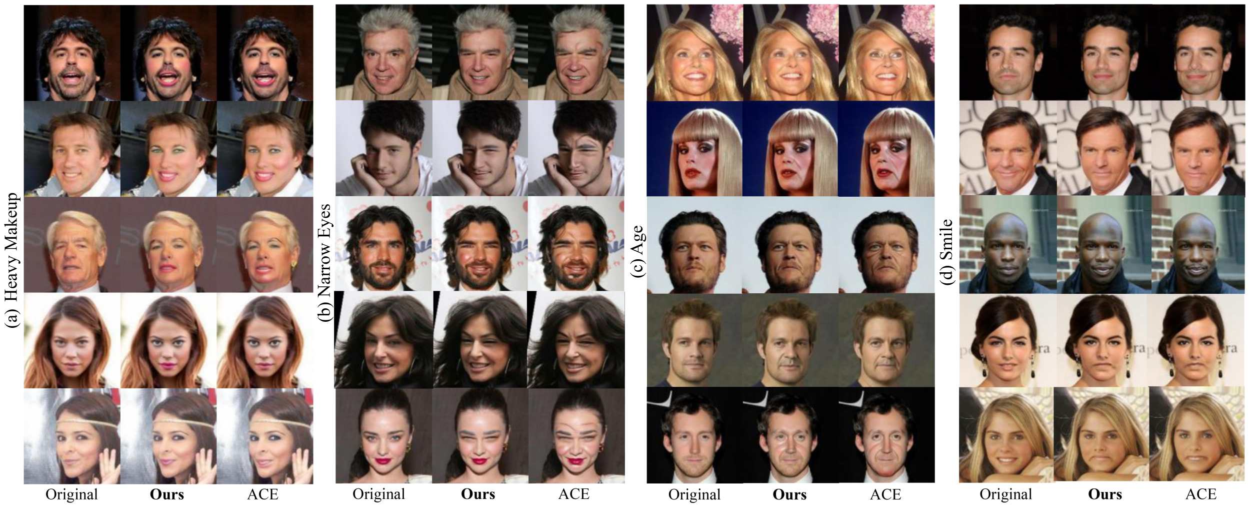

Figure 7: Qualitative counterfactuals results for facial attributes demonstrate the effectiveness of our approach. For "Age", the baseline often generates features correlated with aging, such as eye bags and glasses. However, our causal graph analysis from Figure 5 indicates that "Bags Under Eyes" does not directly influence "Age". For "Smile", our method successfully disentangles correlated attributes like "High Cheekbones" and "Opening Mouth", providing more accurate and realistic counterfactual explanations. For additional examples and higher-resolution images, please refer to the supplementary material. Zoom for best view.

4.4.1 Causal Graph Analysis

The learned causal structures are illustrated in Figure 5 (PAG) and Figure 6 (DAG obtained via the PC algorithm). For the query labels "Young" and "Smile" as child nodes: For "Young," the graph includes a direct edge from "Gray Hair" to "Young," while "Eyeglasses," "Black Hair," and "Double Chin" appear as indirect (upstream) factors. Because "Black Hair" and "Gray Hair" are mutually exclusive and may be confounded with age-related appearance, we avoid direct interventions on these attributes when orientation is uncertain. Accordingly, our masking targets regions associated with "Eyeglasses" and "Double Chin." For "Smile," the graph indicates upstream influences from "Double Chin," "High Cheekbones," and "Gray Hair." We therefore apply masks to the corresponding regions when guiding the generation process, while avoiding edits along edges marked as uncertain or confounded in the PAG.

Attr.	Gender	H. Makeup	N. Eyes
	MNAC (↓)		
ACE	3.95	5.18	3.44
Ours	**3.91**	**5.18**	**2.84**
	sFID (↓)		
ACE	7.56	11.29	**4.04**
Ours	**7.42**	**10.98**	4.44

Table 3: Quantitative Assessment for CelebA. H. Makeup and N. Eyes are heavy makeup and narrow eyes attributes, respectively. These additional attributes are chosen because they are usually related or indirectly related to sensitive attributes such as gender and race, which are not addressed in the attributes of the datasets.

4.4.2 Quantitative Results

As shown in Table 2, our model consistently outperforms prior methods across both CelebA and CelebAHQ datasets. It achieves the lowest FID and sFID scores in most cases, especially for the "smile" attribute, indicating superior image quality and more realistic feature preservation. Moreover, ours obtains significantly better MNAC scores, highlighting enhanced diversity and causal disentanglement in the generated images. Our analysis reveals that the classifier often relies on spurious correlations learned during training rather than causal relationships, resulting in lower MNAC values for these attributes, a limitation also acknowledged by DiME (Jeanneret et al., 2022).

4.4.3 Qualitative Results

Figure 7 illustrates this, showing how the baseline's counterfactuals for "young" often include ageing cues like eye bags and glasses. Similarly, confounding influences are evident for "smile", with high cheekbones and open mouths being incorrectly emphasised. In contrast, our ageing effects appear more realistic. Also, ours generates images with more noticeable heavy makeup than the baseline. In contrast, our approach successfully corrected the classification by generating impactful counterfactuals. Our method integrates causal knowledge to produce realistic, less biased counterfactual images, revealing the classifier's decision process. This enhances interpretability and utility in real-world applications where human understanding is crucial. For example, the lines generated by baseline ACE in Figure 7 for age may be confusing to the general public, as they represent ageing effects.

We present qualitative results for the "Heavy Makeup" and "Narrow Eyes" attributes in Figure 7 (a) and (b), with additional bias evaluations in the supplementary materials. These attributes, often linked to sensitive factors like gender and race, are not explicitly labeled in the dataset. Although not considered in prior work, we find them important for uncovering potential biases and enhancing the fairness of counterfactual explanations.

4.4.4 Fairness Evaluation

Table 4: Quantitative bias metrics comparing ACE and Ours method. For each attribute pair, the first attribute represents the protected group and the second represents the unprivileged group. Higher DI (Disparate Impact) and lower SPD (Statistical Parity Difference) values indicate reduced bias.

CelebA		DI (↑)	SPD (↓)
Age-HighCheek	ACE	2.3715	0.019
	Ours	**2.5213**	**0.011**
Age-H.Makeup	ACE	6.2260	0.159
	Ours	**6.2350**	**0.152**
Gender-H.Makeup	ACE	0.0189	0.016
	Ours	**0.0196**	**0.015**

Table 4 presents bias metrics for different attribute pairs evaluated on the CelebA dataset. We used a ResNeXt-50 classifier trained for 35 epochs to validate attribute intensity in generated counterfactual images. Each row examines a specific attribute pair where the first attribute represents the protected group (e.g., Age, Gender) and the second denotes the unprivileged group (e.g., HighCheek, H.Makeup).

We evaluate fairness using two widely-adopted metrics: **Disparate Impact (DI)** and **Statistical Parity Difference (SPD)**.

Disparate Impact (DI): DI measures the ratio of positive outcomes between the unprivileged and protected groups. It is defined as:

$$\text{DI} = \frac{\Pr(\hat{Y} = 1 \mid A = \text{unprivileged})}{\Pr(\hat{Y} = 1 \mid A = \text{protected})}, \tag{12}$$

where $\hat{Y}$ denotes the predicted attribute, and A represents the group membership. A DI value closer to 1 indicates equitable treatment, with higher values suggesting reduced bias against the unprivileged group.

Statistical Parity Difference (SPD): SPD quantifies the difference in positive prediction rates between the protected and unprivileged groups:

$$\text{SPD} = \Pr(\hat{Y} = 1 \mid A = \text{protected}) - \Pr(\hat{Y} = 1 \mid A = \text{unprivileged}). \tag{13}$$

Lower absolute values of SPD indicate smaller disparities, with SPD = 0 representing perfect statistical parity.

Our method consistently achieves lower SPD values and comparable or higher DI values across all evaluated attribute pairs compared to the ACE baseline. This demonstrates that our approach effectively mitigates bias by reducing prediction disparities between different demographic groups while maintaining equitable performance across attributes in counterfactual explanations. In other words, generated counterfactuals using our method exhibit a fairer distribution of attribute changes, ensuring that the protected and unprivileged groups are treated more equitably.

4.4.5 Logical Interpretation

To quantitatively assess the semantic alignment between images and their corresponding text descriptions, we report two evaluation metrics (Radford et al., 2021; Faghri et al., 2018): *Recall@1* (R@1) and *mean similarity*. R@1 measures the proportion of instances where the correct target category is ranked highest in similarity, while mean similarity denotes the average CLIP similarity score between each image and its descriptive caption. These results (Table 5) demonstrate that CLIP-based similarity metrics effectively quantify semantic alignment across modalities, confirming that the visual modifications introduced by our counterfactual edits are coherent with their intended textual descriptions.

Table 5: CLIP-based semantic alignment evaluation.

Category	**Mean Sim.**	**R@1 (%)**
Smile	5.91	72.95
Age	5.03	61.34

5 Conclusion

We introduced a framework for causal editing of facial images that unifies causal discovery, mask-guided counterfactual generation, and semantic interpretation. Motivated by the need for interpretable and causally grounded visual reasoning, our approach employs the Fast Causal Inference (FCI) algorithm to uncover relationships among facial attributes and uses spatially informed masks to guide a diffusion-based generator, ensuring that only causally relevant regions are modified. Experiments on CelebA and CelebA-HQ show that our method produces realistic, identity-preserving counterfactuals and surpasses state-of-the-art baselines on sFID and MNAC metrics. The integration of CLIP-based semantic explanations further enhances interpretability and supports fairness analysis. While our framework assumes reasonably accurate causal graphs, which may not always hold in noisy real-world settings, it provides a principled foundation for interpretable generative modeling. Future work will focus on handling uncertain causal structures, and we envision adapting our framework beyond facial imagery to broader visual domains such as medical or scene understanding tasks.

References

Saharsh Barve, Andy Mao, Jiayue Melissa Shi, Prerna Juneja, and Koustuv Saha. Can we debias social stereotypes in ai-generated images? examining text-to-image outputs and user perceptions. *arXiv preprint arXiv:2505.20692*, 2025.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 1493–1504, 2023.

Fabio De Sousa Ribeiro, Benoît Ding, Grzegorz Kruszewski, and Julia Kantorovitch. Structural causal interventions on generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7745–7765. PMLR, 2023. URL `https://arxiv.org/abs/2306.08276`.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2018. URL `https://github.com/fartashf/vsepp`.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Paul Jacob, Éloi Zablocki, Hedi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. Steex: steering counterfactual explanations with semantics. In *European Conference on Computer Vision*, pp. 387–403. Springer, 2022.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pp. 858–876, 2022.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16425–16435, 2023.

Yixue Kang, Yuchen Guo, Zhouxing Ren, Feng Zhang, Simon Dib, Vincent Y. F. Lee, and Nicu Sebe. Causal inference-based explanations for fairness in facial image classifiers. *Sensors*, 22(2):654, 2022. URL `https://www.mdpi.com/1424-8220/22/2/654`.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Na Liu, Fan Zhang, Liang Chang, and Fuqing Duan. Scattering-based hybrid network for facial attribute classification. *Frontiers of Computer Science*, 18(3):183313, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Pratik Mazumder and Pravendra Singh. Protected attribute guided representation learning for bias mitigation in limited data. *Knowledge-Based Systems*, 244:108449, 2022.

Alexander Melistas, Nicolas Mayer, Toshihiko Takeuchi, and Severin Suter. Causal inverse graphics: Context-aware modular causal counterfactual generation. In *Advances in Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=uYT1SfrXiS`.

Bo Peng, Siwei Lyu, Wei Wang, and Jing Dong. Counterfactual image enhancement for explanation of face swap deepfakes. In Shiqi Yu, Zhaoxiang Zhang, Pong C. Yuen, Junwei Han, Tieniu Tan, Yike Guo, Jianhuang Lai, and Jianguo Zhang (eds.), *Pattern Recognition and Computer Vision*, pp. 492–508, Cham, 2022. Springer Nature Switzerland.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1056–1065, 2021.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In *First Conference on Causal Learning and Reasoning*, 2022. URL `https://openreview.net/forum?id=LAAZLZIMN-o`.

Dylan Slack, Himabindu Lakkaraju, Julie Shah, and Sorelle A. Friedler. Counterfactual explanations can be manipulated. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29511–29523, 2021.

Peter Spirtes. An anytime algorithm for causal inference. In Thomas S. Richardson and Tommi S. Jaakkola (eds.), *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3 of *Proceedings of Machine Learning Research*, pp. 278–285. PMLR, 04–07 Jan 2001. URL `https://proceedings.mlr.press/r3/spirtes01a.html`. Reissued by PMLR on 31 March 2021.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* MIT press, 2000.

Diego Velazquez, Pau Rodriguez, Alexandre Lacoste, Issam H Laradji, Xavier Roca, and Jordi Gonzàlez. Explaining visual counterfactual explainers. *Transactions on Machine Learning Research*, 2023.

Lan Wang and Vishnu Naresh Boddeti. Do learned representations respect causal relationships? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 264–274, June 2022.

Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 506–523. Springer, 2020.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *arXiv preprint arXiv:2307.16405*, 2023.