Causal-Adapter: Taming Text-to-Image Diffusion for Faithful Counterfactual Generation

Anonymous authors
Paper under double-blind review

Abstract

We present Causal-Adapter, a modular framework that adapts frozen text-to-image diffusion backbones for counterfactual image generation. Our method enables causal interventions on target attributes, consistently propagating their effects to causal dependents without altering the core identity of the image. In contrast to prior approaches that rely on prompt engineering without explicit causal structure, Causal-Adapter leverages structural causal modeling augmented with two attribute regularization strategies: prompt-aligned injection, which aligns causal attributes with textual embeddings for precise semantic control, and a conditioned token contrastive loss to disentangle attribute factors and reduce spurious correlations. Causal-Adapter achieves state-of-the-art performance on both synthetic and real-world datasets, with up to 91% MAE reduction on Pendulum for accurate attribute control and 87% FID reduction on ADNI for high-fidelity MRI image generation. These results show that our approach enables robust, generalizable counterfactual editing with faithful attribute modification and strong identity preservation.

1 Introduction

Answering counterfactual questions (*e.g.* inferring what an event would have happened under an alternative action) requires understanding the cause–effect relationships among variables and performing hypothetical reasoning (Pearl, 2009; Schölkopf et al., 2021; Weinberg et al., 2024). Classical generative models typically tackle counterfactual-style tasks such as image editing or style transfer in a non-causal manner, whereas subsequent work augments such models with explicit structural causal models (SCMs) to implement *abduction–action–prediction* (Pearl, 2013; Pawlowski et al., 2020; De Sousa Ribeiro et al., 2023). This design drives advances in counterfactual image generation (Figure 1) by enforcing edits consistent with an implied causal graph which enables critical domain-specific counterfactual generation applications, such as simulating medical images with fine-grained anatomical changes associated with aging or disease progression to improve clinical interpretation (Starck et al., 2025). Faithful counterfactual generation remains challenging, as real-world attributes are often causally entangled (*e.g.*, only males can grow beards). This entanglement complicates disentangling factors and generalizing edits, making it nontrivial to ensure that interventions yield the intended visual effect while keeping non-intervened attributes invariant and preserving identity-specific details (Komanduri et al., 2024a).

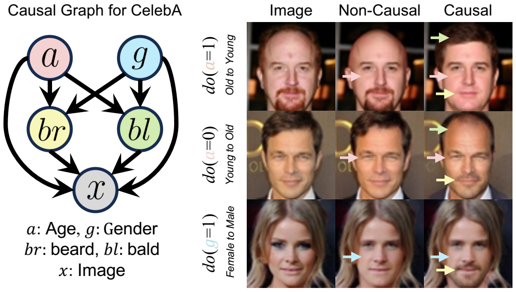


Figure 1: Non-causal editing modifies only the target attribute (*e.g.* age, gender); causal editing propagates changes to related attributes (*e.g.* beard, baldness) enforced by the causal graph.

Counterfactual Image Generation. Early approaches modeled counterfactual image generation using normalizing flows (Papamakarios et al., 2021; Winkler et al., 2020), variational autoencoders (VAEs) (Kingma & Welling, 2013; Yang et al., 2021), hierarchical VAEs (Vahdat & Kautz, 2020; De Sousa Ribeiro et al., 2023), and generative adversarial networks (GANs) (Goodfellow et al., 2014; Kocaoglu et al., 2018), and encouraged attribute disentanglement through variational objectives (Higgins et al., 2017). However, variational optimization inevitably introduces uncertainty into the latent space, which can lead to posterior collapse of meaningful factors, creating a trade-off between image fidelity and controllable attribute manipulation (Figure 2a). Recent works integrate

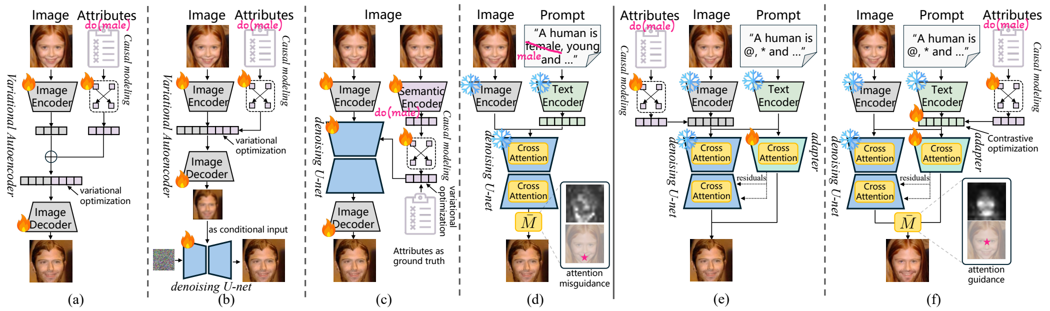


Figure 2: A sketch comparison of counterfactual image generation methods based on: (a) **VAE or GAN**, which fail to achieve high-fidelity results. (b) **Diffusion SCM** and (c) **Diffusion autoencoder**, which are sensitive to spurious correlations. (d) **T2I based editing**, which requires heavy prompt engineering. (e) **Vanilla Causal-Adapter**, which injects causal attributes into image-embedding. (f) **Causal-Adapter with attribute regularization**, which injects causal attributes into learnable textual embeddings with contrastive optimization. Detailed discussion is presented in Appendix A.

diffusion models with the SCM, capitalizing on its high perceptual quality to explore counterfactual identifiability (Sanchez & Tsaftaris, 2022; Rasal et al., 2025; Komanduri et al., 2024c; Pan & Bareinboim, 2024; Xia et al., 2025). Despite domain-specific tuning, previous approaches perform disentanglement only in auxiliary encoders, which has limited effect on diffusion latents and leads to incomplete disentanglement (Figure 2b and c). This makes models prone to spurious correlations, where target factor interventions often cause unintended changes in non-intervened attributes.

Text-to-Image based Editing. An alternative perspective is to view counterfactual image generation as a text-to-image (T2I) based editing, which typically relies on an inversion process (Song et al., 2021). The aim is to generate a target edited image from projected latent states with condition manipulation (Hertz et al., 2023; Ho & Salimans, 2021). To reduce reconstruction error and preserve essential contents, several methods optimize either the unconditional text embedding (Mokady et al., 2023; Xu et al., 2024; Miyake et al., 2025; Ju et al., 2024; Dong et al., 2023), or learn conditional concept embeddings for better attribute disentanglement (Gal et al., 2023; Vinker et al., 2023; Jin et al., 2024). However, generic T2I based editing remains insufficient for counterfactual generation. Existing methods heavily rely on carefully engineered inversion prompts to obtain reliable editing guidance. These approaches lack an explicit, learnable SCM over semantic attributes, making them difficult to guarantee both causal faithfulness and identity preserved counterfactual image generation (Figure 2d). A broader discussion of related works can be found in Appendix A.

Herein, we propose **Causal-Adapter**, an adaptive and modular framework that tames text-to-image diffusion model, such as Stable-Diffusion (SD) (Rombach et al., 2022), for counterfactual generation. Unlike prior diffusion-based methods that require considerable re-training or fine-tuning (Sanchez & Tsaftaris, 2022; Komanduri et al., 2024c; Pan & Bareinboim, 2024; Rasal et al., 2025), our method simply injects causal semantic attributes into a frozen backbone via a pluggable adapter (Figure 2e). Inspired by advances in controllable diffusion (Zhang et al., 2023; Zhao et al., 2023; Mou et al., 2024; Li et al., 2025), we investigate integrating semantic attributes with image embeddings as additional conditions. We find that naive fusion fails to achieve sufficient disentanglement or semantic alignment with spatial features in the diffusion latents. To address this, we introduce two regularization strategies that align causal semantic attributes with textual embeddings, improving disentanglement and causal faithfulness (Figure 2f). Our main contributions are summarized as follows:

- Our motivational studies revealed that relying solely on a frozen text-to-image diffusion model with prompt tuning is insufficient for counterfactual image generation, as it fails to jointly represent causal semantic attributes and image embeddings, leading to imprecise reasoning and edits. This underscores the need to inject causal semantics into diffusion models.
- We propose Causal-Adapter, a framework that employs an adapter encoder to learn causal interactions between semantic variables. These interactions are injected into a frozen diffusion backbone and jointly optimized, introducing a dynamic prior that enables faithful counterfactual generation.
- To enhance causal disentanglement between semantic variables, we introduce two regularization strategies: Prompt-Aligned Injection (PAI) and Conditioned Token Contrastive Loss (CTC). These

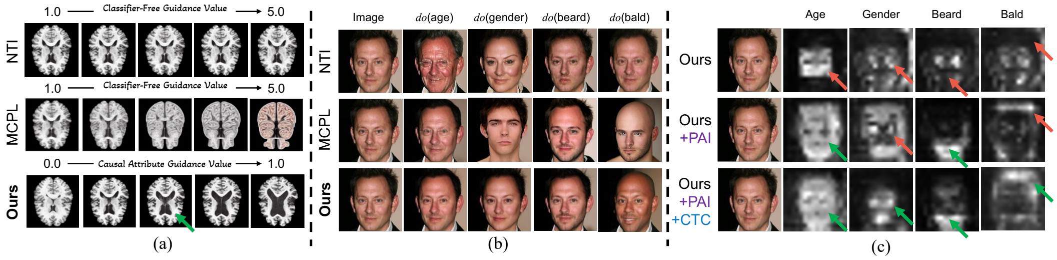

1.0 Classifier-Free Guidance Value 5.0
NTI
1.0 Classifier-Free Guidance Value 5.0
MCPL
0.0 Causal Attribute Guidance Value 1.0
Ours
(a)
Image
do(age)
do(gender)
do(beard)
do(bald)
NTI
MCPL
Ours
(b)
Age
Gender
Beard
Bald
Ours
Ours +PAI
Ours +PAI +CTC
(c)

Figure 3: **Motivational study and preliminary counterfactual generation results between T2I methods and Causal-Adapter.** (a) Fine-grained anatomical counterfactual editing of brain ventricular volume using inversion-based editing (NTI (Mokady et al., 2023)), multi-concept prompt-learning editing (MCPL (Jin et al., 2024)), and our approach. (b) Comparison of counterfactual editing results on human faces. (c) Averaged cross-attention maps from the base Causal-Adapter and the Causal-Adapter with regularizers. Full results and technical details are presented in Appendix B.

strategies separate token embeddings across conditions, enhancing causal representation learning and reducing spurious correlations for more precise counterfactual reasoning.

- We validate Causal-Adapter through extensive experiments on both synthetic and real-world datasets, including human face editing and medical image generation. Our method consistently achieves state-of-the-art performance across key metrics: counterfactual effectiveness (up to 50% MAE reduction on ADNI), realism (81% FID reduction on CelebA), composition (86% LPIPS reduction on CelebA), and minimality (4% CLD reduction on ADNI).

2 Methodology

We first present the preliminaries in Section 2.1, followed by a motivational study in Section 2.2. This study examines the systematic limitations of current T2I based editing methods for counterfactual generation and emphasizes the need to leverage causal semantic attributes for producing faithful counterfactuals (Figure 3). Motivated by these findings, we further introduce the proposed Causal-Adapter in Section 2.3 and describe our regularization strategies in Section 2.4, which further enhance counterfactual generation through semantic disentanglement.

2.1 Preliminaries

Structural Causal Model (SCM). SCM provides a formal framework for modeling causal relationships between variables (Pearl, 2010). An SCM is defined as a triplet $\mathcal{S} = \langle Y, F, U \rangle$, where $Y = \{y_i\}_{i=1}^K$ denotes a set of K endogenous (observed) variables, $U = \{u_i\}_{i=1}^K$ is a set of exogenous (latent) variables, and $F = \{f_i\}_{i=1}^K$ is a set of deterministic functions defining $y_i = f_i(\mathrm{Pa}_i, u_i)$, with $\mathrm{Pa}_i \subseteq Y \setminus \{y_i\}$ denoting the parent variables of y_i. The structural assignments of the SCM induce a directed acyclic graph (DAG) $\mathcal{G}$, where each node represents a variable y_i, and each edge $\mathrm{Pa}_i \to y_i$ represents a direct causal dependency between variables Pa_i and y_i. A SCM is called Markovian if the exogenous variables are mutually independent $p(U) = \prod_{i=1}^K p(u_i)$. A Markovian causal model induces the unique joint observational distribution that satisfies the causal Markov condition $p_{\mathcal{S}}(Y) = \prod_{i=1}^K p(y_i \mid \mathrm{Pa}_i)$.

Counterfactual Reasoning. Counterfactual reasoning aims to answer queries "Given observation Y, what would y_i have been if its parents Pa_i had been different?" This is formalized by the *abduction–action–prediction* procedure (Pearl, 2013; Pawlowski et al., 2020; Shen et al., 2022):

1. *Abduction*: inferring posterior over exogenous variables consistent with observation, $p_{\mathcal{S}}(U \mid Y)$.
2. *Action*: perform an intervention $do(y_i = \tilde{y}_i)$, which replaces the structural function f_i with a constant assignment. This yields a modified SCM $\tilde{\mathcal{S}} = \langle \tilde{Y}, \tilde{F}, U \rangle$, where $\tilde{Y}$ and $\tilde{F}$ denote the modified endogenous variables and structural mechanisms respectively.
3. *Prediction*: compute the counterfactual outcome by evaluating $p_{\tilde{\mathcal{S}}}(\tilde{Y})$ under the modified structural mechanism and the inferred exogenous noise.

Text-to-image Diffusion Model (T2I). T2I models are a class of probabilistic generative models that synthesize images conditioned on textual prompts by progressively denoising Gaussian noise (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022). Given an input image x, a frozen image encoder $\mathcal{E}$ maps it to a latent representation $z_0 = \mathcal{E}(x)$. A conditional text embedding $V = c_\phi(\texttt{Prompt})$ is obtained from a pre-trained text encoder c_ϕ with parameters ϕ, where $\texttt{Prompt}$ denotes the input text. To train the denoising network ϵ_θ, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is added to z_0 to obtain a noisy latent z_t at timestep $t \sim \mathrm{Uniform}\{1, \ldots, T\}$. The denoising loss $\mathcal{L}_{\text{DM}}$ is defined as:

$$\min_{\theta,\phi} \; \mathbb{E}_{\mathcal{E}(x),V,\epsilon\sim\mathcal{N}(0,I),t} \left[\|\epsilon - \epsilon_\theta(z_t, t, V)\|_2^2 \right]. \tag{1}$$

2.2 Motivational Study

To examine the generalization limits of existing text-to-image (T2I) based models for counterfactual image generation, we conducted a motivational study on two real-world datasets: CelebA for human faces and ADNI for brain MRI. Preliminary results are shown in Figure 3, including counterfactual image generation based on different conditional signals in brain volume MRI, as well as interventions on different target attributes in the CelebA dataset. Our study reveals two systematic limitations of directly applying T2I based models to counterfactual generation:

- **Text-only prompting is inadequate for counterfactual generation.** Counterfactual generation should yield consistent visual changes when a single attribute is intervened. However, current T2I based models ignore continuous attributes, making fine-grained edits infeasible, and hence particularly concerning for safety-critical domains such as medical imaging (Figure 3a).
- **Existing T2I based counterfactual generation suffer from attribute entanglement.** Current methods often confuse (entangle) unrelated attributes or require instance-specific fine-tuning, underscoring the need for explicit causal modeling and controllable semantic representations for faithful counterfactual generation (Figure 3b).

2.3 Overview of Causal-Adapter

Recent advances on controllable diffusion methods (Zhang et al., 2023; Zhao et al., 2023; Mou et al., 2024; Li et al., 2025) show that a frozen T2I diffusion backbone can be steered by auxiliary control signals (*e.g.*, segmentation masks or human poses) supplied through a trainable side module. We adopt the same high-level recipe by treating causal semantic attributes as the auxiliary control signals. The causal mechanism between attributes is then learned and injected explicitly into the diffusion backbone via a compact modular encoder that we term as *Causal-Adapter* (Figure 4).

Causal Mechanism Modeling. Let x denote an image and $Y = \{y_i\}_{i=1}^K$ denote a vector of semantic variables, where each y_i represents a scalar value corresponding to a high-level semantic attribute. We assume a known causal graph $\mathcal{G}$ encodes the causal relationships among the variables in Y. Let $A \in \{0,1\}^{K\times K}$ be the binary adjacency matrix of $\mathcal{G}$, where the i-th row $A_i \in \{0,1\}^K$ indicates the parent variables Pa_i of the i-th attribute y_i, *i.e.*, $A_{ij} = 1$ if and only if $y_j \in \mathrm{Pa}_i$. We model each causal mechanism f_i using a nonlinear additive noise model such that:

$$\bar{y}_i := f_i(\mathrm{Pa}_i, u_i) = f_i(A_i \odot Y; \omega_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_i^2), \tag{2}$$

where ω_i are the learnable parameters and $\odot$ denotes element-wise multiplication. $\bar{y}_i$ is the model's prediction of y_i under the change of parent variables Pa_i. For root nodes ($A_i = 0$), we simply set $\bar{y}_i = y_i$ as no causal parents exist. The row vector A_i acts as a binary mask on Y, passing only the true parents of y_i to f_i. To estimate the parameters of mechanisms $F = \{f_i\}_{i=1}^K$ and the noise variances $\{\sigma_i\}_{i=1}^K$, we minimize the negative log-likelihood of the observations:

$$\mathcal{L}_{\text{NLL}} := -\sum_{i=1}^{K} \log p\big(y_i \mid f_i(\mathrm{Pa}_i, u_i)\big) \;=\; \frac{1}{2}\sum_{i=1}^{K} \frac{\|y_i - \bar{y}_i\|_2^2}{\sigma_i^2} + \log(2\pi\sigma_i^2). \tag{3}$$

Unlike prior methods that infer endogenous and exogenous from latent image features (Yang et al., 2021; Komanduri et al., 2024b), we operate directly in the observed semantic attribute space. This reduces reliance on the quality of learned visual representations and enables exact, interpretable *do*-interventions on attributes. The mechanism modeling can be transferred across domains by simply replacing the attribute set and causal graph $\mathcal{G}$. Moreover, it is modular that any differentiable alternatives (*e.g.*, DeepSCM (Pawlowski et al., 2020)) can be seamlessly integrated.

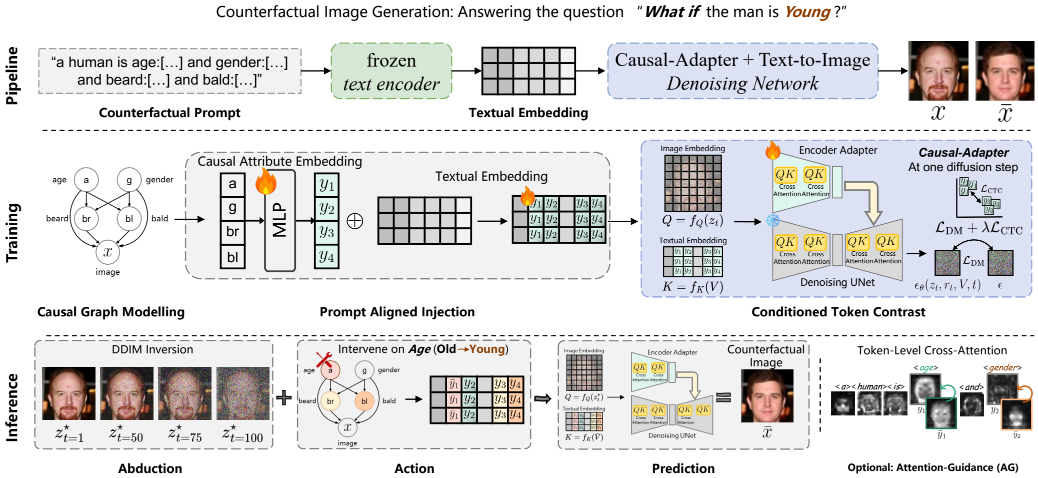


Figure 4: Method overview. A counterfactual prompt and input image x are fed into a pretrained text-to-image diffusion model with a learnable Causal-Adapter $\ddot{\epsilon}_\psi$. Causal mechanisms, modeled over a known causal graph and attributes y_i, are injected into token embeddings via *Prompt-Aligned Injection (PAI)* to align semantic and spatial features. The adapter $\ddot{\epsilon}_\psi$ operates alongside the frozen diffusion U-Net ϵ_θ, optimized with MSE $\mathcal{L}_{\text{DM}}$ and a *Conditioned Token Contrastive (CTC)* loss $\mathcal{L}_{\text{CTC}}$ to enforce disentanglement. At inference, interventions on y_i update token embeddings, and the counterfactual $\bar{x}$ is generated using the abducted exogenous noise $z^\star_t$. Optionally, Attention Guidance (AG) updates the cross-attention map of intervened tokens (*e.g.* age, beard, bald) to achieve localized editing and preserving non-intervened attributes identity (*e.g.* human, gender).

Integrating Causal Conditions into a Frozen T2I Backbone. Prior attempts have treated causal attributes as the conditioning signal to the diffusion backbone, which is computationally expensive as it requires retraining the entire model when switching across datasets (Sanchez & Tsaftaris, 2022; Komanduri et al., 2024c; Rasal et al., 2025). To overcome this limitation, we introduce a branch module that delivers causal guidance without modifying the backbone parameters to preserve the flexibility and scalability of a pretrained T2I model. Given the image latent representation $z_t = \mathcal{E}(x, t)$ at diffusion step t and the text embeddings V, we instantiate a half-scale replica $\ddot{\epsilon}_\psi$ of the denoiser ϵ_θ. This replica specializes $\ddot{\epsilon}_\psi$ in injecting control signals and is parameterized by ψ. The causal attributes Y are added to z_t and processed by $\ddot{\epsilon}_\psi$:

$$r_t := \ddot{\epsilon}_\psi\big(z_t \oplus Y,\, t,\, V\big), \tag{4}$$

the residuals r_t are injected into the mid- and up-sampling blocks of the frozen denoiser ϵ_θ. During training, we keep all backbone parameters fixed and optimize only ψ under the loss in Eqn. 1.

Counterfactual Image Generation. We implement counterfactual generation under the abduction–action–prediction procedure via DDIM inversion and sampling (Song et al., 2021). Let $H_\theta(z_t, V, t)$ denote the DDIM inversion operator and let H_θ^{-1} be its generative inverse. The procedure is as follows: (1) *Abduction:* given an observed image x, we inject the original condition residuals r_t and run $H_\theta(z_{t=0}, r_t, V, t)$ to recover the corresponding exogenous noise $z^\star_T$. (2) *Action:* we intervene y_i by setting $y_i \leftarrow \bar{y}_i$ via $do(y_i = \bar{y}_i)$, propagate the attribute change through Eqn. 2 and Eqn. 4 to obtain the updated $\bar{Y} = \{\bar{y}_i\}_{k=1}^{K}$ and its residuals $\bar{r}_t$. (3) *Prediction:* the counterfactual outcome is generated with $H_\theta^{-1}(z^\star_{t=T}, \bar{r}_t, V, t)$. Optionally, classifier-free guidance (CFG) (Ho & Salimans, 2021) can be applied to amplify the counterfactual signal with guidance weight α:

$$\tilde{\epsilon}_\theta(z_t, \bar{r}_t, t, V, \varnothing) = \alpha\, \epsilon_\theta(z_t, \bar{r}_t, t, V) + (1-\alpha)\, \epsilon_\theta(z_t, r_t^{\varnothing}, t, \varnothing), \tag{5}$$

where $\varnothing = c_\phi(\text{“ ”})$ is the null-text embedding, and $r_t^{\varnothing} := \ddot{\epsilon}_\psi(z_t \oplus \bar{Y},\, t,\, \varnothing)$ are the residuals computed under the null-text condition.

2.4 Regularizing the Causal-Adapter

A key challenge in counterfactual generation is ensuring that interventions modify only the intended attribute while leaving others unaffected. Prior work highlights that achieving such disentanglement in the latent space is crucial for faithful counterfactuals (Yang et al., 2021; Shen et al., 2022; Komanduri et al., 2024b; De Sousa Ribeiro et al., 2023; Komanduri et al., 2024c; Rasal et al., 2025). To assess the degree of disentanglement in a multi-modal setting, we use cross-attention maps to reveal attribute separation in latent space. Figure 3c visualizes averaged cross-attention maps across total denoising time steps among all Causal-Adapter variants. Plain Causal-Adapter, which directly injects causal attributes into the image embedding, fails to align attribute semantics with spatial features and disrupts attribute independence, resulting in entangled, off-target edits. To tackle this issue, we introduce two regularization terms that reinforce the conditional causal attribute disentanglement.

Prompt Aligned Injection. Causal conditions Y are numeric attributes that lack direct alignment with pixel-level representations but are semantically closer to text embeddings. Inspired by prior finding (Yang et al., 2024) that disentangled tokens in the prompt can guide cross-attention to align semantics with spatial structure in diffusion latents, we inject Y through the prompt channel. This allows the cross-attention module to propagate attribute semantics into spatial image features during generation. We introduce a prompt-aligned injection (PAI) mechanism that maps each causal attribute to a learnable token embedding. Let $C = [\,c_1, \ldots, c_K\,]^\top \in \mathbb{R}^{K\times d}$ denote the embeddings of K placeholder tokens with dimension d. Each attribute y_i is mapped into the text space via a small linear projector $g_i : \mathbb{R} \to \mathbb{R}^d$. We form the attribute-injected token embeddings:

$$v_i(y_i) \;=\; c_i + g_i(y_i), \qquad V(Y) \;=\; [\,v_1(y_1), \ldots, v_K(y_K)\,]^\top \in \mathbb{R}^{K\times d}. \tag{6}$$

The set $V(Y)$ serve as the conditioning input to both the frozen denoiser and the adapter via cross-attention. During training, we jointly optimize the adapter ψ, the placeholders token embeddings C, and the projectors $G = \{g_i\}_{i=1}^K$ using the standard MSE:

$$\min_{\psi,\, G,\, C} \; \mathbb{E}_{z,r,Y,\epsilon,t}\left[\left\|\epsilon - \widehat{\epsilon}_{\theta,\psi}\big(z_t, r_t, V(\bar{Y}), t\big)\right\|_2^2\right], \tag{7}$$

where $\widehat{\epsilon}_{\theta,\psi}$ denotes the U-Net prediction modulated by the adapter with PAI. At inference, the learned placeholders C and projectors G are reused to construct $V(\bar{Y})$ for counterfactual query. Notably, PAI only updates token-level embeddings without fine-tuning the CLIP tokenizer or pretrained text encoder, ensuring compatibility between counterfactual queries and existing T2I backbones.

Conditioned Token Contrast (CTC). Motivated by prior efforts that contrastive objectives improve concept separation and reduce cross-factor leakage (Jin et al., 2024; Liu et al., 2025), we introduce a token-conditioned contrastive loss that enforces each placeholder token to capture only a single causal factor. Given a batch with B samples and K attributes (tokens), PAI produces a batch of textual embeddings $\{v_b^k\}_{b=1,k=1}^{B,K}$. For a specific anchor (b,k), positive pairs are same attribute token across samples $\{v_{b'}^k \mid b' \neq b\}$ and negative pairs are different attribute tokens across samples $\{v_b^{k'} \mid k' \neq k\}$. By enforcing inter-token invariance (positive pairs) and intra-token separation (negative pairs), CTC reduces attribute entanglement and suppresses spurious correlations. We implement the objective using InfoNCE (Oord et al., 2018; Chen et al., 2020):

$$\mathcal{L}_{\text{CTC}} = \frac{1}{BK}\sum_{k=1}^{K}\sum_{b=1}^{B}\left[-\,\mathrm{sim}(v_b^k, v_{b'}^k)/\tau + \log\Big(\sum_{k'=1,k'\neq k}^{K}\;\sum_{b'=1,b'\neq b}^{B} \exp\big(\mathrm{sim}(v_b^k, v_{b'}^{k'})/\tau\big)\Big)\right] \tag{8}$$

where sim denotes the cosine similarity and τ is the temperature. Our final training objective becomes:

$$\mathcal{L} = \mathcal{L}_{\text{DM}} + \lambda\,\mathcal{L}_{\text{CTC}}, \tag{9}$$

λ denotes a scaling coefficient. The proposed regularizers improve semantic–spatial alignment in the latent space (Figure 3c) and reduce spurious correlations (Figure 10c). The learned attention maps can be leveraged for localized editing through attention-guided manipulation. Following Ju et al. (2024), interventions are applied only to cross-attention weights corresponding to targeted tokens, while attention for non-intervened tokens is preserved. This ensures that edits remain spatially localized (e.g., gender) without altering unrelated attributes (e.g., hairstyle), improving identity preservation without introducing extra training objectives. Full algorithm is presented in Appendix. C.3.

Table 1: Intervention effectiveness on Pendulum test set. We report MAE from pretrained regressors under four interventions. w/ CM: with causal mechanisms; w/o CM: without causal mechanisms; w/ GT: ground-truth labels injected. "∼" denotes descendant attributes remain unaffected. The table follows the causal graph: e.g., under the "Pendulum (p) MAE", $do(p)$ or $do(l)$ report the pendulum MAE after intervening on p or l.

	Method	**Pendulum (p) MAE ↓**				**Light (l) MAE ↓**				**Shadow Length (sl) MAE ↓**				**Shadow Position (sp) MAE ↓**			
		$do(p)$	$do(l)$	$do(sl)$	$do(sp)$	$do(p)$	$do(l)$	$do(sl)$	$do(sp)$	$do(p)$	$do(l)$	$do(sl)$	$do(sp)$	$do(p)$	$do(l)$	$do(sl)$	$do(sp)$
1.	CausalVAE	24.86	23.03	20.47	11.58	34.20	26.01	35.49	47.06	1.946	1.430	2.020	1.720	52.52	72.50	57.03	32.78
2.	DisDiffAE	0.668	0.648	0.647	0.647	0.656	0.654	0.630	0.651	0.550	0.527	0.560	0.516	0.474	0.475	0.479	0.534
3.	CausalDiffAE	0.297	0.132	**0.031**	**0.034**	0.045	0.434	**0.035**	**0.064**	0.136	0.322	0.492	**0.082**	0.146	0.303	0.064	0.471
4.	Ours$_{\text{w/ CM}}$	**0.014**	**0.035**	0.043	0.259	**0.045**	**0.041**	0.058	0.120	**0.028**	**0.051**	**0.489**	0.110	**0.030**	**0.033**	**0.041**	**0.336**
5.	Ours$_{\text{w/o CM}}$	0.159	0.183	∼	∼	0.060	0.173	∼	∼	0.143	0.235	∼	∼	0.086	0.155	∼	∼
6.	Ours$_{\text{w/ GT}}$	0.013	0.033	∼	∼	0.043	0.039	∼	∼	0.025	0.036	∼	∼	0.028	0.035	∼	∼

3 Experiments

We conduct extensive experiments on four counterfactual image generation datasets across different domains: the synthetic **Pendulum** dataset (Yang et al., 2021), the human-face dataset **CelebA** (Liu et al., 2015) and its high-resolution restoration **CelebA-HQ** (Karras et al., 2017), and the medical imaging dataset **Alzheimer's Disease Neuroimaging Initiative (ADNI)** (Petersen et al., 2010). We follow the benchmarking experimental settings (Melistas et al., 2024; Komanduri et al., 2024c; Rasal et al., 2025) for fair comparison. During evaluation, we follow the official setups for each of the datasets and report: (1) Effectiveness: whether the intervention succeeds, measured by pretrained classifiers using F1-score or MAE depending on the attribute type. (2) Composition: reconstruction quality under null interventions, measured by MAE and LPIPS distance. (3) Realism: visual quality of counterfactual images, evaluated with Fréchet Inception Distance (FID). (4) Minimality: non-intervened attributes remain minimally affected, assessed with the Counterfactual Latent Divergence (CLD) metric. Implementation details are presented in Appendix C.

3.1 Results

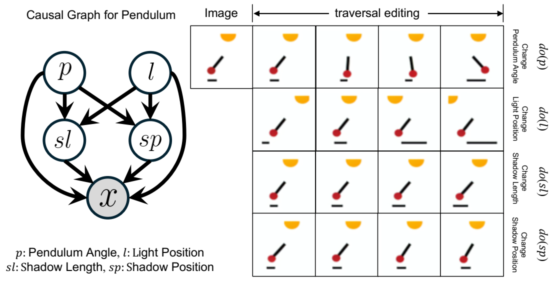


Figure 5: Pendulum counterfactuals with traversal editing along each attribute.

Synthetic Imaging Counterfactuals. We first evaluate our approach on the Pendulum dataset, which consists of four continuous causal variables (pendulum angle, light position, shadow length, and shadow position). We compare against CausalVAE (Yang et al., 2021), as well as diffusion-based methods DisDiffAE (Preechakul et al., 2022) and CausalDiffAE (Komanduri et al., 2024c), using the causal graph shown in Figure 5. As reported in Table 1, our method (row 4) achieves state-of-the-art intervention performance across most attributes. In particular, under $do(l)$ (light intervention), we obtain up to a 91% reduction in MAE for light prediction (from 0.434 to 0.041), indicating highly accurate control over light movement. Moreover, when intervening on light, our approach correctly preserves the pendulum angle while inducing causal changes in the descendant shadow attributes (shadow length and position), consistent with the real-world physical law. We further conduct an investigation into the role of causal mechanisms. Without explicit causal mechanisms (row 5), the model exhibits larger intervention errors, underscoring the necessity of updating semantic variables with causal dependencies during generation. In contrast, injecting synthetic ground-truth labels (row 6) yields performance close to ours (row 4), indicating that our causal mechanism injection provides a principled approximation to ground-truth causal reasoning, generating precise and faithful counterfactuals over all attributes as shown in Figure 5.

Human Face Counterfactuals. Following the benchmarking of Melistas et al. (2024), we evaluate Causal-Adapter on CelebA test set for human face counterfactual generation across four categorical attributes (age, gender, beard, bald) with the causal graph shown in Figure 1. We also incorporate attention guidance to perform localized editing and assess the utility of our learned attention maps. Table 2 reports intervention effectiveness, while Table 3 summarizes composition, realism, and

Table 2: Intervention effectiveness on CelebA test set. Average F1 scores under four interventions.

	Method	**Age** (a) **F1** ↑				**Gender** (g) **F1** ↑				**Beard** (br) **F1** ↑				**Bald** (bl) **F1** ↑			
		$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
1.	VAE	35.0	78.2	81.6	81.9	97.7	90.9	95.9	**97.3**	94.4	82.8	29.6	94.5	2.3	49.6	4.5	41.2
2.	HVAE	**65.4**	89.3	90.8	**89.9**	98.8	94.9	**99.4**	95.0	95.2	95.1	44.1	91.6	2.0	86.0	4.5	**61.1**
3.	GAN	41.3	71.0	81.8	79.9	95.2	98.2	92.0	96.1	90.8	83.8	23.3	90.7	2.1	82.0	5.5	49.2
4.	Ours	58.5	**89.8**	94.0	89.4	**99.6**	**99.9**	74.5	92.5	**99.7**	96.1	**52.1**	**98.1**	**59.2**	**91.2**	**74.5**	58.8
with attention guidance for localized editing																	
5.	Ours	57.1	89.5	**94.5**	81.5	99.6	99.7	73.8	89.2	99.7	**96.8**	48.2	96.6	38.7	78.4	47.7	51.4

Table 3: Composition, realism and minimality results on CelebA test set. "-" indicates attention guidance was not applied for reconstruction.

	Method	**Composition**		**Realism**	**Minimality**
		MAE ↓	LPIPS ↓	FID ↓	CLD ↓
1.	VAE	18.695	0.282	59.397	**0.299**
2.	HVAE	7.143	0.122	35.712	0.305
3.	GAN	60.120	0.276	27.861	0.304
4.	Ours	**3.535**	**0.017**	8.152	0.310
with attention guidance for better realism and minimality					
5.	Ours	-	-	**5.213**	0.301

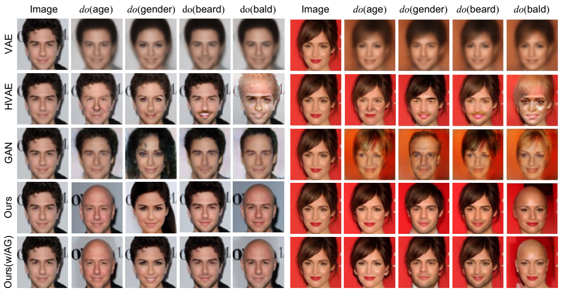


Figure 6: CelebA counterfactuals from Causal-Adapter compared with prior methods.

minimality. Compared with VAE, HVAE, and GAN, our method achieves best performance across most interventions, including up to an 86% reduction in LPIPS (composition, from 0.122 to 0.017) and an 79% reduction in FID (realism, from 27.861 to 8.152). HVAE achieves competitive effectiveness via post-training classifier optimization but produces classifier-biased artifacts and reduced fidelity (Figure 6). Besides, our method delivers significant improvements under $do(a)$ and $do(br)$ on F1 score of bald attribute, demonstrating causal faithfulness in representing baldness across diverse individuals. As illustrated in Figure 6, Causal-Adapter can successfully add baldness to both male and female under $do(bl)$, and under interventions on causal parents such as $do(a)$, it can jointly edit age and baldness, whereas baselines fail to maintain causal consistency. Further localized editing with attention guidance balances intervention effectiveness and identity preservation, achieving the best FID (from 8.152 to 5.213). Qualitative results confirm that the learned attention maps enable precise, localized detailed editing (e.g., modifying gender without altering hairstyle), enable Causal-Adapter to preserve core identity while enforcing causal interventions. Full results are in Appendix E.2.

Brain Imaging Counterfactuals We further evaluate our method on ADNI dataset, which includes six attributes in both categorical variables (ApoE, Sex, Slice) and continuous variables (Age, Brain Volume, Ventricular Volume). Following Melistas et al. (2024), we intervene on three generative conditioning attributes (Brain Volume, Ventricular Volume, Slice) and report the results in Table 4. Our approach achieves best performance in intervention effectiveness (up to 50% MAE reduction in fine-grained edits) and minimality, while also delivering strong realism (87% FID reduction) even without attention guidance. Qualitative results in Figure 7 further show that our model produces sharp and localized interventional changes consistent with the causal graph. For example, edits to ApoE, Age, or Sex appropriately influence Brain and Ventricular Volumes. In particular, fine-grained edits to Ventricular Volume visibly enlarge the ventricle region while faithfully preserving subject identity.

High-Resolution Face Counterfactuals We further compare Causal-Adapter with recent state-of-the-art counterfactual methods including VCI (Wu et al., 2025), HVAE (De Sousa Ribeiro et al., 2023), and DiffCounter (Rasal et al., 2025) on high-resolution human face dataset CelebA-HQ. We follow the same settings in DiffCounter for fair comparison, focusing on three categorical variables (glasses, smile, mouth-open). The quantitative results are presented in Table 5, with an additional reversibility metric that measures how well the generated counterfactuals can be recovered back to the original observations. Our Causal-Adapter (row 4) achieves intervention effectiveness comparable to DiffCounter, while substantially improving both reversibility and identity preservation. For example, under the smile intervention, our method reduces LPIPS by 57% (from 0.66 to 0.028), and for glasses reversal, it reduces L_1 by 73% (from 0.185 to 0.049). Qualitative results in Figure 8 further demonstrate that our model performs causal faithful generation. The reversibility analysis empirically

Table 4: Effectiveness on ADNI test set, evaluated with MAE/F1 from pretrained regressors/classifiers (std. dev. reported). Composition, realism, and minimality metrics are also included.

	Method	Brain volume (b) MAE ↓			Ventricular volume (v) MAE ↓			Slice (s) F1 ↑			Composition		Realism	Minimality
		$do(b)$	$do(v)$	$do(s)$	$do(b)$	$do(v)$	$do(s)$	$do(b)$	$do(v)$	$do(s)$	MAE ↓	LPIPS↓	FID↓	CLD↓
1.	VAE	$0.17_{0.03}$	$0.15_{0.06}$	$0.15_{0.06}$	$0.08_{0.05}$	$0.20_{0.04}$	$0.08_{0.05}$	$0.52_{0.15}$	$0.48_{0.15}$	$0.46_{0.10}$	18.88	0.306	278.245	0.352
2.	HVAE	$0.09_{0.03}$	$0.12_{0.06}$	$0.13_{0.06}$	$0.06_{0.04}$	$0.04_{0.01}$	$0.06_{0.04}$	$0.38_{0.15}$	$0.41_{0.16}$	$0.41_{0.11}$	**3.38**	0.101	74.696	0.347
3.	GAN	$0.17_{0.02}$	$0.16_{0.07}$	$0.16_{0.06}$	$0.12_{0.02}$	$0.22_{0.03}$	$0.12_{0.03}$	$0.14_{0.03}$	$0.16_{0.03}$	$0.05_{0.02}$	24.26	0.268	113.749	0.353
4.	Ours	$\mathbf{0.09}_{0.01}$	$\mathbf{0.11}_{0.03}$	$\mathbf{0.11}_{0.03}$	$\mathbf{0.03}_{0.01}$	$\mathbf{0.03}_{0.01}$	$\mathbf{0.03}_{0.01}$	$0.53_{0.09}$	$0.55_{0.09}$	$\mathbf{0.48}_{0.06}$	3.54	**0.035**	9.130	0.346
with attention guidance for localized editing														
5.	Ours	$0.10_{0.01}$	$0.14_{0.04}$	$0.14_{0.06}$	$0.10_{0.02}$	$0.04_{0.01}$	$0.10_{0.02}$	$\mathbf{0.55}_{0.08}$	$\mathbf{0.57}_{0.08}$	$0.46_{0.08}$	-	-	**9.066**	**0.332**

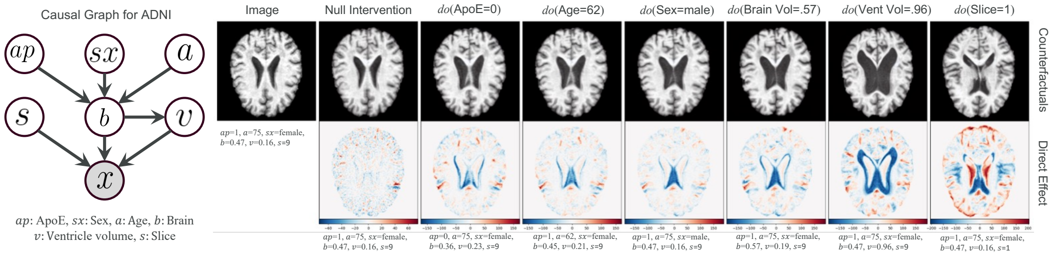

Causal Graph for ADNI
ap
sx
a
s
b
v
x
ap: ApoE, sx: Sex, a: Age, b: Brain
v: Ventricle volume, s: Slice
Image
Null Intervention
do(ApoE=0)
do(Age=62)
do(Sex=male)
do(Brain Vol=.57)
do(Vent Vol=.96)
do(Slice=1)
Counterfactuals
Direct Effect
ap=1, a=75, sx=female, b=0.47, v=0.16, s=9
-60 -40 -20 0 20 40 60
-150 -100 -50 0 50 100 150
-200 -150 -100 -50 0 50 100 150 200
ap=1, a=75, sx=female, b=0.47, v=0.16, s=9
ap=0, a=75, sx=female, b=0.36, v=0.23, s=9
ap=1, a=62, sx=female, b=0.45, v=0.21, s=9
ap=1, a=75, sx=male, b=0.47, v=0.16, s=9
ap=1, a=75, sx=female, b=0.57, v=0.19, s=9
ap=1, a=75, sx=female, b=0.47, v=0.96, s=9
ap=1, a=75, sx=female, b=0.47, v=0.16, s=1

Figure 7: ADNI brain MRI counterfactual results from Causal-Adapter. Direct causal effects are shown (red: increase; blue: decrease). The results show sharp, localized interventional changes consistent with the causal graph (left), while preserving the observation's identity.

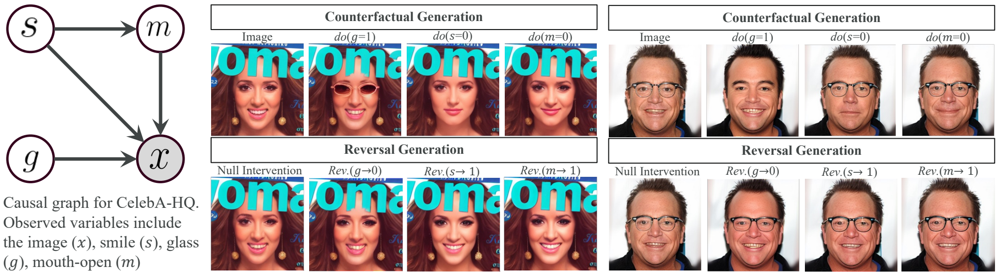
Table 5: Soundness of CelebA-HQ counterfactual images generated by the proposed Causal-Adapter. Effectiveness is evaluated using F1-scores from pre-trained classifiers for eyeglasses (g) and smiling (s), while reversibility (Rev.) and compositional consistency (Comp.) are measured using L_1. Identity preservation (IDP) is assessed using LPIPS.

		Eyeglasses Intervention ($do(g)$)				Smiling Intervention ($do(s)$)				Null
		Effectiveness		Rev.	IDP	Effectiveness		Rev.	IDP	Comp.
	Method	F1(s) ↑	F1(g) ↑	L_1 ↓	LPIPS ↓	F1(s) ↑	F1(g) ↑	L_1 ↓	LPIPS ↓	L_1 ↓
1.	VCI	97.84	3.39	-	-	33.81	**99.85**	-	-	-
2.	HVAE	90.05	65.31	-	-	75.33	95.82	-	-	-
3.	DiffCounter	**99.09**	96.86	0.185	0.096	**94.93**	99.45	0.183	0.066	0.130
4.	Ours	96.89	**99.26**	**0.049**	0.084	94.15	99.15	**0.028**	**0.028**	0.010
5.	Ours (DiT)	98.19	97.39	0.086	**0.060**	94.71	99.71	0.089	0.035	**0.001**

indicates that our model can recover counterfactuals back to their original observations (Appendix F), suggesting potential for achieving counterfactual identifiability (Ribeiro et al., 2025).

Generalization Robustness To further verify the generalization capability of our framework, we additionally test the Causal-Adapter with a different diffusion backbone, Stable Diffusion 3 (Esser et al., 2024), which uses a diffusion transformer (DiT) architecture. The quantitative results (Table 5, row 5) and qualitative examples in Figure 9, demonstrate that our method generalizes effectively across distinct T2I backbones and produces causally faithful, high-resolution counterfactuals, going beyond prior studies that were limited to low-resolution settings (Wu et al., 2025; Rasal et al., 2025; Melistas et al., 2024). Extended generalization results are provided in Appendix E.4

Ablation Study. We conduct an ablation study on the CelebA validation set to evaluate the contribution of each regularizer in Causal-Adapter. As shown in Figures 10a and b, the plain adapter achieves the lowest intervention effectiveness. Adding the PAI module yields a consistent average gain of +11.6% F1 across all attributes with slight increases in FID and CLD, indicating that aligning causal semantics with spatial features in the diffusion latents enables more effective edits. Incorporating CTC further improves intervention effectiveness by enforcing token embedding disentanglement, while also reducing CLD (from 0.317 to 0.310) and keeping FID stable (from 8.453 to 8.643), resulting more faithful counterfactual generation.



Causal graph for CelebA-HQ. Observed variables include the image (x), smile (s), glass (g), mouth-open (m)

Figure 8: CelebA-HQ counterfactuals (256×256) from Causal-Adapter using SD1.5 backbone.

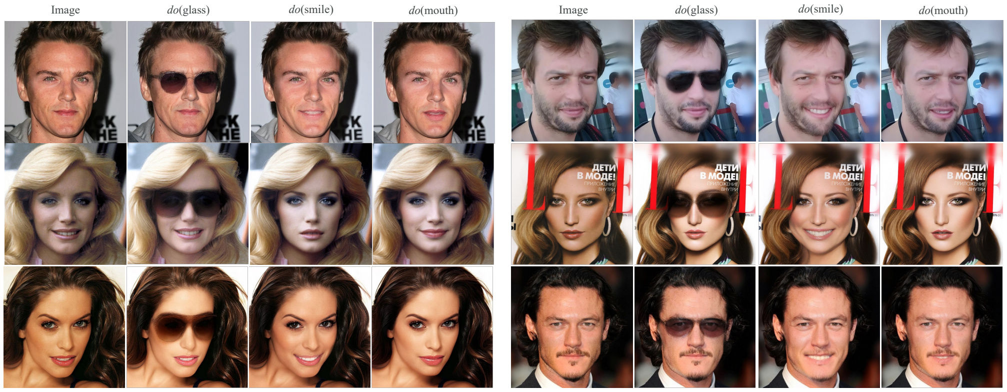


Figure 9: CelebA-HQ counterfactuals (512×512) from Causal-Adapter using SD3 backbone.

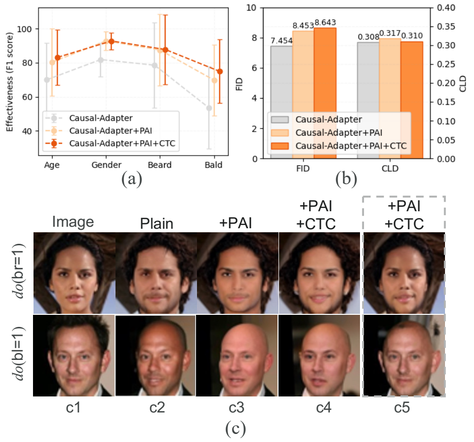


Figure 10: Ablation study on CelebA validation set. (a) Average intervention effectiveness. (b) Realism and minimality. (c) Qualitative examples, with dotted boxes indicating results of localized editing.

Figure 10c presents qualitative ablation results. Spurious correlations arise in plain adapter, *e.g.*, beard interventions in females induce male facial features (c2). PAI alleviates attribute entanglement (c3), while the complete regularization with CTC produces bearded female, demonstrating mitigation of such correlations (c4). Under bald interventions, the baselines alter skin color or age (c2–c3), whereas the full regularization edits baldness with only minor facial changes (c4). Finally, attention guidance enhances identity preservation with localizing edits (c5). Extended results are presented in Appendix D.

4 Conclusion

We introduced Causal-Adapter to tame Text-to-Image diffusion models for counterfactual image generation. Our motivational study revealed that current Text-to-Image diffusion model based editing approaches lack an explicit structural causal model for attribute control and rely heavily on prompt engineering, making it difficult to generate faithful counterfactual images. In contrast, Causal-Adapter is a simple yet effective framework that leverages a frozen diffusion backbone and injects causal semantic attributes through a pluggable adapter network to explicitly learn causal semantics. We further proposed prompt aligned injection and conditioned token contrastive optimization, which align attribute semantics with spatial features and promote disentanglement in the latent space, reducing spurious correlations while preserving identity for generations. Causal-adapter achieves superior counterfactual generation performance on multiple datasets. Extensive evaluation across diverse settings further confirm that Causal-Adapter provides a robust, scalable, and practical alternative for enabling causal editing in modern T2I systems.

Ethics Statement

This work makes use of two real-world datasets: CelebA (Liu et al., 2015) and ADNI (Petersen et al., 2010), together with pretrained text-to-image generation models. We emphasize that our research is conducted strictly for scientific purposes, and we strongly condemn any misuse of generative AI to produce content that harms individuals, violates privacy, or spreads misinformation. While our approach demonstrates capabilities in generating human faces and MRI images, we acknowledge the potential for misuse. To mitigate these risks, we uphold the highest ethical standards, including adherence to applicable legal and institutional frameworks, respect for data privacy, and a commitment to promoting socially beneficial applications of generative models.

Reproducibility Statement

In Section 2, we provide detailed formulations, illustrative examples, and visual demonstrations (e.g., Figure 4) to clarify the model structure and mechanisms. Section 3 and Appendix C describe the training datasets, model parameters, and implementation components. Comprehensive quantitative and qualitative results are reported across multiple domains, including Table 1, Table 2, Figure 6, and Figure 7, with comparisons to existing baselines under standard benchmarks. All experiments are conducted on publicly available datasets (CelebA and ADNI) and evaluated with widely used metrics such as FID and LPIPS. To further promote reproducibility, we will release the experimental code upon acceptance of this work.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International conference on machine learning*, pp. 159–168. PMLR, 2018.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Patrick Chao, Patrick Blöbaum, Sapan Kirit Patel, and Shiva Kasiviswanathan. Modeling causal mechanisms with diffusion models for interventional and counterfactual queries. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=EDHQDsqiSe`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7390–7425, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/de-sousa-ribeiro23a.html`.

Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7430–7440, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=FPnUhsQJ5B`.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=NAQvF08TcyG`.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=_CDixzkzeyb`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL `https://openreview.net/forum?id=qw8AKxfYbI`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13753–13773. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/huang23b.html`.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12469–12478, 2024.

Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Alexander Teare. An image is worth multiple words: Discovering object level concepts using multi-concept prompt learning. In *Forty-first International Conference on Machine Learning*, 2024.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=FoMZ4ljhVw`.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BJE-4xW0W`.

Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL `https://openreview.net/forum?id=PUpZXvNqmb`.

Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 4308–4316. International Joint Conferences on Artificial Intelligence Organization, 8 2024b. doi: 10.24963/ijcai.2024/476. URL `https://doi.org/10.24963/ijcai.2024/476`. Main Track.

Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. *European Conference on Artificial Intelligence*, 2024c.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AumOa10MKG.

Xiutian Li, Siqi Sun, and Rui Feng. Causal representation learning via counterfactual intervention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 3234–3242, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22511–22521, 2023.

Zhihua Liu, Amrutha Saseendran, Lei Tong, Xilin He, Fariba Yousefi, Nikolay Burlutskiy, Dino Oglic, Tom Diethe, Philip Alexander Teare, Huiyu Zhou, and Chen Jin. Segment anyword: Mask prompt inversion for open-set grounded segmentation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9bzgpYtQZn.

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.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios Tsaftaris. Benchmarking counterfactual image generation. *Advances in Neural Information Processing Systems*, 37:133207–133230, 2024.

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2063–2072. IEEE, 2025.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lZOUQQvwI3q.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *International conference on machine learning*, pp. 25733–25754. PMLR, 2023.

Achille Nazaret, Justin Hong, Elham Azizi, and David Blei. Stable differentiable causal discovery. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=JJZBZW28Gn.

Mateusz Olko, Mateusz Gajewski, Joanna Wojciechowska, Mikołaj Morzy, Piotr Sankowski, and Piotr Miłoś. Since faithfulness fails: The performance limits of neural causal discovery. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2nQyYo71ih.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Yushu Pan and Elias Bareinboim. Counterfactual image editing. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=OXzkw7vFIO.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Causal inference. *Causality: objectives and assessment*, pp. 39–58, 2010.

Judea Pearl. Structural counterfactuals: A brief introduction. *Cognitive science*, 37(6):977–985, 2013.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209, 2010.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022.

Lemuel Puglisi, Daniel C Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 173–183. Springer, 2024.

Rajat R Rasal, Avinash Kori, Fabio De Sousa Ribeiro, Tian Xia, and Ben Glocker. Diffusion counterfactual generation with semantic abduction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Wqrqcc8O2v.

Fabio De Sousa Ribeiro, Ainkaran Santhirasekaram, and Ben Glocker. Counterfactual identifiability via dynamic optimal transport. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=h2ttG6HkID.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23 (241):1–55, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.

Nikos Spyrou, Athanasios Vlontzos, Paraskevas Pegios, Thomas Melistas, Nefeli Gkouti, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A Tsaftaris. Causally steered diffusion for automated video counterfactual generation. *arXiv preprint arXiv:2506.14404*, 2025.

Sophie Starck, Vasiliki Sideri-Lampretsa, Bernhard Kainz, Martin J Menten, Tamara T Mueller, and Daniel Rueckert. Diff-def: Diffusion-generated deformation fields for conditional atlases. *IEEE Transactions on Medical Imaging*, 2025.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6):1–13, 2023.

Abraham Itzhak Weinberg, Cristiano Premebida, and Diego Resende Faria. Causality from bottom to top: a survey. *arXiv preprint arXiv:2403.11219*, 2024.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows, 2020. URL `https://openreview.net/forum?id=rJg3zxBYwH`.

Yulun Wu, Louie McConnell, and Claudia Iriondo. Counterfactual generative modeling with variational causal inference. *International Conference on Learning Representations*, 2025.

Tian Xia, Fabio De Sousa Ribeiro, Rajat R Rasal, Avinash Kori, Raghav Mehta, and Ben Glocker. Decoupled classifier-free guidance for counterfactual diffusion models. *arXiv preprint arXiv:2506.14399*, 2025.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9452–9461, 2024.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9593–9602, 2021.

Tao Yang, Cuiling Lan, Yan Lu, and Nanning Zheng. Diffusion model with cross attention as an inductive bias for disentanglement. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7685–7695, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

CONTENTS

A Extended Related Works

Counterfactual Image Generation aims to synthesize images that reflect the visual effect of a hypothetical intervention while preserving non-intervened attributes according to an underlying causal graph and maintaining instance-specific identity details (Komanduri et al., 2024a; Melistas et al., 2024). Existing approaches often augment generative models with explicit structural causal models (SCMs) and follow Pearl's *abduction–action–prediction* paradigm (Pearl, 2009; 2013; Pawlowski et al., 2020; Shen et al., 2022; Sanchez & Tsaftaris, 2022; De Sousa Ribeiro et al., 2023; Wu et al., 2025; Spyrou et al., 2025). Early work predominantly relied on **VAEs or GANs** (Kingma & Welling, 2013; Goodfellow et al., 2014; Kocaoglu et al., 2018; Pawlowski et al., 2020), as their noise-injection and variational sampling mechanisms offered a natural way to represent exogenous uncertainty, while their objectives encouraged disentanglement of causal factors, inspired by β-VAE (Higgins et al., 2017). De Sousa Ribeiro et al. (2023) extends this line by incorporating hierarchical VAEs (Vahdat & Kautz, 2020) to estimate direct, indirect, and total causal effects, whereas Wu et al. (2025) combined variational Bayesian inference with adversarial training to improve abduction and preserve identity. Other works (Yang et al., 2021; Shen et al., 2022; Komanduri et al., 2024b; Li et al., 2024) integrates SCM priors directly into the latent space of VAEs, enabling implicit causal reasoning on image encodings. However, variational optimization inevitably introduces uncertainty into the learned representations, which can lead to posterior collapse or the loss of semantically meaningful factors. This results in an inherent trade-off between high-fidelity image synthesis and flexible attribute control (Higgins et al., 2017; Alemi et al., 2018). Sanchez & Tsaftaris (2022) introduces **DiffSCM**, the first framework to integrate diffusion with SCMs for counterfactual generation. DiffSCM employed DDIM inversion (Song et al., 2021) for abduction and conditioned the generative process on causal labels, but it was limited to small parent sets and simple causal structures. Pan & Bareinboim (2024) extends DiffSCM by combining a VAE for preliminary counterfactual generation with diffusion-based refinement, shifting the conditioning from attribute labels to pre-generated images to produce refined counterfactuals. Chao et al. (2024) models structural equations directly using diffusion processes, enabling counterfactual sampling on a predefined causal graph. While DCM also answers interventional queries, it trains a dedicated diffusion model per causal node. Recent works (Komanduri et al., 2024c; Rasal et al., 2025; Xia et al., 2025) build on **Diffusion Autoencoder** framework (Preechakul et al., 2022), equipping diffusion models with variational encoders to inject semantic attributes into diffusion latents. However, these methods require heavy post-training or fine-tuning and remain prone to spurious correlations. A key limitation is that disentanglement is typically enforced through auxiliary encoders with limited influence on the intrinsic latent space of diffusion models, failing to align semantic attributes with disentangled spatial representations of images and leaving causal disentanglement incomplete (Wu et al., 2025; Yang et al., 2024).

To address these limitations, we propose **Causal-Adapter**, an adaptive and modular framework that employs an adapter encoder to explicitly learn causal interactions between semantic attributes. We further introduce two regularization strategies: Prompt Aligned Injection (PAI) and Conditioned Token Contrastive Loss (CTC). These strategies align semantic attributes with spatial features in the diffusion latents and separate token embeddings across conditions, thereby enhancing causal representation learning and reducing spurious correlations, all while preserving the pre-trained diffusion weights.

Text-to-Image based Editing aims to manipulate existing images according to user-provided natural language instructions (Brooks et al., 2023). Most approaches rely on an inversion process (Song et al., 2021), where the image is projected into a latent state and then synthesized under modified conditions (Hertz et al., 2023; Ho & Salimans, 2021). However, DDIM inversion and classifier-free guidance often often interfere with each other, leading to a trade-off between preserving essential content and achieving faithful edits (Ju et al., 2024; Huberman-Spiegelglas et al., 2024; Kynkäänniemi et al., 2024). To mitigate this, Null-text inversion (Mokady et al., 2023) learns a null embedding to account for reconstruction discrepancies, while Dong et al. (2023) optimizes conditional embeddings to reduce information loss in unconditional guidance. Yet, both methods require costly per-sample optimization. Subsequent works (Ju et al., 2024; Xu et al., 2024; Miyake et al., 2025) improve efficiency by recording residual losses between conditional and unconditional embeddings and re-injecting them during editing, to stabilize edits and preserving content. Textual inversion (Gal et al., 2023) improves content preservation by disentangling single concepts: it learns new text embeddings from a few personalized images to represent objects in novel contexts. Extending this, Vinker et al. (2023) decomposes concept embeddings into sub-concepts via learned vectors injected into the latent space of T2I model, while Jin et al. (2024) performs multi-concept disentanglement using adjective-based prompts and contrastive optimization. Lyu et al. (2024) propose a one-dimensional non-invasive adapter that modulates concept semi-permeability in frozen diffusion models for concept erosion, and Yeganeh et al. (2025) address domain shift in medical imaging, supporting counterfactual-like edits such as aging or disease progression without explicit structural constraints. Despite these advances, generic T2I editing remains insufficient for causal counterfactual generation. Existing methods depend heavily on prompt manipulation and do not incorporate a learnable structural causal model (SCM). They make no use of observed semantic attributes or a predefined causal graph, and therefore cannot enforce that edits follow correct causal dependencies. As a result, current T2I-based editing techniques struggle to simultaneously maintain causal faithfulness and identity preservation in counterfactual image generation.

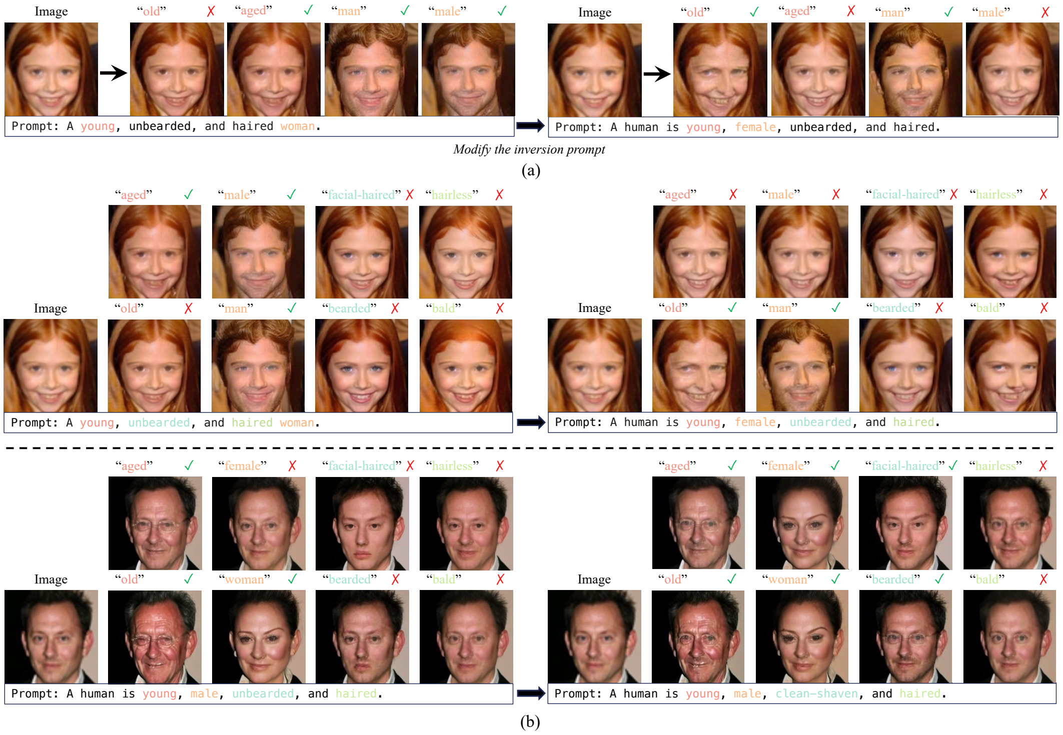
As shown in our motivational study (Appendix B), we evaluate the adaptation of text-to-image diffusion models for counterfactual image generation. Our findings show that relying solely on a frozen diffusion backbone with prompt-tuning is insufficient, as it fails to jointly represent causal semantic attributes and image embeddings. As a result, the model struggles to achieve precise counterfactual reasoning and generation. This highlights the need for an adaptive mechanism within the diffusion model, enhanced with injected causal semantics. Following the standard formulation of counterfactual image generation (Pearl, 2009; De Sousa Ribeiro et al., 2023; Wu et al., 2025; Rasal et al., 2025), where a predefined causal graph is treated as the structural prior, our Causal-Adapter models the SCM directly on the observed semantic attributes. This enforces correct causal dependencies during intervention and enables the generation of counterfactual images that are both visually plausible and causally faithful. When the SCM module is omitted, the framework naturally reduces to a standard conditional generation setting for that attribute, as illustrated in Figure 1.

Controllable Diffusion Models extend T2I diffusion frameworks by incorporating additional user-specified signals to guide generation (Huang et al., 2025). One approach is to train diffusion models from scratch with multi-conditional objectives (Huang et al., 2023; Puglisi et al., 2024), which achieves strong controllability and high-quality synthesis but at huge computational cost. More recently, adapter-based methods (Zhang et al., 2023; Li et al., 2023; Zhao et al., 2023; Mou et al., 2024; Li et al., 2025) have emerged as a scalable alternative. By attaching lightweight, trainable modules to a frozen Stable Diffusion backbone, these methods enable the model to incorporate auxiliary control signals such as segmentation masks or pose skeletons, significantly reducing training overhead while maintaining stability and fidelity. We adopt the same high-level recipe by treating causal semantic attributes as the auxiliary control signals, and employ an adapter encoder to explicitly learn causal interactions between high-level variables. These interactions are then injected into a a frozen diffusion backbone, and jointly optimized with partial textual embeddings. The design of Causal-Adapter introduces a dynamic and learnable prior, enabling the effective adaptation of a frozen diffusion network for realistic and faithful counterfactual image generation.

B Full Motivational Study Results

To assess the feasibility of using a pretrained text-to-image (T2I) model for counterfactual generation, we conduct a motivation study to answer the following three questions.

1. Can existing T2I based editing methods perform counterfactual generation? We begin by examining a representative text-to-image (T2I) editing method, Null-Textual Inversion (NTI) (Mokady et al., 2023). Our findings highlight two fundamental issues: (1) *Heavy reliance on prompt engineering and instability.* As shown in Figure 11a, the success of NTI-based edits is highly sensitive to prompt wording. For example, when editing age, the token "old" may fail while the synonym "aged" succeeds under one inversion prompt, yet the opposite occurs under another prompt. This indicates that editing success and visual quality depend not only on the choice of attribute word but also on the particular inversion prompt. Even minor wording changes that preserve semantic meaning can alter tokenization and cross-attention patterns, leading to inconsistent success rates and perceptual artifacts. Moreover, to preserve identity while achieving the desired edit, practitioners must manually try multiple prompt variants, requiring extensive hand-crafted effort. (2) *Weak counterfactual faithfulness.* Beyond instability, prompt-based editing fails to reliably reflect the intended intervention. In Figure 11b, using the word "man" to edit gender (woman case) or "old" to edit age (man case) yields different counterfactual characteristics depending on the inversion prompt: one "old" edit introduces glasses while another does not. Such variability arises purely from prompt formulation, introducing extraneous variance unrelated to the target attribute. This illustrates weak counterfactual faithfulness: the same intervention should yield coherent edits of the intended factor, without unintended changes in other attributes. Hence, purely text-driven control is inadequate for reliable counterfactual generation.



(a)

(b)

Figure 11: Null-Textual Inversion (NTI) relies heavily on prompt engineering, where minor word changes can determine editing success. (a) Illustration of our study: two inversion prompts are given for the same image, and attribute words are replaced with different synonyms. Successful edits are marked with ✓ and failures with ✗ (human evaluation). Results show that editing outcomes are highly sensitive to prompt wording and reveal weak counterfactual faithfulness. (b) Full results of this investigation.

2. Can multi-concept prompt learning yield disentangled attribute control? Prompt-learning approaches such as Textual Inversion (Gal et al., 2023), Inspiration Tree (Vinker et al., 2023) and Multi-Concept Prompt Learning (MCPL) framework (Jin et al., 2024) to learn conditional concept embeddings for attribute disentanglement. We adopt MCPL framework as our baseline. As shown in Figure 12, MCPL can perform certain edits that NTI fails to capture (e.g., editing baldness for men or adding beards for women) by equipping text embeddings learned from multiple samples. However, we observe two key limitations: (1) *Entangled edits and lack of faithfulness.* Edits often induce changes in unrelated attributes, heavily altering the background or unintended regions. This leads to a loss of fidelity and identity preservation, thus violating the faithfulness requirement of counterfactual generation. (2) *Per-sample optimization is still required.* Similar to NTI, prompt-learning approaches need separate fine-tuning for each image instance. This makes them impractical for scalable, real-world causal reasoning tasks. Our findings show that relying solely on a frozen diffusion backbone with prompt-tuning is insufficient, as it fails to jointly represent causal semantic attributes and image embeddings. As a result, the model struggles to achieve precise counterfactual reasoning and generation. This highlights the need for an adaptive mechanism within the diffusion model, enhanced with causal semantics.

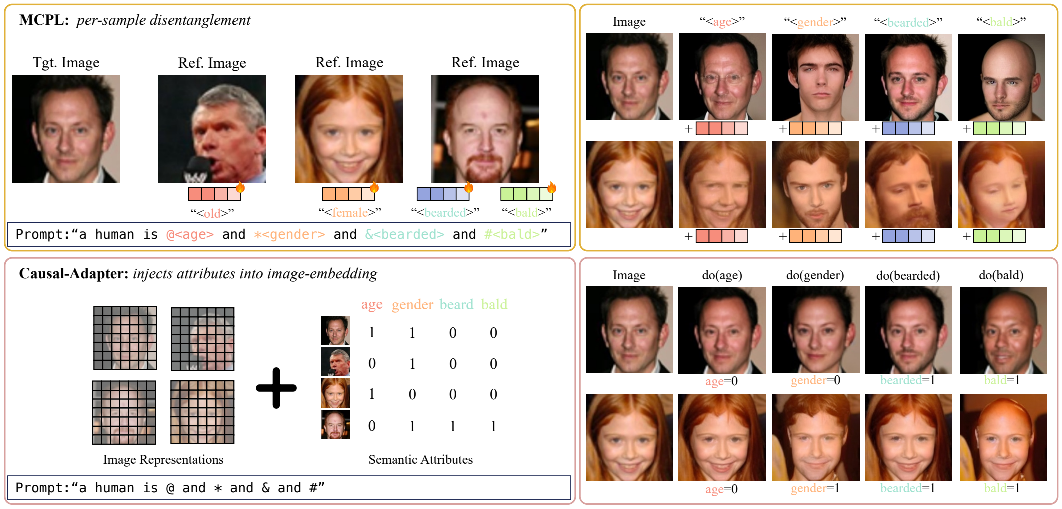
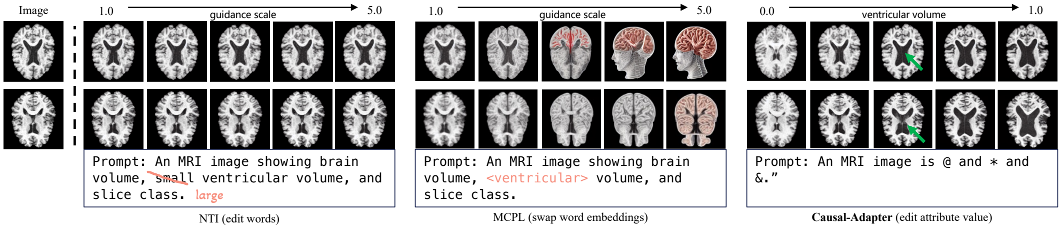


Figure 12: Multi-Concept Prompt Learning (MCPL) as a representative prompt-learning baseline. Each placeholder token (*e.g.*, @ for "young", $*$ for "male") is initialized with a pretrained embedding and jointly optimized across multiple concepts. At test time, counterfactuals are generated by swapping embeddings with target embeddings (*e.g.*, "old", "female"). MCPL can achieve some edits missed by NTI (e.g., baldness, beard), but often entangles unrelated attributes, alters backgrounds, and requires per-instance optimization. In contrast, our vanilla Causal-Adapter injects causal attributes into image embeddings, supporting batch optimization and counterfactual generation via direct attribute interventions.

3. Can existing T2I based methods do fine-grained editing? In causal editing, interventions on a single factor (*e.g.*, increasing object size) should consistently induce predictable and proportional changes in the generated image. However, with current T2I based methods, numerical attributes must be mapped to discrete linguistic tokens, making it difficult to express edits along a continuous range. In practice, the primary way to adjust editing strength is by tuning the classifier-free guidance scale (Ho & Salimans, 2021). As shown in Figure 13, both NTI and MCPL fail to achieve fine-grained anatomical counterfactual editing of brain ventricular volume, even with different guidance scales. Their text-only control mechanisms cannot reliably translate numeric interventions into gradual visual changes. This limitation is particularly problematic in medical imaging domains, where precise, numerically controlled edits (*e.g.*, adjusting brain or ventricular volume in MRI scans) are essential for simulating disease progression.



Prompt: An MRI image showing brain volume, ~~small~~ ventricular volume, and slice class. large

NTI (edit words)

Prompt: An MRI image showing brain volume, <ventricular> volume, and slice class.

MCPL (swap word embeddings)

Prompt: An MRI image is @ and ∗ and &.”

Causal-Adapter (edit attribute value)

Figure 13: Fine-grained anatomical counterfactual editing of brain ventricular volume. NTI and MCPL cannot achieve fine-grained control with text-only prompts. In contrast, our Causal-Adapter produces sharp, localized interventions that smoothly adjust ventricular volume from small to large while preserving subject identity.

C Dataset and Implementation

C.1 Implementation Details

We build our method on the backbone of Stable Diffusion v1.5 (Rombach et al., 2022), using the same hyperparameter settings as the original implementation. The pre-trained weights are trained at a resolution of 256×256. Our adapter network $\ddot{\epsilon}_\psi$ is designed as a half-copy of the diffusion U-Net backbone, consisting of the encoder and bottleneck layers.

Unless specified, experimental setups on CelebA and ADNI follow the benchmark protocol in (Melistas et al., 2024), and CelebA-HQ experiments follow the setup in (Rasal et al., 2025). To ensure fairness, we adopt the same causal mechanism (normalizing flow) as described in the benchmark implementation. Experimental setup on Pendulum follows the configuration in CausalDiffAE (Komanduri et al., 2024c). Since both approaches assume linear causal modeling, we use our constructed simple MLP-based causal mechanism.

We construct prompts in the form of "`a human is @ and ...`" or "`an MRI image is @ and ...`", where the placeholder tokens like "@" is aligned with semantic attributes in the embedding space via PAI. Prefix words (*e.g.* "`a human is`" and "`an image is`") are kept fixed across datasets, as we observed no significant performance differences when switching them.

During training, we apply the proposed CTC loss with a temperature $\tau = 0.2$ and scaling factor $\lambda = 0.0005$ for all datasets. At inference, all images are first generated at 256×256 resolution and then down-sampled via bicubic interpolation to match the resolution of the original benchmarks.

For counterfactual image generation, we employ DDIM inversion (Song et al., 2021). To best preserve identity, we set the inversion guidance scale to 1.0 (no classifier-free guidance (CFG)). For the forward process, we use CFG: CelebA uses $\alpha = 3.0$ and CelebA-HQ uses $\alpha = 1.5$ for effective edits, while ADNI and Pendulum perform well without CFG. Optionally, we incorporate token-level attention guidance manipulation following (Ju et al., 2024), replacing only the intervened token attentions while keeping the others fixed.

All experiments are executed on a single NVIDIA A100 GPU. Training completes within one day, and generating a single counterfactual image takes approximately 5–7 seconds. Table 6 presents detailed training configurations.

Table 6: Experimental settings for our Causal-Adapter across three datasets.

	CelebA	ADNI	Pendulum	CelebA-HQ
Train Set Size	162,770	10,780	5,000	24000
Validation Set Size	19,867	0	500	3000
Test Set Size	19,962	2,240	2,000	3000
Resolution	$256 \times 256 \times 3$	$256 \times 256 \times 3$	$256 \times 256 \times 3$	$256 \times 256 \times 3$
Downsampled Resolution	$64 \times 64 \times 3$	$192 \times 192 \times 1$	$94 \times 94 \times 3$	$64 \times 64 \times 3$
Batch Size	40	40	40	40
Training Steps	200k	100k	50k	100k
Num of Attributes	4	6	4	7
Prompt Template	`"a human is @ and * and & and #"`	`"an MRI image is @ and * and &"`	`"an image is @ and * and & and #"`	`"a human is @ and * and & and # and ! and ? and %"`
CTC Temperature τ	0.2	0.2	0.2	0.2
CTC Scale λ	0.0005	0.0005	0.0005	0.0005
DDIM Steps T	100	100	100	100
Guidance Scale α	3.0	1.0	1.0	1.5
Learning Rate	1e-5			
Optimizer	AdamW (weight decay 1e-2)			
Loss	MSE (noise prediction) + CTC (proposed)			

C.2 METRICS

We detail the counterfactual evaluation metrics used in our experiments below.

Effectiveness. Effectiveness measures how successfully an intervention alters the intended attributes in counterfactual images. To quantitatively evaluate effectiveness, we leverage an anti-causal predictor trained on the observed data distribution for each parent variable of the image x defined in the causal graph (Monteiro et al., 2023). For CelebA and ADNI, we train deep convolutional regressors as anti-causal predictors for continuous attributes (*e.g.*, brain volume, ventricular volume), and deep convolutional classifiers for categorical attributes (*e.g.*, age, gender, beard, bald, slice class). Both models use a ResNet-18 backbone pretrained on ImageNet, following the implementation of (Melistas et al., 2024). For the Pendulum dataset, we train regressors for the four continuous attributes following the implementation of (Komanduri et al., 2024c).

Composition. If an attribute variable y_i is forced to take the same value $\bar{y}_i$ that it would assume without intervention, the intervention should have no effect on any other variables. This corresponds to the *null intervention*, which leaves all variables unchanged and, in the generative setting, can reduce to a standard reconstruction task. To evaluate counterfactual generation under *null intervention*, we perform abduction via DDIM inversion and prediction via DDIM sampling, but skip the action step that edits the original attribute value. For evaluation, we report the MAE distance between the reconstructed and input images, as well as the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), which better reflects perceptual fidelity and content preservation.

Realism. Realism evaluates the perceptual quality of counterfactual images by measuring their similarity to real samples. We adopt the Fréchet Inception Distance (FID) (Heusel et al., 2017), which quantifies the similarity between the distribution of generated counterfactual images and the real dataset. Specifically, real and counterfactual samples are passed through an Inception v3 network (Szegedy et al., 2015) pretrained on ImageNet to extract high-level semantic feature representations, which are then used to compute the FID score.

Minimality. Minimality evaluates whether a counterfactual differs from the factual image only in the intervened parent attribute, ideally leaving all other attributes unaffected. Counterfactual Latent Divergence (CLD) quantifies this by measuring the "distance" between counterfactual and factual images in a latent space (Sanchez & Tsaftaris, 2022). Intuitively, CLD captures a trade-off: the counterfactual should move sufficiently away from the factual class, but not farther than real samples belonging to the counterfactual class. Following Melistas et al. (2024), we compute CLD using an unconditional VAE. Specifically, we measure the KL divergence between the latent distributions of real and counterfactual images. The metric is minimized when both probabilities remain low, reflecting the balance between departing from the factual class while remaining closer to the counterfactual class than unrelated real samples.

C.3 Full Regularization Algorithm

In the following, we present the fully regularized Causal-Adapter in Algorithms 1–2. Below we summarize the training and inference algorithm in high-level description.

Training:

1. *Learn the causal mechanisms F:* Using the predefined adjacency matrix A derived from the causal graph and semantic attributes Y, the model learns how attributes causally influence one another. This produces a standalone SCM that governs how attributes should update when interventions are applied.
2. *Train the conditional adapter:* Each semantic attribute y_i is projected into a token embedding $v_i(y_i)$ through Prompt-Aligned Injection (PAI). With the diffusion model frozen, the adapter $\ddot{\epsilon}_\psi$ is trained to produce residuals r_t conditioned on the attribute embeddings $V(Y)$. A diffusion loss $\mathcal{L}_{\text{DM}}$ ensures correct denoising behaviour, while a contrastive loss $\mathcal{L}_{\text{CTC}}$ encourages disentanglement and suppresses spurious attribute correlations.

Inference (Counterfactual Reasoning):

1. *Abduction:* The input image x is inverted through the diffusion model (via DDIM inversion H_θ) to obtain a latent trajectory $[z_0^\star, \dots, z_T^\star]$ consistent with the observed image.
2. *Action:* A user-specified intervention y'_i is applied to the semantic attributes. The learned causal mechanisms update all causally connected attributes to produce a new attribute set $\bar{y}'_i$ that respects the causal graph.
3. *Prediction:* Starting from the abducted latent $\bar{z}_T$, the model synthesizes a counterfactual image using the conditional adapter $\ddot{\epsilon}_\psi$ and the updated token embeddings $V\left(Y_{\text{do}(y_i=y'_i)}\right)$, yielding the final counterfactual image $\bar{x}$.

Algorithm 1 Causal-Adapter: Training

```
**Input:** Image $x$, semantic attributes $Y = \{y_i\}_{i=1}^K$, binary adjacency matrix $A \in \{0,1\}^{K \times K}$, frozen modules $\{\mathcal{E}, c_\phi, \epsilon_\theta\}$, projector $G = \{g_i\}_{i=1}^K$
**Output:** causal mechanisms $F = \{f_i\}_{i=1}^K$, updated placeholder embeddings $C = \{c_i\}_{i=1}^K$, updated projector $G = \{g_i\}_{i=1}^K$, causal adapter $\ddot{\epsilon}_\psi$.
# Train causal mechanisms $F$
$\bar{y}_i \coloneqq f_i(A_i \odot Y; \omega_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_i^2)$,
$F \coloneqq \arg\min_F \mathcal{L}_{\text{NLL}}(y_i, \bar{y}_i)$
# Construct attribute injected token embeddings
**initialize** placeholder embeddings $C = \{c_i\}_{i=1}^K$,
$v_i(y_i) = c_i + g_i(y_i), i = \{1, \cdots, K\}$
$V(Y) = [v_1(y_1), \ldots, v_K(y_K)]^\top$
# Train conditional adapter
**for** $t = 1$ to $T$ **do**
    Encode latent: $z_t = \mathcal{E}(x, t)$
    Compute residual from causal adapter: $r_t = \ddot{\epsilon}_\psi(z_t, t, V(Y))$
    Update parameters: $\psi, G, C \coloneqq \arg\min_{\psi,G,C} (\mathcal{L}_{\text{DM}} + \mathcal{L}_{\text{CTC}})$
**end for**
**Return** $F = \{f_i\}_{i=1}^K, \ddot{\epsilon}_\psi, V$
```

Algorithm 2 Causal-Adapter: Inference

```
**Input:** Image $x$, semantic attributes $Y$, frozen modules $\{\mathcal{E}, c_\phi, \epsilon_\theta\}$, trained causal mechanisms $F = \{f_i\}_{i=1}^K$, learned placeholder embedding $C = \{c_i\}_{i=1}^K$, trained causal adapter $\ddot{\epsilon}_\psi$, DDIM Inversion operator $H_\theta$ and its generative inverse $H_\theta^{-1}$
**Output:** Counterfactual image $\bar{x}$
$v_i(y_i) = c_i + g_i(y_i), \quad i = \{1, \cdots, K\}$
$V(Y) = [v_1(y_1), \ldots, v_K(y_K)]^\top$
# Abduction - infer inversed latent noise
$z_0^\star = \mathcal{E}(x, 0)$
**for** $t = 0$ to $T - 1$ **do**
    $r_t = \ddot{\epsilon}_\psi(z_t^\star, t, V(Y))$
    $z_{t+1}^\star = H_\theta(z_t^\star, r_t, V(Y), t)$
**end for**
# Action - apply intervention with $do(y_i = y_i')$
$\bar{y}_i' \coloneqq f_i(A_i \odot Y_{do(y_i = y_i')}; \omega_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_i^2), \quad i = \{1, \cdots, K\}$
$v_i(\bar{y}_i') = c_i + g_i(\bar{y}_i'), \quad i = \{1, \cdots, K\}$
$V(Y_{do(y_i = y_i')}) = [v_1(\bar{y}_1'), \ldots, v_K(\bar{y}_K')]^\top$
# Prediction - generate counterfactual
**initialize** $\bar{z}_T \leftarrow z_T^\star$
**for** $t = T$ to 1 **do**
    $\bar{r}_t = \ddot{\epsilon}_\psi(z_t^\star, t, V(Y_{do(y_i = y_i')}))$
    $\bar{z}_{t-1} = H_\theta^{-1}(\bar{z}_t, \bar{r}_t, V(Y_{do(y_i = y_i')}), t)$
**end for**
$\bar{x} = \text{Decode}(\bar{z}_0)$
**Return** $\bar{x}$
```

D Extended Ablation Results

We further investigate the impact of guidance scale (Appendix D.1), DDIM steps (Appendix D.2),, and attention guidance on counterfactual generation(Appendix D.3), and provide additional qualitative evidence (Appendix D.4). To reduce computational cost, these experiments are conducted on the CelebA validation set using 400 random samples.

D.1 Effect of Guidance Scale

We study the influence of the classifier-free guidance scale α on intervention effectiveness, realism, and minimality, as summarized in Table 7 and Figure 14. Increasing α consistently improves intervention effectiveness across attributes, with all variants showing stronger F1-scores as interventions become more easily separable by the anti-causal predictor (classifier). For instance, the fully regularized model (Table 7 rows 11-15) improves Age F1 from 42.0 at $\alpha = 1.0$ to 63.4 at $\alpha = 5.0$, with similar improvement for beard and bald interventions. These results suggest that aligning semantic and spatial features and enforcing disentanglement allow more effective editing, particularly at higher guidance scales. This effectiveness improvement comes at a cost: both FID, reflecting reduced realism and causal minimality (Figure 14). Thus, we adopt $\alpha = 3.0$ as the default setting, where intervention effectiveness is already strong across all variants (Table 7 rows 3,8,13) while FID and CLD remain relatively low. Further increasing α beyond this point will leads to marginal improvements in effectiveness but rapidly increases FID and CLD. Qualitative example is shown in Figure 16.

Table 7: Influence of guidance scale on counterfactual effectiveness. α denotes the classifier-free guidance scale. “Ours” refers to the plain Causal-Adapter, “Ours*” to the regularized Causal-Adapter (+PAI), and “Ours**” to the fully regularized Causal-Adapter (+PAI+CTC).

	Method	**Age (a) F1 ↑**				**Gender (g) F1 ↑**				**Beard (br) F1 ↑**				**Bald (bl) F1 ↑**			
		$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
1.	Ours($\alpha = 1.0$)	38.0	79.9	86.0	77.8	96.1	97.1	65.1	86.6	95.4	75	32.2	93.7	100	39.7	66.7	22.9
2.	Ours($\alpha = 2.0$)	38.1	80.2	95.8	75.5	95.5	96.7	61	82.3	95	78	39.2	93	100	50	85.7	32
3.	Ours($\alpha = 3.0$)	38.2	81.1	85.3	73.9	95.3	95.6	59.7	77.8	93.9	81.3	42.1	92.2	100	61.3	85.7	39.8
4.	Ours($\alpha = 4.0$)	37.2	80.1	84.8	72.6	93.7	95.2	58.7	75.4	92.8	82.4	45.2	91.3	100	66.7	75	50.2
5.	Ours($\alpha = 5.0$)	37.1	80.7	83.1	71	92.8	94.8	57.8	71.5	92.3	84.1	43.4	90.1	80	67	85.7	54.9
6.	Ours*($\alpha = 1.0$)	37.5	78.1	87.6	70.8	96.1	96.8	71.7	86.1	95.1	77.4	29	93.3	100	44.2	85.7	22.5
7.	Ours*($\alpha = 2.0$)	46.8	85.4	93	84.3	99	99.8	72.8	95.5	98.6	93.7	43.8	97	50	84.8	50	47.2
8.	Ours*($\alpha = 3.0$)	48.5	90.8	92.7	89.4	99	100	72.1	97	99.4	97.4	50.7	97.3	50	91.8	66.7	55.8
9.	Ours*($\alpha = 4.0$)	48.8	90.2	92.2	90.6	99.7	100	71.1	97.5	99.7	98.6	53	98.6	50	94.7	54.5	57.7
10.	Ours*($\alpha = 5.0$)	48.6	90	91.6	91.8	100	100	70.5	97.5	100	99.1	54.1	99	66.7	94.3	40	58.2
11.	Ours**($\alpha = 1.0$)	42.0	80.2	86.0	73.6	96.2	97.7	67.8	81.5	94.9	78.0	31.5	92.9	1	45.9	75.0	29.4
12.	Ours**($\alpha = 2.0$)	51.9	88.0	93.4	84.9	98.1	99.6	72.3	90.1	98.8	91.0	46.4	97.0	80.0	84.2	75.0	55.1
13.	Ours**($\alpha = 3.0$)	58.8	91.3	94.4	89.4	100	99.8	74.5	91.1	99.7	96.4	53.4	97.8	80.0	92.4	66.7	59.8
14.	Ours**($\alpha = 4.0$)	61.3	89.9	94.7	91.5	99.7	100	77.4	92.2	100	97.2	55.6	97.6	75.0	92.4	66.7	59.8
15.	Ours**($\alpha = 5.0$)	63.4	90.7	95.2	91.4	99.7	100	79.0	93.5	100	98.0	57.0	97.8	75.0	92.4	66.7	63.0

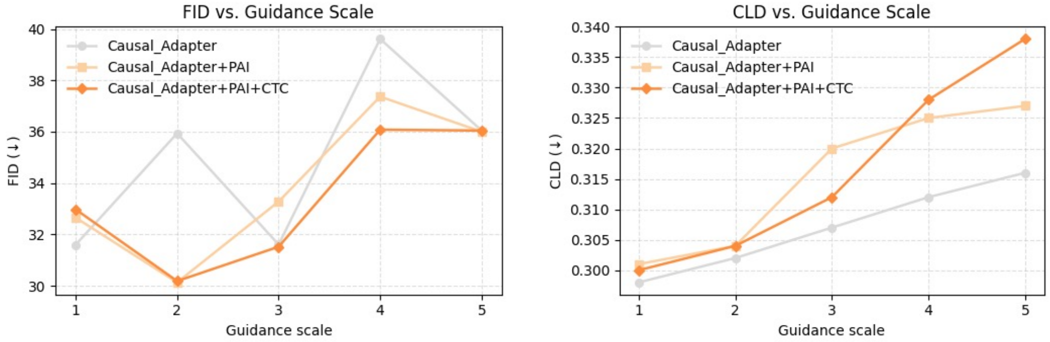


Figure 14: Impact of guidance scale on FID and CLD across three Causal-Adapter variants. Note that FID values here are higher than in the main manuscript due to the smaller evaluation set, which shifts the distribution mean and variance. The implied relative trends in image quality remain consistent.

D.2 Effect of DDIM Steps

We further evaluate the effect of DDIM steps on counterfactual effectiveness using our fully regularized model (Table 8). As T increases, intervention effectiveness remains stable, indicating that performance is insensitive to the number of denoising steps. Realism and minimality also show negligible variation. The main drawback of larger T lies in increased inference time: generating one counterfactual takes 5–7 seconds at $T = 100$, about 20 seconds at $T = 200$, and nearly 40 seconds at $T = 500$. Balancing efficiency and accuracy, we adopt $T = 100$ as the default setting for all experiments.

Table 8: Influence of DDIM steps under $\alpha = 3.0$ on counterfactual effectiveness. T denote the used steps.

	Method	Age (a) F1 ↑				Gender (g) F1 ↑				Beard (br) F1 ↑				Bald (bl) F1 ↑			
		$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
1.	Ours**$(T = 50)$	57.1	90.5	93.8	89.4	99.7	99.6	75.7	91.6	99.7	96.1	49.1	98	80.0	89.4	75.0	60
2.	Ours**$(T = 100)$	58.8	91.3	94.4	89.4	100	99.8	74.5	91.1	99.7	96.4	53.4	97.8	80.0	92.4	66.7	61.8
3.	Ours**$(T = 200)$	60.5	91.7	94.9	89.6	100	99.8	74.1	91.4	99.7	96.4	54.3	97.4	80.0	91.9	66.7	61.3
4.	Ours**$(T = 500)$	60.7	91.6	94.7	90.0	100	99.8	73.8	91.9	99.7	96.1	54.8	98.0	85.7	91.9	66.7	61.3

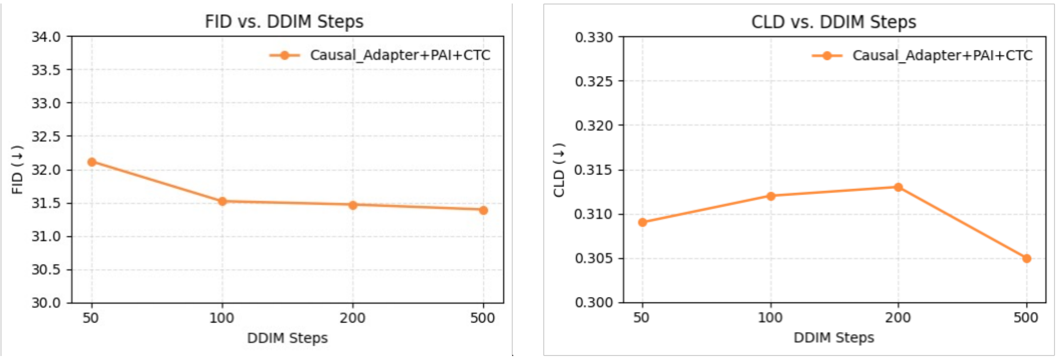


Figure 15: Impact of DDIM steps on FID and CLD

D.3 Investigation of Attention Guidance

Attention guidance (AG) has been proposed as an external mechanism to enforce localized edits. While our primary contribution lies in Causal-Adapter and its regularizers, we further examine whether AG can complement different variants of our model. As shown in Tables 9–10, incorporating AG sometimes reduces intervention effectiveness, particularly for the plain adapter (Table 9 rows 1,3). For example, Age F1 drops sharply from 38.2 to 28.1 when switching from global to local editing, reflecting misaligned attention maps that disrupt counterfactual consistency. In contrast, the regularized variants exhibit only mild decreases (Table 9 rows 2,5), indicating that aligning semantic and spatial features in diffusion latents via PAI is a prerequisite for stable local editing. Our fully regularized model (Table 9 rows 3,6) not only maintains strong intervention effectiveness but also achieves the lowest FID (31.216) with AG, highlighting that PAI and CTC help produce more precise attention maps. By comparison, applying AG to the plain adapter actually worsens FID and CLD (Table 10 row 4), suggesting that noisy or misaligned attention can harm editing quality. Qualitative evidence is provided in Figures 17–18.

We emphasize that while AG can be beneficial for identity preservation in human faces (CelebA), in other domains such as ADNI or Pendulum, our model already produces accurate, faithful, and identity-preserving counterfactuals without AG. This underscores that our contribution lies not in AG itself, but in designing Causal-Adapter such that causal attributes are naturally aligned with token embeddings, enabling both disentanglement and, if desired, effective integration with AG for localized editing.

Table 9: Effectiveness of Causal-Adapter variants with and without attention guidance.

	Method	**Age (a) F1 ↑**				**Gender (g) F1 ↑**				**Beard (br) F1 ↑**				**Bald (bl) F1 ↑**			
		$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
without attention guidance for global editing																	
1.	Ours	38.2	81.1	85.3	73.9	95.3	95.6	59.7	77.8	93.9	81.3	42.1	92.2	100	61.3	85.7	39.8
2.	Ours*	48.5	90.8	92.7	89.4	99	100	72.1	97	99.4	97.4	50.7	97.3	50	91.8	66.7	55.8
3.	Ours**	58.8	91.3	94.4	89.4	100	99.8	74.5	91.1	99.7	96.4	53.4	97.8	80.0	92.4	66.7	61.8
with attention guidance for local editing																	
4.	Ours	28.1	79.4	83.2	63.3	94.7	93.4	62.4	78.1	96.5	80.6	40.4	89.4	66.7	70.6	33.3	34.5
5.	Ours*	49.4	88.8	91.2	73.9	98.8	96.4	71.1	92.8	99.5	96.1	47.4	94.6	22.2	76.2	40	52.4
6.	Ours**	57.1	88.7	94	77	99.4	100	73.3	90.4	99.7	95.6	49.4	95.1	66.7	80.7	50	54.4

Table 10: Realism and minimality of Causal-Adapter variants with and without attention guidance

	Method	**Realism**	**Minimality**
		FID ↓	CLD ↓
without attention guidance for global editing			
1.	Ours	31.604	0.307
2.	Ours*	33.278	0.320
3.	Ours**	31.52	0.312
with attention guidance for localized editing			
4.	Ours	33.849	0.311
5.	Ours**	31.885	0.299
6.	Ours**	31.216	0.300

D.4 Qualitative Evidence

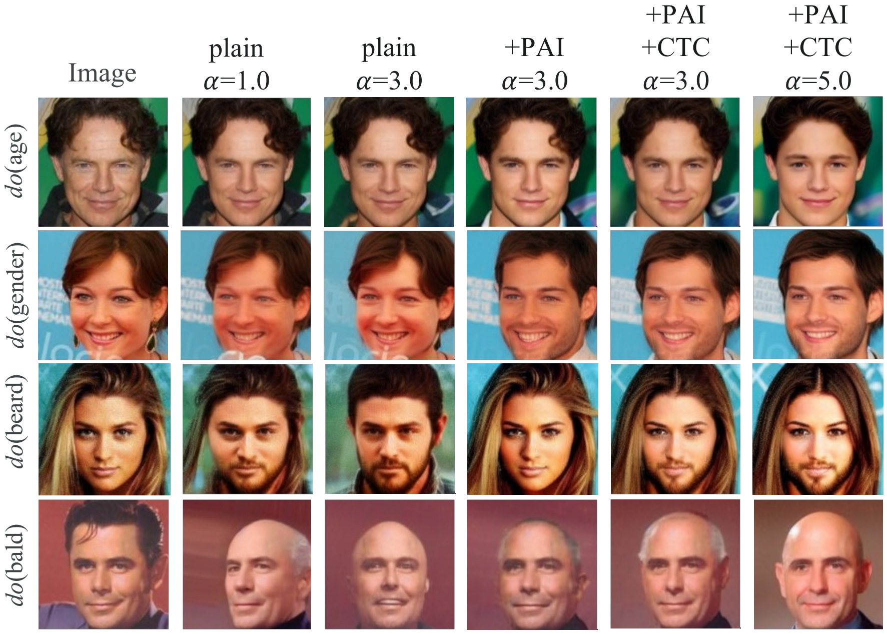


Figure 16: Counterfactuals from Causal-Adapter variants under different guidance scales. The plain variant shows weak effectiveness at $\alpha = 1.0$ and 3.0. Adding PAI strengthens the editing signal, while further incorporating CTC yields the most effective edits with strong identity preservation. At $\alpha = 5.0$, the counterfactuals become slightly over-edited.

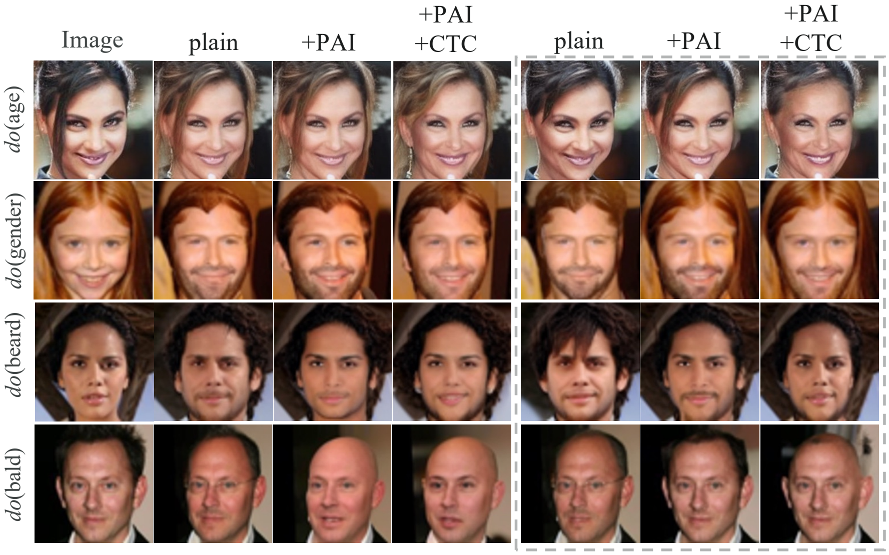


Figure 17: Full ablation visualizations with optional attention guidance (AG). Causal-Adapter with the two regularizers reduces spurious correlations, AG can enhance identity preservation through localized editing. Dotted boxes indicate results with AG-based localized edits.

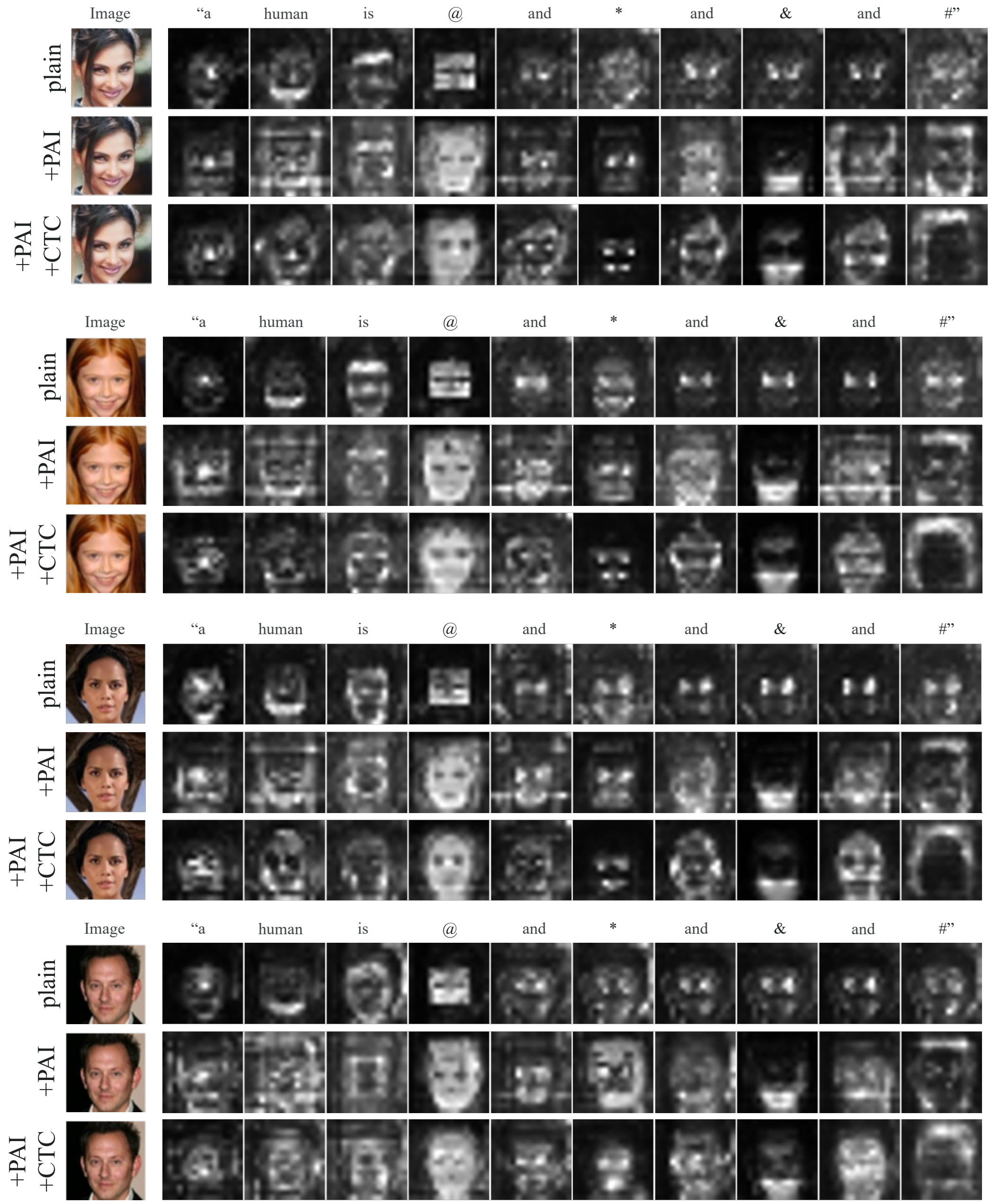


Figure 18: Average cross-attention maps from Causal-Adapter variants. Tokens denote attributes: “@” for age, “*” for gender, “&” for beard, and “#” for bald. The plain adapter fails to align semantics with spatial features, producing poor maps. Adding PAI improves alignment but some tokens (e.g., “*”, “#”) remain entangled. With both PAI and CTC, token embedding disentanglement is enforced, and attentions are clearly localized to the semantic regions.

E Additional Results

E.1 Pendulum

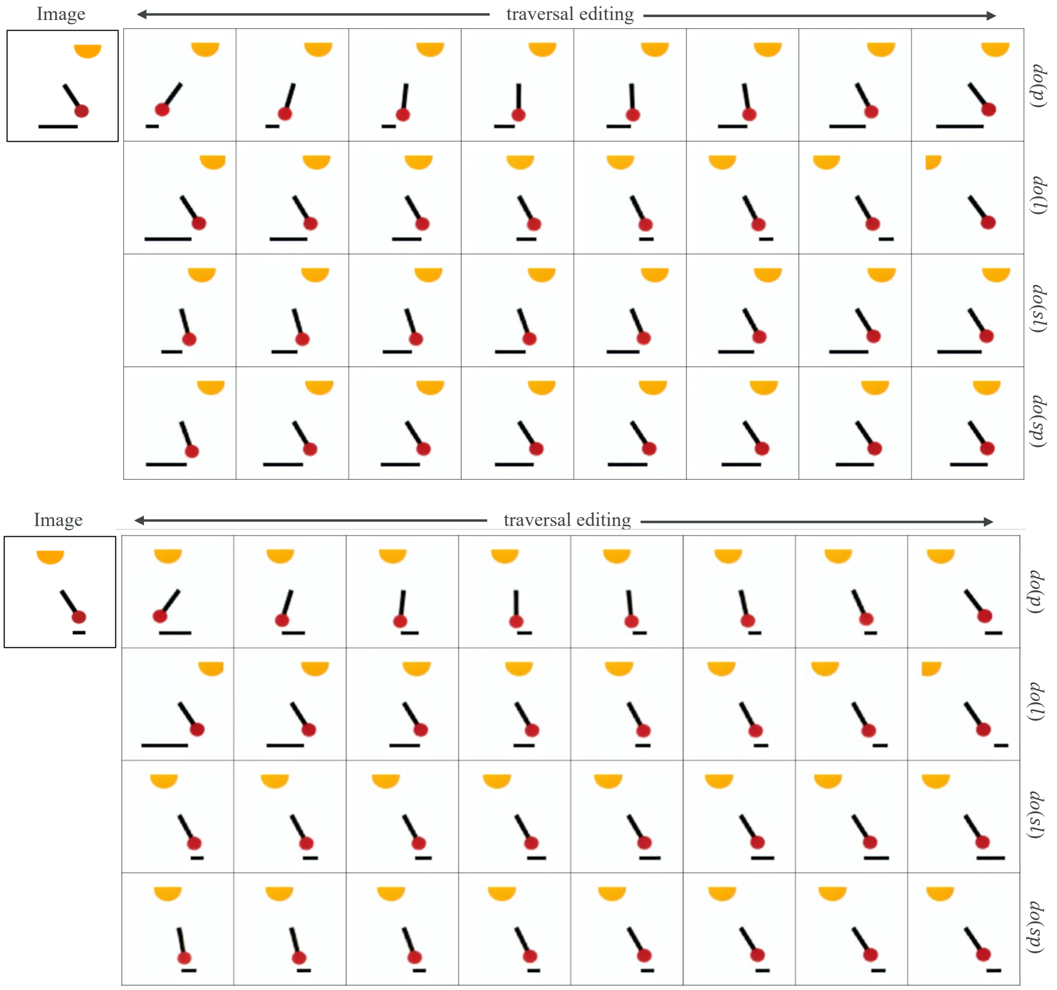


Figure 19: Pendulum counterfactuals from Causal-Adapter. p for pendulum, l for light, sl for shadow length and sp for shadow position.Traversal editing across four attributes demonstrates that our method produces high-quality, fine-grained generations of attribute variations.

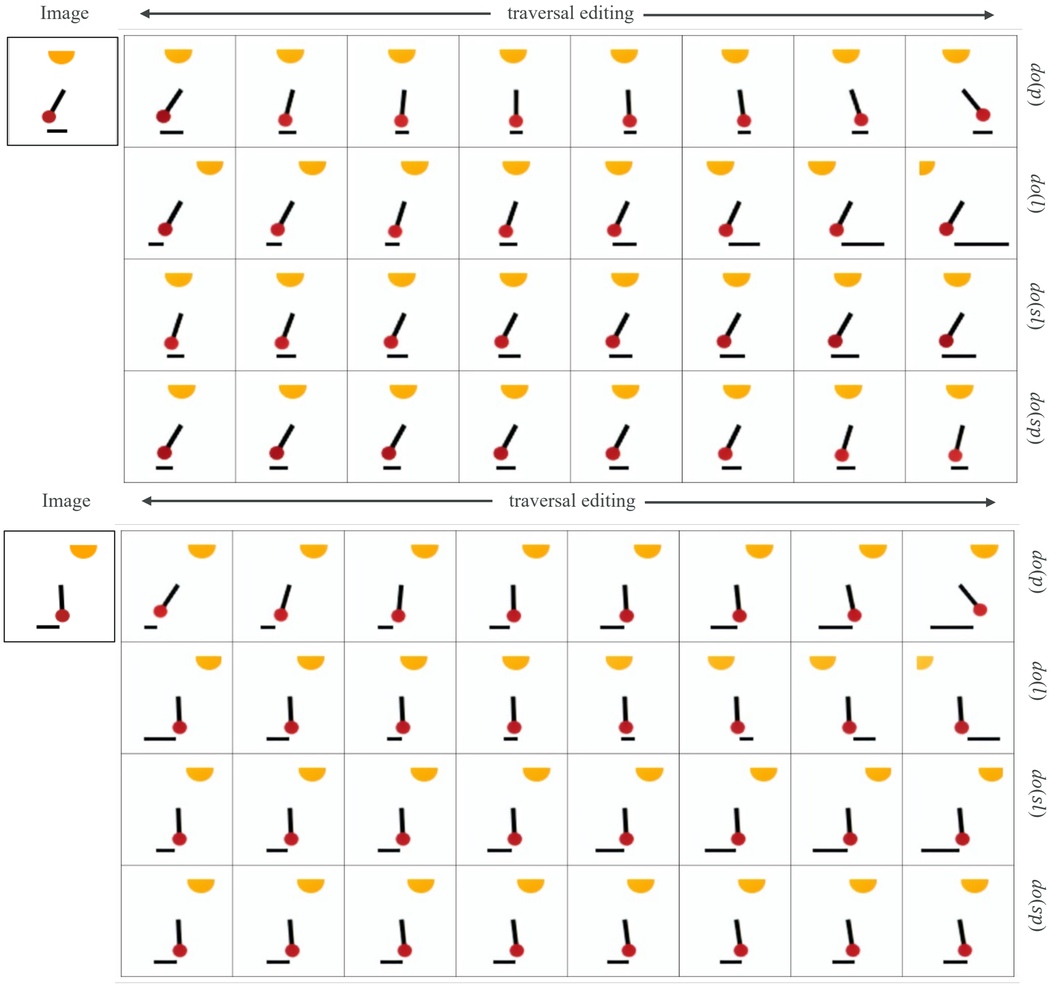


Figure 20: Pendulum counterfactuals from Causal-Adapter. p for pendulum, l for light, sl for shadow length and sp for shadow position.

E.2 CelebA

We report results across three random seeds on the CelebA dataset, with corresponding means and standard deviations shown in Table 11-13. In Table 13, we additionally provide LPIPS distances between each counterfactual and its original image as a measure of identity preservation under different attribute interventions.

Table 11: Intervention effectiveness on CelebA test set (3 seeds). Age and Gender F1 under four interventions.

	Method	**Age** (a) **F1** ↑				**Gender** (g) **F1** ↑			
		$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
1.	VAE	$35.0_{0.04}$	$78.2_{0.02}$	$81.6_{0.02}$	$81.9_{0.02}$	$97.7_{0.01}$	$90.9_{0.02}$	$95.9_{0.02}$	$\mathbf{97.3}_{0.01}$
2.	HVAE	$\mathbf{65.4}_{0.10}$	$89.3_{0.04}$	$90.8_{0.03}$	$\mathbf{89.9}_{0.03}$	$98.8_{0.02}$	$94.9_{0.03}$	$\mathbf{99.4}_{0.01}$	$95.0_{0.03}$
3.	GAN	$41.3_{0.04}$	$71.0_{0.02}$	$81.8_{0.02}$	$79.9_{0.01}$	$95.2_{0.01}$	$98.2_{0.01}$	$92.0_{0.01}$	$96.1_{0.01}$
4.	Ours	$58.5_{0.14}$	$\mathbf{89.9}_{0.35}$	$94.0_{0.07}$	$89.4_{0.00}$	$\mathbf{99.6}_{0.00}$	$\mathbf{99.9}_{0.00}$	$74.5_{0.07}$	$92.5_{0.21}$
with attention guidance for localized editing									
5.	Ours (AG)	$57.1_{0.14}$	$89.5_{1.76}$	$\mathbf{94.5}_{0.00}$	$81.5_{0.63}$	$\mathbf{99.6}_{0.00}$	$99.7_{0.00}$	$73.8_{0.56}$	$89.2_{0.14}$

Table 12: Intervention effectiveness on CelebA test set (3 seeds). Beard and Bald F1 under four interventions.

	Method	**Beard** (br) **F1** ↑				**Bald** (bl) **F1** ↑			
		$do(a)$	$do(g)$	$do(br)$	$do(bl)$	$do(a)$	$do(g)$	$do(br)$	$do(bl)$
1.	VAE	$94.4_{0.01}$	$82.8_{0.03}$	$29.6_{0.05}$	$94.5_{0.02}$	$2.3_{0.03}$	$49.6_{0.05}$	$4.5_{0.04}$	$41.2_{0.03}$
2.	HVAE	$95.2_{0.03}$	$95.1_{0.03}$	$44.1_{0.11}$	$91.6_{0.04}$	$2.0_{0.05}$	$86.0_{0.05}$	$4.5_{0.07}$	$61.1_{0.04}$
3.	GAN	$90.8_{0.01}$	$83.8_{0.02}$	$23.3_{0.03}$	$90.7_{0.01}$	$2.1_{0.02}$	$82.0_{0.02}$	$5.5_{0.02}$	$49.2_{0.02}$
4.	Ours	$\mathbf{99.7}_{0.00}$	$96.1_{0.00}$	$\mathbf{52.1}_{0.21}$	$\mathbf{98.1}_{0.07}$	$\mathbf{59.2}_{0.91}$	$\mathbf{91.2}_{0.28}$	$\mathbf{74.5}_{2.75}$	$\mathbf{58.8}_{0.35}$
with attention guidance for localized editing									
5.	Ours (AG)	$99.7_{0.07}$	$\mathbf{96.8}_{0.00}$	$48.2_{0.14}$	$96.6_{0.07}$	$38.7_{1.90}$	$78.4_{0.49}$	$47.7_{1.76}$	$51.4_{0.49}$

Table 13: Identity preservation, realism, and minimality on the CelebA test set. VAE, HVAE, and GAN are reported as single-run results from prior work, while our methods report mean_{std} over three random seeds. "–" indicates that LPIPS was not reported for the corresponding method.

	Method	**Identity Preservation** (LPIPS ↓)				**Realism**	**Minimality**
		Age	Gender	Bearded	Bald	FID ↓	CLD ↓
1.	VAE	–	–	–	–	59.393	**0.299**
2.	HVAE	–	–	–	–	35.712	0.305
3.	GAN	–	–	–	–	27.861	0.304
4.	Ours	$0.087_{0.0066}$	$0.157_{0.0067}$	$0.0797_{0.0060}$	$0.157_{0.0073}$	$8.152_{0.365}$	$0.310_{0.001}$
Ours with attention guidance for improved realism and minimality							
5.	Ours (AG)	$\mathbf{0.061}_{0.0171}$	$\mathbf{0.072}_{0.0178}$	$\mathbf{0.0330}_{0.0088}$	$\mathbf{0.109}_{0.0192}$	$\mathbf{5.213}_{0.173}$	$0.301_{0.001}$

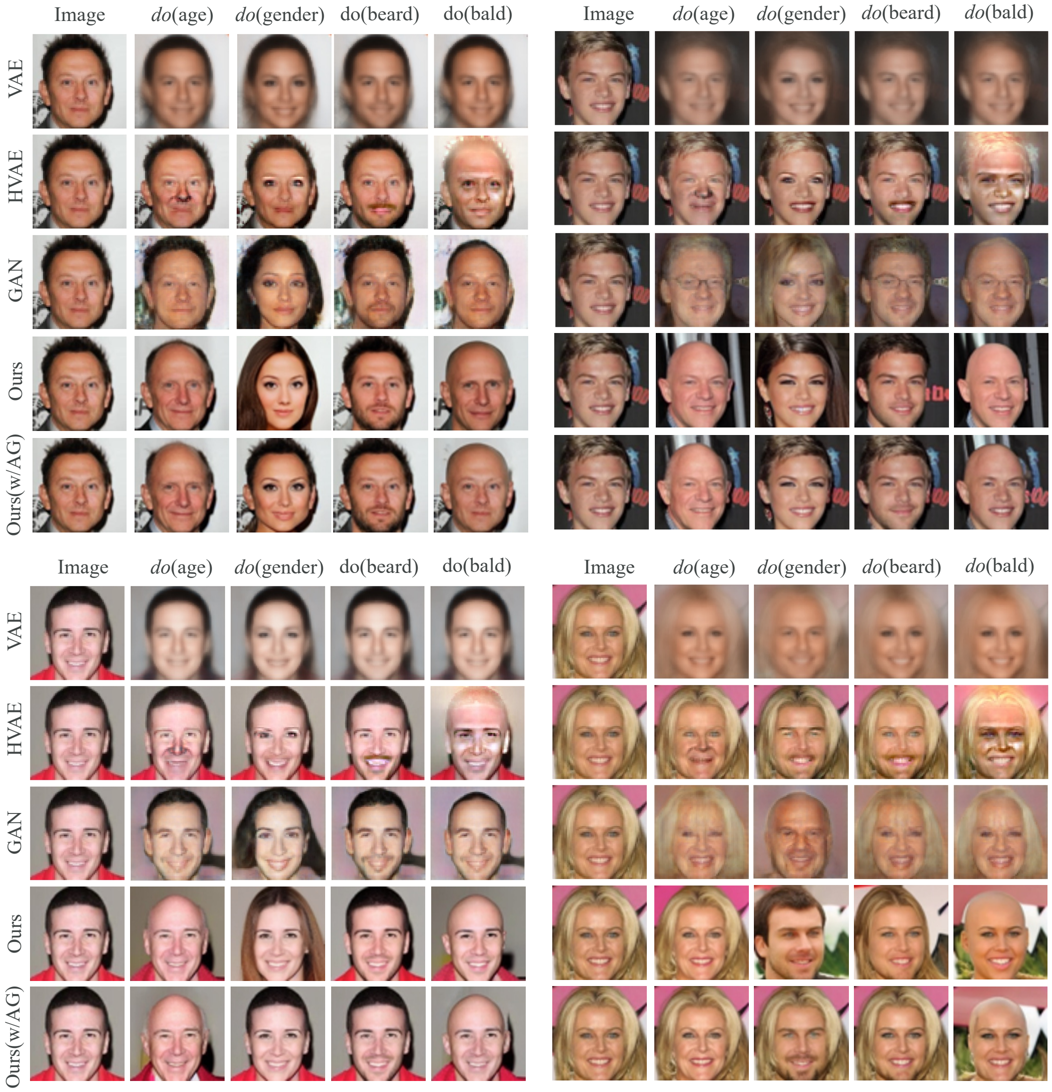


Figure 21: Additional counterfactual results on the CelebA dataset (with edit samples selected in a non-cherrypicked manner following (Melistas et al., 2024)). Our Causal-Adapter effectively disentangles target attributes compared with prior methods and achieve faithful counterfactual generations.

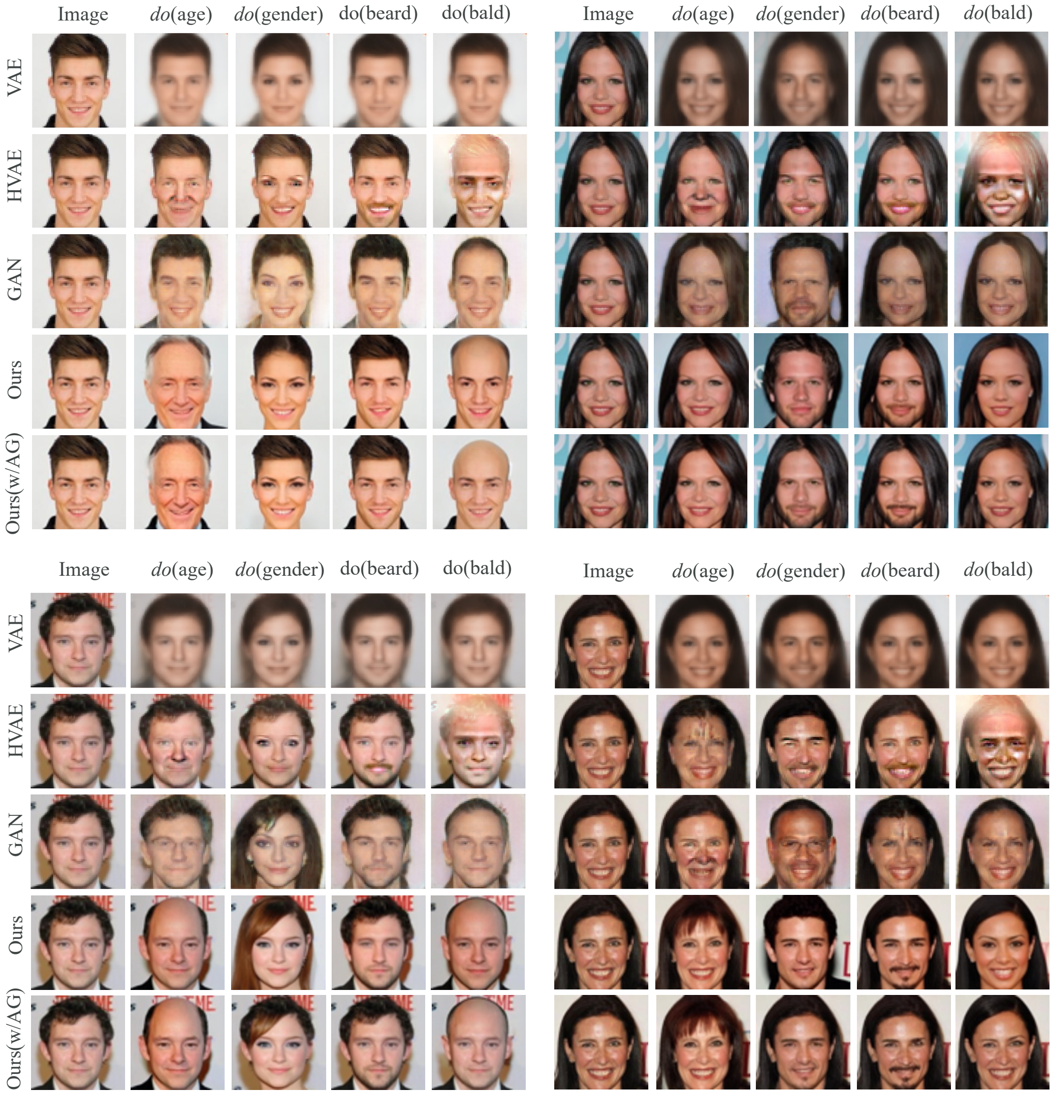


Figure 22: Additional counterfactual results on the CelebA dataset (with edit samples selected in a non-cherrypicked manner following (Melistas et al., 2024)).

E.3 ADNI

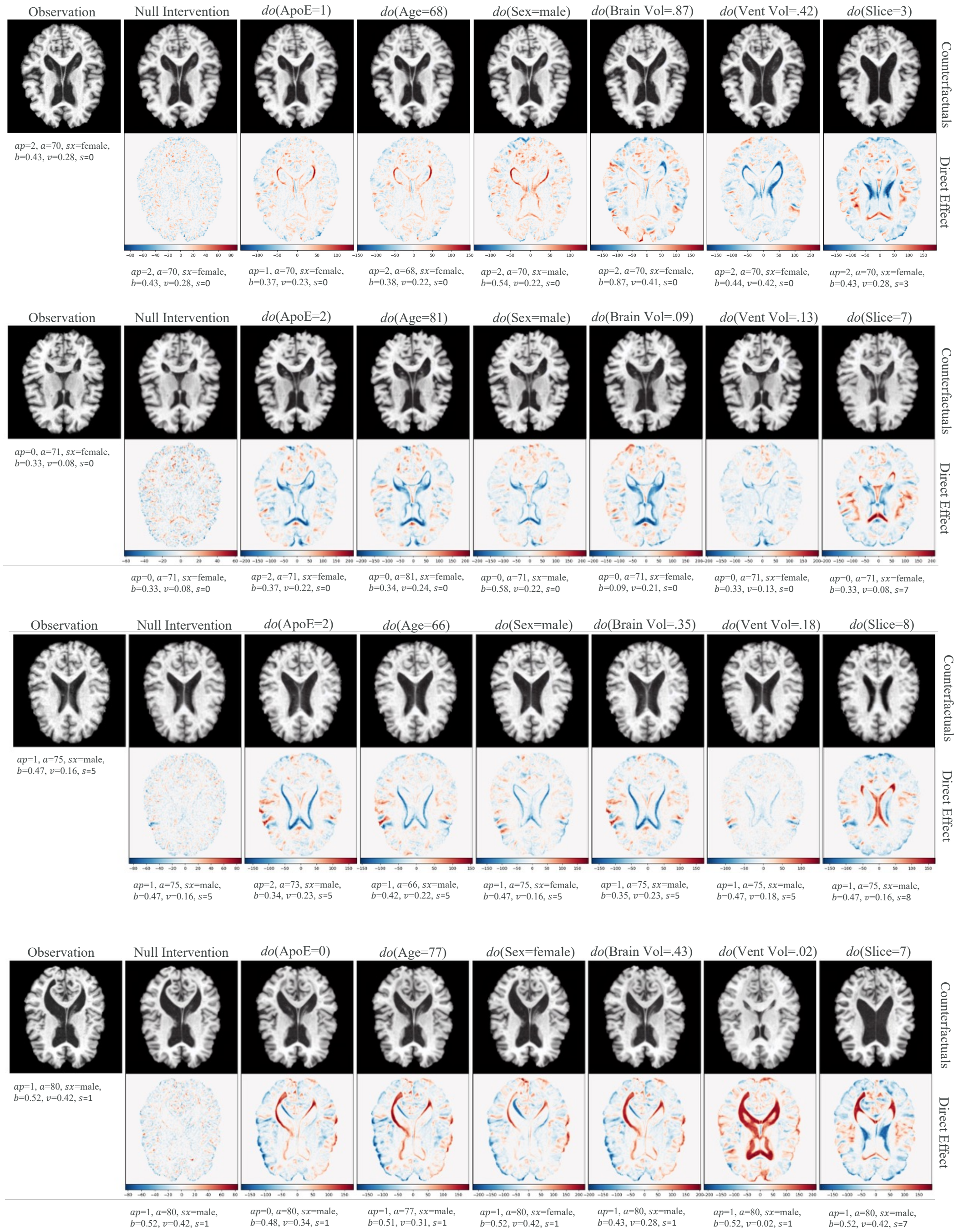

Observation
Null Intervention
do(ApoE=1)
do(Age=68)
do(Sex=male)
do(Brain Vol=.87)
do(Vent Vol=.42)
do(Slice=3)
Counterfactuals
Direct Effect
ap=2, a=70, sx=female, b=0.43, v=0.28, s=0
ap=2, a=70, sx=female, b=0.43, v=0.28, s=0
ap=1, a=70, sx=female, b=0.37, v=0.23, s=0
ap=2, a=68, sx=female, b=0.38, v=0.22, s=0
ap=2, a=70, sx=male, b=0.54, v=0.22, s=0
ap=2, a=70, sx=female, b=0.87, v=0.41, s=0
ap=2, a=70, sx=female, b=0.44, v=0.42, s=0
ap=2, a=70, sx=female, b=0.43, v=0.28, s=3
Observation
Null Intervention
do(ApoE=2)
do(Age=81)
do(Sex=male)
do(Brain Vol=.09)
do(Vent Vol=.13)
do(Slice=7)
Counterfactuals
Direct Effect
ap=0, a=71, sx=female, b=0.33, v=0.08, s=0
ap=0, a=71, sx=female, b=0.33, v=0.08, s=0
ap=2, a=71, sx=female, b=0.37, v=0.22, s=0
ap=0, a=81, sx=female, b=0.34, v=0.24, s=0
ap=0, a=71, sx=male, b=0.58, v=0.22, s=0
ap=0, a=71, sx=female, b=0.09, v=0.21, s=0
ap=0, a=71, sx=female, b=0.33, v=0.13, s=0
ap=0, a=71, sx=female, b=0.33, v=0.08, s=7
Observation
Null Intervention
do(ApoE=2)
do(Age=66)
do(Sex=male)
do(Brain Vol=.35)
do(Vent Vol=.18)
do(Slice=8)
Counterfactuals
Direct Effect
ap=1, a=75, sx=male, b=0.47, v=0.16, s=5
ap=1, a=75, sx=male, b=0.47, v=0.16, s=5
ap=2, a=73, sx=male, b=0.34, v=0.23, s=5
ap=1, a=66, sx=male, b=0.42, v=0.22, s=5
ap=1, a=75, sx=female, b=0.47, v=0.16, s=5
ap=1, a=75, sx=male, b=0.35, v=0.23, s=5
ap=1, a=75, sx=male, b=0.47, v=0.18, s=5
ap=1, a=75, sx=male, b=0.47, v=0.16, s=8
Observation
Null Intervention
do(ApoE=0)
do(Age=77)
do(Sex=female)
do(Brain Vol=.43)
do(Vent Vol=.02)
do(Slice=7)
Counterfactuals
Direct Effect
ap=1, a=80, sx=male, b=0.52, v=0.42, s=1
ap=1, a=80, sx=male, b=0.52, v=0.42, s=1
ap=0, a=80, sx=male, b=0.48, v=0.34, s=1
ap=1, a=77, sx=male, b=0.51, v=0.31, s=1
ap=1, a=80, sx=female, b=0.52, v=0.42, s=1
ap=1, a=80, sx=male, b=0.43, v=0.28, s=1
ap=1, a=80, sx=male, b=0.52, v=0.02, s=1
ap=1, a=80, sx=male, b=0.52, v=0.42, s=7

Figure 23: Additional counterfactual results from random interventions on each attribute in the ADNI dataset (non-cherrypicked). We observe localized changes consistent with the performed interventions and the assumed causal graph. Importantly, the identity of the original observation is well preserved, demonstrating the effectiveness of Causal-Adapter.

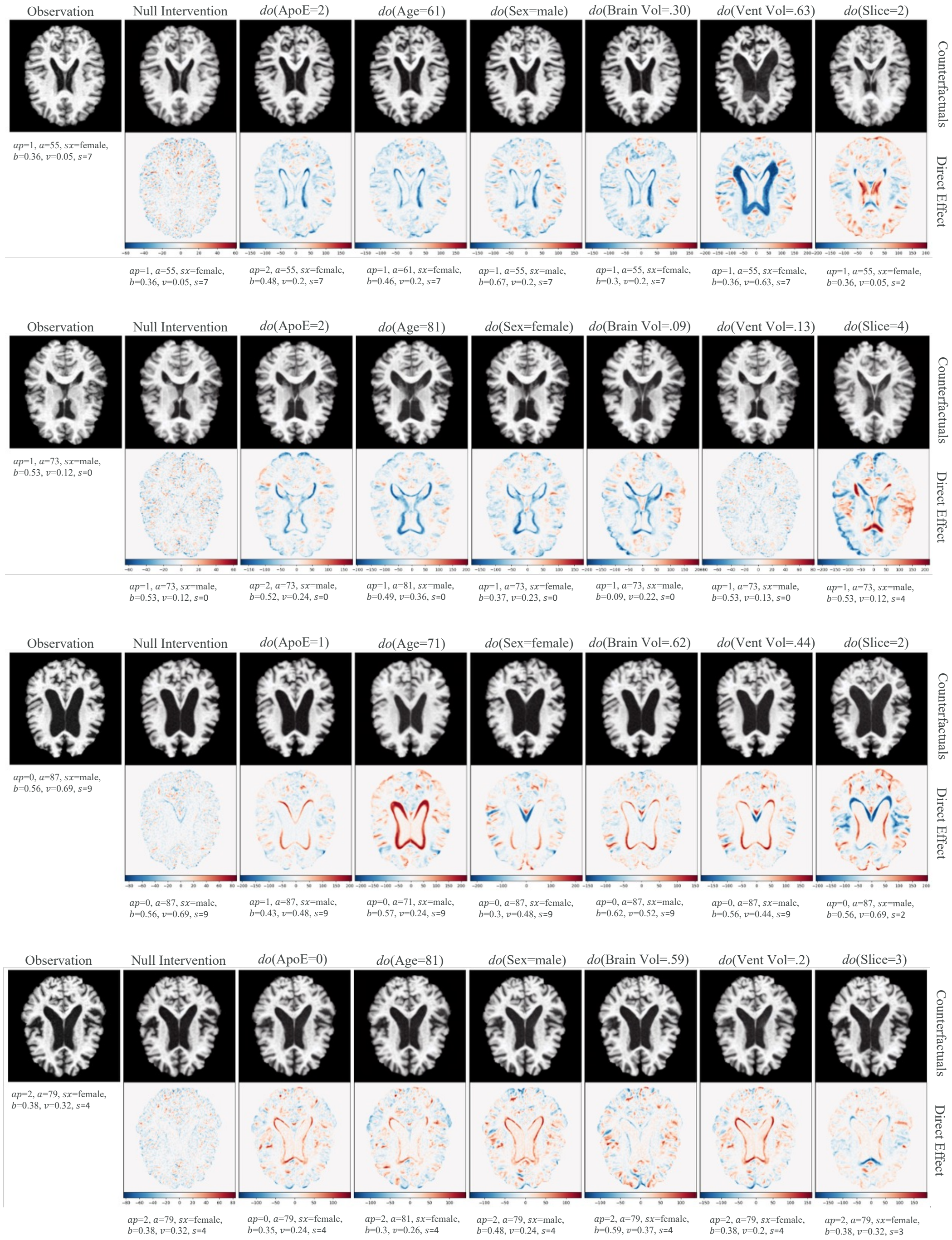

Observation
Null Intervention
do(ApoE=2)
do(Age=61)
do(Sex=male)
do(Brain Vol=.30)
do(Vent Vol=.63)
do(Slice=2)
Counterfactuals
Direct Effect
ap=1, a=55, sx=female, b=0.36, v=0.05, s=7
ap=1, a=55, sx=female, b=0.36, v=0.05, s=7
ap=2, a=55, sx=female, b=0.48, v=0.2, s=7
ap=1, a=61, sx=female, b=0.46, v=0.2, s=7
ap=1, a=55, sx=male, b=0.67, v=0.2, s=7
ap=1, a=55, sx=female, b=0.3, v=0.2, s=7
ap=1, a=55, sx=female, b=0.36, v=0.63, s=7
ap=1, a=55, sx=female, b=0.36, v=0.05, s=2
Observation
Null Intervention
do(ApoE=2)
do(Age=81)
do(Sex=female)
do(Brain Vol=.09)
do(Vent Vol=.13)
do(Slice=4)
Counterfactuals
Direct Effect
ap=1, a=73, sx=male, b=0.53, v=0.12, s=0
ap=1, a=73, sx=male, b=0.53, v=0.12, s=0
ap=2, a=73, sx=male, b=0.52, v=0.24, s=0
ap=1, a=81, sx=male, b=0.49, v=0.36, s=0
ap=1, a=73, sx=female, b=0.37, v=0.23, s=0
ap=1, a=73, sx=male, b=0.09, v=0.22, s=0
ap=1, a=73, sx=male, b=0.53, v=0.13, s=0
ap=1, a=73, sx=male, b=0.53, v=0.12, s=4
Observation
Null Intervention
do(ApoE=1)
do(Age=71)
do(Sex=female)
do(Brain Vol=.62)
do(Vent Vol=.44)
do(Slice=2)
Counterfactuals
Direct Effect
ap=0, a=87, sx=male, b=0.56, v=0.69, s=9
ap=0, a=87, sx=male, b=0.56, v=0.69, s=9
ap=1, a=87, sx=male, b=0.43, v=0.48, s=9
ap=0, a=71, sx=male, b=0.57, v=0.24, s=9
ap=0, a=87, sx=female, b=0.3, v=0.48, s=9
ap=0, a=87, sx=male, b=0.62, v=0.52, s=9
ap=0, a=87, sx=male, b=0.56, v=0.44, s=9
ap=0, a=87, sx=male, b=0.56, v=0.69, s=2
Observation
Null Intervention
do(ApoE=0)
do(Age=81)
do(Sex=male)
do(Brain Vol=.59)
do(Vent Vol=.2)
do(Slice=3)
Counterfactuals
Direct Effect
ap=2, a=79, sx=female, b=0.38, v=0.32, s=4
ap=2, a=79, sx=female, b=0.38, v=0.32, s=4
ap=0, a=79, sx=female, b=0.35, v=0.24, s=4
ap=2, a=81, sx=female, b=0.3, v=0.26, s=4
ap=2, a=79, sx=male, b=0.48, v=0.24, s=4
ap=2, a=79, sx=female, b=0.59, v=0.37, s=4
ap=2, a=79, sx=female, b=0.38, v=0.2, s=4
ap=2, a=79, sx=female, b=0.38, v=0.32, s=3

Figure 24: Additional counterfactual results from random interventions on each attribute in the ADNI dataset

E.4 CelebA-HQ

Additional counterfactual and reversal results. Supplementary counterfactual and reversal examples generated by Causal-Adapter are provided in Figure 25–26. These visualizations demonstrate faithful interventions and strong identity preservation across diverse attribute edits.

Stress test with compositional interventions. Following the settings of Rasal et al. (2025), we concatenated four confounding attributes (Male, Wearing Lipstick, Bald, Wearing Hat) during training to mitigate cross-attribute spurious correlations. Based on the disentangled attributes, we perform a compositional stress test in which we select five non-overlapping attributes (smile, mouth-open, gender, glasses, hat) and progressively apply multiple interventions, i.e., $do(\text{smile})$, $do(\text{smile}, \text{mouth})$, etc. Results in Figure 27–28 show that Causal-Adapter produces plausible counterfactuals under progressively complex intervention sets while maintaining strong identity consistency.

Generalization test (SD3 backbone). To further validate generalizability, we apply Causal-Adapter to a different diffusion backbone, Stable Diffusion 3 (SD3) (Esser et al., 2024), based on the Diffusion Transformer (DiT) architecture. SD3 employs three text encoders that we inject attribute information via PAI across all token-embedding streams. The diffusion loss $\mathcal{L}_{\text{DM}}$ is replaced by a flow-matching loss following the SD3 training recipe. During inference, we adopt FlowEdit (Kulikov et al., 2025), an inversion-free editing method designed for flow-matching models. As shown in the manuscript and in Figure 29–31, Causal-Adapter successfully tames the SD3 backbone and produces causally faithful counterfactuals. This confirms that our method is not tied to DDIM inversion and naturally extends across different T2I generative families. Other inversion techniques, such as null-text inversion (Mokady et al., 2023), can also be integrated, but they require per-sample optimization and are therefore unsuitable for large-scale benchmarks.

Overall, these additional experiments show that Causal-Adapter is a simple yet effective framework for counterfactual image generation. It is modular, generalizable, and scalable across multiple T2I backbones and editing paradigms.

E.4.1 Qualitative Results with the SD1.5 Backbone

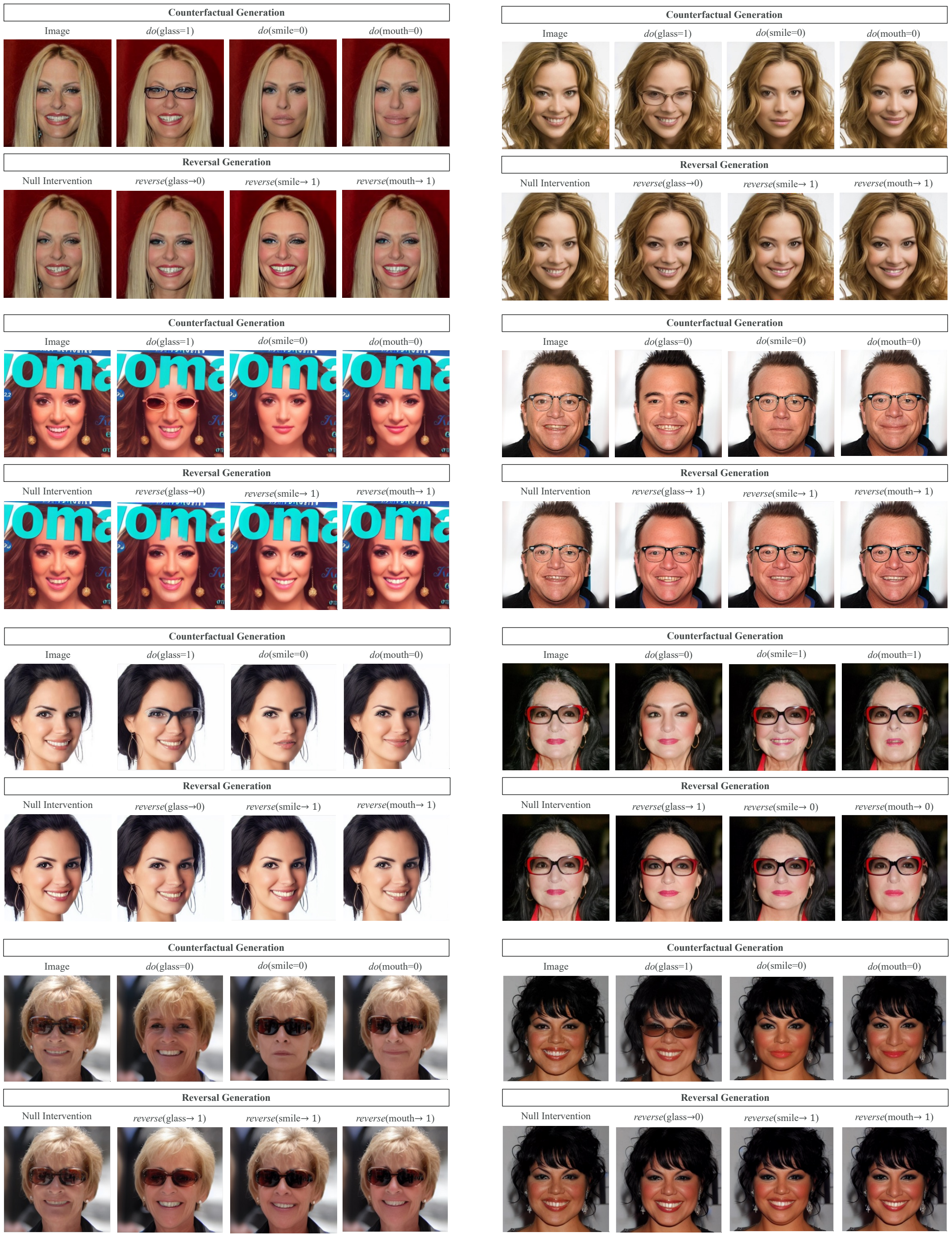


Figure 25: Additional counterfactual and reversal results (256×256) on CelebA-HQ.

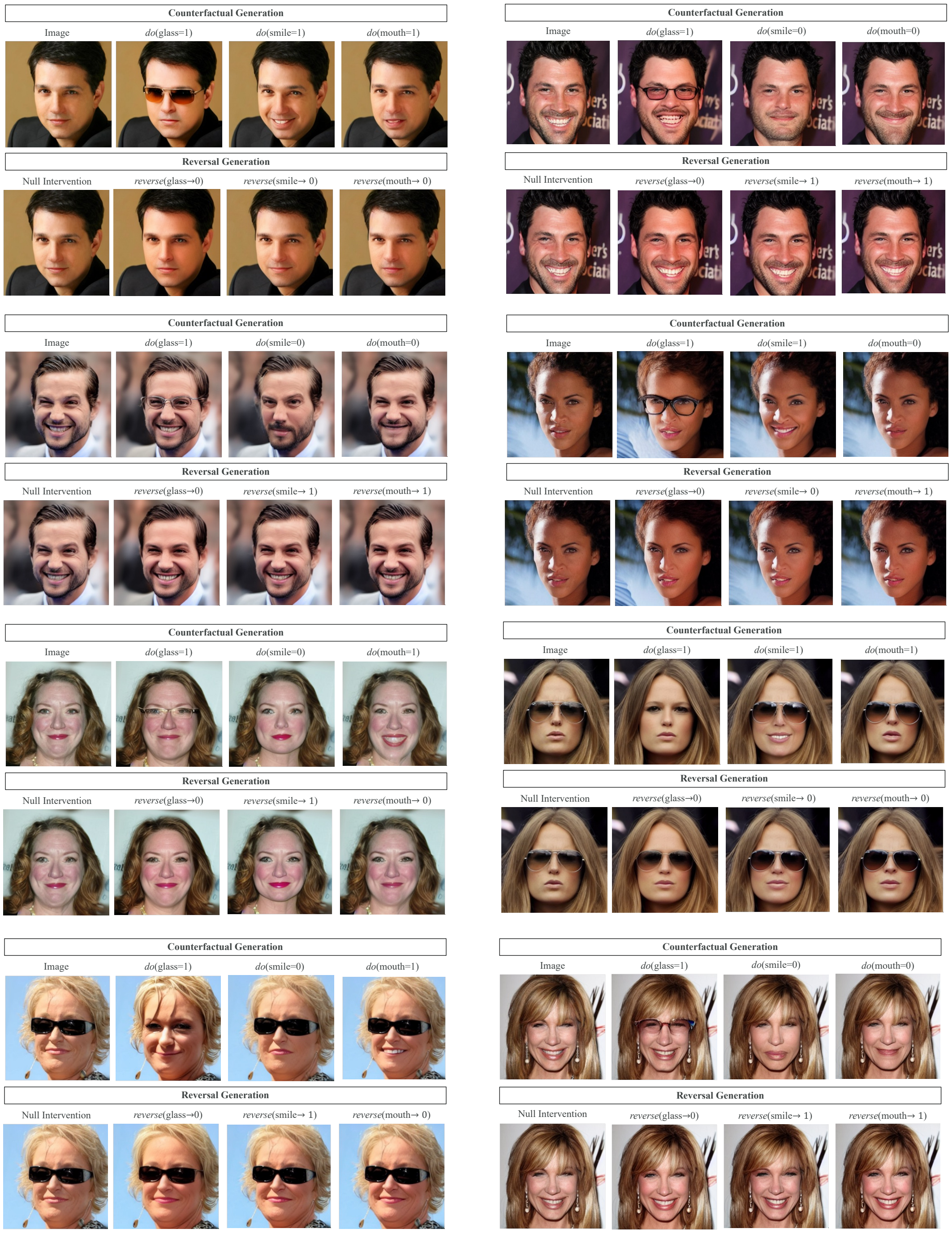


Figure 26: Additional counterfactual and reversal results (256×256) on CelebA-HQ.

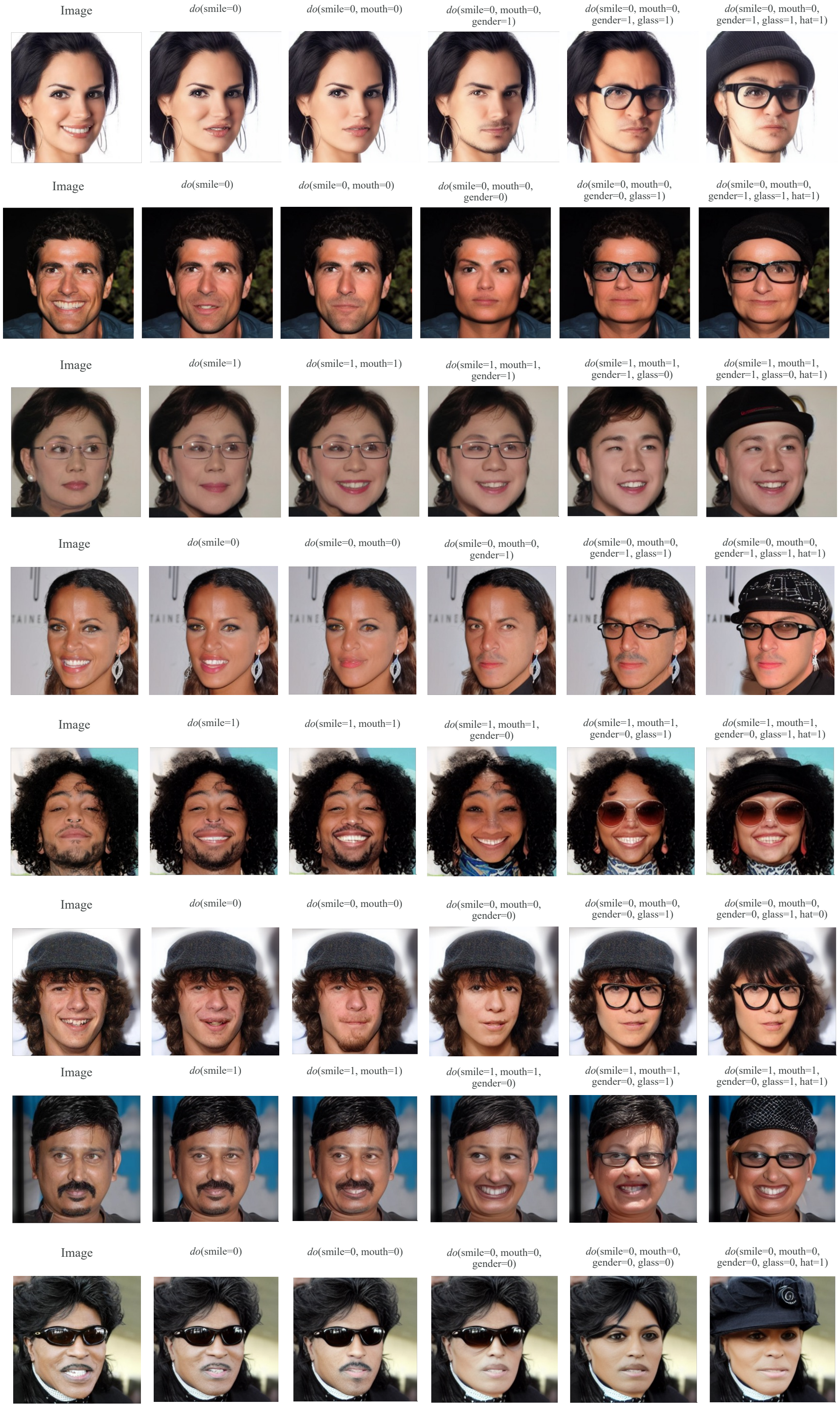


Image
do(smile=0)
do(smile=0, mouth=0)
do(smile=0, mouth=0, gender=1)
do(smile=0, mouth=0, gender=1, glass=1)
do(smile=0, mouth=0, gender=1, glass=1, hat=1)
Image
do(smile=0)
do(smile=0, mouth=0)
do(smile=0, mouth=0, gender=0)
do(smile=0, mouth=0, gender=0, glass=1)
do(smile=0, mouth=0, gender=1, glass=1, hat=1)
Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=1)
do(smile=1, mouth=1, gender=1, glass=0)
do(smile=1, mouth=1, gender=1, glass=0, hat=1)
Image
do(smile=0)
do(smile=0, mouth=0)
do(smile=0, mouth=0, gender=1)
do(smile=0, mouth=0, gender=1, glass=1)
do(smile=0, mouth=0, gender=1, glass=1, hat=1)
Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=0)
do(smile=1, mouth=1, gender=0, glass=1)
do(smile=1, mouth=1, gender=0, glass=1, hat=1)
Image
do(smile=0)
do(smile=0, mouth=0)
do(smile=0, mouth=0, gender=0)
do(smile=0, mouth=0, gender=0, glass=1)
do(smile=0, mouth=0, gender=0, glass=1, hat=0)
Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=0)
do(smile=1, mouth=1, gender=0, glass=1)
do(smile=1, mouth=1, gender=0, glass=1, hat=1)
Image
do(smile=0)
do(smile=0, mouth=0)
do(smile=0, mouth=0, gender=0)
do(smile=0, mouth=0, gender=0, glass=0)
do(smile=0, mouth=0, gender=0, glass=0, hat=1)

Figure 27: Stress test with compositional interventions: five attributes (smile, mouth-open, gender, glasses, hat) are progressively intervened. For example, the second column applies do(smile); the third column applies do(smile, mouth); subsequent columns add further interventions in sequence.

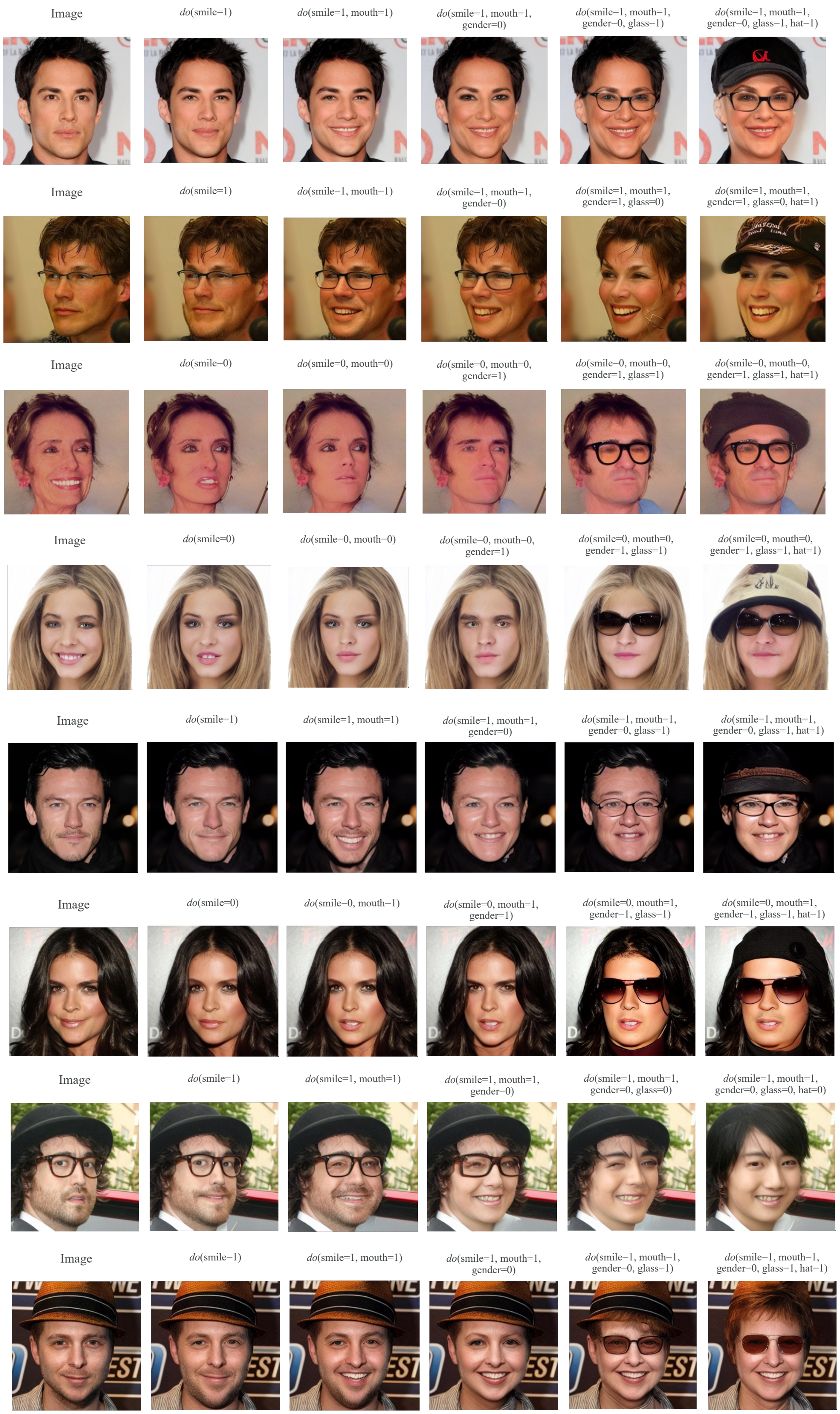


Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=0)
do(smile=1, mouth=1, gender=0, glass=1)
do(smile=1, mouth=1, gender=0, glass=1, hat=1)
Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=0)
do(smile=1, mouth=1, gender=1, glass=0)
do(smile=1, mouth=1, gender=1, glass=0, hat=1)
Image
do(smile=0)
do(smile=0, mouth=0)
do(smile=0, mouth=0, gender=1)
do(smile=0, mouth=0, gender=1, glass=1)
do(smile=0, mouth=0, gender=1, glass=1, hat=1)
Image
do(smile=0)
do(smile=0, mouth=0)
do(smile=0, mouth=0, gender=1)
do(smile=0, mouth=0, gender=1, glass=1)
do(smile=0, mouth=0, gender=1, glass=1, hat=1)
Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=0)
do(smile=1, mouth=1, gender=0, glass=1)
do(smile=1, mouth=1, gender=0, glass=1, hat=1)
Image
do(smile=0)
do(smile=0, mouth=1)
do(smile=0, mouth=1, gender=1)
do(smile=0, mouth=1, gender=1, glass=1)
do(smile=0, mouth=1, gender=1, glass=1, hat=1)
Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=0)
do(smile=1, mouth=1, gender=0, glass=0)
do(smile=1, mouth=1, gender=0, glass=0, hat=0)
Image
do(smile=1)
do(smile=1, mouth=1)
do(smile=1, mouth=1, gender=0)
do(smile=1, mouth=1, gender=0, glass=1)
do(smile=1, mouth=1, gender=0, glass=1, hat=1)

Figure 28: Stress test with compositional interventions: five attributes (smile, mouth-open, gender, glasses, hat) are progressively intervened. For example, the second column applies do(smile); the third column applies do(smile, mouth); subsequent columns add further interventions in sequence.

E.4.2 Qualitative Results with the SD3 Backbone

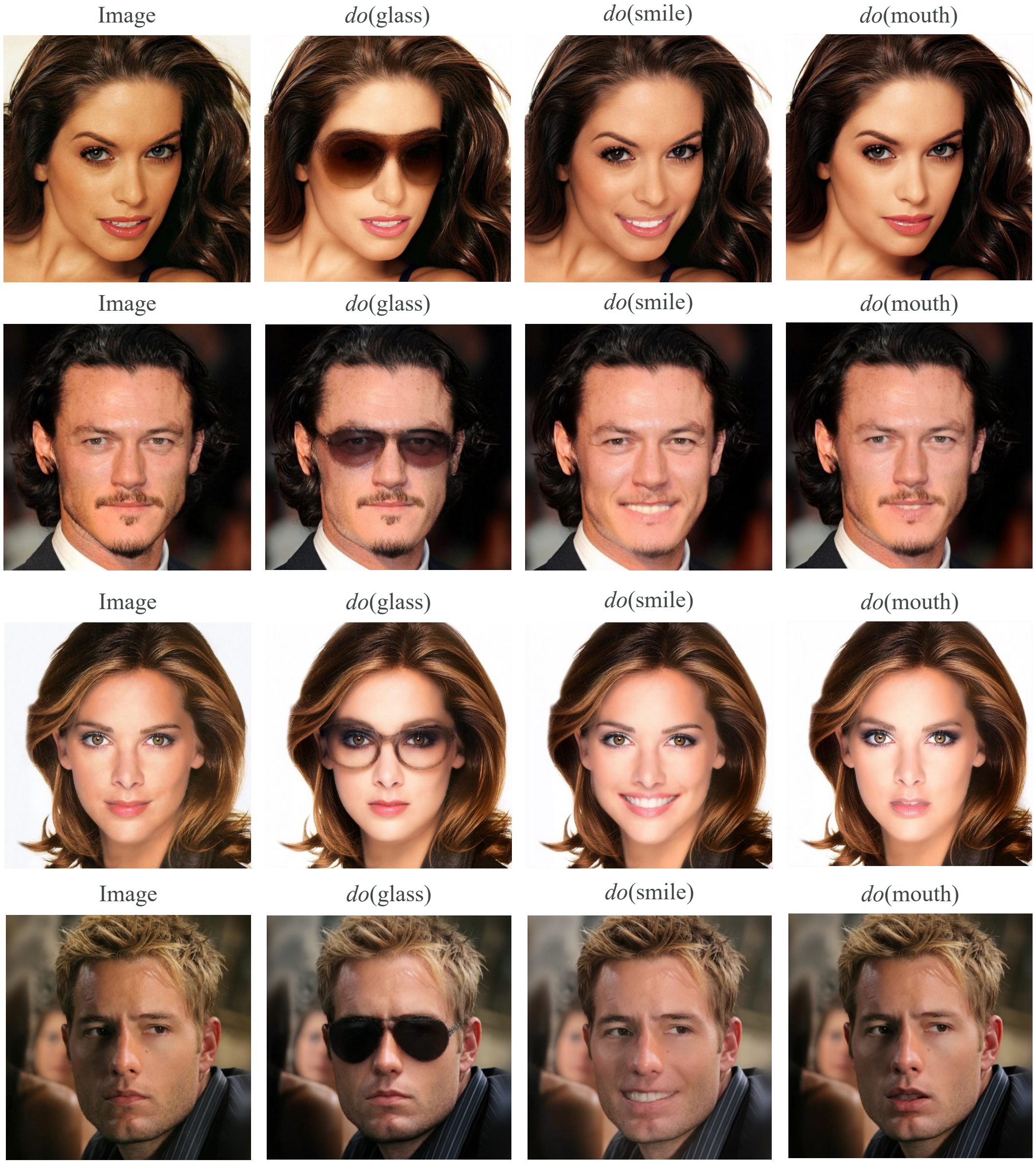


Figure 29: Additional counterfactuals (512×512) generated by using SD3 Backbone.

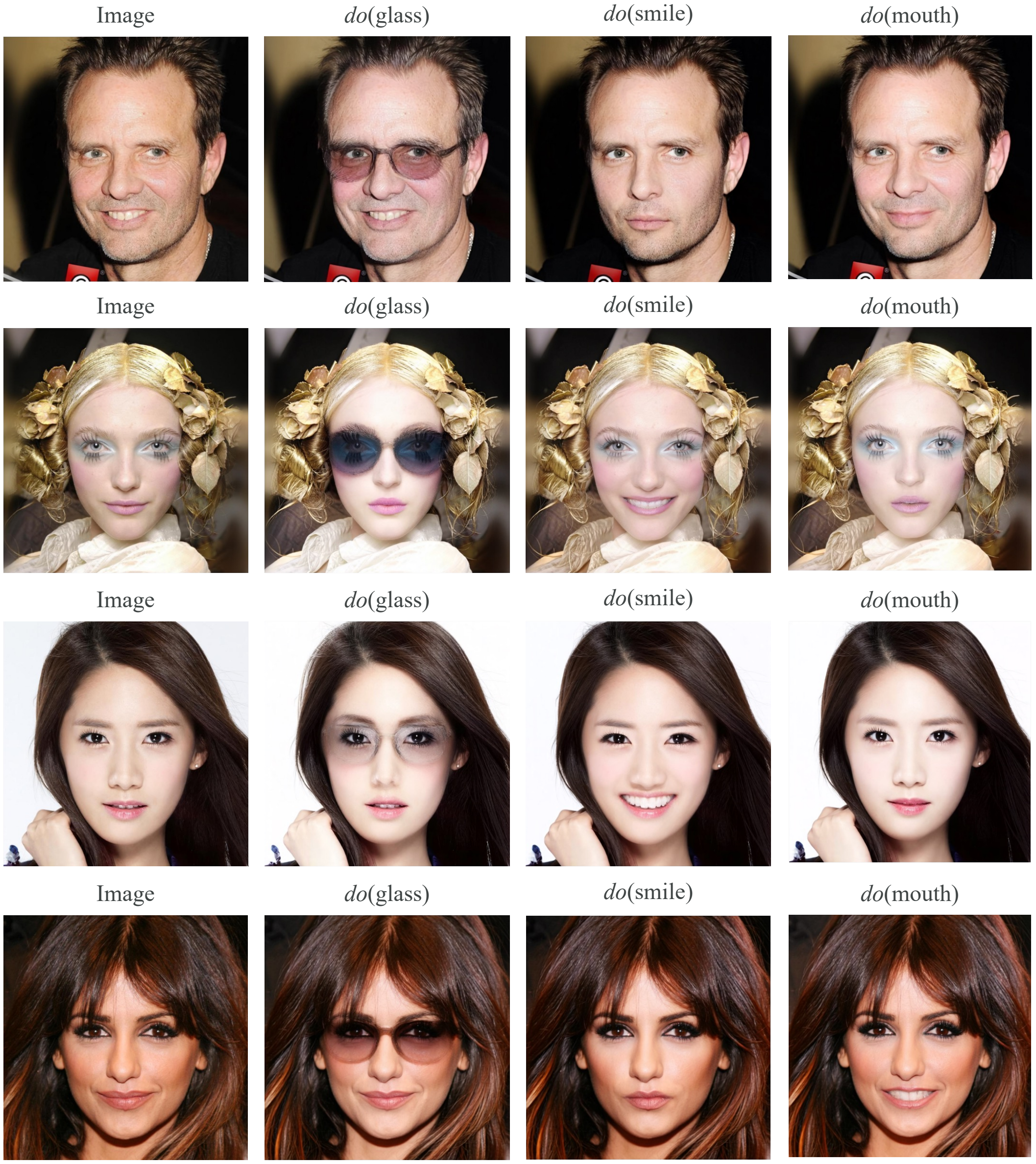


Figure 30: Additional counterfactuals (512×512) generated by using SD3 Backbone.

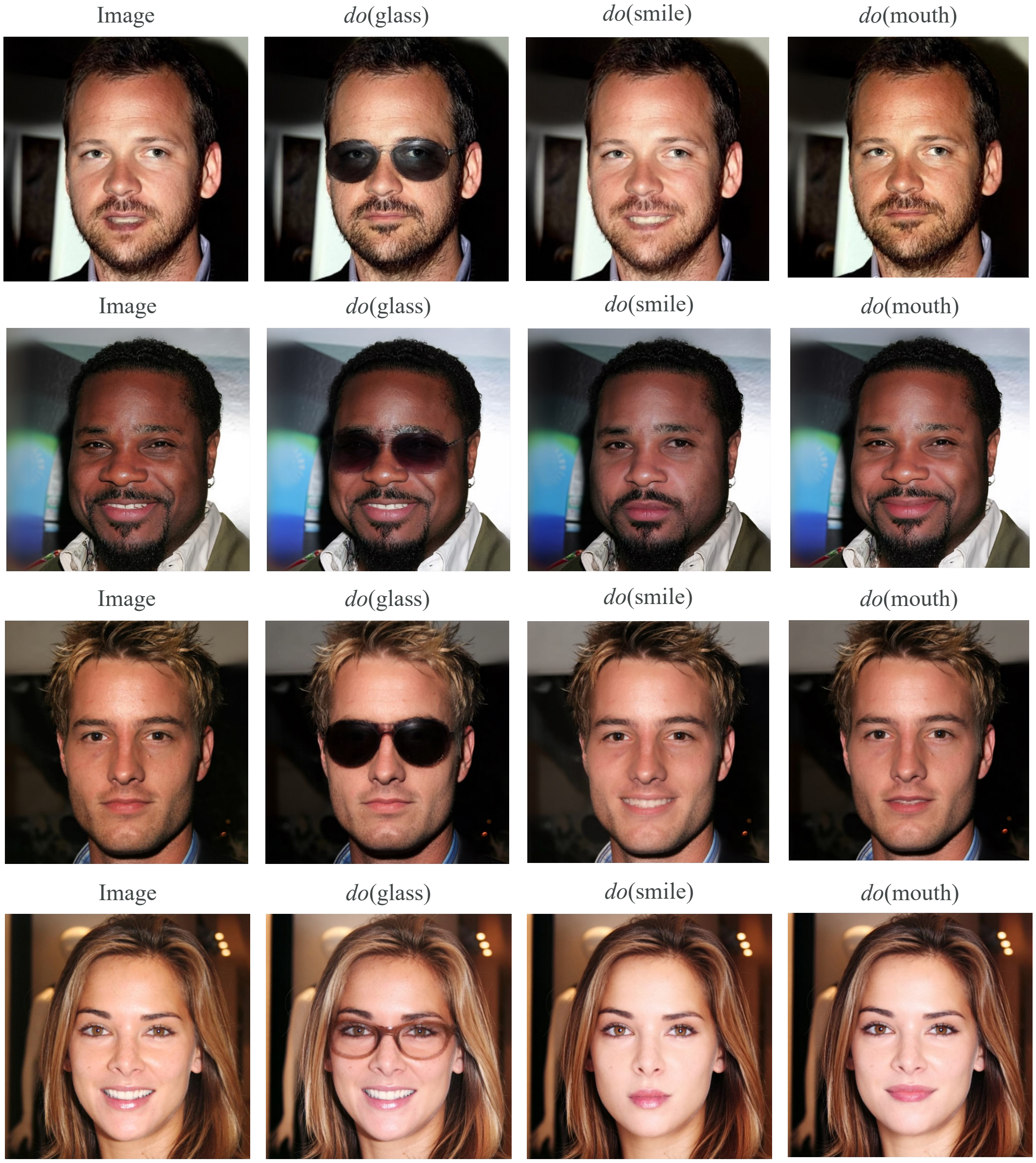


Figure 31: Additional counterfactuals (512×512) generated by using SD3 Backbone.

E.5 Attention Maps

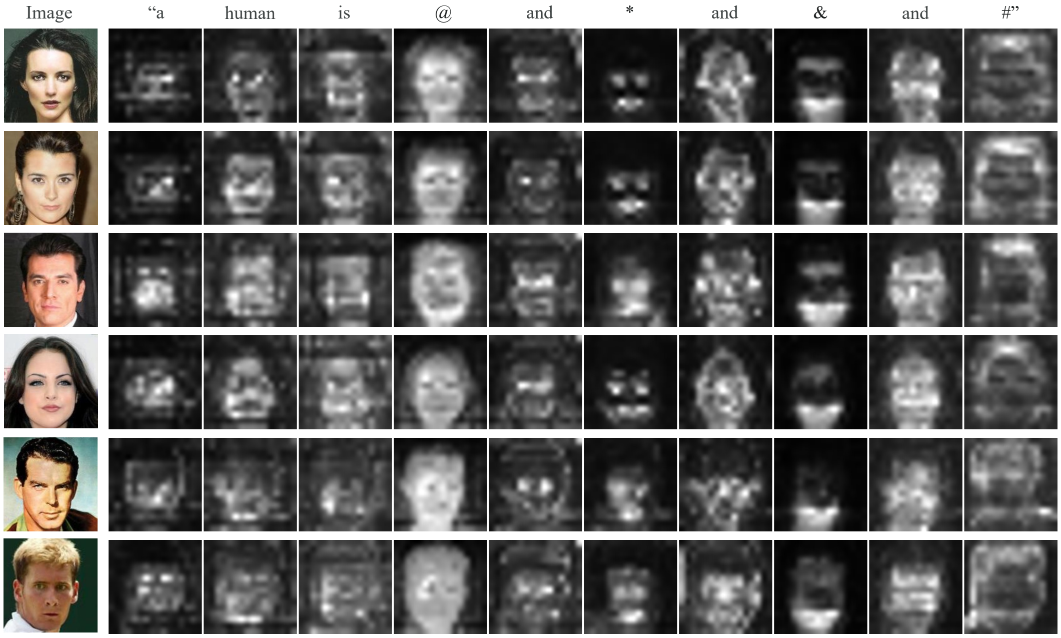


Figure 32: Average cross-attention maps from Causal-Adapter on CelebA dataset. Token denote attributes: “@” for age, “*” for gender, “&” for beard, and “#” for bald.

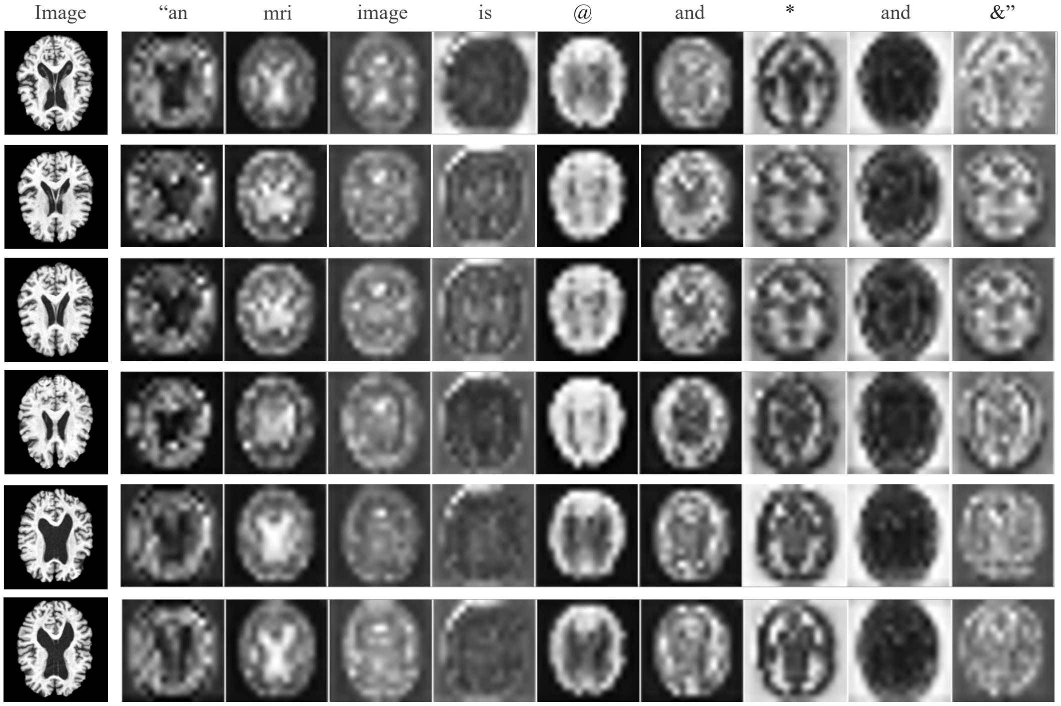


Figure 33: Average cross-attention maps from Causal-Adapter on ADNI dataset. Token denote attributes: “@” for brain volume, “*” for ventricular volume, “&” for slice.

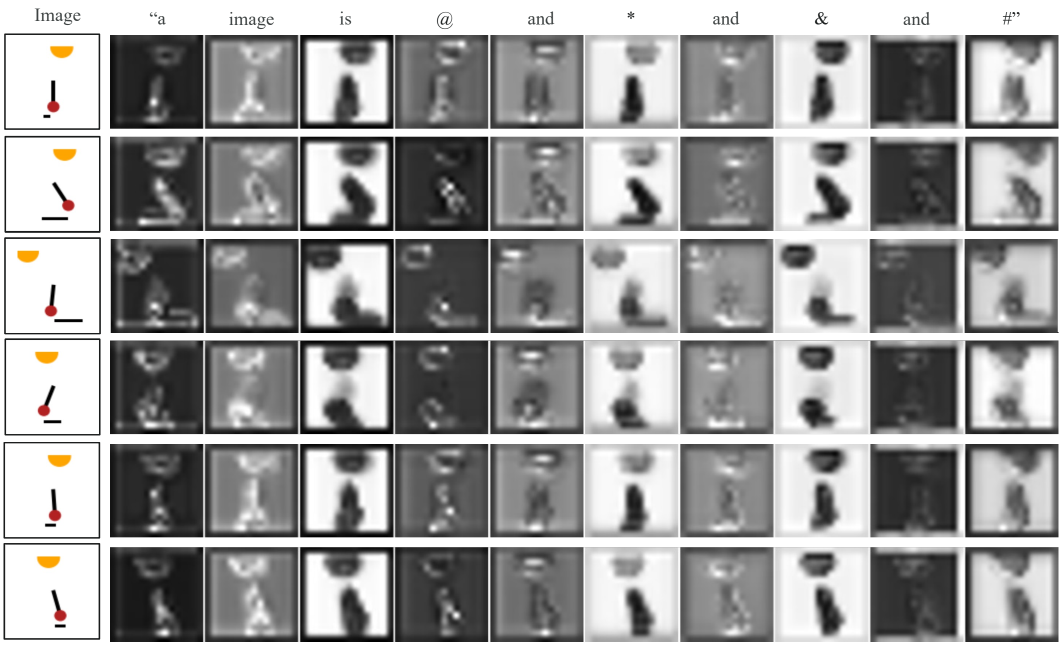


Figure 34: Average cross-attention maps from Causal-Adapter on Pendulum dataset. Token denote attributes: “@” for pendulum, “*” for light, “&” for shadow length, and “#” for shadow position.

F Counterfactual Identifiability

The abduction step in counterfactual generation implicitly assumes that the frozen diffusion model provides a counterfactually valid inverse mapping that DDIM inversion can approximately recover the exogenous noise consistent with the true data-generating process. This relates to the broader challenge of counterfactual identifiability in high-dimensional generative models (Ribeiro et al., 2025; Komanduri et al., 2024a; Nasr-Esfahany et al., 2023), where recovering latent exogenous variables from observations is generally non-trivial and often non-identifiable without additional assumptions.

To empirically assess identifiability, we perform a reversibility analysis, following the principle that counterfactual outputs should be reconstructable from the observed image distribution. In our CelebA-HQ experiments, we compare Causal-Adapter with DiffCounter and observe substantially stronger reversibility under our approach. Representative examples for three interventions (glasses, smile, mouth) are shown in Figure 25–26. These results suggest that our adapter improves the model's ability to produce counterfactuals that remain consistent with the underlying observational manifold.

However, we note that strong empirical recovery does not imply a formal identifiability guarantee. DDIM inversion is approximate, and imperfect reconstruction may cause information loss, especially under complex or compound interventions. Thus, even our Causal-Adapter "tames" causal priors into the frozen T2I backbone and yields practically robust counterfactual behavior, exact theoretical identifiability remains an open challenge.

Achieving formal counterfactual identifiability would require improved inverse operators or diffusion models explicitly designed to recover exogenous noise, may potentially leveraging recent progress in flow matching. We regard this as an important direction for future work, complementary to our empirical findings.

G Limitations and Mitigation Studies

This section describes practical evaluation constraints observed in our experiments and presents corresponding mitigation studies.

G.1 Classifier Bias in OOD Settings

Causal-Adapter demonstrates strong counterfactual generation across domains, but evaluation may fail when generated counterfactuals exceed the representational scope of the intervention classifier, especially in out-of-distribution (OOD) cases. As shown in Figure 35, our method successfully adds beards to female faces while preserving identity. However, intervention classifiers may misclassify these counterfactuals as male, since their training data does not contains examples of bearded females. This illustrates that current evaluation may occasionally fail when counterfactual generation been much stronger than the classifier, especially in OOD settings.

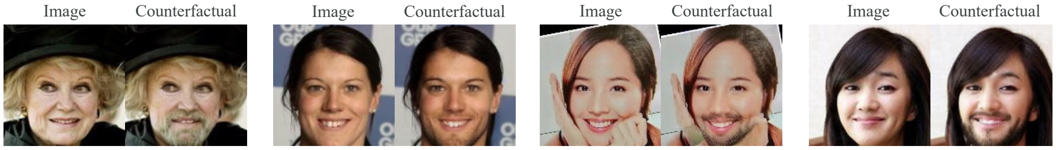


Figure 35: Counterfactuals generated by Causal-Adapter on CelebA under beard interventions. Our method successfully adds beards to female faces while preserving identity. However, since the intervention classifiers were trained without any bearded females, such counterfactuals may occasionally be misclassified as male.

Mitigation Study To obtain a more reliable assessment beyond biased intervention classifiers, we conducted an evaluation on a random subset of CelebA and their beard intervention counterfactuals (Figure 36). We evaluated each counterfactual using three judges: 1. Intervention classifiers (original pre-trained classifiers), 2. Human annotators (three raters; majority vote), 3. GPT-5.1 acting as a vision–language evaluator. Human and GPT followed the same instruction: *“Given an original face and its counterfactual version (intervention: beard), determine (1) whether a beard was successfully added, and (2) whether the person’s gender changed. Respond yes/no for each.”*

Quantitative results are shown in Table 14. We observe a consistent trend: Classifier accuracy $<$ Human accuracy $<$ GPT-5.1 accuracy. This confirms that intervention classifiers mislabel a part of counterfactuals, while human and VLM-based evaluators provide more robust and reliable measurements. These findings suggest that future counterfactual image benchmarks can incorporate human or VLM-based evaluation to eliminate biases arising from classifiers.

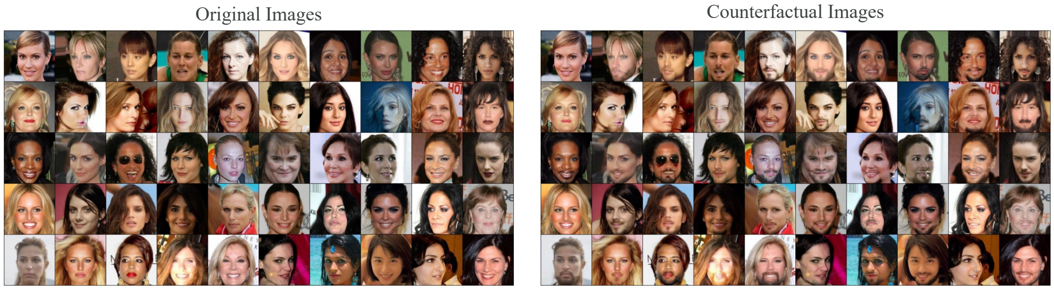


Figure 36: Sampled subset: original images (left) and counterfactuals from Causal-Adapter (right).

Table 14: Bearded-female counterfactual evaluation across classifier, human, and VLM-based judges.

Metric	Cls. Gender Acc	Cls. Beard Acc	Human Gender Acc	Human Beard Acc	GPT-5.1 Gender Acc	GPT-5.1 Beard Acc
Ours	78%	74%	90%	86%	96%	86%

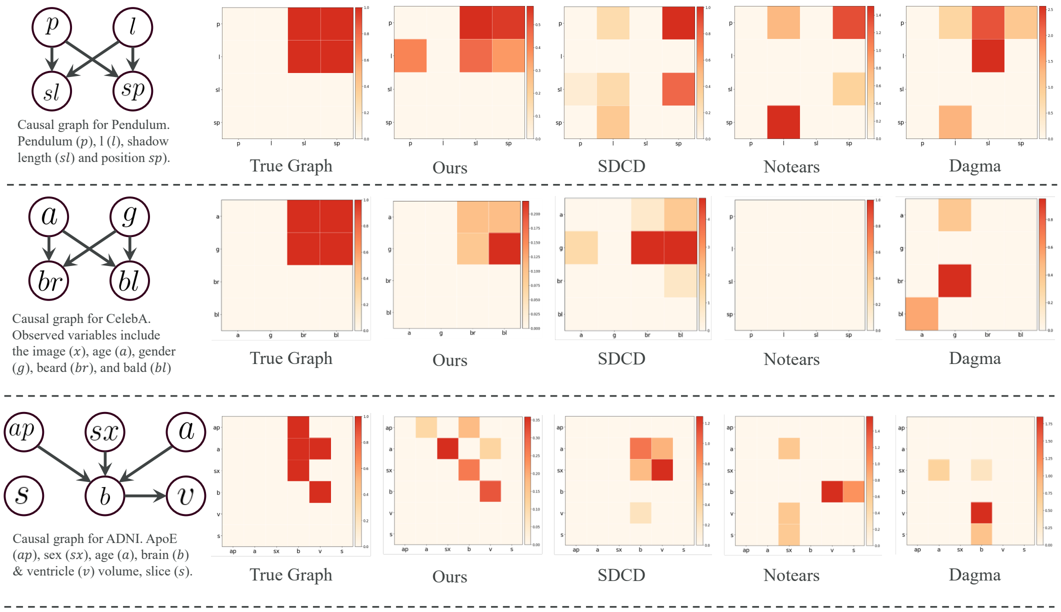
G.2 Assumption of a Known Causal Graph

Counterfactual generation frameworks assume that a pre-defined causal graph is provided as part of the input (Pearl, 2009; De Sousa Ribeiro et al., 2023; Wu et al., 2025; Komanduri et al., 2024a; Rasal et al., 2025). This assumption enables active interventions and makes evaluation comparable across methods. However, it is a relatively strong requirement: in many real-world scenarios, the causal graph may be misspecified, partially known, or even completely unknown. To understand whether our framework can adapt to such settings, we conduct a mitigation study through causal discovery.

Mitigation Study Our Causal-Adapter is modular and can be extended with minimal modification to jointly learn the causal graph. Instead of fixing the adjacency matrix A, we treat it as a learnable parameter initialized to zeros. Following differentiable DAG-based methods such as NOTEARS (Zheng et al., 2018) and DAGMA (Bello et al., 2022), we impose an acyclicity constraint (e.g., log-determinant penalty) directly on A. No additional networks or architectural changes are required; only an extra structural loss term is added.

As shown in Figure 37, this extension allows the causal adapter to perform causal structure learning, and it achieves competitive or superior performance compared with state-of-the-art differentiable causal discovery methods, including SDCD (Nazaret et al., 2024). Across three benchmark settings, our method recovers more true edges than competing baselines. Nevertheless, consistent with prior literatures (Nazaret et al., 2024; Olko et al., 2025), recovering the full causal graph from *purely observational data* (e.g., CelebA or ADNI) remains fundamentally challenging due to the lack of interventional signals.

Additionally, the learning dynamics of A during training are visualized in Figure 38, showing stable convergence toward the truth graph and demonstrating interpretable behavior of our model. Overall, these results indicate that Causal-Adapter has the potential to unify causal discovery and counterfactual generation within a single, simple, and efficient framework.

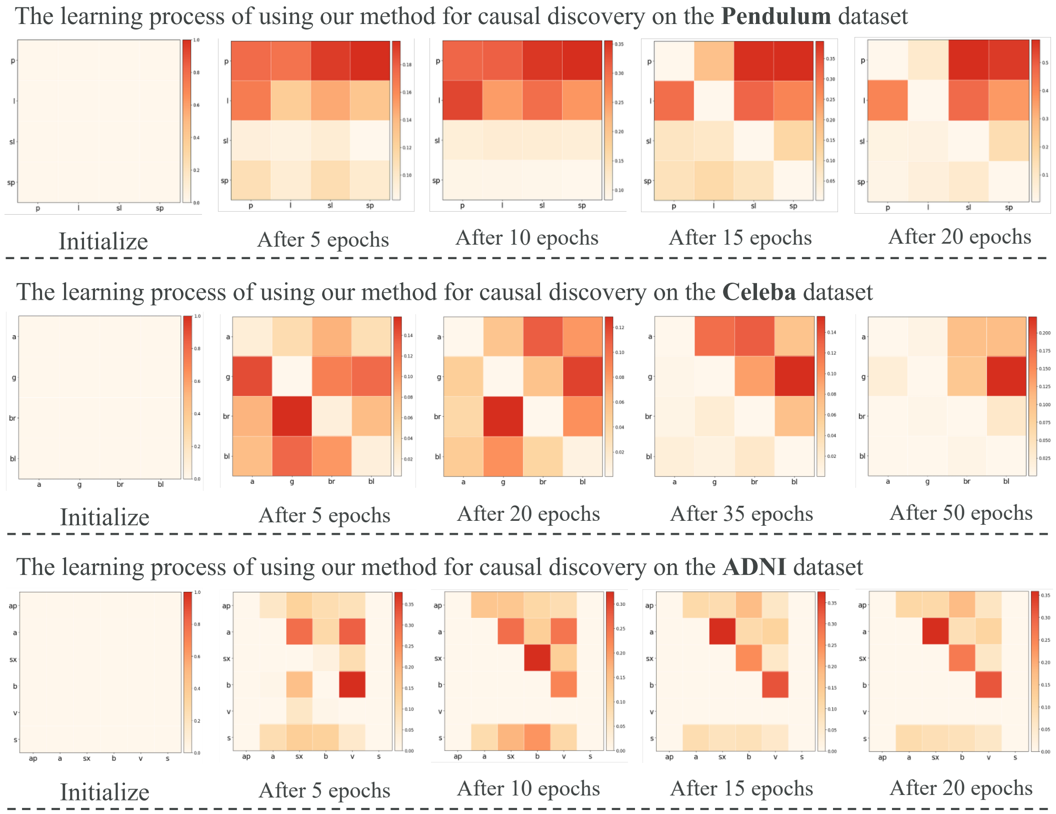


Causal graph for Pendulum. Pendulum (p), l (l), shadow length (sl) and position sp).

Causal graph for CelebA. Observed variables include the image (x), age (a), gender (g), beard (br), and bald (bl)

Causal graph for ADNI. ApoE (ap), sex (sx), age (a), brain (b) & ventricle (v) volume, slice (s).

Figure 37: Causal discovery performance of Causal-Adapter compared with SDCD, NOTEARS, and DAGMA across three benchmarks. Using the predefined graph as ground truth, our method recovers more true edges than competing methods, demonstrating the potential to unify causal discovery and counterfactual generation within a single framework.



The learning process of using our method for causal discovery on the **Pendulum** dataset

Initialize — After 5 epochs — After 10 epochs — After 15 epochs — After 20 epochs

The learning process of using our method for causal discovery on the **Celeba** dataset

Initialize — After 5 epochs — After 20 epochs — After 35 epochs — After 50 epochs

The learning process of using our method for causal discovery on the **ADNI** dataset

Initialize — After 5 epochs — After 10 epochs — After 15 epochs — After 20 epochs

Figure 38: Learning trajectory of the adjacency matrix A. As training progresses, the learned graph progressively converges toward the true structure. A fixed threshold of 0.1 is applied across benchmarks for fair comparison in the end.

H The Use of Large Language Models (LLMs)

To clarify, we only used an LLM (ChatGPT) for grammar checking and text polishing, without any involvement in content generation.