CAUSALSLIDERS: GRAPH-GUIDED LORA INTERVENTIONS FOR CAUSALLY CONSISTENT IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models enable flexible semantic editing but lack causal control: the ability to intervene on a target factor while preserving causally independent attributes during editing. Trained observationally, these models encode both causal and incidental correlations. For example, increasing age should affect wrinkles, a causal descendant, but not camera pose or background, which may be correlated yet causally independent. Existing editing methods entangle these effects because optimizing target accuracy under correlated factors exploits spurious co-occurrences, trading invariance for fidelity and failing to enforce causal mediation. We introduce *CausalSliders*, a parameter-efficient image-editing framework that embeds causal structure directly into diffusion-model adaptation. We represent each semantic factor by a dedicated low-rank LoRA adapter trained to induce a targeted parameter-space effect, enabling reusable edits without per-image optimization. Minimality and conditional-independence losses penalize non-target drift and cross-factor interference, addressing failures of linear, commutative multi-LoRA composition under correlated factors. Factor dependencies are encoded by a directed acyclic graph, and a gated, non-commutative composition operator applies interventions in causal order to enforce mediation under multi-attribute edits. By enforcing graph-ordered parameter interventions, CausalSliders improves multi-factor correctness from 39% (Concept Sliders) to 72% and achieves 81% causal path accuracy, matching the causal accuracy of Deep-SCM while running over 50× faster without per-image optimization.

1 INTRODUCTION

Counterfactual image editing asks whether a model can change a specified semantic factor while holding causally independent factors invariant under the intervention. For example, making a person younger should affect age-related attributes such as wrinkles while preserving identity, pose, and background. This distinction is critical in applications such as fairness auditing, controlled data generation, and scientific imaging, where violations of causal structure can distort conclusions or produce misleading samples Melistas et al. (2024b); Komanduri et al. (2024).

While recent closed-source systems Google DeepMind (2025); Google AI for Developers (2026); Google (2025); OpenAI (2025a;b); Labs et al. (2025) have advanced image editing quality, open-source diffusion models remain large and costly to iterate on, motivating parameter-efficient approaches for reusable editing. However, parameter efficiency alone is insufficient: existing diffusion-based editing methods, including parameter-efficient ones, frequently violate the causal faithfulness requirement. Most approaches are trained on observational correlations Gandikota et al. (2024); Yu et al. (2022); Hertz et al. (2022), which conflate causal effects with incidental co-occurrences and lead to correlation-driven edits rather than targeted interventions. As a result, edits applied to one semantic factor propagate to unrelated factors that are statistically correlated in the training data (e.g., adding eyeglasses changes lighting). Alternatively, a second class of methods enforces semantic invariance by suppressing all non-target changes Couairon et al. (2022); Wallace et al. (2023); Dong et al. (2023), which prevents leakage but also removes valid causal effects (e.g., age edits no longer induce wrinkles). Consequently, methods designed to prevent incorrect changes also suppress correct downstream responses. Thus, without causal structure, existing methods face a fundamental tradeoff: correlation-based edits change too much, while invariance-based edits change too little, and neither yields faithful counterfactuals.

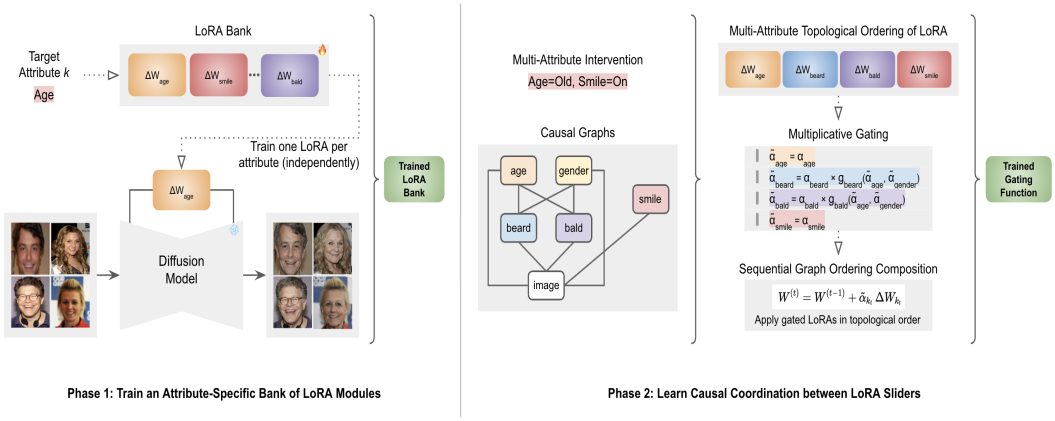


Figure 1: **Overview of CausalSliders.** *Phase 1:* Train a separate low-rank adapter (LoRA) for each semantic attribute on a shared frozen diffusion backbone. *Phase 2:* A causal graph selects and orders attribute LoRAs for a given intervention, which are composed with learned gating and applied in a single forward pass to generate the counterfactual image.

The core challenge is therefore causal coordination under efficiency constraints, assuming a fixed causal graph. Parameter-efficient adaptations such as LoRA are efficient and reusable, but when trained independently and composed linearly (i.e., by commutative summation), they behave as entangled style controls rather than interventions. For example, parameter-efficient methods such as Concept Sliders derive editing directions from textual contrasts or reference images, which inherit dataset correlations and entangle target attributes with correlated factors. In contrast, approaches such as disentangled causal latent spaces have been proposed to scale causal editing to large generative models (Pan & Bareinboim, 2025), but existing methods often rely on domain labels or structure that do not generalize to multi-attribute interventions. Without additional structure, causal faithfulness, compositional correctness, and parameter efficiency cannot be achieved simultaneously within existing LoRA-based or SCM-based approaches.

We address this gap by embedding causal structure directly into parameter-efficient adaptation. We introduce *CausalSliders*, a framework that treats each low-rank adapter as a targeted causal modification of a single semantic factor. As shown in Figure 1, each adapter modifies how a single semantic factor influences generation while the diffusion backbone remains frozen. The resulting adapters are reusable across inputs and coordinated at inference time without retraining the generator. CausalSliders does not assume factors are independent. Dependencies among factors are specified by a directed acyclic graph provided externally, either from known generative processes, lightweight domain knowledge, or prior causal discovery. The graph is fixed during training and inference; learning causal structure from data is outside the scope of this work. It encodes which attributes may change in response to an intervention and which must remain invariant, consistent with structural causal modeling frameworks that underlie controllable generative models (Komanduri et al., 2024; Correa & Bareinboim, 2025). Causal structure governs both training and inference. During training, objectives enforce target correctness, suppress changes to non-descendant factors, and penalize violations of conditional independence. During inference, multiple adapters are composed using a graph-ordered, gated operator. Independent factors commute, while causally related factors exhibit ordered, non-commutative behavior. This prevents descendant attributes from activating without their causal parents and eliminates leakage caused by linear composition.

Across CelebA Liu et al. (2015) and MorphoMNIST Castro et al. (2019), CausalSliders improves causal faithfulness and compositional correctness without sacrificing efficiency. On CelebA-Complex, it raises multi-factor correctness from 39% under linear LoRA composition to 72% and achieves 81% causal path accuracy, matching optimization-based causal models without per-image inference. Single- and continuous-factor experiments confirm that these gains preserve expressivity and image quality while enforcing correct causal mediation. Detailed results and ablations appear in Section 5.

Contributions.

- **Causal coordination of parameter-efficient edits.** We show that linear, commutative composition of LoRA adapters is fundamentally incompatible with causal counterfactual

editing, and propose a non-commutative composition of low-rank parameter interventions grounded in a causal graph.

- **Graph-aware intervention objectives.** We introduce training objectives that make causal violations explicit and optimizable, enforcing target correctness, suppressing non-descendant leakage, and penalizing incorrect causal mediation at the level of parameter-space interventions.
- **Efficient causal editing without per-image optimization.** We demonstrate faithful multi-attribute counterfactual editing that matches optimization-based causal models while eliminating per-image optimization, retaining the efficiency and reusability of parameter-efficient adaptation.

We will release our code and models to support further research in causal image editing.

2 Related Work

Our work lies at the intersection of causal inference, parameter-efficient adaptation, and counterfactual image editing. Prior work addresses these themes independently, but very few existing approach simultaneously achieves causal faithfulness, compositional correctness, and parameter efficiency.

Causal and Counterfactual Editing. Image editing has increasingly been framed as counterfactual inference, requiring edits to respect causal relations rather than observational correlations. Existing causal approaches enforce structure through predefined graphs, factorized mechanisms, latent-space reparameterization, or inference-time optimization, achieving counterfactual consistency at the cost of custom training, specialized representations, or per-image optimization Kocaoglu et al. (2017); Sauer & Geiger (2021); Pan & Bareinboim (2025); Yu et al. (2022); Pawlowski et al. (2020); Tong et al. (2025). Benchmarking studies further show that causal faithfulness and correct downstream propagation require access to ground-truth structure, motivating the use of controlled datasets such as CelebA and MorphoMNIST Melistas et al. (2024b). Application-driven methods emphasize causal isolation in specific settings but do not address efficient multi-factor composition under correlated attributes Alaya et al. (2024); Dong et al. (2023). In contrast, CausalSliders targets efficient, reusable causal composition directly in parameter space.

Parameter-Efficient and Plug-and-Play Controls. Parameter-efficient and plug-and-play methods enable reusable edits by injecting low-rank updates or manipulating prompts, attention maps, intermediate features, or CLIP-based objectives, achieving strong visual quality and usability without retraining Hu et al. (2021); Gandikota et al. (2024); Shah et al. (2023); Gu et al. (2023); Li et al. (2025); Xia et al. (2025); Luo et al. (2024); Sridhar & Vasconcelos (2024); Alekseenko et al. (2026); Huang et al. (2025); Yu et al. (2022); Dong et al. (2023); Hertz et al. (2022); Tumanyan et al. (2022); Couairon et al. (2022); Wallace et al. (2023). However, these approaches compose controls commutatively and rely on observational correlations or invariance constraints, conflating $P(y \mid x)$ with $P(y \mid do(x))$ under correlated attributes.

Benchmarking and Positioning of Our Work. Recent benchmarking and interpretability studies highlight that causal faithfulness and disentanglement require explicit structure and principled coordination when scaling to many concepts Melistas et al. (2024a); He et al. (2026); Luo et al. (2024); Alekseenko et al. (2026). From a broader causal modeling perspective, counterfactual queries are identifiable only under explicit structural assumptions and well-defined intervention semantics Correa & Bareinboim (2025); Lee et al. (2025); Zhang & Bareinboim (2025); Jaimini & Sheth (2022). CausalSliders addresses this gap by embedding causal structure directly into parameter-efficient LoRA adaptations, enabling graph-ordered, non-commutative composition that achieves faithful multi-attribute counterfactuals without per-image optimization or architectural redesign.

3 Method

3.1 Problem Setup and Causal Editing Objective

We study counterfactual image editing under a fixed, pretrained diffusion model. The objective is causal correctness: changing specified semantic factors while preserving all causally independent ones. Let $x \in \mathcal{X}$ denote an input image and $\mathbf{z}(x) = (z_1(x), \ldots, z_K(x))$ a set of interpretable semantic factors, each binary or continuous. We assume a fixed, user-specified set of semantic factors provided via dataset-level supervision or descriptors, which are visually grounded and may be binary or continuous; extending the framework to open-vocabulary or automatically discovered attributes is a natural direction but outside the scope of this work. We assume a directed acyclic graph $G = (\mathcal{V}, \mathcal{E})$

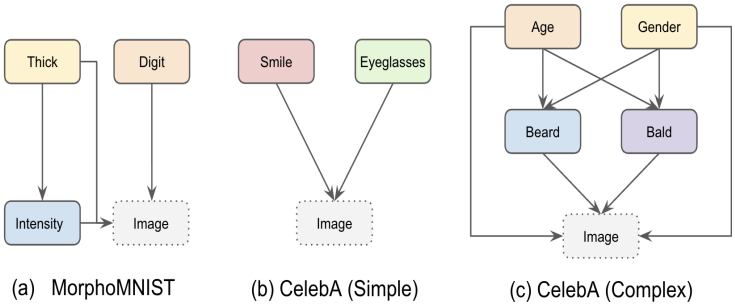


Figure 2: Causal graphs used in our experiments. (a) **MorphoMNIST**: thickness and digit identity influence intensity, which determines the final image. (b) **CelebA (simple)**: smile and eyeglasses directly affect the image and are causally independent. (c) **CelebA (complex)**: age and gender jointly influence beard and baldness, which mediate effects on the image. These graphs define permissible intervention paths and non-descendant invariance constraints.

over factors, where $\mathcal{V} = \{1, \ldots, K\}$ and an edge $i \to j$ indicates that z_i is a direct cause of z_j. The graph specifies which attribute changes are causally permitted under an intervention and which constitute leakage. For a factor k, let $\mathrm{Desc}(k)$ denote its descendants in G, excluding k itself, and for a set T, let $\mathrm{Desc}(T) = \bigcup_{t \in T} \mathrm{Desc}(t)$. We denote by $\pi(k)$ the set of parents of factor k in G, $\pi(k) = \{i \in \mathcal{V} : i \to k \in \mathcal{E}\}$.

Editing queries. An editing query specifies a set of target factors $T \subseteq \mathcal{V}$ and desired values $\{z_t^\star : t \in T\}$, written as $\mathrm{do}(z_t = z_t^\star,\, t \in T)$. We represent the query by intervention magnitudes $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, where α_k controls the strength of intervention on factor z_k and $\alpha_k = 0$ for all $k \notin T$. Binary factors use $\alpha_k \in \{0, 1\}$, while continuous factors use $\alpha_k \in \mathbb{R}$. We write $\boldsymbol{\alpha}^{(k)}$ for a single-factor intervention. More generally, for a set $S \subseteq \mathcal{V}$, $\boldsymbol{\alpha}^{(S)}$ denotes an intervention where $\alpha_k \neq 0$ for $k \in S$ and $\alpha_k = 0$ otherwise. The query intervenes only on the specified targets; any additional changes in the output are allowed only if they arise as causal consequences along directed paths in G.

Causal faithfulness. Given $(x, T, \{z_t^\star\})$, an edited image $\hat{x}$ is causally faithful if it satisfies the following conditions.

Target correctness. $|z_t(\hat{x}) - z_t^\star| \leq \varepsilon_t \; \forall\, t \in T$, with $\varepsilon_t = 0$ for binary factors and ε_t set by descriptor variance for continuous factors.

Non-descendant invariance. $z_i(\hat{x}) = z_i(x) \; \forall\, i \notin \mathrm{Desc}(T)$; violations of this constraint are termed *attribute leakage*.

Causal mediation. For $j \in \mathrm{Desc}(T)$, changes in z_j arise only via active directed paths from T in G. Descendant factors do not change when their parents are inactive and respond consistently under multiple simultaneous interventions. For example, *Wrinkles* may change when intervening on *Age*, but not when activated independently. These conditions rule out edits that exploit statistical correlations rather than isolating the causal effect of an intervention. We additionally require *compositional correctness*: interventions on causally independent factors commute, while interventions on causally related factors exhibit ordered, non-commutative behavior. Finally, we impose *parameter efficiency*: all edits are realized without retraining the diffusion backbone or performing per-image optimization.

Supervision and graphs. We assume access to frozen descriptor functions $h_k : \mathcal{X} \to \mathbb{R}$ used only for training and evaluation. On MorphoMNIST, descriptors correspond to ground-truth generative factors; on CelebA, they are pretrained attribute classifiers. In practice, factor values are evaluated via descriptors, with $z_k(x)$ approximated by $h_k(x)$. The causal graph G is fixed and external, derived from known generative structure or domain knowledge. Examples of the causal graphs used in our experiments are shown in Figure 2. Section 6 evaluates robustness to graph misspecification by perturbing edges in G.

3.2 Causal Editing as Parameter-Space Intervention

The constraints in Section 3.1 restrict where and how interventions act. To satisfy causal faithfulness under composition, an intervention modifies the mechanism by which a factor influences generation rather than conditioning on correlated representations.

Why parameter space. Latent- or prompt-based edits act on entangled representations and cannot guarantee non-descendant invariance or controlled mediation. Parameter-space interventions instead modify a factor-specific generation mechanism while leaving all others unchanged, matching the semantics of causal intervention.

Diffusion editing operator. Let W_0 denote the frozen parameters of a pretrained diffusion model. Given an input image x, editing is performed via image-conditioned diffusion under adapted parameters,

$$\hat{x}_{\boldsymbol{\alpha}} \sim \mathcal{E}(x;\, W,\, \boldsymbol{\alpha},\, \xi), \tag{1}$$

where W denotes parameters after intervention composition and ξ denotes sampling noise. In practice, DDIM inversion maps x to latent space for conditional denoising under adapted parameters. The text prompt is kept empty; all semantic control is exerted through parameter adaptation.

Low-rank parameter interventions. We implement interventions using low-rank adaptation. For a pretrained weight matrix $W_0^{(\ell)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we introduce

$$\Delta W^{(\ell)} = A^{(\ell)} (B^{(\ell)})^\top, \qquad A^{(\ell)} \in \mathbb{R}^{d_{\text{out}} \times r},\ B^{(\ell)} \in \mathbb{R}^{d_{\text{in}} \times r} \tag{2}$$

with $r \ll \min(d_{\text{out}}, d_{\text{in}})$. The adapted weight is $W^{(\ell)} = W_0^{(\ell)} + s \cdot \Delta W^{(\ell)}$. All backbone parameters remain frozen. In the factorized architecture described next, the scalar s is replaced by the factor-specific magnitude α_k that controls activation of adapter ΔW_k.

3.3 Factorized Intervention Architecture

We realize parameter-space interventions using a factorized adapter architecture. Each semantic factor is associated with a dedicated, reusable intervention module.

Factor-indexed adapters. For each factor z_k, we introduce a low-rank adapter ΔW_k. Activating ΔW_k constitutes an intervention on z_k. Adapters are indexed by factor identity rather than by image and are shared across all inputs, requiring K adapters total. Unless stated otherwise, adapters are inserted into the query, key, and value projections of mid-to-late UNet cross-attention layers, which primarily control high-level semantics. All adapters use rank $r = 8$. The diffusion backbone remains frozen.

Slider interface. Each adapter ΔW_k is controlled by a scalar magnitude α_k, enabling continuous control over intervention strength. This motivates the term *slider*. Single-factor edits activate one slider; multi-factor edits activate several. Binary and continuous factors share the same interface.

Parameter independence. Adapters are parameter-independent. Each represents one controllable degree of freedom; interactions arise only through composition rules (Section 3.4), not entangled parameterization.

Limitation of factorization. Factorization alone does not enforce causal correctness. When multiple adapters are active, naive composition leads to interference and attribute leakage. Section 3.4 introduces a graph-ordered composition rule that resolves this limitation.

3.4 Graph-Ordered Composition and Gated Mediation

The factorized adapters from Section 3.3 must compose in a way that respects the causal constraints defined in Section 3.1. Naive composition treats all factors as independent and violates causal mediation under correlated attributes.

Failure of linear composition. A common baseline composes adapters by linear summation,

$$W = W_0 + \sum_{k=1}^{K} \alpha_k \, \Delta W_k. \tag{3}$$

} This operation is commutative. As a result, adapters corresponding to descendant factors may activate even when their causal parents are inactive, producing attribute leakage.

Graph-ordered composition. To enforce causal ordering, we compose adapters according to the structure of G. Let $(k_1, \ldots, k_K)$ denote a topological ordering of the causal graph. For $t = 1, \ldots, K$, starting from $W^{(0)} = W_0$, we apply

$$W^{(t)} = W^{(t-1)} + \tilde{\alpha}_{k_t}\, \Delta W_{k_t}, \qquad W = W^{(K)}. \tag{4}$$

Disconnected components of G are composed independently and therefore commute. For reproducibility, we use a fixed deterministic topological ordering, with ties broken lexicographically by factor index.

Gated mediation. Sequential ordering alone is insufficient when multiple factors are intervened upon jointly, because descendant effects must be conditionally suppressed based on parent activation strength, not order alone. We therefore modulate each intervention by a learnable gate. For factor z_k with parent set $\pi(k)$, the effective intervention magnitude is

$$\tilde{\alpha}_k = \alpha_k \cdot g_k(\{\alpha_p\}_{p\in\pi(k)}), \qquad g_k(\cdot) \in [0,1]. \tag{5}$$

If $\pi(k) = \emptyset$, then $g_k \equiv 1$; otherwise $g_k : \mathbb{R}^{|\pi(k)|} \to [0,1]$. Gating suppresses descendant parameter updates when their causal parents are inactive and permits activation only under valid upstream context. Suppressing a descendant adapter does not prohibit descendant image changes induced by upstream interventions; it prevents spurious parameter updates that bypass causal mediation.

Gating parameterization. Each gate is parameterized as

$$g_k(\{\alpha_p\}_{p\in\pi(k)}) = \sigma\Big(w_k^\top [\alpha_{p_1}, \ldots, \alpha_{p_{|\pi(k)|}}, 1]\Big), \tag{6}$$

where $w_k \in \mathbb{R}^{|\pi(k)|+1}$ and σ denotes the sigmoid. Gating parameters are initialized near pass-through behavior and learned during training.

Non-commutativity and correctness. Graph-ordered composition is generally non-commutative. Interventions on causally independent factors commute, while interventions on causally related factors do not. This behavior instantiates compositional correctness. The contribution is the composition rule itself, which enforces causal semantics independently of the choice of parameterization. We denote the resulting composition as $W = \mathrm{Compose}(W_0, \boldsymbol{\alpha}; G)$.

Losses as causal constraints. We train adapters and gates using losses that correspond to distinct violations of causal faithfulness. Expectations over ξ are estimated using a single diffusion sample per update.

Intervention loss.

$$\mathcal{L}_{\mathrm{int}}^{(k)} = \mathbb{E}_\xi[\ell(h_k(\hat{x}_{\boldsymbol{\alpha}^{(k)}}),\, z_k^\star)], \tag{7}$$

where ℓ is binary cross-entropy for binary factors and squared error for continuous factors.

Minimality loss.

$$\mathcal{L}_{\mathrm{min}}^{(k)} = \sum_{j\notin \mathrm{Desc}(k)\cup\{k\}} \mathbb{E}_\xi\left[\|h_j(\hat{x}_{\boldsymbol{\alpha}^{(k)}}) - h_j(\hat{x}_{\mathbf{0}})\|_1\right]. \tag{8}$$

Conditional independence loss. For each pair $i \perp j$ in G,

$$\mathcal{L}_{\mathrm{CI}}^{(i,j)} = \mathbb{E}_\xi[\|h_j(\hat{x}_{\boldsymbol{\alpha}^{(i,j)}}) - h_j(\hat{x}_{\boldsymbol{\alpha}^{(j)}})\|_1 + \|h_i(\hat{x}_{\boldsymbol{\alpha}^{(i,j)}}) - h_i(\hat{x}_{\boldsymbol{\alpha}^{(i)}})\|_1]. \tag{9}$$

Chain loss. For each edge $i \to j$,

$$\mathcal{L}_{\mathrm{chain}}^{(i\to j)} = \mathbb{E}_\xi\left[\left\|\big(h_j(\hat{x}_{\boldsymbol{\alpha}^{(i)}}) - h_j(\hat{x}_{\mathbf{0}})\big) - \big(h_j(\hat{x}_{\boldsymbol{\alpha}^{(i,j)}}) - h_j(\hat{x}_{\boldsymbol{\alpha}^{(j)}})\big)\right\|_1\right]. \tag{10}$$

This enforces that the edge $i \to j$ represents a stable causal mechanism independent of direct interventions on j. The formulation assumes approximately additive mediated effects; robustness to nonlinear interactions is evaluated in Section 6.

Objective and training. The total objective is

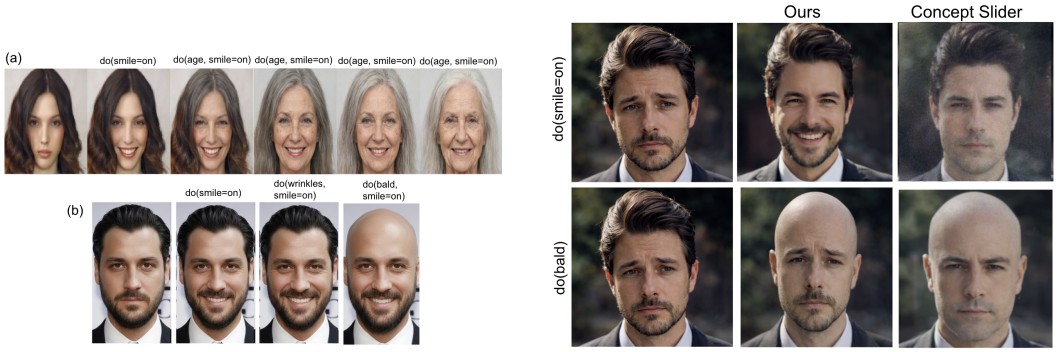
$$\mathcal{L} = \sum_k \Big(\lambda_{\mathrm{int}}\mathcal{L}_{\mathrm{int}}^{(k)} + \lambda_{\mathrm{min}}\mathcal{L}_{\mathrm{min}}^{(k)}\Big) + \sum_{i\perp j} \lambda_{\mathrm{CI}}\mathcal{L}_{\mathrm{CI}}^{(i,j)} + \sum_{i\to j} \lambda_{\mathrm{chain}}\mathcal{L}_{\mathrm{chain}}^{(i\to j)}.$$

We use $\lambda_{\mathrm{int}} = 1.0$ and $\lambda_{\mathrm{min}} = \lambda_{\mathrm{CI}} = \lambda_{\mathrm{chain}} = 0.5$ throughout unless otherwise stated.

Training proceeds in two stages. Stage 1 trains each adapter ΔW_k independently using $\mathcal{L}_{\mathrm{int}}^{(k)} + \mathcal{L}_{\mathrm{min}}^{(k)}$. Stage 2 freezes adapter parameters $\{A_k^{(\ell)}, B_k^{(\ell)}\}$ and trains only gating weights $\{w_k\}$ for five epochs using $\mathcal{L}_{\mathrm{CI}}$ and $\mathcal{L}_{\mathrm{chain}}$. No per-image optimization or inference-time finetuning is performed.

4 Experimental Setup

We evaluate CausalSliders along three axes: causal faithfulness, compositional correctness, and parameter efficiency. All experiments use a frozen diffusion backbone, identical training data across methods, and fixed causal graphs per dataset.



(a) Graph-ordered compositional counterfactual editing.

(b) Non-causal attribute leakage under correlation-driven editing.

Figure 3: Compositional counterfactual editing on CelebA (Complex). (**Left**) Simultaneous interventions on *age* and *smile*, where age increases left to right while *smile* is held fixed; graph-ordered causal composition preserves non-target attributes while allowing valid downstream (mediated) effects. (**Right**) Correlation-driven editing exhibits non-target leakage: Concept Sliders alter facial hair and geometry under *smile* and modify beard and face shape under *bald*, whereas our method preserves causally independent attributes.

Datasets and Causal Graphs. We evaluate on two datasets with complementary causal structure. **MorphoMNIST** Castro et al. (2019) provides continuous generative factors with a known causal chain Thickness $\rightarrow$ Intensity, enabling controlled evaluation of causal mediation under continuous interventions. **CelebA** Liu et al. (2015) provides binary semantic attributes with confounding, for which we use two fixed graphs: *CelebA-Simple*, containing five sparsely connected attributes, and *CelebA-Complex*, containing nine attributes with structures such as Age $\rightarrow$ Bald $\leftarrow$ Gender. All causal graphs are fixed and derived from domain knowledge.

Base Model and Editing Protocol. All methods use the same pretrained *Stable Diffusion v1.4* Rombach et al. (2022) backbone with frozen base parameters. Editing is performed via image-conditioned diffusion using DDIM inversion Song et al. (2022), followed by conditional denoising under adapted parameters. The text prompt is kept empty, and unless stated otherwise, each edit uses a single diffusion sample.

Adapter Configuration and Training. All interventions use low-rank adapters, with each semantic factor associated with a dedicated adapter of rank $r = 8$ by default. Unless stated otherwise, adapters are inserted into the query, key, and value projections of mid-to-late UNet cross-attention layers (depths 6–11). Training follows the two-stage procedure in Section 3.4: Stage 1 trains each adapter independently for 20 epochs using the intervention and minimality losses, while Stage 2 freezes all adapter parameters $\{A_k^{(\ell)}, B_k^{(\ell)}\}$ and trains only gating weights $\{w_k\}$ for five epochs using the conditional independence and chain losses. All models are trained with AdamW using learning rates 10^{-4} (Stage 1) and 10^{-5} (Stage 2), and no per-image optimization or inference-time finetuning is performed.

Baselines. We compare against three representative baselines. **Concept Sliders** Gandikota et al. (2024) trains one adapter per semantic factor and composes multiple edits by linear summation. **Linear Multi-LoRA** uses the same adapter architecture but removes graph ordering, gating, and causal losses. **Deep-SCM** Pawlowski et al. (2020) enforces causal constraints via per-image optimization and serves as a causal correctness reference. All baselines use the same Stable Diffusion v1.4 backbone, training data, and evaluation protocol, and all per-factor LoRAs are trained with comparable rank and identical attribute supervision as Concept Sliders, isolating the effect of causal coordination rather than representational capacity.

Evaluation Metrics. All metrics use descriptor functions h_k (Section 3.1) and are reported as mean $\pm$ standard deviation over three random seeds (paired two-sided t-tests, $p < 0.05$). **Intervention accuracy (IA).** For a single-factor intervention $\boldsymbol{\alpha}^{(k)}$,

$$\mathrm{IA}(k) = \mathbb{E}_x \left[\mathbb{1}\left(|h_k(\hat{x}_{\boldsymbol{\alpha}^{(k)}}) - z_k^\star| \leq \varepsilon_k \right) \right]. \tag{11}$$

Attribute leakage. Unintended changes to non-descendant factors:

$$\mathrm{Leakage}(k) = \mathbb{E}_x \left[\frac{1}{|\mathcal{V} \setminus (\mathrm{Desc}(k) \cup \{k\})|} \sum_{j \notin \mathrm{Desc}(k) \cup \{k\}} |h_j(\hat{x}_{\boldsymbol{\alpha}^{(k)}}) - h_j(\hat{x}_{\mathbf{0}})| \right]. \tag{12}$$

Causal path accuracy (CPA). For each causal edge $i \rightarrow j$,

$$\mathrm{CPA} = \mathbb{E}_{x,\, i \rightarrow j} \left[\mathbb{1}\left(|h_j(\hat{x}_{\boldsymbol{\alpha}^{(i)}}) - h_j(\hat{x}_{\mathbf{0}})| > \varepsilon_j \right) \right]. \tag{13}$$

Compositional commutativity error. For causally independent factors $i \perp j$,

$$\mathrm{CommErr}(i,j) = \mathbb{E}_x \left[\| \hat{x}^{i \prec j}_{\boldsymbol{\alpha}^{(i,j)}} - \hat{x}^{j \prec i}_{\boldsymbol{\alpha}^{(i,j)}} \|_2 \right]. \tag{14}$$

Table 1: Multi-factor composition accuracy (CelebA).

Method	MIA ↑	CommErr ↓	CPA ↑	CLIP ↑
Concept Sliders	0.39 ± 0.04	0.42 ± 0.05	0.52 ± 0.03	0.284 ± 0.008
Linear Multi-LoRA	0.33 ± 0.05	0.48 ± 0.06	0.47 ± 0.04	0.271 ± 0.011
Deep-SCM	0.64 ± 0.03	0.11 ± 0.02	0.74 ± 0.02	0.251 ± 0.009
CausalSliders (ours)	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.09 \pm 0.02}$	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.278 \pm 0.007}$

Table 2: Efficiency comparison with causal and parameter efficient baselines.

Method	Inference (ms) ↓	Params (M) ↓	CPA ↑	MIA ↑
Deep-SCM	2400 ± 180	Full model	0.74 ± 0.02	0.64 ± 0.03
Concept Sliders	45 ± 3	11.8	0.52 ± 0.03	0.39 ± 0.04
Linear Multi-LoRA	43 ± 2	11.8	0.47 ± 0.04	0.33 ± 0.05
CausalSliders (ours)	48 ± 3	12.6	0.81 ± 0.02	0.72 ± 0.03

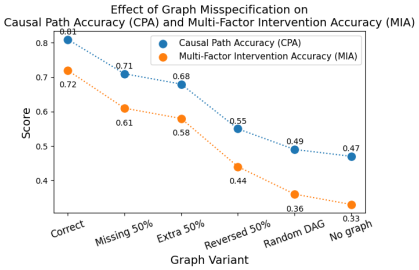


Figure 4: Effect of graph misspecification on causal performance (CPA, MIA).

Table 3: Single-factor intervention results across datasets. CelebA uses binary attribute classifiers (IA, Leakage, FID). MorphoMNIST uses ground-truth continuous factors: Thickness MAE measures intervention accuracy under do(thickness); LPIPS measures composition stability over 10 null-intervention cycles.

Method	CelebA			MorphoMNIST	
	IA ↑	Leak. ↓	FID ↓	Thick. ↓	LPIPS ↓
Concept Sliders	$.84 \pm .02$	$.28 \pm .03$	11.9	$.118 \pm .009$	$.031 \pm .003$
Linear Multi-LoRA	$.78 \pm .03$	$.35 \pm .04$	11.7	$.141 \pm .012$	$.038 \pm .004$
Deep-SCM	$.76 \pm .02$	$\mathbf{.14 \pm .02}$	18.4	$.062 \pm .005$	$.025 \pm .002$
CausalSliders (ours)	$\mathbf{.88 \pm .02}$	$.15 \pm .02$	**12.1**	$\mathbf{.054 \pm .004}$	$\mathbf{.023 \pm .002}$

Multi-factor intervention accuracy (MIA). For a target set T,

$$\mathrm{MIA}(T) = \mathbb{E}_x \left[\prod_{t \in T} \mathbb{1}(|h_t(\hat{x}_{\boldsymbol{\alpha}^{(T)}}) - z_t^\star| \le \varepsilon_t) \cdot \prod_{j \in \mathcal{V} \setminus (\mathrm{Desc}(T) \cup T)} \mathbb{1}(|h_j(\hat{x}_{\boldsymbol{\alpha}^{(T)}}) - h_j(\hat{x}_{\mathbf{0}})| \le \varepsilon_j) \right].$$

Tolerances and efficiency. For binary factors, $\varepsilon_k = 0$; for continuous factors, ε_k equals twice the validation-set standard deviation of h_k. We report trainable parameters, inference time (ms/image), and FID/LPIPS to verify that causal enforcement preserves perceptual quality.

5 Results

We evaluate CausalSliders on causal faithfulness, compositional correctness, and efficiency across CelebA (under both Simple and Complex causal graphs) and MorphoMNIST with known continuous generative factors, with qualitative examples shown in Figures 3a and 3b.

Single-Factor Interventions. Table 3 evaluates whether individual interventions achieve the target edit while preserving causally independent attributes. Across datasets, correlation-based LoRA methods exhibit substantial non-descendant leakage due to correlated attributes. CausalSliders reduces this leakage while maintaining strong intervention accuracy and comparable image quality, matching optimization-based causal baselines without per-image optimization. Results on MorphoMNIST further verify generalization to continuous factors and correct mediation along known causal chains (Table 7).

Multi-Factor Composition. Table 1 shows that graph-guided composition is essential for correctness under simultaneous interventions. Linear composition degrades sharply on CelebA-Complex, exhibiting high commutativity error and reduced causal path accuracy. By enforcing graph-ordered composition with gated mediation, CausalSliders preserves independence between unrelated factors while enabling correct downstream propagation.

Efficiency-Fidelity Tradeoff. Table 2 compares causal fidelity against inference cost. Optimization-based causal models achieve strong correctness but require expensive test-time optimization. In contrast, CausalSliders attains comparable or higher causal accuracy at near-linear inference cost, introducing only minimal parameter overhead relative to standard LoRA composition.

6 Ablation Studies

Unless stated otherwise, ablations are evaluated on CelebA-Complex, which exhibits confounding and multi-parent causal structure. (1) **Graph dependence:** Figure 4 shows that CPA/MIA degrade smoothly under missing-edge or extra-edge variants (50%) and collapse under reversed or random graphs, approaching linear composition, confirming that gains arise from approximate graph correctness rather than architectural bias. (2) **Component contributions:** Table 6 demonstrates that removing graph ordering collapses performance to linear baselines, while disabling gating or causal losses yields intermediate degradation, indicating that causal fidelity

Table 4: **Component ablation** (CelebA-Complex).

Variant	CPA ↑	MIA ↑	Leakage ↓
Full CausalSliders	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.15 \pm 0.02}$
w/o graph ordering	0.47 ± 0.03	0.33 ± 0.05	0.35 ± 0.04
w/o gating	0.63 ± 0.03	0.52 ± 0.04	0.21 ± 0.03
w/o all causal losses	0.58 ± 0.04	0.48 ± 0.05	0.26 ± 0.03
w/o minimality loss	0.71 ± 0.03	0.66 ± 0.04	0.23 ± 0.03
w/o independence loss	0.75 ± 0.02	0.68 ± 0.03	0.18 ± 0.02
w/o chain loss	0.68 ± 0.03	0.61 ± 0.04	0.16 ± 0.02

Table 5: **Adapter placement and training** (CelebA-Complex).

Configuration	CPA ↑	MIA ↑
Early layers (2–5)	0.61 ± 0.04	0.49 ± 0.05
Mid layers (6–9)	0.76 ± 0.03	0.67 ± 0.04
Late layers (10–12)	0.74 ± 0.03	0.64 ± 0.04
Mid+Late (ours)	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.72 \pm 0.03}$
Joint training	0.73 ± 0.03	0.65 ± 0.04
Two-stage (ours)	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.72 \pm 0.03}$

Table 6: **Gating strategy** (CelebA-Complex).

Gating function	CPA ↑	MIA ↑	Infer. (ms) ↓
Learned sigmoid (ours)	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.72 \pm 0.03}$	48 ± 3
Hard threshold	0.67 ± 0.03	0.55 ± 0.04	46 ± 3
Product gating	0.69 ± 0.03	0.58 ± 0.04	48 ± 3
No gating (linear)	0.47 ± 0.03	0.33 ± 0.05	$\mathbf{43 \pm 2}$

Table 7: Causal mediation on MorphoMNIST.

Method	Thickness MAE ↓	Intensity MAE ↓	FID ↓
Concept Sliders	0.118 ± 0.009	4.2 ± 0.3	10.8 ± 0.6
Linear Multi-LoRA	0.141 ± 0.012	5.1 ± 0.4	11.3 ± 0.7
Deep-SCM	0.062 ± 0.005	3.6 ± 0.2	$\mathbf{9.2 \pm 0.5}$
CausalSliders (ours)	$\mathbf{0.054 \pm 0.004}$	$\mathbf{3.3 \pm 0.2}$	9.5 ± 0.5

emerges from their joint enforcement rather than any single component. (3) **Design choices:** Tables 5 and 4 further shows that learned soft gating outperforms hard alternatives, and that mid-to-late layer adapters with two-stage training yield the strongest causal control, consistent with semantic modulation occurring at higher UNet layers.

7 Discussion

Observational Overfitting vs. Structural Intervention. Our results show that state-of-the-art editing methods, such as Concept Sliders degrade under composition not due to limited capacity, but due to *observational overfitting*. These methods optimize for observational associations, $P(y \mid x)$, and therefore learn and amplify spurious correlations present in the data (e.g., Age ↔ Glasses). As a result, visually plausible edits often violate counterfactual intent.

CausalSliders reframes editing as a problem of *structural intervention*. We show that specifying causal relations alone is insufficient: without architectural constraints, correlated attributes remain entangled. Graph-ordered composition, together with gated execution and minimality-based constraints, provides the mechanism required to encourage $P(y \mid do(x))$ behavior. This design promotes interventions that primarily affect the target factor and its causal descendants, while preserving independent attributes such as identity and background.

Evaluation under Controlled Structure. Although CelebA is sometimes viewed as a limited benchmark, it is one of the few datasets that supports quantitative evaluation of causal consistency. Following prior work on benchmarking counterfactual image generation, we use CelebA and MorphoMNIST to enable rigorous measurement of non-descendant leakage, causal path accuracy, and mediation effects. Such metrics are ill-defined on open-ended web-scale datasets, where causal structure and semantic ground truth are unavailable. Our evaluation therefore prioritizes falsifiable evidence of causal control rather than qualitative visual diversity alone.

Limitations and Outlook. A primary limitation of CausalSliders is its reliance on a predefined causal graph; performance degrades under graph misspecification, a limitation shared by other causal editing approaches. Scaling to many attributes introduces challenges: training a separate LoRA for each semantic factor can be costly, and per-attribute adapters increase parameter count and inference overhead. Extending the framework to open-vocabulary or prompt-defined attributes remains unexplored and may require additional mechanisms to avoid correlation-driven entanglement. Future work may mitigate graph dependence by integrating causal discovery or external structural priors.

Beyond image editing, graph-ordered parameter interventions may benefit AI for science domains where causal structure is partially known. For example, in *molecular property optimization*, causal graphs may encode dependencies such as Chirality → Binding Affinity ← Molecular Weight, enabling interventions that preserve upstream factors while editing downstream properties.

Conclusion.

CausalSliders demonstrates that causal reasoning can be embedded into diffusion-based image editing through structured low-rank adaptation. By coordinating low-rank adapters according to a directed acyclic graph, the method bridges the gap between the visual fidelity of modern diffusion models and the semantic requirements of counterfactual reasoning, without per-instance optimization or full retraining.

REFERENCES

Malek Ben Alaya, Daniel M. Lang, Benedikt Wiestler, Julia A. Schnabel, and Cosmin I. Bercea. Mededit: Counterfactual diffusion-based image editing on brain mri, 2024. URL https://arxiv.org/abs/2407.15270.

Grigorii Alekseenko, Aleksandr Gordeev, Irina Tolstykh, Bulat Suleimanov, Vladimir Dokholyan, Georgii Fedorov, Sergey Yakubson, Aleksandra Tsybina, Mikhail Chernyshov, and Maksim Kuprashevich. Vibe: Visual instruction based editor, 2026.

Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: Quantitative assessment and diagnostics for representation learning, 2019. URL https://arxiv.org/abs/1809.10780.

Juan D. Correa and Elias Bareinboim. Counterfactual graphical models: Constraints and inference. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Z1qZoHa6ql.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. URL https://arxiv.org/abs/2210.11427.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining, 2023. URL https://arxiv.org/abs/2208.12262.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, 2024. arXiv:2311.12092.

Gemini Team Google. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Google AI for Developers. Nano banana (image generation) — gemini api documentation. https://ai.google.dev/gemini-api/docs/image-generation, 2026. Accessed: 2026-01-28.

Google DeepMind. Introducing nano banana pro. https://blog.google/technology/ai/nano-bananapro/, November 2025. Published: 2025-11-20, Accessed: 2026-01-28.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models, 2023. URL https://arxiv.org/abs/2305.18292.

Zhenghao He, Guangzhi Xiong, Boyang Wang, Sanchit Sinha, and Aidong Zhang. Casl: Concept-aligned sparse latents for interpreting diffusion models, 01 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Chao Huang, Susan Liang, Yunlong Tang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Scaling concept with text-guided diffusion models, 2025. URL https://openreview.net/forum?id=HafxTJjo6a.

Utkarshani Jaimini and Amit Sheth. Causalkg: Causal knowledge graph explainability using interventional and counterfactual reasoning, 2022. URL https://arxiv.org/abs/2201.03647.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training, 2017. URL https://arxiv.org/abs/1709.02023.

Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling, 2024. URL https://arxiv.org/abs/2310.11011.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

K. Lee, D. Plecko, and E. Bareinboim. Causal explanations through counterfactual variable attributions. Technical Report R-135, Columbia CausalAI Laboratory, May 2025. URL https://causalai.net/r135.pdf.

Mengtian Li, Jinshu Chen, Wanquan Feng, Bingchuan Li, Fei Dai, Songtao Zhao, and Qian He. Hyperlora: Parameter-efficient adaptive generation for portrait synthesis, 2025. URL https://arxiv.org/abs/2503.16944.

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.

Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph E. Gonzalez, Zhifeng Chen, Russ Salakhutdinov, and Ion Stoica. Stylus: Automatic adapter selection for diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3Odq2tGSpp.

Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Benchmarking counterfactual image generation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=0T8xRFrScB.

Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Benchmarking counterfactual image generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, 2024b. ISBN 9798331314385.

OpenAI. The new chatgpt images is here. https://openai.com/index/new-chatgpt-images-is-here/, 2025a. OpenAI Index, Accessed: 2026-01-28.

OpenAI. Gpt image 1.5 model — openai api documentation. https://platform.openai.com/docs/models/gpt-image-1.5, 2025b. OpenAI Platform Documentation, Accessed: 2026-01-28.

Yushu Pan and Elias Bareinboim. Counterfactual image editing with disentangled causal latent space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=u2Lgi4NIe7.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference, 2020. URL https://arxiv.org/abs/2006.06485.

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 (CVPR)*, pp. 10684–10695, June 2022.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras, 2023. URL https://arxiv.org/abs/2311.13600.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Deepak Sridhar and Nuno Vasconcelos. Prompt sliders for fine-grained control, editing and erasing of concepts in diffusion models, 2024. URL https://arxiv.org/abs/2409.16535.

Lei Tong, Zhihua Liu, Chaochao Lu, Dino Oglic, Tom Diethe, Philip Teare, Sotirios A. Tsaftaris, and Chen Jin. Causal-adapter: Taming text-to-image diffusion for faithful counterfactual generation, 2025. URL https://arxiv.org/abs/2509.24798.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022. URL https://arxiv.org/abs/2211.12572.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22532–22541, June 2023.

Siwei Xia, Li Sun, Tiantian Sun, and Qingli Li. DragloRA: Online optimization of loRA adapters for drag-based image editing in diffusion model. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=b74kufhhsk.

Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Junzhe Zhang and Elias Bareinboim. Causal canonical modeling for confounding robust treatment evaluation, 2025. URL https://openreview.net/forum?id=cP2nOl3t3W.

IMPACT STATEMENT.

This work improves the causal reliability of image editing by enforcing graph-structured, parameter-efficient interventions in diffusion models. The approach reduces unintended attribute leakage, supporting more trustworthy use in fairness analysis, controlled data generation, and scientific applications. Like all generative models, it may be misused for deceptive editing, underscoring the need for responsible deployment.