LINA: Learning INterventions Adaptively for Physical Alignment and Counterfactual Generation in Diffusion Models

Shu Yu[1,2,3] and Chaochao Lu[1⋆]

[1] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] Fudan University, Shanghai, China
{yushu, luchaochao}@pjlab.org.cn

Abstract. Diffusion models (DMs) have achieved remarkable success in image and video generation. However, there are still challenges in (1) physical alignment and (2) counterfactual generation. To diagnose the root causes, we introduce the Causal Scene Graph (CSG) to model DM's generative process, and build the Physical Alignment Probe (PAP) dataset to quantify the failure modes. Our analysis yields three key insights: (1) DMs struggle with physical elements not explicitly determined in the prompt; (2) the prompt embedding contains disentangled representations for texture and physics; (3) visual causal structure is disproportionately established during the initial, computationally constrained denoising steps. Based on these findings, we introduce **LINA (Learning INterventions Adaptively)**, a novel framework that learns to predict prompt-specific interventions. LINA employs (1) targeted guidance in the prompt and visual latent spaces, and (2) a reallocated, causality-aware denoising schedule. Without modifying pre-trained DM weights, LINA enforces both physical alignment and counterfactual generation in image and video DMs, achieving state-of-the-art performance on challenging causal generation tasks and the Winoground dataset. Our project page is at https://opencausalab.github.io/LINA

Keywords: Diffusion Models · Causal Reasoning · Physical Alignment · World Simulator

1 Introduction

Diffusion models (DMs) are proposed to achieve both flexibility and tractability in generative modeling [44], by learning to predict the injected noise in the data point at any timestep [16, 45, 46]. These models first achieved significant success in image generation [11, 16, 39, 41, 42, 45], and have since been extended to video [3, 18, 49], audio [6, 22, 26], and text [24, 30].

DMs for visual generation are widely considered a promising path toward realizing world models [9]. However, despite their success, state-of-the-art (SOTA)

⋆ Corresponding author.

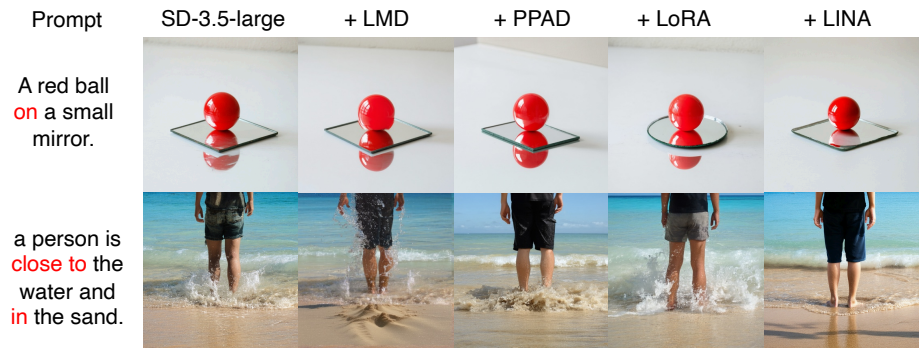

Fig. 1: Failures in DMs [11,25,28,36] and LINA's improvement. (a) Generated with prompt: "A red ball on a small mirror." Baselines generate reflections extending beyond the mirror surface. (b) Generated with Winoground [47] prompt: "a person is *close to* the water and *in* the sand." Baselines incorrectly place the person *in* the water.

DMs still struggle with two key challenges: (1) **physical alignment** [14] and (2) **counterfactual generation** [4,7,17]. We argue that these two issues stem from the DMs' deficient understanding of causality [14,33,43,48], which manifests in two key failures: (1) modeling **symmetric correlations** rather than directional causality, preventing physical alignment, and (2) the **entanglement of causal factors**, hindering the novel recombination of visual elements for counterfactual generation. As shown in Fig. 1, DMs produce redundant reflections in areas even without mirrors (Fig. 1a), or entangle concepts such as "person" and "water", thus depicting the "person in the water" rather than following instructions (Fig. 1b).

To diagnose the root of these failures, first, we introduce the **Causal Scene Graph (CSG)**, a representation modeling the generative process of DMs. Inspired by Causal Graphical Model (CGM) [34] and Scene Graph (SG) [5], CSG unifies causal dependencies and spatial layouts. Second, based on CSG, we construct the **Physical Alignment Probe (PAP) dataset**, which consists of structured prompts, a corresponding set of images generated by SOTA DMs, and segmentation masks for elements in these images. We conduct diagnostic probing using CSG-guided interventions on the PAP dataset, which yields three key insights: (1) DMs struggle with physical elements not explicitly determined in the prompt; (2) the prompt embedding contains disentangled representations for texture and physics; (3) visual causal structure is disproportionately established during the initial, computationally constrained denoising steps.

Based on these findings, we formalize two complementary intervention mechanisms—token-level calibration and latent-level contrastive guidance. We introduce **LINA (Learning INterventions Adaptively)**, which employs a lightweight **Adaptive Intervention Module (AIM)** to predict the strengths of these two interventions adaptively for each prompt. By combining these targeted interventions with a reallocated, causality-aware schedule, LINA enforces causal consistency without modifying pre-trained weights or relying on external

MLLMs [20, 28, 50] during inference, achieving SOTA performance on PAP and Winoground [47] dataset.

Overall, our contributions are as follows:

(1) We introduce the **Causal Scene Graph (CSG)**, a representation that unifies causal dependencies and spatial layouts, providing a basis for diagnostic interventions.
(2) We construct the **Physical Alignment Probe (PAP)**, a dataset consisting of structured prompts, SOTA-generated images, and fine-grained masks to quantify DMs' physical and counterfactual generation failures.
(3) We perform a **diagnostic analysis** via CSG-guided masked inpainting, providing the **first** quantitative evaluation of DMs' multi-hop reasoning failures through **bidirectional** probing of edges in the CSG.
(4) We propose **LINA**, a framework that learns to predict and apply prompt-specific interventions, achieving SOTA alignment on image and video DMs without MLLM inference or retraining.

2 Preliminaries

2.1 Causal Graphical Model

A Causal Graphical Model (CGM) captures the causal structure among variables in a system [34, 53]. It comprises a set of variables $X_1, \ldots, X_n$ along with a joint distribution, represented as a directed acyclic graph (DAG). In this graph, each node corresponds to a variable, and each directed edge $X_i \to X_j$ denotes a direct causal influence from X_i to X_j. CGMs can be used to describe causal relationships in image generation. We denote the prompt as X and the generated image as Y. Both X and Y are partitioned into semantic units, represented as $X = \{x_1, x_2, \ldots\}$ and $Y = \{y_1, y_2, \ldots\}$. We define:

Direct elements $\boldsymbol{Y_D}$: entities $y_i \in Y$ that are exactly one directed edge away from some $x_j \in X$ in the CGM, *i.e.*, elements explicitly determined in the prompt. ***Indirect elements*** $\boldsymbol{Y_I}$: entities $y_k \in Y$ that are causally downstream of some x_j but separated by one or more intermediate nodes, arising from implicit physical laws (*e.g.*, shadows, reflections). Causality in DMs has a clear hierarchy: $\boldsymbol{X \to Y_D \to Y_I}$. We will adhere to these definitions throughout the paper.

2.2 Scene Graph

A Scene Graph (SG) is another form of structured image representation, organizing an image into elements (nodes) and their pairwise spatial relationships (edges) [5]. Unlike CGMs, SG may contain cycles (*e.g.*, three people each "near" one another). We define:

Direct layout: spatial relations explicitly stated in the prompt (*e.g.*, "a lightbulb surrounding some plants"). Such relations can appear in counterfactual generation. ***Indirect layout***: spatial relations not stated in the prompt but implied by physical laws (*e.g.*, density implies "iron sinks in water"). We will adhere to these definitions throughout the paper.

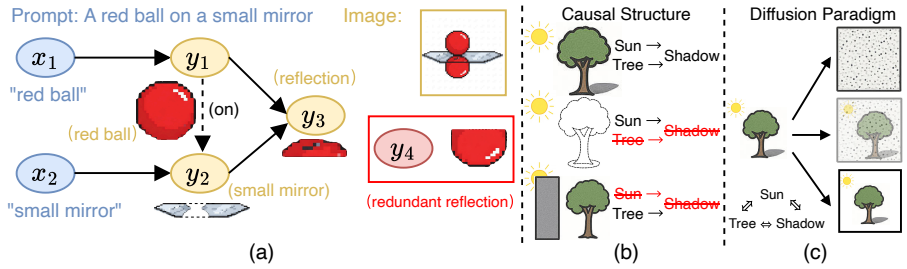

Fig. 2: (a) Causal Scene Graph (CSG). CSG unifies causal dependencies and spatial layouts into a single graph. Nodes represent semantic units from the prompt (X) and generated image (Y). Edges comprise **Causal Edges** (E_C, solid arrows) modeling physical dependencies (*e.g.*, $X \to Y_D \to Y_I$) and **Spatial Edges** (E_S, dashed arrows) modeling spatial relationships. **(b)** In reality, removing a tree causes its shadow to disappear, whereas obscuring the shadow does not remove the tree. The causal relationship has a **clear direction**. **(c)** In DMs, both the tree and its shadow are learned to be denoised **simultaneously** without causal direction.

3 Causal Diagnostics of Diffusion Models

In this section, we diagnose the root of DMs' causal failures. We first introduce the Causal Scene Graph (CSG) to formally define the problem. Then we introduce the Physical Alignment Probe (PAP) dataset to quantify these failures. Finally, we use PAP to conduct two sets of probe experiments to localize the failure point (indirect elements in multi-hop reasoning) and the failure source (guidance miscalibration in the generation process).

3.1 Causal Scene Graph and Problem Formulation

We introduce the Causal Scene Graph (CSG), as shown in Fig. 2(a). Formally, we define a CSG as a graph $G_{\text{CSG}} = (V, E)$, where the node set $V = X \cup Y$ comprises all semantic units from the prompt $X = \{x_i\}$ and the image $Y = \{y_j\}$. The edge set $E = E_C \cup E_S$. E_C contains **Causal Edges** ($v_i \to v_j$), which form a DAG to model the directional flow of physical influence. E_S contains **Spatial Edges**, which model the spatial layout of elements.

In practice, extracting CSGs faces robustness challenges: semantic concepts (*e.g.*, "correct reflections") constitute continuous manifolds where even human annotations lack consistency. We address this by localizing bounding boxes and key points with deterministic thresholds, ensuring reproducible evaluation (Sec. 3.2).

CSG highlights an inherent limitation in DM's paradigm. In conventional DM training, all elements within an image are denoised simultaneously. Consequently, causally hierarchical elements (*e.g.*, a tree as the cause Y_D and its shadow as the effect Y_I) are learned concurrently and symmetrically in a **flattened, non-hierarchical** paradigm, as illustrated in Fig. 2(b)-(c).

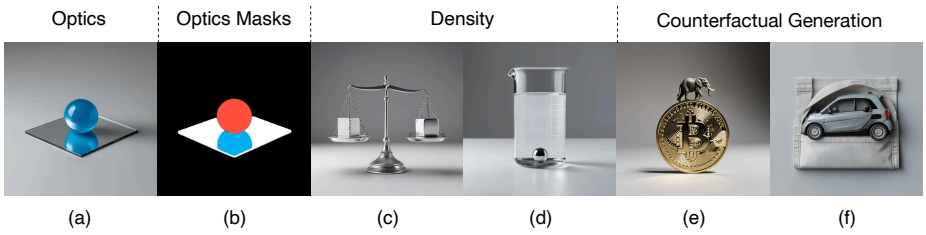

Fig. 3: Samples from the 28,700-image corpus and segmentation masks in our Physical Alignment Probe (PAP) Dataset. (a) Optics (reflection), (b) Segmentation masks for optics, (c-d) Density (balance and buoyancy), (e-f) Counterfactual Generation (size reversal and impossible containment).

3.2 The Physical Alignment Probe (PAP) Dataset

To quantitatively measure the distortion of the generation process, we construct the **Physical Alignment Probe (PAP) dataset**. PAP is a multi-modal corpus based on CSG for diagnosing physical reasoning and counterfactual generation. It comprises three core components: (1) a structured library of prompts, (2) a large-scale image corpus generated from these prompts using SOTA DMs (SD-3.5-large [11] and FLUX.1-Krea-dev [2]), and (3) fine-grained segmentation masks to facilitate diagnostic interventions, as illustrated in Fig. 3.

Specifically, the prompt library consists of 287 prompts, which are divided into three diagnostic subsets: **Optics**, **Density** [14, 29], and **Counterfactual Generation**. We generate 50 images per prompt for each DM, forming a 28,700-image corpus. To assess physical alignment, we employ an MLLM-based evaluator (Qwen2.5-VL-72B [1]) to extract bounding boxes and key points for semantic elements, which enable deterministic rule-based verification of the generated CSG. These annotations also facilitate the generation of precise **segmentation masks** via SAM2 [40] for subsequent diagnostics. Empirically, we have demonstrated the robustness and consistency of this evaluation in Sec. 5.4.

3.3 Probe Diffusion Model's Multi-Hop Reasoning

To systematically diagnose the failure in the causal hierarchy $X \to Y_D \to Y_I$, we use the masks from our PAP dataset to conduct a series of masked inpainting probes. Adopting the notations from Sec. 2, we design five diagnostic experiments on SD-3.5-large, aiming to probe the bidirectional information flow along edges in the CSG, and to distinguish true *directional* reasoning from *superficial* symmetric correlations and thereby assess the DM's causal inference capability.

(I) $P(Y_I|X, Y_D)$, **forward inference**. The success rate (81.1%) shows no significant improvement over $P(Y_I|X)$ (80.4%). **Finding:** DMs fail to leverage visual context (Y_D) to improve physical reasoning for indirect elements.

(II) $P(Y_D|X, Y_I)$, **mediator inference**. The success rate (94.9%) is similar to $P(Y_D|X)$ (94.3%).

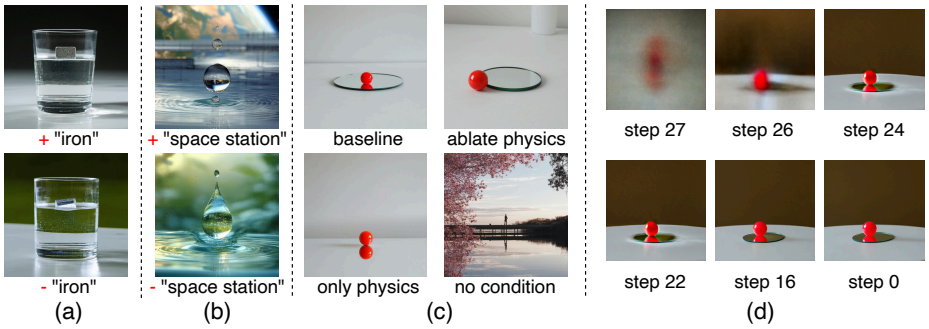

Fig. 4: (a) Iron and ice blocks exhibit identical **floating** behavior. **(b)** Water surface textures remain **gravity-bound** regardless of microgravity specification. **(c)** Ablating relation tokens (*e.g.*, "on") removes spatial layout and physical interactions, while retaining only these tokens produces the core physical texture but loses object semantics. **(d)** The denoising process establishes the causal structure in the initial, computationally constrained steps (step 26 to 24, of 28 total).

(III) $P(Y_D|X, Y_I')$, **mediator inference with conflict**, where Y_I' conflicts with X (*e.g.*, X specifies a "red ball" but Y_I' is a "green reflection"). In 94.0% of cases, the generated Y_D aligns with X, while 0.4% aligns with the conflicting Y_I'. **Finding:** Experiments II and III show that DMs prioritize direct supervision from X when generating Y_D, largely ignoring causal consistency with Y_I.

(IV) $P(Y_I|Y_D)$, **forward inference without prompt**. The success rate plummets to 2.2%.

(V) $P(Y_D|Y_I)$, **backward inference without prompt**. The success rate is 2.4%. **Finding:** Causal relationships cannot be established from visual context alone without the textual prompt X.

Collectively, the causal hierarchy $X \to Y_D \to Y_I$ collapses into a parallelized, correlation-based mapping $(X \to Y_D, X \to Y_I)$, where physical consistency between visual elements is lost, as shown in Fig. 4(a)-(b).

3.4 Causal Representation in Embedding Space

Having localized the causal failure in multi-hop reasoning, we investigate its representational basis within the prompt embedding space. Based on our CSG definitions (Sec. 3.1), we define two metrics: (1) **Texture Alignment**: the success rate for *direct elements* ($\boldsymbol{Y_D}$); and (2) **Physical Alignment**: the success rate for *indirect elements* ($\boldsymbol{Y_I}$).

First, we perform intervention on **relation tokens**. Specifically, they encompass (1) **interaction verbs** (*e.g.*, "hits", "eats") that explicitly denote causal edges E_C, and (2) **spatial prepositions** (*e.g.*, "on", "in") that define the spatial edges E_S. We replace these token embeddings with the unconditional ones at the same positions. The Physical Alignment success rate collapses from its **80.4%** baseline to just **5.2%**, while Texture Alignment remains high (**91.8%** vs. its **94.3%** baseline).

Table 1: Comparison of DM alignment paradigms. Existing methods intervene at different levels of the generative process: $X \mapsto Y$. See supplementary materials for detailed computational cost analysis and further implementation details.

Method	Intervention Target	Overhead	Causal Physics
LMD [25]	Input space (X)	+68%	Layout only
PPAD [28]	Latent trajectory (z_t)	+372%	Limited
LoRA [36]	Model weights (θ)	Baseline (+ training)	Overfitting
LINA (Ours)	**Guidance dynamics** ($\mathcal{F}$)	**+25%**	**Strong**

Second, we perform intervention on **object and attribute tokens** (*e.g.*, "red", "ball"). As a result, the Texture Alignment collapses to **3.1%**. Yet, Physical Alignment is remarkably preserved (**92.5%**), as the model still generates the reflection even without the mirror.

An illustrative example of this disentanglement is shown in Fig. 4 (c), which indicates that the model's internal knowledge is clearly represented, yet lacks proper guidance dynamics $\mathcal{F}$. We can effectively steer $\mathcal{F}$ toward target CSG (G^*_X) by **adaptively** modulating the influence of relation tokens, without altering the underlying model ϵ_θ.

Furthermore, we analyze the denoising process to pinpoint *when* this causal structure emerges. We inspect the intermediate latent states from SD-3.5-large's official 28-step denoising schedule. As shown in Fig. 4 (d), the causal structure is established at the very beginning few steps of the process. In **97.8%** of successful generations on the PAP-Optics subset, the correct structure is identifiable within the initial 2-4 iterations (*i.e.*, at steps 26-24 of the 28-step reverse denoising schedule, corresponding to the highest noise levels). This strongly motivates a computational reallocation on this critical **causal structure formation** period.

4 Method

In this section, we introduce **LINA (Learning INterventions Adaptively)**. Based on our diagnostics in Sec. 3.4, we formalize two targeted interventions (Sec. 4.1): **Token-Embedding-level** Calibration and **Visual-Latent-level** Contrastive Guidance. Leveraging these interventions, LINA adaptively calibrates the sampling dynamics of $\mathcal{F}$ according to different input prompts, as shown in Fig. 5. First, in the offline phase (Sec. 4.2), we utilize an MLLM as a robust evaluator to guide an automated search for optimal intervention strengths. These discovered optima serve as supervision to train a lightweight **Adaptive Intervention Module (AIM)**. Second, in the online phase (Sec. 4.3), the intervention strengths predicted by the AIM are applied in conjunction with a **Causality-Aware Denoising Schedule**, which reallocates computation to the critical causal structure formative period. This process operates without MLLM overhead or DM retraining. See Tab. 1 for a comparison with existing methods.

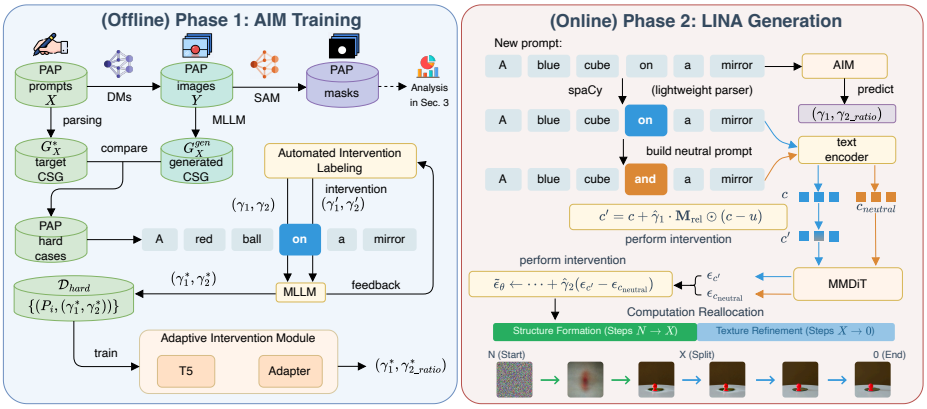

Fig. 5: Overview of the LINA framework, which operates in two phases. (Left) Phase 1: (Offline) AIM Training. We identify baseline failures ("hard cases") from our PAP dataset. An MLLM evaluator performs an automated comparative search to find optimal intervention strengths (γ_1^*, γ_2^*) for these prompts. This creates a dataset $\mathcal{D}_{hard}$, which is used to train the AIM to predict these strengths directly from a prompt. **(Right) Phase 2: (Online) LINA-Guided Generation.** For a new prompt X_{new}, AIM predicts $(\hat{\gamma}_1, \hat{\gamma}_2)$ to modulate (1) **Token-Embedding-level calibration**, (2) **Visual-Latent-level contrastive guidance**, and (3) **computation reallocation** to the initial structure formation phase.

4.1 Causal Intervention Mechanisms

Token-Embedding-level: Causal Edge Calibration ***(γ_1)*** We formulate γ_1 to calibrate the causal edge $Y_D \to Y_I$, which generalizes the token replacement intervention used in our diagnostics (Sec. 3.4) to a continuous, *soft* guidance strength. To implement this, we identify relation tokens via a lightweight language parser (spaCy [19]) and construct a binary mask $\mathbf{M}_{\text{rel}}$ targeting their embeddings. Given a conditional embedding $c = Emb(X)$ and an unconditional embedding $u = Emb(\varnothing)$, we compute the calibrated embedding c':

$$c' = c + \gamma_1 \cdot \mathbf{M}_{\text{rel}} \odot (c - u) \tag{1}$$

where $\odot$ denotes element-wise multiplication. By scaling the semantic direction vector $(c - u)$, this intervention selectively amplifies or attenuates the causal signal strength within the prompt embedding prior to the denoising process.

Visual-Latent-level: Contrastive Causal Guidance ***(γ_2)*** We introduce γ_2 to isolate the causal signal. To achieve this, we automatically construct a "neutral" reference prompt X_{neutral} by relaxing the specific relation tokens in X to generic conjunctions (*e.g.*, substituting "on" with "and"). The guidance term establishes a **contrastive causal signal** of the relation tokens.

The final noise prediction $\tilde{\epsilon}_\theta$ at timestep t is computed by combining the standard Classifier-Free Guidance (CFG) [17] with our contrastive term:

$$\begin{aligned}\tilde{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, u, t) \\ + \gamma_0(\epsilon_\theta(x_t, c', t) - \epsilon_\theta(x_t, u, t)) \\ + \gamma_2(\epsilon_\theta(x_t, c', t) - \epsilon_\theta(x_t, c_{\text{neutral}}, t))\end{aligned} \tag{2}$$

where c' is the calibrated embedding from Eq. 1, $c_{\text{neutral}} = Emb(X_{\text{neutral}})$, and γ_0 is the standard CFG scale.

4.2 Phase 1: AIM Training

The offline phase aims to train an adaptive mapping Φ, which predicts intervention strengths $\Phi(X)$ conditioned on the input prompt X to ensure causal alignment. To construct the training dataset $\mathcal{D}_{hard}$, we identify baseline failures (hard cases) from the PAP image corpus (Sec. 3.2) and determine their optimal intervention strengths (γ_1^*, γ_2^*) via an MLLM-based automated search.

Automated Intervention Labeling. Manually calibrating the target intervention strengths (γ_1, γ_2) for every hard case input is intractable. To address this, we leverage our established composite evaluation framework in Sec. 3.2. This framework identifies the optimal intervention strengths (γ_1^*, γ_2^*) for each hard case X_i, supported by an analysis of the intervention landscape.

(1) **Identify Baseline Failures:** We identify the prompts and seeds from the PAP dataset where baseline DMs fail to satisfy the causal requirements.
(2) **Coordinate Descent Search:** We employ an efficient Coordinate Descent strategy. Empirically, the effects of Visual-Latent-level (γ_2) and Token-Embedding-level (γ_1) interventions are largely **disentangled and monotonic**, which validates the efficacy of our coordinate descent strategy. More details can be found in the supplementary material.
(3) **Construct Target Dataset:** This process yields our structured training dataset, $\mathcal{D}_{hard} = \{(X_i, (\gamma_1^*, \gamma_2^*))\}_{i=1}^N$. Empirically, the optimal intervention strengths are **robust to initialization noise**, clustering tightly for a given prompt across random seeds. This stability confirms the feasibility of training a deterministic mapping $\Phi(X)$ to predict these values. More details can be found in the supplementary material.

AIM Training. We design the **Adaptive Intervention Module (AIM)**, denoted as Φ, to learn the $\mathcal{D}_{hard}$. Φ consists of a pre-trained T5 text encoder [38] followed by a lightweight MLP regression head. It is trained on $\mathcal{D}_{hard}$ using a standard L2 regression loss to predict the intervention strengths:

$$\mathcal{L} = \sum_{i=1}^{N} \|\Phi(X_i) - (\gamma_1^*, \gamma_2^*/\gamma_0)\|_2^2 \tag{3}$$

We predict the ratio γ_2^*/γ_0 (as $\hat{\gamma}_{2\_ratio}$) for normalization, where γ_0 denotes the DM's official classifier-free guidance scale (*e.g.*, 3.5 for SD-3.5-large).

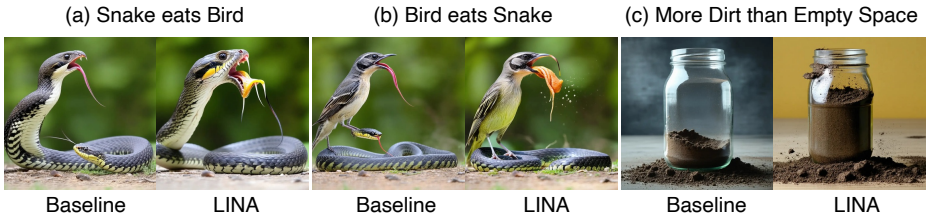

Fig. 6: Qualitative results on Winoground [47]. Left: Baseline (SD-3.5-large); **Right:** LINA. (**a**) "*Snake eats Bird*" (common prior): Baseline exhibits texture artifacts; LINA refines the causal structure. (**b**) "*Bird eats Snake*" (counterfactual): LINA achieves significantly better quality in counterfactual generation. (**c**) "*More Dirt than Empty Space*": Baseline biased toward empty jars; LINA corrects the quantitative prior.

4.3 Phase 2: Causality-Aware Denoising

In the online generation phase, the pre-trained AIM is deployed to guide the generative process for any new user prompt X_{new}, enabling physical alignment, counterfactual generation, and bias correction, as illustrated in Fig. 6.

Pre-processing For a new prompt X_{new}, we first compute the embeddings $u = Emb(\varnothing)$, $c = Emb(X_{new})$, and $c_{\text{neutral}} = Emb(X_{\text{neutral}})$. Simultaneously, the AIM predicts the intervention strengths $(\hat{\gamma}_1, \hat{\gamma}_{2\_ratio}) = \Phi(X_{new})$, and the parser identifies the relation token mask $\mathbf{M}_{\text{rel}}$.

Causality-Aware Denoising Schedule Motivated by our finding (Sec. 3.4) that the causal structure emerges during the initial high-noise timesteps, we implement a Computation Reallocation strategy to prioritize this stage. We build upon the time-shifting employed in modern DMs (*e.g.*, SD-3.5-large [11]), which maps the standard uniform time steps $\tau \in [0, 1]$ to a shifted schedule τ_s:

$$\tau_s = \frac{s \cdot \tau}{1 + (s - 1) \cdot \tau} \tag{4}$$

where s is the shift parameter. We explicitly increase the shift parameter s to further concentrate sampling density in the early structure-formation phase (*e.g.*, we increase s from the official 1.15 for SD-3.5-large to 1.5). The full LINA-guided noise prediction $\tilde{\epsilon}_\theta$ (Eq. 2) is applied throughout this causality-aware schedule.

5 Experiments

5.1 Experimental Setup

We conduct experiments using two SOTA DMs: SD-3.5-large [11] and FLUX.1-Krea-dev [2]. Following the methodology in Sec. 3.2, we employ Qwen2.5-VL-72B [1] to extract **structured representations** (*e.g.*, bounding boxes, key points) for **deterministic rule-based verification** of the Generated CSG. Detailed evaluation thresholds and reproducible criteria are provided in the supplementary materials.

Table 2: Gap between Causal Identification and Generative Correction. We evaluate SOTA closed-source models on $\mathcal{D}_{hard}$ test set. **Identification:** GPT-5's accuracy in detecting physical violations (given image+prompt). **Correction:** Editing models are given the flawed image and correction prompt. "+ CoT" variants perform CoT analysis before correction.

Method	**Optics** (%, ↑)	**Density** (%, ↑)
Nano Banana [12]	2.5	25.0
GPT-Image [31]	2.5	22.5
Nano Banana + CoT	45.0	70.5
GPT-Image + CoT	**67.5**	**82.0**
GPT-5 [32] (Identification Accuracy)	99.5	100.0

5.2 Baselines

Our baselines include (1) **SOTA Closed-Source Models**: Nano Banana [12] and GPT-Image [31]. These serve as black-box baselines and editing systems. (2) **Top-tier Open-Source DMs**: SD-3.5-large (SD-3.5) [11], FLUX.1-Krea-dev (FLUX.1) [2], and Wan2.2-T2V-A14B (Wan-2.2) [49]. These models represent the baseline guidance dynamics $\mathcal{F}$ that we apply corrections to.

We also select three open-source methods representing the three primary correction strategies. (3) **Prompt Engineering Methods**: LLM-grounded Diffusion (LMD) [25]. This strategy corrects the *input* to $\mathcal{F}$, seeking X' such that $\mathcal{F}(G_{X'}) \approx G^*_X$. (4) **MLLM-in-the-loop Methods**: Ping-Pong-Ahead Diffusion (PPAD) [28]. This strategy uses an external corrector M to guide the *output* of $\mathcal{F}$, achieving $M(\mathcal{F}(G^*_X)) \approx G^*_X$. (5) **Finetuning Methods**: LoRA finetuning implemented in Diffusers [36]. This strategy modifies the generation process by retraining the DM's weights θ.

5.3 Main Results

Evaluation on Physical Alignment

Analysis of SOTA Closed-Source Systems Building upon the diagnostic insights in Sec. 3, we investigate the persistent disconnect between *causal identification* and *generative correction*. Even advanced proprietary systems (*e.g.*, GPT-Image [31]) frequently lack the generative control to rectify the physical inconsistencies their identification counterparts (*e.g.*, GPT-5 [32]) correctly diagnose. We quantify this "identification-correction gap" in Tab. 2, utilizing the $\mathcal{D}_{hard}$ test split to ensure a rigorous evaluation and API efficiency. Notably, LINA is not an image editing model and thus cannot be directly compared in Tab. 2. For reference, under identical prompts and seeds, LINA (re-generation) achieves **96.4%** on Optics and **86.0%** on Density.

Table 3: Main Results on Physical Alignment and Counterfactual Generation. We evaluate LINA and SOTA baselines on SD-3.5-large (and FLUX.1 where indicated) across four metrics: Physical Alignment (Optics and Density subsets) and Counterfactual Simulation (Winoground and PAP-OOD subsets). The metric is Success Rate (% ↑). Best results are in **bold**.

Method	Optics (↑)	Density (↑)	Winoground (↑)	PAP-OOD (↑)
SD-3.5 (Baseline)	80.4±4.0	54.2±4.5	54.4±3.4	69.3±3.8
FLUX.1 (Baseline)	86.9±3.3	64.3±4.1	65.5±4.2	80.6±3.2
SD-3.5+LMD [25]	80.5±3.7	81.5±3.4	73.1±4.0	75.2±3.5
SD-3.5+PPAD [28]	91.7±2.6	76.2±3.3	62.6±4.3	74.1±3.4
SD-3.5+LoRA [36]	95.9±4.0	91.3±3.9	57.3±4.5	72.0±3.8
LINA (SD-3.5)	**97.4±2.1**	**92.3±2.0**	**79.5±2.0**	**84.3±2.4**
LINA (FLUX.1)	**96.8±1.9**	**94.0±2.1**	**83.0±2.5**	**86.1±2.2**

Open-Source Methods All baselines are applied to SD-3.5-large for a fair comparison. As shown in Tab. 3, LINA significantly outperforms all baseline methods. LMD [25] enforces spatial layout but struggles to capture physical laws. PPAD [28] suffers from the "identification-correction gap" observed in the closed-source systems; its MLLM identifies flaws, but its corrections fail to override the DM's flawed internal dynamics. LoRA finetuning [36] overfits to correlations rather than fixing the underlying symmetric paradigm. In contrast, LINA succeeds by fixing causal guidance miscalibration (Sec. 3.4).

Evaluation on Counterfactual Generation We evaluate counterfactual generation using two challenging benchmarks. The first is our **PAP-OOD subset**, which contains Out-of-Distribution (OOD) spatial and causal instructions. The second is the **Winoground** [47] dataset. Its paired, contrastive prompts probe model biases against uncommon distributions. We adopt the 171-sample subset identified by [10], which excludes ambiguous samples. As shown in Tab. 3, baseline DMs fail by collapsing to prior correlations. While LMD [25] improves performance by generating an explicit scene layout, it cannot fully override the DM's bias. Other methods, such as PPAD [28] and LoRA [36], provide only marginal gains. LINA achieves SOTA performance by directly targeting this bias.

Evaluation on Video Generation To evaluate the generalizability of LINA, we extend our framework to the video generation domain using the SOTA model Wan-2.2 [49]. Video generation introduces distinct challenges compared to images: causality manifests as dynamic processes, and models exhibit heightened sensitivity to action verbs. Consequently, we utilize the **PAP-Density** subset (*e.g.*, sinking, floating) and the **Winoground** [47] dataset to probe physical consistency across temporal sequences (*i.e.*, initial state, motion dynamics, and final state), using Qwen2.5-VL-72B as the evaluator.

Table 4: Ablation Study. Analysis of LINA's components on SD-3.5-large. We evaluate on PAP-Optics (Opt.), PAP-Density (Dens.), and Winoground (Wino.). Metric is Success Rate (% ↑).

Method	Opt. (% ↑)	Dens. (% ↑)	Wino. (% ↑)
LINA	**97.4±2.1**	**92.3±2.0**	**79.5±2.0**
w/o γ_1	85.1±3.6	80.5±3.8	60.2±3.5
w/o γ_2	81.3±3.9	78.0±4.1	74.9±3.4
Fixed γ	90.5±3.1	85.2±3.3	68.4±3.0
Std. Schedule	92.3±2.8	88.1±3.0	74.3±2.7
Baseline	80.4±4.0	54.2±4.5	54.4±3.4

Quantitatively, LINA significantly improves temporal physical alignment, boosting the success rate from a baseline of **29.5%** to **58.0%**. Qualitatively, consider the Winoground prompt: "a person is *close to* the water and *in* the sand." The baseline model suffers from spatial confusion, incorrectly placing the person *inside* the water. In contrast, LINA enforces the correct causal structure, generating a temporally coherent sequence where the subject interacts with the sand while remaining adjacent to the water. This confirms that LINA's adaptive interventions transfer effectively to the temporal domain. See supplementary materials for more details.

5.4 Ablation Study

We conduct an ablation study to validate LINA's core components using SD-3.5-large. As shown in Tab. 4, we evaluate on PAP-Optics (Opt.), PAP-Density (Dens.), and Winoground (Wino.).

Effectiveness of Intervention Mechanisms Removing either the token-level γ_1 intervention (Eq. 1) or the latent-level γ_2 intervention (Eq. 2) leads to significant performance degradation (rows 2-3). Notably, removing γ_1 ("w/o γ_1") severely impacts OOD performance (Wino.) and complex physical reasoning (Dens.). Removing γ_2 ("w/o γ_2") broadly undermines physical alignment across both PAP subsets. This confirms that both interventions are complementary and essential.

Necessity of Adaptive Intervention Module (AIM) We test the necessity of our adaptive AIM module by replacing it with a "Fixed γ" baseline, which applies a static, averaged intervention strength to all prompts. The sharp performance drop (row 4) validates our hypothesis that a prompt-specific, adaptive guidance is critical.

Effectiveness of Causality-Aware Schedule Finally, reverting to a "Std. Schedule" (row 5) from our reallocated, front-loaded schedule also degrades performance. This confirms our finding (Sec. 3.4) that reallocating computation to the critical, early structure-formation phase is essential.

Analysis of Evaluator Robustness The AIM training labels (Sec. 4.2) are obtained via automated search with **deterministic rule-based verification** of geometric relationships (containment, reflection bounds, *etc.*). While we use Qwen2.5-VL-72B to extract bounding boxes and key points, the final success labels are decided by threshold-based rules. An inter-evaluator study across Qwen2.5-VL, GPT-5 [32], Gemini 2.5 Pro [8], and Gemini 3.1 Pro [13] ($n = 500$) confirms this robustness: raw MLLM judgments agree substantially ($\alpha = 0.92$), reaching near-perfect after geometric consistency checks ($\alpha = 0.97$). A human evaluation on a 200-image subset (4 annotators *vs.* post-rule verification) yields the same $\alpha = 0.97$. A side-by-side perceptual comparison on the same subset (no tie) shows LINA achieves 61% winning rate *vs.* Baseline and 67% *vs.* LoRA. All evaluation criteria and prompts are provided in the supplementary materials.

6 Related Work

Prompt Engineering and Sampling Guidance. This path seeks a pre-distorted input X' (or a modified $G_{X'}$) such that the flawed dynamics $\mathcal{F}$ produce the correct output: $\mathcal{F}(G_{X'}) \approx G^*_X$. Methods include training LLMs to rewrite naive prompts [21, 51], or parsing prompts into structured layouts (*e.g.*, bounding boxes) to guide attention [23, 25, 35]. Others use LLM reasoning to generate negative guidance [15]. While effective for spatial layout, these methods fundamentally fail to address implicit physical and causal laws, as they only alter the input X without correcting the flawed dynamics of $\mathcal{F}$.

MLLM-in-the-loop Refinement. This path uses an external MLLM as a post-hoc corrector, M, which is applied iteratively to guide the generation process, achieving $M(\mathcal{F}(G^*_X)) \approx G^*_X$. This includes "outer-loop" systems that refine prompts between generations [50], and "inner-loop" methods that correct intermediate latents during denoising [20, 28]. These approaches suffer from prohibitive inference costs and the identification-correction gap we demonstrate in our experiments.

Finetuning and Adapter-based Methods. This path attempts to correct the generation by modifying the DM's weights from θ to θ'. Methods include preference alignment [52, 54], lightweight adapters [37, 48], and reinforcement learning [27]. These methods fundamentally assume a *knowledge deficit* in θ that must be retrained, without explicitly modeling the causal structure of the generative process.

7 Conclusion

In this work, we introduce LINA, a novel framework for adaptive causal intervention in diffusion models. Based on diagnostic insights from our Causal Scene Graph (CSG), LINA learns to perform prompt-specific interventions, significantly enhancing the physical alignment and counterfactual generation capabilities of DMs. Without relying on MLLMs during inference or requiring DM retraining, LINA demonstrates strong generalizability across both image and video DMs.

Our work sets a foundation for developing generative models that can function as robust world simulators, understanding and rendering complex causal structures.

Acknowledgements

We thank all the anonymous reviewers and area chair for their valuable feedback throughout the review process. This work is supported by the Shanghai Artificial Intelligence Laboratory.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Black Forest Labs: Flux.1 krea [dev]: An 'opinionated' text-to-image model (2025), https://bfl.ai/announcements/flux-1-krea-dev, accessed: 2026-06-26
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Bradley, A., Nakkiran, P.: Classifier-free guidance is a predictor-corrector. arXiv preprint arXiv:2408.09000 (2024)
5. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: A comprehensive survey of scene graphs: Generation and application. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1–26 (2021)
6. Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., Chen, X.: F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. arXiv preprint arXiv:2410.06885 (2024)
7. Chung, H., Kim, J., Park, G.Y., Nam, H., Ye, J.C.: Cfg++: Manifold-constrained classifier free guidance for diffusion models. arXiv preprint arXiv:2406.08070 (2024)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al.: Understanding world or predicting future? a comprehensive survey of world models. ACM Computing Surveys **58**(3), 1–38 (2025)
10. Diwan, A., Berry, L., Choi, E., Harwath, D., Mahowald, K.: Why is winoground hard? investigating failures in visuolinguistic compositionality. arXiv preprint arXiv:2211.00768 (2022)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
12. Google: Image generation (2025), https://ai.google.dev/gemini-api/docs/image-generation/, accessed: 2026-06-26
13. Google: Gemini 3.1 pro (2026), https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, accessed: 2026-06-26
14. Han, Y., Han, A., Huang, W., Lu, C., Zou, D.: Can diffusion models learn hidden inter-feature rules behind images? International Conference on Machine Learning (2025)
15. Hao, Y., Chen, C., Mian, A.S., Xu, C., Liu, D.: Enhancing physical plausibility in video generation by reasoning the implausibility. arXiv preprint arXiv:2509.24702 (2025)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

18. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)
19. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303
20. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8362–8371 (2024)
21. Ji, Y., Zhang, J., Wu, J., Zhang, S., Chen, S., Ge, C., Sun, P., Chen, W., Shao, W., Xiao, X., et al.: Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18725–18735 (2025)
22. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
23. Li, P., Yu, P., Liu, Z., He, W., Pan, X., Rao, X., Wei, T., Chen, W.: Ldgen: Enhancing text-to-image synthesis via large language model-driven language representation. arXiv preprint arXiv:2502.18302 (2025)
24. Li, X., Thickstun, J., Gulrajani, I., Liang, P.S., Hashimoto, T.B.: Diffusion-lm improves controllable text generation. Advances in neural information processing systems **35**, 4328–4343 (2022)
25. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023)
26. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503 (2023)
27. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. Advances in neural information processing systems **38**, 40783–40818 (2026)
28. Lv, Z., Chen, J., Tian, Q., Yin, K., Zhang, S., Wu, F.: Multimodal llm-guided semantic correction in text-to-image diffusion. arXiv preprint arXiv:2505.20053 (2025)
29. Meng, F., Shao, W., Luo, L., Wang, Y., Chen, Y., Lu, Q., Yang, Y., Yang, T., Zhang, K., Qiao, Y., et al.: Phybench: A physical commonsense benchmark for evaluating text-to-image models. arXiv preprint arXiv:2406.11802 (2024)
30. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models. arXiv preprint arXiv:2502.09992 (2025)
31. OpenAI: Gpt image api (2025), https://help.openai.com/en/articles/11128753-gpt-image-api, accessed: 2026-06-26
32. OpenAI: Introducing gpt-5 (2025), https://openai.com/index/introducing-gpt-5/, accessed: 2026-06-26
33. Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic books (2018)
34. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press (2017)
35. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7932–7942 (2024)

36. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Liu, S., Berman, W., Xu, Y., Wolf, T.: Diffusers: State-of-the-art diffusion models, `https://github.com/huggingface/diffusers`, accessed: 2026-06-26
37. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems **36**, 79320–79362 (2023)
38. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
39. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
40. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
42. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
43. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
44. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
46. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
47. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
48. Tong, L., Liu, Z., Lu, C., Oglic, D., Diethe, T., Teare, P., Tsaftaris, S.A., Jin, C.: Causal-adapter: Taming text-to-image diffusion for faithful counterfactual generation. arXiv preprint arXiv:2509.24798 (2025)
49. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
50. Wan, X., Zhou, H., Sun, R., Nakhost, H., Jiang, K., Sinha, R., Arık, S.Ö.: Maestro: Self-improving text-to-image generation via agent orchestration. arXiv preprint arXiv:2509.10704 (2025)
51. Wang, L., Xing, X., Cheng, Y., Zhao, Z., Li, D., Hang, T., Tao, J., Wang, Q., Li, R., Chen, C., et al.: Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. arXiv preprint arXiv:2509.04545 (2025)

52. Wu, S., Si, X., Xing, C., Wang, J., Jin, G., Cheng, G., Zhang, L., Huang, X.: Preference alignment on diffusion model: A comprehensive survey for image generation and editing. arXiv preprint arXiv:2502.07829 (2025)
53. Yu, S., Lu, C.: ADAM: an embodied causal agent in open-world environments. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), https://openreview.net/forum?id=Ouu3HnIVBc, accessed: 2026-06-26
54. Zhao, H., Chen, H., Guo, Y., Winata, G.I., Ou, T., Huang, Z., Yao, D.D., Tang, W.: Fine-tuning diffusion generative models via rich preference optimization. arXiv preprint arXiv:2503.11720 (2025)