Dual Inversion for Text-to-Image Diffusion Models: From Both Prompt and Noise Perspectives

Xiaolong Liu*
University of Technology Sydney
Sydney, Australia
xiaolongliu2000@outlook.com

Junjian Li*
Geely
Hangzhou, China
ljj1016cc@163.com

Yuan Xiao
City University of Hong Kong
Hong Kong, China
yuan.xiao.cs@outlook.com

Jiaqi Deng
University of Technology Sydney
Sydney, Australia
jiaqi.deng@student.uts.edu.au

Dayong Ye
City University of Macau
Macau, China
dayongye@outlook.com

Tianqing Zhu
City University of Macau
Macau, China
tqzhu@cityu.edu.mo

Huan Huo†
University of Technology Sydney
Sydney, Australia
huan.huo@uts.edu.au

arXiv:2607.26735v1 [cs.CV] 29 Jul 2026

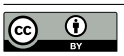
Abstract

Prompt inversion, as a typical reverse engineering technique, enables text-to-image (T2I) diffusion models to generate the desired target images without extensive prompt engineering. However, existing prompt inversion methods suffer from significant limitations: (1) gradient-based methods are unstable and uninterpretable, often resulting in generated images with severe artifacts; (2) gradient-free methods yield human-readable prompts but still fail to preserve visual fidelity due to the lack of fine-grained detail alignment. We contend that the limitations stem from treating prompt inversion as a sufficient condition for reverse engineering, ignoring the critical role of the latent noise that encodes structural information. Consequently, we propose *Dualin* (**Dual in**version), a two-stage method that jointly recovers both the semantic prompt and latent noise of the target image. In the first stage, we integrate vision-language model, CLIP and large language model to invert a faithful, human-interpretable hard prompt. In the second stage, unconditional DDIM inversion reconstructs the exact latent noise of the target image, guaranteeing the consistency at the structural information level. Theoretically, we prove that the inverted noise enables flexible image editing without re-optimization. Extensive experiments on diverse datasets demonstrate that *Dualin* simultaneously generates high-quality inverted prompts and achieves state-of-the-art image fidelity. Additionally, *Dualin* can establish a robust foundation for the precise and controllable image editing.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

*Both authors contributed equally to this research.
†Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil.
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835579

Keywords

Text-to-Image Diffusion Model, Prompt Inversion, Noise Inversion

ACM Reference Format:
Xiaolong Liu, Junjian Li, Yuan Xiao, Jiaqi Deng, Dayong Ye, Tianqing Zhu, and Huan Huo. 2026. Dual Inversion for Text-to-Image Diffusion Models: From Both Prompt and Noise Perspectives. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835579

1 Introduction

Reverse engineering [2, 21, 27, 31–33] of text-to-image (T2I) diffusion models [10, 40, 43] has emerged as a pivotal research direction, aiming to decipher the black-box generation process and achieve controllable image generation and editing [35, 38]. As a typical reverse engineering technique, prompt inversion [9, 12] seeks to recover the specific text descriptions that faithfully reproduce a given target image by mapping visual content back into the discrete textual space [16, 20, 23, 37, 42]. Moreover, prompt inversion serves as a cornerstone for various downstream applications, including semantic editing [14], model safety auditing [41], and automated prompt discovery [25]. Unlike manual prompt inference relying on trial-and-error, prompt inversion can automatically discover the prompts aligned with the target visual distribution [18, 19, 39].

However, simultaneously achieving high prompt interpretability and image fidelity remains a persistent challenge for existing inversion methods. Gradient-based prompt inversion methods [20, 37] typically rely on gradient-based optimization or iterative search in the text embedding space, guided by similarity metrics such as CLIP cosine similarity. While these methods demonstrate promising capability in reconstructing image-level semantics, they often suffer from issues including optimization instability, limited interpretability, and low image fidelity (Figure 1(a)). In an attempt to address the aforementioned limitations of gradient-based prompt inversion, recent works [9, 12] have investigated gradient-free prompt inversion methods that operate directly in the discrete textual space.

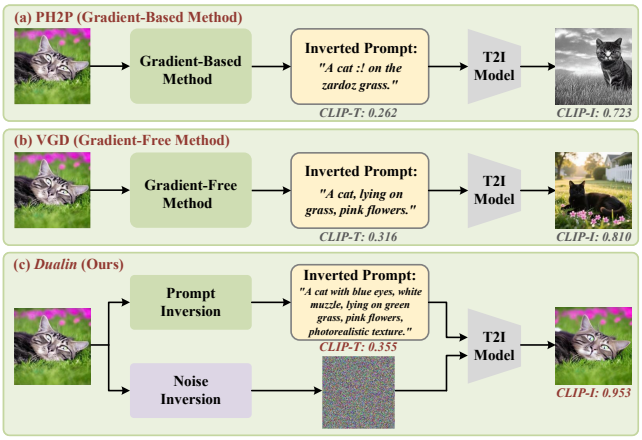


Figure 1: The comparison between the proposed *Dualin* and the existing prompt inversion methods, where "CLIP-T" refers to the CLIP cosine similarity between the inverted prompt and the target image and "CLIP-I" denotes the CLIP cosine similarity between the generated and target image.

By leveraging the black-box optimization, these methods generate prompts that are more interpretable and semantically coherent. Despite the improved quality of the inverted prompts, the CLIP cosine similarity between the generated and target images is still limited due to the lack of fine-grained visual details (Figure 1(b)).

To address the above limitations, we rethink the reverse engineering for text-to-image diffusion models from perspectives of prompt and noise inversion. We argue that prompt inversion alone is inherently insufficient for reliable reverse engineering of T2I diffusion models. Instead, high-quality reconstruction requires jointly accounting for both the semantic control exerted by the prompt and the structural information encoded in the latent noise. In the process, the inverted noise serves as a structural encoding of the original image, enabling the decoupling of spatial layout from semantic content, thereby effectively enhancing the fidelity of the reconstruction. Therefore, we propose **Dual in**version called *Dualin* (Figure 1(c)), a two-stage method that decouples the reverse engineering process into two complementary stages: prompt inversion and noise inversion. Specifically, we first recover a semantically faithful and human-readable hard prompt via a gradient-free inversion strategy. In this stage, we utilize the image semantic extraction capability of vision-language model (VLM), the text-image matching capability of CLIP, and the text processing capability of large language model (LLM) to achieve accurate prompt inversion. Then, we perform unconditional DDIM inversion to reconstruct the latent noise corresponding to the target image, thereby stabilizing the generation process and enabling faithful image reconstruction. By jointly inverting prompt and noise, *Dualin* effectively bridges the gap between prompt interpretability and reconstruction fidelity, achieving substantially higher CLIP cosine similarity at both the text–image and image–image levels, while naturally supporting precise and controllable image editing. Our main contributions are as follows:

- We demonstrate the importance of noise for reverse engineering of T2I diffusion models, as high-fidelity image reconstruction cannot be reliably achieved by prompt inversion alone.
- We propose *Dualin* decoupling the reverse engineering process into both prompt and noise inversion, and theoretically prove *Dualin* enables controllable image editing through unconditional DDIM noise inversion.
- We conduct extensive experiments on diverse datasets, demonstrating that *Dualin* achieves state-of-the-art performance compared with the existing methods of prompt quality and image fidelity.
- We further validate that *Dualin* can be applied to precise and controllable image editing tasks, such as subject replacement, background editing, and style transfer.

2 Related Work

2.1 Gradient-Based Prompt Inversion

Following the prompt-tuning paradigm [15], textual prompt inversion methods [8, 11] augment the vocabulary of model with dedicated tokens to embed visual objects from input images. However, soft prompts are continuous, high-dimensional vectors lacking human interpretability [5, 29]. Therefore, it is imperative to explore hard prompt inversion. To mitigate the limitations of textual inversion, gradient-based prompt inversion techniques [13, 20, 37] has been proposed. PEZ [37] robustly optimize hard prompts through efficient gradient-based optimization and automatically generates hard text-based prompts for both text-to-image and text-to-text applications. PH2P [20] utilize a delayed projection scheme to optimize for prompts representative of the vocabulary space in the diffusion model and identify semantically meaningful prompts for target images. EDITOR [13] initializes embeddings using a pre-trained image captioning model and converts them to texts using an embedding-to-text model. However, these gradient-based prompt inversion methods optimize continuous pseudo-token embeddings rather than directly operating in the discrete vocabulary space. Consequently, the optimized embeddings often lie in regions that do not correspond to valid tokens.

2.2 Gradient-Free Prompt Inversion

To address the aforementioned problem, the gradient-free approaches [9, 12, 25] have recently been proposed. In prompt inversion, BLIP-2 [12] functions as a powerful vision–language encoder that provides high-quality, image-grounded semantic representations to guide the optimization of the textual prompt. VGD [9] leverages large language models and CLIP-based guidance to generate coherent and semantically aligned prompts. ARPO [25] iteratively refines an initial prompt through an imitative gradient-based optimization process. In each iteration, the current prompt generates a recreated image, textual gradients are estimated to reduce its discrepancy with the reference image, and a greedy search selects the update that maximizes CLIP similarity. Although gradient-free methods can generate prompts that are highly interpretable and easy to read, the images synthesized from these inverted prompts often exhibit substantial deviation from the target image, resulting in suboptimal CLIP cosine similarity.

3 Preliminary

3.1 Diffusion Generation and Inversion

Existing text-to-image generation models primarily utilize the framework of latent diffusion models. Latent diffusion models transform the noise $x_T \sim \mathcal{N}(0,1)$ into the latent representation x_0 through T steps of denoising iteratively. The latent x_0 is then decoded by VAE-decoder $\mathcal{D}$ to generate the realistic image $I \sim q(\mathbf{I})$. Moreover, the VAE-encoder $\mathcal{E}$ maps I into the latent x_0 at the beginning of the forward process. Formally, the forward process diffuses the sample x_0 by adding random noise:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I}), \tag{1}$$

where $\{\beta_t\}_{t=1}^T$ is the time-dependent variance hyperparameter. Also, x_t can be obtained from x_0:

$$x_t = \sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}}\epsilon, \tag{2}$$

where $\bar{\alpha} = \prod_0^T (1-\beta_t)$ and $\epsilon \sim \mathcal{N}(0,1)$. Then, a network ϵ_θ is trained to predict the added noise in each step through minimizing the following objective:

$$\min_\theta \mathbb{E}_{t\sim Uniform(1,T),\epsilon\sim\mathcal{N}(0,\mathrm{I})} \|\epsilon - \epsilon_\theta(x_0, t, \psi(p))\|_2^2, \tag{3}$$

where x_t is the noise latent at t and $\psi(p)$ is the embedding of text prompt p.

DDIM enables deterministic sampling through an ordinary differential equation (ODE) solver by reformulating the original diffusion process as a non-Markovian chain. DDIM computes x_{t-1} from x_t by predicting the estimation of x_0 and the direction pointing to x_t:

$$x_0' = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, \psi(p))}{\sqrt{\bar{\alpha}_t}}, \tag{4}$$

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}x_0' + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon(x_t, t, \psi(p)). \tag{5}$$

The deterministic properties of DDIM can reconstruct the noise from the latent x_0:

$$\hat{x}_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, \psi(p)). \tag{6}$$

3.2 Motivation: The Impact of Sampling Noise

In T2I diffusion models, the sampling noise serves as a pivotal mechanism for enabling diversity in image generation [7, 34, 45]. The stochastic nature of noise ensures that, even under identical textual prompts, each generated image exhibits significant variation, as can be seen in the first row of Figure 2. However, this poses a challenge to reverse engineering of T2I diffusion models. In the second and third rows of Figure 2, although the text prompt obtained through inversion is semantically consistent with the target image, differences in noise sampling can lead to considerable variability in CLIP-I between the generated and target images. This may not only result in significant visual-style discrepancies between the generated and target ones, but also introduce deviations in fine-grained details and overall structure. This uncertainty introduced by noise has a significant impact on the accuracy and stability of T2I reverse engineering. Therefore, to achieve a more accurate reconstruction effect, the reverse engineering of noise is also indispensable.

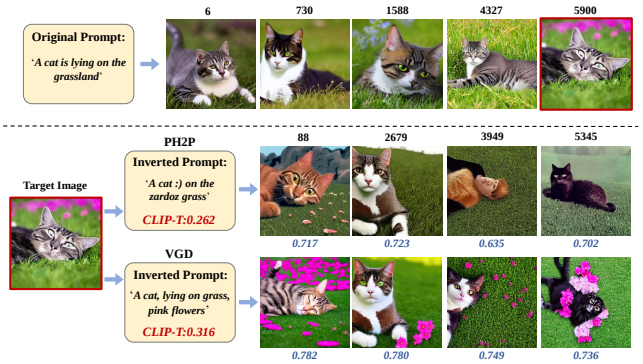


Figure 2: The impact of different sampling noises on the generation results of Stable Diffusion V1.5, where black numbers indicate the random seeds, while blue numbers indicate the results of CLIP-I.

4 Methodology

Previous studies focus solely on optimizing prompts while neglecting the impact of noise. Motivated by above analysis, we rethink the reverse engineering for T2I diffusion models and propose *Dualin* with the dual objectives of obtaining an accurate text prompt and a more matching noise corresponding to the target image. The dual inversion not only ensures interpretable inverted prompt and high image fidelity but also provides a robust foundation for precise and controllable image editing. The framework of *Dualin* along with its verification and application is illustrated as Figure 3.

4.1 Stage I: Prompt Inversion

The goal of prompt inversion aims to invert a faithful, human-interpretable hard prompt P^* that can most effectively reflect the semantics of the target image I. We decompose the prompt inversion method into three steps: **Semantic Content Extraction**, **Zero-Shot Concept Retrieval** and **Prompt Generation and Selection**.

Semantic Content Extraction. To obtain an accurate and interpretable initialization, we leverage a pre-trained VLM to extract the semantic content P' from the target image I. By inputting the target image I and the instruction prompt P_{VLM} to the VLM, we can obtain the semantic content of target image:

$$P' = V(I, P_{\text{VLM}}), \tag{7}$$

where V indicates the vision-language model adopted in this paper. To prevent the hallucination of artistic styles (e.g., "oil painting" vs. "photograph") at this stage, the instruction prompt P_{VLM} explicitly constrains the model to focus solely on visual content.

Zero-Shot Concept Retrieval. While VLMs excel at scene description, they often struggle to precisely identify specific aesthetic attributes (e.g., rendering engines, lighting types, specific art styles) that are crucial for generative models. To address this, we introduce a CLIP-based Interrogator module. We define a "**Concept Bank**":

$$C = \{C_{\text{medium}}, C_{\text{style}}, C_{\text{lighting}}, C_{\text{quality}}\}, \tag{8}$$

where each category contains a pre-defined vocabulary of aesthetic tags (e.g., "Cyberpunk", "Volumetric Lighting"). We utilize the CLIP image encoder E_I and text encoder E_T to project the image and

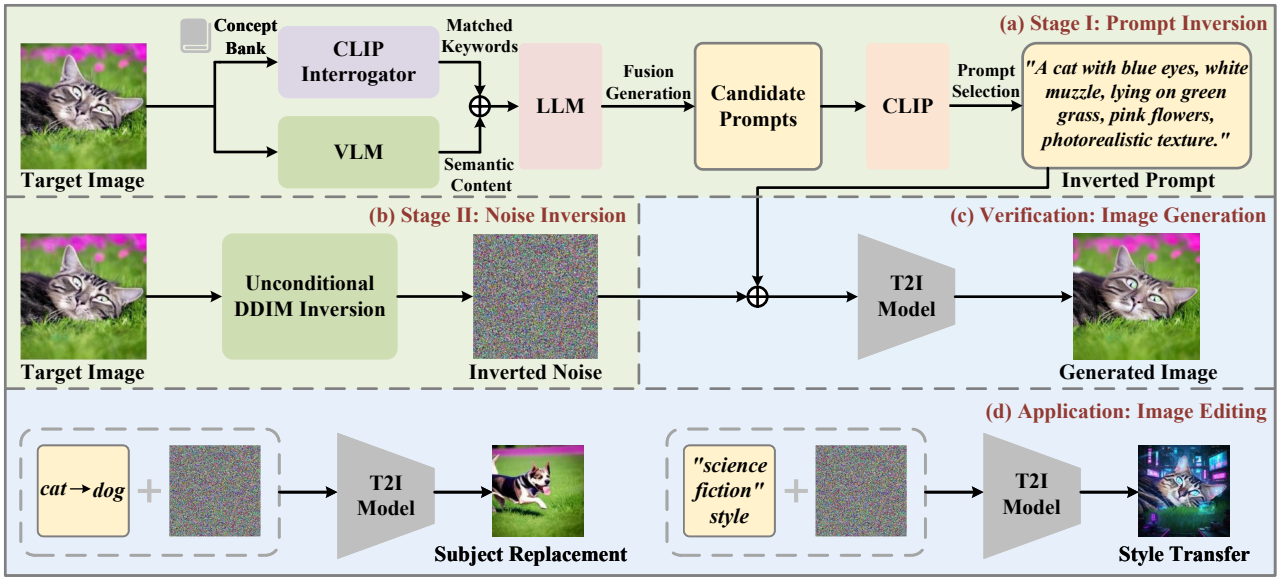


Figure 3: The framework of *Dualin* along with its verification and application. (a) Prompt inversion: inverting the prompt of target image using the image semantic extraction capability of VLM, the text-image matching capability of CLIP, and the text processing capability of LLM. (b) Noise inversion: performing unconditional DDIM inversion to reconstruct the latent noise corresponding to target image. (c) Image generation: combining the inverted prompt and noise to generate a reconstructed image similar to the target image. (d) Image editing: utilizing the inverted noise and different modified prompts based on the inverted prompt to achieve controllable image editing.

concept tags into a shared latent space. For each concept category set $c \subseteq C$, we first define the CLIP cosine similarity score $S(I, p)$ between the target image I and a candidate tag $p \in c$:

$$S(I,p) = \frac{E_I(I) \cdot E_T(p)}{\|E_I(I)\| \|E_T(p)\|}. \tag{9}$$

We then retrieve the subset of keywords K by aggregating the top-k candidates from each category that maximize this similarity score:

$$K = \bigcup_{c \subseteq C} \underset{p \in c}{\text{argtopk}} \left(S(I,p)\right), \tag{10}$$

where argtopk denotes the selection of the k tokens $p \in c$ yielding the highest similarity values. This retrieval-based approach ensures that the extracted stylistic keywords are grounded in the visual feature space of CLIP, effectively mitigating the bias of the language model.

Prompt Generation and Selection. We fuse the semantic content P' and the stylistic keywords K into a coherent prompt. The LLM, as the synthesizer Φ, is prompted to act as a "Prompt Engineer", merging the disparate inputs into a set of candidate prompts $\mathcal{P} = \{P_1, P_2, ..., P_N\}$. The model is instructed to perform stylistic transfer, applying the attributes in K to the subject described in P':

$$\mathcal{P} = \Phi(\{P', K\}, P_{\text{LLM}}), \tag{11}$$

where P_{LLM} is the system prompt used to induce LLM.

To obtain the optimal prompt, we employ a discriminator-guided selection mechanism. We compute the CLIP similarity score $S(I, P_i)$ for each candidate prompt P_i against the target image I. We select the highest one as the final output prompt P^* as follows:

$$P^* = \underset{P_i \in \mathcal{P}}{\text{argmax}} \left(S(I, P_i)\right). \tag{12}$$

4.2 Stage II: Noise Inversion

Section 3.2 demonstrates the impact of noise for the T2I reverse engineering. In the following part, we discuss the noise inversion method of the target image in detail.

Building upon the inverted prompt obtained in Stage I, we aim to recover the latent noise of the target image I. In contrast to the standard DDIM sampling pipeline, inversion seeks to reverse-engineering the latent noise that underlies a target image within the diffusion trajectory, such that forward generation initialized from this noise reconstructs the target image as illustrated in Figure 3(b). Specially, we solve the following reverse-time recurrence to obtain the noise through unconditional inversion:

$$\begin{aligned} x_{t+1} =& \sqrt{\beta_{t+1}} \left(\frac{x_t - \sqrt{1-\beta_t}\, \epsilon_\theta(x_t, t, \phi)}{\sqrt{\beta_t}} \right) \\ &+ \sqrt{1-\beta_{t+1}}\, \epsilon_\theta(x_t, t, \phi), \end{aligned} \tag{13}$$

where ϕ represents the unconditional input. The inversion process starts from $x_0 = \mathcal{E}(I)$ and $\mathcal{E}(\cdot)$ is VAE-encoder. Notably, the unconditional inversion satisfies the following theorem.

Theorem 4.1 (Decoupling Property). *Let x_T^* be obtained through unconditional inversion of x_0^*, while $G_{0:T}$ and $G_{0:T}^{\phi}$ represent conditional and unconditional generation processes, respectively. Then, for any prompt P, the following two statements hold:*

- *The value of x_T^* is independent of P, i.e., $x_T^* = \mathcal{I}^\phi(x_0^*)$, $\frac{\partial x_T^*}{\partial P} = 0$, where $\mathcal{I}^\phi$ is the mapping of the unconditional inversion;*
- *When using any condition P to generate an image from x_T^*, i.e., $x_0(P) = G_{0:T}(x_T^*, P)$, the generated result can be decomposed into:*

$$x_0(P) = x_0^* + \triangle_{uncond} + \triangle_{cond}(P),$$

where $\triangle_{uncond} = G_{0:T}^\phi(x_T^*) - x_0^*$ denotes the unconditional reconstruction error independent of P, whereas $\triangle_{cond}(P) = G_{0:T}(x_T^*, P) - G_{0:T}^\phi(x_T^*)$ represents the offset introduced by the conditional prompt P.

4.3 Image Generation and Editing

Image Generation. To validate our inversion quality, we reconstruct the target image via $\hat{I} = \mathcal{G}(x_T^*, P^*)$. As shown in Figure 3(c), jointly utilizing the inverted prompt and noise ensures both high semantic alignment and structural fidelity, effectively bridging the gap left by prompt-only methods.

Image Editing. According to Theorem 4.1, unconditional inversion ensures that the noise encodes the image content and prompt controls the generation direction. Consequently, the following corollary holds:

Corollary 4.2 (Editable Property). *Perform an unconditional inversion on the original image to obtain x_T^*, then fix it and generate various edited results using different modified prompts P based on the inverted prompt P^* without requiring reinversion.*

The proposed *Dualin* can further enables controllable editing. Leveraging Corollary 4.2, we fix the unconditionally inverted noise x_T^* to preserve spatial layout while modifying specific tokens in P^* for semantic manipulation. As illustrated in Figure 3(d), this allows for precise operations like subject replacement and style transfer without re-optimization.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on three datasets with diverse distributions: MS COCO [17], LAION [28], and DiffusionDB [36]. Notably, we utilize the provided prompts in DiffusionDB as the sources to generate the target images.

T2I Diffusion Models. To ensure the fairness and diversity of the generated images, we use three different T2I diffusion models to generate images based on the inverted prompts and noises: Stable Diffusion V1.5 (SD-V1.5) [26], Stable Diffusion XL (SDXL) [22], and Pixart-α [4].

Settings of Prompt Inversion. In this first stage of our method, we adopt Qwen3-VL-7B-Instruct [3] to extract semantic content, and use Llama-3-8B-Instruct [6] to generate candidate prompts set. In addition, we utilize CLIP-ViT-Large Patch14 [24] to match target images and keywords of Concept Bank.

Baselines We compare the proposed *Dualin* with with state-of-the-art prompt inversion methods: PEZ [37], PH2P [20], BLIP [12], and VGD [9].

Evaluation Metrics. For prompt fidelity, we measure the cosine similarity between the CLIP [24] embeddings of the inverted prompts and the target images, denoted as CLIP-T. For image fidelity, we measure the cosine similarity called CLIP-I, the LPIPS [44] image similarity and SSIM between the generated images and the target images. To further present the performance comparison, we employe GPT-4o [1] and Gemini 3.0 [30] as the expert evaluators. Each model quantifies the alignment between (i) the inverted prompt and (ii) the generated image with respect to the target image, denoted as $\mathcal{S}_p$ and $\mathcal{S}_i$.

5.2 Qualitative Evaluation

We qualitatively compare the inversion results of *Dualin* with baseline methods. Comparative results are shown in Figure 4, focusing on both the interpretability of the inverted prompts and the visual fidelity of the generated images.

Prompt Interpretability. Gradient-based methods (PEZ, PH2P) optimize continuous embeddings, which, upon being projected back into the discrete vocabulary, yield gibberish prompts fraught with disjointed tokens and other artifacts (e.g., "A tiger !!, gtrth big" in Figure 4). While these prompts may trigger specific visual features, they lack human readability and semantic coherence, making them unsuitable for downstream image editing tasks. In contrast, gradient-free methods (BLIP, VGD) generate readable prompts but often oversimplify the scene of target images, missing crucial stylistic nuances. *Dualin* generates prompts that are not only linguistically coherent but also rich in descriptive detail, accurately capturing attributes such as lighting ("cinematic lighting"), texture ("photorealistic"), and style ("cyberpunk").

Image Fidelity. As observed in Figure 4, baseline methods struggle to reproduce the fine-grained visual details of the target images. For instance, in the "tiger mural" example, VGD and BLIP capture the subject but fail to reproduce the specific neon artistic style and the purple wall background. Gradient-based methods like PH2P generate images with severe artifacts and color distortions. *Dualin*, by jointly inverting the prompt and the latent noise, achieves pixel-level consistency. The reconstructed images retain the exact composition, lighting, and texture of the targets, indistinguishable from the originals to the naked eye.

5.3 Quantitative Evaluation

We provide a comprehensive quantitative analysis using standard metrics and LLM-based assessments to validate the effectiveness of *Dualin*. Table 1 summarizes the performance on different datasets. *Dualin* outperforms all baselines by a substantial margin across all metrics and models.

Prompt Alignment. In terms of CLIP-T (measuring CLIP cosine similarity between the inverted prompt and target image), *Dualin* achieves the highest scores (e.g., 0.351 on MS COCO), significantly surpassing the second-best method, VGD (0.282). This indicates that our retrieval-augmented prompt inversion strategy captures the semantic essence of the image more effectively than pure optimization or captioning approaches.

Image Fidelity. For image reconstruction, *Dualin* achieves state-of-the-art results. On the MS COCO dataset using SD-V1.5, we achieve a CLIP-I score of 0.906, compared to 0.804 for VGD and 0.791 for PH2P. Furthermore, our method demonstrates superior structural preservation, evidenced by a dramatic improvement in SSIM (0.627 vs. 0.212 for VGD) and perceptual similarity (LPIPS).

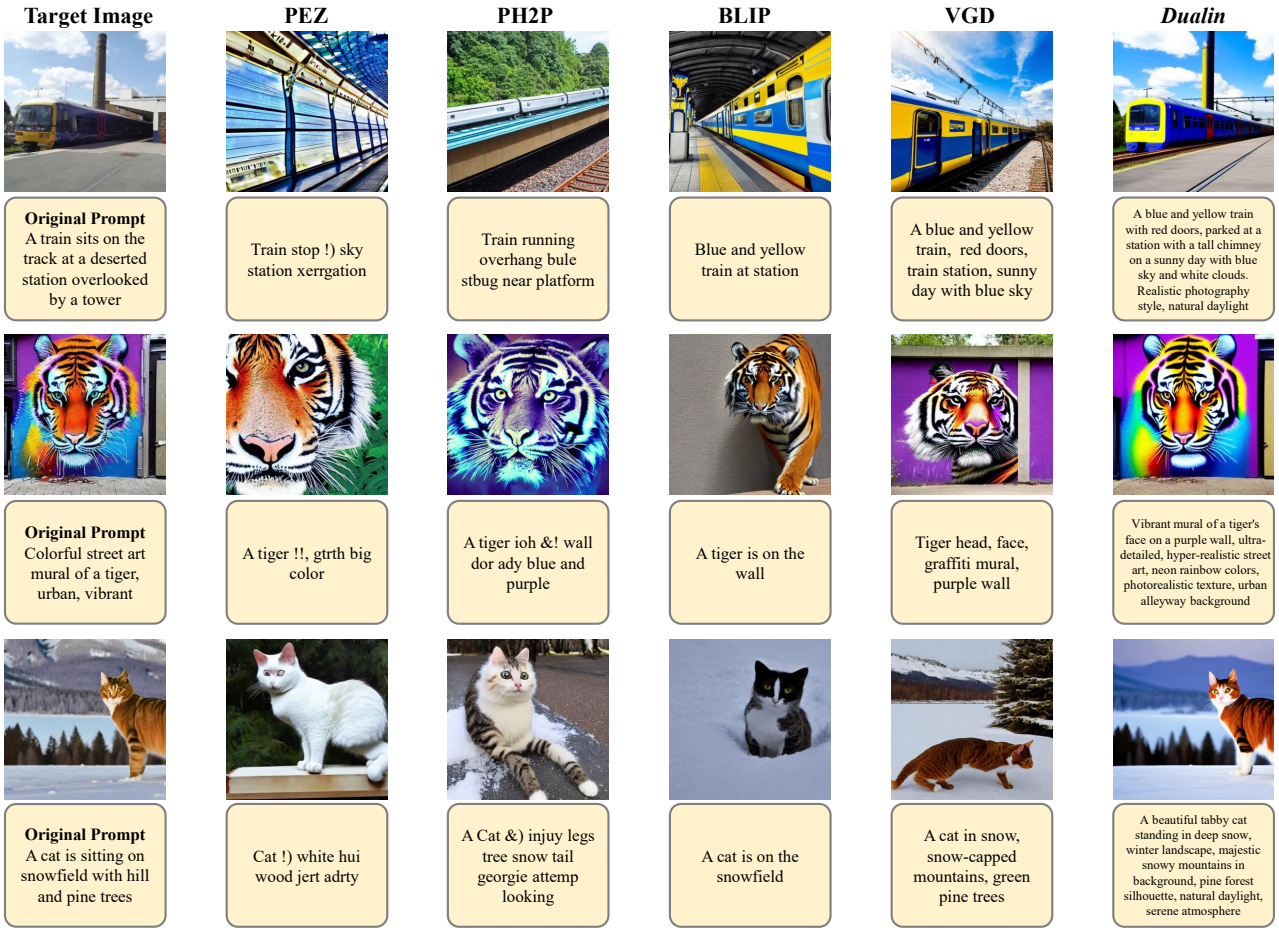


Figure 4: Qualitative Evaluation of the quality of inverted prompts and generated images for the proposed *Dualin* and baseline methods based on Stable Diffusion V1.5.

This trend holds consistent for SDXL and PixArt-α, confirming that decoupling prompt and noise inversion is model-agnostic and robust.

LLM-Based Evaluation. To overcome the limitations of automated metrics in capturing semantic nuances, we employed GPT-4o and Gemini 3.0 as expert evaluators as shown in Table 2. In this experiment, we randomly select 100 target images. For each method, we randomly adopt different diffusion models to reconstruct the images based on inverted prompts. The models evaluated the alignment of the inverted prompt ($\mathcal{S}_p$) and the generated image ($\mathcal{S}_i$) with the target image. *Dualin* achieves the highest ratings from both evaluators. For instance, Gemini 3.0 assigns *Dualin* a prompt score of 8.7 and an image score of 8.2, whereas the strongest baseline (VGD) scores 8.1 and 7.5, respectively.

Inversion Efficiency. We further investigate the efficiency of *Dualin* in comparison with other baseline methods. As shown in Table 3, the time required by our method is significantly less than that of the gradient-based optimization methods. However, it is slightly longer than other gradient-free methods because *Dualin* has an additional noise inversion stage compared to these methods.

5.4 Ablation Study

We conduct ablation studies to verify the contribution of each component in our framework. The results are reported in Table 4 and Figure 6.

Impact of Dual Inversion Stages. We analyze the necessity of the two-stage approach by removing the prompt inversion and noise inversion separately. **(1) w/o Stage I:** Removing the prompt inversion stage (relying solely on noise with a null or random prompt) leads to a catastrophic drop in CLIP-I (0.928 $\rightarrow$ 0.263) and SSIM. This confirms that the semantic direction provided by the prompt is prerequisite for the noise inversion to function correctly and noise alone cannot encode the full image semantics. **(2) w/o Stage II:** Relying only on the inverted prompt (Stage I) without noise inversion results in a CLIP-I score of 0.807. While the prompt captures the semantic content, the lack of exact noise guidance leads to variations in layout and fine details, causing a significant gap compared to the full model (0.928). This validates our motivation that prompt inversion alone is insufficient for precise reverse engineering. As shown in Figure 5, whether in the first stage or the second stage, the image reconstruction effect has significantly declined. After the removal

Table 1: Quantitative evaluation of the quality of inverted prompts and generated images for the proposed *Dualin* and baseline methods on different datasets.

Datasets	Methods	CLIP-T ↑	SD-V1.5			SDXL			Pixart-α		
			CLIP-I ↑	LPIPS ↓	SSIM ↑	CLIP-I ↑	LPIPS ↓	SSIM ↑	CLIP-I ↑	LPIPS ↓	SSIM ↑
MS COCO	PEZ	0.209	0.725	0.496	0.205	0.729	0.489	0.180	0.712	0.488	0.197
	PH2P	0.230	0.791	0.482	0.165	0.812	0.471	0.174	0.790	0.485	0.201
	BLIP	0.264	0.788	0.491	0.244	0.792	0.478	0.230	0.807	0.473	0.218
	VGD	0.282	0.804	0.475	0.212	0.810	0.469	0.248	0.818	0.461	0.198
	***Dualin* (ours)**	**0.351**	**0.906**	**0.402**	**0.627**	**0.909**	**0.398**	**0.659**	**0.917**	**0.405**	**0.653**
LAION	PEZ	0.215	0.746	0.475	0.187	0.756	0.470	0.195	0.741	0.482	0.192
	PH2P	0.246	0.765	0.466	0.183	0.776	0.467	0.178	0.768	0.475	0.190
	BLIP	0.273	0.782	0.453	0.220	0.779	0.469	0.241	0.804	0.467	0.226
	VGD	0.289	0.797	0.445	0.278	0.815	0.457	0.292	0.820	0.459	0.306
	***Dualin* (ours)**	**0.348**	**0.892**	**0.392**	**0.618**	**0.903**	**0.387**	**0.673**	**0.908**	**0.390**	**0.662**
DiffusionDB	PEZ	0.218	0.720	0.491	0.203	0.763	0.482	0.198	0.760	0.481	0.209
	PH2P	0.238	0.763	0.478	0.234	0.779	0.479	0.227	0.765	0.476	0.217
	BLIP	0.269	0.758	0.480	0.298	0.797	0.468	0.275	0.802	0.462	0.280
	VGD	0.297	0.806	0.458	0.303	0.810	0.461	0.324	0.814	0.455	0.319
	***Dualin* (ours)**	**0.353**	**0.928**	**0.371**	**0.643**	**0.937**	**0.358**	**0.665**	**0.931**	**0.352**	**0.658**

Table 2: Evaluation results of GPT-4o and Gemini3.0.

Methods	GPT-4o		Gemini 3.0	
	$\mathcal{S}_p$ ↑	$\mathcal{S}_i$ ↑	$\mathcal{S}_p$ ↑	$\mathcal{S}_i$ ↑
PEZ	3.8	2.4	3.5	2.6
PH2P	4.9	3.7	4.2	3.6
BLIP	7.5	3.5	8.1	3.2
VGD	8.6	6.7	8.1	7.5
***Dualin* (Ours)**	**8.9**	**8.1**	**8.7**	**8.2**

Table 3: Inversion time of *Dualin* and baseline methods.

Methods	PEZ	PH2P	BLIP	VGD	*Dualin* (ours)
Time (sec)	190.68	35293.38	10.80	15.97	18.73

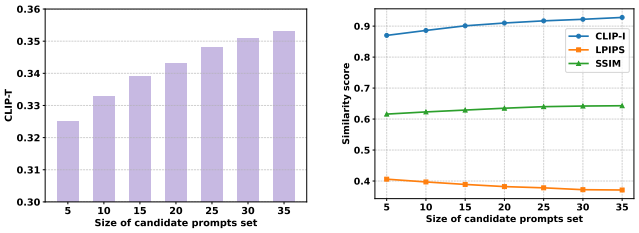
Figure 5: Qualitative comparison with *Dualin*, w/o Stage I and w/o Stage II.

of the first stage, the main body of the images become distorted and damaged; After removing the second stage, the semantics of the generated images cannot be fully consistent with the target images. **(3) DDPM Inversion:** We replace the DDIM inversion in Stage I with the DDPM inversion. DDIM models the diffusion process as a system of ordinary differential equations, while the generation process of DDPM relies on Markov chains and is a form of random sampling. As shown in Table 4, the image similarity metrics obtained through DDIM inversion are significantly worse than the proposed *Dulian*.

Impact of Prompt Inversion Modules. We further dissect Stage I by systematically removing the CLIP retrieval, VLM, or LLM components to understand the individual contribution of each module to the overall prompt generation pipeline. **(1) w/o CLIP:** Removing the CLIP-based style retrieval causes the CLIP-T score to drop from 0.353 to 0.320, indicating that the VLM alone struggles to capture specific aesthetic tags. **(2) w/o VLM:** Excluding the VLM results in a lower CLIP-T (0.307), as the system loses the foundational semantic description of the scene, including object identities, spatial relationships, and contextual details. **(3) w/o LLM:** Removing the LLM fusion step yields the lowest performance (CLIP-T 0.292), demonstrating the critical role of the LLM in synthesizing the heterogeneous signals from VLM and CLIP into a coherent and accurate prompt. Overall, the progressive performance degradation across these ablations validates our design choice of a three-stage collaborative architecture for prompt generation.

Table 4: Quantitative comparison under different ablation settings.

Methods	CLIP-T ↑	CLIP-I ↑	LPIPS ↓	SSIM ↑
Dualin	0.353	0.928	0.371	0.643
w/o Stage I	–	0.739	0.263	0.680
w/o Stage II	0.353	0.807	0.452	0.334
DDPM Inversion	0.353	0.824	0.494	0.367
w/o CLIP	0.320	0.867	0.395	0.619
w/o VLM	0.307	0.840	0.412	0.593
w/o LLM	0.292	0.831	0.417	0.578



(a) CLIP-T (b) CLIP-I, LPIPS, and SSIM

Figure 6: Ablation study on the size of candidate prompts set.

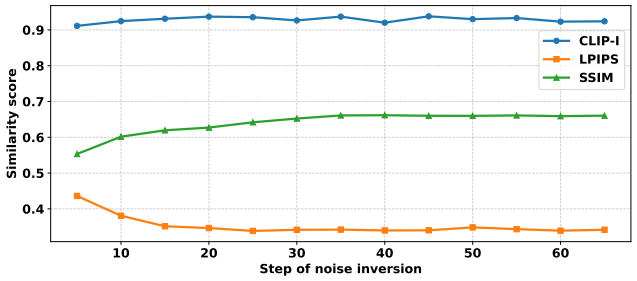


Figure 7: Ablation study on the step of noise inversion.

Hyperparameter Sensitivity. We analyze the impact of the candidate prompts set size and noise inversion step on the experimental results. As for impact of candidate prompts set size, Figure 6 illustrates the impact of the candidate prompts set size in Stage I. As the number of candidate prompts increases, the CLIP-T improves steadily, saturating around $k = 30$. Consequently, the image similarity metrics (CLIP-I, LPIPS, and SSIM) also show a positive correlation with the candidate set size, suggesting that a larger search space allows the discriminator to find a prompt that better aligns with the nuances of target image. In term of impact of noise inversion step, as shown in Figure 7, with the increase in inversion steps, SSIM of the generated images initially exhibits a steady rise before saturating, while the LPIPS drops sharply in the early stages and subsequently stabilizes at a low level. Meanwhile, the CLIP-I score remains consistently high. These observations confirm that the the introduction of noise inversion efficiently recovers structural details and perceptual authenticity while strictly maintaining semantic consistency. However, marginal gains diminish significantly beyond 50 steps. Thus, limiting the inversion process to 35–50 steps ensures the best compromise between computational efficiency and output quality.

5.5 Image Editing Applications

We further explore the application of *Dualin* in controllable image editing, leveraging the Theorem 4.1 (Decoupling Property) to manipulate semantic attributes while preserving structural layout. By fixing the unconditionally inverted noise x_T^* and modifying specific tokens in the inverted prompt P^*, our framework enables precise editing operations as illustrated in Figure 8, including subject replacement (e.g., swapping a corgi for a pig while retaining the original pose), background editing (e.g., relocating an elephant to snowfield without altering the lighting or perspective of subject), and style transfer (e.g., rendering a photorealistic scene as vector art). These results confirm that *Dualin* effectively disentangles high-level semantics from low-level structural information, providing a robust and flexible foundation for diverse downstream editing tasks without requiring extensive re-optimization. In summary, these applications evidence that *Dualin* does not merely memorize the image, but effectively disentangles structural information from semantic information, offering a flexible and robust foundation for controllable downstream image editing tasks.

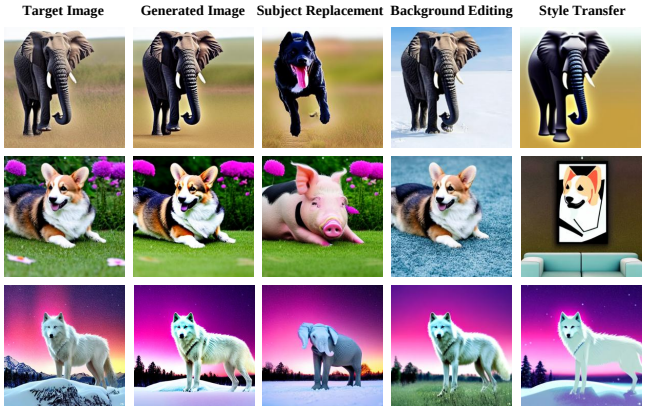


Figure 8: Image editing results of subject replacement, background editing and style transfer for the proposed *Dualin*.

6 Conclusion

In this paper, we rethink the reverse engineering of T2I diffusion models and demonstrate that prompt inversion alone without considering the noise is inherently insufficient. Thus, we propose a two-stage method that includes both prompt and noise inversion called *Dualin*. Specially, we integrate the capabilities of VLM, CLIP and LLM to achieve the precise prompt inversion. Then, we perform unconditional DDIM inversion to obtain the latent noise corresponding to the target image, thereby stabilizing the generation process and enabling the image reconstruction. Furthermore, we theoretically demonstrate that controllability in image editing can be achieved through unconditional noise inversion.

Acknowledgments

This work is supported by the Top Education Group Ltd under Grant PRO26-25541.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Lorenzo Agnolucci, Alberto Baldrati, Alberto Del Bimbo, and Marco Bertini. 2025. iSEARLE: Improving textual inversion for zero-shot composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 11 (2025), 10801–10817.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631* (2025).

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *Proceedings of the Twelfth International Conference on Learning Representations*.

[5] Yiyi Chen, Heather Lent, and Johannes Bjerva. 2024. Text embedding inversion security for multilingual language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 7808–7827.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The LLaMA 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[7] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. 2024. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems* 37 (2024), 125487–125519.

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *Proceedings of the Eleventh International Conference on Learning Representations*.

[9] Donghoon Kim, Minji Bae, Kyuhong Shim, and Byonghyo Shim. 2025. Visually Guided Decoding: Gradient-Free Hard Prompt Inversion with Language Models. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

[10] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22691–22702.

[11] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. 19730–19742.

[13] Mingzhe Li, Gehao Zhang, Zhenting Wang, Shiqing Ma, Siqi Pan, Richard Cartwright, and Juan Zhai. 2025. EDITOR: Effective and Interpretable Prompt Inversion for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2506.03067* (2025).

[14] Ruibin Li, Ruihuang Li, Song Guo, and Lei Zhang. 2024. Source prompt disentangled inversion for boosting image editability with diffusion models. In *European Conference on Computer Vision*. 404–421.

[15] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 4582–4597.

[16] Jiayi Lin, Jiabo Huang, Jian Hu, and Shaogang Gong. 2025. InvSeg: Test-Time Prompt Inversion for Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5245–5253.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*. 740–755.

[18] Zhihua Liu, Amrutha Saseendran, Lei Tong, Xilin He, Fariba Yousefi, Nikolay Burlutskiy, Dino Oglic, Tom Diethe, Philip Alexander Teare, Huiyu Zhou, et al. 2025. Segment anyword: Mask prompt inversion for open-set grounded segmentation. In *Proceedings of the 42nd International Conference on Machine Learning*. 39420–39454.

[19] Zhe Ma, Qingming Li, Xuhong Zhang, Tianyu Du, Ruixiao Lin, Zonghui Wang, Shouling Ji, and Wenzhi Chen. 2025. An Inversion-based Measure of Memorization for Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16959–16969.

[20] Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. 2024. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6808–6817.

[21] John Xavier Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. 2024. Language Model Inversion. In *Proceedings of the Twelfth International Conference on Learning Representations*.

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *Proceedings of the Twelfth International Conference on Learning Representations*.

[23] Wenjie Qu, Yuguang Zhou, Yongji Wu, Tingsong Xiao, Binhang Yuan, Yiming Li, and Jiaheng Zhang. 2025. Prompt inversion attack against collaborative inference of large language models. In *Proceedings of the IEEE Symposium on Security and Privacy*. 1695–1712.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763.

[25] Zhiyao Ren, Yibing Zhan, Baosheng Yu, and Dacheng Tao. 2025. Reverse Prompt: Cracking the Recipe Inside Text-to-Image Generation. *arXiv preprint arXiv:2503.19937* (2025).

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[27] Danila Rukhovich, Elona Dupont, Dimitrios Mallis, Kseniya Cherenkova, Anis Kacem, and Djamila Aouada. 2025. Cad-recode: Reverse engineering cad code from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9801–9811.

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[29] Jeremy Styborski, Mingzhi Lyu, Jiayou Lu, Nupur Kapur, and Adams Wai-Kin Kong. 2025. When and Where do Data Poisons Attack Textual Inversion?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19439–19449.

[30] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[31] Xinpeng Ti, Wentao Ye, Zhifang Zhang, Junbo Zhao, Chang Yao, Lei Feng, and Haobo Wang. 2025. Towards Reverse Engineering of Language Models: A Survey. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025*. 7483–7502.

[32] Corban Villa, Shujaat Mirza, and Christina Pöpper. 2025. Exposing the Guardrails: Reverse-Engineering and Jailbreaking Safety Filters in DALL·E Text-to-Image Pipelines. In *Proceedings of the 34th USENIX Security Symposium*. 897–916.

[33] Kai Wang, Fei Yang, Bogdan Raducanu, and Joost van de Weijer. 2025. Multi-class textual-inversion secretly yields a semantic-agnostic classifier. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4400–4409.

[34] Ruoyu Wang, Huayang Huang, Ye Zhu, Olga Russakovsky, and Yu Wu. 2025. The silent assistant: Noisequery as implicit guidance for goal-driven image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17618–17628.

[35] Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. 2025. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8213–8224.

[36] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 893–911.

[37] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems* 36 (2023), 51008–51025.

[38] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. 2025. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

[39] Bing Xiong, Yue Peng, RanRan Zhang, Fuqiang Chen, JiaYe He, and Wenjian Qin. 2025. Unpaired multi-domain histopathology virtual staining using dual path prompted inversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8780–8787.

[40] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *Proceedings of the 41st International Conference on Machine Learning*. 56704–56721.

[41] Dayong Ye, Tianqing Zhu, Feng He, Bo Liu, Minhui Xue, and Wanlei Zhou. 2025. Cross-Modal Prompt Inversion: Unifying Threats to Text and Image Generative AI Models. In *Proceedings of the 34th USENIX Security Symposium*. 2303–2322.

[42] Collin Zhang, John Xavier Morris, and Vitaly Shmatikov. 2024. Extracting prompts by inverting LLM outputs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 14753–14777.

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595.

[45] Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. 2025. Golden noise for diffusion models: A learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17688–17697.