# *FreeInpaint*: Tuning-free Prompt Alignment and Visual Rationality Enhancement in Image Inpainting

**Chao Gong**[1*], **Dong Li**[2], **Yingwei Pan**[2], **Jingjing Chen**[1,3†], **Ting Yao**[2], **Tao Mei**[2]
[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]HiDream.ai Inc.
[3]Institute of Trustworthy Embodied AI, Fudan University
cgong24@m.fudan.edu.cn, {lidong, pandy}@hidream.ai, chenjingjing@fudan.edu.cn, {tiyao, tmei}@hidream.ai

**Abstract**

Text-guided image inpainting endeavors to generate new content within specified regions of images using textual prompts from users. The primary challenge is to accurately align the inpainted areas with the user-provided prompts while maintaining a high degree of visual fidelity. While existing inpainting methods have produced visually convincing results by leveraging the pre-trained text-to-image diffusion models, they still struggle to uphold both prompt alignment and visual rationality simultaneously. In this work, we introduce *FreeInpaint*, a plug-and-play tuning-free approach that directly optimizes the diffusion latents on the fly during inference to improve the faithfulness of the generated images. Technically, we introduce a prior-guided noise optimization method that steers model attention towards valid inpainting regions by optimizing the initial noise. Furthermore, we meticulously design a composite guidance objective tailored specifically for the inpainting task. This objective efficiently directs the denoising process, enhancing prompt alignment and visual rationality by optimizing intermediate latents at each step. Through extensive experiments involving various inpainting diffusion models and evaluation metrics, we demonstrate the effectiveness and robustness of our proposed *FreeInpaint*.

**Code** — https://github.com/CharlesGong12/FreeInpaint
**Extended version** — https://github.com/CharlesGong12/FreeInpaint/blob/main/FreeInpaint.pdf


Figure 1: Comparisons between our *FreeInpaint* and existing methods. *FreeInpaint* simultaneously enhances prompt alignment and visual rationality. Zoom in for better view.

## 1 Introduction

Recent years have witnessed the success of large text-to-image (T2I) diffusion models (Rombach et al. 2022; Saharia et al. 2022; Cai et al. 2025; Yao et al. 2025) and their exceptional ability to produce high-quality images. Despite the effectiveness of these generative models, users with specific design requirements often face the need for iterative interactions to achieve their desired outcomes. Text-guided image editing (Hertz et al. 2022; Brooks, Holynski, and Efros 2023; Han et al. 2025) stands out as a transformative approach that enriches the image generation process by facilitating interactive refinement. Particularly, text-guided image inpainting empowers users to create new content within specified regions of existing images, guided by textual prompts. This capability opens up various applications, including object removal (Suvorov et al. 2022), photo restoration (Lugmayr et al. 2022), and virtual try-on (Choi et al. 2024; Wan et al. 2024; Li et al. 2025).

One straightforward approach to extending pre-trained T2I diffusion models for text-guided image inpainting is to blend diffused known regions with denoised unknown regions during the reverse diffusion process (Lugmayr et al. 2022; Avrahami, Lischinski, and Fried 2022). However, this technique frequently produces incoherent inpainting results due to limited global scene comprehension and constrained perceptual understanding of mask boundaries. To tackle this issue, recent methods (Rombach et al. 2022; Nichol et al. 2021; Podell et al. 2023) have shifted their focus towards finetuning T2I diffusion models by randomly removing portions of the input image and compelling the model to reconstruct the missing areas based on the corresponding caption. Despite reasonable performance, this training strategy leads to weak image-prompt alignment since randomly selected regions can often be feasibly inpainted purely based on image context, with limited consideration for the provided prompt. To improve prompt alignment, SmartBrush

[*]This work was performed at HiDream.ai.
[†]Corresponding author.

(Xie et al. 2023), ImagenEditor (Wang et al. 2023), and PowerPaint (Zhuang et al. 2024) propose object-aware masking strategies and leverage object descriptions as prompts. Beyond that, BrushNet (Ju et al. 2024) employs a dual-branch architecture to reduce the training burden. While these methods alleviate prompt misalignment, the visual rationality of generated images has been compromised (See Fig. 1).

To achieve both prompt alignment and visual rationality concurrently, our investigation dives into the direct optimization of diffusion latents within the reverse diffusion process. Notably, we argue that two key ingredients should be taken into account during inference. One is ***initial noise optimization*** at the beginning of the denoising process. Recent studies (Wallace et al. 2023; Samuel et al. 2024) highlight that pre-trained T2I diffusion models are sensitive to the initial random noise, with not all randomly sampled noise leading to visually consistent images. We reveal that this sensitivity extends to image inpainting models as well. As shown in Fig. 3, when different noise inputs are fed into the inpainting diffusion model with identical text prompts, a substantial inconsistency is observed in the image-prompt alignment. Therefore, we posit that a well-optimized initial noise can significantly enhance prompt alignment. The other is ***task-specific guidance*** throughout the denoising process. While leveraging the generative capabilities of T2I models proves efficient across various downstream tasks, the generative data distributions derived from diverse training data often deviate from the specific requirements of individual tasks. Specifically, for the task of image inpainting, this discrepancy will degrade the visual rationality of generated images. Training-free guidance has proven effective in aligning generated images with task-specific objectives by utilizing differentiable rewards to optimize diffusion latents during inference (Dhariwal and Nichol 2021; Chefer et al. 2023). Nevertheless, an effective guidance objective tailored for the inpainting task remains underexplored.

By consolidating the idea of optimizing both initial noise and intermediate latents during the denoising process, we novelly present a plug-and-play tuning-free method, *FreeInpaint*, to enhance prompt alignment and visual rationality in image inpainting. On the one hand, we introduce a Prior-Guided Noise Optimization method, dubbed *PriNo*, to direct attention maps towards valid inpainting regions. Through contrastive analysis, we observe an obvious attention direction gap between prompt-aligned and unaligned inpainting results. Recognizing that object presence and low-frequency characteristics are predominantly established in the early stages of denoising (Yang et al. 2023; Park et al. 2023), we propose an attention loss to steer the attention maps of the first denoising step towards concentrating on the masked region, and then optimize the initial noise with respect to the gradients of the attention loss. This approach effectively directs the randomly sampled noise towards a direction that better aligns with the semantic context outlined in the prompt. On the other hand, we exquisitely craft a Decomposed Training-free Guidance objective, dubbed *DeGu*, customized for the inpainting task to guide the denoising process. Specifically, we decompose the conditional distribution of the inpainting process into three distinct objectives: text alignment, visual rationality, and human preference. Such decomposition enables the seamless integration of off-the-shelf differentiable models to comprehensively enhance the inpainting model's perception and generation quality, without any training or fine-tuning requirements.

The main contribution of this work is the proposal of *FreeInpaint*, a tuning-free approach that facilitates the text-guided image inpainting task by directly optimizing diffusion latents during inference. Our method elegantly explores how to optimize the initial noise with the prior knowledge tailored to image inpainting, and how to improve the denoising process with task-specific guidance. By conducting comprehensive experiments with a range of inpainting diffusion models on the EditBench and MSCOCO datasets, we demonstrate the effectiveness and robustness of *FreeInpaint*.

## 2 Related Work

**Image Inpainting** Image inpainting (Bertalmio et al. 2000), the task of filling missing regions in images, has advanced significantly with deep learning. While early methods employing Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) can complete contexts, they often struggle to generate novel objects based on textual descriptions. A recent breakthrough comes with diffusion models (Dhariwal and Nichol 2021). Early blending-based diffusion approaches (Avrahami, Lischinski, and Fried 2022) blend the edited region with the unchanged parts of the image at different noise levels. However, the edited areas often exhibit visible artifacts and lack global consistency with the overall scene. To address this limitation, current research is predominantly focused on fine-tuning pre-trained diffusion U-Nets. For example, Stable Diffusion Inpainting (Rombach et al. 2022) refines a pre-trained U-Net by concatenating the noise latent, mask, and known image latent as input. ControlNet (Zhang, Rao, and Agrawala 2023) and BrushNet (Ju et al. 2024) use dedicated branches to diminish the training load. Despite these advancements, these methods typically rely on randomly scribbled or prompt-irrelevant masks during training, leading to potential neglect of the provided prompt. To enhance prompt alignment, subsequent methods have utilized paired mask-description data (Wang et al. 2023; Xie et al. 2023) or introduced learnable prompts for specific scenarios (Zhuang et al. 2024). In contrast, HD-Painter (Manukyan et al. 2023) adopts a training-free approach. It reweights the self-attention scores according to the alignment of the pixels with the prompt. Nevertheless, its singular focus on prompt alignment often compromises visual rationality, resulting in a decline in overall quality. Our *FreeInpaint* resolves this trade-off by decoupling the objectives: it prioritizes prompt alignment through optimizing initial noise before denoising, and then guides each denoising step to balance between prompt alignment and visual coherence. This approach yields significant improvements in both prompt adherence and overall quality.

**Initial Noise Optimization** Initial noise optimization refines the initial noise latent to enhance the quality of images generated by diffusion models. First introduced by DOODL (Wallace et al. 2023) to improve guidance in T2I models, this technique has since been applied to diverse tasks like rare-concept generation (Samuel et al. 2024), improving synthesis fidelity (Guo et al. 2024; Weimin, Jieke, and Meng 2025), creating 3D motion priors (Karunratanakul et al. 2024), audio generation (Novack et al. 2024), and solving inverse problems (Ben-Hamu et al. 2024). Other works have focused on improving its efficiency (Samuel et al. 2024) or using reward-based optimization (Eyring et al. 2024). Despite these advancements, how to optimize the initial noise specifically for the inpainting task has not yet been studied.

**Post-Training Guidance** In the T2I task, post-training guidance manipulates the denoising process of pre-trained models to steer generation towards specific objectives without retraining. This area is pioneered by Classifier Guidance (Dhariwal and Nichol 2021) and later advanced by CLIP guidance (Nichol et al. 2021) and classifier-free guidance (Ho and Salimans 2022) for improved text-guided sample quality. Subsequent works leverage attention mechanisms to influence object attributes (Chefer et al. 2023; Epstein et al. 2023; Xie and Gong 2024) under complex prompts. However, directly applying T2I guidance to image inpainting is ineffective. Unlike the T2I task, which primarily aligns the entire image with the prompt, image inpainting concentrates on the relevance between the masked local area and the prompt, alongside ensuring consistency between the masked and non-masked regions in the image. Therefore, crafting a tailored guidance objective for image inpainting is essential.

# 3 Preliminaries

**Stable Diffusion Model** Our work builds on the Stable Diffusion (SD) model (Rombach et al. 2022), which handles the T2I task in a latent space managed by a variational autoencoder (VAE) (Kingma and Welling 2013). The VAE's encoder $\mathcal{E}(\cdot)$ maps an image $I \in \mathbb{R}^{h\times w\times 3}$ to a latent $z_0$, and its decoder $\mathcal{D}(\cdot)$ reconstructs the image from the latent. SD's forward diffusion process (Ho, Jain, and Abbeel 2020) gradually adds Gaussian noise to $z_0$ over timesteps $t \in [0, T]$:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \tag{1}$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is noise and $\bar{\alpha}_t$ is a noise schedule that decreases from 1 to 0 as $t$ increases. The reverse process uses a denoising model, $\epsilon_\theta$, to predict and remove the noise at each step, trained with the objective:

$$\mathcal{L} = \mathbb{E}_{z_t,t,\epsilon_t\sim\mathcal{N}(\mathbf{0},\mathbf{1})} \|\epsilon_t - \epsilon_\theta(z_t, t)\|_2^2. \tag{2}$$

For text-guided inpainting, the model is fine-tuned by conditioning on a text prompt embedding $c$, a mask $M \in \mathbb{R}^{h\times w}$, and the masked image latent $z^m = \mathcal{E}(I \odot (1 - M))$ to redraw the specified region. It estimates the noise $\epsilon_\theta(z_t, t, c, z^m, M')$, where $M'$ is a downsampled mask matching the latent's dimensions. These conditions are typically injected either by concatenation with the latent $z_t$ (Rombach et al. 2022; Zhuang et al. 2024) or through a dedicated dual-branch architecture (Ju et al. 2024).

**Diffusion Guidance** Conditional diffusion is often steered by guidance. In score-based diffusion models (Song et al. 2020), the score function is directly related to the noise $\epsilon_t$:

$$\nabla_{z_t} \log p(z_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t. \tag{3}$$

To introduce a condition $c$, Bayes' Theorem (Jøsang 2016) is used to derive a conditional score function:

$$\begin{aligned}\nabla_{z_t} \log p(z_t|c) &\propto \nabla_{z_t} \log p(z_t) + \nabla_{z_t} \log p(c|z_t)\\ &= -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t + \nabla_{z_t} \log p(c|z_t).\end{aligned} \tag{4}$$

This leads to a modified noise prediction with a guidance term (Dhariwal and Nichol 2021):

$$\hat{\epsilon}_t = \epsilon_\theta(z_t, t) - \sqrt{1-\bar{\alpha}_t}\nabla_{z_t} \log p(c|z_t), \tag{5}$$

where the guidance term $\nabla_{z_t} \log p(c|z_t)$ acts as a corrective gradient from a conditioning model (e.g., a classifier) to guide the output towards satisfying condition $c$.

# 4 Method

We propose *FreeInpaint*, an innovative training-free framework for image inpainting, as depicted in Fig. 2. Our framework comprises two key components: Prior-Guided Noise Optimization (PriNo, Sec. 4.1) and Decomposed Training-Free Guidance (DeGu, Sec. 4.2). Initially, prior to the denoising process, we leverage prior knowledge of image inpainting to optimize the initial noise latent, enhancing prompt alignment. Subsequently, during denoising, we decompose the inpainting process into three conditional distributions and apply specific guidance strategies to each, leading to improved visual rationality and human preference.

## 4.1 Prior-Guided Noise Optimization

A primary limitation of existing inpainting methods is the failure to generate objects specified in the text prompt, a problem we denote as *prompt-unaligned*. Since a model's output is highly sensitive to the initial random noise (Wallace et al. 2023; Samuel et al. 2024; Guo et al. 2024), we propose **PriNo**, a novel Prior-Guided Noise Optimization strategy leveraging the attention mechanisms within inpainting models. PriNo optimizes the initial noise latent $z_T$ to steer the model's attention, promoting the generation of the desired object during the early, structure-defining stages of denoising (Yang et al. 2023; Yi et al. 2024; Park et al. 2023).

**Misdirected Attention** Our approach is motivated by an analysis of the attention mechanism (Vaswani et al. 2017) in inpainting diffusion models. In SD, the attention map is computed as $A = \mathrm{softmax}\left(QK^T/\sqrt{d}\right)$, where $d$ is the feature dimension. The query $Q$ is a projection of the visual features, while the key $K$ is a projection of either the prompt embedding $c$ (for cross-attention) or the visual features themselves (for self-attention). Thus, cross-attention maps $A^c$ indicate relevance between text tokens and visual patches, while self-attention maps $A^s$ indicate relevance between different visual patches.

To visualize attention maps across different inpainting results, we aggregate a cross-attention map $A^c$ by averaging

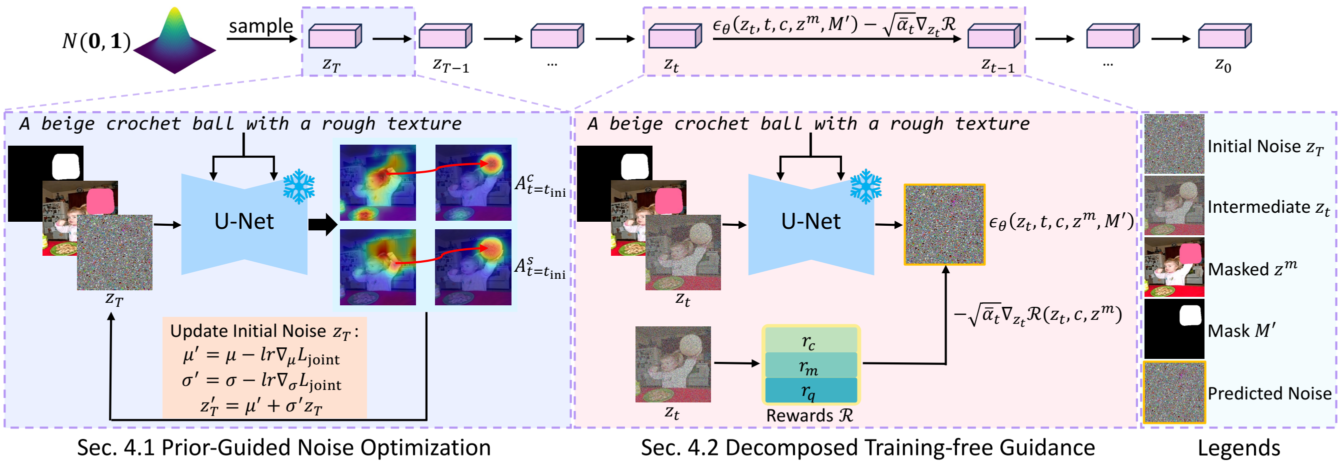



Figure 2: Our tuning-free *FreeInpaint* framework consists of two key stages: (1) Prior-Guided Noise Optimization (Sec. 4.1), which optimizes the initial noise $z_T$ at the first denoising step $t = t_{\text{ini}}$ to concentrate cross-attention ($A^c_{t=t_{\text{ini}}}$) and self-attention ($A^s_{t=t_{\text{ini}}}$) maps within the masked region, improving prompt alignment; and (2) Decomposed Training-Free Guidance (Sec. 4.2), which decomposes the conditional distribution of inpainting and leverages text alignment $r_c$, visual rationality $r_m$, and human preference $r_q$ reward models to guide the predicted noise $\epsilon_\theta(z_t, t, c, z^m, M')$ during denoising, enhancing visual rationality.

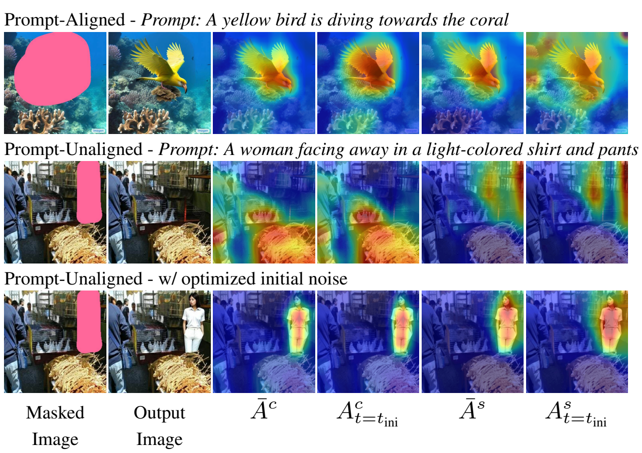



Figure 3: Visualization of cross-attention (cols 3-4) and self-attention (cols 5-6). Row 2 shows misdirected attention causing an unaligned result, while row 3 shows our optimized noise concentrates attention for a successful alignment. The similarity between the first-step ($t = t_{\text{ini}}$) and averaged maps validates our efficient optimization.

them across all text tokens at the $h' \times w'$ spatial resolution of the model's feature maps. Similarly, a self-attention map $A^s$ is derived by averaging across all visual patches within the masked region. Columns 3 and 5 in Fig. 3 display the attention maps using BrushNet (Ju et al. 2024), averaged across all timesteps ($\bar{A}^c$ and $\bar{A}^s$), for a prompt-aligned and unaligned inpainting example. We observe that prompt-aligned inpainting results (row 1) exhibit highly concentrated attention within the mask, whereas unaligned instances (row 2) suffer from misdirected attention that fails to concentrate on the masked region. We hypothesize that this is a side effect of training on global image captions with randomly scribbled masks (Wang et al. 2023; Xie et al. 2023), which teaches the model to prioritize background context.

**Noise Optimization for Attention Steering** To counteract attention misdirection, we propose an attention-steering objective: **minimize attention outside the mask while maximizing it inside**. This principle is applied distinctly to each attention type to resolve their specific failure modes:

- Cross-attention $A^c$ measures text-visual patch relevance. Misdirected $A^c$ incorrectly associates prompt content with the background, failing to generate the desired object within the mask. We correct this by forcing relevance between the prompt and visual patches inside the mask, ensuring prompt concepts manifest in the correct place.
- Self-attention $A^s$ measures inter-patch visual relevance. Misdirected $A^s$ causes the inpainted region to be over-influenced by the surrounding context. Our objective mitigates this, guiding the new content to focus on its own structure, not the background.

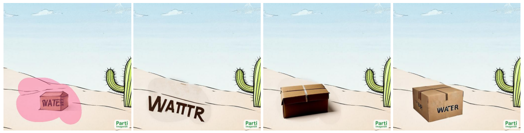
To implement this objective, we design two corresponding loss functions. The cross-attention loss encourages relevance between the prompt and the masked region:

$$\mathcal{L}_{\text{c}} = \sum_{i=1}^{h'} \sum_{j=1}^{w'} \left[ (1 - M'_{ij}) \cdot A^c_{ij} - M'_{ij} \cdot A^c_{ij} \right], \qquad (6)$$

where $M' \in \mathbb{R}^{h' \times w'}$ is the downsampled input mask. Similarly, the self-attention loss encourages the process to focus on the masked region:

$$\mathcal{L}_{\text{s}} = \sum_{i=1}^{h'} \sum_{j=1}^{w'} \left[ (1 - M'_{ij}) \cdot A^s_{ij} - M'_{ij} \cdot A^s_{ij} \right]. \qquad (7)$$

*Prompt: The letters "water" on a brown paper box.*




| Input | BrushNet | BrushNet + PriNo | BrushNet + PriNo + DeGu |
|---|---|---|---|

Figure 4: Sec. 4.1 PriNo improves prompt alignment and Sec. 4.2 DeGu further refines image details.

Optimizing the initial noise latent using the attention map averaged across all denoising steps is computationally prohibitive. However, we observe that attention maps at the initial denoising step $t = t_{\text{ini}}$ closely resemble the full average attention maps, as depicted in cols. 4 and 6 of Fig. 3. Thus, we compute $\mathcal{L}_{\text{c}}$ and $\mathcal{L}_{\text{s}}$ only at this first step.

We employ an effective update mechanism proposed by (Guo et al. 2024), which tunes the mean $\mu$ and standard deviation $\sigma$ of the noise distribution $z_T \sim \mathcal{N}(\mu, \sigma^2)$, initialized with $\mu = \mathbf{0}$ and $\sigma = \mathbf{1}$. Our full objective is shown as:

$$\mathcal{L}_{\text{joint}} = \lambda_1 \mathcal{L}_{\text{c}} + \lambda_2 \mathcal{L}_{\text{s}} + \lambda_3 \mathcal{L}_{\text{KL}}, \tag{8}$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are weighting hyperparameters, and $\mathcal{L}_{\text{KL}}$ is a KL divergence (Kullback and Leibler 1951) to constrain the optimized distribution to remain close to $\mathcal{N}(\mathbf{0}, \mathbf{1})$. We iteratively update $\mu$ and $\sigma$ with a learning rate $lr$ for $\tau_{\text{iter}}$ iterations to minimize $\mathcal{L}_{\text{joint}}$, then sample a new noise latent $z'_T = \mu' + \sigma' z_T$ for the inpainting process. The complete procedure is detailed in the pseudocode in Appendix A. The steered attention maps and refined inpainting results are illustrated in row 3 of Fig. 3.

## 4.2 Decomposed Training-free Guidance

While our PriNo strategy improves prompt alignment, it can leave finer details suboptimal. To enhance image detail and visual rationality, we introduce **DeGu**, a Decomposed Training-free Guidance approach. DeGu refines results by decomposing the inpainting process into three explicit objectives, each guided by an off-the-shelf reward model.

**Decomposing Inpainting Process** As illustrated in Fig. 4, even when the main object "paper box" is generated correctly via PriNo, finer details (e.g., the drawn letter) and image quality (e.g., brightness) can be lacking. To address these limitations in detail and quality, we revisit the inpainting process from the perspective of conditional probability. Then we decompose the inpainting process to tailor a fine-grained guidance mechanism that targets text alignment, visual rationality, and human preference.

Existing guidance mechanisms lack fine-grained control specifically tailored for the nuances of image inpainting (Dhariwal and Nichol 2021; crowsonkb 2022; Jack000 2022; Xie and Gong 2024). We address this by formulating the inpainting process as modeling the conditional probability $p(z_t|c, z^m)$, and decompose it to explicitly address the key objectives for high-quality inpainting:

$$\begin{aligned} p(z_t|c, z^m) &\overset{①}{\propto} p(c, z^m|z_t) \cdot p(z_t) \\ &\overset{②}{=} p(c|z_t) \cdot p(z^m|z_t) \cdot p(z_t). \end{aligned} \tag{9}$$

Here, proportionality ① follows from Bayes' Theorem (Jøsang 2016), while equality ② assumes conditional independence between the text $c$ and masked latent $z^m$ (Gut 2013). To further enhance human preference, we consider a human preference condition $q$, with $p(q|z_t) \propto \exp(r_{\text{q}}(z_t))$ based on a human preference reward function $r_{\text{q}}$ (Song and Kingma 2021). This allows us to augment the posterior $p(z_t|c, z^m)$ with condition $q$ for a more refined distribution:

$$p(z_t|c, z^m, q) \propto p(c|z_t) \cdot p(z^m|z_t) \cdot p(q|z_t) \cdot p(z_t). \tag{10}$$

Instead of treating text-guided inpainting as a monolithic process, we explicitly decompose this process into three distinct and guided objectives: (1) *text alignment* $p(c|z_t)$, enforcing alignment of the generated content to the text prompt, (2) *visual rationality* $p(z^m|z_t)$, ensuring coherence between the generated and known regions, and (3) *human preference* $p(q|z_t)$, promoting the overall aesthetic appeal and plausibility of the generated image.

**Reward-Guided Denoising** Building upon the guidance techniques in Sec. 3, we achieve multi-conditional guidance by employing a weighted combination of score functions (Song et al. 2020) during denoising:

$$\begin{aligned} \nabla_{z_t} \log p(z_t|c, z^m, q) &= -\frac{\epsilon_t}{\sqrt{1-\bar{\alpha}_t}} + \gamma_c \nabla_{z_t} \log p(c|z_t) \\ &+ \gamma_m \nabla_{z_t} \log p(z^m|z_t) + \gamma_q \nabla_{z_t} \log p(q|z_t), \end{aligned} \tag{11}$$

where $\gamma$ terms control the guidance strength. This novel decomposition enables us to use specialized, off-the-shelf reward models for each objective, leading to more refined and controllable inpainting results. Specifically, we employ a reward model $r_c$ for text alignment, $r_m$ for visual rationality, and $r_q$ for human preference. Since the probability of each objective is proportional to the exponential of its reward score (i.e., $p(\cdot|z_t) \propto \exp(r(z_t))$) (Song and Kingma 2021), we can directly refine the predicted noise:

$$\begin{aligned} \hat{\epsilon}_t = \epsilon_\theta(z_t, t, c, z^m, M') &- \gamma_c \sqrt{\bar{\alpha}_t} \nabla_{z_t} r_c(z_t, c) \\ - \gamma_m \sqrt{\bar{\alpha}_t} \nabla_{z_t} r_m(z_t, z^m) &- \gamma_q \sqrt{\bar{\alpha}_t} \nabla_{z_t} r_q(z_t, c), \end{aligned} \tag{12}$$

Notably, we use $\sqrt{\bar{\alpha}_t}$ as the reward modulator instead of the conventional $\sqrt{1-\bar{\alpha}_t}$ in Sec. 3. As $\sqrt{\bar{\alpha}_t}$ is monotonically increasing during denoising, it down-weights the influence of unreliable predictions from initial noisy steps, which we find empirically yields better performance. Finally, to keep the unmasked region unchanged, we blend the inpainted image with the original input. See pseudocode in Appendix A.

# 5 Experiment

## 5.1 Experimental Settings

**Baselines** We compare *FreeInpaint* against a suite of competitive baselines, including SD1.5-Inpainting (SDI) (Rombach et al. 2022), PowerPaint (PPT) (Zhuang et al. 2024), BrushNet-Random (BN) (Ju et al. 2024), SDXL-Inpainting (SDXLI) (Podell et al. 2023), and the state-of-the-art DiT-based (Peebles and Xie 2023) SD3-ControlNet-Inpainting (SD3I) (alimama creative 2024). PPT uses paired mask-description data for training, while others do not. Our key competitor is HD-Painter (HDP) (Manukyan et al. 2023), a training-free method focused only on prompt alignment, while the other baselines are training-based. The same random seeds are used across all experiments.

**Implementation Details** We evaluate our method and HDP by integrating them with SDI, PPT, BN, SDXLI, SD3I. In DeGu, the reward models are: local CLIPScore (Radford et al. 2021) $r_c$ to assess the alignment between the prompt $c$ and the content generated within the mask $M$, InpaintReward (Liu et al. 2024b) $r_m$ to evaluate the visual coherence, and ImageReward (Xu et al. 2023) $r_q$ to measure human preference. Other details are in Appendix B.

**Benchmarks** We employ two benchmarks: EditBench (Wang et al. 2023) and MSCOCO (Lin et al. 2014). EditBench provides handcrafted, scribble-like free-form masks. For extended evaluation, we also use precise, object-aligned layout masks from the MSCOCO validation set. As original MSCOCO labels are single words, we generate detailed mask captions using LLaVA (Liu et al. 2024a) to better assess prompt alignment. More details are in Appendix B.

We define two prompt types: the *global prompt*, which describes the entire image (the ”Full prompt” in EditBench and the caption of an entire image in MSCOCO), and the *local prompt*, which is a detailed, mask-level description (the ”Mask-Rich prompt” in EditBench and our annotated mask caption in MSCOCO). For inference, we use the *local prompt* as it better simulates real-world scenarios, such as adding content to a photograph.

**Metrics** To prevent metric leakage, we employ multiple metrics beyond DeGu's rewards. We assess: human preference with ImageReward and HPSv2 (Wu et al. 2023), using *global prompts* to be consistent with prior works (Ju et al. 2024) and better gauge overall image quality (distinct from the local prompts used during inference); prompt alignment with local and global CLIPScore; and visual rationality with InpaintReward and LPIPS (Zhang et al. 2018). A user study was also conducted for thorough, unbiased comparison.

## 5.2 Quantitative Results

**Free-form Mask Inpainting** On EditBench with scribble-like free-form masks (Tab. 1), *FreeInpaint* comprehensively improves all baselines. As our guidance directly targets ImageReward, local CLIPScore, and InpaintReward, these metrics show substantial gains as expected. Notably, our method boosts the global-prompt-based ImageReward even while using only local prompts in DeGu. Non-guidance metrics like HPSv2, global CLIP, and LPIPS also improve obviously, indicating holistic enhancement.

Furthermore, *FreeInpaint* consistently outperforms HDP. This superiority is especially clear in cases where HDP degrades image quality. For instance, with the SDI baseline, HDP worsens HPSv2, InpaintReward, and LPIPS scores,

| | | ImageReward | HPSv2 | L.CLIP | G.CLIP | InpaintReward | LPIPS ↓ |
|---|---|---|---|---|---|---|---|
| SDI | Base | -0.1341 | 23.36 | 23.90 | 26.13 | -0.1732 | 0.2073 |
| | +HDP | -0.0110 | 22.99 | **26.09** | 26.85 | -0.3075 | 0.2419 |
| | +Ours | **0.1753** | **23.65** | 25.92 | **27.59** | **-0.0595** | **0.2011** |
| PPT | Base | 0.0842 | 23.72 | 25.12 | 26.47 | -0.2133 | 0.2244 |
| | +HDP | 0.0554 | 23.30 | 25.51 | 26.58 | -0.2445 | 0.2436 |
| | +Ours | **0.2361** | **23.94** | **25.69** | **27.52** | **-0.1262** | **0.2116** |
| BN | Base | 0.2729 | 25.34 | 26.45 | 27.18 | -0.1791 | **0.1947** |
| | +HDP | 0.3836 | 25.20 | 27.08 | 27.69 | -0.2124 | 0.2135 |
| | +Ours | **0.5006** | **25.64** | **27.81** | **28.28** | **-0.0878** | 0.2005 |
| SDXLI | Base | 0.2318 | 25.10 | 25.87 | 27.93 | -0.2792 | 0.2826 |
| | +HDP | 0.1543 | 24.34 | **26.62** | **28.38** | -0.3463 | 0.3206 |
| | +Ours | **0.3455** | **25.14** | **26.62** | 28.29 | **-0.1888** | **0.2087** |
| SD3I | Base | 0.2993 | 25.48 | 26.26 | 27.43 | -0.2170 | 0.2155 |
| | +HDP | -0.5020 | 21.56 | 22.83 | 25.36 | -0.2988 | 0.2516 |
| | +Ours | **0.5248** | **25.70** | **26.98** | **28.31** | **-0.0694** | **0.2057** |

Table 1: Quantitative comparisons on EditBench dataset with free-form masks. L/G.CLIP refer to Local/Global CLIP. **Bold** indicates the best performance.

whereas our method provides substantial improvements. This highlights *FreeInpaint*'s ability to simultaneously enhance prompt alignment and visual rationality, which we attribute to our strategy of decoupling and separately optimizing for these two objectives. Furthermore, unlike our approach, HDP is incompatible with the DiT-based SD3I, highlighting *FreeInpaint*'s superior adaptability.

**Layout Mask Inpainting** To evaluate generalization, we test on MSCOCO's object-aligned layout masks (Tab. 2). Our method's strong performance extends to this dataset, confirming its robustness across different mask types. *FreeInpaint* demonstrates remarkable effectiveness even on strong, specialized baselines, achieving top performance across all metrics for both PPT (which is trained on paired mask-description data) and the DiT-based SD3I.

While HDP is competitive on some metrics (e.g., Local CLIP on BN), our method provides a better overall trade-off. Take BN for example, *FreeInpaint* achieves better ImageReward, InpaintReward, and LPIPS at the cost of a narrow margin on CLIP score. This confirms that our decoupled optimization strategy achieves more robust enhancements in both prompt alignment and visual rationality.

**User Study** Considering that quantitative metrics may not perfectly align with human preferences in image generation tasks, we conducted a user study. We randomly selected 30 samples from EditBench and MSCOCO datasets. 59 evaluators with diverse educational backgrounds independently assessed the generated results from SDI, SDI + HDP, and SDI + Ours, selecting the best based on prompt alignment and overall quality. The win rates of the three models are 16.16%, 19.32%, and 64.52%, respectively, which demonstrate a significant user preference for our method.

| | | ImageReward | HPSv2 | L.CLIP | G.CLIP | InpaintReward | LPIPS ↓ |
|---|---|---|---|---|---|---|---|
| | Base | 0.2574 | 25.50 | 24.09 | 24.63 | -0.0206 | 0.1126 |
| SDI | +HDP | **0.2601** | 24.43 | **24.53** | 24.36 | -0.0346 | 0.1231 |
| | +Ours | 0.2591 | **25.69** | 24.41 | **24.94** | **0.0112** | **0.0844** |
| | Base | 0.2593 | 25.52 | 24.67 | 24.72 | -0.0149 | 0.1188 |
| PPT | +HDP | 0.2167 | 25.51 | 24.91 | 24.80 | -0.0201 | 0.1180 |
| | +Ours | **0.2671** | **25.79** | **25.09** | **24.91** | **0.0171** | **0.0861** |
| | Base | 0.2701 | **26.02** | 24.51 | 24.76 | 0.0140 | 0.0867 |
| BN | +HDP | 0.2766 | 25.98 | **24.87** | **24.98** | 0.0097 | 0.0876 |
| | +Ours | **0.2782** | **26.02** | 24.69 | 24.80 | **0.0568** | **0.0862** |
| | Base | 0.2665 | 26.10 | 24.38 | **24.95** | **-0.0439** | 0.1802 |
| SDXLI | +HDP | 0.1810 | 24.22 | 23.73 | 24.54 | -0.0565 | 0.2640 |
| | +Ours | **0.3191** | **26.89** | **24.66** | 24.89 | -0.0504 | **0.0683** |
| | Base | 0.2795 | 26.51 | 24.62 | 24.84 | 0.0093 | 0.1008 |
| SD3I | +HDP | -0.0010 | 24.99 | 23.52 | 24.62 | -0.1005 | 0.1072 |
| | +Ours | **0.3422** | **27.10** | **25.14** | **25.13** | **0.0273** | **0.0680** |

Table 2: Quantitative comparisons on MSCOCO dataset with layout masks. L/G.CLIP refer to Local/Global CLIP. **Bold** indicates the best performance.

| | ImageReward | HPSv2 | L.CLIP | G.CLIP | InpaintReward | LPIPS ↓ |
|---|---|---|---|---|---|---|
| BN | 0.2729 | 25.34 | 26.45 | 27.18 | -0.1791 | 0.1947 |
| BN + PriNo | 0.3785 | 25.57 | 26.96 | 27.72 | -0.2124 | 0.2021 |
| BN + DeGu | 0.3908 | 25.36 | 27.17 | 27.87 | **-0.0643** | 0.1944 |
| Constant | 0.3533 | 25.41 | 26.92 | 27.56 | -0.1088 | 0.1936 |
| $\sqrt{1-\bar{\alpha}_t}$ | 0.3454 | 25.27 | 26.85 | 27.53 | -0.1146 | **0.1922** |
| Ours | **0.5006** | **25.64** | **27.81** | **28.28** | -0.0878 | 0.2005 |

Table 3: Quantitative ablation study on EditBench dataset.

### 5.3 Qualitative Results

Qualitative results are shown in Fig. 5 and 6 (prompts simplified for presentation). In Fig. 6, both baseline methods BN and HDP struggle to render text (I) or generate the correct content (II, IV), while HDP produces an unnatural boundary (III). In Fig. 5, the outputs of the baselines are either unaesthetic (I) or visually flawed (III). In contrast, our method resolves all these shortcomings.

### 5.4 Ablation Study

We conduct an ablation study on EditBench using BrushNet to analyse each component's effect (Tab. 3 and Fig. 4). The results confirm that our main components, Sec. 4.1 PriNo and Sec. 4.2 DeGu, are individually effective and their benefits are additive. PriNo improves prompt-related metrics but reduces the visual rationality metric InpaintReward, while DeGu compensates for this. We also analyze our guidance modulator, $\sqrt{\bar{\alpha}_t}$, against alternatives $\sqrt{1-\bar{\alpha}_t}$ and a constant 0.5. Our chosen modulator demonstrates superior performance across most metrics. For more ablation analyses, please refer to Appendix C.

## 6 Conclusion

In this work, we present *FreeInpaint*, a plug-and-play tuning-free framework for text-guided image inpainting. By optimizing the initial noise with priors tailored for inpainting and guiding intermediate latents with a task-specific composite objective, our method achieves both superior prompt alignment and visual rationality. Specifically, the proposed noise optimization steers attention toward masked regions in the initial denoising process, while the composite reward—combining text alignment, visual rationality, and human preference—effectively guides the inpainting process without training. Experiments on multiple inpainting diffusion models and benchmarks demonstrate the effectiveness and generalizability of our *FreeInpaint*.

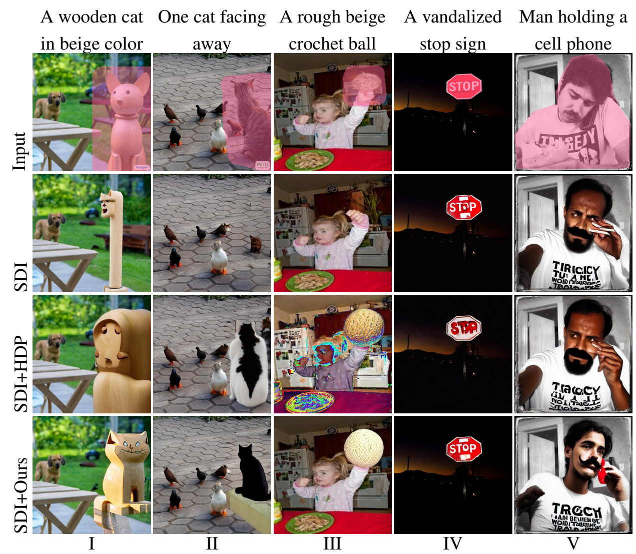



Figure 5: Comparison against SDI-based approaches.


Figure 6: Comparison against BN-based approaches.

## Acknowledgments

This work was supported by NSFC project (No. 62232006).

## References


alimama creative. 2024. SD3-Controlnet-Inpainting. https://huggingface.co/alimama-creative/SD3-Controlnet-Inpainting.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*.

Ben-Hamu, H.; Puny, O.; Gat, I.; Karrer, B.; Singer, U.; and Lipman, Y. 2024. D-flow: Differentiating through flows for controlled generation. *arXiv preprint arXiv:2402.14017*.

Bertalmio, M.; Sapiro, G.; Caselles, V.; and Ballester, C. 2000. Image inpainting. In *SIGGRAPH*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

Cai, Q.; Chen, J.; Chen, Y.; Li, Y.; Long, F.; Pan, Y.; Qiu, Z.; Zhang, Y.; Gao, F.; Xu, P.; et al. 2025. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. *arXiv preprint arXiv:2505.22705*.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*

Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; and Shin, J. 2024. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*.

crowsonkb. 2022. k-diffusion. https://github.com/crowsonkb/k-diffusion.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. *NeurIPS*.

Eyring, L.; Karthik, S.; Roth, K.; Dosovitskiy, A.; and Akata, Z. 2024. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *NeurIPS*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Commun. ACM*.

Guo, X.; Liu, J.; Cui, M.; Li, J.; Yang, H.; and Huang, D. 2024. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*.

Gut, A. 2013. *Probability: A Graduate Course (Second ed.)*.

Han, F.; Wang, Y.; Li, C.; Liang, Z.; Wang, D.; Jiao, Y.; Wei, Z.; Gong, C.; Jin, C.; Chen, J.; et al. 2025. UniREditBench: A Unified Reasoning-based Image Editing Benchmark. *arXiv preprint arXiv:2511.01295*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Jack000. 2022. glid-3-xl. https://github.com/Jack000/glid-3-xl.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *ECCV*.

Jøsang, A. 2016. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*.

Karunratanakul, K.; Preechakul, K.; Aksan, E.; Beeler, T.; Suwajanakorn, S.; and Tang, S. 2024. Optimizing diffusion noise can serve as universal motion priors. In *CVPR*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*.

Li, D.; Zhong, W.; Yu, W.; Pan, Y.; Zhang, D.; Yao, T.; Han, J.; and Mei, T. 2025. Pursuing Temporal-Consistent Video Virtual Try-On via Dynamic Pose Interaction. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, K.; Zhu, Z.; Li, C.; Liu, H.; Zeng, H.; and Hou, J. 2024b. Prefpaint: Aligning image inpainting diffusion model with human preference. *NeurIPS*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*.

Manukyan, H.; Sargsyan, A.; Atanyan, B.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. In *ICLR*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Novack, Z.; McAuley, J.; Berg-Kirkpatrick, T.; and Bryan, N. J. 2024. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*.

Park, Y.-H.; Kwon, M.; Choi, J.; Jo, J.; and Uh, Y. 2023. Understanding the latent space of diffusion models through the lens of riemannian geometry. *NeurIPS*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *CVPR*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*.

Samuel, D.; Ben-Ari, R.; Raviv, S.; Darshan, N.; and Chechik, G. 2024. Generating images of rare concepts using pre-trained diffusion models. In *AAAI*.

Song, Y.; and Kingma, D. P. 2021. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*.

Wallace, B.; Gokul, A.; Ermon, S.; and Naik, N. 2023. End-to-end diffusion latent optimization improves classifier guidance. In *ICCV*.

Wan, S.; Li, Y.; Chen, J.; Pan, Y.; Yao, T.; Cao, Y.; and Mei, T. 2024. Improving virtual try-on with garment-focused diffusion models. In *ECCV*.

Wang, S.; Saharia, C.; Montgomery, C.; Pont-Tuset, J.; Noy, S.; Pellegrini, S.; Onoe, Y.; Laszlo, S.; Fleet, D. J.; Soricut, R.; et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*.

Weimin, Q.; Jieke, W.; and Meng, T. 2025. Self-Cross Diffusion Guidance for Text-to-Image Synthesis of Similar Subjects. In *CVPR*.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*.

Xie, X.; and Gong, D. 2024. DyMO: Training-Free Diffusion Model Alignment with Dynamic Multi-Objective Scheduling. *arXiv preprint arXiv:2412.00759*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*.

Yang, X.; Zhou, D.; Feng, J.; and Wang, X. 2023. Diffusion probabilistic model made slim. In *CVPR*.

Yao, T.; Li, Y.; Pan, Y.; Qiu, Z.; and Mei, T. 2025. Denoising token prediction in masked autoregressive models. In *ICCV*.

Yi, M.; Li, A.; Xin, Y.; and Li, Z. 2024. Towards understanding the working mechanism of text-to-image diffusion model. *arXiv preprint arXiv:2405.15330*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2024. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *ECCV*.