Token Painter: Training-Free Text-Guided Image Inpainting via Mask Autoregressive Models

Longtao Jiang[1], Jie Huang[1], Mingfei Han[2], Lei Chen[1], Yongqiang Yu[2], Feng Zhao[1], Xiaojun Chang[1], Zhihui Li[1]*

[1]University of Science and Technology of China
[2]Department of Computer Vision, MBZUAI
{taotao707,hj0117}@mail.ustc.edu.cn, mingfei.han,yongqiang.yu@mbzuai.ac.ae
chenlei.hfut@gmail, {fzhao956,xjchang,lizhihuics}@ustc.edu.cn

Abstract

Text-guided image inpainting aims to inpaint masked image regions based on a textual prompt while preserving the background. Although diffusion-based methods have become dominant, their property of modeling the entire image in latent space makes it challenging for the results to align well with prompt details and maintain a consistent background. To address these issues, we explore Mask AutoRegressive (MAR) models for this task. MAR naturally supports image inpainting by generating latent tokens corresponding to mask regions, enabling better local controllability without altering the background. However, directly applying MAR to this task makes the inpainting content either ignore the prompts or be disharmonious with the background context. Through analysis of the attention maps from the inpainting images, we identify the impact of background tokens on text tokens during the MAR generation, and leverage this to design **Token Painter**, a training-free text-guided image inpainting method based on MAR. Our approach introduces two key components: (1) Dual-Stream Encoder Information Fusion (DEIF), which fuses the semantic and context information from text and background in frequency domain to produce novel guidance tokens, allowing MAR to generate text-faithful inpainting content while keeping harmonious with background context. (2) Adaptive Decoder Attention Score Enhancing (ADAE), which adaptively enhances attention scores on guidance tokens and inpainting tokens to further enhance the alignment of prompt details and the content visual quality. Extensive experiments demonstrate that our training-free method outperforms prior state-of-the-art methods across almost all metrics.

Code — https://github.com/longtaojiang/Token-Painter

1 Introduction

Image inpainting (Li et al. 2022; de Jorge et al. 2024; Liu et al. 2022; Dong, Cao, and Fu 2022) aims at filling masked regions and keeping harmonious with context. With the rapid development of text-to-image (T2I) generation (Esser et al. 2024; Saharia et al. 2022; Sun et al. 2025; Liu et al. 2025; Yang et al. 2025), text-guided image inpainting has gained significant attention, with approaches like Stable Diffusion Inpainting (SDI) (Rombach et al. 2022; Ho and Jain 2020; Song et al. 2020) leading the field.

*Zhihui Li is the corresponding author.

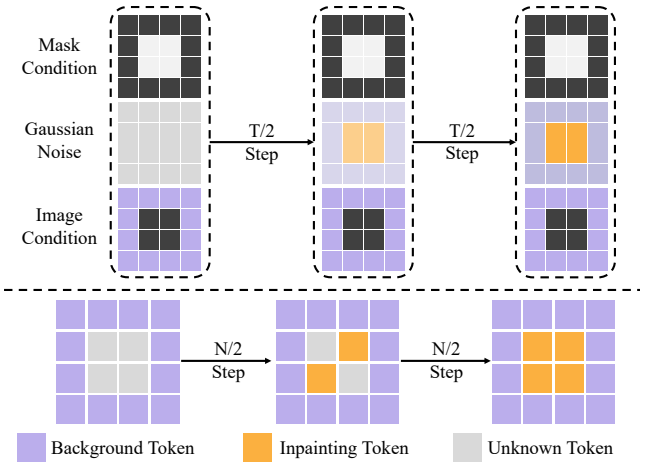


Figure 1: Comparison of the inpainting process of diffusion models (above) and our MAR-based method (below).

However, the diffusion-based text-guided inpainting faces potential challenges. The property of diffusion models that denoises the entire image in the latent space, causing the structure of the inpainting region to be guided more by context rather than text prompts (Manukyan et al. 2023; Hsiao et al. 2024). *This leads to the inpainting content either being poorly aligned with the prompt or lacking harmony due to the conflict with context, resulting in low visual quality.* Some methods like HD-Painter (Manukyan et al. 2023) and FreeCond (Hsiao et al. 2024) made training-free improvements based on SDI. Other methods like BrushNet, PowerPaint (Zhuang et al. 2023; Ju et al. 2024), attempt to finetune stable diffusion (SD). Although these methods show some improvement in text alignment, the quality of inpainting content is still unsatisfactory. Furthermore, *the generative property of denoising the entire image also inherently disrupts the consistency of the background*, as shown in Figure 1. And simple blending operations (Ju et al. 2024) hardly resolve issues such as lighting or color mismatches.

Recently, Autoregressive (AR) models (Vaswani et al. 2017; Radford et al. 2018; Sun et al. 2024; Team 2024) have garnered increasing attention in T2I generation. In AR models, each token in latent space corresponds to a part of the

Figure 2: Our method Token Painter faithfully follows the prompt details and seamlessly connect with the image context.

image spatially, and unknown tokens are directly predicted based on known tokens at each step during the generation process. Compared to the joint denoising of all tokens in diffusion, the property of AR leads to better controllability of local content generation. But traditional AR models predict tokens through a raster-order, limiting their applicability in inpainting, where the mask positions are typically random. Recently, variants of AR models (Tian et al. 2024; Pang et al. 2025; Yu et al. 2025; Li et al. 2024) have emerged, among which the mask autoregressive (MAR) (Li et al. 2024) stands out. Unlike traditional AR models, MAR generates image tokens at arbitrary locations. In addition, MAR inherits the property of AR that generates tokens at each step, making it have strong local controllability, and also keeping background tokens unchanged. As shown in Figure 1, MAR is naturally suitable for inpainting. Therefore, our work focuses on applying recent T2I MAR model NOVA (Deng et al. 2024) with an encoder-decoder architecture to text-guided image inpainting.

Our study starts by setting both text prompt and background tokens (T&B) as the input to generate the masked region. However, the inpainting content ignores the text prompt and relies solely on the context, as shown in Figure 3(a). Then, we only use the text prompt as input and disregard all background tokens (T-only). Although the inpainting content follows the prompt this time, it is highly disharmonious with the surrounding context. To further investigate the reasons for these failures, we visualize two types of self-attention maps in the decoder stage for both approaches, as shown in Figure 3(b). The results reveal that in the T&B case, the attention scores of both types are dispersed to background regions, while in the T-only case, the attention scores are concentrated within the inpainting region. Therefore, we infer that in the T&B approach, the semantic information of text tokens is overwhelmed by the context information of background tokens during encoder interaction. In the T-only approach, due to the lack of background tokens, the text tokens fully retain their semantic information, but the generated tokens are highly disharmonious with image context.

To address the above issues, we propose Token Painter, a training-free text-guided image inpainting method based on MAR. We design it at both encoder and decoder stages. At the encoder stage, we present the Dual-Stream Encoder Information Fusion (DEIF) module, which aims to produce novel guidance tokens, *i.e.*, the updated text tokens after encoder interaction. With those guidance tokens, the model can generate inpainting content that follows the text prompt and keeps in harmony with the image context. The module first aligns the two rough guidance tokens from T&B and T-only statistically, and then fuses them in the frequency domain. At the decoder stage, we design the Adaptive Decoder Attention Score Enhancing (ADAE) module to further improve the prompt detail alignment and the content visual quality. This module enhances the attention of inpainting tokens to guidance tokens adaptively, and strengthens the attention interaction within the inpainting region. Our contributions are summarized as follows:

- We improve the T2I MAR model specifically for the text-guided image inpainting task, and conduct a detailed analysis of MAR text-guided inpainting process, revealing the interactions between text tokens and background tokens, as well as their impact on the inpainting content.
- We introduce the Dual-Stream Encoder Information Fusion (DEIF) module to obtain novel guidance tokens that guide MAR to generate inpainting content that follows the text prompt while keeping harmonious with context.
- To further enhance the alignment of prompt details and the visual quality of inpainting, we introduce the Adaptive Decoder Attention Score Enhancing (ADAE) module, which adaptively enhances the attention scores.
- Token Painter is a training-free method based on the T2I MAR model, yet it outperforms previous methods across nearly all metrics, including SOTA, even though they are based on models fine-tuned on inpainting datasets.

2 Related Work

2.1 Text-guided Image Inpainting

In recent years, many diffusion-based text-guided image inpainting works (Nichol et al. 2021; Avrahami et al. 2023, 2022; Wang et al. 2023) have emerged. However, the property of diffusion limits visual quality and prompt alignment of these methods, and also disrupts background consistency. There are some T2I MAR models (Chang et al. 2023; Bai et al. 2024) that involve this task, but their naive approaches lead to low-quality inpainting and inconsistent background.

2.2 Text-to-Image Generation

Diffusion and Autoregressive Models. Text-to-image generation, has been dominated by diffusion-based methods (Betker et al. 2023; Chen et al. 2023; Esser et al. 2024) in recent years. However, as AR models (Fan et al. 2024; Sun et al. 2024; Yu et al. 2022) gradually enter the field of visual generation, it has been found that AR-based T2I models tend to follow text prompts better to generate spatial structure due to the higher independence between different image tokens.

Mask Autoregressive Model. The classic AR model generates image tokens according to the raster-order, which conflicts with intuition. As a result, variants of AR models have emerged, among which MAR models (Li et al. 2024; Chang et al. 2022) stands out for its ability to generate tokens at arbitrary positions. The T2I MAR models (Deng et al. 2024; Chang et al. 2023; Bai et al. 2024) has also gained attention due to their superior text alignment. Considering visual generation quality and background region consistency, we choose the T2I MAR model NOVA(Deng et al. 2024) as base model to achieve text-guided image inpainting task.

3 Analysis of MAR Model for Inpainting

3.1 Vanilla Solutions for MAR-based Inpainting

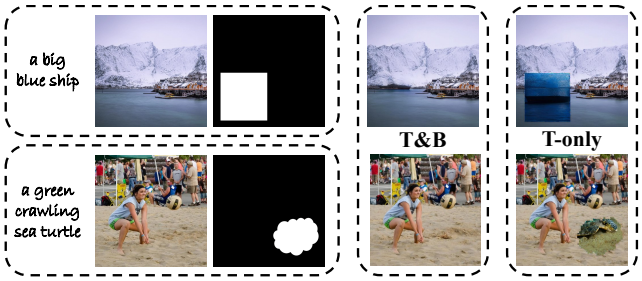
We firstly revisit the main generation process of the T2I MAR model NOVA (Deng et al. 2024) with an encoder-decoder architecture. Given a text prompt, the text encoder converts it into the fixed-length text tokens $T \in \mathbb{R}^{L\times D}$. The model then initializes a group of unknown image tokens $I \in \mathbb{R}^{H\times W\times D}$, and divides them into V sets $\{S^1, S^2, ..., S^V\}$. Those sets are predicted in order based on text tokens and known image tokens. This paradigm is written as:

$$p(S^1, ..., S^V) = \prod_v^V p(S^v \mid T, S^1, ..., S^{v-1}), \qquad (1)$$

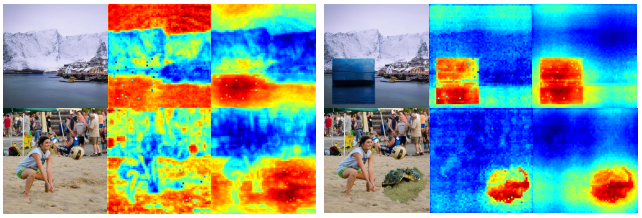
where S^v is a set to be predicted at v-th step, with $\cup_v S^v = I$. During this process, the text tokens T and the known image tokens interact within the encoder, resulting in the updated text tokens, i.e., guidance tokens T_g. The decoder then predicts the next set of image tokens mainly based on them.

To apply this paradigm to the text-guided image inpainting task, we make a simple modification. We label the set of inpainting tokens $I_p \in \mathbb{R}^{N\times D}$ as unknown, while keeping the background tokens $I_b \in \mathbb{R}^{M\times D}$ as predicted tokens, where $M+N = HW$. The modified paradigm is as follows:

$$p(S^1, ..., S^k) = \prod_k^K p(S^k \mid T, I_b, S^1, ..., S^{k-1}), \qquad (2)$$



(a) Result of two approaches. **T** means text prompt, **B** means background.



(b) Attention maps of Image2Text and Image2Inpaint. The image tokens as query, and updated text tokens (guidance tokens), inpainting tokens as key.

Figure 3: Comparison of vallia T&B and T-only approaches.

where S^k is a set to be predicted at k-th step, with $\cup_k S^k = I_p$. However, we find that the inpainting region generated by this approach (T&B) does not follow the prompt at all, as shown in Figure 3(a). The content seems to only rely on the image context. Then we completely mask the background tokens (T-only). The modified paradigm is as follows:

$$p(S^1, ..., S^k) = \prod_k^K p(S^k \mid T, S^1, ..., S^{k-1}). \qquad (3)$$

As shown in Figure 3(a), though the inpainting region follows the prompt now, it is disharmonious with context.

3.2 The Mechanism behind MAR Inpainting

To better understand the impact of the guidance tokens in T&B and T-only approaches, we visualize two types of attention maps for both, as shown in Figure 3(b). In the T&B case, the attention scores of the guidance tokens are distributed across the entire image, especially in the background. Meanwhile, the inpainting tokens show high similarity with both themselves and the surrounding image context. In the T-only case, the attention scores for both token types are strictly limited to the inpainting region. Based on these observations, we propose that in the T&B case, where background tokens are fully visible and interact with text tokens during encoder stage, the semantic information in the text tokens is overwhelmed by the context information. As a result, the inpainting region generated under these guidance tokens tends to align with image context. Conversely, when background tokens are masked, the lack of context interaction means the guidance tokens contain only semantic information, leading to the disharmonious inpainting content.

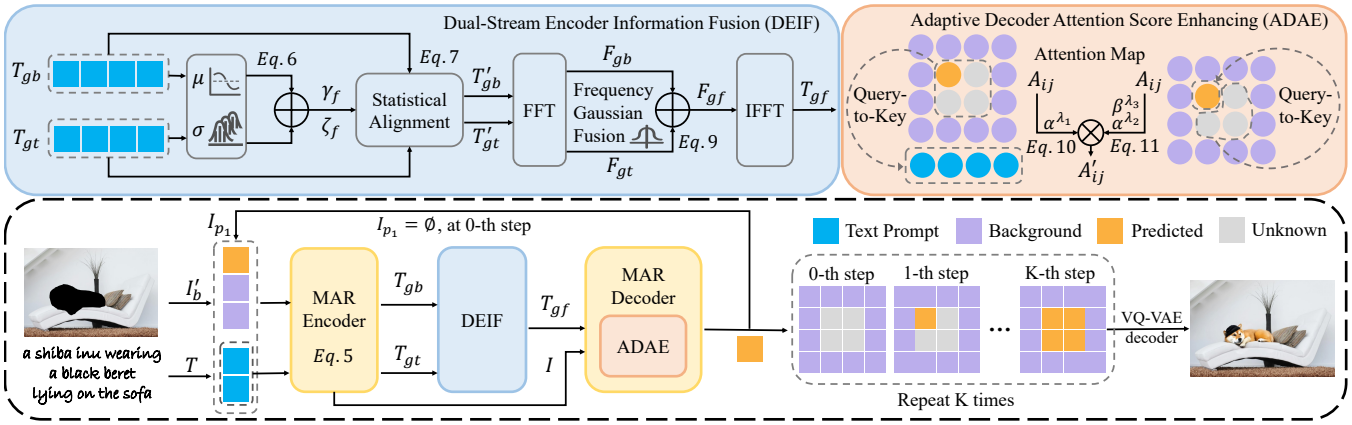


Figure 4: Overview of Token Painter, which includes the DEIF at encoder stage and the ADAE at decoder stage. DEIF produces novel guidance tokens T_{gf} that contain both text and context information through information fusion in frequency domain. ADAE enhances two parts of attention map A to further improve prompt detail alignment and content visual quality.

4 Token Painter

4.1 Overview of Our Method

Overview of Token Painter is presented in the Figure 4. The text prompt is converted into fixed-length text tokens T. Then a VAE encode the image to I. The mask is downsampled to M and multiplied with I to produce I_b and unknown I_p. Next, we input those tokens to the MAR encoder in the way of T&B and T- only to get T_{gb} and T_{gt}, and DEIF module fuse them to get the novel guidance tokens T_{gf}. It is then fed into the decoder to guide the generation of the inpainting region, with ADAE module adaptively enhancing two parts of the attention map A. After repeating K times, we decode latent tokens by VAE to obtain final image.

4.2 Dual-Stream Encoder Information Fusion

Dual-Stream Guidance Tokens. To obtain a guidance tokens that includes both semantic and context information, we need to fuse the two rough guidance tokens, $T_{gb} \in \mathbb{R}^{L\times D}$ and $T_{gt} \in \mathbb{R}^{L\times D}$, from the T&B and T-only approaches mentioned in Section 3. In practical application, we find that T_{gb} only needs to interact with the background tokens surrounding the inpainting region, and it effectively captures contextual information. Therefore, we use the mask $M \in \mathbb{R}^{H\times W\times 1}$ and its dilated version M_d to choose the background tokens for interaction. The process is as follows:

$$I'_b = (I \odot (M_d - M))[: p \cdot N, :], \tag{4}$$

where p is a proportionality coefficient and N is the number of inpainting tokens. Next, we feed the text tokens with the known image tokens, as well as the text tokens alone, into the MAR encoder for interaction. The process is as follows:

$$\begin{aligned} T_{gb} &= ME(Concat(T, I'_b, I_{p_1}))[: L, :] \in \mathbb{R}^{L\times D}, \\ T_{gt} &= ME(T)[:, :] \in \mathbb{R}^{L\times D}, \end{aligned} \tag{5}$$

where $Concat(\cdot)$ means the operation of concatenation, and $ME(\cdot)$ represents the interaction of encoder in MAR.

Adaptive Statistical Alignment. To better fuse T_{gb} and T_{gt} in the frequency domain, we first need to align them statistically in the space domain, which involves normalizing both T_{gb} and T_{gt}, then shifting them to a common distribution. Since the alignment process must account for the statistics of both semantic and context information of each instance, the mean γ_f and variance ζ_f for the shift are adaptively obtained in each instance, as described by the following equation:

$$\begin{aligned} \gamma_f &= a \cdot \mu(T_{gb}) + (1-a) \cdot \mu(T_{gt}), \\ \zeta_f &= a \cdot \sigma(T_{gb}) + (1-a) \cdot \sigma(T_{gt}), \end{aligned} \tag{6}$$

where $\mu(\cdot)$ represents the mean of tokens along L dimension, and $\sigma(\cdot)$ represents the variance. a is a proportional coefficient. Then we align T_{gb} and T_{gt} as follows:

$$\begin{aligned} T'_{gb} &= \gamma_f \cdot \left(\frac{T_{gb} - \mu(T_{gb})}{\sigma(T_{gb})}\right) + \zeta_f, \\ T'_{gt} &= \gamma_f \cdot \left(\frac{T_{gt} - \mu(T_{gt})}{\sigma(T_{gt})}\right) + \zeta_f. \end{aligned} \tag{7}$$

Frequency Information Fusion. Inspired by (Kwon et al. 2024; Gao et al. 2024), we need to fuse the high-frequency context style information from T'_{gb} with the low-frequency semantic structure information from T'_{gt}. We first transform them into the frequency domain using fast fourier transform (FFT) and shift their zero-frequency components to the center of the frequency spectrum, i.e., at $L/2$, to obtain $F_{gb} \in \mathbb{R}^{L\times D}$ and $F_{gt} \in \mathbb{R}^{L\times D}$. Next, we use a modified Gaussian function to fuse the two frequency spectra, yielding F_{gf}. The entire process is illustrated below:

$$MG(l) = exp\left(-(\frac{|l - L/2|}{\varphi})^{\tau}\right), \tag{8}$$

$$F_{gf}(l) = (1 - MG(l)) \cdot F_{gb} + MG(l) \cdot F_{gt}, \tag{9}$$

where $MG(l)$ is the value of the modified Gaussian function at position $l \in \{0, 1, ..., L-1\}$. φ and τ are coefficients used to control the shape. Afterward, we shift the F_{gf} to the original frequency spectrum, and apply the IFFT to it to obtain the novel guidance tokens $T_{gf} \in \mathbb{R}^{L\times D}$.

Dataset	Methods	IR$_{\times 10}$ ↑	PS$_{\times 10^2}$ ↑	HPS$_{\times 10^2}$ ↑	AS ↑	PSNR ↑	LPIPS$_{\times 10^3}$ ↓	SSIM$_{\times 10}$ ↑	CLIP-S ↑
EditBench	SDI	-8.72	41.90	21.39	3.79	<u>24.47</u>	28.70	8.50	24.39
	HD-Painter	-6.16	48.64	22.73	3.79	24.08	<u>25.64</u>	8.46	25.62
	PowerPaint	<u>-5.11</u>	<u>50.13</u>	<u>22.94</u>	3.86	24.45	25.69	<u>8.69</u>	26.05
	BrushNet	-7.50	46.37	22.45	3.80	22.71	30.73	8.31	**26.14**
	FreeCond	-7.13	47.16	22.90	3.88	22.38	36.88	8.13	25.15
	Meissonic	-6.66	47.02	22.70	<u>3.90</u>	22.24	35.66	8.31	24.31
	Token Painter (Ours)	**-2.49**	**55.37**	**23.00**	**3.92**	**28.03**	**24.92**	**9.41**	<u>26.06</u>
BrushBench	SDI	11.89	41.30	27.20	4.20	22.82	<u>43.54</u>	7.68	14.44
	HD-Painter	10.81	37.25	26.96	4.21	20.98	49.34	7.61	14.37
	PowerPaint	11.96	42.46	<u>27.65</u>	4.13	<u>23.43</u>	51.46	<u>7.96</u>	<u>14.45</u>
	BrushNet	12.42	41.15	27.51	**4.25**	21.84	47.87	7.59	14.44
	FreeCond	11.97	40.44	27.61	4.21	21.49	50.03	7.43	14.40
	Meissonic	<u>12.51</u>	<u>44.34</u>	27.56	4.21	22.35	64.81	7.72	14.44
	Token Painter (Ours)	**13.01**	**47.90**	**28.36**	<u>4.22</u>	**26.39**	**42.27**	**8.78**	**14.46**

Table 1: Quantitative results on the EditBench and BrushBench datasets. The **best results** and the <u>second best results</u> are marked in bold and underline, respectively. Results are all re-evaluated based on the released code, models and setting.

4.3 Adaptive Decoder Attention Score Enhancing

Adaptive Enhancement Coefficient. To further enhance the prompt detail alignment and content visual quality, we apply adaptive enhancement to two parts of the attention map. Inspired by (Jin et al. 2023), we propose that this coefficient primarily depends on the difference in the number of tokens generated during the training and inference stages. Therefore, we set this coefficient $\alpha = \log_N HW$.

Guided Tokens Enhancement. In the stage of the MAR decoder, guidance tokens T_{gf} and all image tokens I are concatenated together as input $X \in \mathbb{R}^{(L+HW)\times D}$. Then these tokens are projected as queries, keys, values, denoted as $Q, K, V \in \mathbb{R}^{(L+HW)\times D'}$, respectively. And the attention map is defined as $A = \frac{QK^T}{\sqrt{D'}} \in \mathbb{R}^{(L+HW)\times(L+HW)}$. Firstly, we enhance the attention of the inpainting region to guidance tokens, allowing the inpainting content to better align with prompt details. The enhanced attention map is as follows:

$$A'_{ij} = \begin{cases} \alpha^{\lambda_1} \cdot A_{ij} & X_i \in I_p \text{ and } X_j \in T_{gf}, \\ A_{ij} & \text{otherwise}, \end{cases} \tag{10}$$

where λ_1 is a hyperparameter for the power of α.

Dynamic Inpainting Tokens Enhancement. Next, to further enhance the inpainting visual quality, we enhance the attention scores of unknown inpainting tokens $I_{p_1} \in \mathbb{R}^{N_1\times D}$ to the predicted inpainting tokens $I_{p_2} \in \mathbb{R}^{N_2\times D}$, where $N_1 + N_2 = N$. This aims to enable the content of unknown tokens to be more guided by the predicted tokens during the generation process. Unlike guidance tokens, the number of unknown and predicted tokens, *i.e.* N_1 and N_2, dynamically changes at each step. Therefore, we add an extra coefficient $\beta = \log_{N_2+1} N_1$, which dynamically changes at each step. The enhanced attention map is as follows:

$$A'_{ij} = \begin{cases} \beta^{\lambda_3} \cdot \alpha^{\lambda_2} \cdot A_{ij} & X_i \in I_{p_1} \text{ and } X_j \in I_{p_2}, \\ A_{ij} & \text{otherwise}, \end{cases} \tag{11}$$

where λ_2 and λ_3 is the hyperparameters for the power of α and β. At the beginning of the generation, the number of predicted inpainting tokens is small, so they are assigned higher weights. As the number of predicted tokens gradually increases, the weights gradually decrease. Specifically, since there are no predicted tokens at the start of generation, we enhance the attention scores of the entire inpainting region.

5 Experiments

5.1 Experimental Settings

Baseline. We select recent and competitive methods. SDI is a model fine-tuned on random mask inpainting datasets based on stable diffusion (SD) (Rombach et al. 2022). HD-Painter (Manukyan et al. 2023) and FreeCond (Hsiao et al. 2024) are training-free improved methods based on SDI. PowerPaint (Zhuang et al. 2023) and BrushNet (Ju et al. 2024) are fine-tuned on inpainting datasets derived from image segmentation based on SD. Meissonic (Bai et al. 2024) is a MAR model with feature compression layers that employ the naive approach for text-guided inpainting. All diffusion-based methods have been trained on inpainting datasets.

Evaluation Benchmarks. We adopt the two most commonly used text-guided image inpainting benchmarks (Ju et al. 2024; Wang et al. 2023). First, we evaluate on EditBench, which has loose masks for the inpainting objects. Unlike BrushNet, we use the richest captions of the inpainting regions, rather than the annotations of entire images. Next, we evaluate on BrushBench, which contains 600 text-image pairs. Its captions describe the entire images, and the masks are tight, similar to segmentation masks.

Evaluation Metrics. For the choice of metrics, we follow the previous works (Manukyan et al. 2023; Ju et al. 2024), considering three aspects: image visual quality, background region consistency, and text alignment. First, we use metrics aligned with human preferences, including Image Reward (IR) (Xu et al. 2023), HPS v2 (HPS) (Wu et al. 2023), PickScore (PS) (Kirstain et al. 2023), and Aesthetic Score (AS) (Schuhmann et al. 2022). For PickScore, we input both the generated image and the original image, and calculate

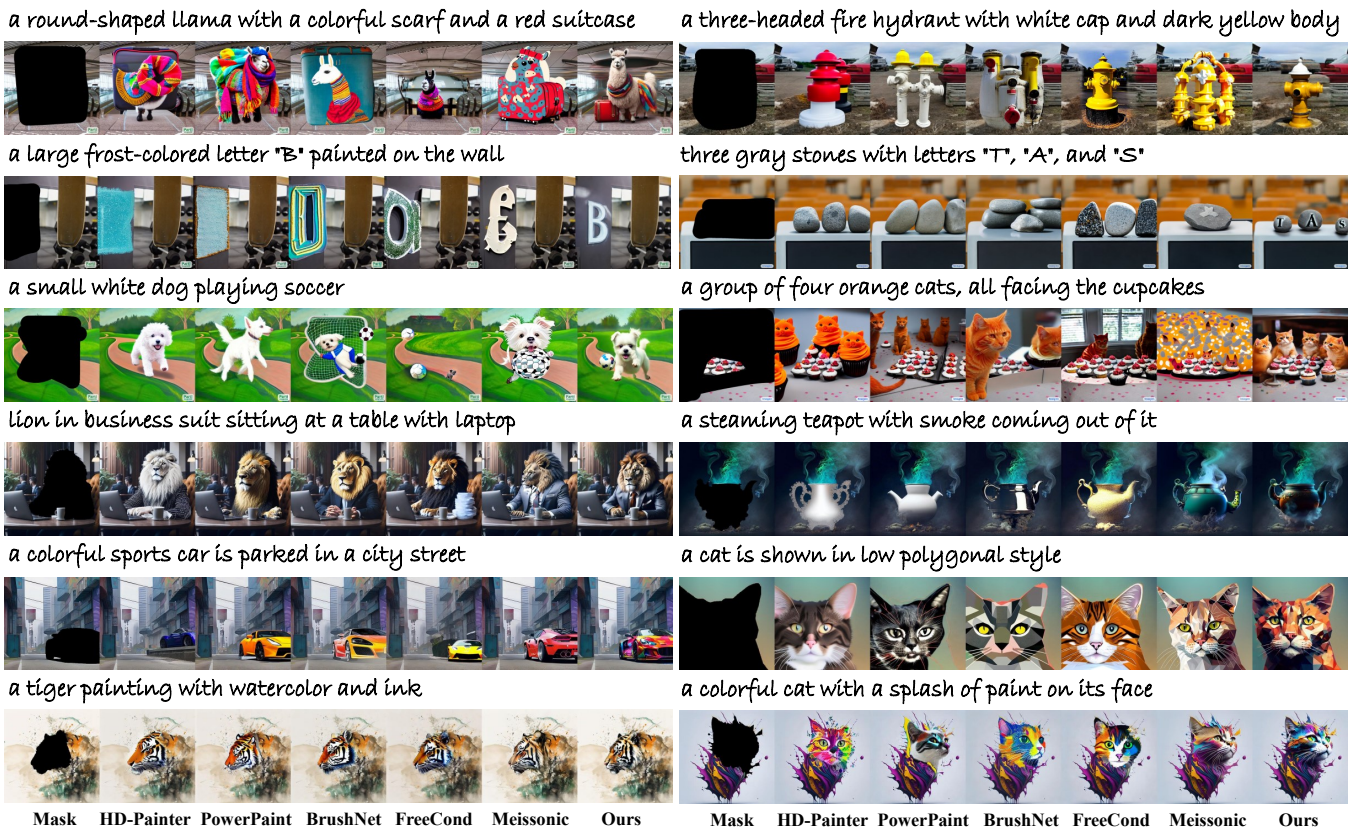


Figure 5: Qualitative results of our Token Painter with previous text-guided inpainting methods. The first three rows of samples are from EditBench with loose masks, and the last three rows of samples are from BrushBench with tight masks.

the average scores for the generated images. Next, for background region consistency, we select Peak Signal-to-Noise Ratio (PSNR) (Korhonen and You 2012), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), and Structural Similarity (SSIM) (Hore and Ziou 2010) to measure the similarity between the generated image and the original image in unmasked regions. Finally, we use CLIP Similarity (CLIP-S) (Wu et al. 2021) to evaluate the text alignment between the generated inpainting content and the text prompt. unlike BrushNet (Ju et al. 2024), we crop the mask region to compute the scores of text alignment.

Implementation Details. To ensure the rigorism of the evaluation, we re-evaluate all baseline methods on the two benchmarks using the officially released code, settings, and models. For fairness, all diffusion-based methods use 0.9B parameter SD-1.5 or SDI-1.5 as base model, and Meissonic use 1B model, while Token Painter uses NOVA-0.6B as base model. In DEIF, the background ratio coefficient p is set to 1, the alignment coefficient a is set to 0.3, and the coefficients controlling the function shape φ and τ are set to 250 and 6. In the ADAE module, the hyperparameters of exponents λ_1, λ_2, and λ_3, are 0.3, 0.1, and 0.03.

5.2 Comparisons with State-of-the-Art

Quantitative Comparison. The Table 1 shows that Token Painter demonstrates highly competitive results, achieving state-of-the-art results on almost all metrics across two benchmarks. FreeCond and HD-Painter are training-free methods improved upon SDI, showing better visual quality on the loose masks of EditBench, while the visual quality of these three methods is almost the same on the BrushBench with tight masks. PowerPaint and BrushNet, based on SD fine-tuning, perform well across baselines. However, due to the inherent limitations of diffusion, they still lag behind Token Painter in local generation visual quality and background consistency. Meissonic is also competitive, but its naive approach and feature compression layers compromise performance, especially in background consistency.

Qualitative Comparison. The qualitative comparison with other inpainting methods is shown in Figure 5, with the basic SDI excluded due to space limitations. Token Painter demonstrates exceptional performance in color, style, structure, and alignment with prompt details. In the first row, Token Painter accurately follows the prompt details, while Other methods either miss or mix elements. In the second row, Token Painter is the only method that generates the correct letter shapes, showcasing its superior local structural control capability. In the sixth row, Token Painter completes missing parts based on image styles, resembling a real painting. In other examples, Token Painter also shows significant improvements in prompt detail alignment and visual content.

Figure 6: Visualization of effects of each component. From left to right, we progressively add each proposed component.

Components	IR$_{\times 10}$ ↑	PS$_{\times 10^2}$ ↑	PSNR ↑	CLIP-S ↑
Baseline	4.23	19.47	26.26	6.42
+DEIF	12.41	44.26	26.35	14.42
+ADAE-G	12.76	46.28	26.27	14.45
+ADAE-I	**13.01**	**47.90**	**26.39**	**14.46**

Table 2: Effects of each component on BrushBench. ADAE-G represents the guidance token enhancement, and ADAE-I is the inpainting token enhancement.

5.3 Ablation Studies

Effects of Each Component. As shown in the Table 2 and Figure 6, we use the T&B approach as the baseline. After adding DEIF, we observe a significant improvement in model performance, particularly in image quality and CLIP score. After adding ADAE-G, the prompt detail alignment of the inpainting content improves further. We then add ADAE-I to enhance the interaction within inpainting tokens, which leads to a further enhancement in the visual quality. We notice that the PSNR remains almost unchanged across different components, as MAR does not alter background tokens.

Effects of Frequency Fusion Function. Table 3 shows the impact of different frequency fusion functions, all of which are centered at L/2 and decrease symmetrically towards both sides. First, we find that the simple linear function introduces too much context information, leading to a decrease in both image quality and alignment with the prompt. Next, we try a constant function, which performs a 0-1 transformation at L/4. While this simple approach improves prompt alignment, the crude frequency stitching results in unsatisfactory visual quality. We then try a quadratic function, which emphasizes semantic information more, but the lack of high-frequency context information limits further improvement in visual quality. Finally, we adopt a modified Gaussian function, which preserves more semantic information in low-frequency and more context information in high-frequency, while smoothly transitioning at L/4. This results in improvements in both visual quality and prompt alignment. PSNR remains largely unchanged due to the MAR property.

Effects of Hyperparameters of Power. The ADAE module includes three hyperparameters, λ_1, λ_2 and λ_3, used to control the exponents of α and β. These hyperparameters are divided into two types. The parameter λ_1 is used to enhance the attention of inpainting tokens towards the guidance token. The parameters λ_2 and λ_3 are used to increase the interaction with inpainting tokens. As shown in the Figure 7, the inpainting region gradually incorporates more details from the prompt as λ_1 increases. When λ_2 and λ_3 increase, the visual quality of the objects in the inpainting region improves, and the structure becomes more coherent. However, in practical applications, we find that excessively large power values lead to distortion or even chaotic colors in inpainting regions. This is likely due to the over-enhancement of attention scores, causing the attention mechanism to fail. Therefore, we ultimately set λ_1, λ_2, and λ_3 to 0.3, 0.1, and 0.03.

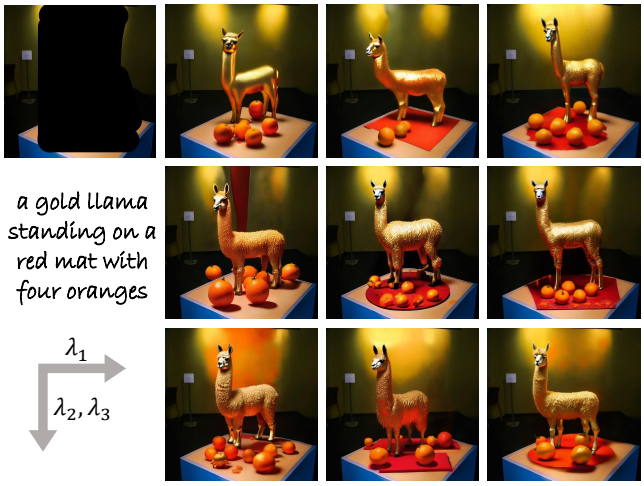


Figure 7: Effects of two types of power hyperparameters, λ_1 and (λ_2, λ_3), varying from 0.1 to 0.3 and (0.03, 0.01) to (0.1, 0.03). The λ_1 primarily controls the prompt detail alignment, while (λ_2, λ_3) enhance the content visual quality.

Functions	IR$_{\times 10}$ ↑	PS$_{\times 10^2}$ ↑	PSNR ↑	CLIP-S ↑
Linear	12.52	44.84	26.25	14.42
Constant	12.71	45.65	26.28	14.46
Quadratic	12.79	46.42	26.27	14.44
M-Gaussian	**13.01**	**47.90**	**26.39**	**14.46**

Table 3: Effects of different functions on BrushBench.

6 Conclusion

In this paper, we improve the T2I MAR model specifically for the text-guided image inpainting task. After conducting analyses of the MAR generation process, we propose a training-free framework, Token Painter. It introduces the Dual-Stream Encoder Information Fusion (DEIF) module at the encoder stage and the Adaptive Decoder Attention Score Enhancing (ADAE) module at the decoder stage. Extensive experiments show that Token Painter outperforms all previous methods, including SOTA methods, across nearly all metrics. We hope that this work will promote the future development of AR models in the image inpainting domain.

Acknowledgments

This project is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62573399.

References

Avrahami; et al. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

Avrahami; et al. 2023. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4): 1–11.

Bai, J.; Ye, T.; Chow, W.; Song, E.; Chen, Q.-G.; Li, X.; Dong, Z.; Zhu, L.; and Yan, S. 2024. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 8.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11315–11325.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

de Jorge, P.; Volpi, R.; Dakania, P. K.; Torr, P. H. S.; and Gregory, R. 2024. Placing Objects in Context via Inpainting for Out-of-distribution Segmentation. In *The European Conference on Computer Vision (ECCV)*.

Deng, H.; Pan, T.; Diao, H.; Luo, Z.; Cui, Y.; Lu, H.; Shan, S.; Qi, Y.; and Wang, X. 2024. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*.

Dong, Q.; Cao, C.; and Fu, Y. 2022. Incremental Transformer Structure Enhanced Image Inpainting With Masking Positional Encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11358–11368.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Fan, L.; Li, T.; Qin, S.; Li, Y.; Sun, C.; Rubinstein, M.; Sun, D.; He, K.; and Tian, Y. 2024. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens, 2024. *URL https://arxiv. org/abs/2410.13863*.

Gao, X.; Xu, Z.; Zhao, J.; and Liu, J. 2024. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1824–1832.

Ho, J.; and Jain, A. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, 2366–2369. IEEE.

Hsiao, T.-F.; Ruan, B.-K.; Tsai, S.-L.; Wu, Y.-L.; and Shuai, H.-H. 2024. Freecond: Free lunch in the input conditions of text-guided inpainting. *arXiv preprint arXiv:2412.00427*.

Jin, Z.; Shen, X.; Li, B.; and Xue, X. 2023. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36: 70847–70860.

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. arXiv:2403.06976.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36: 36652–36663.

Korhonen, J.; and You, J. 2012. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, 37–38. IEEE.

Kwon, J.; Kim, S.; Lin, Y.; Yoo, S.; and Cha, J. 2024. Aesfa: an aesthetic feature-aware arbitrary neural style transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 13310–13319.

Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445.

Li, X.; Guo, Q.; Lin, D.; Li, P.; Feng, W.; and Wnag, S. 2022. MISF: Multi-level Interactive Siamese Filtering for High-Fidelity Image Inpainting. *CVPR*.

Liu, H.; Sun, W.; Di, D.; Sun, S.; Yang, J.; Zou, C.; and Bao, H. 2025. Moee: Mixture of emotion experts for audio-driven portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26222–26231.

Liu, Q.; Tan, Z.; Chen, D.; Chu, Q.; Dai, X.; Chen, Y.; Liu, M.; Yuan, L.; and Yu, N. 2022. Reduce Information Loss in Transformers for Pluralistic Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11347–11357.

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 *The Thirteenth International Conference on Learning Representations*.

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*.

Pang, Z.; Zhang, T.; Luan, F.; Man, Y.; Tan, H.; Zhang, K.; Freeman, W. T.; and Wang, Y.-X. 2025. Randar: Decoder-only autoregressive visual generation in random orders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 45–55.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving Language Understanding by Generative Pre-Training. *https://api.semanticscholar.org/CorpusID:49313245*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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. *Advances in neural information processing systems*, 35: 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

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*.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Sun, W.; Li, X.; Di, D.; Liang, Z.; Zhang, Q.; Li, H.; Chen, W.; and Cui, J. 2025. Uniavatar: Taming lifelike audio-driven talking head generation with comprehensive motion and lighting control. *arXiv preprint arXiv:2412.19860*.

Team, C. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37: 84839–84865.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18359–18369.

Wu, C.; Huang, L.; Zhang, Q.; Li, B.; Ji, L.; Yang, F.; Sapiro, G.; and Duan, N. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*.

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*.

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. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Yang, M.; Lin, S.; Li, C.; and Chang, X. 2025. Let LLM Tell What to Prune and How Much to Prune. In *Proceedings of the 42nd International Conference on Machine Learning*, 70833–70849.

Yu, H.; Luo, H.; Yuan, H.; Rong, Y.; and Zhao, F. 2025. Frequency autoregressive image generation with continuous tokens. *arXiv preprint arXiv:2503.05305*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2023. A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting. arXiv:2312.03594.