Eliminating the Cross-Domain Misalignment in Text-guided Image Inpainting

Muqi Huang[1] , **Chaoyue Wang**[2] , **Yong Luo**[1] and **Lefei Zhang**[1,3*]
[1]Institute of Artificial Intelligence, School of Computer Science, Wuhan University
[2]JD Explore Academy
[3]Hubei Luojia Laboratory
{huangmuqi, luoyong, zhanglefei}@whu.edu.cn, chaoyue.wang@outlook.com

Abstract

Text-guided image inpainting has rapidly garnered prominence as a task in user-directed image synthesis, aiming to complete the occluded image regions following the textual prompt provided. However, current methods usually grapple with issues arising from the disparity between low-level pixel data and high-level semantic descriptions, which results in inpainted sections not harmonizing with the original image (either structurally or texturally). In this study, we introduce a Structure-Aware Inpainting Learning (SAIL) scheme and an Asymmetric Cross Domain Attention (ACDA) to address these cross-domain misalignment challenges. The proposed structure-aware learning scheme employs features of an intermediate modality as structure guidance to bridge the gap between text information and low-level pixels. Meanwhile, asymmetric cross-domain attention enhances the texture consistency between inpainted and unmasked regions. Our experiments show exceptional performance on leading datasets such as MS-COCO and Open Images, surpassing state-of-the-art text-guided image inpainting methods. Code is released at: https://github.com/MucciH/ECDM-inpainting.

1 Introduction

Image inpainting aims to restore damaged images to make them appear realistic and harmonious [Criminisi *et al.*, 2003; Barnes *et al.*, 2009]. The advent of deep learning [Liu *et al.*, 2018; Cao and Fu, 2021; Vaswani *et al.*, 2017; Li *et al.*, 2022a; Suvorov *et al.*, 2022] has brought significant advancements in related fields. Nevertheless, image inpainting, being an ill-posed task, yields unpredictable outcomes that may not align with user expectations. As a response, text-guided image inpainting has emerged [Zhang *et al.*, 2020b; Zhang *et al.*, 2020a], empowering users with control over restoration through textual descriptions, resulting in customized outcomes (Figure 1). Recently, Denoising Diffusion Models [Ho *et al.*, 2020; Rombach *et al.*, 2022; Zhang *et al.*, 2023b], especially the Stable Diffusion (SD) family, have demonstrated outstanding performance across tasks involving image generation, editing, and inpainting.

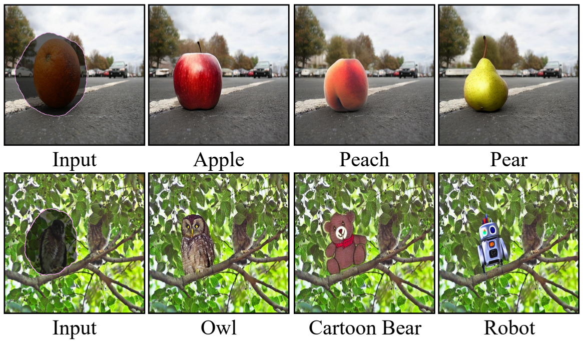


Figure 1: Examples of text-guided image inpainting. Input indicates the masked image. The prompt of the first row is “A/An (*) lying near the white stripe of a highway”, and that of the second row is “There is a/an (*) on dense tree branches”.

However, we find that most text-guided image inpainting methods primarily focus on the semantic alignment between the generated results and the provided textual prompts, yet pay little attention to the authenticity and correctness of the generated images. Therefore, when the task of text-guided image inpainting necessitates the dual objectives of adhering to semantic constraints while also ensuring the creation of visually cohesive images, the majority of current algorithmic models encounter considerable challenges. To a significant extent, these challenges arise from the substantial disparity and misalignment that exists between the guidance provided by textual semantic cues and the fine-grained pixel-level information. These challenges manifest concretely as follows: (a) *Structural Inconsistency*: Despite the semantics of the inpainted image being roughly correct, there could be artifacts at the boundaries of masks, as well as fragmentation of object structures. This problem results from the emphasis on synthesizing content based on text prompts while disregarding the preserved structural information present in the original image. (b) *Texture Detail Disparity*: The reconstructed areas may lack critical high-frequency information, leading to distinct clarity disparities when compared to adjacent retained regions. These issues stem from the underlying misalignment between the high-level semantic information con-

[*] Corresponding author

veyed through textual descriptions and the inherent low-level pixel information in the images.

In this paper, we propose a Structure-Aware Inpainting Learning scheme and an Asymmetric Cross Domain Attention to address the aforementioned misalignment. First, we leverage a pre-trained ControlNet branch as a teacher to add structure guidance during the training process. Specifically, we leverage the edge features extracted from the training image serving as an intermediate modality connecting semantic features of the text prompt with the low-level information of the image. The pre-trained ControlNet branch guides the training process of the inpainting network by supervising the weight updates of the decoding stage, which directs the network's focus toward capturing the overarching semantic structure implicitly, yielding results that exhibit integrated structural coherence. Within this architecture, we introduce an Asymmetric Cross Domain Attention module, where the masked input is mapped to the frequency domain through Fourier transformation. Subsequently, high-pass filtering is employed to extract high-frequency components before computing the similarities of the objective in the image domain with features in the frequency domain and the textual domain, respectively. This process facilitates the fusion of high-frequency image information with text features, which dynamically influences the generated results to intricately restore texture details while ensuring semantic correctness. The contributions of this work are summarized as follows:

- We devise a Structure-Aware Inpainting Learning mechanism, integrating a structure-guided pre-trained controlnet branch, to direct the network to consciously restore comprehensive semantic information.
- We introduce an Asymmetric Cross Domain Attention module that employs attention on high-frequency features and textual prompts, enabling the model to emphasize and incorporate the texture information.
- Experimental results conducted on two prominent datasets MS-COCO and Open Images demonstrate the superiority of our method over the state-of-the-art text-guided inpainting techniques.

2 Related Work

2.1 Image Inpainting

A multitude of deep learning-based methodologies have been explored for tackling image inpainting tasks, frequently adopting frameworks such as Generative Adversarial Networks (GANs) [Goodfellow *et al.*, 2020] or encoder-decoder architectures. Certain approaches have decomposed the task into multiple stages [Yu *et al.*, 2019; Huang and Zhang, 2022], employing sequential networks to gradually restore coarse elements and intricate textures within images. To ensure the coherence of the whole image, attention mechanisms have gained prominence to align the generated content with the surrounding context. RFR [Li *et al.*, 2020] adopts an iterative strategy to propagate inpainting and calculate attention score from mask edges inward. LBAM [Sun *et al.*, 2021] introduces bidirectional attention maps for enhanced long-range contextual modeling. However, the filled content often lacks control and exhibits significant randomness, leading to uncertain quality of reconstructed images that might not fulfill user expectations in practical applications.

2.2 Text-guided Image Inpainting

To offer more controlled restoration options, text-guided image inpainting methods enable users to input both masks and accompanying text, facilitating the reconstruction of missing content based on textual cues. LSAI [Xie *et al.*, 2022] employs semantic alignment to repair patches of the damaged image. ALMR [Wu *et al.*, 2021] employs adversarial learning and introduces a dual-attention module for both coarse- and fine-grained stages, spanning from semantic to textural image recovery. MMFL [Lin *et al.*, 2020] introduces a multimodal fusion approach that leverages both textual and image information. However, these methods often fall short in terms of image diversity, struggling to provide ample choices when confronted with low-quality generations.

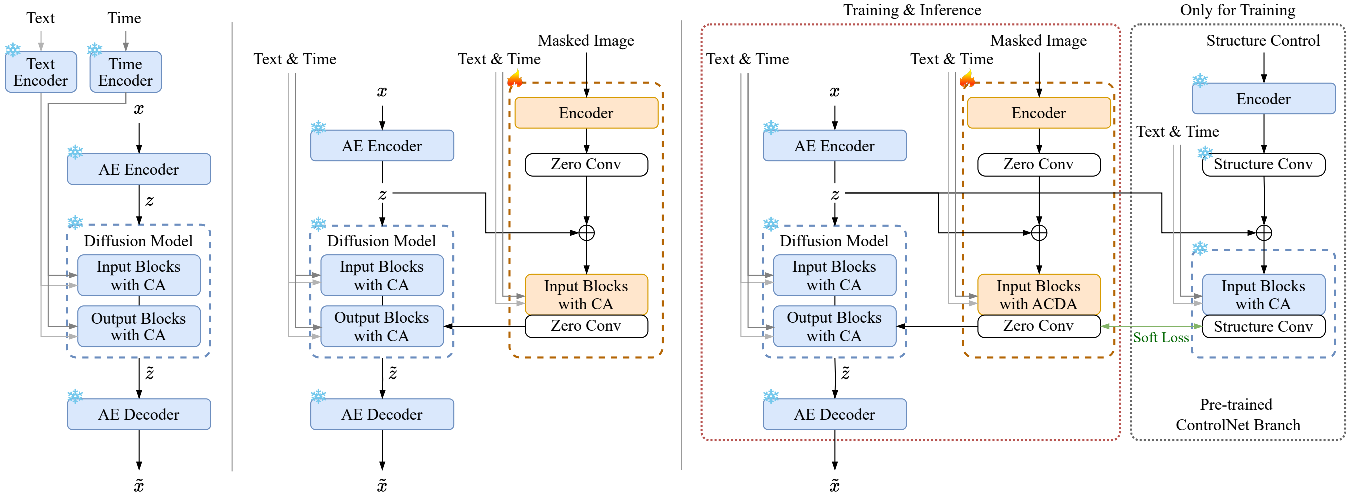
2.3 Diffusion Model in Image Inpainting

The diffusion model has emerged as a prominent framework for generating high-quality images in a controlled manner. Originating from DDPM [Ho *et al.*, 2020], this paradigm involves iteratively diffusing noise levels to transform latent representations into images. By modulating noise diffusion, this approach elegantly balances the trade-off between image generation and control. The inherent attributes of diffusion models also find resonance in the domain of image restoration. SDM [Li *et al.*, 2022b] introduces a spatial diffusion model to restore images with large holes, while RePaint [Lugmayr *et al.*, 2022] employs fine-tuning through resampling based on pre-trained DDPM, enhancing image realism. Furthermore, EditBench [Wang *et al.*, 2023], Blended Diffusion [Avrahami *et al.*, 2022], GLIDE [Nichol *et al.*, 2022], and SmartBrush [Xie *et al.*, 2023] propose diffusion-based text-guided image editing methods, producing controlled high-quality images that are also applicable to inpainting tasks. To maintain the high performance and variability of DDPM under limited computational resources, LDM (Stable Diffusion) [Rombach *et al.*, 2022] incorporates a pre-trained autoencoder, enabling the diffusion model to operate on smaller latent feature maps. Building upon this, ControlNet [Zhang *et al.*, 2023b] introduces additional control branches, exclusively training control branches while keeping the weights of Stable Diffusion branches fixed, offering a more flexible approach to image generation.

These methods inherit the iterative nature of diffusion, enabling them to adaptively restore missing regions while adhering to textual cues. However, while augmenting the image restoration with guidance has its merits, it also presents new challenges. Due to gaps between different modalities, there can be misalignment or even conflicts between the textual information and the preserved image details.

3 Preliminary: Diffusion Model

In this paper, we employ the diffusion model as our generator. The Denoising Diffusion Probabilistic Models (DDPM) consist of a forward process and a reverse process, with the former being also referred to as the diffusion process.



(a) LDM for Inpainting (b) ControlNet for Inpainting (c) Ours

Figure 2: Structure-Aware Inpainting Learning scheme. (a) and (b) are Latent Diffusion Model (LDM) and ControlNet with cross attention (CA) for text-guided image inpainting. The snowflakes signify fixed weights, while the flames mean the modules are trainable.

The diffusion process involves incrementally adding Gaussian noise to the data until it becomes random noise. For the original data $x_0 \sim q(x_0)$, each step of the diffusion process generates a sample $x_t \sim q(x_t|x_0)$ by iteratively adding Gaussian noise to the previous step's data over T steps, ultimately resulting in a random noisy image $x_t \sim q(x_t|x_0)$ that completely loses the original information:

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

where $\{\beta_t\}_{t=1}^{T}$ represents the variance at each noise addition step.

Given $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, by reparameterization we have:

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

The variable x_t can be regarded as a linear combination of the original data x_0 and random noise ϵ, where $\sqrt{\bar{\alpha}_t}$ and $\sqrt{1-\bar{\alpha}_t}$ act as the blending coefficients, with their squares summing up to 1.

The reverse process, on the other hand, is a denoising process. It starts with a noisy image $x_T \sim \mathcal{N}(0, I)$ at step T and gradually denoises it using known real distributions $q(x_{t-1}|x_t)$ for each step:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3}$$

Here, $p(x_T) = \mathcal{N}(x_T; 0, I)$, and $p_\theta(x_{t-1}|x_t)$ represents a parameterized Gaussian distribution, with its mean and variance determined by the trained networks $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$. After reparameterizing , μ_θ can be expressed as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right), \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, I)$, and ϵ_θ is a fitting function based on neural networks, implying the use of predicted noise ϵ instead of predicted mean. The simplified optimization objective is:

$$L_{t-1}^{\text{simple}} = E_{x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta\|^2\right]. \tag{5}$$

4 Method

Given an original image I_{gt}, a binary mask m, and a text prompt c, the masked image I_{input} is obtained by overlaying the mask on the original image.

$$I_{input} = I_{gt} \odot (1 - m), \tag{6}$$

The corresponding latent feature z_0 is derived through encoding. Passing through the overall network yields the latent feature map $\tilde{z}$, which is subsequently decoded into the final output I_{output}. The generation of $\tilde{z}_{t-1}$ and the composition with restored region in each denoising step is as follows:

$$\tilde{z}_{t-1} \sim \mathcal{N}(\mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \tag{7}$$

$$\tilde{z}_{t-1} = z_{t-1} \odot (1 - m) + \tilde{z}_{t-1} \odot m. \tag{8}$$

In this work, we introduce a novel Structure-Aware Inpainting Learning scheme that offers semantic guidance to the inpainting network, enabling the generation of structurally coherent images. Furthermore, we devise an Asymmetric Cross Domain Attention module that blends high-frequency texture details from the preserved region of the image with text embeddings, which dynamically constrains the generation of texture details.

4.1 Structure-Aware Inpainting Learning Scheme

To narrow the disparity between the semantic content of the text and the pixel-level information in the image, we introduce the use of image structure as an intermediary between the two domains. This strategy associates high-level textual features with low-level image features, aiding the inpainting network to implicitly focus on global image semantics during training, thus leading to the generation of cohesive images in accordance with textual descriptions.

The training strategy is shown as Figure 2. (a) depicts a simplified diagram of the latent diffusion model, where images are encoded into the latent space through an AutoEncoder, reducing the feature map dimensions. Training

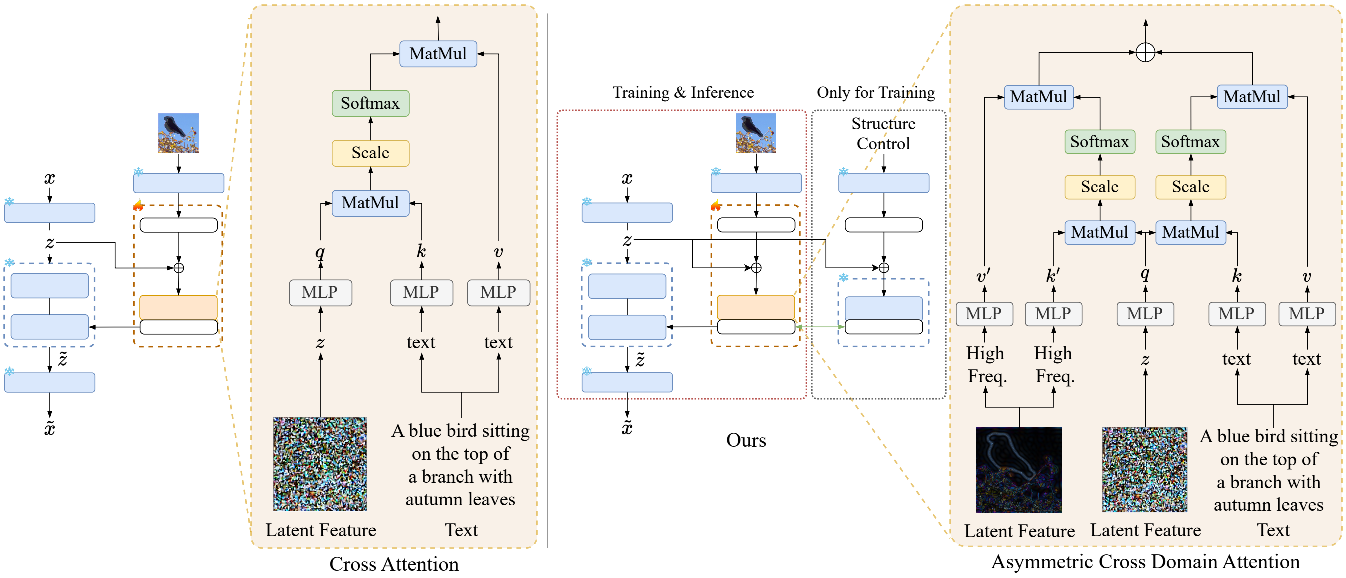


Figure 3: The Asymmetric Cross Domain Attention mechanism in trainable input blocks.

of the diffusion model then occurs within the latent feature space, utilizing a U-net architecture for the generator network. The output of the diffusion model is finally mapped back into image space through a decoder. Figure 2(b) illustrates ControlNet with the LDM. It introduces a control branch between the input x and output blocks of the diffusion model. The condition signal added with noise z is propagated through input blocks, and each input block is processed with a zero-initialized convolution, effectively connecting with the main LDM. This enables ControlNet to accept various types of conditional signals, such as Canny edge and segmentation maps, thereby flexibly guiding the generation process of the latent diffusion model without compromising the performance of the pre-trained Stable Diffusion model.

Figure 2(c) showcases our training scheme. We employ a pre-trained stable diffusion model as the backbone of the generation process. The inpainting branch receives the concatenation of the masked image and mask as inputs, which are then passed through trainable input blocks and zero convolutions before being fed back into the main network. Notably, a pre-trained ControlNet, employing structure information as a condition, is used as supervision during network training. This is achieved by computing the feature differences of each output block in the inpainting branch and the ControlNet branch using a soft loss mechanism, implicitly guiding the network to focus more on image structural information. "Zero Conv" refers to 1×1 convolution initialized with zero, while "Structure Conv" stands for the zero-initialized convolution that has already been activated. The computation of the soft loss involves employing an intuitive mean error loss:

$$\mathcal{L}_{soft} = \left\| f_{control} - f_{inpaint} \right\|_1 , \tag{9}$$

where $f_{control}$ is the feature map during the structure convolution block of the ControlNet, and $f_{inpaint}$ is that of the inpainting network branch.

4.2 Asymmetric Cross Domain Attention

In the context of inconsistencies between regions of image reconstruction, beyond discrepancies in content and structure, differences in image clarity and resolution also exist. Many existing text-guided image inpainting algorithms predominantly focus on adhering to textual prompts while disregarding the coherence between the reconstructed objects and the preserved regions. To address this issue, we further propose a network architecture based on Asymmetric Cross Domain Attention, facilitating the interaction between information from the image's spatial domain, frequency domain, and the textual domain, to integrate the impact of high-frequency image texture information into text-guided object generation. The Asymmetric Cross Domain Attention calculation is as follows:

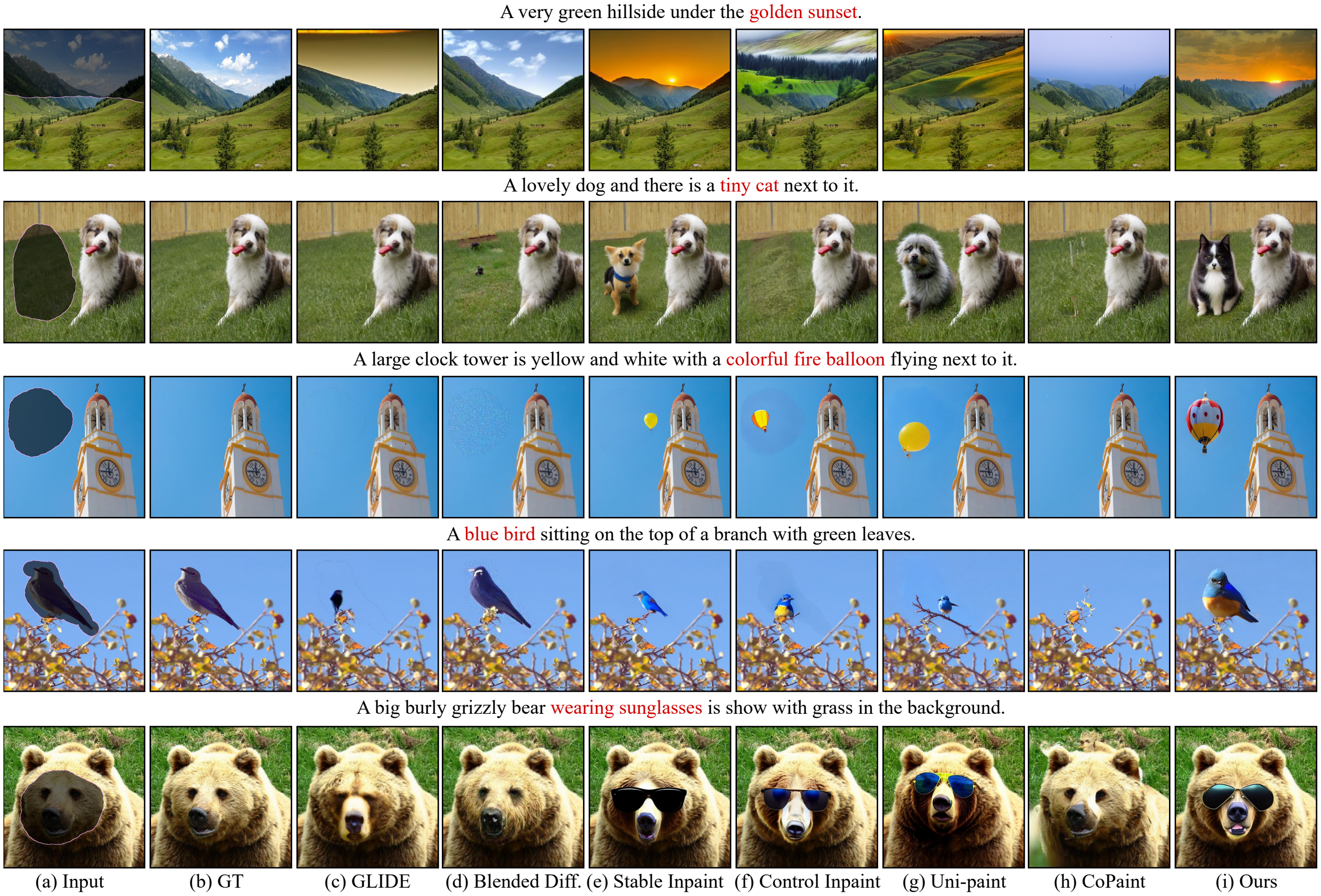
$$Attn = \mathrm{softmax}(\frac{qk^\top}{\sqrt{d}} + b)v + \mathrm{softmax}(\frac{q'k'^\top}{\sqrt{d}} + b)v, \tag{10}$$

where $Attn$ denotes $Attention(q, k, v, k', v')$, d represents the dimension of the input vector, and b is the bias.

The overall framework of the network is illustrated in Figure 3. The text prompt is encoded through CLIP and successively introduced into the transformer blocks within the U-net structure of the diffusion model. This process involves the integration of Asymmetric Cross Domain Attention with the high-frequency feature map of the image encoded into the latent feature space, as depicted on the right side of Figure 3. The high-frequency feature map I_h is obtained by transforming the masked image into the frequency domain using Fourier transformation, followed by extraction via a high-pass filter with a radius of 4. Then we get the feature f_{freq} in frequency domain by:

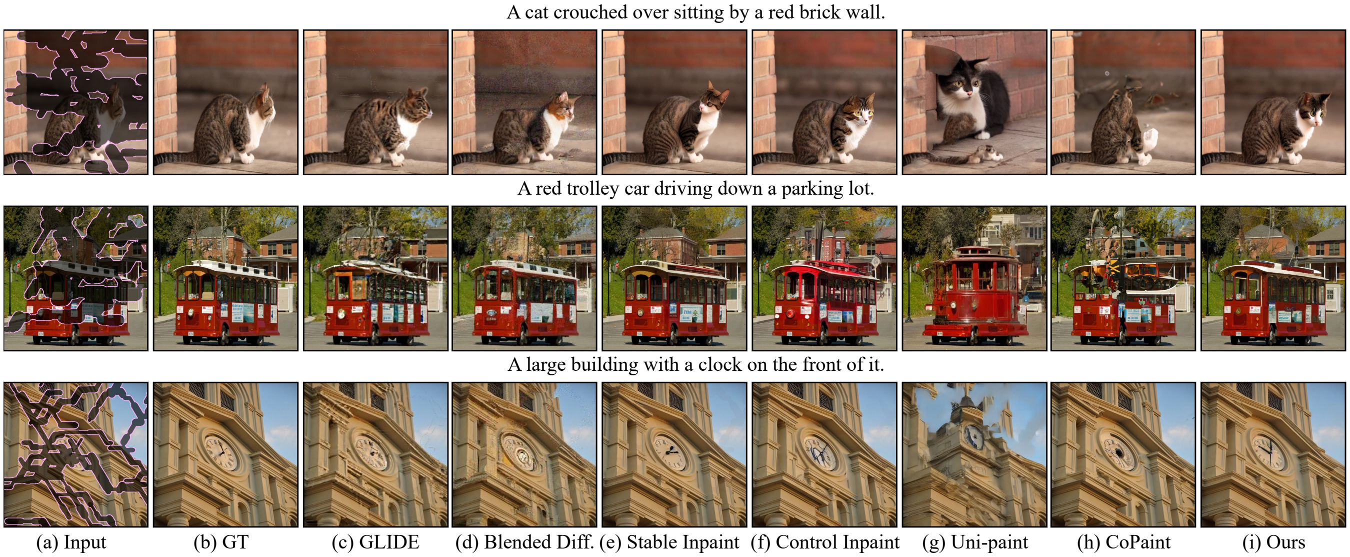
$$f_{freq} = Concat(Norm(\mathcal{F}_{abs}(I_h)) + \mathcal{F}_{angle}(I_h)), \tag{11}$$

where $Concat$ denotes concatenate, $Norm$ means normalization, $\mathcal{F}_{abs}(\cdot)$ and $\mathcal{F}_{angle}(\cdot)$ are used to compute the ampli-


A very green hillside under the golden sunset.
A lovely dog and there is a tiny cat next to it.
A large clock tower is yellow and white with a colorful fire balloon flying next to it.
A blue bird sitting on the top of a branch with green leaves.
A big burly grizzly bear wearing sunglasses is show with grass in the background.

(a) Input (b) GT (c) GLIDE (d) Blended Diff. (e) Stable Inpaint (f) Control Inpaint (g) Uni-paint (h) CoPaint (i) Ours

Figure 4: Qualitative Comparison over datasets MS-COCO and Open Images with customized masks and prompts. (a) is the input corrupted image, (b) is the ground truth, (c)-(h) are results of other approaches. (i) is the result of the proposed method.


A cat crouched over sitting by a red brick wall.
A red trolley car driving down a parking lot.
A large building with a clock on the front of it.

(a) Input (b) GT (c) GLIDE (d) Blended Diff. (e) Stable Inpaint (f) Control Inpaint (g) Uni-paint (h) CoPaint (i) Ours

Figure 5: Qualitative Comparison over datasets MS-COCO with wide/narrow masks. (a) is the input corrupted image, (b) is the ground truth, (c)-(h) are results of other approaches. (i) is the result of the proposed method.

Datasets		MS-COCO				Open Images			
Mask Ratio	Methods	PSNR ↑	SSIM ↑	Mean l_1 ↓	LPIPS ↓	PSNR ↑	SSIM ↑	Mean l_1 ↓	LPIPS ↓
10%-20%	GLIDE	27.4670	0.8720	0.0223	0.0523	27.9267	0.8780	0.0212	0.0530
	Blended Diff.	27.5872	0.8683	0.0226	0.0523	28.0664	0.7784	0.0377	0.1085
	Stable Inpaint	29.2245	*0.9321*	0.0139	0.0228	29.7867	*0.9359*	0.0133	0.0211
	Control Inpaint	*28.9814*	0.9333	*0.0140*	*0.0238*	*29.4835*	0.9374	0.0133	*0.0220*
	Uni-paint *	21.8992	0.8701	0.0287	0.0981	22.3254	0.8807	0.0262	0.0925
	CoPaint	28.3742	0.9244	0.0154	0.0280	28.9633	0.9336	*0.0137*	0.0258
	Ours	**29.8000**	**0.9379**	**0.0132**	**0.0203**	**30.3400**	**0.9415**	**0.0126**	**0.0187**
30%-40%	GLIDE	21.1245	0.6892	0.0462	0.1590	23.3936	0.7811	0.0337	0.1100
	Blended Diff.	23.6027	0.7609	0.0350	0.1218	*24.1145*	0.7715	0.0334	0.1171
	Stable Inpaint	*23.5053*	0.8366	0.0281	0.0734	24.1427	0.8461	0.0266	0.0676
	Control Inpaint	22.7211	*0.8315*	*0.0303*	*0.0826*	23.3488	*0.8414*	*0.0286*	*0.0760*
	Uni-paint *	16.3802	0.6877	0.0733	0.2489	16.6082	0.6931	0.0711	0.2503
	CoPaint	22.4218	0.8208	0.0313	0.0894	22.9441	0.8287	0.0297	0.0851
	Ours	**23.6437**	**0.8482**	**0.0271**	**0.0696**	**24.1812**	**0.8566**	**0.0258**	**0.0643**
50%-60%	GLIDE	21.1304	0.6893	0.0461	0.1591	21.5083	0.6972	0.0448	0.1587
	Blended Diff.	*21.6013*	0.6567	0.0481	0.1823	*21.9916*	0.6680	0.0465	0.1763
	Stable Inpaint	21.8510	0.7649	0.0387	0.1107	22.2850	0.7743	0.0376	0.1040
	Control Inpaint	21.0475	*0.7524*	*0.0427*	*0.1247*	21.4741	*0.7633*	*0.0410*	*0.1165*
	Uni-paint *	14.0914	0.5203	0.1182	0.3904	14.3470	0.5216	0.1153	0.3922
	CoPaint	20.6561	0.7390	0.0435	0.1374	20.9275	0.7480	0.0425	0.1329
	Ours	**22.0327**	**0.7804**	**0.0373**	**0.1039**	**22.4339**	**0.7889**	**0.0363**	**0.0976**

Table 1: Quantitative comparisons with the state-of-the-art approaches, namely GLIDE, Blended Diffusion, Stable Diffusion inpainting, ControlNet inpainting, Uni-paint, CoPaint, over datasets MS-COCO and Open Images with different ratios of random mask. ↑ denotes higher is better. ↓ denotes lower is better. The optimum result is highlighted in **bold**, the second-ranked result is underlined, and the third-ranked result is *italicized*. Uni-paint marked with * is only effective for inpainting with customized masks.

tude and phase, respectively. We concatenate the amplitude spectrum and phase spectrum of the high-frequency component along the channel dimension, presenting them in a frequency domain form for input into the Asymmetric Cross Domain Attention module.

The overall optimization objective is

$$\mathcal{L}_{total} = L_{t-1}^{\text{simple}} + \lambda \mathcal{L}_{soft}. \tag{12}$$

where λ is a hyper-parameter controlling the weight of the soft loss and is set to 0.1 in our experiments.

5 Experiments

5.1 Experiment Settings

We employ our proposed Structure-Aware Inpainting Learning (SAIL) approach for image inpainting under the architecture of ControlNet and it is finetuned from Controlnet - v1.1 - InPaint Version. The learning rate is set at $5e^{-5}$, and the batch size is configured to be 4. Each experiment necessitate the utilization of one A100 GPU. After a series of systematic experiments, it has been shown that choosing Canny edge as the structure information of the auxiliary ControlNet branch during training can effectively correlate high-level semantic information with low-level pixel details. Therefore, all subsequent experimental setups utilize Canny edge as the condition for the ControlNet branch.

Datasets. We fine-tune our model on the standard MS-COCO dataset [Lin *et al.*, 2014], which comprises over 100k images in the training set. For testing, we utilize 5k image-text pairs from the MS-COCO validation set. To assess the robustness of our model to diverse data, we further validate its performance on 1.5k images from the Open Images dataset [Kuznetsova *et al.*, 2020].

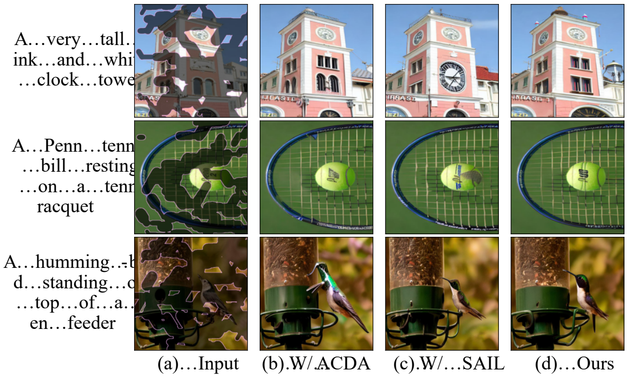


Figure 6: Visualization of ablation study.

Baselines. We select six state-of-the-art approaches based on the diffusion model as our baselines: GLIDE [Nichol *et al.*, 2022], Blended Diffusion [Avrahami *et al.*, 2022], Stable Diffusion inpainting [Rombach *et al.*, 2022], ControlNet inpainting [Zhang *et al.*, 2023b], Uni-paint [Yang *et al.*, 2023], and CoPaint [Zhang *et al.*, 2023a].

Metrics. We compare the proposed method with other approaches on four widely-adopted metrics, namely PSNR, SSIM, Mean l_1, and LPIPS, to evaluate the clarity, structure similarity, and diversity of the results comprehensively.

5.2 Qualitative Comparison

We conduct image inpainting tests with various mask ratios on both the MS-COCO dataset and the Open Images dataset. As depicted in Figure 4, for custom masks and prompt inputs, our proposed method generates high-quality images that align both internally and externally with the textual descriptions. In contrast, GLIDE and Blended Diffusion often fail to generate content aligned with textual descriptions. While the generated results of Stable Diffusion, ControlNet, and Unipaint incorporate textual descriptions, the overall semantic coherence of the images may be compromised, resulting in artifacts or noticeable boundaries along the mask edges. CoPaint is the only comparative method that performs inpainting without text guidance. It can produce reasonable results but falls short in meeting user-customized content generation demands. Figure 5 showcases the comparative results of our approach against other methods on the MS-COCO dataset using randomly generated masks. It is evident that our method consistently produces coherent and semantically sound clear images, both for wide and narrow masks. In contrast, other methods struggle to maintain the structural consistency when dealing with objects that are partially occluded, such as the cat in the first row of the figure. Note that Uni-paint, being a task-specific training-free approach, relies solely on the generative capacity inherited from the pre-trained stable diffusion model. While it performs well with regularly shaped masks, it exhibits a lack of robustness in situations involving free-form masks or cases where the mask does not entirely cover the object.

5.3 Quantitative Comparison

We conduct quantitative evaluations on both the MS-COCO and Open Images validation sets, selecting 5k and 1.5k images respectively. We apply same irregular masks to different methods. Table 1 presents numerical outcomes with various mask ratios. As observed, the images generated by our proposed approach outperform other SOTA methods across all metrics, objectively showcasing our method's superiority in terms of image structure, clarity, and diversity.

5.4 Ablation Study

In this section, we meticulously validate the efficacy of the proposed Asymmetric Cross Domain Attention module (ACDA) and the Structure-Aware Inpainting Learning (SAIL) scheme. The baseline is the naive fine-tuning scheme as ControlNet with masked image as condition. The notation "w/ ACDA" signifies the adoption of solely the ACDA module without utilizing the pre-trained ControlNet model with Canny edge as structural guidance, while "w/ SAIL" denotes the exclusive use of the SAIL training mode, substituting the ACDA module with a standard cross attention mechanism.

As discerned from Figure 6, employing only the ACDA module yields images with distinct texture details, but semantic structural inaccuracies are observed. Conversely, employing solely the SAIL strategy enhances the overall structural coherence of the images, but at the expense of losing high-frequency information in the reconstructed missing regions. Table 2 objectively quantifies the effectiveness of our proposed network architecture and training methodologies.

Model	PSNR ↑	SSIM ↑	Mean l_1 ↓	LPIPS ↓
Baseline	21.6907	0.7569	0.0417	0.1594
w/ SAIL	21.7218	0.7764	0.0383	0.1072
w/ ACDA	21.9833	0.7784	0.0377	0.1086
Ours	**22.0327**	**0.7804**	**0.0373**	**0.1039**

Table 2: Ablation study on MS-COCO with mask ratio 50%-60%. ↑ denotes higher is better. ↓ denotes lower is better.

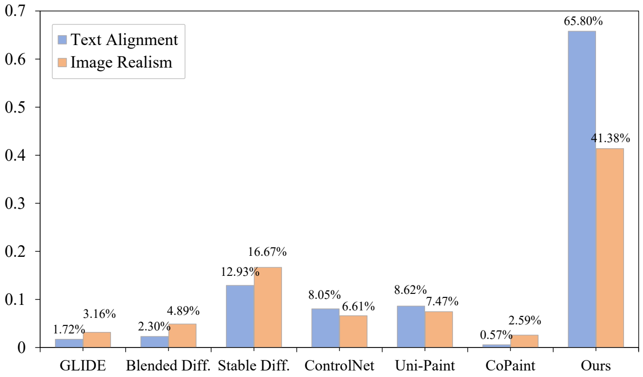


Figure 7: User study. We have 348 samples voted over text alignment and image realism of the inpainted results of different methods. Our approach outperforms the baselines by a large margin.

5.5 User Study

To investigate human perception of generated results, we conducted user study with a random selection of 29 participants. A total of 348 answers for several image-text pairs were presented. They were asked two questions for each pair, focusing on the generated images from GLIDE, Blended Diff., Stable Diff., ControlNet, Uni-paint, CoPaint and our method: 1) Which image best matches the textual description? 2) Which result appears the most realistic and natural?

The results of the study, as shown in Figure 7, demonstrate significant advantages of our method in terms of textual fidelity and image realism.

6 Conclusion

In this paper, we propose Structure-Aware Inpainting Learning scheme for text-guided image inpainting, aiming at enhancing the efficacy of conditioned image synthesis model based on LDM for inpainting tasks. By leveraging Canny edge as intermediate modality, our method offers crucial structural guidance to the inpainting network training procedure. Moreover, we introduce an Asymmetric Cross Domain Attention mechanism within the diffusion model to achieve harmonious alignment between and high-frequency image textures. This strategic fusion ensures both semantic consistency in composition and retention of fine-grained texture details. Experimental results validate the capability of our approach to reconstruct globally semantically coherent high-quality images across different datasets, achieving remarkable performance compared to state-of-the-art diffusion model-based methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62122060 and the Special Fund of Hubei Luojia Laboratory under Grant 220100014.

References

[Avrahami *et al.*, 2022] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18187–18197, 2022.

[Barnes *et al.*, 2009] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[Cao and Fu, 2021] Chenjie Cao and Yanwei Fu. Learning a sketch tensor space for image inpainting of man-made scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14489–14498, 2021.

[Criminisi *et al.*, 2003] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 721–728, 2003.

[Goodfellow *et al.*, 2020] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33*, 2020.

[Huang and Zhang, 2022] Muqi Huang and Lefei Zhang. Atrous pyramid transformer with spectral convolution for image inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4674–4683, 2022.

[Kuznetsova *et al.*, 2020] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[Li *et al.*, 2020] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7757–7765, 2020.

[Li *et al.*, 2022a] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10758, 2022.

[Li *et al.*, 2022b] Wenbo Li, Xin Yu, Kun Zhou, Yibing Song, Zhe Lin, and Jiaya Jia. SDM: spatial diffusion model for large hole image inpainting. *arXiv preprint arXiv:2212.02963*, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proceedings of European Conference on Computer Vision*, volume 8693, pages 740–755, 2014.

[Lin *et al.*, 2020] Qing Lin, Bo Yan, Jichun Li, and Weimin Tan. MMFL: multimodal fusion learning for text-guided image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1094–1102, 2020.

[Liu *et al.*, 2018] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision*, pages 85–100, 2018.

[Lugmayr *et al.*, 2022] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11451–11461, 2022.

[Nichol *et al.*, 2022] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning*, volume 162, pages 16784–16804, 2022.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022.

[Sun *et al.*, 2021] Liujie Sun, Qinghan Zhang, Wenju Wang, and Mingxi Zhang. Image inpainting with learnable edge-attention maps. *IEEE Access*, 9:3816–3827, 2021.

[Suvorov *et al.*, 2022] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3172–3182. IEEE, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008, 2017.

[Wang *et al.*, 2023] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023.

[Wu *et al.*, 2021] Xingcai Wu, Yucheng Xie, Jiaqi Zeng, Zhenguo Yang, Yi Yu, Qing Li, and Wenyin Liu. Adversarial learning with mask reconstruction for text-guided image inpainting. In *Proceedings of the 29th ACM Multimedia Conference*, pages 3464–3472, 2021.

[Xie *et al.*, 2022] Yucheng Xie, Zehang Lin, Zhenguo Yang, Huan Deng, Xingcai Wu, Xudong Mao, Qing Li, and Wenyin Liu. Learning semantic alignment from image for text-guided image inpainting. *Vis. Comput.*, 38(9):3149–3161, 2022.

[Xie *et al.*, 2023] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023.

[Yang *et al.*, 2023] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-Paint: a unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 3190–3199, 2023.

[Yu *et al.*, 2019] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4470–4479, 2019.

[Zhang *et al.*, 2020a] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1302–1310, 2020.

[Zhang *et al.*, 2020b] Zijian Zhang, Zhou Zhao, Zhu Zhang, Baoxing Huai, and Jing Yuan. Text-guided image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4079–4087, 2020.

[Zhang *et al.*, 2023a] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. In *Proceedings of the International Conference on Machine Learning*, pages 41164–41193, 2023.

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