MMGInpainting: Multi-Modality Guided Image Inpainting Based On Diffusion Models

Cong Zhang , Wenxia Yang , Xin Li , *Fellow, IEEE*, Huan Han

***Abstract*—Proper inference of semantics is necessary for realistic image inpainting. Most image inpainting methods use deep generative models, which require large image datasets to predict and generate content. However, limited control makes predicting the missing regions and generating coherent content difficult. Existing approaches include image-guided or text-guided image inpainting, but none of them has taken both image and text as the guidance signals, as far as we know. We propose a multi-modality guided (MMG) image inpainting approach based on the diffusion model to fill this gap. This MMGInpainting method uses image and text as guidance for generating content within the target area for inpainting, effectively integrating the semantic information conveyed by the guiding image or text into the content of the inpainted region. To construct MMGInpainting, we start by enhancing the U-Net backbone with a customized Nonlinear Activation Free Network (NAFNet). This adapted NAFNet incorporates an *Anchored Stripe Attention* mechanism, which utilizes anchor points to model global contextual dependencies effectively. To regulate inpainting, we use a Semantic Fusion Encoder to guide the inverse process of the diffusion model. The process is iteratively executed to denoise and generate the desired inpainting result. Additionally, we explore how different modes of meaning interact and coordinate to offer users helpful guidance for a more manageable inpainting procedure. Experimental results demonstrate that our approach produces faithful results adhering to the guiding information while significantly improving computational efficiency. Github Repository: https://github.com/skipper-zc/MMGInpainting/**

***Index Terms*—Image inpainting, multi-modality guidance, diffusion models, NAFNet, controllable inpainting.**

I. Introduction

IMAGE inpainting is a computer vision technique that aims to intelligently fill in the missing parts of an image seamlessly and coherently. Its primary objective is to generate visually plausible content with reasonable semantics that harmonizes with the surrounding regions, which not only focuses on the image's appearance but also considers the semantic content to ensure that the inpainted areas make sense within the overall image. At present, the most advanced methods are mainly based on Generative Adversarial Networks(GANS) [1] [2] [3], or diffusion models [4] [5] [6] [7] [8]. Among them, GAN-based inpainting models typically involve the construction of a generator that takes incomplete images as input. The generator is then trained to produce reasonable content for the missing regions, while concurrently, the discriminator is trained to differentiate between the generated images and the ground truth images. This approach guides the generator toward generating progressively more realistic and persuasive content.

In contrast, inpainting methods based on diffusion models operate by generating images from noise and iteratively removing noise to approximate the data distribution of natural images. These models seek to predict the content of missing regions and generate completed images through the assimilation of knowledge from an extensive dataset [5] [9] [10]. Pioneering diffusion-based generative models, such as Denoising Diffusion Probabilistic Models (DDPM) [9], Denoising Diffusion Implicit Models (DDIM) [11], and Latent Diffusion Models (LDM) [6], have demonstrated their superiority over state-of-the-art GAN-based methods in image synthesis [4]. Consequently, they have found widespread application in various vision tasks, including image editing [12] [13], image translation [14], and text-to-image synthesis [15]. DDPM progressively introduces noise into an image until it reaches a state of pure noise. Subsequently, it samples from this stochastic noise distribution and applies a predetermined number of denoising steps iteratively, culminating in the final image sample [16]. DDPM employs the U-Net architecture to predict the noise added in the forward process, allowing the model to learn the potential distribution of target image data. By gradual denoising, it generates image samples that conform to the distribution of the target image [17] [18].

As a classical inverse problem, one challenge in image inpainting is that the outcomes often lack diversity, and visual quality can significantly degrade when missing regions are large [19]. LaMa [20] perceives that both the inpainting network and the loss function lack effective receptive fields; it introduces a novel inpainting network architecture that employs Fast Fourier Convolution (FFC) with a fully optimized graph receptive field, guided by a high-receptive field-aware loss. FT-TRD [21] aims to detect corrupted areas in face images and generate visually reasonable content within the masked regions. For the joint inpainting task with other tasks (such as mosaic removal), the staged training often performs better [22] [23].

Another major challenge is the **lack of guidance and controllability** over results, which restricts the practicality of image inpainting in achieving desired outcomes [24] [25]. To address this, several methods attempt to introduce additional prompts [26] [6] for fine-grained control over generated images. Previous work typically incorporated text [27] [28] [29],

Manuscript received October 18, 2023; This work is partially supported by the National Natural Science Foundation of China under Grants 11971024 and 11901443. (*Corresponding author:Wenxia Yang*).

Cong Zhang, Wenxia Yang, and Huan Han are with the Department of Mathematics, the School of Science, Wuhan University of Technology, Wuhan, Hubei 430070, China (e-mail:zhangcong_, wenxiayang, hanhuan11@whut.edu.cn).

Xin Li is with the Department of Computer Science at the University at Albany, State University of New York, Albany, New York 12222, USA (e-mail:xli48@albany.edu).

scene graphs [30] [31], sketches [32] or doodles [33] [34]. The fundamental idea involves extracting semantic information from this guidance and embedding it into diffusion generation with reference information [35].

Despite the above improvements, most existing methods are often confined to a single modality. When dealing with complex semantics, such as images containing multiple objects or scenes, depending solely on image or text information may limit the feasibility of leveraging multiple modalities for more robust completions. The advantages of image-based guidance are evident, as images provide intuitive reference information such as human facial features, postures, expressions, and makeup styles, to name a few. Meanwhile, text-based guidance offers greater flexibility, given its convenience in providing concise and precise textual descriptions. When used together, the interaction between text and image guidance can supply richer prompt clues, helping better understand the missing regions' semantics and structures. In particular, cross-domain guidance enriches the diversity and enhances the consistency and controllability of the inpainting outcome.

To fully harness the advantages of both text and image guidance and their interaction, we present MMGInpainting, a multi-modality guided image inpainting model. MMGInpainting supports image-guided or text-guided inpainting independently and facilitates a **hybrid guidance** approach that combines both modalities for improved robustness and fine-granularity control. We use the diffusion model to fill in the missing regions and integrate the inverse diffusion process with CLIP [36] to encode semantic information. Additionally, to make the results more realistic, we design a cyclic iterative denoising strategy to incorporate reference information effectively. Furthermore, to enhance the generation of realistic inpainting results while incorporating reference semantic information, we modify the original U-Net architecture [37] with an improved NAFNet [38]. This modification streamlines the model parameters and significantly improves the inpainting outcomes. The new architecture enhances model performance by introducing the Anchor Strip Attention mechanism(ASA), which leverages anchor points to model global image structural features, leading to a higher quality of image inpainting. To our knowledge, MMGInpainting is the first inpainting model capable of simultaneously integrating image and text guidance during the inpainting process. Our extensive qualitative and quantitative experiments demonstrate that, compared to single-modality guidance or unconditional inpainting approaches, MMGInpainting excels in achieving fine texture details while providing desirable controllability.

The contributions of this paper are summarized as follows.

- We propose MMGInpainting, a comprehensive framework for guided image inpainting that accommodates the guidance of text, image, or a combination of both modalities. This unified approach provides versatility in inpainting tasks, especially for large regions.
- We enhance the baseline U-Net architecture by replacing it with a modified NAFNet within the diffusion model. We further incorporate a stripe self-attention mechanism centered on anchor points. This addition facilitates global dependency modeling, improving inpainting results and computational efficiency.
- We thoroughly analyze the semantic relationship between images and texts used in inpainting guidance. We conclude with a general recommendation on effective strategies for combining these modalities for multi-modality guidance.
- The experimental results conducted on CelebA-HQ [39] and Places2 [40] datasets demonstrate the remarkable guidance performance of our proposed model MMGInpainting. A closed-loop process with image captioning was designed to validate the semantic consistency of inpainting results.

The remaining part of this article is organized as follows. Section II introduces the related work, including image inpainting based on GANs, the Diffusion Model, and guided image synthesis. Section III provides a detailed description of the proposed method. Section IV presents the experimental results and analysis. Finally, Section V concludes with the results and discussions.

II. Related Work

Image Inpainting based on GANs. GAN-based image inpainting approaches have demonstrated significant advancements in generating meaningful content within masked holes. A typical pipeline involves employing deep neural networks to create semantically reasonable content and subsequently using this content to fill the corrupted regions. Context Encoder(CE) [3] first adopts an encoder-decoder architecture with adversarial learning to generate new content for masked areas. Early improvements focus on refining network designs and learning strategies, including Partial Convolutions [2], Gated Convolutions [41], and Contextual Attention [42]. Liu et al. [43] focuses on improving the quality of image inpainting by detecting and preserving dominant linear structures, defined as a set of lines. To generate finer details of the structure within the filled regions, EdgeConnect [44] and CTSDG [45] use edge information to guide the reconstruction of image structures. However, these methods require additional networks to extract the corresponding edge for each dataset image, and the performance is sensitive to the quality of the edge images. Prior-guided GAN [46] introduces a data-driven parametric network to predict the matching prior directly for an occluded image. To enable scene inpainting, Zhang et al. [47] proposed a general framework for facade inpainting and automatic object removal using object detection and image inpainting. Multi-GAN [48] uses a LBP(Local-binary-patterns [49])-based loss function to minimize differences between generated and natural textures. BSN [50] introduces an improved objective function for deep style transfer (DST) and an enhanced Shift-Net with multiscale feature connectivity and depthwise separable convolution (DSC) to capture local details and global semantics effectively. With the increasing applications of transformer models [51], MAT [52] tailors a transformer block for inpainting, where the attention module aggregates non-local information exclusively from partially valid tokens indicated by dynamic masks. These methods

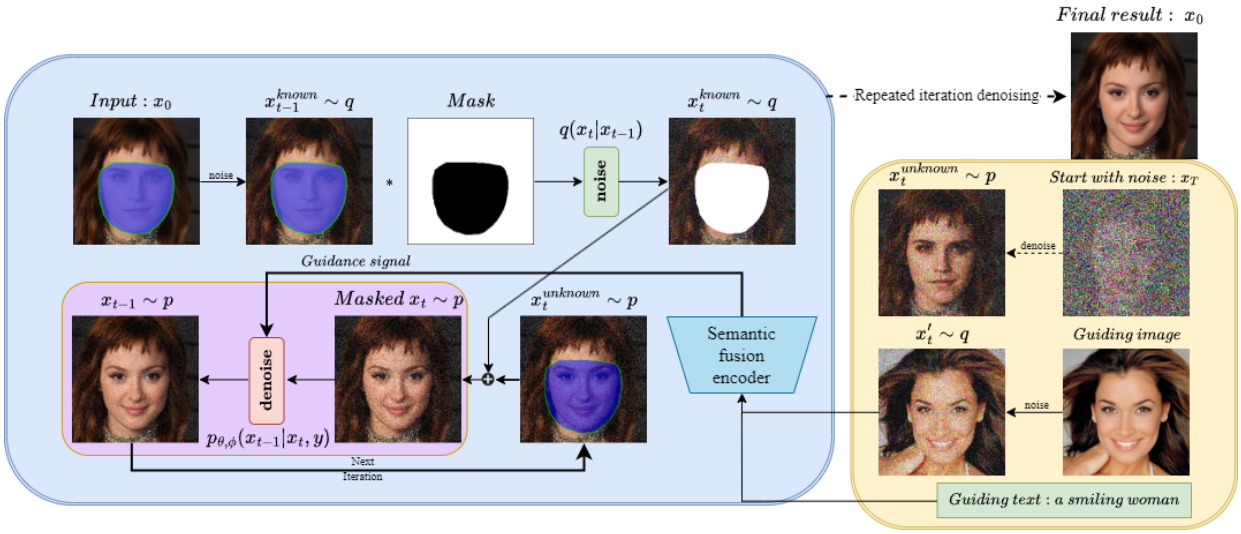


Fig. 1: **Overview of MMGInpainting.** Semantic Fusion Encoder encodes the reference image/text into a latent space. It jointly combines the inverse diffusion process to generate the mask region for inpainting. Simultaneously, it achieves the final semantically consistent result through repeated iteration denoising.

achieve visually realistic results. However, their inpainting capabilities are limited in generating specific targets or desired facial attributes, primarily due to the absence of guidance during the inpainting process.

Image Inpainting based on Diffusion Models. Diffusion models have shown remarkable performance across diverse domains [53], including image generation, image inpainting, image-to-image translation and image editing, etc. [54] [55]. For inpainting tasks, addressing issues like edge artifacts and the lack of complete contextual information provided to the model has been a research focus. Palette [14] trains the inpainting model on freely generated masks and enhances it using simple rectangular masks. Repaint [5] uses a pre-trained unconditional Denoising Diffusion Probabilistic Model (DDPM) as the generative prior. Unmasked regions are extracted from the forward records during the backward denoising process, while masked areas are filled with noise. Repaint also incorporates *resampling*, which involves adding noise to the denoised spliced result during each denoising step. By repeating this process multiple times, a semantically consistent output is obtained. DiffEdit [7] utilizes a text-conditioned diffusion model for semantic image editing tasks, enabling image edits based on textual queries. To address fidelity and realism concerns, SDEdit [8] introduces a novel image synthesis and editing method based on the diffusion model. It synthesizes realistic images by iterative denoising using a stochastic differential equation. By fine-tuning Imagen [56] into a text-conditioned image editor on a base image, UniTune [57] synthesizes images using simple prompts while maintaining fidelity to the input image. While these methods for editing images can be adapted for inpainting tasks, they often rely on random generation for filling or are constrained to single-modality guidance. Unlike previous approaches, we propose the adoption of image-text blended guidance to produce results that are more controllable with the desired outcomes.

Image Synthesis with Guidance. Image synthesis aims to create content that is not only realistic but also diverse and visually distinctive. Several innovative approaches have emerged to enhance the quality and details of synthetic content. Ren et al. [58] propose a mask embedding mechanism to facilitate efficient initial feature projection in the generator. TediGAN [59] propose a novel framework for multi-modality image generation and manipulation based on text descriptions. Dhariwal et al. [4] demonstrate that diffusion models outperform current state-of-the-art generative models in the quality of generated images. By incorporating a cross-attention layer into the model architecture, the Latent Diffusion Model (LDM) [6] transforms the diffusion model into a flexible generator capable of handling various conditional inputs, such as text or bounding boxes, enabling guided synthesis through convolution. SDG [35] introduces a novel unified framework for semantic diffusion guidance, allowing for text or image guidance, or both. The pretraining-based image-to-image translation (PITI) [60] framework adjusts a pre-trained diffusion model to accommodate various types of image-to-image translation by using a pre-trained neural network to capture the natural image manifold. Singh et al. [34] propose a novel guided image synthesis framework that models the generated image to solve a constrained optimization problem, aiming to generate an image with more vivid details. It is important to note that these methods are primarily geared to enhance the overall fineness and naturalness of synthesized images, which often involve global changes in the synthesized images. This differs from the purpose of the proposed MMGInpainting, which aims to introduce guided semantics exclusively within specific regions designated for inpainting.

III. Proposed Method

The pipeline of MMGInpainting is shown in Figure 1. This section begins with a concise overview of guided image-generation methods based on explicit classifiers. Since this

paper will use some notation, we will also introduce the guided image synthesis method based on a fine-tuned CLIP model. Then, we will present the enhanced Nonlinear Activation-Free Blocks (NAFBlocks) bolstered with attention mechanisms. Finally, we elaborate on the MMGInpainting model that can accommodate mixed guidance from textual descriptions and images.

A. Class-guided synthesis

In diffusion models, the forward process is to add noise to the initial input image x_0 via a Markov Chain over T time steps, ultimately resulting in pure noise,

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

where $\{\beta_{t=1:T}\}$ denotes a fixed or learned variance schedule that regulates the size of the noise step. The reparameterization trick is then used to sample x_t directly,

$$q\left(x_t \mid x_0\right)=\sqrt{\bar{\alpha}_t} x_0+\sqrt{1-\bar{\alpha}_t} \varepsilon, \varepsilon \sim N(0,1), \tag{2}$$

where $\alpha_t=1-\beta_t$ and $\bar{\alpha}_t=\Pi_{s=1}^t \alpha_s$.

The reverse process of diffusion models is designed to reconstruct the original data from Gaussian noise. It is reasonable to assume that the reverse process follows a series of Gaussian distributions. However, it is not feasible to gradually fit the distribution. Consequently, a parametric distribution must be constructed for estimation [61],

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

To learn the reverse process, neural networks are trained to predict μ_θ and Σ_θ can be fixed or trained as a neural network.

If an extra guidance signal y is introduced, the conditioned sampling distribution becomes:

$$p_{\theta, \phi}\left(x_{t-1} \mid x_t, y\right)=Z p_\theta\left(x_{t-1} \mid x_t\right) p_\phi\left(y \mid x_t\right), \tag{4}$$

where Z is a normalization constant. It has been proven [4] that after incorporating the guidance, the new sampling distribution in the reverse process can be approximated by a Gaussian distribution with shifted mean [62]:

$$p_\theta\left(x_{t-1} \mid x_t\right) p_\phi\left(y \mid x_t\right)=\mathcal{N}(\mu+\Sigma g, \Sigma), \tag{5}$$

where $\mu=\mu_\theta,\ \Sigma=\sigma_\theta^2 \mathbf{I},\ g=\nabla_{x_t} \log p_\phi(y|x_t)$.

B. Improved NAFBlocks based Anchored Stripe Attention

The predominant architecture framework in diffusion models is U-Net based on residual blocks and attention mechanisms. While this architecture performs well on numerous downstream tasks, fine-tuning the U-Net architecture is often required. For image inpainting tasks, efficient modeling of global contextual dependencies in high-dimensional images is crucial due to the anisotropy of the regions to be inpainted [63]. To achieve this goal, we first introduce an Anchor-based Stripe self-attention mechanism (ASA) [64] to capture global-scale dependencies. Currently, to reduce parameters and enhance model performance, we replace U-Net with NAFNet and incorporate ASA into modified nonlinear activation-free blocks (NAFBlocks), as illustrated in Figure 2. ASA first

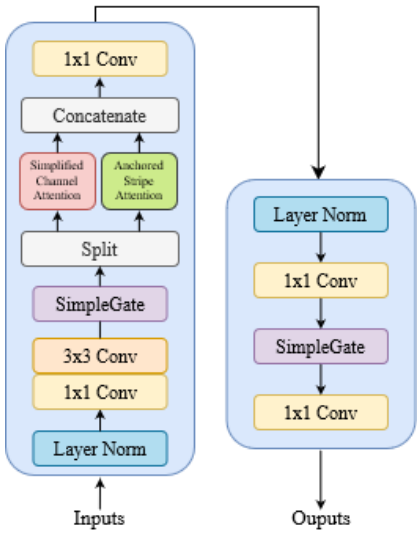


Fig. 2: **Modified NAFBlock**. The NAFBlock has an additional attention module. Here, "simple channel attention" aggregates global information and channel-wise interaction, while "Anchored Stripe Attention" models global image structural features.

introduces Anchors as an intermediary to reduce the number of tokens. When the image is summarized in a low-dimensional space using anchors, the overall image structure is preserved. The specific operations of ASA can be expressed as [64]

$$\mathbf{Y}=\mathbf{M}_e \cdot \mathbf{Z}=\mathbf{M}_e \cdot\left(\mathbf{M}_d \cdot \mathbf{V}\right), \tag{6}$$

$$\mathbf{M}_d=\operatorname{Softmax}\left(\mathbf{A} \cdot \mathbf{K}^T / \sqrt{d}\right), \tag{7}$$

$$\mathbf{M}_e=\operatorname{Softmax}\left(\mathbf{Q} \cdot \mathbf{A}^T / \sqrt{d}\right), \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{M \times d}$ denotes the anchor, $\mathbf{M}_e \in \mathcal{R}^{N \times M}$ and $\mathbf{M}_d \in \mathcal{R}^{M \times N}$ denote the attention maps of query-anchor and anchor-key, respectively.

By introducing anchors to capture the spatial relation of image features, ASA can adaptively model features separately within vertical and horizontal stripes and automatically adjust attention weights based on the content of images, thus capturing the local structure and contextual information in images. Since the masks in inpainting are typically anisotropic because they vary significantly in different directions, the anisotropic image features offered by ASA make it more efficient to exploit the anisotropic property of the inpainting domain.

To reduce the parameter count, we replace all nonlinear activation functions with "SimpleGate" in NAFNet, which involves splitting the feature map into two parts along the channel dimension and applying multiplication. LayerNorm is used instead of BatchNorm because small batches may introduce unstable statistical data that hamper image details inpainting [38]. By introducing ASA, our network can capture the image structural features from local and global regions while effectively modeling the local features using a simplified channel attention mechanism. As shown later in our ablation study, the introduction of ASA to NAFNet can improve both image fidelity and computational efficiency of MMGInpainting.

C. Multi-modality Guided Inpainting

In this paper, the ground truth image is denoted by x, and the mask is a binary matrix m where 0 indicates the known regions and 1 indicates the missing pixels to be inpainted. Thus, $(1-m)\odot x$ denotes the known pixels, $m\odot x$ is the masked unknown pixels, and $\odot$ is the Hadamard product. The inpainted result at the $t-1$ step is obtained by

$$x_{t-1}=m\odot x_{t-1}^{\text{known}}+(1-m)\odot x_{t-1}^{\text{unknown}}. \tag{9}$$

Our inpainting method is built on a diffusion model. Specifically, the diffusion model generates the missing part and seamlessly combines it with the noisy original image. Then, iterative denoising is performed to ultimately inpaint the complete result. The primary objective is to control the reverse diffusion process to achieve multimodality-guided inpainting.

The optimization goal of diffusion models is essentially to fit an optimal gradient direction $\nabla\log P(x_t)$ towards the target data distribution in the data space [65] [18]. To facilitate the guided generation, the Bayesian theorem can be applied to decompose the gradient for conditional generation into two components, including a regular, unconditional generation gradient and a gradient based on an explicit classifier [54] [66]:

$$\nabla\log p(x_t\mid y)=\underbrace{\nabla\log p(x_t)}_{\text{unconditional score}}+\underbrace{\nabla\log p(y\mid x_t)}_{\text{adversarial gradient}}, \tag{10}$$

Among them, y can be text, image, or multi-modality guidance. For clarity, we refer to the classifier as:

$$F_\phi(x_t,y,t)=\log P_\phi(y\mid x_t). \tag{11}$$

Next, we utilize the alignment representation between texts and images in the Contrastive Language-Image Pre-Training (CLIP) model [36] to compute specific loss values. During each step of text-guided image generation, the distance between the current image representation and the text representation is calculated, typically using the inner product distance [57] or cosine similarity (a.k.a. CLIPScore [67]):

$$F(x_t,l,t)=E_I'(x_t,t)\cdot E_L(l), \tag{12}$$

where E_I' denotes the image encoder trained on noisy images with an additional time step input [35], and E_L denotes the text encoder. In the case of image-guided image generation, inner distance does not account for spatial information. To incorporate spatial context information, we consider the L2 norm difference at the corresponding positions in the feature maps. By considering the spatial layout, we obtain the following accumulation for an entire image:

$$F(x_t,x_t',t)=-\sum_j\frac{1}{C_jH_jW_j}\left\|E_I'(x_t,t)_j-E_I'(x_t',t)_j\right\|_2^2, \tag{13}$$

Algorithm 1 Multi-Modality Guided Inpainting

Input : Guidance y, Editing strength t_{edit}, scaling factor s
Given: diffusion model($\mu_\theta(x_t,t),\Sigma_\theta(x_t,t)$), Guidance function $F_\phi(x_t,y,t)$

```
x_T ~ N(0, I)
for t = T, ..., 1 do
    for u = 1, ..., U do
        x_{t-1}^{known} ~ N(sqrt(ᾱ_t) x_0, (1 − ᾱ_t) I)
        μ, Σ ← μ_θ(x_t, t), Σ_θ(x_t, t)
        if t > t_edit then
            x_{t-1}^{unknown} ~ N(μ + sΣ∇_{x_t} F_φ(x_t, y, t), Σ)
        else
            x_{t-1}^{unknown} ~ N(μ, Σ)
        end
        x_{t-1} = m ⊙ x_{t-1}^{known} + (1 − m) ⊙ x_{t-1}^{unknown}
        if u < U and t > 1 then
            x_t ~ N(sqrt(1 − β_{t-1}) x_{t-1}, β_{t-1} I)
        end
    end
end
return x_0
```

where $E_I'()_j\in\mathcal{R}^{C_j\times H_j\times W_j}$ denotes the spatial feature maps of E_I'. As for multi-modality guided synthesis, we assign specific weights s_1 and s_2 ($s_1+s_2=1$) to the image-guided and text-guided functions, respectively.

$$F_{\phi_0}(x_t,y,t)=s_1F_{\phi_1}(x_t,y,t)+s_2F_{\phi_2}(x_t,y,t) \tag{14}$$

where F_{ϕ_1},F_{ϕ_2} denote the loss functions corresponding to image-guided and text-guided inpainting, respectively.

To integrate the high-level semantic information of the reference image (or text) into the reverse generation process, in each denoising step, we extract the noisy image from the forward process and concatenate it with the missing region generated with reference information guidance, followed by a reverse denoising step. To prolong the sampling process to generate overall semantically consistent content, the denoised image is subjected to noise addition repeatedly. Inpainting results can be obtained with reference semantics when this process is repeated multiple times. The Semantic Fusion Encoder is a fine-tuned CLIP model that can accept noisy images as input, which encodes supplementary reference information into the latent semantic space and iteratively integrates the reference semantics into every inpainting process.

Inspired by Wang et al. [60], we adopt a pre-trained image-to-image translation framework with an encoder that transforms the input into a task-independent latent space and a decoder that performs diffusion modeling. The main idea is to fix the pretrained decoder, update only the encoder, and then jointly fine-tune the entire network. This staged training approach maximizes the utilization of pre-trained knowledge while ensuring practical guidance for inpainting. Unlike [60], our pretext condition involves text and image guidance. The two hyperparameters in Eq. (14) control the weight adjustment between text and image guidance. This way, our conditional fine-tuning works in the pre-trained semantic

space of multi-modality. Algorithm 1 summarizes the proposed multi-modality guided inpainting model. Our model can also operate without guidance(guidance-free case), with the signal y being empty in such cases.

IV. Experimental Results

A. Dataset and Implementation Details

We conducted guided image inpainting experiments on CelebA-HQ [39] and Places2 [40] datasets, using the same experimental setup as NAFNet [38] with the image size 256×256, and the batch size 8. We use the AdamW [68] optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The initial learning rate was set to 3×10^{-3}, which was decayed to 1e-7 using a cosine scheduler. The weight parameters s_1 and s_2 in Eq.(14) were set at 0.4 and 0.6 by the experiment discussed in section IV, part D. The Semantic Fusion Encoder is CLIP ResNet50x16 fine-tuned on the noisy images of each dataset, with the initial learning rate set to 10^{-4}, weight decay to 10^{-3}, and the batch size to 64. The denoising steps of the diffusion model were set to $T = 250$, with the resampling setting consistent with Repaint [5]. It was trained on a single 3090 GPU for approximately five days. To ensure semantic consistency throughout the image, the injection procedure of guiding semantics was confined within a predefined range of steps, denoted as $[T, t_{edit}]$, without being introduced into the resampling loop. The hyperparameter t_{edit} was set to 75 by experiment.

B. Experimental Result Analysis

1) Experimental settings and Quantitative Comparisons: To evaluate the performance of our model, we first conducted quantitative experiments on CelebA-HQ and Places2. We compared the proposed MMGInpainting with five state-of-the-art image-inpainting models from the literature, including LAMA [20], LDM [6], RePaint [5], MAT [52], and BLD [69]. Among them, LDM can accept either image or text guidance, BLD only supports text-driven, while LAMA, Repaint, and MAT are guidance-free. For fairness, we use the official pre-trained models provided by the corresponding authors.

The experiment was conducted on 100 images selected from the test sets of CelebA-HQ and Places2, respectively. We masked the most critical components for each image using a 64×64 center mask. Because the text or image used for inpainting guidance can be diverse, the experimental setup was as follows: for CelebA-HQ, if the model is image-guided, the ground-truth image is used as guidance; if it is text-guided, the text is generated using the labels corresponding to the ground-truth image. For example, if the attribute labels of the original image indicate "Smiling = 1" and "Male = -1", the text used for guidance would be "a smiling woman". In Places2, we use additional object cues to guide different scenes, such as outdoors, buildings, and vegetation, as labeled in the dataset. In this way, the semantic information of the text and image is consistent with the original image. This setup was designed to ensure uniformity in the guidance semantics for different models when conducting experiments on many samples, thereby making the comparison of experimental results meaningful. Since our proposed MMGInpainting supports text guidance, image guidance, and simultaneous guidance from both, we conducted three sets of experiments, obtaining results with no guidance, guided solely by text, solely by images, and by both simultaneously.

Six evaluation metrics, namely *Fréchet Inception Distance* (FID) [70], *Learned Perceptual Image Patch Similarity* (LPIPS) [71], *Peak Signal to Noise Ratio* (PSNR), *Paired/Unpaired Inception Discriminative Score* (P/U-IDS) [72], and *CLIP-based scoring function*(PickScore) [73] are calculated. Among them, FID serves as a quantitative metric to assess the similarity between the statistical distributions of real images and generated images, and P/U-IDS reflects the fidelity of the generated images by calculating the linear separability between the generated images and the real images in the feature space of the perceptron. The P/U-IDS is given by

$$\textbf{P-IDS}(\boldsymbol{X}) = \Pr_{(\boldsymbol{x},\boldsymbol{x}')\in\boldsymbol{X}} \left\{ f\left(\mathcal{I}\left(\boldsymbol{x}'\right)\right) > f(\mathcal{I}(\boldsymbol{x})) \right\}, \tag{15}$$

$$\begin{aligned} \textbf{U-IDS}\left(X, X'\right) = & \frac{1}{2} \Pr_{\boldsymbol{x}\in X} \{ f(\mathcal{I}(\boldsymbol{x})) < 0 \} + \\ & \frac{1}{2} \Pr_{\boldsymbol{x}'\in X'} \left\{ f\left(\mathcal{I}\left(\boldsymbol{x}'\right)\right) > 0 \right\} \end{aligned}, \tag{16}$$

where x denotes the real image and x' denotes the corresponding generated fake image. $\mathcal{I}(\cdot)$ is the pre-trained Inception v3 model that maps the input images to the output features of 2048 dimensions. And $f(\cdot)$ denotes the (linear) decision function of the SVM. Compared to CLIPScore [67], PickScore is a text-to-image metric that measures the fidelity of the generated content based on learned human preferences. The matching score between text x and image y is calculated as:

$$s(x, y) = E_{\text{txt}}(x) \cdot E_{\text{img}}(y) \cdot T \tag{17}$$

where T denotes the learned scalar temperature parameter of CLIP. The objective function L_{pref} aims to optimize the parameters of the scoring function by minimizing the KL-divergence between the preference distribution p and the softmax-normalized scores of y_1 and y_2:

$$\begin{aligned} \hat{p}_i &= \frac{\exp s\left(x, y_i\right)}{\sum_{j=1}^{2} \exp s\left(x, y_j\right)} \\ L_{\text{pref}} &= \sum_{i=1}^{2} p_i \left(\log p_i - \log \hat{p}_i\right) \end{aligned} \tag{18}$$

Table I presents a quantitative comparison of MMGInpainting with other methods on Celeb-HQ and Place2 test sets containing 100 images, respectively. For CelebA-HQ, our primary focus is on the fidelity of the inpainted facial results and the presence of additional semantic information. We use the FID, LPIPS, PSNR and PickScore metrics to assess this. On the other hand, for Place2, given the complexity of scene data, our emphasis lies on the overall cohesiveness of the images. Therefore, we use the FID, LPIPS, PSNR, and P/U-IDS metrics for evaluation. As shown in Table I, our model performs favorably regarding image fidelity compared to the unconditional inpainting model. The fidelity of our model's inpainting results guided by the original image exhibits the

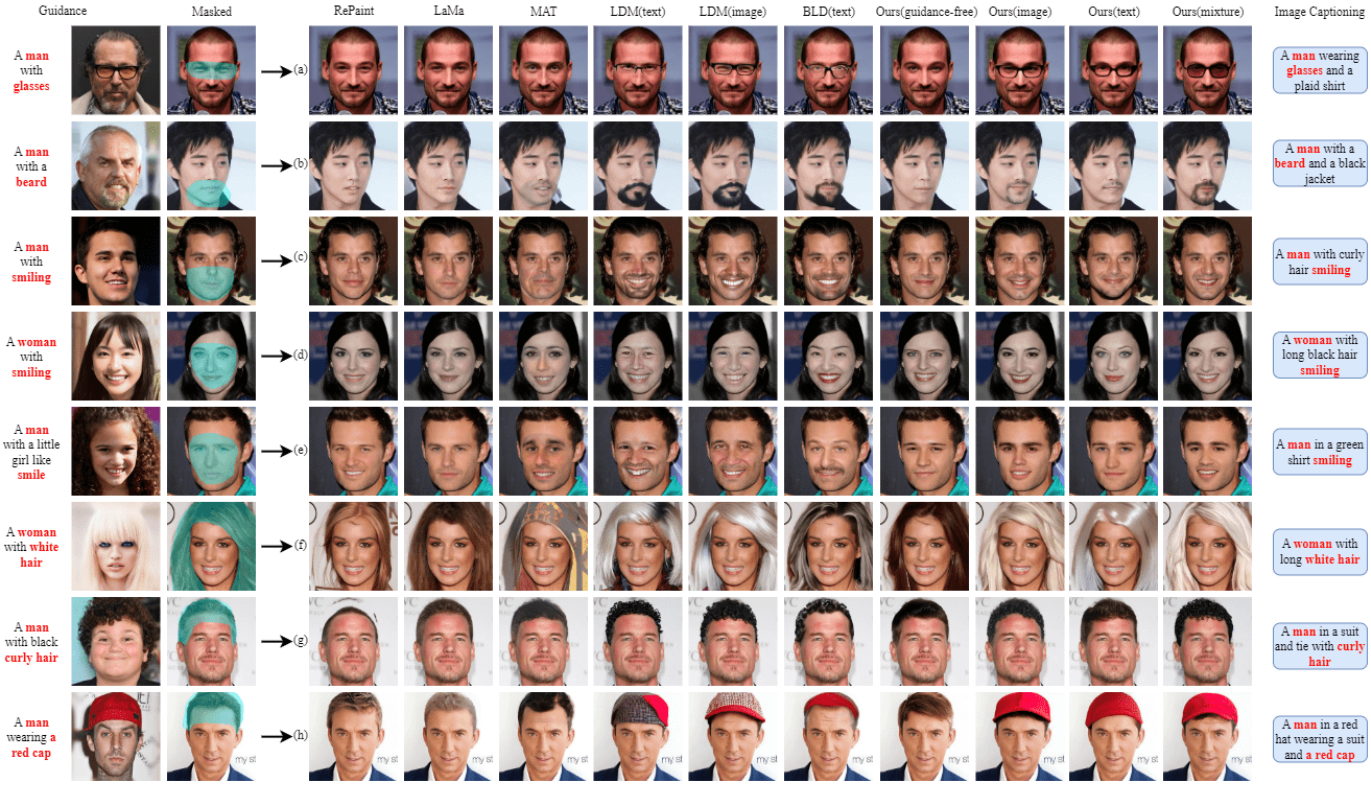


Fig. 3: **Qualitative Results on CelebA-HQ**. Comparison of MMGInpainting against other state-of-the-art face inpainting models, where the semantics of the guiding texts and images are consistent.

Method	#Params[M]	CelebA-HQ				Place2				
		FID ↓	LPIPS ↓	PSNR ↑	PickScore ↑ /%	FID ↓	LPIPS ↓	PSNR ↑	U-IDS ↑ /%	P-IDS ↑ /%
RePaint	55	10.32	0.219	22.82	26.75	15.38	0.231	21.64	18.67	7.56
LaMa	51/27	9.87	0.207	23.45	29.97	13.82	0.198	23.81	23.03	8.57
MAT	62	12.72	0.231	19.27	25.53	17.04	0.252	20.36	19.89	7.20
LDM(image)	387	10.24	**0.185**	23.71	29.63	13.47	0.227	23.42	22.98	9.28
LDM(text)	387	11.48	0.192	23.62	28.75	14.23	0.238	22.31	22.54	8.97
BLD(text)	160	10.12	0.201	21.48	28.62	14.72	0.246	22.02	23.79	9.02
Ours(guidance-free)	52	9.21±0.08	0.211	21.58	27.32±0.11	14.74±0.05	0.229	21.85	19.82±0.25	8.17±0.19
Ours(image)	67	**8.47±0.12**	0.187	23.62	36.14±0.09	**11.93±0.02**	0.218	23.25	23.63±0.23	9.74±0.24
Ours(text)	67	9.03±0.13	0.190	22.79	35.93±0.12	13.02±0.09	0.231	22.69	23.87±0.24	9.36±0.26
Ours(hybrid)	67	8.73±0.15	0.186	**24.31**	**36.41±0.08**	12.06±0.04	**0.196**	**24.17**	**25.93±0.17**	**10.85±0.18**

TABLE I: **CelebA-HQ and Place2 Quantitative Results**. ↓ indicates that a smaller value is better.

lowest FID, even slightly surpassing the state-of-the-art LaMa. LDM achieves the optimal LPIPS, but its PSNR is somewhat lower than our model. Furthermore, our model, guided by a combination of image and text, generates inpainting results with a higher PickScore than the guided inpainting model. This indicates that our inpainting results share more semantic similarities with the reference information. Regarding model parameters, our parameter count is lower than other guided methods and slightly higher than unguided methods. This difference is attributed to incorporating an additional feature extraction module and an attention mechanism to enhance the final guided inpainting effect.

2) Qualitative results and Comparisons: We further experimented with various texts and images to assess different models' guiding effects and inpainting quality. In this experiment, the guiding image is entirely different from the ground truth, and the semantics of the text used for guidance are set to be consistent with the semantics of the guiding image. Figure 3 illustrates some examples of CelebA-HQ. We purposely masked some critical facial features such as eyes, overall facial features, hair, or chins to verify whether the inpainted images contained additional guidance information.

Visual inspection of Figure 3(a)-(h) reveals that all unconditional inpainting methods like RePaint, LaMa, MAT, and our guidance-free case successfully obtain good texture details with natural facial expressions, despite there may be some color discrepancies around the borders in MAT. Nevertheless, there are comparatively noticeable flaws and blurred details in inpainting the hair details in Figure3(f). In Figure 3(c)-(e), MAT may generate shadows when there are occlusions of prominent facial features. LaMa generates content with overall semantic consistency but lacks guidance during the generation

Fig. 4: **Qualitative Results on Place2**. Comparison against state-of-the-art methods over different mask settings.

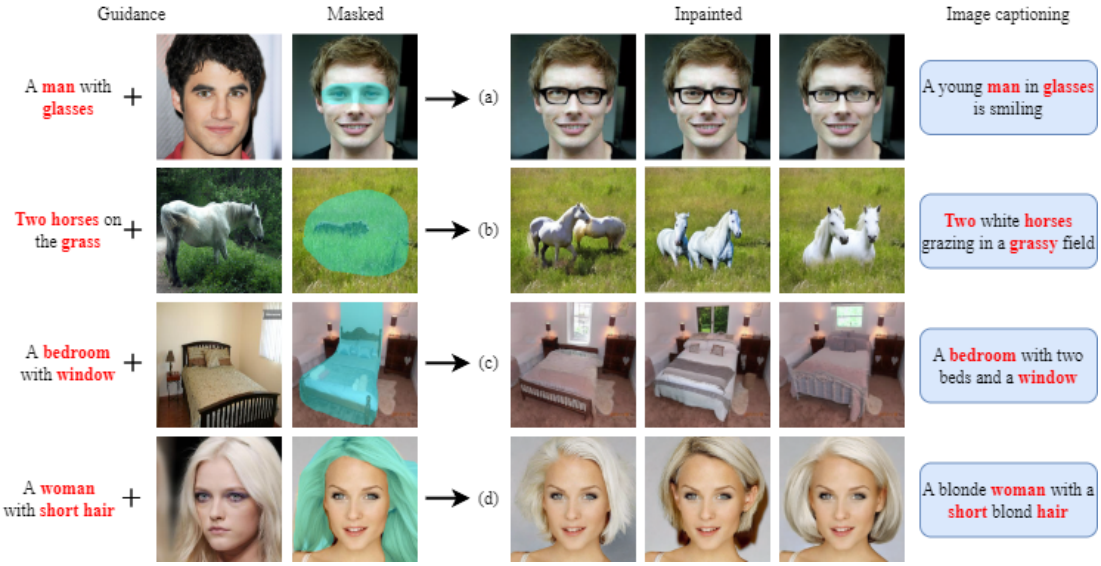


Fig. 5: **Inpainting results when the guiding semantics complement each other**. The combination of image and language guidance provides complementary information, and our model inpaints images that match both sources of guidance.

process. LDM and BLD exhibit inadequate coherence in overall semantics for guided inpainting models. In contrast, our model efficiently models global feature information in the images to be inpainted, considering both the unmasked regions and the guidance information. Specifically, our model excels in inpainting texture details of hair and beard in Figure 3(f) and (b), highlighting its advantages in preserving texture details and conducting guided region inpainting.

We observed that LDM performs well in image-guided generation, while our model achieves the best results with mixed guidance. Therefore, for the Place2 dataset, we use image guidance for LDM and composite guidance for our model. As shown in Figure 4, for object-guided generation, our model effectively inpaints the details of the reference object and integrates it into the target image without producing artifacts like BLD. LDM successfully incorporates the reference object but loses some details. Regarding unrestricted inpainting, the focus is on maintaining overall semantic consistency, where LaMa performs the best.

To further evaluate the effectiveness of text guidance, we conducted a closed-loop experiment. Specifically, we used an off-the-shelf image captioning API [74] to generate the descriptions for the inpainted results. In Figure 3 and Figure 4, we have highlighted the identical words found in both the guiding texts and the generated captions. The semantic information from the text guidance is well embedded into the inpainted regions. Additionally, we observed that the surplus text obtained after image captioning captures semantic details related to the guiding image. This closed-loop experiment shows the effectiveness of semantic guidance through text and image elements.

C. Analysis of the relationship of the semantics between text and image guidance

In the preceding subsections, we demonstrate that our model can achieve richer details and more desirable visual outcomes

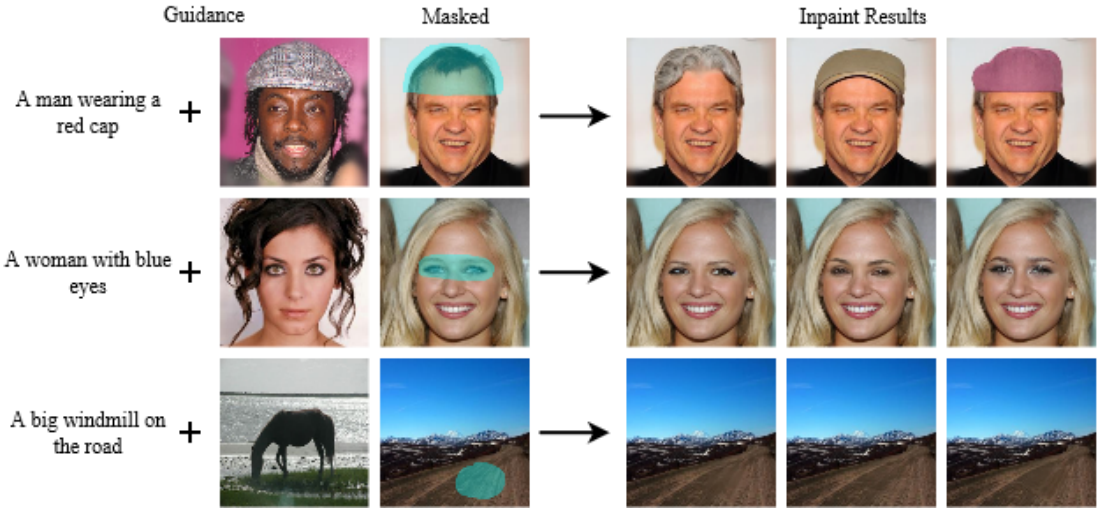


Fig. 6: **Inpainting Results with contradictory semantics of guiding image and text** . When there is a semantic conflict, it may cause the inpainting results to uncertain directions.

when the semantics of the text and image used for inpainting guidance are consistent. However, what happens when the two semantics are contradictory or complementary? In this section, we conduct experiments to explore these scenarios. Semantic complementary guidance involves using an image while providing textual descriptions of significant attributes absent in the guiding image. Figure 5 illustrates some results for each image; it can be observed that although the inpainted images are diverse, the texts of image captioning have subtle differences. Furthermore, we make the following observations:

(I) For face inpainting, when guided by semantic complementary information, the overall appearance is determined by the guiding image. In contrast, the complementary text introduces diversity and finer details into the inpainting results, such as attributes like "short hair" or "glasses."

(II) For Scene Inpainting, image guidance ensures that the inpainting result contains objects guided by the input image. In contrast, text guidance allows for specifying object layouts and the number of objects, adding extra content, and introducing variations.

We argue that complementary semantics enhance diversity and boost finer details in inpainting results. When the semantics are identical, more precise guidance is provided, emphasizing specific semantics through repetition. However, when semantics are opposed or in apparent contradiction with the original image, conflicting information can lead to randomly generated results, some of which are visually unreasonable. This may produce misleading and even ethically concerning results, as shown in Figure 6. Therefore, we conclude with the following recommendations. For face inpainting, prioritize image guidance with text as a supplementary source. For Place2 inpainting, consider using a specific object as the guiding image, ideally with a structure similar to the original image. Text guidance can be employed for object layout and specific morphological changes.

D. Analysis of the weights s_1 and s_2

In Eq.(14), the ratio of weights s_1 and s_2 signifies the proportion of semantics embedded by the guiding image and text during the inpainting, respectively. When the guiding image and text semantics are consistent enough, their ratio has little impact on the results. However, when the semantics of the image and text complement each other, different settings of s_1 and s_2 significantly influence the results. In this section, we experimented to explore the impact of varying weight settings.

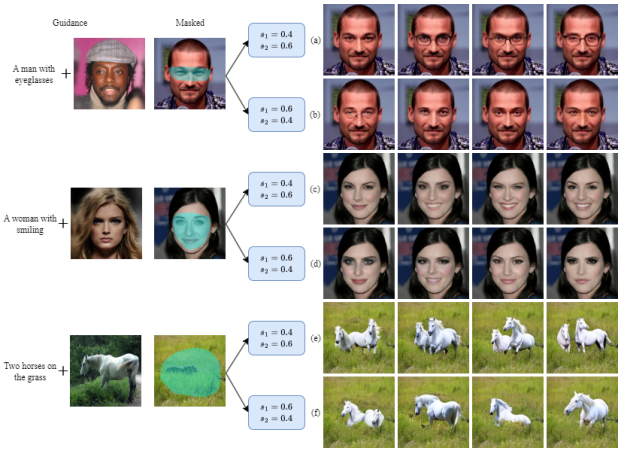


Fig. 7: **Selection of weights for s_1 and s_2**. Where s_1 and s_2 are the weights of image and text guidance respectively.

As shown in Figure 7, when the text weight is smaller, although the results still contain the semantics guided by the text, the probability is much lower. In contrast, facial appearances or animal postures in the results are more similar to those of the guiding images, as illustrated in Figure 7(b), (d), and (f). On the contrary, when the text weight is increased, the probability of combining complementary semantics increases significantly, obtaining results with more semantics indicated by the text, such as "eyeglasses" in Figure 7(a), "smiling"

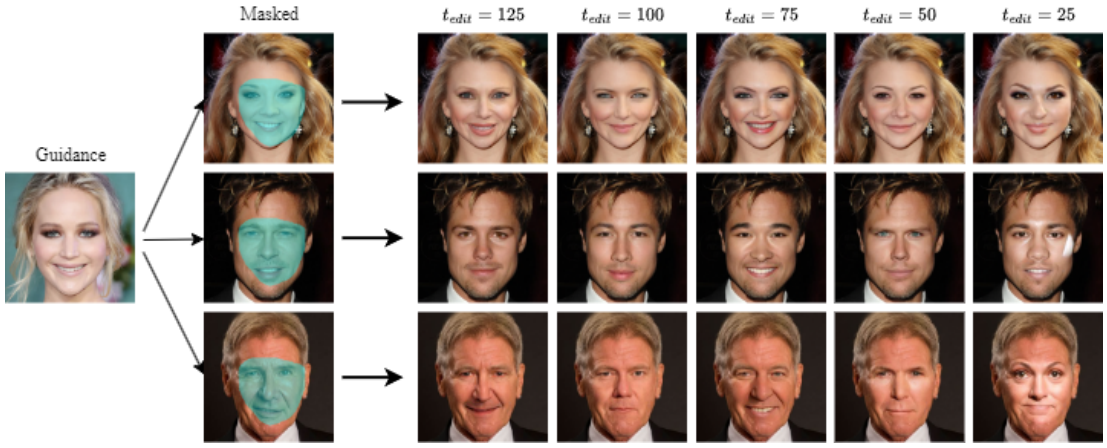


Fig. 8: **Impact of different t_{edit} on the inpainting results**. When t_{edit} is too small, the "excessive guidance" problem will occur with excessive guiding semantics embedded, leading to inconsistent skin color. Conversely, when t_{edit} is too large, it can cause "insufficient guidance". Setting t_{edit} to be 75 achieves a relatively proper balance.

in Figure7(c), and "Two horses" in Figure7(e). Based on the experiment, we set $s_1 = 0.4$ and $s_2 = 0.6$ in this paper.

E. Discussion on the setting of t_{edit}

The editing strength t_{edit} is a user-controllable hyperparameter that regulates guidance strength. If the guidance steps are excessively long, it may disrupt the global semantics of the image, leading to an inconsistent result. Conversely, additional prompt injection may be insufficient if the steps are too short. In Figure 8, we investigate the impact of varying t_{edit} on the inpainting results. We used the weighted combination of FID and PickScore, PS-FID (PickScore-based FID), to strike a balance between the global semantic coherence of the inpainted image and its similarity to the guiding semantics. PS-FID is calculated by :

$$\text{PS-FID} = \alpha\text{FID} - \beta\text{PickScore} \tag{19}$$

In this case, a smaller PS-FID represents a better inpainting result. We select images from Figure 3 in the CelebA-HQ dataset with $\alpha = 5, \beta = 1$ and vary t_{edit} with an interval length of 25. Then, we obtain the polyline graph of the mean PS-FID as a function of t_{edit}, as shown in Figure 9. Notably, when t_{edit} is too large, it leads to "insufficient guidance", which means that the guiding semantics are subtly embedded into the inpainting procedure, and when t_{edit} is too small, the "excessive guidance" effect will cause overall semantic incoherence such as inconsistent skin color. Both scenarios increase the PS-FID. Therefore, we recommend that a $t_{edit} = 75$ strikes an acceptable balance.

F. Ablation Study

To assess the effectiveness of our proposed modified NAFNet and ASA module, we conducted ablation studies on the CelebA-HQ dataset. In comparison, we evaluated our approach against the diffusion model based on U-Net. We used an extra evaluation metric termed Runtime [63] to quantify the time required for computing the inpainting of a single image. Runtime measures the computational efficiency gains achieved by our method relative to the baseline. The experimental results are presented in Table II and Figure 10. The results show that replacing the U-Net backbone with NAFNet significantly reduces the inference time and slightly improves image quality. Our MMGInpainting, combined with ASA, showcases exceptional performance in terms of image fidelity and guiding content measurement, surpassing the baseline in computational efficiency.

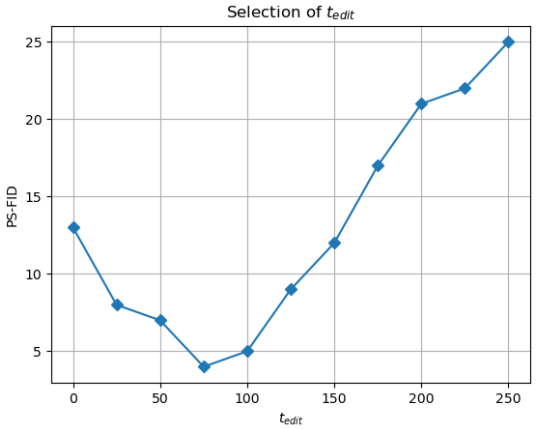


Fig. 9: **Changing trend of PS-FID with respect to t_{edit}**. The relatively optimal value of t_{edit} is around 75.

NAFBlocks	ASA	Method	FID↓	PickScore↑	Params↓	Runtime↓
		U-Net baseline	10.12	32.62	70M	296s
✓		NAFNet	9.86	28.75	66M	160s
✓	✓	MMGInpainting	**8.45**	**36.5**	67M	171s

TABLE II: The quantitative results of different models in ablation study.

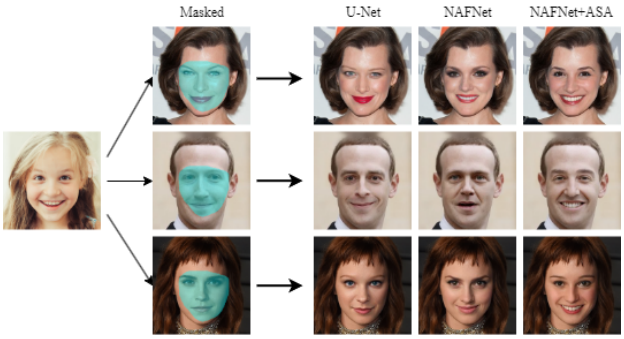


Fig. 10: **The inpainting results of different models in ablation study**. NAFNet and ASA, in combination, show the best performance.

V. Conclusion and Discussions

In this paper, we present a multi-modality guided image inpainting method aimed at enhancing the controllability of the inpainting process. Specifically, we replace the U-Net backbone with a modified NAFNet to achieve superior inpainting results and boost computational efficiency. This modified architecture incorporates an Anchored Stripe Attention mechanism, leveraging anchor points for comprehensive global contextual modeling. Then, we design a novel inpainting method capable of intelligently filling in image regions through a hybrid-guided approach. Experimental results demonstrate that our method significantly improves the controllability of inpainting, resulting in higher-quality inpainting results and faster computational efficiency. Our model seamlessly integrates text and image guidance within a unified framework, offering flexibility for diverse applications. While the long sampling steps are designed to generate detailed and high-quality inpainting results, their time-consuming nature hinders practical applications, especially for large-scale image inpainting tasks or real-time scenarios. In future work, we plan to explore methods to divide large-scale image inpainting tasks into multiple sub-tasks, enabling parallel processing and acceleration of the overall speed. Cross-modal alignment [75] is an underresearched topic in MMGInpainting, especially when text and image guidances are inconsistent. How to control the multi-modality guided diffusion in a collaborative manner (e.g., using bilateral connections [76]) deserves a more systematic study.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

[2] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image Inpainting for Irregular Holes Using Partial Convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100, 2018.

[3] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, June 2016.

[4] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.

[5] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[7] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," *arXiv preprint arXiv:2210.11427*, 2022.

[8] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations," in *International Conference on Learning Representations*, Mar. 2022.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

[10] K. Pandey, A. Mukherjee, P. Rai, and A. Kumar, "Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents," *arXiv preprint arXiv:2201.00308*, 2022.

[11] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[12] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

[13] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007–6017, 2023.

[14] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

[15] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, *et al.*, "ediffi: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[16] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *arXiv preprint arXiv:2112.07804*, 2021.

[17] Y. Song and S. Ermon, "Improved Techniques for Training Score-Based Generative Models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 12438–12448, Curran Associates, Inc., 2020.

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

[19] H. Sun, W. Li, Y. Duan, J. Zhou, and J. Lu, "Learning adaptive patch generators for mask-robust image inpainting," *IEEE Transactions on Multimedia*, 2022.

[20] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022.

[21] J. Wang, S. Chen, Z. Wu, and Y.-G. Jiang, "Ft-tdr: Frequency-guided transformer and top-down refinement network for blind face inpainting," *IEEE Transactions on Multimedia*, 2022.

[22] Y. Guo, Q. Jin, J.-M. Morel, T. Zeng, and G. Facciolo, "Joint demosaicking and denoising benefits from a two-stage training strategy," *Journal of Computational and Applied Mathematics*, p. 115330, 2023.

[23] J. Sun, F. Xue, J. Li, L. Zhu, H. Zhang, and J. Zhang, "Tsinit: a two-stage inpainting network for incomplete text," *IEEE Transactions on Multimedia*, 2022.

[24] Y. Yu, H. Wang, T. Luo, H. Fan, and L. Zhang, "Magic: Multi-modality guided image completion," *arXiv preprint arXiv:2305.11818*, 2023.

[25] J. Jain, Y. Zhou, N. Yu, and H. Shi, "Keys to better image inpainting: Structure and texture go hand in hand," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 208–217, 2023.

[26] C. Cao, Q. Dong, Y. Wang, Y. Cai, and Y. Fu, "A unified prompt-guided in-context inpainting framework for reference-based image manipulations," *arXiv preprint arXiv:2305.11577*, 2023.

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

[28] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

[29] Z. Zhang, L. Han, A. Ghosh, D. N. Metaxas, and J. Ren, "Sine: Single image editing with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6027–6037, 2023.

[30] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," *arXiv preprint arXiv:2108.02938*, 2021.

[31] Y. Zeng, Z. Lin, J. Zhang, Q. Liu, J. Collomosse, J. Kuen, and V. M. Patel, "SceneComposer: Any-Level Semantic Image Synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22468–22478, 2023.

[32] Y. Peng, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "Difffacesketch: High-fidelity face image synthesis with sketch-guided latent diffusion model," *arXiv preprint arXiv:2302.06908*, 2023.

[33] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

[34] J. Singh, S. Gould, and L. Zheng, "High-Fidelity Guided Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5997–6006, 2023.

[35] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, "More control for free! image synthesis with semantic diffusion guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 289–299, 2023.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[38] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*, pp. 17–33, Springer, 2022.

[39] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

[40] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[41] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4471–4480, 2019.

[42] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

[43] J. Liu, S. Yang, Y. Fang, and Z. Guo, "Structure-guided image inpainting using homography transformation," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3252–3265, 2018.

[44] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[45] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14134–14143, 2021.

[46] A. Lahiri, A. K. Jain, S. Agrawal, P. Mitra, and P. K. Biswas, "Prior guided gan based semantic inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13696–13705, 2020.

[47] J. Zhang, T. Fukuda, and N. Yabuki, "Automatic object removal with obstructed façades completion using semantic segmentation and generative adversarial inpainting," *IEEE Access*, vol. 9, pp. 117486–117495, 2021.

[48] M. A. Hedjazi and Y. Genc, "Efficient texture-aware multi-gan for image inpainting," *Knowledge-Based Systems*, vol. 217, p. 106789, 2021.

[49] H. Wu, J. Zhou, and Y. Li, "Deep generative model for image inpainting with local binary pattern learning and spatial attention," *IEEE Transactions on Multimedia*, vol. 24, pp. 4016–4027, 2021.

[50] C. Long, X. Li, Y. Jing, H. Shen, *et al.*, "Bishift networks for thick cloud removal with multitemporal remote sensing images," *International Journal of Intelligent Systems*, vol. 2023, 2023.

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

[52] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10758–10768, 2022.

[53] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," *arXiv preprint arXiv:2210.10960*, 2022.

[54] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[55] O. Avrahami, D. Lischinski, and O. Fried, "Blended Diffusion for Text-Driven Editing of Natural Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

[56] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36479–36494, 2022.

[57] D. Valevski, M. Kalman, E. Molad, E. Segalis, Y. Matias, and Y. Leviathan, "Unitune: Text-driven image editing by fine tuning a diffusion model on a single image," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023.

[58] Y. Ren, Z. Zhu, Y. Li, D. Kong, R. Hou, L. J. Grimm, J. R. Marks, and J. Y. Lo, "Mask embedding for realistic high-resolution medical image synthesis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pp. 422–430, Springer, 2019.

[59] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-Guided Diverse Face Image Generation and Manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2256–2265, 2021.

[60] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen, "Pretraining is all you need for image-to-image translation," *arXiv preprint arXiv:2205.12952*, 2022.

[61] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022.

[62] Y. Benny and L. Wolf, "Dynamic Dual-Output Diffusion Models," *arXiv preprint arXiv:2203.04304*, 2022.

[63] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1680–1691, 2023.

[64] Y. Li, Y. Fan, X. Xiang, D. Demandolx, R. Ranjan, R. Timofte, and L. Van Gool, "Efficient and explicit modelling of image hierarchies for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18278–18289, 2023.

[65] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *International Conference on Machine Learning*, pp. 8857–8868, PMLR, 2021.

[66] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, 2022.

[67] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[68] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[69] O. Avrahami, O. Fried, and D. Lischinski, "Blended latent diffusion," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–11, 2023.

[70] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in

Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586–595, 2018.

[72] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, "Large scale image completion via co-modulated generative adversarial networks," *arXiv preprint arXiv:2103.10428*, 2021.

[73] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *arXiv preprint arXiv:2305.01569*, 2023.

[74] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.

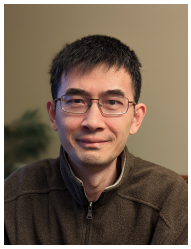
[75] Y. Zhou and G. Long, "Improving cross-modal alignment for text-guided image inpainting," *arXiv preprint arXiv:2301.11362*, 2023.

[76] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6080–6090, 2023.

VI. Biography Section

Cong Zhang was born in 1999. He received the bachelor's degree in information and computing science from Wuhan University of Technology in 2021. He is currently pursuing the M.S. degree with the Department of Mathematics, Wuhan University of Technology. His research interests include image inpainting and deep learning.

Wenxia Yang received the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology, Wuhan, China, in 2009. She is an Associate Professor of the Department of Mathematics, Wuhan University of Technology, Wuhan, China. Her current research interests include image processing and machine learning.



Xin Li received the B.S. degree (Hons.) in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 1996 and 2000, respectively. He was a member of Technical Staff with Sharp Laboratories of America, Camas, WA, USA, from 2000 to 2002. He has been a faculty member at the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, USA, from 2003 to 2023. He is now a professor with the Department of Computer Science at the University at Albany, State University of New York, Albany, USA. His current research interests include image and video coding and processing. He was elected a Fellow of IEEE in 2017.

Huan Han received the Ph.D. degree in applied mathematics from University of Chinese Academy of Sciences, China. He is currently an Associate Professor in the Department of Mathematics, Wuhan University of Technology, China. His interests include image processing and numerical methods for partial differential equations.