Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Few-shot exemplar-driven inpainting with parameter-efficient diffusion fine-tuning*

Shiyuan YANG[1], Zheng GU[2], Wenyue HAO[1], Yi WANG[1], Huaiyu CAI[1], Xiaodong CHEN [1‡]

[1] *Key Laboratory of Opto-Electronics Information Technology of Ministry of Education, School of Precision Instruments and Opto-Electronic Engineering, Tianjin University, Tianjin 300072, China*

[2] *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210008, China*

E-mail: yangshiyuan@tju.edu.cn; guzheng@smail.nju.edu.cn; wy_hao, koala_wy, hycai, xdchen@tju.edu.cn

Received May 14, 2024; Revision accepted Oct. 25, 2024; Crosschecked

Abstract: Text-to-image diffusion models have demonstrated impressive capabilities in image generation that have been effectively applied to image inpainting. While text prompt provides an intuitive guidance for conditional inpainting, users often seek the ability to inpaint a specific object with customized appearance by providing an exemplar image. Unfortunately, existing methods struggle to achieve high-fidelity in exemplar-driven inpainting. To address this, we utilized a plug-and-play low-rank adaptation(LoRA) module based on a pretrained text-driven inpainting model. The LoRA module is dedicated to learn the exemplar-specific concepts through few-shot fine-tuning, bringing improved fitting capability to customized exemplar images, without intensive training on large-scale datasets. Additionally, we introduced GPT-4v prompting and prior noise initialization techniques to further facilitate the fidelity in inpainting results. In brief, the denoising diffusion process first starts with the noise derived from a composite exemplar-background image, and is subsequently guided by an expressive prompt generated from the exemplar using the GPT-4v model. Extensive experiments demonstrate that our method achieves state-of-the-art performance, both qualitatively and quantitatively, offering users an exemplar-driven inpainting tool with enhanced customization capability.

unedited

Key words: Diffusion model; Image inpainting; Exemplar-driven; Few-shot fine-tuning
https://doi.org/10.1631/FITEE.2400395 **CLC number:** TP

1 Introduction

Image inpainting is a typical image editing technique commonly used to modify local areas within an image, including object removal and replacement. Traditional inpainting algorithms based on patchmatch (Criminisi et al., 2004; Barnes et al., 2009) or generative adversarial networks(GAN) (Nazeri et al., 2019; Li et al., 2020) often do not support user-provided guidance signal, and the lack of controllability limits its further application. Recent years have seen breakthrough progress in artificial intelligence-generated content (AIGC) (Zhang et al., 2024), especially in image generation with the advent of large-scale text-to-image(T2I) diffusion models (e.g., Stable Diffusion (Rombach et al., 2022), DALLE2 (Ramesh et al., 2022), and Imagen (Saharia et al., 2022)). These models can generate high-quality and highly diverse images from the user-provided text. Due to the simple and intuitive nature of natural language, text-driven image editing has also evolved rapidly (e.g., p2pEdit (Hertz et al., 2022) and instructPix2pix (Brooks et al., 2023)). The use of text guidance has also been successfully applied in image inpainting (Wang et al., 2023; Xie et al., 2023), with the most notable method being the stable inpainting (SD-inpaint) (Rombach et al., 2022), allowing users

‡ Corresponding author
* project supported by the National Natural Science Foundation of China (No. 82027801)
© Zhejiang University Press 2024

to fill in local areas of an image using a simple natural prompt, which has become the most widely used and state-of-the-art text-driven inpainting tool.

Fig. 1 Text-driven inpainting (top) struggles to accurately describe the object's details, while exemplar-driven inpainting (bottom) can make it easier.

While natural language provides an intuitive approach to image editing, as the saying goes "a picture is worth a thousand words", even a detailed language description struggles to precisely convey detailed object features, such as in custom scenarios where users wish to inpaint specific items like their own toy (see the example in Fig. 1). Therefore, beyond conventional text-driven inpainting, a more effective solution would be the exemplar-driven inpainting - which allows users to provide a reference image (exemplar), enabling the model to insert the object from the exemplar into a background image. However, exemplar-driven inpainting is an under-explored topic, with the following two main existing strategies: (1) Textual Inversion(TxtInv) (Gal et al., 2022), which learns textual embedding from exemplar images and reuses it during inference, and (2) Paint-by-Example(PbE) (Yang et al., 2023), which relies on a dataset to train a model that directly accepts exemplar as input conditions during inference. Both methods face challenges in achieving high-fidelity inpainting. The limitation of TxtInv lies in the fact that merely learning a textual embedding still cannot adequately represent the reference image as the textual embedding contains very limited parameters that can be less expressive. Moreover, TxtInv employs the background blending technique to maintain the known area, which often leads to visually noticeable boundary artifacts. The limitation of PbE is its dependency on the dataset, which cannot cover all cases. When users provide personalized exemplars that are not present in the dataset, this method fails to produce high-fidelity results. Additionally, training on large-scale datasets is also labor-intensive.

To address these challenges, this study introduces a novel few-shot exemplar-driven inpainting framework via a parameter-efficient finetuned diffusion model, while retaining model's original text-driven inpainting capability. Essentially, we upgraded a pretrained SD-inpaint model with plug-and-play low-rank adaptation (LoRA) modules (Hu et al., 2021), enabling few-shot learning of user-provided samples with much less computational cost, and without compromising the original model. In contrast to TxtInv that encodes concepts into textual embeddings, our method encodes subject concepts directly into model weights, which learns the concept in a larger parameter space, providing stronger fitting capabilities and improved faithfulness to the inpainting results. Since our method is built on SD-inpaint without using external techniques to maintain the known area, thus our results do not show edge artifacts, unlike TxtInv. On the contrary, our approach offers advantages over paint-by-example's(PbE) large-scale dataset training instead of tuning a lightweight exemplar-specific LoRA module, and this module is dedicated to learn a specific subject concept from the given exemplar image, eliminating the need for large-scale dataset collection as well as the restriction to its training domain and bringing improved fidelity for customized inpainting task, which we will demonstrate in our comparison experiment (Section 4.2).

Building on few-shot LoRA fine-tuning, we proposed the following two additional techniques to further enhance the fidelity of the results: GPT-4v prompting and prior noise initialization. On the one hand, GPT-4v prompting involves generating detailed prompts for the exemplar using the GPT-4v model (Achiam et al., 2024), which is used for both fine-tuning and inference phases. This is inspired by a consensus in the field of large language models (LLMs): the quality of prompts is crucial for model performance (Lei et al., 2024; Zhou et al., 2024). We seek this as an improvement compared to the use of plain or rare word tokens as previously employed in the study by (Ruiz et al., 2023) and (Kumari et al., 2023). On the other hand, prior noise initialization refers to that instead of sampling from a random Gaussian noise (which is a standard practice in backward-diffusion sampling), and we opted to sample from the noise that contains exemplar prior information. Specifically, such prior noise is obtained via one-step denoising diffusion probabilistic mod-

unedited

els(DDPM) noising on a composite input sample, where the exemplar image is pasted into the masked region of the background. This operation ensures that the latent input retains information from the exemplar. Our ablation studies (Section 4.4) demonstrate that our GPT-4v prompting and prior noise initialization techniques can further help to improve the inpainting fidelity.

Our contributions are summarized as follows:

- We introduced a new few-shot exemplar-driven inpainting method to the AIGC community, which is achieved through parameter-efficient finetuned diffusion model.
- We proposed the techniques of GPT-4v prompting and prior noise initialization to foster more robust and high-fidelity inpainting results.
- Experimental results demonstrate that our method outperforms existing state-of-the-art baselines both qualitatively and quantitatively.

2 Related work

2.1 T2I diffusion model

Recently, T2I diffusion models have dominated the field of image synthesis, showcasing superior generation quality and diversity compared to classical GAN models (Dhariwal and Nichol, 2021). These models are formally referred as DDPM (Ho et al., 2020), which have been extensively trained on large image-text datasets such as LAION (Schuhmann et al., 2021), and are capable of reconstructing clean images that conform to the data distribution from Gaussian noise, guided by text prompts. Early T2I diffusion models mainly work in pixel space, such as GLIDE (Yang et al., 2020), DALLE2 (Ramesh et al., 2022) and ImagGen (Saharia et al., 2022), which are computationally expensive when generating high resolution images. In contrast, other models, such as Stable Diffusion (Rombach et al., 2022), Pixart-α (Chen et al., 2023), operate in a more compact and low-dimensional latent space, making them more memory-efficient and widely used. Among these, the early release of Stable Diffusion has enabled the development of numerous derivative models and a wide array of downstream tasks beyond T2I generation. For example, DreamBooth (Ruiz et al., 2023) and TxtInv (Gal et al., 2022) focus on subject-driven generation by optimizing model weights or text embeddings. Methods, such as ControlNet (Zhang et al., 2023), T2Iadapter (Mou et al., 2023), GLIGEN (Li et al., 2023), and IP-adapter (Ye et al., 2023) incorporate additional plug-in modules and train on task-specific datasets to achieve controllable image generation by using additional conditions, such as sketch maps, depth maps, skeletal poses, bounding boxes, and reference images, significantly enhancing the controllability and flexibility of the generative models.

2.2 Image inpainting

Early inpainting techniques primarily relied on the patch-match for texture synthesis (Criminisi et al., 2004; Barnes et al., 2009; He and Sun, 2014), which could only generate low-level textures and are inadequate for completion of high-level information. Subsequently, a series of models based on GAN backbones (Pathak et al., 2016) were developed, typically incorporating modules such as partial convolution (Liu et al., 2018), contextual attention (Yu et al., 2018), and a coarse-to-fine pipeline (Zhang et al., 2018; Yang et al., 2022). While these data-driven, learning-based approaches facilitated the handling of high-level semantic completion tasks, such as face inpainting, they are restricted to unconditional inpainting. The lack of support for user-guided conditions has constrained their further application.

Fortunately, T2I diffusion models have emerged as a powerful generative paradigm and have been applied to image inpainting. Notable text-driven inpainting models include zero-shot methods such as Blended Diffusion (Avrahami et al., 2022) and its latent-space variant (Avrahami et al., 2023), which achieve tuning-free inpainting through background blending. Training-based methods, such as GLIDE (Yang et al., 2020), SD-inpaint (Rombach et al., 2022), and Imagen editor (Wang et al., 2023), are finetuned from their base T2I models with modified structure on inpainting datasets. These methods allow for controllable inpainting through the use of text prompts.

In contrast, research on exemplar-driven inpainting has been relatively limited. PbE (Yang et al., 2023) is the first to implement this topic, but it alters the base model's text embedding interface to accommodate image embedding inputs, resulting in a loss of text-guiding capability. Additionally, its generaliz-

unedited

ability is limited by the scope of its training dataset. Another notable approach, TxtInv (Gal et al., 2022), employs text embedding learning with background blending (Avrahami et al., 2023) for subject-driven inpainting, yet this technique often leads to boundary artifacts during the blending process. IP-Adapter (Ye et al., 2023) adopts an image-to-image(I2I) model capable of performing local edits on source object when used in conjunction with the SDEdit (Meng et al., 2022) technique. However, it struggles to get satisfactory results when source object is absent. In contrast, our method is based on SD-inpaint, which inherently mitigates the introduction of boundary artifacts. By additionally fine-tuning the plug-and-play LoRA module on user-specific exemplar, our method can achieve high-fidelity exemplar-driven inpainting without the need for extensive training on large-scale datasets or modifications to the base model structure.

3 Methods

In this section, we first revisited preliminaries of the foundational model used in our approach, i.e., the SD-inpaint in Section 3.1. We then provided an overview of our method in Section 3.2, followed by description of the technical details from both fine-tuning and inference stages in Sections 3.3 and 3.4.

3.1 Preliminaries

The SD-inpaint model (Rombach et al., 2022) utilizes a 2D U-Net architecture, parameterized by $\boldsymbol{\theta}$. The standard training procedure involves adding Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to the input latent sample $\boldsymbol{z}$ to obtain a noised latent $\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}$ at timestep t, where α_t and σ_t are time-dependent DDPM hyperparameters. The U-Net considers the timestep t, text guidance $\boldsymbol{c}$, a mask $\boldsymbol{m}$ and a masked image latent $\boldsymbol{z}_m$ as conditions and predicts the noise $\boldsymbol{\epsilon}_\theta$. The model is optimized using the following denoising loss:

$$\mathcal{L} = \mathbb{E}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \boldsymbol{c}, \boldsymbol{z}_m, \boldsymbol{m}, t)\|_2^2\right]. \tag{1}$$

During inference, the model initiates with a random Gaussian noise $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ as the latent input. Conditioned on guidance signal $\boldsymbol{c}$, $\boldsymbol{m}$ and $\boldsymbol{z}_m$, it progressively denoises the latent representation for T timesteps until it reaches $\boldsymbol{z}_0$ through the following denoising diffusion implict model(DDIM) (Song et al., 2022) denoising process:

$$\boldsymbol{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left[\boldsymbol{z}_t - \frac{1-\alpha_t}{1-\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \boldsymbol{c}, \boldsymbol{z}_m, \boldsymbol{m}, t)\right], \tag{2}$$

where $\bar{\alpha}_t = \prod_1^t \alpha_t$. Finally, the denoised $\boldsymbol{z}_0$ is decoded back into the pixel space to obtain the inpainted image $\boldsymbol{x}_0 = \mathrm{Decode}(\boldsymbol{z}_0)$.

3.2 Overview

3.2.1 Task formulation

In this study, we focused on exemplar-driven inpainting. In this setup, the user provides an exemplar image $\boldsymbol{I}_{\mathrm{fg}}$ containing a specific foreground object, a background image $\boldsymbol{I}_{\mathrm{bg}}$, and a binary mask $\boldsymbol{m}$ (1 for area to be inpainted and 0 for area to be preserved). We aimed to train LoRA modules $\boldsymbol{\phi}$ for the foreground object on top of the frozen pretrained SD-inpaint model $\boldsymbol{\theta}$. The combined model $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ should be able to seamlessly integrate the object concept from $\boldsymbol{I}_{\mathrm{fg}}$ into the designated inpainted area of $\boldsymbol{I}_{\mathrm{bg}}$ while keeping the rest region unchanged. The output image should maintain a visually coherent effect in the filled area, avoiding a simple direct copy of the foreground object and allowing for a certain degree of variations in gesture, or orientation. Furthermore, the model still retains the capability of the original text-driven inpainting.

3.2.2 Overall pipeline

To achieve this purpose, we designed a two-stage fine-tuning-inference pipeline as shown in Fig. 2. During fine-tuning, the user provides an exemplar image. Based on the pretrained SD-inpaint base model with frozen weights, we introduced additional trainable LoRA modules. These modules are designed to learn the concept of the exemplar image by fine-tuning on its augmented variants (detailed in Section 3.3). Concurrently, text conditions are generated using GPT-4v, which we found beneficial for preserving details. During the inference, for each exemplar, we loaded its pretrained LoRA modules onto the SD-inpaint base model, thereby equipping the model with the acquired knowledge of the exemplar concept. When given with a background image to be inpainted inpainting and its corresponding mask, we then allowed the model to sample from the noise determined by our proposed prior noise initialization (detailed in Section 3.4). This further helps the model to inpaint

unedited

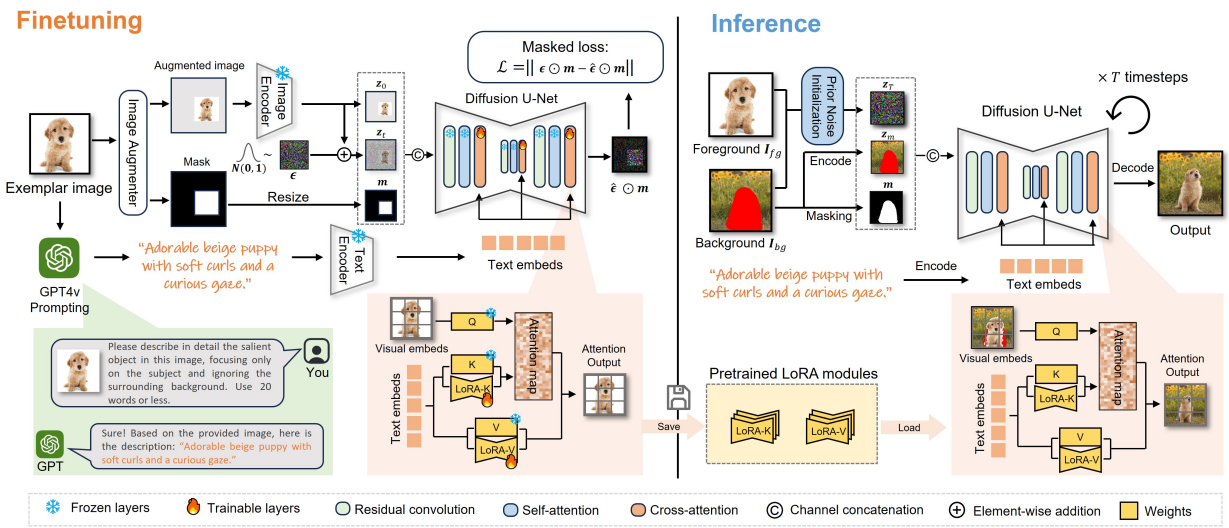

Fig. 2 Pipeline overview of our method. Based on frozen SD-inpaint model, the fine-tuning stage (left) involves fine-tuning learnable LoRA modules on a given exemplar image with GPT-4v generated prompt. The inference stage (right) first loads exemplar-specific pretrained LoRA modules and then samples from the prior noise to facilitate high-fidelity exemplar-driven inpainting within the masked region of the provided background image. LoRA, low-rank adaptation.

unedited

the learned exemplar concept with higher fidelity.

3.3 Fine-tuning stage

The fine-tuning stage involves learning the concept from given exemplar. This process includes the following three components: exemplar-specific LoRA, GPT-4v prompting, and masked fine-tuning loss.

3.3.1 Exemplar-specific LoRA

LoRA (Hu et al., 2021) is a method designed for efficiently modifying large models across various tasks, utilizing a unique approach for weight modification. In our application, LoRA learns the customized features from the exemplar image provided by the user. It operates on the principle that modifications $\boldsymbol{\phi}$, to the base model's weights $\boldsymbol{\theta} \in \mathbb{R}^{m \times n}$, have a "low intrinsic rank" and can be decomposed into two lower-rank matrices, $\boldsymbol{\phi}_B \in \mathbb{R}^{m \times r}$ and $\boldsymbol{\phi}_A \in \mathbb{R}^{r \times n}$. This factorization enables efficient parameterization with $\boldsymbol{\phi} = \boldsymbol{\phi}_B \boldsymbol{\phi}_A$, where $r \ll \min(m, n)$. During fine-tuning, only $\boldsymbol{\phi}_A$ and $\boldsymbol{\phi}_B$ are updated and $\boldsymbol{\theta}$ is kept unchanged, resulting in merged weight $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \boldsymbol{\phi}_B \boldsymbol{\phi}_A$. LoRA's efficiency makes it popular for fine-tuning large models, such as SD-inpaint in our case.

Now we determined which part of the model to apply the LoRA. A diffusion U-Net consists of various layers, including residual convolution layers, self-attention, and cross-attention layers. It has been well-established that cross-attention layers are crucial for textual responses (Hertz et al., 2022), as textual tokens $\boldsymbol{e}_{\text{txt}}$ (providing key and value) and vision tokens $\boldsymbol{e}_{\text{vis}}$ (providing query) attend to each other within these layers. Therefore, we opted to apply LoRA to cross-attention layers, specifically to the key and value projection matrices. Formally, the cross-attention operation in our method is defined as follows:

$$\text{CrossAttn}(\boldsymbol{e}_{\text{vis}}, \boldsymbol{e}_{\text{txt}}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{e}_{\text{vis}}[(\boldsymbol{K} + \boldsymbol{\phi}_K)\boldsymbol{e}_{\text{txt}}]^T}{\sqrt{d}}\right)[\boldsymbol{V} + \boldsymbol{\phi}_V]\boldsymbol{e}_{\text{txt}} \tag{3}$$

Here, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the original query, key, and value weight matrices, respectively, while $\boldsymbol{\phi}_K$ and $\boldsymbol{\phi}_V$ are the learnable LoRA key and value weight matrices, respectively. d is the dimension of the attention features.

3.3.2 GPT-4v prompting

Prompting is an extremely important factor in the output quality of generative models. Previous subject-driven generation tasks normally used "rare token + class noun" (Kumari et al., 2023; Ruiz et al., 2023) as the prompt template, which lacks detailed

information. In our task, we believe it is beneficial to select an expressive prompt for the exemplar image, which can greatly aid the model in producing the desired results. As such, we utilized GPT-4v for generating prompts for the exemplar image. The usage of GPT-4v prompting is shown in the green box in Fig. 2. Additionally, we also employed negative text prompts (e.g., blurry image, disfigured, bad anatomy, low resolution, deformed body features, poorly drawn face, and bad composition) to prevent the diffusion model from generating unwanted and chaotic results. These two types of prompts work cooperatively to inject semantic-level guidance into the model.

3.3.3 Masked fine-tuning loss

To learn the exemplar concept while preserving the model's existing knowledge, we froze the original weights $\boldsymbol{\theta}$ and trained only the newly added LoRA modules $\boldsymbol{\phi}$. The whole model is conditioned on the prompt $\boldsymbol{c}$ provided by GPT-4v. During fine-tuning, to mitigate overfitting with the few-shot data, we augmented the given exemplar image with random flips, scaling, and shifting, and we only supervised the predicted noise $\hat{\boldsymbol{\epsilon}}$ in the valid region, which denoted by a binary mask $\boldsymbol{m}$. Consequently, the fine-tuning process employs the following masked diffusion noise-prediction loss:

$$\mathcal{L} = \mathbb{E}\left[\|\boldsymbol{\epsilon} \odot \boldsymbol{m} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, \boldsymbol{c}, \boldsymbol{z}_m, \boldsymbol{m}, t\right) \odot \boldsymbol{m}\|_2^2\right]. \quad (4)$$

3.4 Inference stage

We now explain how to inpaint the background with the exemplar concept. This involves prior noise initialization and sampling.

3.4.1 Prior noise initialization

Commonly, the diffusion inference stage starts from a random noise. However, this scheme sometimes results in missing details in the output. We hypothesized that this issue arises due to a discrepancy between fine-tuning and inference phases. During fine-tuning, the model processes a noised latent input that blends noise and signal. This mixture retains some prior information about the input signal, even at high noise levels. In contrast, the inference typically involves sampling from purely random noise, creating a knowledge gap. To bridge this gap, we proposed the prior noise initialization, instead of sampling from a random noise, which samples from the noise that aligns with the fine-tuning scheme. This is achieved by applying one-step forward DDPM noising (Ho et al., 2020) to a composited input sample $\hat{\boldsymbol{z}}_0$, as shown below:

$$\boldsymbol{z}_T = \alpha_T \hat{\boldsymbol{z}}_0 + \sigma_T \boldsymbol{\epsilon} \quad (5)$$

where T is the last timestep. The composited input sample $\hat{\boldsymbol{z}}_0$ is created by copying the exemplar image and pasting it onto the bounding box area of the masked region in the background image. As a result, $mathbf z_T$ will contain the information from the exemplar. Note that this initialization still incorporates randomness; thus, the model is still able to generate diverse results. Fig. 3 illustrates the process of prior noise initialization.

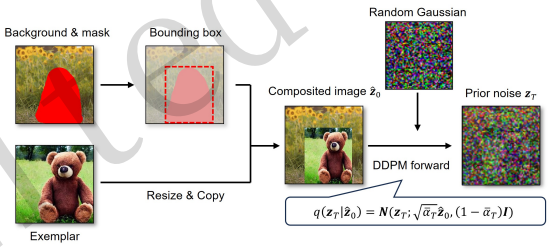

Fig. 3 Illustration of prior noise initialization.

3.4.2 Sampling

With the finetuned LoRA, the input background image, and its mask, the user can then engage in either exemplar-driven or text-driven inpainting:
(1) Exemplar-driven inpainting: with LoRA enabled, reuse of GPT-4v prompt, and prior noise obtained by Eq.(5).
(2) Text-driven inpainting: with LoRA disabled, customized prompt, and random noise.

In either case, the iterative sampling process is formulated by Eq.(2), and the final output result is decoded from the denoised latent $\boldsymbol{z}_0$.

4 Experiment

4.1 Experimental setup

4.1.1 Implementation details

Our method is based on Stable-inpaint-v1.5. We chose rank=16 for LoRA module, and fine-tuned it

unedited

for 300 iterations using the Adam optimizer (Kingma and Ba, 2017) for each exemplar image, with a batch size of 1 and a learning rate of 5e-5. For inference, we used the DDIM sampler (Song et al., 2022) with 50 steps and a classifier-free guidance of 8 (Ho and Salimans, 2022).

4.1.2 Dataset

We manually curated a benchmark dataset containing 192 pairs of exemplar-background-mask samples, which are constituted by 24 foreground exemplars, 23 background images and 23 masks, covering different categories of object type and scenes.

4.1.3 Baselines

We compared our method with recent state-of-the-art methods, including:
(1) SD-inpaint (Rombach et al., 2022): we used GPT-4v to generate prompts that simulate a user's detailed description for the desired inpainting results.
(2) PbE (Yang et al., 2023): a feed-forward model that incorporates exemplar embeddings as input conditions, facilitating example-driven inpainting tasks.
(3) TxtInv (Gal et al., 2022): this method focuses on customized generation through textual embedding optimization. It requires integration with latent blending (Avrahami et al., 2023) technique for localized inpainting.
(4) IP-adapter (Ye et al., 2023): A feed-forward model designed for image-guided generation, needs to be combined with SDEdit (Meng et al., 2022) for local editing purpose.
We used the official source code and models to reproduce the results for these baseline methods.

4.1.4 Metrics

We used the following popular metrics for evaluating the performance of the different methods.
(1) CLIP image-image similarity (CLIP-I) (Hessel et al., 2022): measures how output image is like the exemplar image by computing image-image similarity in CLIP image embedding space (Radford et al., 2021).
(2) CLIP text-image similarity (CLIP-T): measures how output image aligns with the text prompt by computing text-image cosine similarity in CLIP joint embedding space.
(3) Fréchet Inception Distance (FID) (Heusel et al., 2017): measures the technical quality of generated images by comparing with 1600 images from the LAION-art dataset (Schuhmann et al., 2021).
(4) Aesthetic score: measures the aesthetic quality of generated images by using the pretrained aesthetic evaluation model (Schuhmann, 2023).

4.2 Qualitative comparison

In Fig. 4, we presented a side-by-side comparison with these baselines, and the exemplar images contain objects and human faces. Our method generally achieves the most faithful results to the exemplar images, and our results exhibit the fewest artifacts. While all methods can achieve comparable results for commonly seen concepts (e.g., the cat and the dog examples), feed-forward-based methods such as PbE and IP-adapter struggle with less common or customized objects, such as the cartoon character and human faces, which may not be present in their training datasets. Additionally, since IP-adapter is originally designed for I2I generation rather than inpainting, it requires integration with SDEdit for local editing based on the source object. This can lead to more artifacts when the source object is unavailable, causing the inpainted object to conform unnaturally to the mask shape.

While SD-inpaint does not show artifacts, it was trained specifically for inpainting task that can inherently produce natural inpainting results. However, it fails to produce exemplar-like results even with detailed prompts, underscoring the importance of exemplar-driven inpainting task, which is challenging to achieve through text guidance alone.

In contrast, optimization-based approaches like our method and TxtInv demonstrate more aligned results compared to other baselines, indicating the significance of few-shot fine-tuning for customization purposes. Furthermore, our results preserve more details than TxtInv, as evidenced by the backpack, clock, and human face examples. We attribute this superiority to our method's encoding of exemplar concepts within LoRA weights, which possess a stronger fitting capability. This is due to the fact that the optimization space of the weights is much larger than TxtInv's text embedding space. Moreover, TxtInv also suffers from edge artifacts, i.e., the unnatural stitching along the object boundary, which is a consequence of the latent blending process where boundary transitions may not be fully harmonized. Our method avoids

unedited

Fig. 4 Qualitative comparison with related baselines. Each image grid shows results obtained from one exemplar image and three different background images. Zoom-in for best view.

Table 1 Qualitative comparison with related baselines. ↑ Higher is better. ↓ Lower is better.

Method	I2I score ↑	T2I score ↑	FID score ↓	Aesthetic score ↑
SD-Inpaint (Rombach et al., 2022)	65.59	28.10	149.62	5.0836
TxtInv (Gal et al., 2022)	70.86	**29.45**	146.89	5.0765
PbE (Yang et al., 2023)	69.47	28.79	150.45	**5.0864**
IP-adapter (Ye et al., 2023)	64.77	27.74	152.45	4.9558
Ours	**72.05**	29.31	**146.14**	5.0854

FID, Fréchet inception distance; I2I, Image-to-image; PbE, paint-by-example.

such issue by building on the SD-inpaint model.

4.3 Quantitative comparison

We reported I2I, T2I, FID, and aesthetic scores, as detailed in Table 1. Among these, the I2I score is particularly crucial for our exemplar-driven study because it measures the similarity between the output and the exemplar image. Notably, our method achieves the highest I2I score compared to all baselines, indicating that it preserves more details from the exemplar image, aligning well with the visual results as shown in Fig. 4.

For the T2I score, which assesses the alignment between the text prompt and the output, TxtInv emerges as the leader. We hypothesize that this is due to TxtInv's tendency to generate oversized objects that completely fill the mask shape, resulting in a stronger response to the text prompt. However, this approach often introduces edge artifacts caused by latent blending, as observed in cat and clock examples as shown in Fig. 4.

Regarding the FID and aesthetic scores, the performance across all baselines is relatively consistent, with no discernible significant differences. These findings imply that the baseline models may exhibit similar performance levels with respect to image quality and aesthetics.

4.4 Ablation studies

We conducted ablation studies to validate the components of our approach.

4.4.1 The effect of GPT-4v prompting

In Section 3.3.2, we introduced the use of GPT-4v prompting to achieve a more accurate initial textual description of the exemplar image, which can benefit the few-shot learning process. To demonstrate the advantages of utilizing GPT-4v prompting, we compare this approach with a pain prompt, "a photo of <x>", where <x> denotes the subject category noun. Our findings indicate that using detailed GPT-4v prompts retains more details from the exemplar image. As illustrated in Fig. 5, for the backpack example, the GPT-4v prompt not only specifies the object but also includes color constraint, indicating that the backpack should be red. Similarly, for the vase example, the GPT-4v prompt includes additional contextual details about the flowers inside the vase, successfully guiding the model to reproduce these elements in the output. This enhanced detailing in the prompts significantly contributes to the accuracy and richness of the inpainting results. We also reported quantitative ablation results in Table 2.

Table 2 Impact of GPT-4v prompting and prior noise initialization.

GPT-4v prompt	Prior noise	I2I score	T2I score
×	×	68.01	28.98
×	✓	69.20	29.14
✓	×	71.26	**29.34**
✓	✓	**72.05**	29.31

I2I, Image-to-image.

4.4.2 The effect of prior noise initialization

In Section 3.4.1, we proposed sampling from prior noise derived from the exemplar image instead of random noise. This prior noise incorporates some information from the exemplar, leading to more accurately aligned results. We illustrated this effect in Fig. 6, where we compared batches of results sampled with and without prior noise. When sampling without prior noise, the outputs are more prone to missing certain customized features from the exemplar, such as the color of the backpack and the clock is shifted in some examples. Conversely, initializing with prior

unedited

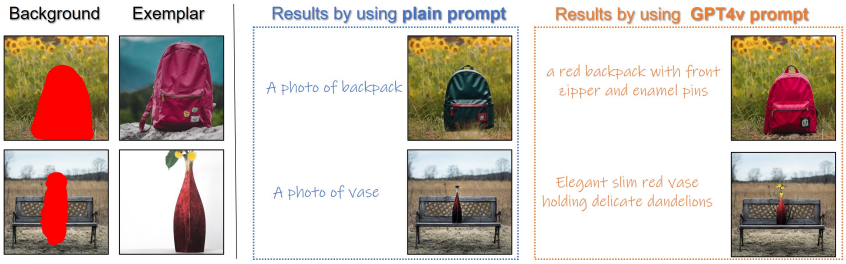

Fig. 5 Visual comparison between using plain prompt (blue box) vs. GPT-4v prompt (orange box).

noise results in more stable and consistent outputs. We also provided quantitative statistics in Table 2 to support our findings.

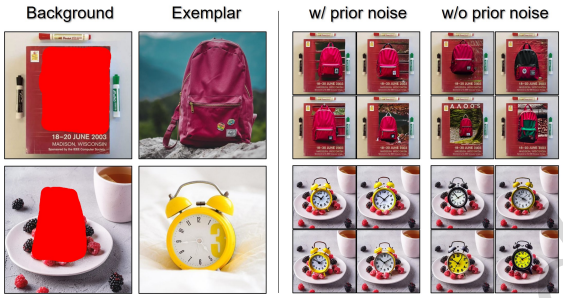

Fig. 6 Batched visual comparison between sampling with / without prior noise initialization.

4.4.3 Fine-tuning hyper-parameters

There are the following two key hyper-parameters to choose from during the LoRA fine-tuning process: LoRA rank and fine-tuning iterations. To determine a proper choice of them, we also conducted experiments with different LoRA ranks and fine-tuning iterations. The objective statistics are presented in Tables 3 and 4, respectively. For subjective results, we observed that a higher LoRA rank can lead to more aligned results in some instances, while in others, the differences are less pronounced. We selected a rank of 16 by default to balance effectiveness and LoRA size. On the contrary, the number of fine-tuning iterations has a more significant impact on the results. Increased fine-tuning steps result in more aligned outputs, as demonstrated in Fig. 7. We opted for 300 iterations, which we found to be sufficient for most cases.

Table 3 Performance analysis with different LoRA ranks.

LoRA rank	4	16	64
I2I score	71.94	72.05	**72.13**
# of params	**397k**	1591k	6365k
Speed* (iter/s)	1.49	1.49	1.49
GPU memory (MB)	**6380**	6394	6488

I2I, Image-to-image; LoRA, low-rank adaptation.
* ——Tested on single NVIDIA RTX3090 GPU.

Table 4 Performance analysis with different fine-tuning steps.

# of iteration	100	200	300	400
I2I score	71.26	71.73	72.05	**72.12**
time* (s)	**152**	305	457	611

I2I, Image-to-image.
* ——Tested on single NVIDIA RTX3090 GPU.

5 Limitations

While our method shows the capability for few-shot exemplar-driven inpainting, we have identified several limitations in some specific scenarios:

5.1 Failure in small masked areas

When the masked area is small, our method sometimes fails to inpaint the object effectively, as illustrated in Fig. 8a. For example, when we attempt to insert a can into a small area on the table, the result contains unexpected artifacts. We believe this issue is inherited from the base SD-inpaint model.

Previous studies (Wang et al., 2023; Xie et al., 2023) revealed that the SD-inpaint model sometimes ignores the text prompt, particularly when the masked area is small. The underlying reason is that SD-inpaint was originally trained with randomly generated masks, which often cover image regions unrelated to the text prompt. Training on such

unedited

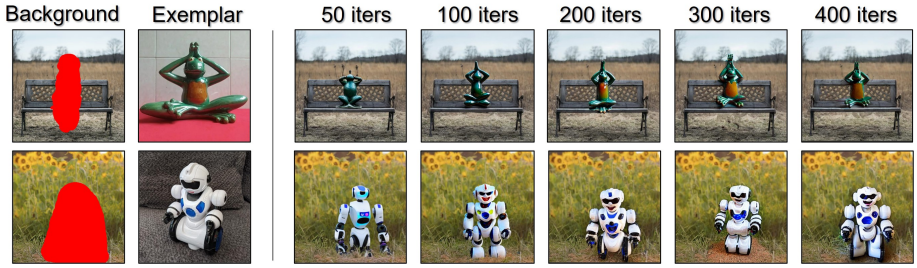

Fig. 7 The effect of fine-tuning iteration. More fine-tuning steps result in more aligned outputs.

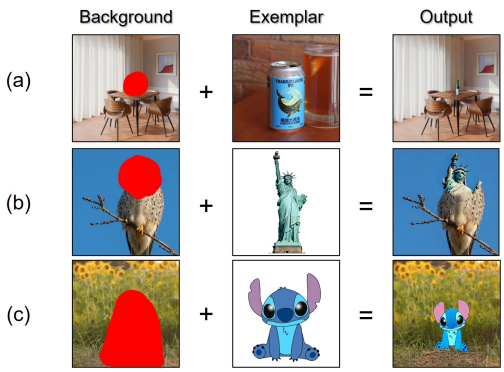

Fig. 8 Failure cases: (a) failed to inpaint small masked areas. (b) Failed to inpaint the exemplar image that is semantic incompatible with the background. (c) Failed to achieve harmonized inpainting across style gap.

image-mask pairs encourages the model to ignore the prompt, leading to reduced textual responsiveness.

We proposed two potential solutions for this issue, which we leave for future work: (1) regional attention amplification, as suggested in the study by (Chefer et al., 2023). During the sampling process, we explicitly increased the cross-attention values of the text prompt within the masked area to enhance the presence of the target object. (2) cropping and upsampling the masked area to increase its size before inpainting. After inpainting, the result can be resized back and inserted into the original background.

5.2 Exemplar-background semantic incompatibility

Another limitation arises when the exemplar image is semantically incompatible with the background context. As illustrated in Fig. 8b, attempting to inpaint the Statue of Liberty onto a bird's body results in unrealistic outputs. This is because the base model was pretrained on natural images and has not been exposed to such unusual combinations, making it less capable of generating uncommon or semantically inconsistent outputs.

This issue reflects a fundamental limitation of exemplar-driven methods, where the model cannot handle large semantic mismatches between the exemplar and background. In such cases, user guidance becomes essential. This suggests that exemplar-driven methods may require human-in-the-loop systems to guide users toward feasible inputs. For example, an interactive feedback system could remind users to provide reasonable inputs and assist in adjusting or selecting suitable exemplars.

5.3 Exemplar-background style incompatibility

When there is a significant style difference between the exemplar and the background image, our method struggles to harmonize them (see Fig. 8c). For example, trying to inpaint a cartoon-style exemplar into a real-world scene can result in incoherent outputs. This issue is common across many generative models.

Possible mitigation strategies include the following: (1) applying style transfer techniques to pre-align the styles of the exemplar and background before inpainting. (2) Integrating a style-aware fine-tuning step into the LoRA module. Specifically, we could finetune style-specific LoRAs, similar to how we handle exemplars. During inference, we could mix a subject LoRA with a style LoRA to achieve style-aligned exemplar-driven inpainting.

6 Conclusions

We presented a novel approach for few-shot exemplar-driven inpainting via parameter-efficient finetuned diffusion model, allowing users to provide a reference image for customized inpainting. Specifi-

unedited

cally, our approach enhances a pretrained SD-inpaint model with learnable LoRA modules, which are dedicated to learning customized concepts from the exemplar, facilitating efficient fine-tuning with less computational cost while preserving base model's integrity. We also introduced GPT-4v prompting and prior noise initialization to further assist the detail preservation. Our experimental results and ablation studies confirm that these innovations lead to more robust and high-fidelity inpainting outcomes, outperforming existing state-of-the-art methods both qualitatively and quantitatively. We believe this work will contribute to the field of AIGC by providing an adaptable, customized, and effective exemplar-driven image editing solution.

Contributors

Shiyuan YANG and Zheng GU designed the research. Shiyuan YANG and Wenyue HAO processed the data. Shiyuan YANG and Zheng GU drafted the manuscript. Yi WANG, Huaiyu CAI and Xiaodong CHEN helped organize the manuscript. Shiyuan YANG, Wenyue HAO and Xiaodong CHEN revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

References

Achiam J, Adler S, Agarwal S, et al., 2024. Gpt-4 technical report. https://doi.org/10.48550/arXiv.2303.08774

Avrahami O, Lischinski D, Fried O, 2022. Blended diffusion for text-driven editing of natural images. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18187-18197, https://doi.org/10.1109/CVPR52688.2022.01767

Avrahami O, Fried O, Lischinski D, 2023. Blended latent diffusion. *ACM Trans Graph*, 42(4):article 149. https://doi.org/10.1145/3592450

Barnes C, Shechtman E, Finkelstein A, et al., 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans Graph*, 28(3):article 24. https://doi.org/10.1145/1531326.1531330

Brooks T, Holynski A, Efros AA, 2023. Instructpix2pix: Learning to follow image editing instructions. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18392-18402, https://doi.org/10.1109/CVPR52729.2023.01764

Chefer H, Alaluf Y, Vinker Y, et al., 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans Graph*, 42(4):article 148. https://doi.org/10.1145/3592116

Chen JS, Yu JC, Ge CJ, et al., 2023. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. https://doi.org/10.48550/arXiv.2310.00426

Criminisi A, Pérez P, Toyama K, 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans Image Process*, 13(9):1200-1212. https://doi.org/10.1109/TIP.2004.833105

Dhariwal P, Nichol A, 2021. Diffusion models beat gans on image synthesis. Proc 35th Int Conf on Neural Information Processing Systems, p.8780-8794.

Gal R, Alaluf Y, Atzmon Y, et al., 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. https://doi.org/10.48550/arXiv.2208.01618

He KM, Sun J, 2014. Image completion approaches using the statistics of similar patches. *IEEE Trans pattern anal mach intell*, 36(12):2423-2435. https://doi.org/10.1109/TPAMI.2014.2330611

Hertz A, Mokady R, Tenenbaum J, et al., 2022. Prompt-to-prompt image editing with cross attention control. https://doi.org/10.48550/arXiv.2208.01626

Hessel J, Holtzman A, Forbes M, et al., 2022. Clipscore: A reference-free evaluation metric for image captioning. https://doi.org/10.48550/arXiv.2104.08718

Heusel M, Ramsauer H, Unterthiner T, et al., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Proc 31st Int Conf on neural information processing systems, p.6629-6640.

Ho J, Salimans T, 2022. Classifier-free diffusion guidance. https://doi.org/10.48550/arXiv.2207.12598

Ho J, Jain A, Abbeel P, 2020. Denoising diffusion probabilistic models. Proc 34th Int Conf on Neural Information Processing Systems, p.6840-6851.

Hu EJ, Shen YL, Wallis P, et al., 2021. Lora: Low-rank adaptation of large language models. https://doi.org/10.48550/arXiv.2106.09685

Kingma DP, Ba J, 2017. Adam: A method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980

Kumari N, Zhang B, Zhang R, et al., 2023. Multi-concept customization of text-to-image diffusion. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1931-1941, https://doi.org/10.1109/CVPR52729.2023.00192

Lei YM, Li JQ, Li ZL, et al., 2024. Prompt learning in computer vision: a survey. *Front Inform Technol Electron Eng*, 25(1):42-63. https://doi.org/10.1631/FITEE.2300389

Li JY, Wang N, Zhang LF, et al., 2020. Recurrent feature reasoning for image inpainting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7757-7765, https://doi.org/10.1109/CVPR42600.2020.00778

Li YH, Liu HT, Wu QY, et al., 2023. Gligen: Open-set grounded text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22511-22521, https://doi.org/10.1109/CVPR52729.2023.02156

Liu GL, Reda FA, Shih KJ, et al., 2018. Image inpainting for irregular holes using partial convolutions. Proc 15th European Conf on computer vision, p.89-105, https://doi.org/10.1007/978-3-030-01252-6_6

unedited

Meng CL, He YT, Song Y, et al., 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations.
https://doi.org/10.48550/arXiv.2108.01073

Mou C, Wang XT, Xie LB, et al., 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models.
https://doi.org/10.48550/arXiv.2302.08453

Nazeri K, Ng E, Joseph T, et al., 2019. Edgeconnect: Generative image inpainting with adversarial edge learning.
https://doi.org/10.48550/arXiv.1901.00212

Pathak D, Krähenbühl P, Donahue J, et al., 2016. Context encoders: Feature learning by inpainting. Proc IEEE Conf on computer vision and pattern recognition, p.2536-2544,
https://doi.org/10.1109/CVPR.2016.278

Radford A, Kim JW, Hallacy C, et al., 2021. Learning transferable visual models from natural language supervision. Proc 38th Int Conf on machine learning, p.8748-8763.

Ramesh A, Dhariwal P, Nichol A, et al., 2022. Hierarchical text-conditional image generation with clip latents.
https://doi.org/10.48550/arXiv.2204.06125

Rombach R, Blattmann A, Lorenz D, et al., 2022. High-resolution image synthesis with latent diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10674-10685,
https://doi.org/10.1109/CVPR52688.2022.01042

Ruiz N, Li YZ, Jampani V, et al., 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22500-22510,
https://doi.org/10.1109/CVPR52729.2023.02155

Saharia C, Chan W, Saxena S, et al., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Proc 36th Int Conf in Neural Information Processing Systems, p.36479-36494.

Schuhmann C, Vencu R, Beaumont R, et al., 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.
https://doi.org/10.48550/arXiv.2111.02114

Schuhmann C, 2023. christophschuhmann/improved-aesthetic-predictor.
https://github.com/christophschuhmann/improved-aesthetic-predictor

Song JM, Meng CL, Ermon S, 2022. Denoising diffusion implicit models.
https://doi.org/10.48550/arXiv.2010.02502

Wang S, Saharia C, Montgomery C, et al., 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18359-18369,
https://doi.org/10.1109/CVPR52729.2023.01761

Xie S, Zhang Z, Lin Z, et al., 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22428-22437,
https://doi.org/10.1109/CVPR52729.2023.02148

Yang BX, Gu SY, Zhang B, et al., 2023. Paint by example: Exemplar-based image editing with diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18381-18391,
https://doi.org/10.1109/CVPR52729.2023.01763

Yang H, Zhang RM, Guo XB, et al., 2020. Towards photo-realistic virtual try-on by adaptively generating ⟷ preserving image content. Proc of the IEEE/CVF Conf on computer vision and pattern recognition, p.7847-7856,
https://doi.org/10.1109/CVPR42600.2020.00787

Yang SY, Wang Y, Cai HY, et al., 2022. Residual inpainting using selective free-form attention. *Neurocomputing*, 510:149-158.
https://doi.org/10.1016/j.neucom.2022.09.041

Ye H, Zhang J, Liu SB, et al., 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.
https://doi.org/10.48550/arXiv.2308.06721

Yu JH, Lin Z, Yang JM, et al., 2018. Generative image inpainting with contextual attention. Proc of the IEEE Conf on computer vision and pattern recognition, p.5505-5514,
https://doi.org/10.1109/CVPR.2018.00577

Zhang HR, Hu ZZ, Luo CZ, et al., 2018. Semantic image inpainting with progressive generative networks. Proc of the 26th ACM int Conf on Multimedia, p.1939-1947,
https://doi.org/10.1145/3240508.3240625

Zhang JP, Sun LY, Jin C, et al., 2024. Recent advances in artificial intelligence generated content. *Front Inform Technol Electron Eng*, 25(1):1-5.
https://doi.org/10.1631/FITEE.2410000

Zhang LM, Rao AY, Agrawala M, 2023. Adding conditional control to text-to-image diffusion models. Proc IEEE/CVF Int Conf on Computer Vision, p.3813-3824,
https://doi.org/10.1109/ICCV51070.2023.00355

Zhou J, Ke P, Qiu XP, et al., 2024. Chatgpt: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, 25(1):6-11.
https://doi.org/10.1631/FITEE.2300089

unedited