# ByteEdit: Boost, Comply and Accelerate Generative Image Editing

Yuxi Ren*, Jie Wu*†, Yanzuo Lu*, Huafeng Kuang, Xin Xia,
Xionghui Wang, Qianqian Wang, Yixing Zhu, Pan Xie, Shiyin Wang,
Xuefeng Xiao, Yitong Wang, Min Zheng, and Lean Fu

ByteDance
Project Page: https://byte-edit.github.io

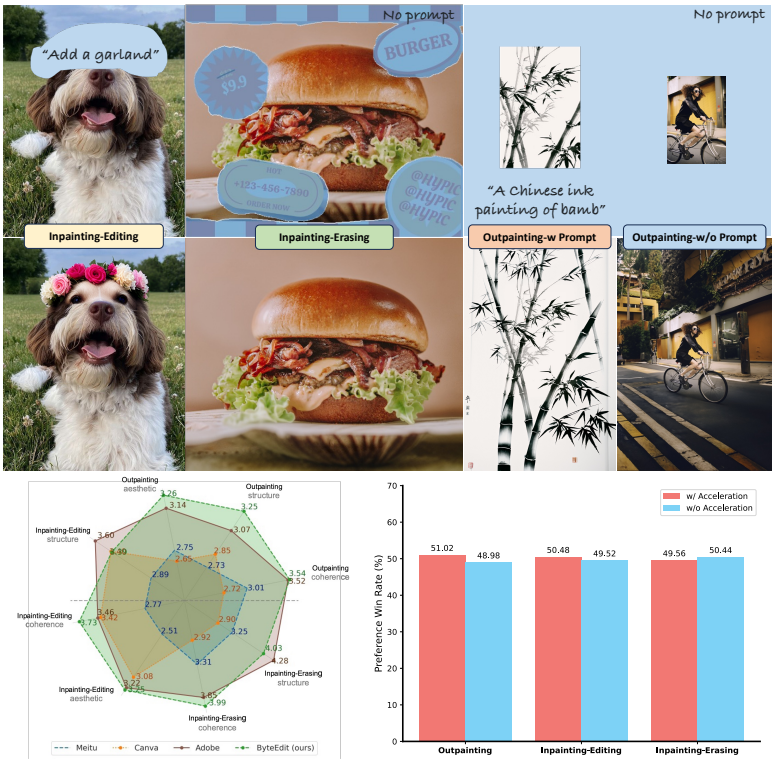



**Fig. 1:** We introduce ***ByteEdit***, a novel framework that utilizes feedback learning to enhance generative image editing tasks, resulting in outstanding generation performance, improved consistency, enhanced instruction adherence, and accelerated generation speed. To the best of our knowledge, ByteEdit emerges as the most *superior* and the *fastest* solution currently in the field of generative editing.

*Equal contribution. †Corresponding author.

**Abstract.** Recent advancements in diffusion-based generative image editing have sparked a profound revolution, reshaping the landscape of image outpainting and inpainting tasks. Despite these strides, the field grapples with inherent challenges, including: i) inferior quality; ii) poor consistency; iii) insufficient instrcution adherence; iv) suboptimal generation efficiency. To address these obstacles, we present ***ByteEdit***, an innovative feedback learning framework meticulously designed to ***B***oost, Compl***y***, and Accelera***te*** Generative Image ***Edit***ing tasks. ByteEdit seamlessly integrates image reward models dedicated to enhancing aesthetics and image-text alignment, while also introducing a dense, pixel-level reward model tailored to foster coherence in the output. Furthermore, we propose a pioneering adversarial and progressive feedback learning strategy to expedite the model's inference speed. Through extensive large-scale user evaluations, we demonstrate that ByteEdit surpasses leading generative image editing products, including Adobe, Canva, and MeiTu, in both generation quality and consistency. ByteEdit-Outpainting exhibits a remarkable enhancement of **388%** and **135%** in quality and consistency, respectively, when compared to the baseline model. Experiments also verfied that our acceleration models maintains excellent performance results in terms of quality and consistency.

**Keywords:** Outpainting · Intpainting · Feedback Learning

## 1 Introduction

The field of generative image editing has experienced remarkable advancements in recent years [5, 6, 20, 21, 23, 34, 37, 38, 42, 44, 45], propelled by the development of diffusion models [12, 25, 27, 28, 47]. This progress has led to the emergence of influential products that have reshaped the landscape of image editing. A notable example is Adobe Firefly [1], which has revolutionized the creative process by enabling users to seamlessly incorporate, extend, or remove content from images through simple text prompts, thereby transcending the traditional boundaries of Photoshop. In our paper, we focus on the domain of generative image editing, with particular emphasis on two key aspects: 1) *Outpainting*, involving the expansion of the surrounding scene in an image based on provided input or even without explicit prompts, and 2) *Inpainting*, encompassing the random masking of specific image regions followed by the generation of corresponding content guided by given prompts (Inpainting-Editing) or erase certain objects (Inpainting-Erasing). Despite the notable advancements achieved through diffusion-based algorithms, several challenges persist within this field:

***Inferior Quality***: the quality of generated images frequently falls short in terms of realism, aesthetic appeal, and fidelity to minute details.

***Insufficient Instruction Adherence***: The existing models grapple with the arduous task of faithfully adhering to provided instructions, resulting in a lack of alignment between the generated image and the input text;

***Poor Consistency***: The generated regions exhibit an unsatisfactory level of coherence with the original image, manifesting as a deficiency in terms of color, style, texture, and other salient visual attributes;
***Suboptimal Generation Efficiency***: The generation process is characterized by sluggish speeds and inadequate efficiency, thereby imposing significant obstacles when confronted with large-scale image editing endeavors.

Recently, various efforts have been made to address the aforementioned challenges in the field. For instance, Imagen Editor [34] has employed an object detection approach to extract inpainting masks, while simultaneously capitalizing on original high-resolution images to faithfully capture intricate details. SmartBrush [37] has adopted a multi-task training strategy coupled with precision controls, encompassing both text and shape guidance, to enhance visual quality, mask controllability, and preserve the background. Additionally, RePaint [21] has utilized a pre-trained unconditional DDPM [12] prior and ingeniously modified the reverse diffusion iterations to generate high-quality and diverse output images. However, these approaches have primarily focused on addressing singular problems and have yet to achieve a more comprehensive solution. Large Language Models (LLMs) has made a notable surge in incorporating learning based on human feedback, and initial endeavors have been undertaken in the Text-to-Image (T2I) domain [9,17,39,43,46,48]. Inspired by these developments, we pose the question: *Can we leverage human feedback to guide generative image editing to unleash the potential for superior generation outcomes?*

This paper introduces ByteEdit, an innovative framework for optimizing generative image editing through the incorporation of feedback learning. ByteEdit builds multiple reward models, namely the Aesthetic reward model, Alignment reward model, and Coherent reward model, to achieve exceptional generation effects, improved instruction adherence and enhanced consistency, respectively. These carefully designed reward models serve as the foundation for our proposed perceptual feedback learning (PeFL) mechanism, which provides task-aware and comprehensive supervision signals. Moreover, ByteEdit introduce an adversarial feedback learning strategy that employs the trainable reward model as the discriminator. This strategy ensures that the model benefits from the PeFL supervision and provide clear images even during high-noise stages, further improves both the performance and speed of our model. To expedite the sampling process, a progressive training strategy is empolyed to gradually reduce the optimization time steps and facilitate model inference in a low-steps regime.

- *New Insight*: To the best of our knowledge, we offer the first attempt to incorporate human feedback into the field of generative image editing. ByteEdit significantly enhances the overall performance of the model across various key aspects, opening new horizons in this field of study.
- *Comprehensive Framework*: By designing complementary global-level and pixel-level reward models, we effectively guide the model towards achieving improved beauty, enhanced consistency, and superior image-text alignment.
- *Efficiency and Pioneering*: Progressive feedback and adversarial learning techniques are introduced to accomplish a remarkable acceleration in the

model's inference speed, all while maintaining a minimal compromise on output quality. Notably, ByteEdit stands as the first successful endeavor in accelerating generative editing models.
– *Outstanding Performance*: Extensive user studies show that ByteEdit exhibits obvious advantages in terms of quality, consistency, efficiency, and speed, compared to the most competitive products. ByteEdit emerges as the fastest and most superior solution currently available in image editing.

## 2 Related Work

**Generative Image Editing.** Generative Image Editing is a research area focused on filling user-specified regions of interest with desired contents. GLIDE [23] is the pioneering work that introduced text-to-image diffusion for editing purposes, and Repaint [21], on the other hand, conditions an unconditionally trained model (e.g. DDPM [12]) and leverages visible pixels to fill in missing areas. To enable more precise editing control, Blended Diffusion [5] incorporates multimodal embeddings and enforces similarity between images and text using CLIP [26]. SmartBrush [37] pushes the boundaries of mask generation by utilizing instance and panoptic segmentation datasets instead of random generation. Further improvements include the introduction of the powerful Segment Anything (SAM) [15] model by [45], which achieves mask-free removal, filling, and replacing of multiple pipelines. Inst-Inpaint [44] specializes in text-driven object removal without the need for explicit masking. Additionally, this method proposes the GQA-Inpaint dataset, which comprises image pairs with and without the object of interest, facilitating effective object removal guided by textual input. In addition to comparing our proposed method with existing academic approaches, we also benchmark against industry methods like Adobe [1], Canva [2], and MeiTu [3], providing a comprehensive evaluation across different domains and highlighting the strengths of our approach.

**Human Feedback Learning.** Foundation models for text-to-image diffusion often rely on pre-training with large-scale web datasets, such as LAION-5B [31], which may result in generated content that deviates from human ethical values and legal compliance requirements. Previous approaches [9,17] attempted to address this issue by constructing preference datasets using hand-crafted prompts or expert generators. However, these methods suffered from over-fitting due to their limited real-world scenarios and generator capabilities. To overcome these limitations, researchers proposed various reward models trained with expert annotations [35,39] or feedback from web users [13,16] to enforce alignment with human preferences. Drawing inspiration from reinforcement learning with human feedback (RLHF) utilized in natural language processing (NLP), researchers explored the application of human feedback learning in text-to-image diffusion [39, 43, 46, 48] to achieve more realistic, faithful, and ethical outcomes. Among these efforts, ImageReward [39] primarily focused on overall image quality and overlooked the complexity of human perception. In our work, we extend the concept of human feedback learning by introducing three fine-grained indepen-

dent reward models tailored for generative image editing: aesthetics, image-text alignment, and pixel-level coherence.

## 3 ByteEdit: Boost, Comply and Accelerate

ByteEdit, focuses on generative image editing tasks that enable users to manipulate image content within a specific region of interest using textual descriptions. With an input image $x$, a region-of-interest mask $m$, and a user-provided textual description $c$, our primary objective is to generate an output image $y$ that preserves the unmasked region in the input image $x$, while aligning the masked region well with both the description of $c$ and visual attributes in $x$. In this study, we introduce two key functionalities within ByteEdit: ByteEdit-Outpainting and ByteEdit-Inpainting. ByteEdit-Outpainting extends the image by generating content beyond the boundaries of the input image, while ByteEdit-Inpainting fills or erases in arbitrary areas of the input image.

The ByteEdit pipeline is presented in Fig 2, providing an overview of the system's workflow. In the subsequent subsections, we delve into the details of two crucial components: Boost (Sec. 3.1) and Comply (Sec. 3.2). Furthermore, we elaborate on the Accelerate scheme in Sec. 3.3, which illustrates an approach to expedite the processing time and improve the efficiency of the ByteEdit system.

### 3.1 Boost: Perceptual Feedback Learning

In the field of generative image editing, the persistent challenge of subpar quality has impelled us to propose a pioneering approach that introduces human feedback, hitherto unexplored in this domain. Our novel pipeline comprises three key components: feedback data collection, reward model training, and perceptual feedback learning.

**Feedback Data Collection.** We first randomly extract more than 1,500,000 text prompts from the Midjourney Discord [33] and MS-COCO Caption [7] datasets. To ensure the diversity, a clustering algorithm, namely K-Means, was employed, leveraging the similarities derived from state-of-the-art large language models [19]. Further, the features were visualized in lower dimensions using t-SNE [22], enabling the identification of data points possessing the largest average distance from their k-nearest neighbors. We also manually eliminate less informative and decorative-dominanted prompts such as "unbelivable", "fantastic" and "brilliant" to improve the prompt quality. This meticulous procedure yielded approximately 400,000 candidate prompts, exhibiting diverse distributions, which were subsequently subjected to multiple text-to-image diffusion models, including SD1.5 [28] and SDXL [24]. Those images which excessively inferior quality or ambiguous characteristic are manually removed. Accompanying each prompt, a set of four generated images was presented to experts, who were tasked with selecting the best and worst images based on various aspects, encompassing aesthetic appeal, color composition, structural coherence, and brightness. The

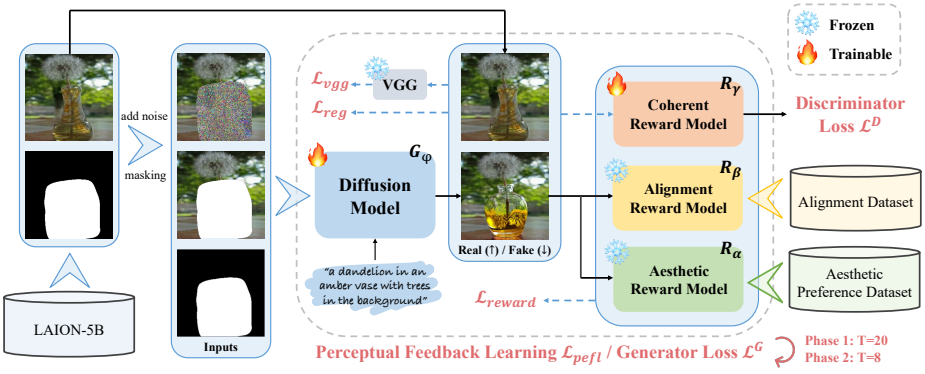



**Fig. 2:** ByteEdit formulates a comprehensive feedback learning framework that facilitating aesthetics, image-text matching, consistency and inference speed.

resulting dataset, herein referred to as the aesthetic preference dataset $\mathcal{D}_{aes}$, encompasses a collection of massive triplets $(c, x_p, x_n)$, where $x_p$ and $x_n$ correspond to the preferred and non-preferred generated images of prompt $c$, respectively.

**Reward Model Training.** Building upon this dataset, we follow the training techniques in [39] to learn an aesthetic reward model $R_\alpha(\cdot)$ of trainable parameters $\alpha$, which we briefly summarize here. The image and text features of the input are extracted from the BLIP [18] backbone, combined with cross attention, and fed into an MLP to obtain an aesthetic score. The training objective can be formulated as,

$$\mathcal{L}(\alpha) = -\mathbb{E}_{(c,x_p,x_n)\sim\mathcal{D}_{aes}}[\log \sigma(R_\alpha(c, x_p) - R_\alpha(c, x_n))], \tag{1}$$

where $\sigma(\cdot)$ represents the Sigmoid function used for normalization.

**Perceptual Feedback Learning.** Leveraging the power of our crafted reward model, we specifically introduce Perceptual Feedback Learning (PeFL) to fine-tune diffusion models with human feedback for generative image editing. Departing from the conventional practice of sequentially refining the predictions from the final step $x_T$ to the initial step $x'_0$ ( $x_T \rightarrow x_{T-1} \rightarrow \cdots \rightarrow x'_0$), we adopt an innovative perspective by performing optimization on the direct prediction outcomes $x_t \rightarrow x'_0$ at various intermediate steps $t \in [1, T]$ ($T = 20$ in this case). Through this dynamic exploration across different stages of denoising, we uncover the following distinctive observations:

- During the initial stages of denoising ($t \in [11, 20]$), the generative model (i.e. U-Net [29]) struggles to seamlessly complement the full image. Consequently, the reward model encounters challenges in accurately evaluating images that are hindered by obstacle masking.
- As the denoising process progresses ($t \in [1, 10]$), the reward model is able to identify and reward visually appealing images even in the presence of mild noise.

Drawing upon these insightful observations, we present an innovative stage-wise approach, to amplify the potential of generative image editing. Our proposed method encompasses the following key stages: 1) In stage 1 ($t \in [16, 20]$), we simply skip these steps with extremely heavy noise by diffusing the masked input $x \odot (1-m)$ into noisy latents at a fixed step $T_1 = 15$. This strategy is motivated by the fact that the generative model's ability to fill in intricate details within the masked region is limited in this timestep, thus rendering the training overhead unnecessary. This is the central difference between our approach and the ReFL proposed in [39], whose training paradigm relies solely on prompts and starts inference from pure noise. We aim to bring about a more pronounced correlation between the output and input image, thus facilitating the preservation of details in the unmasked areas; 2) In stage 2 ($t \in [t', 15]$), we randomly select a denosing step $t' \sim [1, T_2]$ ($T_2 = 10$ in this case) and perform inference without gradient starting from the noisy latent, i.e. $x_{T_1} \rightarrow x_{T_1-1} \rightarrow \cdots \rightarrow x_{t'}$. This method ensures that the complemented image generated at this stage exhibits a sufficiently high level of quality, rendering it more amenable to evaluation by the reward model; 3) In stage 3 ($x_{t'} \rightarrow x'_0$), the final stage entails the direct prediction of the complemented image $x'_0$. We leverage the aesthetic score obtained from the reward model as invaluable human feedback to refine the generative model $G_\phi(\cdot)$. This refinement process is achieved through the utilization of the following loss function:

$$\mathcal{L}_{reward}(\phi) = -\mathbb{E}_{(x,m,c)\sim\mathcal{D}_{train},t'\sim[1,T_2]}[\log \sigma(R_\alpha(c, G_\phi(x, m, c, t')))], \quad (2)$$

where $\mathcal{D}_{train}$ represents the fine-tuning dataset (i.e. LAION-5B [31]). The term $G\phi(x, m, c, t')$ denotes the decoded output image $x'_0$ generated by the generative model at step $t'$, given the masked input $x \odot (1-m)$ and the prompt $c$. To further maintain the consistency and detail fidelity of the generated area and the original image area,we introduce pixel-level regularization (i.e., L1 loss) and a perceptual loss, which captures the discrepancy in VGG features [32]. Collectively, these regularization techniques can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{reg}(\phi) &= \mathbb{E}_{(x,m,c)\sim\mathcal{D}_{train},t'\sim[1,T_2]}\|x - G_\phi(x, m, c, t')\|_1, \\ \mathcal{L}_{vgg}(\phi) &= \mathbb{E}_{(x,m,c)\sim\mathcal{D}_{train},t'\sim[1,T_2]}\|V(x) - V(G_\phi(x, m, c, t'))\|_1, \end{aligned} \quad (3)$$

where $V(\cdot)$ represents the VGG network. The overall training objective of our PeFL can be summarized as,

$$\mathcal{L}_{pefl}(\phi) = \mathcal{L}_{reward} + \eta(\mathcal{L}_{reg} + \mathcal{L}_{vgg}), \quad (4)$$

where $\eta$ is a hyperparameter for balancing loss weights.

### 3.2 Comply: Image-Text Alignment with Coherence

Diverging from the text-to-image synthesis focus of [39], our method encompasses an additional emphasis on assessing the **alignment** between the generated content of the masked area and the user-specified prompt, as well as

ensuring **coherence** with the unmasked region at the pixel level. To achieve this, we introduce two further components in this section, which complement the aesthetic reward model $R_\alpha$ proposed earlier.

**Image-Text Alignment.** We note that the presence of numerous poorly matched image-text pairs within the LAION dataset [31]. Exploiting these pairs as non-preferred samples for training reward models allows us to reduce manual annotation costs significantly. Initially, we employ the CLIP model [26] to identify image-text pairs with lower CLIPScore [11] from the LAION dataset. Subsequently, we leverage advanced multi-modal large language models such as LLAVA [19] to generate more informative and descriptive captions for the input images. These captions are considered more accurate than the original prompts. This process yields approximately 40,000 triplets $(c_p, c_n, x)$ as alignment dataset $\mathcal{D}_{align}$, where $c_p$ and $c_n$ correspond to the preferred and non-preferred textual descriptions of the image $x$, respectively. These triplets are utilized for training the image-text alignment reward model, denoted as $R_\beta(\cdot)$. The architecture of $R_\beta$ mirrors that of $R_\alpha$, while the training objective is similar to Eq. 1:

$$\mathcal{L}(\beta) = -\mathbb{E}_{(c_p,c_n,x)\sim\mathcal{D}_{align}}[\log\sigma(R_\beta(c_p,x) - R_\beta(c_n,x))], \tag{5}$$

**Pixel-Level Coherence.** The issue of coherence arises from the presence of inconsistent content within and outside the regions of interest, characterized by subtle visual cues such as color variations, stylistic discrepancies, and textural differences. To tackle this challenge, a coherent reward model, denoted as $R_\gamma(\cdot)$, is specifically designed for pixel-level discrimination, as opposed to the holistic evaluation performed by $R_\alpha(\cdot)$ and $R_\beta(\cdot)$. Our approach entails training a ViT-based [10] backbone network, followed by a prediction MLP head, to assess the authenticity and assign a score to each pixel in the input image. By formulating the loss function as follows:

$$\mathcal{L}(\gamma) = -\mathbb{E}_{(x,m,c)\sim\mathcal{D}_{train},t'\sim[1,T_2]}[\log\sigma(R_\gamma(z)) + \log(1-\sigma(R_\gamma(z')))], \tag{6}$$

where $z \sim x \in \mathbb{R}^{H\times W\times 3}$, $z' \sim G_\phi(x, m, c, t') \in \mathbb{R}^{H\times W\times 3}$ are pixels of the corresponding image and $H, W$ represent the height and weight respectively.

### 3.3 Accelerate: Adversarial and Progressive Training

**Adversarial training.** Concurrent works such as UFOGen [41] and SDXL-Turbo [30] proposed to introduce adversarial training objective into fine-tuning diffusion models, which dramatically speeds up the sampling process and allows for one-step generation. They supposed that the Gaussian assumption of diffusion process does not hold anymore when the inference steps are extremely low, and therefore enabling the generative model to output samples in a single forward step by adversarial objective [36, 40]. We note that the functionality of our coherent reward model $R_\gamma(\cdot)$ is very similar to that of the discriminator in adversarial training, except for the different granularity in prediction. To this end, unlike the aesthetic and alignment reward models, which necessitate offline learning prior to fine-tuning, the coherent reward model can be learned online

and seamlessly integrated into the fine-tuning process. The adversarial objective of **generator** that raises the score of output image is also in compliance with our feedback learning in Eq. (2), we can simply achieve adversarial training by incorporating the optimization of $R_\gamma(\cdot)$ into fine-tuning to serve as a **discriminator**. Thus the Eq. (2) can be reformulated as follows:

$$\mathcal{L}_{reward}(\phi) = - \mathop{\mathbb{E}}_{\substack{(x,m,c)\sim\mathcal{D}_{train} \\ t'\sim[1,T_2]}} \sum_{\theta\in\{\alpha,\beta,\gamma\}} \log\sigma(R_\theta(c, G_\phi(x,m,c,t'))). \tag{7}$$

For completeness, we also rewrite the overall training objective as,

$$\begin{aligned} &\mathcal{L}^G(\phi) = \mathcal{L}_{pefl}(\phi) = \mathcal{L}_{reward} + \eta(\mathcal{L}_{reg} + \mathcal{L}_{vgg}), \\ &\mathcal{L}^D(\gamma) = -\mathbb{E}_{(x,m,c)\sim\mathcal{D}_{train},t'\sim[1,T_2]}[\log\sigma(R_\gamma(z)) + \log(1-\sigma(R_\gamma(z')))]. \end{aligned} \tag{8}$$

**Progressive training.** To expedite the sampling process, we employ a progressive training strategy where we gradually reduce the optimization time steps $T$. Surprisingly, we find that the quality of the generated images does not significantly degrade under the supervisor of reward models. This approach strikes a fine balance between performance and speed, leading to compelling results. In our experiments, we adopt a two-phase progressive strategy. During phase 1, we set the optimization time steps as $T = 20$, $T_1 = 15$, and $T_2 = 10$. In phase 2, we further decrease the time steps to $T = 8$, $T_1 = 6$, and $T_2 = 3$. Notably, we achieve remarkable outcomes without employing any distillation operations, relying solely on the inheritance of model parameters.

## 4 Experiments

### 4.1 Implementation Details

**Dataset.** The fine-tuning dataset, denoted as $\mathcal{D}_{train}$, consists of a substantial collection of 7,562,283 images encompassing diverse domains such as real-world scenes, authentic portraits, and computer-generated (CG) images. To enhance the model's generalization ability and generation quality, we adopted a meticulous fine-grained masking strategy inspired by StableDiffusion [28]. Our masking strategy encompasses four distinct types of masks: global masks, irregular shape masks, square masks, and outward expansion masks. Each mask type corresponds to a specific probability value, which is randomly selected and applied to images during the training process. Moreover, we devised a specialized masking strategy tailored for Inpainting-Editing tasks. Leveraging instance-level data, which identifies specific objects within images, we introduced random dilation operations to coarsen the masks during training. These coarsened masks were then integrated with randomly generated masks surrounding the instances. This approach not only enhances the model's performance in instruction-based image editing tasks but also facilitates more accurate instance generation, ultimately leading to superior quality outputs.

To evaluate the performance of our approach, we conducted comprehensive qualitative and quantitative assessments using two distinct datasets. The first dataset, UserBench, was meticulously curated by gathering a vast amount of user-customized image-text matching data from various online sources. This dataset proved invaluable for evaluating image inpainting and outpainting tasks. From this extensive collection, we judiciously handpicked 100 high-quality image-text matching pairs to serve as our test data. We also leverage the experimental results from this dataset to collect and report human preferences. The second dataset, EditBench [34], presents a novel benchmark specifically tailored for text-guided image inpainting. Consisting of 240 images, each image within EditBench is paired with a corresponding mask that precisely delineates the region within the image to be modified through inpainting.

**Training Setting.** To facilitate the perceptual feedback learning stage, we employed a relatively small learning rate of 2e-06, complemented by a learning rate scheduling strategy that encompassed a warm-up phase consisting of 1000 iterations. Furthermore, to ensure stability in model parameter updates, we incorporated an exponential moving average (EMA) decay parameter set to 0.9999. Instead of employing 100% noise as in ReFL [39], we introduced a 50The weight assigned to the perceptual feedback loss was set to 0.01. During the adversarial acceleration stage, we maintained similar settings to the perceptual feedback learning stage, with an additional adversarial loss weighted 0.05.

### 4.2 Evaluation Principles and Criteria

**Subjective metrics.** To assess the robustness of our proposed method, we conducted a comprehensive subjective evaluation involving both expert evaluations and a large number of volunteer participants. Expert evaluators were tasked with individually assessing each generated image and assigning scores based on three key aspects: *coherence*, *structure*, and *aesthetics*. These aspects were rated on a scale of 1 to 5, with higher scores indicating superior generation quality: 1) *Coherence* focused on evaluating the consistency of texture, style, and color between the generated region and the original image. 2) *Structure* emphasized the clarity, sharpness, and absence of deformations or mutilations, particularly in human body parts. 3) *Aesthetics* gauged the overall level of creativity and diversity exhibited by the generated images. In addition to expert evaluations, we also sought the opinions of a large number of volunteers, with over 36,000 samples collected. These volunteers were presented with pairs of generated images and asked to choose between "Good", "Same", or "Bad", representing their preference in terms of GSB (Good-Same-Bad).

**Objective metrics.** In this study, we also incorporate objective text-image alignment metrics, specifically CLIPScore [11, 26] and BLIPScore [18], to comprehensively evaluate the alignment of our models.

**Table 1:** Comparisons with state-of-the-art generative image editing systems in terms of coherence, structure and aesthetic scored by experts. More than 6000 image-text pairs are randomly sampled for each task and we report the average scores.

| Method | Outpainting | | | Inpainting-Editing | | | Inpainting-Erasing | |
|---|---|---|---|---|---|---|---|---|
| | *coherence* | *structure* | *aesthetic* | *coherence* | *structure* | *aesthetic* | *coherence* | *structure* |
| MeiTu [3] | 3.01 | 2.73 | 2.75 | 2.77 | 2.89 | 2.51 | 3.31 | 3.25 |
| Canva [2] | 2.72 | 2.85 | 2.65 | 3.42 | 3.40 | 3.08 | 2.92 | 2.90 |
| Adobe [1] | 3.52 | 3.07 | 3.14 | 3.46 | **3.60** | 3.22 | 3.85 | **4.28** |
| **ByteEdit** | **3.54** | **3.25** | **3.26** | **3.73** | 3.39 | **3.25** | **3.99** | 4.03 |

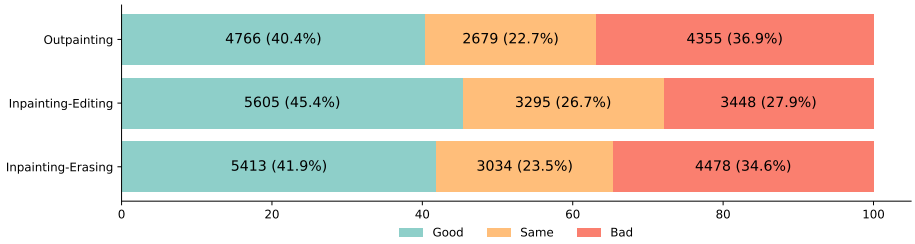



**Fig. 3:** Comparisons with state-of-the-art generative image editing systems in terms of human preference (i.e. GSB). More than 12,000 samples are collected for each task. For simplicity and to minimize the difficulty of collecting a large number of user opinions, we only offer the generated images by Adobe and our ByteEdit to the volunteers. "Good" indicates the generated images by our ByteEdit is preferred and vice versa.

### 4.3 Comparisons with State of the arts

We compare our method with concurrent state-of-the-art generative image editing systems such as Adobe [1], Canva [2] and MeiTu [3]. The comparisons cover three different tasks, including outpainting, inpainting-editing and inpainting-erasing. The inpainting editing will specify the content to be generated for the region of interest in the prompt. In contrast, inpainting-erasing requires the model to remove content within it and be as consistent as possible. Since the erased image has little change, experts were not asked to score aesthetics for user study.

**User study.** The average scores evaluated by experts are shown in Tab. 1. From the results, our ByteEdit significantly outperforms the state-of-the-art across different metrics in the outpainting task. It demonstrates that our method works well to expand images based on existing content and maintains superior consistency, structural integrity, and creativity. As for the inpainting tasks, our method also can provide the most coherent edited or erased images. To further investigate the gap between Adobe and our proposed ByteEdit, we solicited feedback from a large number of volunteers on the images generated by both, and the results are illustrated in Fig. 3. The results show that users generally found the images we generated to be more natural in overall perception. Our GSB

**Table 2:** The quantitative results of ByteEdit and recent state-of-the-art approaches.

| Metrics | UserBench | | | | EditBench | | | |
|---|---|---|---|---|---|---|---|---|
| | *Meitu* [3] | *Canva* [2] | *Adobe* [1] | ***ByteEdit*** | *DiffEdit* [8] | *BLD* [4] | *EMILIE* [14] | ***ByteEdit*** |
| CLIPScore | 0.235 | 0.241 | 0.237 | **0.255** | 0.272 | 0.280 | 0.311 | **0.329** |
| BLIPScore | 0.174 | 0.467 | 0.450 | **0.687** | 0.582 | 0.596 | 0.620 | **0.691** |

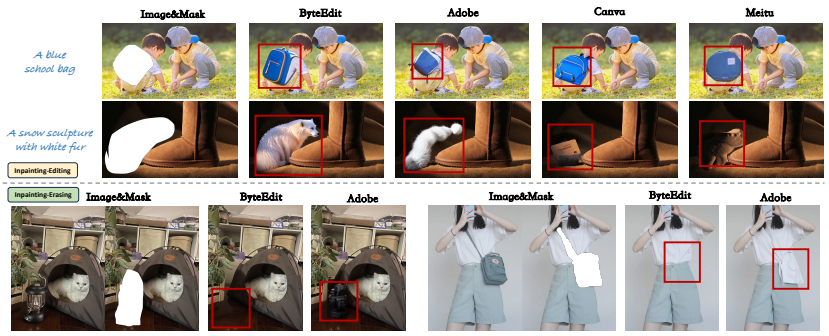



**Fig. 4:** Qualitative comparison in inpainting. We highlight key areas with red boxes.

superiority percentages (i.e. (G+S)/(S+B) * 100%) on three different tasks are 105%, 163%, and 112%, respectively.

**Quantitative comparison.** To quantitatively evaluate the performance of our method compared to other approaches, we conduct a quantitative evaluation of how well the edited image can capture the semantics of the edit instruction successfully by measuring the CLIPScore and BLIPScore. We conduct the experiment in inpainting-editing task and the results are provided in Table 2. From the UserBench against state-of-the-art generative image editing systems, we noticed that the score results are not completely consistent with user opinion. Nevertheless, our method is still ahead of the second-place Canva by **5.8%**(+0.014) and **47.1%**(+0.22) in terms of CLIPScore and BLIPScore, respectively. As for the EditBench, we follow [14] to compare our method with several concurrent editing approaches, i.e. DiffEdit [8], BLD [4] and EMILIE [14]. It demonstrates that the ByteEdit consistently yields the state-of-the-art performance, which shows our superior quality, consistency and instruction adherence.

**Qualitative comparison.** In Figure 4 and 5, we visualize samples produced by different systems under different tasks. It clearly shows that our method exhibits a superior performance for learning both coherence and aesthetic. For the inpainting task, the ByteEdit consistently follows the user-specified instructions and generates coherent images with better image-text alignment. It is worth noting that our system allows both prompt and prompt-free generation when it comes to outpainting, which has broader application scenarios in reality.

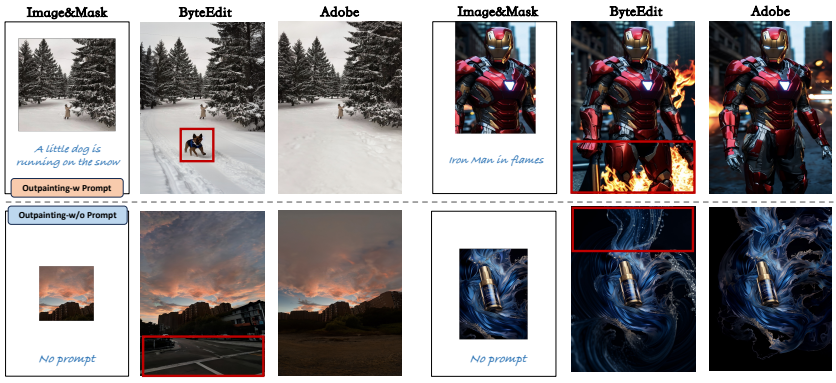
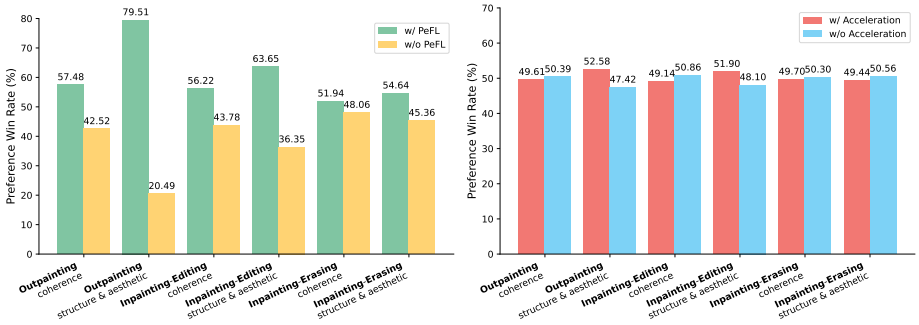



**Fig. 5:** Qualitative comparison in outpainting. We highlight key areas with red boxes.




**(a)** w/ vs w/o PeFL **(b)** w/ vs w/o Acceleration

**Fig. 6:** Human Perference Evaluation on our proposed PeFL and Acceleration.

### 4.4 Ablation Studies

In Figure 6, we conduct ablation experiments on both our proposed PeFL and acceleration strategy. The experts were asked to choose GSB preference and we report the human preference rates in the figure, which are calculated as (G+S)/[(G+S)+(S+B)] * 100% for win and (S+B)/[(G+S)+(S+B)] * 100% for lose, respectively. The evaluation is similar to the user study, except that we combine structure and aesthetics to reduce evaluation costs. More visualizations are also included in Figure 7.

**PeFL preference.** From the results in Figure 6(a), our proposed PeFL significantly improves the generation quality, outperforming the baseline on all different tasks. Especially in the outpainting task with PeFL, our method exceeds the baseline by about 60% in terms of structure and aesthetic, which is consistent with the edited image shown at the top of Figure 7 that is more realistic and conforms to the rules of realistic physics.

**Acceleration preference.** In Figure 6(b), we demonstrate that our model has no significant loss in either consistency or structure and aesthetic with the pro-

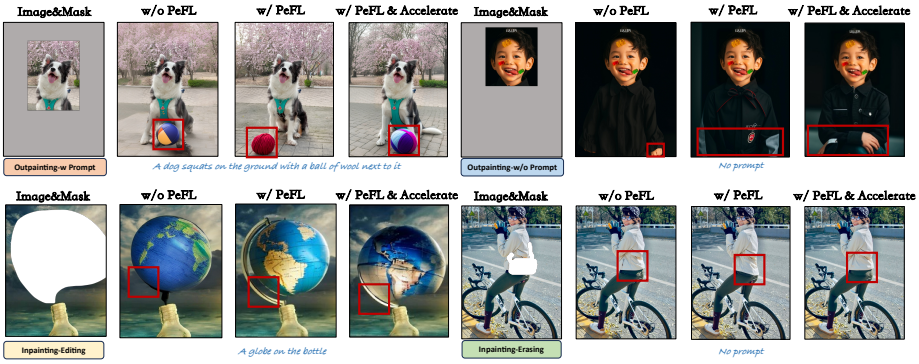



**Fig. 7:** Ablation Studies Visualization.

gressive training strategy. To our surprise, we have even achieved both increasing speed and quality in the outpainting and inpainting-editing tasks. Based on our experimental observations, this phenomenon can be attributed to two underlying factors: i) Stable Training: By considering the discriminator as a reward model, trainable reward model offers flexible and robust supervision for PeFL, alleviating issues related to model over-optimization; ii) Expand Supervision Scope: The incorporation of adversarial supervision enables us to extend the time step of PEFL optimization. Consequently, even at high-noise stages, such as step 999, the model can still benefit from PeFL supervision, further driving improvements in model performance. The visualization at the bottom of Figure 7 also verifies this, where the outputs become more realistic and natural after acceleration.

## 5 Discussion

ByteEdit has demonstrated remarkable performance across various editing tasks. However, several promising directions warrant further exploration:

- *Performance*: One avenue for improvement lies in developing more targeted reward models tailored to specific editing tasks. By refining the reward models, we can potentially unlock even higher levels of performance and generate more precise and desirable output.
- *Acceleration*: Another area of interest is investigating how ByteEdit can be further integrated with advanced techniques such as LCM and SDXL-turbo to achieve accelerated processing speeds.
- *Task*: Expanding the capabilities of ByteEdit beyond image editing to domains like video editing or instruction editing holds significant potential.

By incorporating human feedback to optimize generative image editing, ByteEdit can greatly enhance the practicality and usability in real-world scenarios. We hope that our work will provide valuable insights and inspire deeper reflections in this field, propelling its future development.

## References


1. Adobe firefly - free generative ai for creatives. https://www.adobe.com/products/firefly.html 2, 4, 11, 12
2. Free ai image generator: Online text to image app | canva. https://www.canva.com/ai-image-generator/ 4, 11, 12
3. Miraclevision. https://ai.meitu.com/index/ 4, 11, 12
4. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023) 12
5. Avrahami, O., Lischinski, D., Fried, O.: 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) 2, 4
6. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023) 2
7. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 5
8. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022) 12
9. Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767 (2023) 3, 4
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 8
11. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 8, 10
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 2, 3, 4
13. Isajanyan, A., Shatveryan, A., Kocharyan, D., Wang, Z., Shi, H.: Social reward: Evaluating and enhancing generative ai through million-user feedback from an online creative community. arXiv preprint arXiv:2402.09872 (2024) 4
14. Joseph, K., Udhayanan, P., Shukla, T., Agarwal, A., Karanam, S., Goswami, K., Srinivasan, B.V.: Iterative multi-granular image editing using diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8107–8116 (2024) 12
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 4
16. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024) 4
17. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023) 3, 4
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 6, 10

19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 5, 8
20. Lu, Y., Zhang, M., Ma, A.J., Xie, X., Lai, J.H.: Coarse-to-fine latent diffusion for pose-guided person image synthesis. arXiv preprint arXiv:2402.18078 (2024) 2
21. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022) 2, 3, 4
22. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 5
23. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 2, 4
24. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 5
25. Qin, J., Wu, J., Chen, W., Ren, Y., Li, H., Wu, H., Xiao, X., Wang, R., Wen, S.: Diffusiongpt: Llm-driven text-to-image generation system. arXiv preprint arXiv:2401.10061 (2024) 2
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4, 8, 10
27. Ren, Y., Wu, J., Zhang, P., Zhang, M., Xiao, X., He, Q., Wang, R., Zheng, M., Pan, X.: Ugc: Unified gan compression for efficient image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17281–17291 (2023) 2
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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) 2, 5, 9
29. Ronneberger, O., Fischer, P., Brox, T.: 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) 6
30. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023) 8
31. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 4, 7, 8
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 7
33. Turc, I., Nemade, G.: Midjourney user prompts & generated images (250k) (2022). https://doi.org/10.34740/KAGGLE/DS/2349267 5
34. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., et al.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18359–18369 (2023) 2, 3, 10

35. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2096–2105 (2023) 4
36. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. arXiv preprint arXiv:2112.07804 (2021) 8
37. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023) 2, 3, 4
38. Xie, S., Zhao, Y., Xiao, Z., Chan, K.C., Li, Y., Xu, Y., Zhang, K., Hou, T.: Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. arXiv preprint arXiv:2312.03771 (2023) 2
39. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. arXiv preprint arXiv:2304.05977 (2023) 3, 4, 6, 7, 10
40. Xu, Y., Gong, M., Xie, S., Wei, W., Grundmann, M., Hou, T., et al.: Semi-implicit denoising diffusion models (siddms). arXiv preprint arXiv:2306.12511 (2023) 8
41. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. arXiv preprint arXiv:2311.09257 (2023) 8
42. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023) 2
43. Yang, S., Chen, T., Zhou, M.: A dense reward view on aligning text-to-image diffusion with preference. arXiv preprint arXiv:2402.08265 (2024) 3, 4
44. Yildirim, A.B., Baday, V., Erdem, E., Erdem, A., Dundar, A.: Inst-inpaint: Instructing to remove objects with diffusion models. arXiv preprint arXiv:2304.03246 (2023) 2, 4
45. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023) 2, 4
46. Yuan, H., Chen, Z., Ji, K., Gu, Q.: Self-play fine-tuning of diffusion models for text-to-image generation. arXiv preprint arXiv:2402.10210 (2024) 3, 4
47. Zhang, M., Wu, J., Ren, Y., Li, M., Qin, J., Xiao, X., Liu, W., Wang, R., Zheng, M., Ma, A.J.: Diffusionengine: Diffusion model is scalable data engine for object detection. arXiv preprint arXiv:2309.03893 (2023) 2
48. Zhang, Z., Zhang, S., Zhan, Y., Luo, Y., Wen, Y., Tao, D.: Confronting reward overoptimization for diffusion models: A perspective of inductive and primacy biases. arXiv preprint arXiv:2402.08552 (2024) 3, 4