Energy-Guided Optimization for Personalized Image Editing with Pretrained Text-to-Image Diffusion Models

Rui Jiang[1]*, Xinghe Fu[1]*, Guangcong Zheng[1], Teng Li[1], Taiping Yao[2], Xi Li[1]†

[1]College of Computer Science and Technology, Zhejiang University
[2]Youtu Lab, Tencent
{jrss, xinghefu, guangcongzheng, tengli19}@zju.edu.cn, taipingyao@tencent.com, xilizju@zju.edu.cn

Abstract

The rapid advancement of pretrained text-driven diffusion models has significantly enriched applications in image generation and editing. However, as the demand for personalized content editing increases, new challenges emerge especially when dealing with arbitrary objects and complex scenes. Existing methods usually mistakes mask as the object shape prior, which struggle to achieve a seamless integration result. The mostly used inversion noise initialization also hinders the identity consistency towards the target object. To address these challenges, we propose a novel training-free framework that formulates personalized content editing as the optimization of edited images in the latent space, using diffusion models as the energy function guidance conditioned by reference text-image pairs. A coarse-to-fine strategy is proposed that employs text energy guidance at the early stage to achieve a natural transition toward the target class and uses point-to-point feature-level image energy guidance to perform fine-grained appearance alignment with the target object. Additionally, we introduce the latent space content composition to enhance overall identity consistency with the target. Extensive experiments demonstrate that our method excels in object replacement even with a large domain gap, highlighting its potential for high-quality, personalized image editing.

Introduction

With the rapid advancement of pretrained text-driven diffusion models (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022), the field of image generation has experienced unprecedented growth in guided image manipulation techniques through textual directives. The development of pretrained text-to-image (T2I) generation models has significantly enriched various domains, particularly image editing (Mou et al. 2024, 2023; Mokady et al. 2023), offering novel approaches to understanding and manipulating images. However, as the demand for personalized content editing increases, new challenges arise, particularly when dealing with arbitrary objects and complex scenes. Existing methods primarily focus on personalized image generation (Gal et al. 2022; Ruiz et al. 2023), aiming to create new images with customized content.

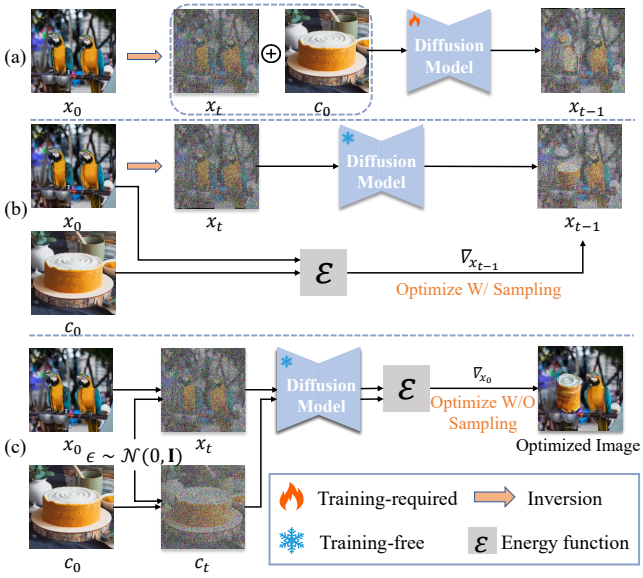

Figure 1: Comparisons among different methods in personalized content editing. (a) Inpainting-based methods usually require fine-tuning the diffusion model with reference images as the condition. (b) Sampling-based methods initialize the noise with inversion to maintain the background information from the source image. (c) The proposed method iteratively optimizes the latent code to perform training-free and inversion-free editing.

Personalized image editing involves placing a target object into a desired position within a scene image or replacing objects in a source image with those from a reference image. The primary challenges in this task are accurately integrating personalized content harmoniously into the target image and maintaining editing flexibility.

To address these challenges, there are two main categories of personalized image editing methods: inpainting-based and sampling-based (as shown in Fig. 1). These methods attempt to remove and regenerate objects to edit specific image regions. Other methods like SelfGuidance (Epstein et al. 2023) and Diffeditor (Mou et al. 2024) provide guidance in the sampling process and manipulate the sampling direction based on reference images.

*These authors contributed equally.
†Corresponding author.

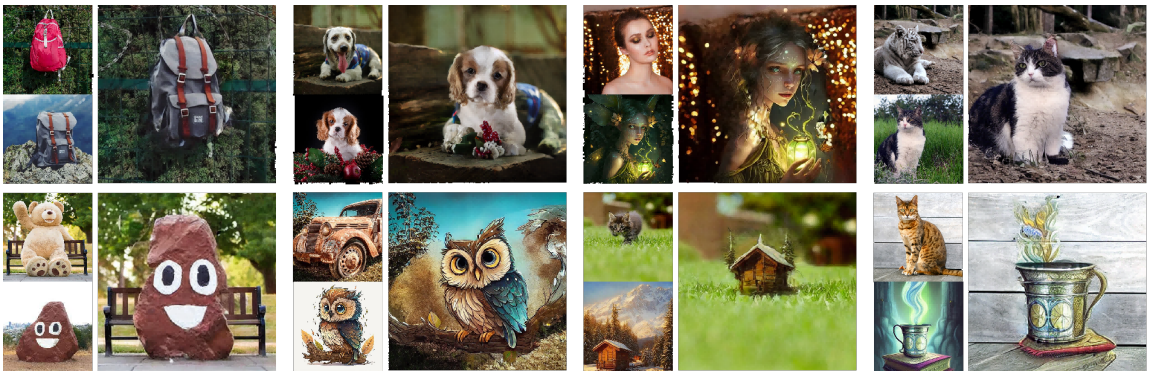

Figure 2: Performance overview of the proposed method in image customization editing. Our method generates edited images by integrating contextual guidance with a reference image. The first row demonstrates object replacement within the same category, while the second row shows object replacement across different categories.

Previous methods achieve impressive performance in some applications (*e.g.*, appearance replacement). However, some hard cases in personalized editing (e.g., cross-class object replacement) remain challenging for previous methods. Most inpainting-based methods like Anydoor (Chen et al. 2024) require extra training and treat the source mask as the target object shape prior. These cause unstable editing results when the shape mismatches between source and target objects or the test domain shifts. Sampling-based methods rely on the inversion of latent codes to maintain the background information from the source image. The sampling process (ODE or SDE) encounters the tradeoff between maintenance and flexibility, and makes it hard to present the target object in the edited image. Therefore, developing a training-free and inversion-free algorithm for personalized editing is in demand.

Unlike previous methods, we propose optimizing the latent code directly to obtain the edited image. Score Distillation Sampling (SDS) (Poole et al. 2022) and Delta Denoising Score (DDS) (Hertz, Aberman, and Cohen-Or 2023) utilize the pre-trained T2I diffusion model to control the optimization of the latent code and achieve text-conditioned editing. Inspired by this, the reference image in personalized editing can also be treated as a condition in optimization. The optimization is training-free and allows more flexibility in the edited area even for hard cases.

Our approach formulates personalized content editing as an energy-based optimization problem conditioned by a reference text-image pair. First, we use the reference text and image as queries and the diffusion model as a conditioned energy function. The corresponding feature-level differences between the edited and reference images within the target object are minimized during optimization. Second, to achieve higher identity consistency for the target object, we design a replacement-based content composition operation that integrates the target information into the latent variables during optimization. Third, to avoid blurred output and achieve coarse-to-fine optimization, we propose scheduling the timesteps for the diffusion model in descending order and using text guidance only in early iterations. This allows a natural transition of the object shape towards the target and achieves stable refining of appearance at the late stage. Additionally, we utilize truncation and smoothing techniques for the gradients to maintain the background information and ensure a harmonious integration of the target object.

In summary, the contributions of this paper are as follows:

- We first conceptualize the problem of personalized image editing as a conditioned optimization task using diffusion-based text-image energy guidance.
- We propose an energy-guided optimization framework (EGO-Edit) involving coarse-to-fine strategies along with latent space content composition for stability and higher consistency in optimization.
- Extensive experiments demonstrate the effectiveness of our method in personalized image editing, such as object swapping and inpainting, and can produce desired edited results in hard cases.

Related Work

Text-to-image Model. The field of text-to-image synthesis has been profoundly influenced by conditional diffusion models (Dhariwal and Nichol 2021; Ho and Salimans 2022; Nichol et al. 2021), which have ushered in a new era of image generation. These models have demonstrated the ability to produce high-quality images that are conditioned on textual descriptions, significantly advancing the capabilities of generative systems. However, the sensitivity of these models to the quality of the text prompt has become a recognized limitation, often necessitating careful and deliberate prompt design (Hao et al. 2024) to achieve satisfactory results (Witteveen and Andrews 2022). Recent works have begun to explore the integration of image prompts to guide the generation process. DALL-E 2 (Ramesh et al. 2022) marked a pivotal step with its pioneering approach to image-guided image generation. Subsequent research, such as ELITE (Wei et al. 2023), BilpDiffusion (Li, Li, and Hoi 2024), ProFusion(Zhou et al. 2023), Domain-Agnostic(Arar et al. 2023) and IP-Adapter (Ye et al. 2023), has focused on learning from image prompts to enable object customization, offering a new dimension in the control over generative models

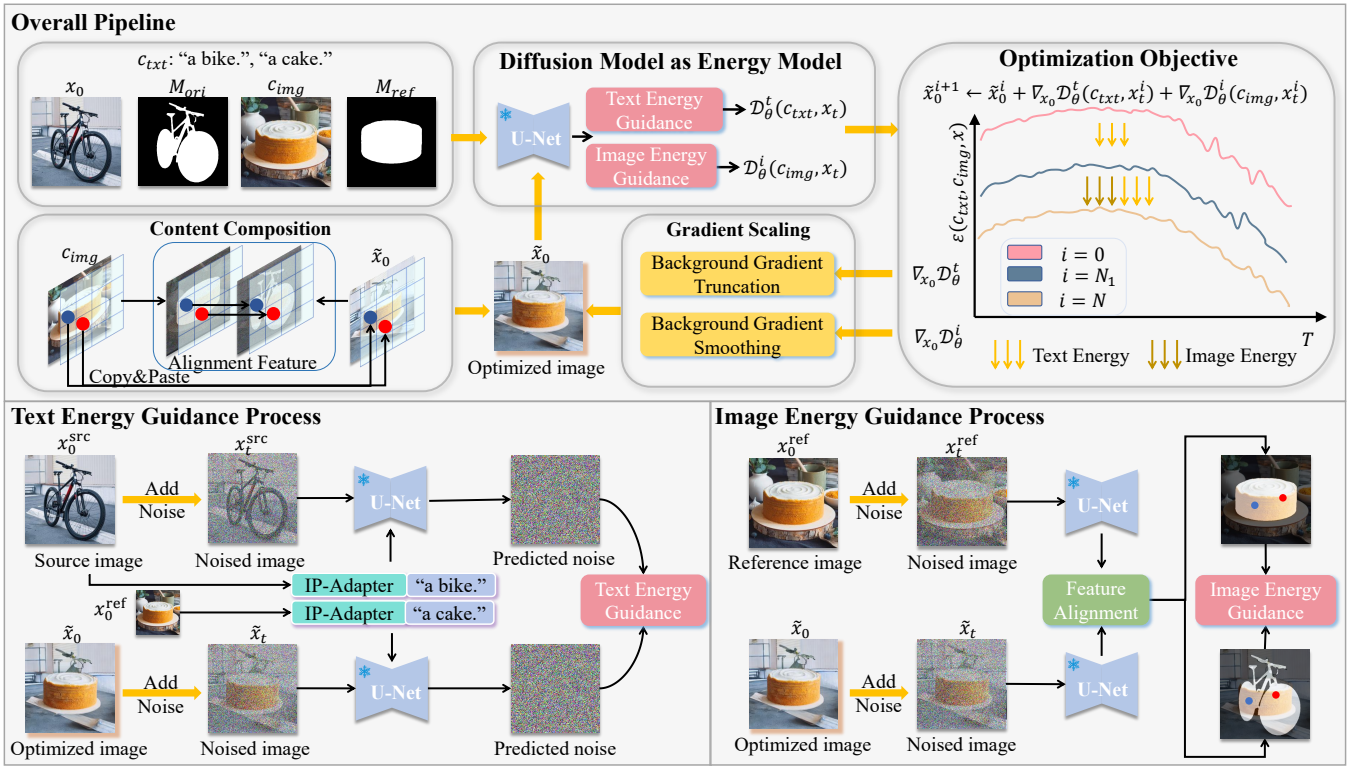

Figure 3: Pipeline Overview of the Proposed Method: The illustration above outlines the pipeline for our energy-guided optimization method. We construct the energy function derived from the diffusion model, aiming to minimize the energy of the edited image, $\tilde{\mathbf{x}}_t$, to progressively align its distribution with that of the reference image. The diffusion-based energy function is composed of two key components: Text Energy Guidance (TEG) and Image Energy Guidance (IEG). TEG is applied throughout the entire process, ensuring consistent semantic alignment, while IEG is specifically employed during the N2 optimization step to refine visual details, enhancing the fidelity of the edited image to the reference. The processes for both TEG and IEG are detailed below the main pipeline.

(Jiang et al. 2024). Despite the recent strides in conditional image generation, leveraging the combined power of text and image prompts to enhance the versatility and precision of diffusion models in image editing tasks continues to be an open challenge.

Image Editing. The image editing methodologies can be categorized into three directions: training-based approaches (Kim, Kwon, and Ye 2022; Wang et al. 2023; Yang et al. 2024; Li, Singh, and Grover 2023; Sheynin et al. 2023), test-time fine-tuning approaches (Choi et al. 2023; Mokady et al. 2023; Dong et al. 2023), and traing-free approaches (Hertz et al. 2022; Parmar et al. 2023; Tumanyan et al. 2023; Lu, Liu, and Kong 2023). Most of the previous editing (Ju et al. 2024; Huberman-Spiegelglas, Kulikov, and Michaeli 2024; Brooks, Holynski, and Efros 2023) focus on editing local image regions conditioned by text prompts.

Paint-by-Example (PbE) (Yang et al. 2023) proposed a training-based approach for exemplar-guided personalized image editing. Subsequent training-based methods Custom-Edit (Choi et al. 2023), ObjectStitch (Song et al. 2023), Unipaint (Yang, Chen, and Liao 2023), DreamInpainter (Xie et al. 2023), Photoswap (Gu et al. 2024) and Anydoor (Chen et al. 2024) also implemented personalized object replacement by inpainting. Blip-diffusion (Li, Li, and Hoi 2024) used a pre-trained multimodal encoder for efficient, zero-shot generation and rapid fine-tuning for customized editing. Recently, Dragon (Mou et al. 2023) and Diffeditor (Mou et al. 2024) designed the energy-guided sampling with pre-trained SD, which can achieve accurate personalized editing while the backbone remains training-free. However, generalized cross-category object replacement remains challenging for these methods and is worth further exploration.

Preliminary

Score Distillation Sampling (SDS). Given input image $\mathbf{x}_0$, text embedding y, noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, timestep $t \sim \mathcal{U}(0, 1)$, and noise estimater ϵ_ϕ parameterized by ϕ, the diffusion loss is defined by:

$$\mathcal{L}_{\text{diff}} = w(t) \left\| \epsilon_{\boldsymbol{\phi}}(\mathbf{x}_t, y, t) - \boldsymbol{\epsilon} \right\|_2^2,$$

where $w(t)$ is a weighting function, $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon$. Classifier-free Guidance (CFG) (Ho and Salimans 2022) achieves high-quality text-conditioned generation via guidance scale ω: $\epsilon_{\boldsymbol{\phi}}^{\omega}(\mathbf{x}_t, y, t) = \omega\epsilon_{\boldsymbol{\phi}}(\mathbf{x}_t, y, t) +$

$(1-\omega)\epsilon_{\boldsymbol{\phi}}(\mathbf{x}_t, t)$. SDS (Poole et al. 2022) and DDS (Hertz, Aberman, and Cohen-Or 2023) utilize the diffusion loss to optimize the image or model parameters θ towards the text condition y:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = (\epsilon^{\omega}_{\boldsymbol{\phi}}(\mathbf{x}_t, y, t) - \boldsymbol{\epsilon})\frac{\partial \mathbf{x}_t}{\partial \boldsymbol{\theta}},$$

$$\nabla_\theta \mathcal{L}_{\text{DDS}} = (\epsilon^{\omega}_{\boldsymbol{\phi}}(\mathbf{x}_t, y, t) - \epsilon^{\omega}_{\boldsymbol{\phi}}(\hat{\mathbf{x}}_t, \hat{y}, t))\frac{\partial \mathbf{x}_t}{\partial \boldsymbol{\theta}}.$$

Energy-based Guidance. Song et al. (Song et al. 2020) introduces conditional generation by decomposing score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{c})$ into:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{c}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}|\mathbf{x}_t),$$

The challenge lies in modeling the correction gradient $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}|\mathbf{x}_t)$. A solution involves using energy function:

$$p(\mathbf{c}|\mathbf{x}_t) = \frac{\exp\{-\lambda \mathcal{E}(\mathbf{c}, \mathbf{x}_t)\}}{Z},$$

where λ is a positive temperature coefficient, Z is a normalizing factor, and $\mathcal{E}(\mathbf{c}, \mathbf{x}_t)$ is an energy function measuring compatibility between condition $\mathbf{c}$ and noised image $\mathbf{x}_t$. Lower energy values indicate higher compatibility. This formulation approximates the correction gradient as

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{c}|\mathbf{x}_t) \propto -\nabla_{\mathbf{x}_t} \mathcal{E}(\mathbf{c}, \mathbf{x}_t).$$

The SDS gradient can be interpreted within this energy-based framework. It can be decomposed into two components: $\epsilon_{\boldsymbol{\phi}}(\mathbf{x}_t, t)$ corresponding to $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$, and $\omega(\epsilon_{\boldsymbol{\phi}}(\mathbf{x}_t, y, t) - \epsilon_{\boldsymbol{\phi}}(\mathbf{x}_t, t))$ associated with $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}|\mathbf{x}_t)$.

Method

Energy-Guided Optimization for Image Editing

Classifier-based methods (Dhariwal and Nichol 2021; Zhao et al. 2022; Liu et al. 2023) use time-dependent distance measuring functions to approximate energy functions:

$$\mathcal{E}(c, \mathbf{x}_t) \approx \mathcal{D}_\phi(c, \mathbf{x}_t, t),$$

where ϕ denotes the pre-trained parameters. $\mathcal{D}_\phi(c, \mathbf{x}_t, t)$ denotes the distance between condition c and noised latent $\mathbf{x}_t$. While off-the-shelf pre-trained networks (e.g., classification, segmentation) can be used, they often rely on one-step denoising approximations (Yu et al. 2023), leading to inaccuracies, especially in the early stage of generation. We propose using the pre-trained diffusion model itself as energy function to leverage its comprehensive image prior knowledge:

$$\mathcal{E}(\mathbf{c}, \tilde{\mathbf{x}}_0)) \approx \mathbb{E}_{p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_0))}[\mathcal{D}_\theta(\mathbf{c}, \tilde{\mathbf{x}}_t)],$$

measuring the distance between the target image $\tilde{\mathbf{x}}_0$ and the condition $\mathbf{c}$ at various noise levels t, based on the intuition that the distance between a noised image $\tilde{\mathbf{x}}_t$ and the condition $\mathbf{c}$ reflects the distance between its corresponding target image $\tilde{\mathbf{x}}_0$ and the same condition $\mathbf{c}$.

Our training-free approach (see Fig. 3) regards image editing as a conditional optimization problem. Instead of relying on iterative sampling, we directly update the target image $\tilde{\mathbf{x}}_0$ under the guidance from both text and image energy, which can be formulated as:

$$\mathcal{E}(\mathbf{c}_{txt}, \mathbf{c}_{img}, \tilde{\mathbf{x}}_0) \approx \eta_1 \mathbb{E}_{p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_0)}[\mathcal{D}^t_\theta(\mathbf{c}_{txt}, \tilde{\mathbf{x}}_t)] + \eta_2 \mathbb{E}_{p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_0)}[\mathcal{D}^i_\theta(\mathbf{c}_{img}, \tilde{\mathbf{x}}_t))],$$

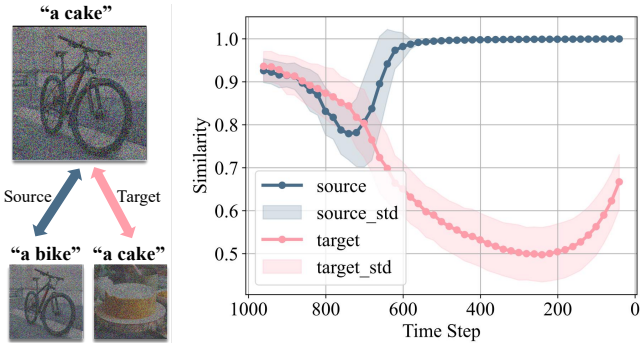

Figure 4: Visualization of the feature similarity. Given two text-image pairs, referred as source and target. We query the source image with the target text and compute the feature similarity with the source and target under noise addition to different times t. We analyze 140 images from different categories and calculate the mean and standard deviation of the feature similarity.

where η_1 and η_2 are the weight coefficients. The text energy provides high-level conceptual guidance particularly in the early stage of optimization, ensuring that the editing result aligns with the text prompt. While the image energy provides fine-grained control in the later stage by enhancing appearance details.

Text Energy. Preliminarily, distillation loss proposed by SDS and DDS can be regarded as text energy. We further extend the formulation to incorporate with negative prompts:

$$\mathcal{D}^t_\theta(\mathbf{c}_{txt}, \tilde{\mathbf{x}}_t)) = ||\epsilon^{\omega}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_t, y_{ref}, t) - \epsilon^{\omega}_{\boldsymbol{\phi}}(\mathbf{x}^{src}_t, y_{src}, t)||^2_2 + ||\epsilon_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_t, t) - \epsilon_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_t, y_{neg}, t)||^2_2.$$

where y_{src}, y_{ref} and y_{neg} represents two positive prompt for source and target objects respectively, along with an extra negative prompt. By leveraging these prompts, we achieve precise control over object attributes. However, it brings a trade-off between quality and fidelity, as negative prompts may cause deviations from the target prompt, leading to visible artifacts. Thus we assign a smaller weight to the energy gradient from negative prompt.

Given that text prompts are usually ambiguous and lack the detailed description necessary for precise guidance in personalized image editing, they may conflict with the details recovered through image energy. The IP-Adapter (Ye et al. 2023) is compatible with the text energy function and can further enhance the gradient precision for editing with image tokens as prompts. However, the image background can be inadvertently altered after introducing the IP-Adpater since background information is also captured by image tokens. We employ Gradient Scaling (GS) to adjust the text energy gradient. Specifically, for the text energy function, we use Background Gradient Truncation to limit the energy gradient outside the mask M_{ori}: $\nabla_{x_0} \mathcal{D}^t_\theta \odot M_{ori}$.

Image Energy. Image energy focuses primarily on the transfer of reference features. At each timestep t, the UNet denoiser ϵ_ϕ is utilized to extract intermediate features $\mathcal{F}^{opt}_t$

from the edited image $\tilde{\mathbf{x}}_t$, so do the reference features $\mathcal{F}_t^{ref}$ from the reference image $\mathbf{x}_t^{ref}$.

Following the methodology outlined in DIFT (Tang et al. 2023), $\mathcal{F}_t^{opt}$ and $\mathcal{F}_t^{ref}$ preserve the high-level semantic consistency required for precise point-to-point correspondence. To ensure coherent edits across the image, we employ two binary masks to constrain the region for editing. Within the mask, we focus on transferring the features by identifying the points in $\mathcal{F}_t^{ref}$ that have the shortest distance to the corresponding points in $\mathcal{F}_t^{opt}$.

Calculating the shortest distance ensures that the transferred features from the reference image maintain their semantic alignment in the optimized image, resulting in precise and contextually consistent edits. The distance metric $d(\cdot,\cdot)$ can be but not limited to Euclidean distance:

$$p_o = \underset{p \in M_{ori}}{\operatorname{argmin}}\, d(\mathcal{F}_t^{ref}[p_r], \mathcal{F}_t^{opt}[p]). \tag{1}$$

Here, $\mathcal{F}_t^{opt}[p]$ and $\mathcal{F}_t^{ref}[p_r]$ represent the feature vectors at the point $p \in \mathcal{M}_{ori}$ and $p_r \in \mathcal{M}_{ref}$. After finding the matching point p_o for each p_r. The image energy function is thus defined to minimize the point-to-point feature distance inside the mask:

$$\mathcal{D}_\theta^i(\mathbf{c}_{img}, \tilde{\mathbf{x}}_t) = \sum_{p_r \in M_{ref}} ||\mathcal{F}^{opt}[p_o] - \mathcal{F}^{ref}[p_r]||_2^2.$$

Unlike text energy which provides the coarse semantic guidance, we aim for image energy to provide fine-grained transition control upon the masked regions. If the gradient update is restricted within the region, it may reduce editing flexibility and result in visible artifacts along the object boundary. Therefore, another Gradient Scaling operation is introduced, we apply Background Gradient Smoothing to ensure a more natural transition at the boundary: $\nabla_{x_0}\mathcal{D}_\theta^i \odot M_{ori} + (\nabla_{x_0}\mathcal{D}_\theta^i * k_s) \odot (1 - M_{ori})$, where k_s is the Gaussian smoothing kernel.

Latent Space Content Composition

Although conditional energy function optimization can effectively integrate the target object into the source image, there still exists inconsistency in appearance due to the information loss and large domain gaps in certain scenarios. It demands a more efficient feature transferring method to achieve higher consistency.

Notice that the feature matching results in Eq. 1 and the structural information become stable after some iterations, we design a Content Composition operation accordingly. This operation directly manipulates latent variables in a copy-paste manner, transferring features between the matching point pairs from the reference image to $\tilde{x}_0$ within the mask. Given a point pair (p_o, p_r) in Eq. 1, we have:

$$\tilde{x}_0[p_o] \leftarrow x_0^{ref}[p_r].$$

By directly transferring in the latent space, local appearance information from the target is transferred to the edited results. Moreover, unlike the one-step copy-paste in the pixel space, we apply the operation at an interval during optimization. This allows the subsequent optimization with energy guidance to repair the integrity and semantics of the result.

Coarse-to-fine Optimization

In the original SDS and DDS methods, optimization is performed by randomly sampling timesteps per iteration. However, this strategy can lead to blurry edited results when optimizing for the energy function $\mathcal{E}$ due to instability in feature matching (Tang et al. 2023). Furthermore, the high similarity between the edited and source images in early optimization iterations makes the feature matching inaccurate, causing misleading image energy guidance D_θ^i.

To address these limitations, we introduce a coarse-to-fine scheduling strategy, which narrows the timestep t to a predefined sequence $\{t_1, t_2, t_3, \ldots, t_N\}$, satisfying $t_1 > t_2 > t_3 > \cdots > t_N$. Larger t captures coarse-grained concepts, while smaller t focuses on fine-grained details, resulting in more effective edits.

We emphasize text energy during early optimization stages with large noise for narrowing the semantic gaps. As optimization progresses, image energy becomes increasingly relevant to the editing result. Fig. 4 illustrates the variation in feature similarity between edited and source images with respect to timestep t. At large timestep $t > 700$, the edited images resemble target images, indicating the dominant influence of text energy. After timestep $t = 600$, the similarity between edited and source images surpasses that with the target image, indicating a shift towards image feature dominance. Therefore, we opt to image energy gradients to guide detail restoration during the middle and later stages of optimization.

This coarse-to-fine strategy effectively employs text and image energy functions: large time steps (early stage) for establishing structure via text energy, medium time steps (middle stage) for refining features with increasing image energy influence, and small time steps (later stage) for fine details jointly guided by text-image energy.

Experiments

Settings

Benchmarks. We evaluated our method using established benchmarks. For object swapping tasks, we utilized DreamEditBench (Li et al. 2023), which features 22 themes aligned with the DreamBooth framework (Ruiz et al. 2023). We performed two-by-two exchanges using 10 images from the same theme but different environments within DreamEditBench. Additionally, we selected 50 images from PIE-Bench (Ju et al. 2024), representing distinct conceptual categories, and paired them randomly for object swapping.

Implementation Details. We implemented our proposed method using the Stable Diffusion 1.5 model as the pretrained text-to-image diffusion model. All experiments were conducted on images with a resolution of 512 x 512 pixels, striking a balance between image quality and computational efficiency. The number of optimization steps is set to 50 for all experiments. More details can be found in Appendix A.

Evaluation Metrics. To comprehensively assess the performance of our proposed method, we employed a diverse set of evaluation metrics: CLIP Score (text and image) (Radford et al. 2021) and DINO Score (Caron et al. 2021). These

Figure 5: Qualitive results of cross-class replacement. The source object and the target object are sampled from different classes.

Method	DINO (↑)	CLIP-T (↑)	CLIP-I (↑)
PbE	47.666	27.149	71.322
Anydoor	57.093	26.410	72.133
PAIR	48.786	25.901	68.575
Dragon	48.176	26.361	69.949
EGO-Edit(Ours)	**62.749**	**28.764**	**76.624**

Table 1: Quantitative comparison results. DINO and CLIP scores are reported to evaluate the quality of replacement.

Figure 6: Ablation on the components. The baseline uses the text guidance and the IP-adapter. "IEG": image-energy guidance. "CC": content composition. "GS": gradient scaling.

metrics effectively reflect the similarity between the generated region and the target object or the matching degree with the text. The user study is presented in Appendix B.

Comparison with Previous Methods

We present the quantitative results of our method and the competing baselines on the customized image swapping tasks in Table 1. Our approach demonstrates superior performance in both DINO and CLIP metrics, indicating higher fidelity in the edited images compared to existing methods.

We provide a qualitative comparison of cross-class replacement between the proposed method and competing approaches in Fig. 5. When there is a significant size difference between the source object and the target object (see second row), inpainting-based methods (PbE (Yang et al. 2023), Graphit (Gu et al. 2023), Anydoor (Chen et al. 2024)) often introduce irrelevant details within the mask area. Appearance editing-based methods (PAIR (Goel et al. 2023), Dragon (Mou et al. 2023)) tend to fill the entire mask area with the reference object. In contrast, our method follows the size of the reference image and can make a smooth transition from the source image. In Fig. 7, we further visualize the results of in-class replacement. Our method can transfer the reference image features while maintaining the source image pose. More applications are shown in Appendix C.

Ablation Study

Components. To understand the contributions of different components in our method, we conducted an ablation study, with the results presented in Table 2. For a clearer comparison, Fig. 6 illustrates the effects of each component. The last column presents results from our method, while the preceding three columns demonstrate the impact of systematically removing key components. Without IEG, the generated images lost distinctive identity features, maintaining only coarse semantic consistency. Excluding Content Composition (CC) led to significant issues: while image energy guidance transferred relevant features, it often missed key details and inadequately preserved color information, causing noticeable degradation of fine details like the duck's head color (row 1) and the bag's logo (row 2). Removing Gradient Scaling (GS) results in the influence of semantic guidance on the entire image, which adversely impacts editing quality.

Timestep Selection for Content Composition. We investigated the key factors for the insertion of the content composition. In Fig. 8, we compare edited results with different timesteps t to insert CC. We concluded that performing content composition early in the optimization process (upper left) results in generated images that exhibit high consistency with the reference image in terms of pose and relative position. However, early content composition can lead

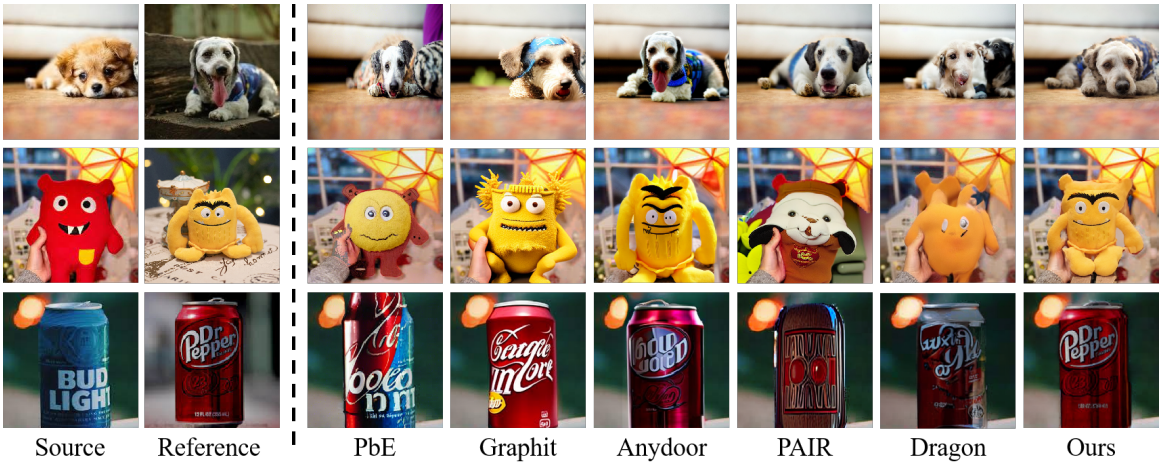

Figure 7: Qualitative results of in-class replacement. The target object in the reference image belongs to the source object class.

Method	DINO (↑)	CLIP-T (↑)	CLIP-I (↑)
w/o IEG	56.510	28.660	75.664
w/o CC	57.531	28.753	75.457
EGO-Edit	**62.749**	**28.764**	**76.624**

Table 2: Quantitative ablation results of components. We compare the results without image-energy guidance (IEG) or content composition (CC).

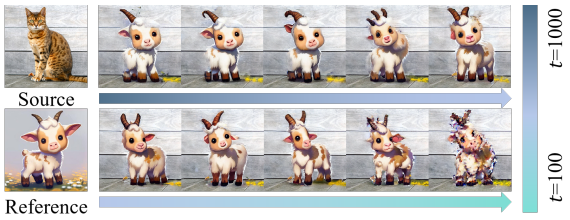

Figure 8: Ablation on the insertion condition t for CC. We start CC when the timestep is t and stop it with the timestep $t-100$ in optimization.

to excessive adherence to the reference, potentially sacrificing flexibility in the final output. Conversely, conducting content composition at the late stage of optimization (lower right) fails to adequately correct artifacts caused by discrete matching, leading to image blurriness and a lack of coherent details. It appears that late iterations (smaller t) do not provide sufficient opportunity to refine the transferred features, resulting in lower quality in the generated image. Optimal results are achieved when the content composition is applied at the medium stages of the optimization process. This choice balances the trade-off between adherence to the reference image and flexibility in the final output.

Optimization Strategy. We compare three optimization schedules of timesteps, namely random, ascending, and descending in Fig. 9. The visual comparison reveals that the random schedule suffer from noticeable blurriness. This is

Figure 9: Ablation on the optimization schedule of timesteps. We compare three schedules of time steps during optimization: random, ascending order, and descending order. The descending schedule in our method performs best on the consistency of both details and semantics.

due to the inherent instability of the random strategy in feature matching, resulting in inconsistent feature point alignment across optimization iterations. The ascending schedule optimizes the fine-grained appearance first (start from small timestep t). However, the semantic gap at the early stage causes a shape mismatch with the target object. The mismatch makes the edited result exhibit obvious shape characteristics of the source object (cat). We employ the descending schedule in the coarse-to-fine optimization that aligns the coarse shape information with the target first. This allows more accurate and stable matching for refining the appearance details and achieves better quality.

Conclusion

This paper introduces a paradigm shift in personalized image editing through the innovative application of pretrained diffusion models. We address the limitations of existing methods by proposing a training-free and inversion-free approach that harnesses the conditional optimization of latent codes, guided by reference images and text. Our method minimizes feature discrepancies between edited and reference images, ensuring a seamless integration of personalized content. Extensive experiments have shown that our method has achieved good results in both in-class and cross-class customized object replacement.

Acknowledgements

This work is supported in part by National Science Foundation for Distinguished Young Scholars under Grant 62225605, Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, Natural Science Foundation of Shanghai under Grant 24ZR1425600, Project 12326608 supported by NSFC, "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01020), National Natural Science Foundation of China under Grant No.62441602, the Ningbo Science and Technology Innovation Project (No.2024Z294), and the Fundamental Research Funds for the Central Universities.

References

Arar, M.; Gal, R.; Atzmon, Y.; Chechik, G.; Cohen-Or, D.; Shamir, A.; and H. Bermano, A. 2023. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*, 1–10.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

Choi, J.; Choi, Y.; Kim, Y.; Kim, J.; and Yoon, S. 2023. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dong, W.; Xue, S.; Duan, X.; and Han, S. 2023. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7430–7440.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36: 16222–16239.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goel, V.; Peruzzo, E.; Jiang, Y.; Xu, D.; Sebe, N.; Darrell, T.; Wang, Z.; and Shi, H. 2023. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*.

Gu, G.; Chun, S.; Kim, W.; Jun, H.; Yun, S.; and Kang, Y. 2023. Graphit: A Unified Framework for Diverse Image Editing Tasks. https://github.com/navervision/Graphit.

Gu, J.; Wang, Y.; Zhao, N.; Fu, T.-J.; Xiong, W.; Liu, Q.; Zhang, Z.; Zhang, H.; Zhang, J.; Jung, H.; et al. 2024. Photoswap: Personalized subject swapping in images. *Advances in Neural Information Processing Systems*, 36.

Hao, Y.; Chi, Z.; Dong, L.; and Wei, F. 2024. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Hertz, A.; Aberman, K.; and Cohen-Or, D. 2023. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2328–2337.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huberman-Spiegelglas, I.; Kulikov, V.; and Michaeli, T. 2024. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12469–12478.

Jiang, R.; Zheng, G.-C.; Li, T.; Yang, T.-R.; Wang, J.-D.; and Li, X. 2024. A Survey of Multimodal Controllable Diffusion Models. *Journal of Computer Science and Technology*, 39(3): 509–541.

Ju, X.; Zeng, A.; Bian, Y.; Liu, S.; and Xu, Q. 2024. PnP Inversion: Boosting Diffusion-based Editing with 3 Lines of Code. *International Conference on Learning Representations (ICLR)*.

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

Li, D.; Li, J.; and Hoi, S. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

Li, S.; Singh, H.; and Grover, A. 2023. InstructAny2Pix: Flexible Visual Editing via Multimodal Instruction Following. *arXiv preprint arXiv:2312.06738*.

Li, T.; Ku, M.; Wei, C.; and Chen, W. 2023. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*.

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

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2305.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2023. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2024. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8488–8497.

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

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695. IEEE.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sheynin, S.; Polyak, A.; Singer, U.; Kirstain, Y.; Zohar, A.; Ashual, O.; Parikh, D.; and Taigman, Y. 2023. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*.

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

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; and Aliaga, D. 2023. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18310–18319.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wang, S.; Saharia, C.; Montgomery, C.; Pont-Tuset, J.; Noy, S.; Pellegrini, S.; Onoe, Y.; Laszlo, S.; Fleet, D. J.; Soricut, R.; et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18359–18369.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Witteveen, S.; and Andrews, M. 2022. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*.

Xie, S.; Zhao, Y.; Xiao, Z.; Chan, K. C.; Li, Y.; Xu, Y.; Zhang, K.; and Hou, T. 2023. DreamInpainter: Text-Guided Subject-Driven Image Inpainting with Diffusion Models. *arXiv preprint arXiv:2312.03771*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yang, S.; Chen, X.; and Liao, J. 2023. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3190–3199.

Yang, Y.; Peng, H.; Shen, Y.; Yang, Y.; Hu, H.; Qiu, L.; Koike, H.; et al. 2024. ImageBrush: Learning Visual In-Context Instructions for Exemplar-Based Image Manipulation. *Advances in Neural Information Processing Systems*, 36.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, J.; Wang, Y.; Zhao, C.; Ghanem, B.; and Zhang, J. 2023. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23174–23184.

Zhao, M.; Bao, F.; Li, C.; and Zhu, J. 2022. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35: 3609–3623.

Zhou, Y.; Zhang, R.; Sun, T.; and Xu, J. 2023. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*.