VersaFusion: A Versatile Diffusion-Based Framework for Fine-Grained Image Editing and Enhancement

Haocun Ye[1,2,4*], **Xinlong Jiang**[1,2,4,5*], **Chenlong Gao**[1,2,4], **Bingyu Wang**[1,2,4], **Wuliang Huang**[1,2,4,5], **Yiqiang Chen**[1,2,3,4,5]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Peng Cheng Laboratory
[4]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing, China
[5]AIER Eye Hospital Group Co., Ltd., Changsha 410015, China
{yehaocun23s, jiangxinlong, gaochenlong, wangbingyu, huangwuliang19b, yqchen}@ict.ac.cn

Abstract

Text-to-image (T2I) diffusion models have achieved remarkable progress in generating realistic images from textual descriptions. However, ensuring consistent high-quality image generation with complete backgrounds, object appearance, and optimal texture rendering remains challenging. This paper presents a novel fine-grained pixel-level image editing method based on pre-trained diffusion models. The proposed dual-branch architecture, consisting of Guidance and Generation branches, employs U-Net Denoisers and Self-Attention mechanisms. An improved DDIM-like inversion method obtains the latent representation, followed by multiple denoising steps. Cross-branch interactions, such as KV Replacement, Classifier Guidance, and Feature Correspondence, enable precise control while preserving image fidelity. The iterative refinement and reconstruction process facilitates fine-grained editing control, supporting attribute modification, image outpainting, style transfer, and face synthesis with Click-and-Drag style editing using masks. Experimental results demonstrate the effectiveness of the proposed approach in enhancing the quality and controllability of T2I-generated images, surpassing existing methods while maintaining attractive computational complexity for practical real-world applications.

Introduction

Generative models(Isola et al. 2017; Goodfellow et al. 2020; Hertz et al. 2022), particularly diffusion models for text-to-image (T2I) synthesis, have transformed the landscape of generative image creation on a grand scale. However, achieving fine-grained control over the output remains challenging, especially in image-editing tasks. Traditional text-guided approaches(Dong et al. 2017; Li et al. 2020; Nam, Kim, and Kim 2018; Reed et al. 2016; Xu et al. 2018) struggle to establish accurate correspondences between text and specific objects within the image, limiting their effectiveness in complex scenarios. Point-based manipulation techniques(Hinz et al. 2021; Ling et al. 2021; Shaham, Dekel, and Michaeli 2019; Xu et al. 2022), such as DragGAN(Pan et al. 2023), have shown potential for more intuitive and precise editing, but are constrained by the underlying generative models. DragDiffusion(Shi et al. 2023) and DragonDiffusion(Mou et al. 2023), built upon pre-trained T2I diffusion models(Ramesh et al. 2021), leverage diverse generation capabilities for detailed manipulations on general images, but lack flexibility in certain editing tasks.

To improve the adaptability and enhance the output quality of image editing using diffusion models, we suggest integrating image prompts and solvers for stochastic differential equations (SDEs)(Song et al. 2020). Our approach enables diverse and adaptable modifications while preserving the coherence and realism of the edited image. In contrast to parallel research endeavors(Karras et al. 2022), our proposed technique does not necessitate supplementary fine-tuning of the model architecture or the incorporation of novel structural components.

Based on the strong correspondence between the image features in T2I diffusion models (Avrahami, Lischinski, and Fried 2022; Kim, Kwon, and Ye 2022; Ramesh et al. 2022a), we present an innovative approach to image editing that utilizes proximal-negative inversion (PNI) to transform the input image into its corresponding latent space representation. During the denoising process, the model incorporates task-specific gradient guidance at every step to facilitate precise image editing, reconstructing the image to match the user's intent. By formulating image editing as a modification of feature correspondences and translating it into gradient guidance via energy functions in score-based diffusion (Song et al. 2020; Karras et al. 2022), we enable precise and flexible manipulation without relying on text-to-image correspondence.

In summary, our main contributions are:

- A novel dual-branch architecture for diffusion-based image editing, consisting of a Guidance Branch and a Generation Branch, which maintains fidelity to the original image while enabling targeted edits.
- Cross-branch interactions through K V Replacement, Classifier Guidance, and Feature Correspondence, allowing for precise control over the editing process.
- Extension of image editing capabilities, enabling various functions such as outpainting, resizing, style transfer.

*Corresponding author: Xinlong Jiang.

Figure 1: VersaFusion, a tuning-free diffusion-based framework, enables fine-grained image editing operations such as canvas extension, drag-and-gen editing, object relocation, resizing, style transfer, face synthesis, and seamless object integration using reference images.

- Demonstration of state-of-the-art performance and computational efficiency across a wide range of fine-grained image editing tasks.

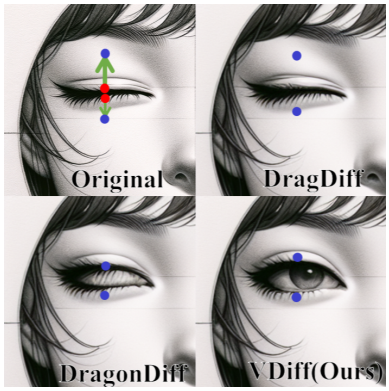

Figure 2: DragDiff and DragonDiff face limitations in constrained image editing, such as transforming closed eyes to open or generating novel content. Our approach demonstrates superior adaptability and advanced editing capabilities compared to existing methods.

Related Work

Diffusion Models

As highly effective generative models, diffusion models master the inversion of a forward process in which data undergoes gradual transformation into noise through iterative steps (Ho, Jain, and Abbeel 2020). A neural network defines the parameters of the reverse process, which predicts the noise term $\epsilon_\theta(x_t, t, y)$ based on the noisy input x_t, the corresponding time step t, and an optional conditioning variable y, such as an embedded text prompt (Nichol et al. 2022). The estimated noise term is subsequently used to incrementally de-noise the input data point using different update schemes, including DDPM (Ho, Jain, and Abbeel 2020) or DDIM (Song, Meng, and Ermon 2022). During the diffusion process, Gaussian noise is introduced incrementally to an image $\mathbf{x}_0$ according to $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$, where α_t decreases linearly from 1 to a small value, ensuring that $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$(Sohl-Dickstein et al. 2015). The reverse process involves training a denoiser to iteratively reconstruct $\mathbf{x}_0$ from $\mathbf{x}_T$, conditioned on the noisy image $\mathbf{x}_t$ and the time step t:

$$\mathbb{E}_{\mathbf{x}_0,t,\epsilon_t\sim\mathcal{N}(0,1)}\left[\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right] \tag{1}$$

where ϵ_θ represents the denoiser function. DDIM formulates the diffusion sampling as $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\delta_t^2}\cdot\frac{\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}}, \alpha_t^2\mathbf{I}\right)$, a non-Markovian process that can be expressed as:

$$\mathbf{x}_{t-1} = \underbrace{\sqrt{\alpha_{t-1}}\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}}}_{\text{“predicted } \mathbf{x}_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}}\sqrt{1-\frac{\alpha_t}{\alpha_{t-1}}}\epsilon_\theta(\mathbf{x}_t)}_{\text{“direction pointing to } \mathbf{x}_t\text{”}} + \underbrace{\sigma_t\epsilon_t}_{\text{“noise”}} \tag{2}$$

The parameter σ_t is defined as $\eta\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$. When η is set to 1 for all values of t, the process is equivalent to DDPM, which can be described as a stochastic differential equation (SDE). Conversely, setting η to 0 results in a deterministic sampling process, akin to an ordinary differential equation (ODE). While most diffusion-based image editing techniques rely on ODE for improved content consistency, investigating the potential of SDE in this domain is an area that merits further exploration(Meng et al. 2021).

Image Editing

Image editing aims to precisely alter an image's content. Conventional approaches, such as DragGAN(Pan et al. 2023), focus on inverting images into the latent space of

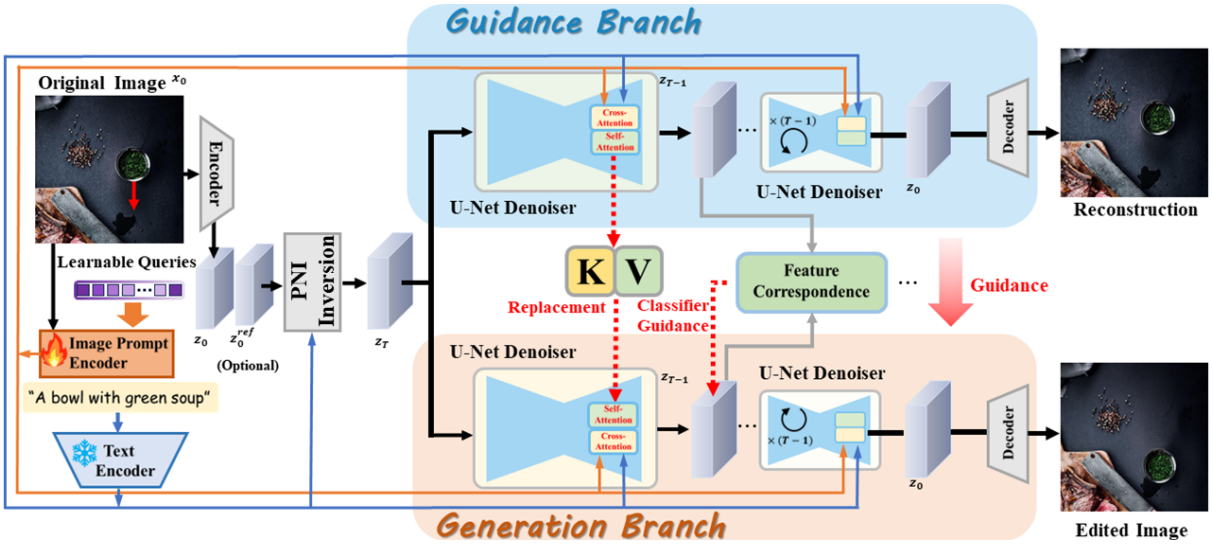

Figure 3: Overview of the dual-branch VersaFusion framework. The framework consists of two primary branches: the Guidance Branch (top) and the Generation Branch (bottom). The Guidance Branch focuses on reconstructing the input image with high fidelity, ensuring the preservation of original features through iterative denoising with a U-Net and a self-attention mechanism. In contrast, the Generation Branch introduces controlled modifications to the image, utilizing a similar U-Net architecture but augmented with classifier guidance and feature replacement strategies.

Generative Adversarial Networks and manipulating latent vectors to achieve desired modifications(Abdal, Qin, and Wonka 2019, 2020; Alaluf et al. 2022). However, these GAN-based methods are limited by the generalization abilities of the foundational models and the resulting image quality. Text-to-image diffusion models have introduced various text-guided image editing techniques(Avrahami, Lischinski, and Fried 2022), employing strategies like corrupting and denoising images based on target descriptions, using cross-attention maps for manipulation, or leveraging text as instructions(Hertz et al. 2022). While promising, the limited correspondence between text and image in these models hinders their effectiveness in fine-grained editing. Recently, DragDiff(Shi et al. 2023) and DragonDiff(Mou et al. 2023) have explored fine-grained image editing by capitalizing on the feature correspondence in pretrained StableDiffusion (SD) models(Rombach et al. 2022). DragDiff uses low-rank adaptation (LoRA)(Ryu 2023) to maintain content consistency while optimizing latent representations, while DragonDiff employs score-based gradient guidance(Song et al. 2020) and a visual cross-attention mechanism(Dhariwal and Nichol 2021) for drag-style editing without model fine-tuning. These advancements have opened up new possibilities for precise and intuitive image editing, enabling granular manipulation while preserving overall quality.

Guidance

Guidance is a powerful technique utilized in diffusion models, enabling the denoising process to be influenced towards a desired outcome(Dhariwal and Nichol 2021; Nichol et al. 2022; Meng et al. 2021). This influence can stem from multiple sources, including the diffusion model itself through classifier-free guidance(Ho and Salimans 2022), a separate classifier such as the ImageNet classifier guidance introduced by Dhariwal and Nichol, or more broadly, the gradients of an energy function(Nichol et al. 2022). Researchers have explored a wide array of guide functions, ranging from CLIP(Ramesh et al. 2022b) embedding distance and LPIPS similarity to bilateral filters, internal representations of diffusion models, and "readout heads"(Avrahami, Lischinski, and Fried 2022). Ho et al. put forth the notion of guiding the denoising process based on a single-step approximation of the clean data, termed reconstruction guidance(Geng and Owens 2024). Building upon this concept, Bansal et al. demonstrate the effectiveness of integrating guidance from various pre-existing models, such as those designed for segmentation, detection, facial recognition, and style transfer(Bansal et al. 2023).

Method

In this section, we formally introduce and present our proposed VERSAFUSION approach. Our study is based on the DragonDiff(Mou et al. 2023) architecture.

Preliminaries: Inversion and Score-Based Editing Guidance

DDIM inversion, a technique that maps an image back to its corresponding latent representation in the diffusion model's latent space, serves as a crucial component in our proposed VersaFusion approach. The standard DDIM inversion process can be formulated as follows:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t) \tag{3}$$

$\mathbf{z}_t$ and $\mathbf{z}_0$ denote latent vectors at timestep t and initially, $\bar{\alpha}_t$ is a timestep coefficient, ϵ_θ is the learned denoising model, and $\mathbf{x}_t$ is the noisy image at t. **Score-based editing guidance** is a technique that uses the gradient of the score func-

tion(Ho and Salimans 2022) to guide the image editing process in diffusion models. The score function, denoted as $\nabla_{\mathbf{x}} \log p(\mathbf{x}_t)$, represents the gradient of logarithmic probability density with respect to the input image $\mathbf{x}$ at the time step t. The score-based editing guidance can be formulated as:

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \alpha_t \nabla_{\mathbf{x}} \log p(\mathbf{x}_t) + \beta_t \mathbf{z} \tag{4}$$

where $\mathbf{x}_{t-1}$ is the edited image at timestep $t-1$, $\mathbf{x}_t$ is the input image at timestep t, α_t and β_t are scaling coefficients, and $\mathbf{z}$ is a noise term. The gradient term $\nabla_{\mathbf{x}} \log p(\mathbf{x}_t)$ guides the editing process by providing information about the direction that maximizes the log-probability density of the image. By incorporating score-based editing guidance(Ho and Salimans 2022), the diffusion model can generate more coherent and visually pleasant edited images, as the gradient information helps to maintain consistency with the original image content and structure.

Dual-branch Network Architecture of VersaFusion

Fig.3 illustrates the overall pipeline of our image editing approach. The framework comprises two branches: the guidance branch and the generation branch. The primary objective of the guidance branch is to reconstruct the original image while injecting relevant information into the generation branch. Concurrently, the generation branch focuses on guiding the editing process of the original image based on the provided information while maintaining consistency with the core content of the source image. Initially, the image to be edited, $\mathbf{x}_0$, and the reference image, $\mathbf{x}_{ref}$ (if available), are mapped to their representations in the diffusion space through an enhanced diffusion inversion process. These representations serve as inputs to both branches. Then, we transform the latent variables z_t of both branches into the feature domain using a shared UNet denoiser at each diffusion step. Two masks, $\mathbf{m}_{gud}$ and $\mathbf{m}_{gen}$, specify the locations of the dragged content in the original and edited images, constraining the content of $\mathbf{m}_{gud}$ to appear within the $\mathbf{m}_{gen}$ region. The similarity between the two regions is measured using cosine distance and normalized:

$$\mathcal{S}(\mathbf{m}^{gen}, \mathbf{m}^{gud}) = \tfrac{1}{2}\big(\cos(\mathbf{F}_t^{gen}[\mathbf{m}^{gen}], \mathcal{S}g(\mathbf{F}_t^{gud}[\mathbf{m}^{gud}])) + 1\big) \tag{5}$$

where $\mathcal{S}(\mathbf{m}^{gen}, \mathbf{m}^{gud})$ denotes our similarity measure that quantifies the alignment between the regions specified by the masks $\mathbf{m}^{gen}$ and $\mathbf{m}^{gud}$. $\mathbf{F}_t^{gen}$ and $\mathbf{F}_t^{gud}$ refer to the feature representations of the generation and guidance branches. The total loss function balances these constraints to guide the editing process:

$$\mathcal{L} = \frac{w_e}{\alpha + \beta \cdot \mathcal{S}(\mathbf{m}^{gen}, \mathbf{m}^{gud})} + \frac{w_p}{\alpha + \beta \cdot \mathcal{S}(\mathbf{m}^{share}, \mathbf{m}^{share})} \tag{6}$$

where w_e and w_p are weights that control the importance of the edited and preserved regions. α and β are parameters that adjust the influence of the similarity measures. $\mathbf{m}^{share}$ represents the mask indicating the shared regions between the original and generated images. To effectively inject editing information into the generation branch, we treat the conditional diffusion process as a joint score function(Ho and Salimans 2022) based on score-based diffusion(Ho and Salimans 2022). The editing signal is transformed into gradients through the score function, leveraging the strong correspondence of features. These gradients are then used to update the latent variable z_t during the diffusion process. To further enhance the alignment between semantic and graphical elements, a multi-scale guided alignment design is introduced on top of this guidance strategy.

$$\begin{aligned}\nabla_{\mathbf{z}_t^{\text{gen}}} \log q(\mathbf{z}_t^{\text{gen}}, \mathbf{m}^{\text{gen}}, \mathbf{m}^{\text{share}}) = \nabla_{\mathbf{z}_t^{\text{gen}}} \log q(\mathbf{z}_t^{\text{gen}}) \\ + \nabla_{\mathbf{z}_t^{\text{gen}}} \log q(\mathbf{m}^{\text{gen}}, \mathbf{m}^{\text{share}} | \mathbf{z}_t^{\text{gen}})\end{aligned} \tag{7}$$

The cross-branch self-attention mechanism, illustrated in Fig. 3, maintains coherence between the modified outputs and the initial image. By substituting the Key and Value components from the guidance branch's self-attention module into the corresponding elements of the generation branch, the model effectively incorporates reference information at the feature level. Moreover, we utilize a pair of learnable encoders—one for images and another for text—to incorporate both semantic and visual features into the cross-attention module of the SD(Rombach et al. 2022). This integration ensures the alignment of these features across multiple scales. By harnessing the complementary nature of textual and visual information, our method produces outputs that exhibit semantic consistency and visual coherence.

Proximal Negative-Prompt Inversion (PNI)

Inspired by proximal guidance(Han et al. 2023), we introduce PNI to address this issue, PNI incorporates an additional loss term that encourages the classifier-free guidance (CFG)(Ho and Salimans 2022) noise $\tilde{\epsilon}$ to align with $\tilde{\epsilon}_{src}$. Moreover, PNI introduces a regularization term to restrict the magnitude of $(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src})$. This regularization is achieved through the application of a proximal function,

$$\text{prox}_{\lambda, L_p}(x) = \underset{z}{\text{argmin}} \frac{1}{2}\|z - x\|_2^2 + \lambda \|z\|_p \tag{8}$$

This approach promotes desirable characteristics during the editing phase. In the case where p equals 1 (which corresponds to L_1 regularization), the solver assumes the form of a soft-thresholding function,

$$[\text{prox}_{\lambda, L_1}(x)]_i = [S_\lambda(x)]_i = \begin{cases} x_i - \lambda & \text{if } x_i > \lambda \\ 0 & \text{if } -\lambda \le x_i \le \lambda \\ x_i + \lambda & \text{if } x_i < -\lambda \end{cases} \tag{9}$$

where the i-th element is represented by $[\cdot]_i$. When p is set to 0 (corresponding to L_0 regularization), the solver assumes the form of a hard-thresholding function,

$$[\text{prox}_{\lambda, L_0}(x)]_i = \begin{cases} x_i & \text{if } |x_i| > \sqrt{2\lambda} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Inversion Guidance. PNI incorporates an inversion guidance mechanism that involves executing a single gradient descent step on the present latent $\tilde{z}_{t-1}$. This process aims to achieve alignment between $\tilde{z}_{t-1}$ and the inversion latent z^*_{t-1}. The gradient descent step is exclusively applied

Algorithm 1: Proximal Negative-Prompt Inversion

Require: Let z_0 be the original source sample, C and C' the source and target conditions, ϵ_θ the denoising model, and $\text{prox}_\lambda(\cdot)$ the proximal function.

$\tilde{z}_T \leftarrow \text{DDIMInvert}(z_0, C, w = 1)$
$\tilde{z}_T \leftarrow \tilde{z}_T$
for $t = T$ to 1 **do**
  $\tilde{\epsilon}_{src} \leftarrow \epsilon_\theta(\tilde{z}_t, t, C)$
  $\tilde{\epsilon}_{tar} \leftarrow \epsilon_\theta(\tilde{z}_t, t, C')$
  $\tilde{\epsilon} \leftarrow \tilde{\epsilon}_{src} + w \cdot \text{prox}_\lambda(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src})$
  $M \leftarrow |\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src}| \leq \lambda$
  $\tilde{z}_0 \leftarrow \frac{1}{\sqrt{\alpha_t}}\tilde{z}_t - \sqrt{\frac{1}{\alpha_t} - 1}\tilde{\epsilon}$
  $\tilde{z}_{t-1} \leftarrow \sqrt{\alpha_{t-1}}\tilde{z}_0 + \sqrt{1 - \alpha_{t-1}}\tilde{\epsilon}$
  if inversion guidance **and** $t < T_{inv}$ **then**
    $\tilde{z}_{t-1} \leftarrow \tilde{z}_{t-1} - \eta M \odot (\tilde{z}_{t-1} - z^*_{t-1})$
  end if
end for
return $\tilde{z}_0$

to the "unedited" area, which is determined by the mask $M = |\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src}| \leq \lambda$, with λ being reused to denote the threshold value. The update can be formulated as $\tilde{z}_{t-1} \leftarrow \tilde{z}_{t-1} - \eta M \odot (\tilde{z}_{t-1} - z^*_{t-1})$, where the step size is denoted by η, and a value of 1 for η signifies a total replacement. Algorithm 1 provides a comprehensive outline of the entire procedure. The presented algorithm can be regarded as an ADMM-like(Mokady et al. 2022; Song, Meng, and Ermon 2022) approach that applies Null-text Inversion (NTI)(Mokady et al. 2022) to the inversion trajectory obtained through DDIM(Song, Meng, and Ermon 2022):

$$\min_{\theta_t} \|\tilde{z}_{t-1}(\tilde{z}_t, \theta_t, C') - z_{t-1}\|_2^2 \quad \text{s.t.} \quad \tilde{z}_{t-1} = z^*_{t-1} \tag{10}$$

where the objective is solved by Negative-prompt Inversion(NPI)(Miyake et al. 2023) and the constraint is enforced by inversion guidance.

Visual and Textual Prompts for Image Editing

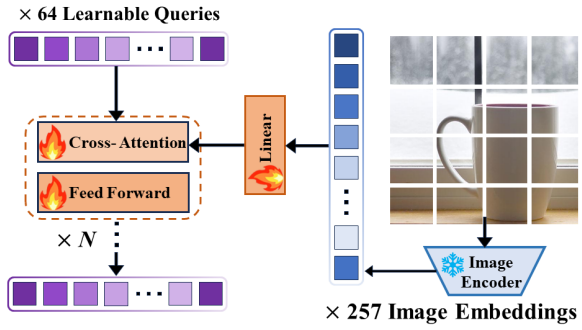

Figure 4: A depiction of our image prompt encoder's architecture.

Inspired by the architectural design of IP-Adapter(Ye et al. 2023), we have developed an innovative image prompt encoder, the conceptual framework of which is depicted in Fig. 4. The initial step in the process involves embedding an input image $\mathbf{x}_0$ into a sequence of 257 tokens utilizing a pre-trained CLIP(Ramesh et al. 2022b) image encoder. Subsequently, a linear layer is employed to modify the channel dimension, which is then followed by the application of a QFormer module(Zhang et al. 2024) (excluding the self-attention layer) to adjust the number of tokens to 64 through the utilization of 64 learnable queries. The QFormer module comprises N submodules (with a default value of 8), each of which incorporates a cross-attention layer followed by a feedforward network (FN) to process the input data. The information extraction process is facilitated by the 64 learnable queries, which interact with the 257 image tokens that function as keys and values within the QFormer module. The 257 image tokens are then aggregated and transformed into a set of 64 embedding tokens ($\mathbf{c}_{im}$), which are subsequently inputted into the same cross-attention module that processes the text tokens ($\mathbf{c}$) within the SD(Rombach et al. 2022) architecture. To facilitate the implementation of classifier-free guidance(Ho and Salimans 2022), we employ a training strategy that involves the simultaneous training of conditional and unconditional image prompts, similar to the approach used for text conditions. During the training process, random dropping is applied, which entails setting the image to zero, to enhance the model's ability to generate diverse and coherent outputs. In the final stage, the image tokens and text tokens are processed independently using the query $\mathbf{Q}$ within the cross-attention module. The resultant outputs from these separate processing streams are then combined through an element-wise addition operation to produce the final representation.

$$\text{Att}(\mathbf{Q}, \mathbf{K}', \mathbf{V}', \mathbf{K}'', \mathbf{V}'') = S(\frac{\mathbf{Q}\mathbf{K}'^T}{\sqrt{d}})\mathbf{V}' + \gamma \cdot S(\frac{\mathbf{Q}\mathbf{K}''^T}{\sqrt{d}})\mathbf{V}'' \tag{11}$$

In this equation, $(\mathbf{K}', \mathbf{V}')$ and $(\mathbf{K}'', \mathbf{V}'')$ represent the keys and values obtained from the text and image prompt, respectively. The parameter γ serves as a weight to balance the contributions of these two terms, while S denotes the Softmax function. It is crucial to note that in tasks involving reference images, such as object pasting and appearance replacing, $\mathbf{K}''$ and $\mathbf{V}''$ are formed by concatenating image tokens from both the source image and the reference image. During the training phase, the parameters in the pre-trained SD(Rombach et al. 2022) and CLIP(Ramesh et al. 2022b) image encoder remain fixed, and only the linear embedding and QFormer are optimized using the loss function $\mathcal{L}_2$, which is defined as:

$$\mathbb{E}_{\mathbf{x}_0, t, \epsilon_t \sim \mathcal{N}(0,1)} \left[\left\| \epsilon_t - \epsilon^t_\theta(\mathbf{z}_t, \mathbf{c}, \mathbf{c}_{im}) \right\|_2^2 \right] \tag{12}$$

Following a one-time training process, the module can be seamlessly incorporated into pre-trained SD(Rombach et al. 2022) to facilitate a wide range of image editing tasks, as showcased throughout this paper.

Experiments

Implementations

As the foundation for our image editing framework, we select Stable Diffusion V1.5(Rombach et al. 2022). For the

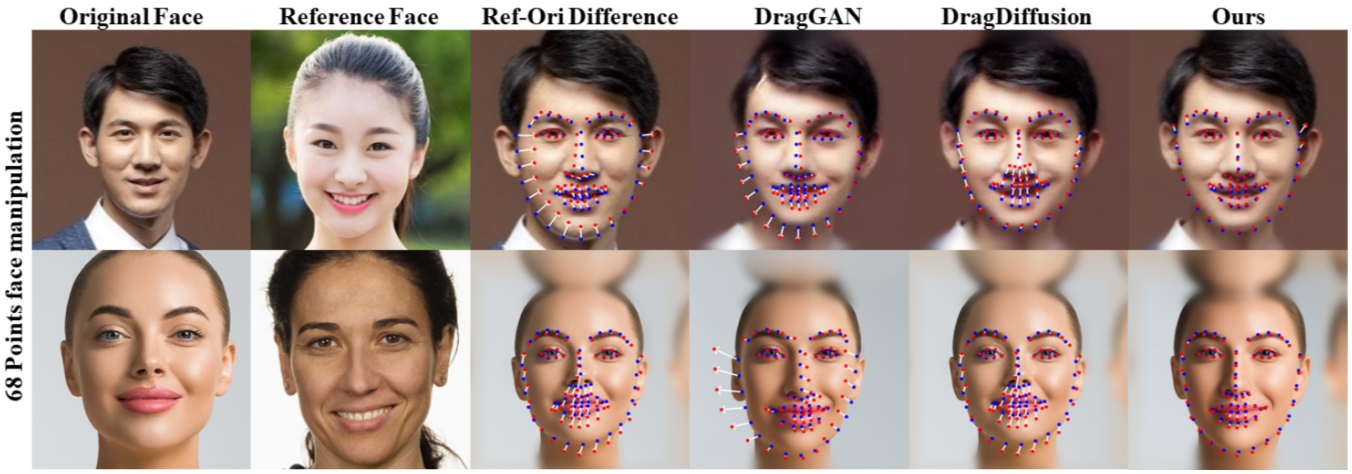

Figure 5: our VersaFusion approach to other face manipulation methods. Red and blue markers indicate current and target landmark points, respectively, with a white line depicting the distance between them.

	Complexity Preparation	Complexity during Inference	Face Aligned	From 36.36: 68 Points	FID
DragGAN(Pan et al. 2023)	75.33s	10.23s	×	**10.81**	39.38
DragDiff(Shi et al. 2023)	63.56s	29.28s	✓	16.99	36.60
DragonDiff(Shi et al. 2023)	5.23s	22.50s	✓	16.90	34.20
VersaFusion(Ours)	5.23s	20.82s	✓	11.45	**32.98**

Table 1: Quantitative Evaluation on Face Manipulation

training of image prompts, we utilize data from the LAION dataset and resize the images to a resolution of 512×512. We employ the Adam(Kingma and Ba 2014) optimizer with an initial learning rate set to 1×10^{-5}. During the training phase, we use a batch size of 16 and conduct the training process for 1×10^{6} iterations on a system equipped with 4 NVIDIA 3090 GPUs. Across different applications, we consistently use the same embedding module to process the image prompts. For inference, we adopt the DDIM sampling technique with 50 steps and set the scale for classifier-free guidance(Ho and Salimans 2022) to 5.

Comparison and Ablation Study

Our study presents a comparative examination of the time complexity associated with different methods, concentrating on the preparing and inference phases. The preparing phase involves the inversion of Diffusion/GAN models and the tuning of model parameters, whereas the inference phase focuses on generating the edited output from the latent representation. For a fair and unbiased comparison, we evaluate the time complexity of each method by considering a single point dragging operation performed on an image with a resolution of 512×512 pixels. The findings, summarized in Table. 1, Underscore the appealing complexity of our proposed method during the preparing stage. The sheet presents a quantitative assessment of face manipulation techniques using 68 landmark points. The evaluation metric used is the mean squared error (MSE) distance between the edited and target points, which measures the accuracy of the manipulation. The initial distance of the 36.36 serves as the upper bound, representing the baseline without any editing applied. To quantitatively assess the editing quality across various methods, the Fréchet Inception Distance (FID)(Seitzer 2020) is utilized as an additional metric. Moreover, we note that the inference complexity of our approach is comparatively lower than that of current diffusion-based techniques, including DragDiff(Shi et al. 2023) and DragonDiff(Mou et al. 2023). The results underscore the efficacy and feasibility of our method, considering both the preparation and inference time complexities, rendering it a compelling choice for image editing applications. While DragGAN(Pan et al. 2023) exhibits superior editing accuracy on aligned facial images, it is important to note that its base model is exclusively trained for this specific task and does not possess the ability to edit general face images, as depicted in Fig. 5. The qualitative comparison presented in Fig. 5 underscores the superior performance of our approach, which attains high levels of editing precision and content coherence while preserving a commendable degree of adaptability. For instance, when transferring a smile to a target face, our VersaFusion can generate more natural and realistic results, seamlessly integrating the desired expression. However, DragDiff(Shi et al. 2023) and DragonDiff(Mou et al. 2023) struggle to effectively transfer and imagine the smile on the target face, often resulting in inconsistencies or artifacts in the edited image.

Performance.As shown in Fig. 6, there are three types of inversion methods. In this paper, we employ the Proximal Negative-Prompt guidance(Miyake et al. 2023) method to optimize the DDIM inversion(Song, Meng, and Ermon 2022) process, providing guidance for image editing. To verify its effectiveness, we compare it with methods that have

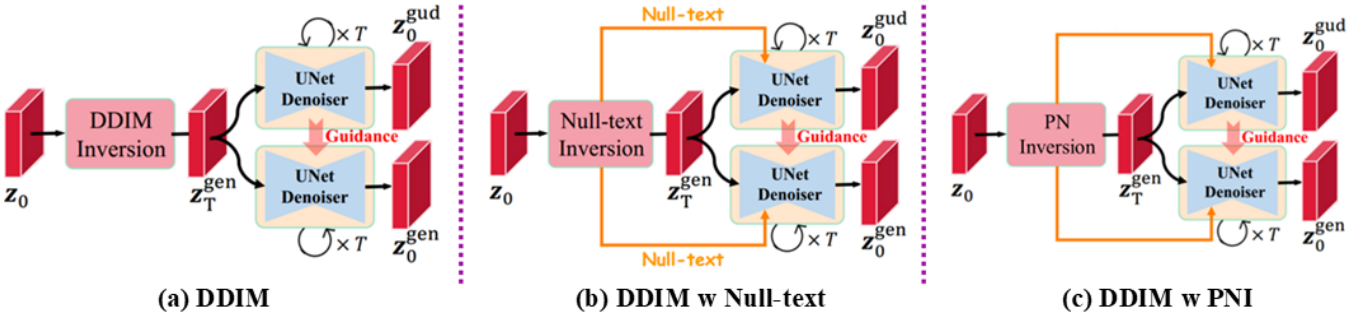

Figure 6: Different strategies for generating inversion prior ($i.e.$, $\mathbf{z}_T$) and guidance information ($i.e.$, $\mathbf{K}_i^{gud}, \mathbf{V}_i^{gud}$). (a) DDIM inversion; (b) NT inversion Mokady et al. (2023); (c) our PN inversion design.

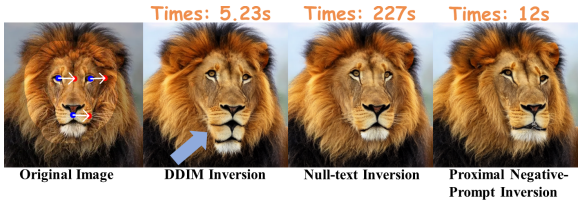

Figure 7: The editing quality of different guidance strategies.

the same function. Specifically, (a) guidance information is generated by a separate generation branch from z_T ; (b) null-text optimization(Mokady et al. 2022) is added based on method (a); (c) using PNI inversion. As Fig. 7 shows, extracting guidance information from z_T using DDIM inversion(Song, Meng, and Ermon 2022) can lead to deviations. This is due to the approximation bias in the DDIM inversion(Song, Meng, and Ermon 2022). Although incorporating null-text optimization(Mokady et al. 2022) can yield more accurate results, it comes with higher time complexity. The PNI method cleverly leverages the intermediate information stored during the DDIM inversion(Song, Meng, and Ermon 2022) process, achieving accurate results while maintaining a time complexity of only 12 seconds.

Fig. 8 illustrates the impact of visual and textual prompts on our image dragging editing approach using a kitten image as the base. The first row demonstrates the effectiveness of our method in generating semantically aligned edits guided by the text prompt "A lion", even in the absence of an image prompt. The second row provides insights into the individual contributions of image and text prompts. Without both prompts, the generated image deviates from the original kitten, while removing only the image prompt results in an edited image closely resembling the original. Conversely, omitting the text prompt generates an image similar to the lion cub, underscoring the influence of textual guidance in the editing process. These results highlight the complementary roles of visual and textual prompts in our framework, with the image prompt ensuring visual coherence and the text prompt guiding the semantic direction of the edits. This ablation study demonstrates the effectiveness and flexibility of our methodology in generating compelling and user-guided edits by balancing these components to achieve fine-grained control over the editing process.

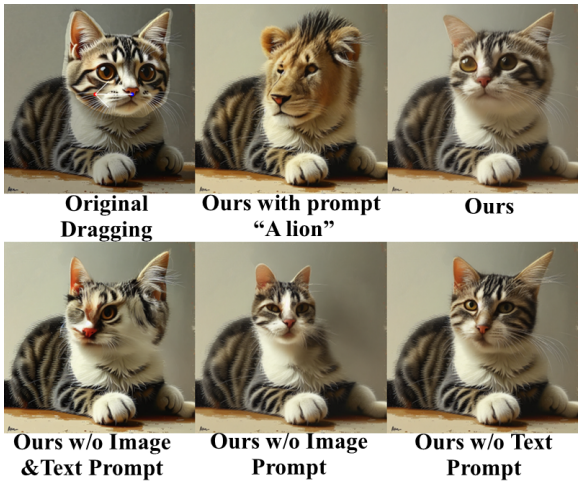

Figure 8: Image dragging editing. Shows the effectiveness of our text prompt-guided cross-attention and the absence of both visual and textual prompts.

Conclusion

We present VersaFusion, a method that demonstrates significant versatility in its ability to seamlessly integrate into a wide range of fine-grained image editing tasks without the need for task-specific training. Despite the flexibility offered by existing diffusion-based image editing projects, we observed that their functions remain insufficient, often resulting in editing inaccuracies and unexpected artifacts. To address these challenges, we propose a novel framework for the diffusion-based image editing pipeline, incorporating image prompts and a new inversion approach.

We acknowledge that certain editing scenarios, especially those demanding significant content imagination, may still pose challenges. Future research endeavors will focus on investigating sophisticated methods, including the integration of large-scale pre-trained language models and extensive user studies, to enhance the imaginative capabilities and usability of our method, while also investigating its potential application in other domains, such as video editing and 3D object manipulation.

Acknowledgements

This work was supported in part by the National Key Research and Development Plan of China (No. 2023YFC3604802), the National Natural Science Foundation of China (No. 62472406), the Youth Innovation Promotion Association of the Chinese Academy of Sciences, the Science and Technology Innovation Program of Hunan Province (No. 2022RC4006) and the Hunan Provincial Natural Science Foundation of China (No. 2023JJ70009).

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, 4432–4441.

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8296–8305.

Alaluf, Y.; Tov, O.; Mokady, R.; Gal, R.; and Bermano, A. 2022. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, 18511–18521.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

Bansal, A.; Chu, H.-M.; Schwarzschild, A.; Sengupta, S.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 843–852.

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

Dong, H.; Yu, S.; Wu, C.; and Guo, Y. 2017. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE international conference on computer vision*, 5706–5714.

Geng, D.; and Owens, A. 2024. Motion guidance: Diffusion-based image editing with differentiable motion estimators. *arXiv preprint arXiv:2401.18085*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Han, L.; Wen, S.; Chen, Q.; Zhang, Z.; Song, K.; Ren, M.; Gao, R.; Stathopoulos, A.; He, X.; Chen, Y.; et al. 2023. Improving Tuning-Free Real Image Editing with Proximal Guidance. *arXiv preprint arXiv:2306.05414*.

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*.

Hinz, T.; Fisher, M.; Wang, O.; and Wermter, S. 2021. Improved techniques for training single-image gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1300–1309.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

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

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577.

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.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. H. 2020. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7880–7889.

Ling, H.; Kreis, K.; Li, D.; Kim, S. W.; Torralba, A.; and Fidler, S. 2021. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34: 16331–16345.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Miyake, D.; Iohara, A.; Saito, Y.; and Tanaka, T. 2023. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. arXiv:2211.09794.

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*.

Nam, S.; Kim, Y.; and Kim, S. J. 2018. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems*, 31.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741.

Pan, X.; Tewari, A.; Leimkühler, T.; Liu, L.; Meka, A.; and Theobalt, C. 2023. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. arXiv:2305.10973.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022b. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, 1060–1069. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ryu, S. 2023. Low-rank adaptation for fast text-to-image diffusion fine-tuning. *Low-rank adaptation for fast text-to-image diffusion fine-tuning*.

Seitzer, M. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

Shaham, T. R.; Dekel, T.; and Michaeli, T. 2019. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4570–4580.

Shi, Y.; Xue, C.; Pan, J.; Zhang, W.; Tan, V. Y.; and Bai, S. 2023. DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing. *arXiv preprint arXiv:2306.14435*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

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*.

Xu, C.; Yang, S.; Galanti, T.; Wu, B.; Yue, X.; Zhai, B.; Zhan, W.; Vajda, P.; Keutzer, K.; and Tomizuka, M. 2022. Image2point: 3d point-cloud understanding with 2d image pretrained models. In *European Conference on Computer Vision*, 638–656. Springer.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

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.

Zhang, Q.; Zhang, J.; Xu, Y.; and Tao, D. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.