This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FireEdit: Fine-grained Instruction-based Image Editing via Region-aware Vision Language Model

Jun Zhou[1,2*], Jiahao Li[2], Zunnan Xu[3*], Hanhui Li[1], Yiji Cheng[2], Fa-Ting Hong[4*],
Qin Lin[2], Qinglin Lu[2], Xiaodan Liang[1†]
[1]Shenzhen Campus of Sun Yat-sen University [2]Hunyuan, Tencent [3]Tsinghua University [4]HKUST

Abstract

*Currently, instruction-based image editing methods have made significant progress by leveraging the powerful cross-modal understanding capabilities of vision language models (VLMs). However, they still face challenges in three key areas: 1) complex scenarios; 2) semantic consistency; and 3) fine-grained editing. To address these issues, we propose FireEdit, an innovative **F**ine-grained **I**nstruction-based image editing framework that exploits a **RE**gion-aware VLM. FireEdit is designed to accurately comprehend user instructions and ensure effective control over the editing process. Specifically, we enhance the fine-grained visual perception capabilities of the VLM by introducing additional region tokens. Relying solely on the output of the LLM to guide the diffusion model may lead to suboptimal editing results. Therefore, we propose a Time-Aware Target Injection module and a Hybrid Visual Cross Attention module. The former dynamically adjusts the guidance strength at various denoising stages by integrating timestep embeddings with the text embeddings. The latter enhances visual details for image editing, thereby preserving semantic consistency between the edited result and the source image. By combining the VLM enhanced with fine-grained region tokens and the time-dependent diffusion model, FireEdit demonstrates significant advantages in comprehending editing instructions and maintaining high semantic consistency. Extensive experiments indicate that our approach surpasses the state-of-the-art instruction-based image editing methods.*

1. Introduction

As a critical area of research in multimedia interaction, image editing has consistently garnered significant attention from both academia and industry. The emergence of diffusion models [15, 16, 48] and autoregressive models [50, 51] has reshaped the fields of image generation [7, 19, 32, 33] and video generation [23, 26, 28, 56, 57]. This advancement provides enhanced control and flexibility for generative image editing, enabling the development of numerous commercial products. By transferring the discrepancies between input and target text descriptions to a cross-modal implicit mapping of intermediate latent features, the desired regions of real images can be modified [3, 14, 24, 34, 35, 40]. Precise editing is achievable when region masks are available. However, fine-grained text descriptions and region masks are often difficult to obtain.

Instruction-based image editing methods eliminate the reliance on these seemingly essential conditions and are widely adopted in the community for their simplicity, user-friendliness, and efficiency. IP2P [4] is a pioneering work in instruction-based editing, allowing users to express their modification intentions for a specified image using straightforward natural language. To achieve precise and high-fidelity image editing, recent methods focus on large-scale generation of high-quality input-target-instruction triples



Figure 1. Our framework leverages a vision language model (VLM) to guide instruction-based image editing. Our primary innovation is the introduction of region tokens, which enable the VLM to accurately identify edited objects or areas in complex scenarios while preserving high-frequency details in unintended regions during image decoding.

*Work done during the internship at Tencent.
†Corresponding author.

[12, 22, 49, 61, 63, 66] and exploit the remarkable reasoning and comprehension capabilities of LLMs [11, 21, 62]. Despite the significant progress made by these methods in advancing instruction-based image editing, several shortcomings persist in the following scenarios: 1) complex instructions; 2) intricate scenes (e.g., multi-object environments and urban streets); and 3) fidelity degradation, which encompasses undesirable changes in style, texture, color, and other visual details. As illustrated in Figure 1, current methods either demonstrate inaccurate localization or inadvertently alter unintended objects. These challenges arise from two primary factors. First, the CLIP text encoder [46] employed by Stable Diffusion [48] struggles to extract the desired modifications from the instructions provided by general users. In complex scenarios, even when users invest considerable time crafting their instructions, aligning the text with the editing regions in the input images remains a challenge. Second, the data pairs used for training are predominantly synthesized by text-to-image (T2I) generation models [9, 14, 45, 48], which inevitably introduce implicit biases [10, 27]. Consequently, models trained on ambiguous or imprecise data often struggle to accurately identify the editing object and may lose high-frequency details in unintended areas of the input image.

To address these limitations, we propose FireEdit, a novel end-to-end, instruction-based image editing method. FireEdit primarily comprises a Vision Language Model (VLM) and a diffusion model. The VLM converts user-provided editing instructions into an edit representation that can be comprehended by the generative model. Previous methods [11, 21] have employed VLM to improve the model's understanding of editing instructions. MGIE [11] utilizes VLM to generate concise expression instructions, enhancing simple instructions with greater visual imagination. SmartEdit [21] addresses complex instructions by implementing a bidirectional interaction module that promotes mutual understanding between visual features and VLM outputs. Although these methods have achieved commendable results, they frequently encounter a trade-off between adherence to instructions and semantic consistency. Unlike the standard practice of adopting image tokens alongside text tokens as input to VLMs, we provide fine-grained visual features to ground textual instructions, thereby enhancing the reasoning capabilities of the VLM. Our insight is that directly utilizing LLMs to interpret editing intentions is not sufficiently effective. Consequently, we decouple the process of understanding instructions into two distinct tasks: locating the region of interest and associating the relevant text. We observe that the perceptual skills required for localization are not the strong suit of LLMs. By integrating fine-grained region features into the tokenization process of image features, we can complement the capabilities of LLMs. Specifically, we employ an open-vocabulary object detector [68] to identify potential regions of interest and tokenize each region into region tokens, which serve as a supplement to standard image tokens. The enhanced visual tokens, along with instruction tokens, form a new input for the LLM, which implicitly locates the editing regions and articulates the modification intentions through the LLM's powerful reasoning capabilities. Introducing region tokens improves the LLM's ability to perceive and understand the visual world.

Once the editing representation is obtained from the LLM response, a direct approach is to employ it for interaction with the diffusion model. However, due to the semantic gap between the output of VLM and the diffusion model, this approach may result in insufficient guidance. To address this issue, we introduce a Time-Aware Target Injection (TATI) module. As previously noted [18, 65], the diffusion model emphasizes low-frequency structures (e.g., edges and layouts) in the early stages of the denoising process, while it focuses on high-frequency details (e.g., textures and colors) in the later stages. Therefore, we propose TATI to associate the desired editing region and object shape at the low-frequency semantic level. Subsequently, during the later stages of denoising, TATI captures high-frequency details, linking the attributes of the target. To maintain spatial detail consistency between the source image and the edited image, we propose the Hybrid Visual Cross Attention (HVCA) module. HVCA integrates multi-scale visual features to deliver rich visual details, which may play a crucial role in preserving the content of non-edited areas during the intermediate feature evolution process in diffusion models. We utilize learnable queries to filter out the detailed features that need to be preserved through interaction with the text features. These two modules complement each other and enhance the controllability of the denoising process.

By integrating a VLM with fine-grained visual perception and a time-dependent diffusion model, FireEdit achieves high levels of semantic consistency and editability. Extensive quantitative and qualitative results demonstrate that our proposed method outperforms other instruction-based editing techniques in complex scenarios, semantic consistency, and fine-grained editing. Our main contributions can be summarized as follows:

- We propose FireEdit, a novel, versatile, and precise instruction-based image editing method that leverages a region-aware VLM to comprehend editing instructions in complex scenarios.
- We propose a time-aware target injection module and a hybrid visual cross-attention module to adaptively control the editing targets and preserve high-frequency details in non-edited regions.
- Extensive experiments demonstrate that our method outperforms other instruction-based editing methods, establishing a new state-of-the-art in this field.

2. Related Work

2.1. Instruction-based Image Editing

The instruction-based image editing task involves using human natural language to modify specified content in the input image while minimizing the distortion of undesired areas. Compared to the text-to-image generative task, the editing task typically entails targeted alterations to specific entities or characteristics within the image. Recently, there are some excellent works [2, 4, 8, 14, 30, 31, 53, 63] on image editing based on diffusion model. For instance, Prompt-to-Prompt [14] adjust the cross-attention features according to the difference between the source caption and the target caption. Building on this approach, IP2P [4] fine-tuned stable diffusion on a large-scale dataset of source-target-instruction triplets constructed by Prompt-to-Prompt [14], enabling the model to perform editing under the guidance of simple natural instructions. Subsequently, MagicBrush [63] further improves the model's editing capabilities with a real-world dataset. HQ-Edit [22] utilize DALL-E3 and GPT-4V [42] to construct a high-quality synthetic dataset for fine-tuning IP2P. InstructDiffusion [13] extends this instruction editing paradigm, trains a general model capable of handling various visual tasks across multiple datasets. UltraEdit [66] scales up data by combining SDXL with advanced inpainting methods [2, 41] and trains on a more powerful generative diffusion model [9]. To enhance the editing capabilities for complex instructions, SmartEdit [21] and MGIE [11] introduce VLMs to assist in training diffusion models, facilitate understanding and reasoning for complex instructions. Our method aims to achieve precise localization of editing targets in complex scenarios while preserves high-frequency details in non-edited regions.

2.2. VLMs with Diffusion Models

Large language models (LLMs) have demonstrated impressive generalization capability in text generation [6, 52]. Providing visual inputs to assist text generation has shown outstanding performances in a range of downstream language-to-vision tasks [20, 29, 37, 38, 47, 55, 58]. Bridging the gap between LLMs and visual inputs is a key focus of the latest Vision-Language Models (VLMs). LLaVA [37] fine-tune LLMs on an instruction-following dataset constructed by GPT-4. xGen-MM [58] replaces the Q-Former in BLIP-2 with a scalable visual token sampler, trains LLMs on large-scale, high-quality, and diverse datasets to advance various vision-language tasks. Leveraging the strong reasoning capabilities of vision-language models, researchers have begun to introduce them into the field of image generation. GILL [25] is a pioneering work in this attempt, guides diffusion models to generate images consistent with text by using text representations inferred from vision-language models. MGIE [11] extends this to image editing, leveraging VLMs to learn concise expressive instructions. SmartEdit [21] introduces a bidirectional interactive module to facilitate mutual understanding between the text representations outputted by VLMs and the input image features, addressing the issue of CLIP's text encoder's insufficient understanding of complex instructions. ILLUM [54] establishes a unified VLM with the capability of understanding images and charts, as well as image generation and editing. Unlike previous methods that train a custom VLM, PromptFix [62] utilizes the extensive world knowledge and understanding capabilities of the existing LLaVA to enhance text prompts. Our approach also focuses on training a dedicated VLM for editing tasks. Our insight is that a general VLM often struggles to effectively understand overly simple or complex instructions, necessitating the decoupling of the instruction understanding process into locating and associating target objects. We integrate fine-grained region tokens into standard image tokenization, providing the LLM with more visual imagination possibilities, thereby obtaining more robust and accurate text representations of user editing intentions.

3. Preliminary

The instruction-based editing method aims to modify a specific area of a given image x according to the provided natural language instruction c_T to obtain the edited image. InstructPix2Pix (IP2P), a latent diffusion model, is a pioneering work in this field. IP2P fine-tunes the pre-trained Stable Diffusion (SD) model on a synthesized large-scale dataset consisting of input-target-instruction triplets. Given a target image y and an image encoder $\mathcal{E}$, the diffusion process adds noise to the encoding latent $z = \mathcal{E}(y)$, resulting in a noisy latent z_t, with the noise level increasing over timestep $t \in T$. A diffusion model ϵ_δ, given the image condition c_I and the text instruction condition c_T, is trained to minimize

$$\mathcal{L}_{diff} = \mathbb{E}_{\mathcal{E}(y),\mathcal{E}(c_I),c_T,\epsilon,t}||\epsilon - \epsilon_\delta(z_t, t, \mathcal{E}(c_I), c_T)||_2^2, \quad (1)$$

where ϵ is the noise added to the noisy latent z_t. IP2P extends the classifier-free guidance (CFG) strategy to obtain score estimates in dual condition pattern and introduces an image guidance scale s_I and an instruction guidance scale s_T to balance the guidance strengths from both conditions:

$$\begin{aligned}\tilde{\epsilon}_\delta(z_t, t, c_I, c_T) = \epsilon_\delta(z_t, t, \emptyset, \emptyset) \\ + s_I \cdot (\epsilon_\delta(z_t, t, c_I, \emptyset) - \epsilon_\delta(z_t, \emptyset, \emptyset) \\ + s_T \cdot (\epsilon_\delta(z_t, t, c_I, c_T) - \epsilon_\delta(z_t, t, c_I, \emptyset)).\end{aligned} \quad (2)$$

During inference, our method also derives score estimates in this manner. An increase in s_I retains semantic details but weakens the instructions, while an increase in s_T strengthens the instructions, potentially leading to over-editing.

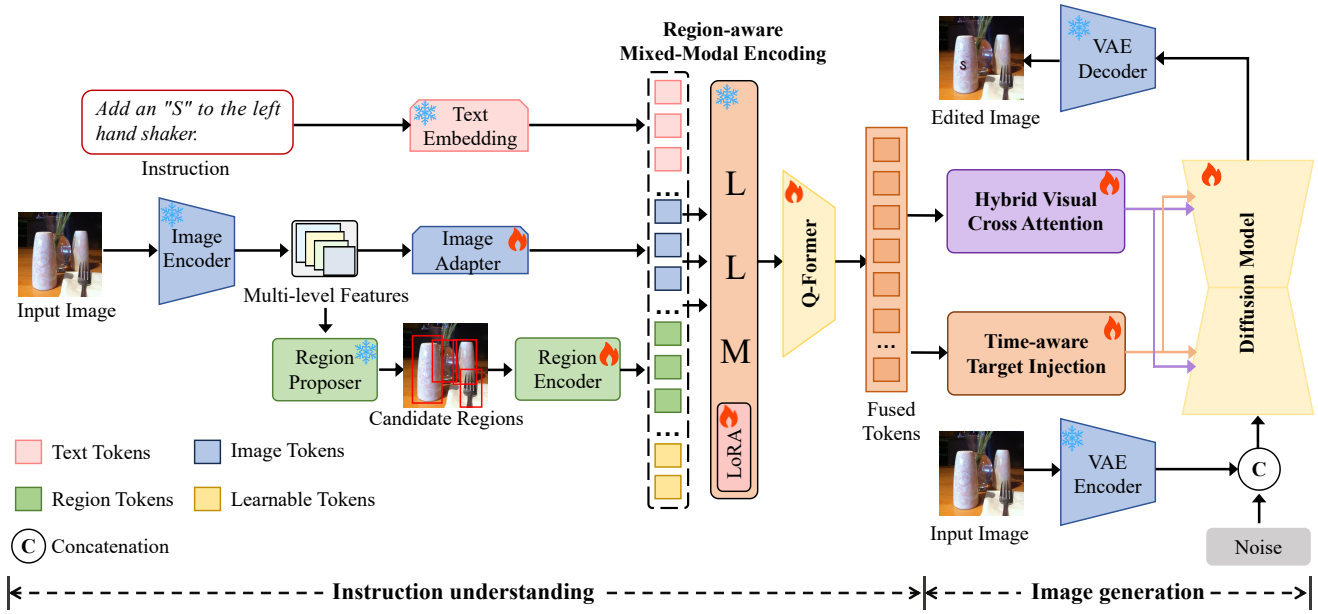


Figure 2. The overall framework of FireEdit. The core of FireEdit is to conduct region-aware fusion of multi-modal tokens to promote VLMs and facilitate fine-grained, localized alignments between editing instructions and images. It also introduces a hybrid visual cross-attention module to better preserve image details and a time-aware target injection module to edit targets adaptively.

4. Methodology

In this section, we provide a detailed introduction to the proposed FireEdit framework. FireEdit is designed to achieve precise and high-quality image editing tasks based on text instructions. Unlike existing methods that mainly rely on text encoders to understand user intentions, we propose to fine-tune a VLM by region-aware fusion of multi-modal tokens, so that we can exploit the extensive world knowledge of VLM and achieve model fine-grained alignments between images and instructions. As shown in Figure 2, the core components of FireEdit include a region-aware mixed-model encoder for token fusion (Sec. 4.1), a hybrid visual cross-attention module for preserving image details (Sec. 4.2), and a time-aware target injection module for incorporating target information into images adaptively (Sec. 4.3).

4.1. Region-aware Mixed-Modal Encoding

Given an input image x, a text instruction c, which is tokenized into a sequence $(s_1, s_2, \ldots, s_T)$, our goal is to achieve precise modification to x under the guidance of c without impacting other undesired regions. As shown in Figure 2, our model consists of a VLM and a diffusion model. Unlike the previous method that only uses the visual features extracted by the image encoder $\mu_\nu(\cdot)$ along with the text embeddings from $(s_1, s_2, \ldots, s_T)$ as the input of the LLM, $\mathcal{V}(\cdot;\theta)$, our method incorporates fine-grained visual tokens. Specifically, we obtain multi-level features from $\mu_\nu(\cdot)$ and use a sophisticated detector $\mathcal{D}(\cdot)$ to obtain potential candidate regions. A region encoder $\mathcal{R}(\cdot)$ is deployed to extract features of these regions of interest, which will serve as region tokens to provide fine-grained visual information for LLM. We define the process for obtaining the embeddings of region tokens as follows:

$$\mathcal{P} = \mathcal{R}(\mathcal{D}(x)), \tag{3}$$

here $\mathcal{P} \in \mathbb{R}^{L_r \times D}$, L_r indicates the length of the region tokens and D is the dimension of the embeddings.

The response of VLM is represented as a sequence of discrete tokens, which cannot be directly utilized by diffusion models. Current methods [21, 25] address this by expanding the vocabulary of LLaVA [37] with special [IMG] tokens. Similarly, our method introduces r additional special tokens, denoted as $\mathcal{Q} = \{[\text{IMG}_1], \ldots, [\text{IMG}_r]\}$. Based on this, we incorporate a learnable embedding matrix E into the LLM's embeddings. The learning objective of the VLM is thus transformed to predict these r tokens and extract their corresponding latent states as editing representations. Therefore, we restructure the input of the VLM to include a cascade of holistic image feature tokens $\mu_\nu(x)$, region tokens $\mathcal{P}$, text embeddings c, and learnable query tokens $\mathcal{Q}$. The process of understanding the input instructions can be formalized as follows:

$$\begin{aligned} \mathcal{H} &= [c, \mu_\nu(x), \mathcal{P}, \mathcal{Q}], \\ e &= \mathcal{V}(\mathcal{H};\theta), \end{aligned} \tag{4}$$

$[\cdot,\cdot]$ refers to the concatenation operation, and $e \in \mathcal{R}^{r \times D}$ corresponds to the hidden embeddings of r special tokens. Due to the semantic gap between the text embedding space

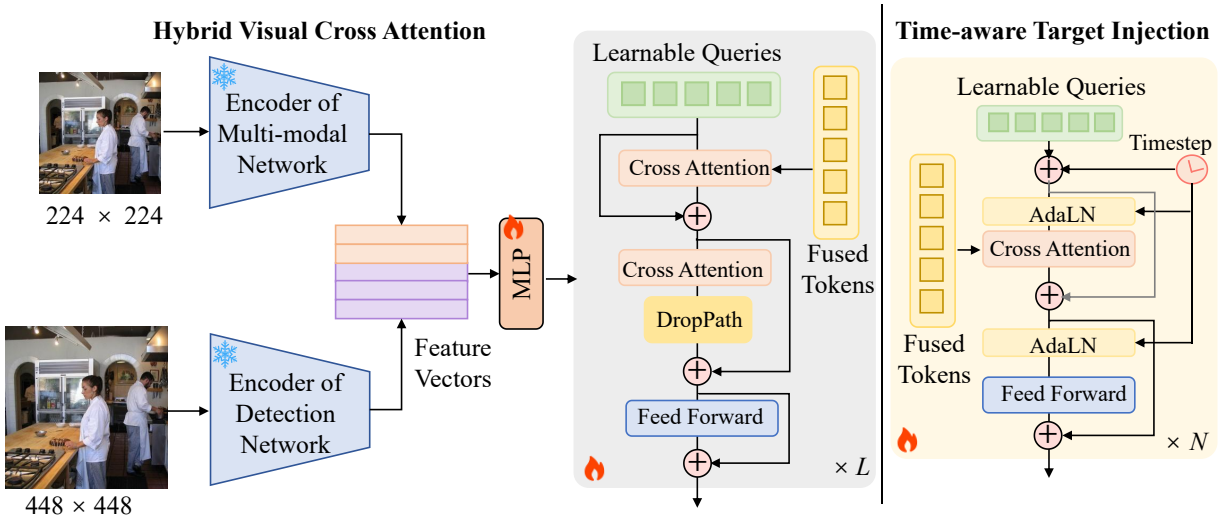


Figure 3. The proposed hybrid visual cross attention module (HVCA, left) and the time-aware target injection module (TATI, right). HVCA exploits hybrid encoders of pre-trained networks to balance between global semantic information (e.g., multi-modal alignment networks like CLIP) and local details. TATI utilizes timestep embeddings to incorporate target information into the denoising process.

of the diffusion model and the output of the LLM, an additional Q-Former module is required to bridge them. Finally, we obtain the edited representation e_t, which is derived from e through the Q-Former, is used to guide the diffusion model in the subsequent steps.

To effectively train the VLM, we freeze most of the parameters of the LLM and employ LoRA [17] for fine-tuning. Additionally, we freeze the image encoder and train an adapter W to adapt the image features to the latent space of the LLM. We use a pre-trained Deformable DETR [69] to extract region proposals and a region encoder transforms the region features into region tokens. Consequently, the i-th [IMG$_i$] token is generated by minimizing the negative log-likelihood based on the previously generated tokens

$$\mathcal{L}_{VLM} = -\sum_{i=1}^{r} \log p(\theta \cup \Delta\theta)([\text{IMG}_i]|\mu_\nu(x)^T W, \mathcal{P}, c, [\text{IMG}_1], \ldots, [\text{IMG}_{i-1}]), \quad (5)$$

where $\Delta\theta$ denotes the parameter of LoRA.

4.2. Hybrid Visual Cross Attention

To prevent the loss of viusal details in non-target areas, we introduce a Hybrid Visual Cross Attention (HVCA) module. As shown in Figure 3, we cascade the visual features extracted by the CLIP image encoder [46] and DINOv2 [43] as visual inputs. The use of different image encoders is intended to capture rich visual details at various pixel levels. These features are then enhanced through interaction with text-related features to obtain reinforced visual features. Our HVCA consists of L attention blocks, each comprising two cross-attention layers and a feed-forward network. The IP-Adapter [59] directly injects features extracted by the image encoder into the denoising network, which makes it challenging to maintain high-frequency details, especially in the field of image editing. By introducing visual content with context-aware capabilities, we ensure that the generated images have higher fidelity while maintaining editability.

Following the IP-Adapter [59], we decouple the cross-attention layers to inject the refined visual embeddings v and the edited representation e_t into the cross-attention layers of the denoising network. Specifically, given the query features Z, the edited representation e_t, and the visual embeddings v, the output Z' of the cross-attention can be defined as follows:

$$Z' = \text{Softmax}(\frac{QK_1^T}{\sqrt{d}})V_1 + \lambda \cdot \text{Softmax}(\frac{QK_2^T}{\sqrt{d}})V_2, \quad (6)$$

where $Q = ZW_q$, $K_1 = e_t W_{k_1}$, $V_1 = e_t W_{v_1}$, and $K_2 = vW_{k_2}$, $V_2 = vW_{v_2}$ are query, key, and value matrices. λ is weight factor, and d represents the feature dimension.

4.3. Time-aware Target Injection

Previous research has demonstrated that diffusion models for text-to-image generation tend to focus on low-frequency structures during the early stages of denoising, while concentrating on high-frequency details in the later stages. Instruction-based image editing retains this characteristic, attending to different transformations at various stages of the denoising process [67]. Unlike [67], which optimizes instruction features by splitting different denoising stages, we associate sampling timesteps for editing instructions using a resampler that incorporates time embeddings. We refer to the proposed module as the Time-Aware Target Injection (TATI) module, which is inspired by [18]. Although structurally similar, the motivations are distinct. TATI in-

Table 1. Results on the Emu Edit test. We compare our approach with other instruction-based image editing baselines. ↑ indicates the higher the better, while ↓ indicates the lower the better. The best results are in bold.

Method	L1↓	CLIP-I↑	DINO↑	LPIPS↓	CLIP-T↑	User Study↑
IP2P	0.1225	0.8570	0.7616	0.2723	0.2715	1.83%
MagicBrush	0.0991	0.8670	0.7810	0.2551	0.2744	5.00%
HQ-Edit	0.2541	0.7095	0.5404	0.5397	0.2566	4.92%
UltraEdit	0.0838	0.8120	0.7406	0.2826	0.2773	5.00%
MGIE	0.1628	0.7456	0.5944	0.3801	0.2314	1.25%
SmartEdit	0.1186	0.8592	0.7705	0.2717	0.2740	3.33%
Ours	**0.0574**	**0.9140**	**0.8829**	**0.1373**	**0.2783**	**78.67%**

corporates timestep information across the diffusion process, optimizing text representation utilization. Specifically, our TATI consists of N stacked resamplers [1], where temporal information is integrated into adaptive Layer Normalization [44] within each resampler to establish time dependencies.

4.4. Optimization Objectives

During the diffusion process, we follow the design principles of IP2P [4]. We concatenate the encoded image latent $\mathcal{E}(x)$ and the noisy latent z_t as input to the denoising network ϵ_δ. The difference lies in that we substitute the text embeddings obtained by the CLIP text encoder with e_t, and use it along with the visual feature v through the decoupled cross-attention layers of the UNet to participate in the denoising process. The optimization process for ϵ_δ can be expressed as follows:

$$\mathcal{L}_{diff} = \mathbb{E}_{\mathcal{E}(y),\mathcal{E}(x),c,\epsilon,t}||\epsilon - \epsilon_\delta(t, \mathcal{C}(z_t, \mathcal{E}(x)), e_t, v)||_2^2, \tag{7}$$

where $\epsilon \sim \mathcal{U}(0,1)$, t is the sampling time, y is the target edited images and $\mathcal{C}$ denotes the operation of concatenation along the channel dimension. The combined total loss function for the VLM and the diffusion model is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{VLM} + \mathcal{L}_{diff}. \tag{8}$$

5. Experiments

5.1. Implementation Details

Model Architecture: Following [11, 21], our visual large language model retains the large language model component of LLAVA-7B, with the visual encoder being DINOv2 and a two-layer linear layer connecting the LLM and the image encoder. We use Deformable DETR [69] as the region detector and ROIAlign [47] as the region encoder, and their pre-trained weights are derived from [38]. Our Q-Former uses the same structure as [21]. TATI contains 2 units and 77 learnable queries, while the HVCA module has 2 units and 16 learnable queries. For LLM, we freeze its parameters and use LoRA for fine-tuning, with rank and alpha of 8 and 16, respectively. We implement experiments on Stable Diffusion 1.5, and its parameters are initialized by IP2P [4].

Table 2. Results on the MagicBrush test set. We include both single-turn and multi-turn settings. The results demonstrate that our method outperforms other VLM-guided editing techniques.

Settings	Method	L1↓	L2↓	CLIP-I↑	DINO↑
Single-turn	Global Description-guided				
	SDEdit	0.1014	0.0278	0.8526	0.7726
	NTI	0.0749	0.0197	0.8827	0.8206
	GLIDE	3.4973	115.8347	0.9487	0.9206
	BleDiff	3.5631	119.2813	0.9291	0.9644
	Instruction-guided				
	HIVE	0.1092	0.0380	0.8519	0.7500
	IP2P	0.1141	0.0371	0.8512	0.7437
	MagicBrush	0.0625	0.0203	0.9332	0.8987
	MGIE	0.1624	0.0669	0.7454	0.5771
	SmartEdit	0.0936	0.0377	0.8945	0.8201
	Ours	**0.0701**	**0.0238**	**0.9131**	**0.8619**
Multi-turn	Global Description-guided				
	SDEdit	0.1616	0.0602	0.7933	0.6212
	NTI	0.1057	0.0335	0.8468	0.7529
	GLIDE	11.7487	1079.5997	0.9094	0.8498
	BleDiff	14.5439	1510.2271	0.8782	0.7690
	Instruction-guided				
	HIVE	0.1521	0.0557	0.8004	0.6463
	IP2P	0.1345	0.0460	0.8304	0.7018
	MagicBrush	0.0964	0.0353	0.8924	0.8273
	MGIE	0.1912	0.0836	0.7008	0.5208
	SmartEdit	0.1218	0.0510	0.8601	0.7538
	Ours	**0.0911**	**0.0326**	**0.8819**	**0.8010**

Training Dataset: Our training data can be divided into three categories: (1) segmentation dataset, which comes from COCOStuff [5], RefCOCO [60], GRefCOCO [36]; (2) image editing dataset, including InstructPix2Pix [4], MagicBrush [63], Ultraedit [66], and ReasonEdit [21]; (3) visual question answering dataset, LLAVA-instruct-150k [37].
Benchmark and Metrics: We validate our approach on two popular image editing test sets: MagicBrush test [63] and Emu Edit test [49]. MagicBrush evaluates the model across multiple metrics by comparing the generated edited images to the ground truth images and the corresponding captions. Following [63, 66], we use L1 distance, L2 distance, CLIP image similarity (CLIP-I), and DINO similarity as metrics. The Emu Edit test set verifies the edited images against the source images and the target captions. In addition to the above metrics, we also use CLIP text-image similarity and LPIPS.
Baseline Models: We make comparisons with the state-of-the-art instruction-based image editing methoss, including IP2P [4], MagicBrush [63], HIVE [64], HQ-Edit [22], UltraEdit [66], MGIE [11],and SmartEdit [21]. In addition, we compare with description-based image editing methods, including SDEdit [39], NTI [40], GLIDE [41], and BleDiff [2] on the MagicBrush test set.

5.2. Comparisons with SOTA models

Quantitative Evaluation. We conducted quantitative evaluations on the Emu Edit and MagicBrush test sets. Table 1 shows the quantitative results on the Emu Edit test set,

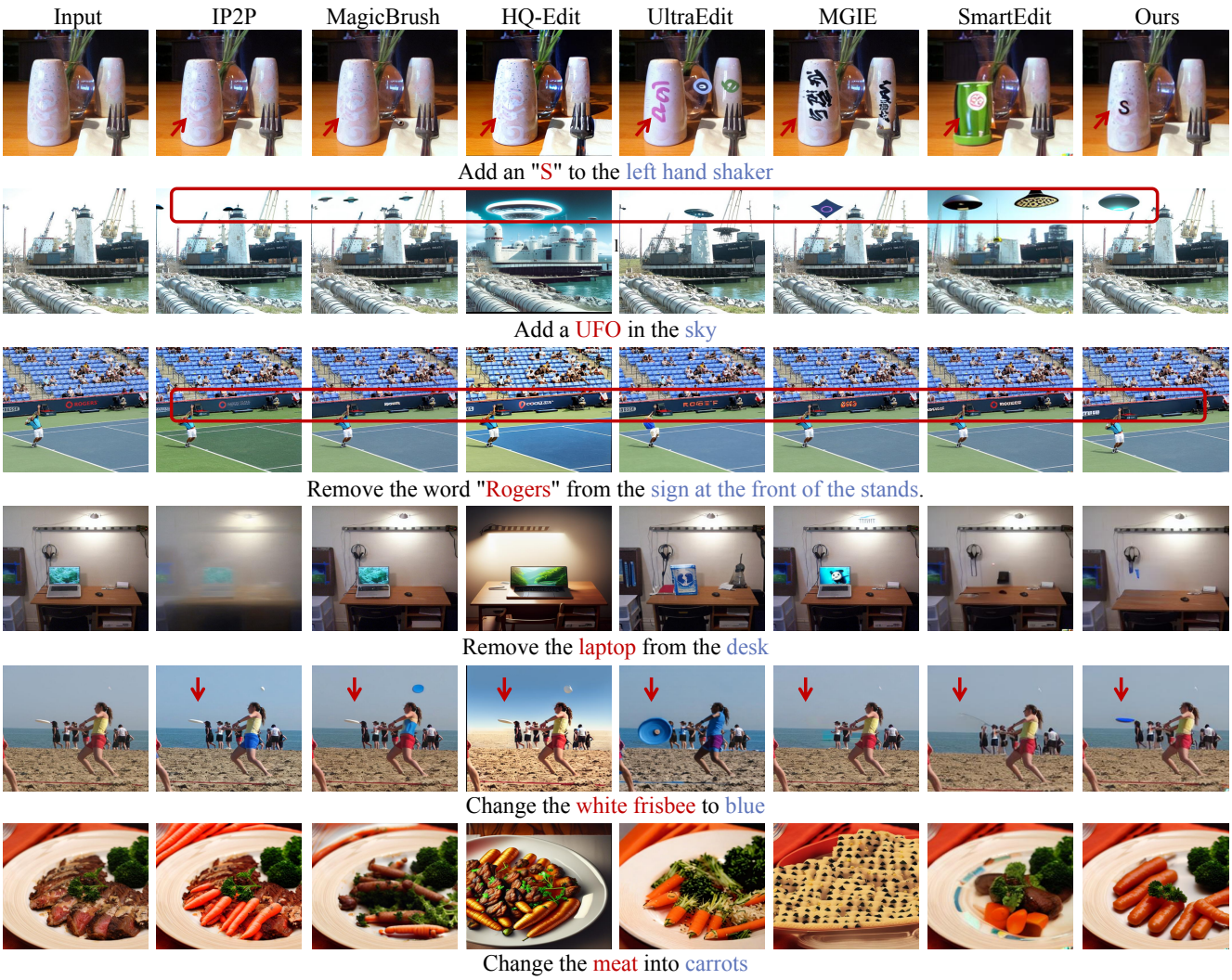


Figure 4. Qualitative comparison. We compare the editing performance of FireEdit with SOTA methods on the Emu Edit test set. Each editing instruction is written below each row of images. Compared with other SOTA methods, our approach is superior in accurately locating the edited objects or regions and preserving the detailed information of the input image.

revealing that our method outperforms competitors in metrics related to semantic structure and quality, suggesting that FireEdit more effectively preserves non-targeted editing regions. In addition, compared with VLM-based editing methods such as SmartEdit and MGIE, our approach significantly enhances semantic consistency without compromising its robust editing capabilities.

Table 2 shows the quantitative results on the MagicBrush benchmark, where we conducted a comprehensive comparison across both single-turn and multi-turn editing scenarios. Our method demonstrates a distinct advantage over VLM-guided counterparts in preserving non-editing areas. Notably, MagicBrush performs exceptionally well across various metrics, likely due to biases introduced from its training set. To further demonstrate the effectiveness of our approach, we conduct user studies from three perspectives: editing instruction alignment, semantic consistency, and image quality. Table 1 shows the user preference for our approach, with higher user scores. More details of the User Study can be found in the supplementary materials.

Qualitative Comparison. In Figure 4, we illustrate the editing results of the SOTA method and our method. We selected 6 sets of edited images across 3 types of editing actions (add, remove, change), and the inputs were all from real images. Our method achieves precise localization of the edited area while preserving the remaining pixels, which is particularly important for perceiving high-frequency details (texture, color, etc.). For example, in the first row of Figure 4, our method can accurately add the letter ”S” while recognizing the location of the target pointed to by ”the left hand shaker” in the input image. In contrast, other methods either fail to perceive the correct location or do not accu-

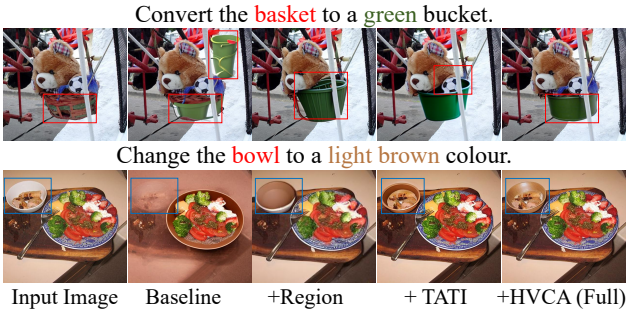


Figure 5. Ablation studies for components in our method.

rately add the "S" symbol. In the second row, HQ-Edit correctly adds the UFO, but causes the degradation of the semantic structure, similar issues are observed with UltraEdit and SmartEdit. Our method, however, achieves local editing without changing the semantic structure. In the fifth row, our method can locate the "white Frisbee" from the complex scene and modify it to blue. IP2P fails to locate the desired editing object, and UltraEdit modifies irrelevant objects (the color of the lady's top next to the Frisbee is changed).

5.3. Ablation Study

The quantitative results of the ablation study we conducted are presented in Table 3, and some representative visualizations can be seen in Figure 5.

Effect of Region-aware Mixed-Modal Encoding (Region). To validate the effectiveness of region-aware mixed-modal encoding, we remove this component and make comparison on the Emu edit test set. As illustrated in Figure 5, the model struggles to accurately identify the editing targets without the region tokens, leading to the unintended alteration of content in undesired areas. This is supported by the metrics assessing semantic structure presented in Table 3. The incorporation of region tokens significantly improves the VLM's capability to locate the intended targets.

Effect of Time-aware Target Injection (TATI). We confirm the crucial role of the TATI module in balancing semantic consistency and editability. As shown in Table 3, we observe a significant improvement in the metrics measuring semantic consistency, while the instruction-following metrics remained at a satisfactory level. The TATI module facilitates adaptive adjustments to the guidance strength at different denoising stages. As shown in the first row of Figure 5, TATI mitigates the erosion of non-editing regions. The model's enhanced instruction-following capability allows for precise modifications to the editing targets while preserving the content of non-editing regions.

Effect of and Hybrid Visual Cross Attention (HVCA). To demonstrate the effectiveness of HVCA, we integrate it with region tokens and the TATI module. As shown in Table 3, after implementing HVCA, our method demonstrates improvements in the L1, CLIP-T, and LPIPS metrics, indicating its effectiveness in retaining semantic details. As illustrated in Figure 3, HVCA integrates visual features across multiple scales, thereby enhancing visual detail. The interaction between the fused visual features and the editing instructions strengthens the preservation of high-frequency details, effectively incorporating visual content into the denoising process of the diffusion model.

Table 3. Ablation study on the Emu Edit test. The model's performance benefits from all of the components.

Region	TATI	HVCA	L1↓	CLIP-I↑	LPIPS↓	CLIP-T↑
-	-	-	0.0897	0.8501	0.3174	0.2770
✓	-	-	0.0859	0.8709	0.2391	0.2762
✓	✓	-	0.0609	**0.9258**	0.1580	0.2746
✓	✓	✓	**0.0574**	0.914	**0.1373**	**0.2783**

6. Limitations

Although FireEdit outperforms other baseline methods in comprehending simple text instructions within complex scenarios, it does not support reference images as input. Like other instruction-based image editing techniques, our approach aims to tackle challenging tasks such as altering poses and repositioning objects. In future work, a promising direction is to incorporate reference images to enhance control over the editing target and leverage visual autoregression [51] to push the boundaries of image editing.

7. Conclusion

In this paper, we propose FireEdit, a novel instruction-based image editing method that achieves precise localization of desired editing regions in complex scenarios through a region-aware vision language model. We introduce a region-aware multi-modal encoder in FireEdit to interpret editing instructions in a fine-grained manner, without the need for manually selected masks or reference regions. Furthermore, FireEdit introduces a hybrid visual cross-attention module and a time-aware target injection module to capture the characteristics of non-target areas and control the editing process adaptively. Extensive experimental results on two public datasets demonstrate that FireEdit outperforms state-of-the-art methods in terms of instruction adherence and semantic consistency.

Acknowledgements

This work is supported by National Key Research and Development Program of China (2024YFE0203100), Shenzhen Science and Technology Program No.GJHZ20220913142600001, National Natural Science Foundation of China (NSFC) under Grants No.62476293 and No.62372482, Nansha Key R&D Program under Grant No.2022ZD014 and General Embodied AI Center of Sun Yat-sen University.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 6

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 3, 6

[3] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 1

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 3, 6

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 3

[7] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 1

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023. 3

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3

[10] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 2

[11] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 6

[12] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 2

[13] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12709–12720, 2024. 3

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[18] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 2, 5

[19] Jiehui Huang, Xiao Dong, Wenhui Song, Hanhui Li, Jun Zhou, Yuhao Cheng, Shutao Liao, Long Chen, Yiqiang Yan, Shengcai Liao, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024. 1

[20] Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Densely connected parameter-efficient tuning for referring image segmentation. *arXiv preprint arXiv:2501.08580*, 2025. 3

[21] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 2, 3, 4, 6

[22] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024. 2, 3, 6

[23] Xiaoyu Jin, Zunnan Xu, Mingwen Ou, and Wenming Yang. Alignment is all you need: A training-free augmentation strategy for pose-guided video generation. *arXiv preprint arXiv:2408.16506*, 2024. 1

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1

[25] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4

[26] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1

[27] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[28] Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Chen Long. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*, 2024. 1

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[30] Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, et al. Zone: Zero-shot instruction-guided local editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6254–6263, 2024. 3

[31] Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, Ying Shan, Yuexian Zou, and Qiang Xu. Brushedit: All-in-one image inpainting and editing. *arXiv preprint arXiv:2412.10316*, 2024. 3

[32] Yaowei Li, Lingen Li, Zhaoyang Zhang, Xiaoyu Li, Guangzhi Wang, Hongxiang Li, Xiaodong Cun, Ying Shan, and Yuexian Zou. Blobctrl: A unified and flexible framework for element-level image generation and editing. *arXiv preprint arXiv:2503.13434*, 2025. 1

[33] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1

[34] Xiaojian Lin, Hanhui Li, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. Gdrag: Towards general-purpose interactive editing with anti-ambiguity point diffusion. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[35] Yukang Lin, Hokit Fung, Jianjin Xu, Zeping Ren, Adela SM Lau, Guosheng Yin, and Xiu Li. Mvportrait: Text-guided motion and emotion control for multi-view vivid portrait animation. *arXiv preprint arXiv:2503.19383*, 2025. 1

[36] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 6

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 4, 6

[38] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2025. 3, 6

[39] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 6

[40] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 1, 6

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3, 6

[42] R OpenAI. Gpt-4v(ision) system card. *View in Article*, 2(5), 2023. 3

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 6

[45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[47] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 3, 6

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[49] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 2, 6

[50] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 1

[51] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 1, 8

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[53] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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*, pages 1921–1930, 2023. 3

[54] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024. 3

[55] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023. 3

[56] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[57] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, Qin Lin, Xiu Li, and Qinglin Lu. Hunyuanportrait: Implicit condition control for enhanced portrait animation. *arXiv preprint arXiv:2503.18860*, 2025. 1

[58] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 3

[59] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 5

[60] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 6

[61] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024. 2

[62] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024. 2, 3

[63] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 6

[64] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024. 6

[65] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 2

[66] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *arXiv preprint arXiv:2407.05282*, 2024. 2, 3, 6

[67] Ruoyu Zhao, Qingnan Fan, Fei Kou, Shuai Qin, Hong Gu, Wei Wu, Pengcheng Xu, Mingrui Zhu, Nannan Wang, and Xinbo Gao. Instructbrush: Learning attention-based instruction optimization for image editing. *arXiv preprint arXiv:2403.18660*, 2024. 5

[68] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 2

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5, 6