

# GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS

🍏 Tsu-Jui Fu<sup>1</sup>, Wenze Hu<sup>2</sup>, Xianzhi Du<sup>2</sup>, William Yang Wang<sup>1</sup>, Yinfei Yang<sup>2</sup>, Zhe Gan<sup>2</sup>

<sup>1</sup>UC Santa Barbara, <sup>2</sup>Apple

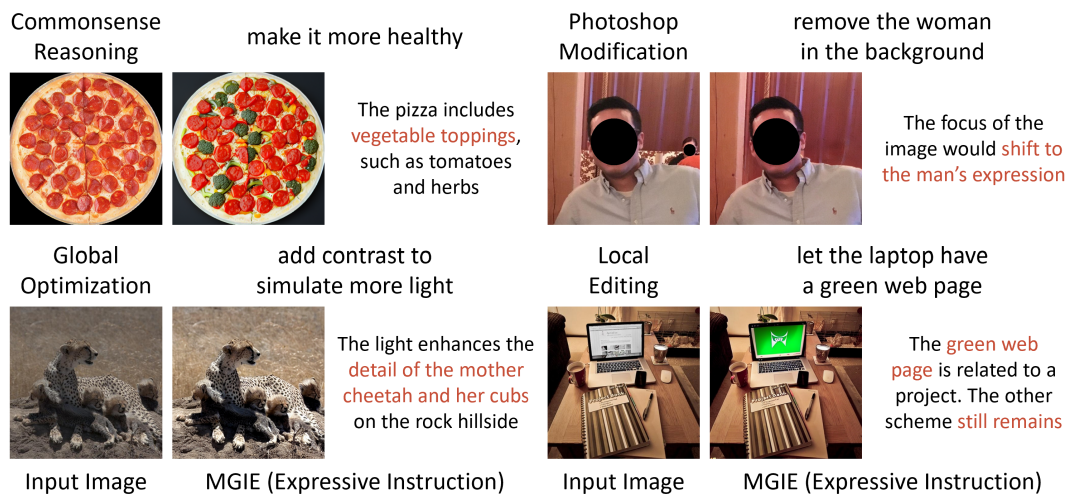


Figure 1: We introduce MLLM-Guided Image Editing (MGIE) to improve instruction-based image editing for various editing aspects. The top is the input instruction, and the right is the jointly derived expressive instruction by MGIE.

## ABSTRACT

Instruction-based image editing improves the controllability and flexibility of image manipulation via natural commands without elaborate descriptions or regional masks. However, human instructions are sometimes too brief for current methods to capture and follow. Multimodal large language models (MLLMs) show promising capabilities in cross-modal understanding and visual-aware response generation via LMs. We investigate how MLLMs facilitate edit instructions and present MLLM-Guided Image Editing (MGIE). MGIE learns to derive expressive instructions and provides explicit guidance. The editing model jointly captures this visual imagination and performs manipulation through end-to-end training. We evaluate various aspects of Photoshop-style modification, global photo optimization, and local editing. Extensive experimental results demonstrate that expressive instructions are crucial to instruction-based image editing, and our MGIE can lead to a notable improvement in automatic metrics and human evaluation while maintaining competitive inference efficiency.

## 1 INTRODUCTION

Visual design tools are widely adopted in various multimedia fields nowadays. Despite considerable demand, they require prior knowledge to operate. To enhance controllability and accessibility, text-guided image editing has obtained popularity in recent studies (Li et al., 2020; Patashnik et al., 2021; Crowson et al., 2022; Gal et al., 2022). With an attractive ability to model realistic images, diffusion models (Ho et al., 2020) are also adopted in image editing (Kim et al., 2022). By swapping the latent cross-modal maps, models can perform visual manipulation to reflect the alteration of the input-goal

🍏 Work done during an internship at Apple. Project website: <https://mllm-ie.github.io>

caption (Hertz et al., 2023; Mokady et al., 2022; Kawar et al., 2023). They can further edit a specific region by a guided mask (Nichol et al., 2022; Avrahami et al., 2022). Instead of relying on elaborate descriptions or regional masks, instruction-based editing (El-Nouby et al., 2019; Li et al., 2020; Fu et al., 2020) allows human commands that directly express how and which aspect of an image to edit. This flexibility also benefits practicality as such guidance is more aligned with human intuition.

Due to the data scarcity of the input-goal-instruction triplet, InsPix2Pix (Brooks et al., 2023) collects a curated IPr2Pr dataset. The instruction is generated by GPT-3 (Brown et al., 2020), and the input-goal image pair is synthesized from Prompt-to-Prompt (Hertz et al., 2023). InsPix2Pix then applies a pre-trained CLIP text encoder (Radford et al., 2021) to lead the diffusion model along with the input image. Although having feasible results, CLIP is trained for static descriptions, which is challenging to capture the essential visual transformation in editing. Furthermore, the instruction is too brief but ambiguous and insufficient to guide toward the intended goal. The deficiency limits the effectiveness of InsPix2Pix in instruction-based image editing.

Large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023) have shown significant advancement in diverse language tasks, including machine translation, text summarization, and question answering. Learning from large-scale corpora with diverse content, LLMs contain latent visual knowledge and creativity, which can assist various vision-and-language tasks (Wu et al., 2023; Feng et al., 2023; Chakrabarty et al., 2023). Upon LLMs, multimodal large language models (MLLMs) can treat images as input naturally and provide visual-aware responses to serve as multimodal assistants (Zhang et al., 2023b; Liu et al., 2023; Zhu et al., 2023; Koh et al., 2023).

Inspired by MLLMs, we incorporate them to deal with the insufficient guidance issue of instructions and introduce MLLM-Guided Image Editing (MGIE). As demonstrated in Fig. 2, MGIE consists of an MLLM and a diffusion model. The MLLM learns to derive concise expressive instructions and offers explicit visual-related guidance. The diffusion model is jointly updated and performs image editing with the latent imagination of the intended goal via end-to-end training. In this way, MGIE benefits from the inherent visual derivation and addresses ambiguous human commands to achieve reasonable editing. For the example in Fig. 1, it is difficult to capture what “*healthy*” means without additional context. Our MGIE can precisely connect “*vegetable toppings*” with the pizza and lead to the related editing as human expectation.

To learn instruction-based image editing, we adopt IPr2Pr as our pre-training dataset. We consider different editing aspects in EVR (Tan et al., 2019), GIER (Shi et al., 2020), MA5k (Shi et al., 2022), and MagicBrush (Zhang et al., 2023a). MGIE performs Photoshop-style modification, global photo optimization, and local object alteration. All should be guided by human instructions. Experimental results indicate that our MGIE significantly strengthens instruction-based image editing with reasonable expressive instructions in automatic metrics and human evaluation, and visual-aware guidance is crucial to this improvement. In summary, our contributions are three-fold:

- We introduce MLLM-Guided Image Editing (MGIE), which jointly learns the MLLM and editing model with visual-aware expressive instructions to provide explicit guidance.
- We conduct comprehensive studies from various editing aspects, including Photoshop-style modification, global photo optimization, and local editing, along with qualitative comparisons.
- Extensive experiments demonstrate that visual-aware expressive instructions are crucial for image editing, and our MGIE effectively enhances editing performance.

## 2 RELATED WORK

**Instruction-based Image Editing.** Text-guided image editing can significantly improve the controllability and accessibility of visual manipulation by following human commands. Previous works built upon the GAN frameworks (Goodfellow et al., 2015; Reed et al., 2016) to alter images but are limited to unrealistic synthesis or specific domains (Nam et al., 2018; Li et al., 2020; El-Nouby et al., 2019; Fu et al., 2020; 2022). With promising large-scale training, diffusion models (Ho et al., 2020; Ramesh et al., 2022; Sahari et al., 2022; Rombach et al., 2022) can accomplish image transformation via controlling the cross-modal attention maps for the global caption (Meng et al., 2022; Hertz et al., 2023; Kawar et al., 2023; Gu et al., 2023). Local image editing allows fine-grained manipulation by inpainting target regions with user-provided (Nichol et al., 2022; Avrahami et al., 2022; Wang et al., 2023b) or predicted masks (Bar-Tal et al., 2022; Couairon et al., 2023) while preserving the remain-

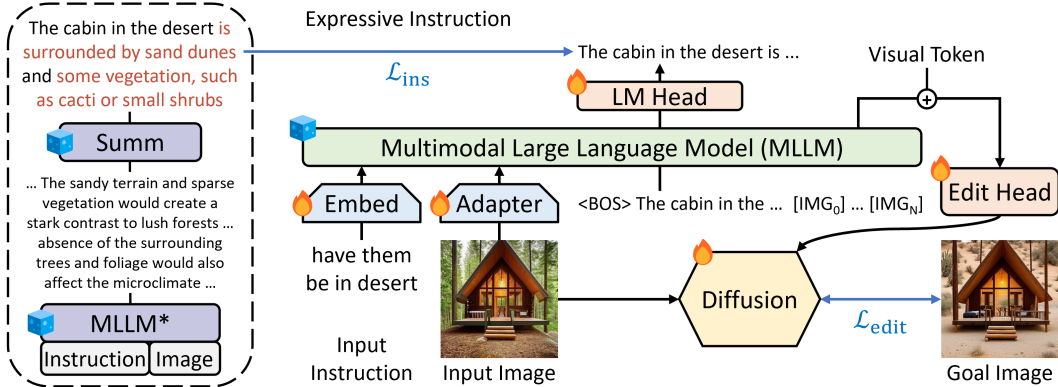


Figure 2: Overview of MLLM-Guided Image Editing (MGIE), which leverages MLLMs to enhance instruction-based image editing. MGIE learns to derive concise expressive instructions and provides explicit visual-related guidance for the intended goal. The diffusion model jointly trains and achieves image editing with the latent imagination through the edit head in an end-to-end manner. 🔥 and ❄️ show the module is trainable and frozen<sup>1</sup>, respectively.

ing areas. Different from them, instruction-based image editing accepts straight commands, such as “*add fireworks to the sky*”, which is not restricted to elaborate descriptions or regional masks. Recent methods learn from synthetic input-goal-instruction triples (Brooks et al., 2023) and with additional human feedback (Zhang et al., 2023c) to follow editing instructions. However, the frozen CLIP text encoder is pre-trained for static descriptions but not the crucial transformation in editing. Moreover, the instructions are sometimes ambiguous and imprecise for the editing goal. In this paper, we learn with multimodal large language models to perceive images along with given prompts for expressive instructions, which provides explicit yet detailed guidance, leading to superior editing performance.

**Large Language Models for Vision.** Large language models (LLMs) have demonstrated impressive capabilities for text generation and generalizability in various tasks (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023). With robust text understanding, previous works adapt LLMs for input prompts and reason downstream vision-and-language tasks (Zhang et al., 2023d; Wu et al., 2023; Lu et al., 2023; Yang et al., 2023; Chakrabarty et al., 2023). They further produce pseudocode instructions or executable programs by LLMs (Huang et al., 2022; Gupta & Kembhavi, 2023; Surís et al., 2023; Feng et al., 2023; Lian et al., 2023). Through visual feature alignment with instruction tuning, multimodal large language models (MLLMs) can perceive images and provide adequate responses (Li et al., 2023b; Zhang et al., 2023b; Liu et al., 2023; Zhu et al., 2023). Recently, studies also adopt MLLMs for generating chat-related images (Koh et al., 2023; Sun et al., 2023). However, they can only produce images from scratch, which are distinct from inputs. Our proposed MGIE is the first to leverage MLLMs and improve image editing with derived expressive instructions.

### 3 METHOD

#### 3.1 BACKGROUND: MULTIMODAL LARGE LANGUAGE MODELS (MLLMs)

Large language models (LLMs) have shown impressive capabilities for natural language generation. Multimodal large language models (MLLMs) empower LLMs to perceive images and provide reasonable responses. Initialized from a pre-trained LLM, the MLLM contains a visual encoder (*e.g.*, CLIP-L (Radford et al., 2021)) to extract the visual features  $f$ , and an adapter  $\mathcal{W}$  to project  $f$  into the language modality. We follow the training of LLaVA (Liu et al., 2023), which is summarized as:

$$\begin{aligned}
 \mathcal{C} &= \{x_1, x_2, \dots, x_l\}, \\
 f &= \text{Enc}_{\text{vis}}(\mathcal{V}), \\
 x_t &= \text{MLLM}(\{x_1, \dots, x_{t-1}\} \mid \mathcal{W}(f)),
 \end{aligned} \tag{1}$$

<sup>1</sup>We adopt Flan-T5-XXL (Chung et al., 2022), which has been specifically fine-tuned for summarization, as our summarization model for the original MLLM (MLLM\*).

where  $l$  is the length of the word token in  $\mathcal{C}$ .  $\mathcal{C}$  can be the image caption (Features Alignment) or the multimodal instruction-following data (Instruction Tuning). The MLLM follows the standard autoregressive training for the next token prediction and then can serve as a visual assistant for various tasks such as visual question answering and complex reasoning. Although the MLLM is capable of visual perception via the above training, its output is still limited to text.

### 3.2 MLLM-GUIDED IMAGE EDITING (MGIE)

As illustrated in Fig. 2, we propose MLLM-Guided Image Editing (MGIE) to edit an input image  $\mathcal{V}$  into a goal image  $\mathcal{O}$ , by a given instruction  $\mathcal{X}$ . To handle imprecise instructions, MGIE contains the MLLM and learns to derive explicit yet concise expressive instructions  $\mathcal{E}$ . To bridge the language and visual modality, we add special [IMG] tokens after  $\mathcal{E}$  and adopt the edit head  $\mathcal{T}$  to transform them. They serve as the latent visual imagination from the MLLM and guide our diffusion model  $\mathcal{F}$  to achieve the intended editing goal. MGIE is then able to comprehend ambiguous commands with visual-related perception for reasonable image editing.

**Concise Expressive Instruction.** From features alignment and instruction tuning, the MLLM can offer visual-related responses with its cross-modal perception. For image editing, we use this prompt “*what will this image be like if [instruction]*” as the language input with the image and derive a detailed explanation of the editing command. However, those explanations are always too lengthy and involve redundant descriptions, which even mislead the intention. To obtain succinct narrations, we apply a pre-trained summarizer<sup>1</sup> and make the MLLM learn to generate the summarized outputs. We treat this explicit yet concise guidance as expressive instruction  $\mathcal{E}$ :

$$\begin{aligned}\mathcal{E} &= \text{Summ}(\text{MLLM}^*(\text{[prompt, } \mathcal{X}] \mid \mathcal{W}(f))) \\ &= \{w_1, w_2, \dots, w_l\}, \\ w'_t &= \text{MLLM}(\{w_1, \dots, w_{t-1}\} \mid \mathcal{W}(f)), \\ \mathcal{L}_{\text{ins}} &= \sum_{t=1}^l \text{CELoss}(w'_t, w_t),\end{aligned}\tag{2}$$

where we apply the cross-entropy loss (CELoss) to train the MLLM via teacher forcing.  $\mathcal{E}$  can provide a more concrete idea than  $\mathcal{X}$  such as linking “*dessert*” with “*sand dunes*” and “*cacti or small shrubs*”, which mitigates the comprehension gap for reasonable image editing. This strategy further enhances our efficiency. During inference, the trained MGIE straightforwardly derives concise  $\mathcal{E}$  instead of rolling out lengthy narrations (22.7 vs. 64.5 tokens) and relying on external summarization. MGIE now can acquire a visual imagination of the editing intention but is confined to the language modality. To bridge the gap, we append  $N$  visual tokens [IMG] after  $\mathcal{E}$ , where their word embeddings are trainable, and the MLLM also learns to generate them through its language modeling (LM) head. Inspired by GILL (Koh et al., 2023), the visual tokens are treated as visual-related instruction understanding in  $\mathcal{E}$  and establish a connection between the language and vision modalities.

**Image Editing via Latent Imagination.** We adopt the edit head  $\mathcal{T}$  to transform [IMG] into actual visual guidance.  $\mathcal{T}$  is a sequence-to-sequence model, which maps the sequential visual tokens from the MLLM to the semantically meaningful latent  $\mathcal{U} = \{u_1, u_2, \dots, u_L\}$  as the editing guidance:

$$u_t = \mathcal{T}(\{u_1, \dots, u_{t-1}\} \mid \{e_{\text{[IMG]}} + h_{\text{[IMG]}}\}),\tag{3}$$

where  $e$  is the word embedding and  $h$  is the hidden state (from the last layer of MLLM before the LM head) of [IMG]. Specifically, the transformation over  $e$  can be treated as a general representation in the visual modality, and  $h$  is an instance-aware visual imagination for such editing intention. Our  $\mathcal{T}$  is similar to GILL and BLIP-2 (Li et al., 2023b;a) for extracting visual features.

To guide image editing with the visual imagination  $\mathcal{U}$ , we consider a latent diffusion model  $\mathcal{F}$  (Romach et al., 2022), which includes the variational autoencoder (VAE) and addresses denoising diffusion in the latent space. Our goal of  $\mathcal{F}$  is to generate the latent goal  $o = \text{Enc}_{\text{VAE}}(\mathcal{O})$  from preserving the latent input  $v = \text{Enc}_{\text{VAE}}(\mathcal{V})$  and following the editing guidance  $\{u\}$ . The diffusion process keeps adding noises to  $o$  as  $z_t$ , where the noise level is increasing over timesteps  $t$ . We then learn the UNet  $\epsilon_\theta$  to predict the added noise (Ho et al., 2020). As LDM, we inject the visual imagination into  $\epsilon_\theta$  via the cross-attention layer  $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{\text{dim}}}) \cdot V$  with

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), K = W_K^{(i)} \cdot \{u\}, V = W_V^{(i)} \cdot \{u\},\tag{4}$$

Method	EVR			GIER			MA5k			MagicBrush			
	L1↓	DINO↑	CVS↑	L1↓	SSIM↑	CVS↑	L1↓	SSIM↑	LPIPS↓	L1↓	DINO↑	CVS↑	CTS↑
InsPix2Pix	0.189	67.82	81.38	0.144	<u>57.51</u>	86.63	0.176	58.92	0.359	0.101	71.46	85.22	29.34
LGIE	<b>0.159</b>	<u>69.71</u>	<b>82.04</b>	0.152	56.86	<u>86.99</u>	<u>0.144</u>	<u>64.60</u>	<u>0.327</u>	<u>0.084</u>	<u>80.90</u>	<u>88.87</u>	<u>30.10</u>
MGIE	<u>0.163</u>	<b>71.49</b>	<u>81.73</u>	<b>0.135</b>	<b>59.24</b>	<b>88.59</b>	<b>0.133</b>	<b>66.25</b>	<b>0.298</b>	<b>0.082</b>	<b>82.22</b>	<b>91.14</b>	<b>30.40</b>

Table 1: **Zero-shot editing results.** All models are only pre-trained on IPr2Pr (Brooks et al., 2023).

Method	EVR			GIER			MA5k			MagicBrush			
	L1↓	DINO↑	CVS↑	L1↓	SSIM↑	CVS↑	L1↓	SSIM↑	LPIPS↓	L1↓	DINO↑	CVS↑	CTS↑
InsPix2Pix	0.166	70.79	82.76	0.111	64.86	<u>91.49</u>	0.122	67.12	0.267	0.063	87.99	93.83	30.93
LGIE	<u>0.147</u>	<u>74.71</u>	<u>85.06</u>	<b>0.104</b>	<u>65.30</u>	90.61	<u>0.094</u>	<u>71.47</u>	<u>0.246</u>	<u>0.058</u>	<u>88.09</u>	<u>93.57</u>	<u>31.33</u>
MGIE	<b>0.146</b>	<b>75.65</b>	<b>85.28</b>	<u>0.105</u>	<b>68.68</b>	<b>92.42</b>	<b>0.082</b>	<b>72.91</b>	<b>0.235</b>	<b>0.057</b>	<b>90.65</b>	<b>95.28</b>	<b>31.73</b>

Table 2: **Fine-tuned editing results.** All models are further fine-tuned and adapted to each dataset.

where  $\varphi$  is the flattened operation,  $W_Q^{(i)}$ ,  $W_K^{(i)}$ , and  $W_V^{(i)}$  are learnable attention matrices. Following InsPix2Pix, we also concatenate  $v$  with  $z_t$ . In this way, our  $\mathcal{F}$  can condition both  $\mathcal{V}$  and  $\mathcal{U}$  to perform image editing. We take classifier-free guidance (Ho & Salimans, 2021), and the score estimation  $s_\theta$  is extrapolated to keep away from the unconditional  $\emptyset$ , where the editing loss  $\mathcal{L}_{\text{edit}}$  is calculated as:

$$\begin{aligned}
 s_\theta(z_t, v, \{u\}) &= s_\theta(z_t, \emptyset, \emptyset) \\
 &\quad + \alpha_{\mathcal{V}} \cdot (s_\theta(z_t, v, \emptyset) - s_\theta(z_t, \emptyset, \emptyset)) \\
 &\quad + \alpha_{\mathcal{X}} \cdot (s_\theta(z_t, v, \{u\}) - s_\theta(z_t, v, \emptyset)), \\
 \mathcal{L}_{\text{edit}} &= \mathbb{E}_{o, v, \{u\}, \epsilon \sim \mathcal{N}(0, 1), t} [\|\epsilon - \epsilon_\theta(z_t, t, v, \{u\})\|_2^2],
 \end{aligned} \tag{5}$$

where  $\alpha_{\mathcal{V}}$  and  $\alpha_{\mathcal{X}}$  are the weights of the guidance scale for the image and the instruction. Similar to InsPix2Pix, we randomly make  $v = \emptyset$ ,  $\{u\} = \emptyset$ , or both  $= \emptyset$  for 5% of data during training. After we have the generated latent  $o'$  through the denoising process by  $\epsilon_\theta$ , we can obtain the editing result  $O' = \text{Dec}_{\text{VAE}}(o')$ . During inference, we use  $\alpha_{\mathcal{V}} = 1.5$  and  $\alpha_{\mathcal{X}} = 7.5$ .

### 3.3 LEARNING OF MGIE

Algo. 1 presents the learning process of the proposed MGIE. The MLLM learns to derive concise  $\mathcal{E}$  via the instruction loss  $\mathcal{L}_{\text{ins}}$ . With the latent imagination from  $[\text{IMG}]$ ,  $\mathcal{T}$  transforms their modality and guides  $\mathcal{F}$  to synthesize the resulting image. The editing loss  $\mathcal{L}_{\text{edit}}$  is applied for diffusion training. Most weights can be frozen (self-attention blocks inside the MLLM), leading to parameter-efficient end-to-end training. Overall optimization of  $\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{ins}} + 0.5 \cdot \mathcal{L}_{\text{edit}}$  can be:

$$\min_{\text{MLLM}, \mathcal{W}, \mathcal{T}, \mathcal{F}} \mathcal{L}_{\text{all}}. \tag{6}$$

#### Algorithm 1 MLLM-Guided Image Editing

```

1: while TRAIN_MGIE do
2:    $\mathcal{V}, \mathcal{X}, \mathcal{O} \leftarrow$  input/instruction/goal triple
3:    $\{w\} \leftarrow$  summarized explanation
4:    $\{w'\} = \text{MLLM}(\mathcal{V} | \mathcal{X})$ 
5:    $\mathcal{L}_{\text{ins}} \leftarrow$  instruction loss ▷ Eq. 2
6:    $\mathcal{U} = \mathcal{T}(\{[\text{IMG}]\})$ 
7:    $\mathcal{O}' = \mathcal{F}(\mathcal{V}, \mathcal{U})$ 
8:    $\mathcal{L}_{\text{edit}} \leftarrow$  editing loss ▷ Eq. 5
9:    $\mathcal{L}_{\text{all}} \leftarrow$  overall training loss
10: end while

```

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

**Datasets and Evaluation Metrics.** We use **IPr2Pr** (Brooks et al., 2023) as our pre-training data. It contains 1M CLIP-filtered data, where instructions are extracted by GPT-3 (Brown et al., 2020), and images are synthesized by Prompt-to-Prompt (Hertz et al., 2023). For a comprehensive evaluation, we consider various editing aspects. **EVR** (Tan et al., 2019) collects 5.7K triples from PhotoshopRequest. We treat the standard pixel difference (L1) and visual feature similarity from DINO (Caron et al., 2021) or the CLIP visual encoder (CVS) between generated images and ground-truth goals as the evaluation metrics. **GIER** (Shi et al., 2020) crawls a larger-scale 29.9K triples also from online forums. Since there are more examples about global optimization, we apply L1, CVS, and Structural

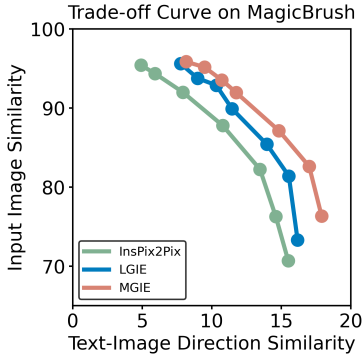


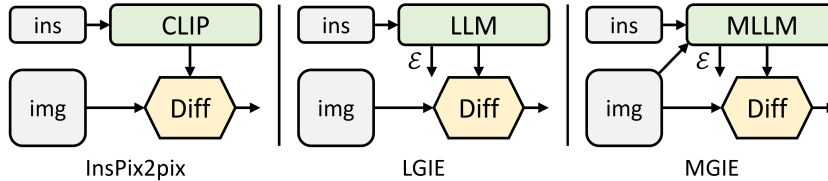
Figure 3: **Trade-off curve for image editing.** We set  $\alpha_X$  as 7.5 and vary  $\alpha_Y$  in  $[1.0, 2.2]$ . For both edit (X-axis) and input consistency (Y-axis), higher is better.

Arch. Method	MA5k			MagicBrush			
	L1↓	SSIM↑	LPIPS↓	L1↓	DINO↑	CVS↑	CTS↑
InsPix2Pix	0.176	<b>58.92</b>	<b>0.359</b>	<b>0.101</b>	71.46	85.22	29.34
FZ LGIE	0.178	57.26	0.372	0.133	67.53	82.49	28.79
FZ MGIE	<b>0.163</b>	<u>57.54</u>	0.366	<u>0.128</u>	<b>71.65</b>	<b>86.00</b>	<b>29.43</b>
FT LGIE	0.166	60.11	0.357	0.124	71.04	85.47	29.37
FT MGIE	<b>0.163</b>	<b>61.38</b>	<b>0.348</b>	<b>0.101</b>	<b>74.79</b>	<b>87.12</b>	<b>29.68</b>
E2E LGIE	0.144	64.60	0.327	0.084	80.90	88.87	30.10
E2E MGIE	<b>0.133</b>	<b>66.25</b>	<b>0.298</b>	<b>0.082</b>	<b>82.22</b>	<b>91.14</b>	<b>30.40</b>

Table 3: **Ablation study.** We attempt FZ, FT, or E2E to utilize expressive instructions. **FZ** directly treats expressive instructions as the inputs to frozen InsPix2Pix. **FT** further fine-tunes InsPix2Pix and makes it adapt to expressive instructions. Our **E2E** learns expressive instructions along with the MLLM and trains the diffusion model in an end-to-end manner.

Similarity Index (SSIM). **MA5k** (Shi et al., 2022) consists of 24.8K triples and aims at changing the contrast, brightness, or saturation of a whole photo. We leverage L1, SSIM, and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) as the photo difference<sup>2</sup>. **MagicBrush** (Zhang et al., 2023a) annotates 10.5K triples. We follow them to use L1, DINO, CVS, and text-visual feature similarity (CTS) (Hessel et al., 2021) between goal captions and resulting images. We treat the same training/validation/testing split as the original settings. Without specific mention, all evaluations are averaged from 5 random seeds in a zero-shot manner, where models are only trained on IPr2Pr.

**Baselines.** We treat InsPix2Pix (Brooks et al., 2023), built upon the CLIP text encoder with a diffusion model for instruction-based image editing, as our baseline. We consider a similar LLM-guided image editing (LGIE) model, where LLaMA-7B (Touvron et al., 2023) is adopted for expressive instructions  $\mathcal{E}$  from instruction-only inputs but without visual perception.



**Implementation Details.** The MLLM and diffusion model  $\mathcal{F}$  are initialized from LLaVA-7B (Liu et al., 2023) and StableDiffusion-v1.5 (Rombach et al., 2022). We jointly update both for the image editing task. Note that only word embeddings and LM head in the MLLM are trainable. Following GILL (Koh et al., 2023), we use  $N=8$  visual tokens. The edit head  $\mathcal{T}$  is a 4-layer Transformer, which transforms language features into editing guidance. We adopt AdamW (Loshchilov & Hutter, 2019) with the batch size of 128 to optimize MGIE. The learning rates of the MLLM and  $\mathcal{F}$  are  $5e-4$  and  $1e-4$ , respectively. All experiments are conducted in PyTorch (Paszke et al., 2017) on 8 A100 GPUs.

## 4.2 QUANTITATIVE RESULTS

Table 1 shows the zero-shot editing results, where models are trained only on IPr2Pr. For EVR and GIER that involve Photoshop-style modifications, expressive instructions can reveal concrete goals instead of brief but ambiguous commands, which makes the editing results more similar to intentions (e.g., higher 82.0 CVS on EVR by LGIE and higher 59.2 SSIM on GIER by MGIE). For global photo optimization on MA5k, InsPix2Pix is hard to deal with due to the scarcity of related training triples. Though trained from the same source, LGIE and MGIE can offer detailed explanations via learning with the LLM, but LGIE is still confined to its single modality. With access to images, MGIE derives explicit instructions such as *which regions should brighten* or *what objects are more distinct*. It can bring a significant performance boost (e.g., higher 66.3 SSIM and lower 0.3 photo distance). Similar

<sup>2</sup>As there is no object alteration in MA5k, feature-based DINO and CVS cannot clearly tell the difference.

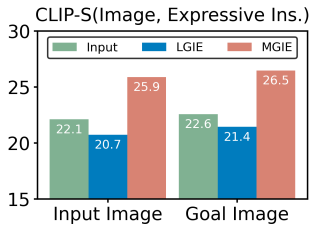


Figure 4: **CLIP-S** across images (input / goal) and expressive instructions.

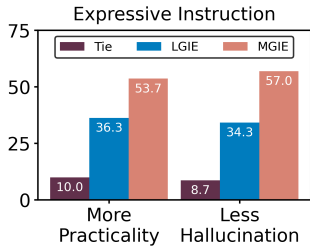


Figure 5: **Human eval** of expressive instructions quality.

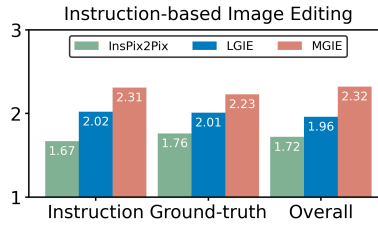


Figure 6: **Human eval** of image editing results in terms of instruction following, ground-truth relevance, and overall quality.

results are found on MagicBrush. MGIE also achieves the best performance from the precise visual imagination and modifies the designate targets as the goals (e.g., higher 82.2 DINO visual similarity and higher 30.4 CTS global caption alignment).

To investigate instruction-based image editing for the specific purpose, Table 2 fine-tunes models on each dataset. For EVR and GIER, all models obtain improvements after the adaptation to Photoshop-style editing tasks. Since fine-tuning makes expressive instructions more domain-specific as well, our MGIE increases the most via learning with domain-related guidance. This also helps our diffusion model to demonstrate concrete edited scenes from the fine-tuned MLLM, which benefits both global optimization and local modification (e.g., notably lower 0.24 LPIPS on MA5k and higher 95.3 CVS on MagicBrush). MGIE is consistently superior to LGIE in all aspects of editing since our visual-aware guidance is more aligned with the intended goal. From the above experiments, we illustrate that learning with expressive instructions can effectively enhance image editing, and visual perception plays a crucial role in deriving explicit guidance for the greatest enhancements.

**Trade-off between  $\alpha_X$  and  $\alpha_Y$ .** There are two goals in image editing: manipulate the target as the instruction and preserve the remaining as the input image. Fig. 3 plots the trade-off curves between the instruction ( $\alpha_X$ ) and input consistency ( $\alpha_Y$ ). We fix  $\alpha_X$  as 7.5 and vary  $\alpha_Y$  in [1.0, 2.2]. Higher  $\alpha_Y$  will make an editing result more similar to the input but less aligned with the instruction. X-axis calculates the CLIP directional similarity as how much the editing follows the instruction; Y-axis is the feature similarity to the input image from the CLIP visual encoder. Through concrete expressive instructions, we surpass InsPix2Pix in all settings. Our MGIE additionally results in comprehensive enhancements by learning with explicit visual-related guidance. This supports robust improvement, whether requiring higher input correlation or edit relevance.

### 4.3 ABLATION STUDY

MLLM-Guided Image Editing exhibits encouraging improvement in both zero-shot and fine-tuning scenarios. Now, we investigate different architectures to use expressive instructions. Table 3 considers **FZ**, **FT**, and our **E2E**. FZ directly uses the derived expressive instructions<sup>3</sup> as the input prompts to the frozen InsPix2Pix. In spite of having additional guidance, the scenario still differs from the trained editing instructions, which makes it difficult to deal with. LGIE even hurts the performance as it may mislead due to the shortage of visual perception. FT fine-tunes InsPixPix and adapts it to expressive instructions. These results support that image editing can benefit from explicit guidance along the derivation of instructions from the LLM/MLLM. E2E updates the editing diffusion model in conjunction with the LM, which learns to extract applicable guidance and discard irrelevant narration simultaneously through the end-to-end hidden states. In addition, our E2E can also avoid the potential error that may be propagated from the expressive instructions. Hence, we observe the most enhancements in both global optimization (MA5k) and local editing (MagicBrush). Among FZ, FT, and E2E, MGIE consistently surpasses LGIE. This indicates that expressive instructions with crucial visual perception are always advantageous across all ablation settings.

**Why MLLM Guidance is Helpful?** Fig. 4 presents the CLIP-Score between input or ground-truth goal images and expressive instructions. A higher CLIP-S to input images indicates that instructions

<sup>3</sup>During the ablation study, we employ concise summarized expressive instructions for a fair comparison.

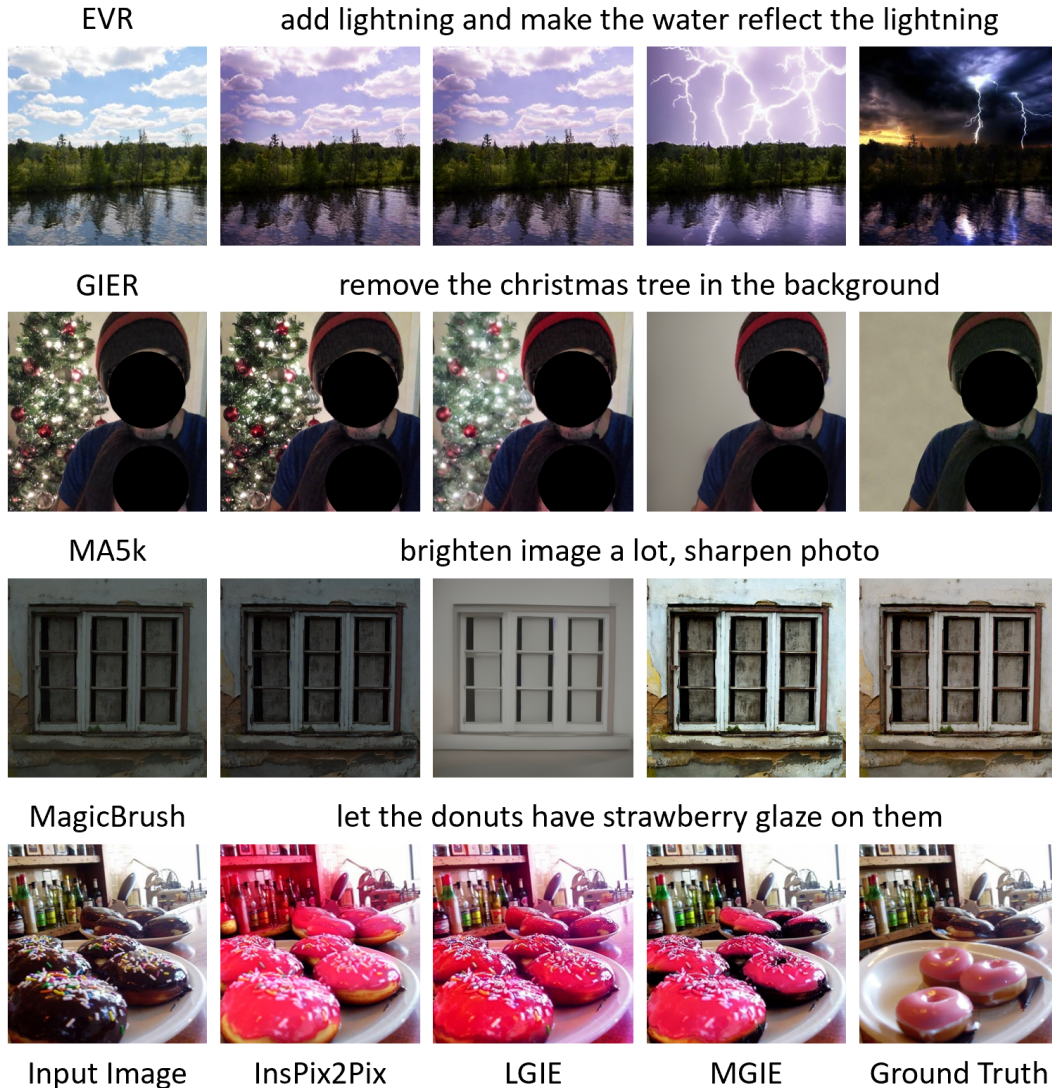


Figure 7: **Qualitative comparison** between InsPix2Pix, LGIE, and our MGIE. For the 1st example, MGIE can showcase the clear “lightning” in the sky and its reflection on the water. For the 2nd one, although LGIE accurately targets the Christmas tree, only MGIE removes it in the background. For photo optimization (the 3rd example), InsPix2Pix fails to adjust the brightness, and LGIE makes the whole photo white and obviously distinct. In contrast, MGIE follows the instruction to brighten as well as sharpen it. Moreover, in the 4th one, MGIE puts the “glaze” only on the donuts, but baselines even draw the entire image in strawberry pink.

are relevant to the editing source. Better alignment with goal images provides explicit and correlated edit guidance. Without access to visual perception, expressive instructions from LGIE are limited to general language imagination, which is not tailored to the source image. The CLIP-S are even lower than the original instructions. By contrast, MGIE is more aligned with inputs/goals, which explains why our expressive instructions are helpful. With a clear narration of the intended result, our MGIE can achieve the greatest improvements in image editing.

**Human Evaluation.** Apart from automatic metrics, we conduct a human evaluation to study generated expressive instructions and image editing results. We randomly sample 25 examples for each dataset (100 in total) and consider humans to rank across baselines and MGIE. To avoid potential ranking bias, we hire 3 annotators for each example. Fig. 5 plots the quality of generated expressive instructions. Precise guidance is informative and aligns with the intended goal (More Practicality). At the same time, it should avoid incorrect or unrelated explanations (Less Hallucination). Firstly,





Figure 8: **Qualitative comparison** of expressive instructions by LGIE and our MGIE. Due to the limitation of the single modality, LGIE can only have language-based insight but may derive irrelevant or even wrong explanations for image editing (e.g., “two people still in the foreground” for GIER). With access to images, MGIE provides explicit visual imagination after the editing such as “baby on the beach with a shark” or “bring out details of leaves and trunk”. More surprisingly, we can link “lightsaber or spaceship” from Star Wars and describe “chewing on the stick” for the dog, which is aligned with the intended goal.

over 53% support that MGIE provides more practical expressive instructions, which facilitates the image editing task with explicit guidance. Meanwhile, 57% of labelers indicate that our MGIE can prevent irrelevant descriptions from language-derived hallucinations in LGIE since it perceives the image to have a precise goal for editing. Fig. 6 compares the image editing results by InsPix2Pix, LGIE, and our MGIE in terms of instruction following, ground-truth relevance, and overall quality. The ranking score is ranging from 1 to 3, higher is better. With derived expressive instructions from the LLM or MLLM, LGIE and MGIE both outperform the baseline and perform image editing that is correlated with the instruction as well as similar to the ground-truth goal. Additionally, since our expressive instructions can provide concrete and visual-aware guidance, MGIE has the best human preference in all aspects, including the overall editing quality. These performance trends also align with automatic evaluations, which support our usage of metrics.

**Inference Efficiency.** Despite relying on MLLM to facilitate image editing, MGIE only rolls out concise expressive instructions (less than 32 tokens) and contains feasible efficiency as InsPix2Pix. Table 4 presents the inference time cost on an NVIDIA A100 GPU. For a single input, MGIE can accomplish the editing task in 10 seconds. With greater data parallelization, we take a similar amount of time (e.g., 37 seconds when batch size 8). The entire process can be affordable in one GPU (40GB). In summary, our MGIE surpasses the baseline on quality yet maintains competitive efficiency, leading to effective and practical image editing.

	BS	InsPix2Pix	MGIE
1	6.8	9.2	
4	16.5	20.6	
8	31.5	36.9	

Table 4: Time cost.

**Qualitative Comparisons.** Fig. 7 illustrates the visualized comparison on all used datasets. Fig. 8 further compares the expressive instructions by LGIE or MGIE. Our superior performance benefits from the explicit guidance of visual-related expressive instructions. Please visit our project website<sup>4</sup> for more qualitative results.

## 5 CONCLUSION

We propose MLLM-Guided Image Editing (MGIE) to enhance instruction-based image editing via learning to produce expressive instructions. Instead of brief but ambiguous guidance, MGIE derives explicit visual-aware intention and leads to reasonable image editing. We conduct extensive studies from various editing aspects and demonstrate that our MGIE effectively improves performance while maintaining competitive efficiency. We also believe the MLLM-guided framework can contribute to future vision-and-language research.

<sup>4</sup>Project website: <https://mllm-ie.github.io>

## REFERENCES

- Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended Diffusion for Text-driven Editing of Natural Images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2LIVE: Text-Driven Layered Image and Video Editing. In *European Conference on Computer Vision (ECCV)*, 2022.
- Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *International Conference on Computer Vision (ICCV)*, 2021.
- Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors. In *Annual Meetings of the Association for Computational Linguistics (ACL)*, 2023.
- Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways. In *arXiv:2204.02311*, 2022.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. In *arXiv:2210.11416*, 2022.
- Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In *International Conference on Learning Representations (ICLR)*, 2023.
- Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Casciato, and Edward Raff. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance. In *European Conference on Computer Vision (ECCV)*, 2022.

- Alaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In *International Conference on Computer Vision (ICCV)*, 2019.
- Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. In *arXiv:2305.15393*, 2023.
- Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. SSCR: Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-Driven Artistic Style Transfer. In *European Conference on Computer Vision (ECCV)*, 2022.
- Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. In *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2022.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.
- Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, Hyunjoon Jung, and Xin Eric Wang. Photoswap: Personalized Subject Swapping in Images. In *arXiv:2305.18286*, 2023.
- Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional visual reasoning without training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. In *International Conference for Learning Representations (ICLR)*, 2023.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
- Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In *International Conference on Machine Learning (ICML)*, 2022.
- 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating Images with Multimodal Language Models. In *arXiv:2305.17216*, 2023.
- Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. ManiGAN: Text-Guided Image Manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

- Dongxu Li, Junnan Li, and Steven Hoi. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. In *arXiv:2305.14720*, 2023a.
- 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 (ICML)*, 2023b.
- Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. In *arXiv:2305.13655*, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *arXiv:2304.08485*, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference for Learning Representations (ICLR)*, 2019.
- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models. In *arXiv:2304.09842*, 2023.
- 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 for Learning Representations (ICLR)*, 2022.
- Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.
- 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. In *International Conference on Machine Learning (ICML)*, 2022.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *International Conference on Learning Representations Workshop (ICLRW)*, 2017.
- Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *International Conference on Computer Vision (ICCV)*, 2021.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv:2204.06125*, 2022.
- Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning (ICML)*, 2016.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- Chitwan Sahari, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- Jing Shi, Ning Xu, Trung Bui, Franck Deroncourt, Zheng Wen, and Chenliang Xu. A Benchmark and Baseline for Language-Driven Image Editing. In *Asian Conference on Computer Vision (ACCV)*, 2020.
- Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Deroncourt, and Chenliang Xu. Learning by Planning: Language-Guided Global Image Editing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative Pretraining in Multimodality. In *arXiv:2307.05222*, 2023.
- Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *arXiv:2303.08128*, 2023.
- Hao Tan, Franck Deroncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing Visual Relationships via Language. In *Annual Meetings of the Association for Computational Linguistics (ACL)*, 2019.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. In *arXiv:2302.13971*, 2023.
- Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A Generative Image-to-text Transformer for Vision and Language. In *Transactions on Machine Learning Research (TMLR)*, 2022.
- Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. InstructEdit: Improving Automatic Masks for Diffusion-based Image Editing With User Instructions. In *arXiv:2305.18047*, 2023a.
- Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.
- Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective Image Filter: Reflecting Emotions from Text to Images. In *International Conference on Computer Vision (ICCV)*, 2023.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. In *arXiv:2303.04671*, 2023.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action. In *arXiv:2303.11381*, 2023.
- Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. In *arXiv:2306.10012*, 2023a.
- Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. In *arXiv:2303.16199*, 2023b.

- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. HIVE: Harnessing Human Feedback for Instructional Visual Editing. In *arXiv:2303.09618*, 2023c.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models. In *arXiv:2205.01068*, 2022.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. In *arXiv:2302.00923*, 2023d.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *arXiv:2304.10592*, 2023.

Method	MA5k		MagicBrush		
	SSIM↑	LPIPS↓	DINO↑	CVS↑	CTS↑
InsPix2Pix	58.92	0.359	71.46	85.22	29.34
+ Enc <sub>LLaMA</sub>	59.08	<b>0.334</b>	72.38	85.99	29.29
+ Enc <sub>LLaVA</sub>	<b>60.94</b>	<u>0.352</u>	<b>74.10</b>	<b>87.21</b>	<b>29.37</b>
HIVE	<u>65.17</u>	<u>0.302</u>	78.95	88.23	29.42
InsEdit	59.59	0.364	<b>83.26</b>	<b>91.16</b>	<u>29.80</u>
MGIE	<b>66.25</b>	<b>0.298</b>	<u>82.22</u>	<u>91.14</u>	<b>30.40</b>

Table 5: **Zero-shot editing comparison** to different instruction encoders (Enc), human feedback (HIVE), and mask-then-inpaint (InsEdit).

Method	Size	MA5k		MagicBrush		
		SSIM↑	LPIPS↓	DINO↑	CVS↑	CTS↑
InsPix2Pix		58.92	0.359	71.46	85.22	29.34
LGIE	7B	<b>64.60</b>	0.327	<b>80.90</b>	<b>88.87</b>	30.10
	13B	63.50	<b>0.308</b>	80.18	88.77	<b>30.31</b>
MGIE	6.7B	63.78	0.300	78.82	90.01	29.47
	7B	<b>66.25</b>	<u>0.298</u>	<b>82.22</b>	<u>91.14</u>	<u>30.40</u>
	13B	<u>65.91</u>	<b>0.279</b>	<u>82.15</u>	<b>91.52</b>	<b>30.75</b>

Table 6: **Zero-shot editing comparison** of different LM sizes. We treat the visual-tuned OPT-6.7B in our used MGIE-6.7B.

## A ADDITIONAL RESULTS

**Comparison to More Baselines.** InsPix2Pix (Brooks et al., 2023) applies the CLIP encoder (Radford et al., 2021), which is insufficient to capture the transformation for editing. We treat the stronger LLM/MLLM as the instruction encoder (Enc) and follow the same training strategy. Table 5 presents that adopting LLaMA (Touvron et al., 2023)/LLaVA (Liu et al., 2023) can slightly outperform CLIP, and the visual-aware encoding is also crucial in the original InsPix2Pix. However, they still contain a performance gap with our MGIE, which indicates that merely replacing the instruction encoder is not enough for their limitation. We further consider HIVE (Zhang et al., 2023c) and InsEdit (Wang et al., 2023a) for the additional baselines. HIVE collects human preference and enhances InsPix2Pix via reward feedback learning. InsEdit depends on an external segmentation model to provide the target mask and performs inpainting as the editing result. The results demonstrate that MGIE consistently surpasses HIVE without extra human feedback, which is more data-efficient for training. InsEdit is superior in local editing with its mask-then-inpaint but not in global optimization. The mask should always be the entire photo, and the inpainting is not capable of adjusting the brightness or saturation. In contrast, through learning with the derivation of the MLLM, our MGIE performs robustly in both.

increase the brightness



it should be a pizza on the tray



Input Image

HIVE

InsEdit

MGIE

Ground Truth

**Does Larger LM Help?** Our MGIE leverages LLMs/MLLMs to enhance instruction-based image editing. We investigate that if stronger LMs can bring more improvement. We consider the visual-tuned OPT-6.7B (Zhang et al., 2022) and the larger LLaVA-13B in Table 6. We also adopt LLaMA-13B for LGIE. Even though MGIE-7B has a similar size to MGIE-6.7B, its LLaVA is more powerful than OPT, which leads to an accurate visual imagination for better editing. The 13B obtains further performance gain for both LGIE and MGIE. Fig. 9 plots the CLIP-Score of expressive instructions by different sizes of MGIE. This indicates that the guidance from larger LMs is more alignment with the vision, and thus can benefit image editing more.

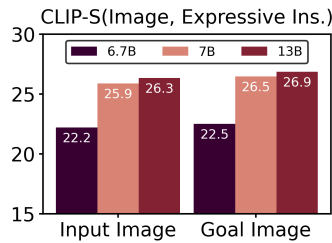


Figure 9: **CLIP-S** across images and expressive instructions by different sizes of MGIE.

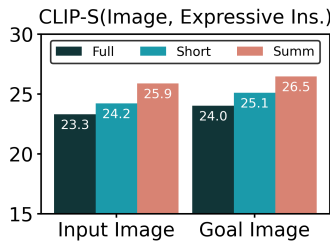


Figure 10: **CLIP-S** across images and expressive instructions (full / short / summarized).

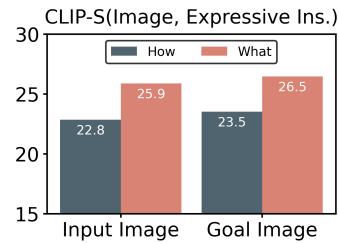
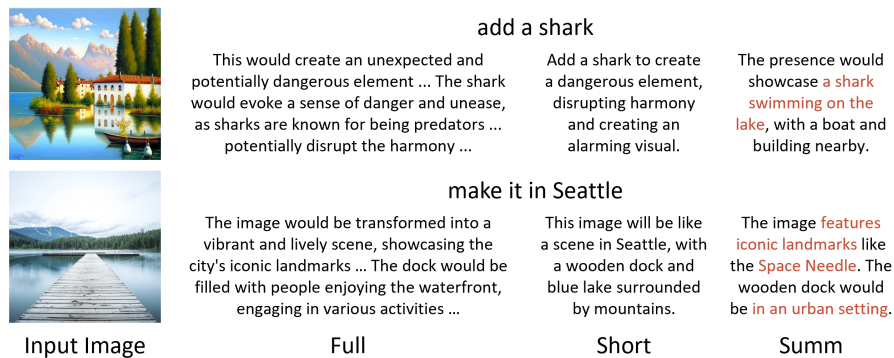


Figure 11: **CLIP-S** across images and expressive instructions by the “how” or “what” prompt.

**Learning with Summarized Expressive Instruction.** By default, MGIE learns with summarized expressive instructions for better performance and inference efficiency. We compare our form to the full description and the one making “*what will this image be like if [INS] (in short)*” as the prompt. Fig. 10 illustrates that Full is not that aligned with images due to its irrelevant narrations (e.g., “*filled with people enjoying the waterfront*”). Although Short can derive brief statements (21.1 tokens), our Summ (22.7 tokens) is still more aligned with input or goal images. In the qualitative aspect, Short’s “*create a dangerous element*” is not explicit for “*add a shark*”. Short even merely captions the photo but without “*in Seattle*”. In contrast, our Summ provides concise yet concrete guidance, such as “*a shark swimming on the lake*” or “*iconic Space Needle, urban setting*”.



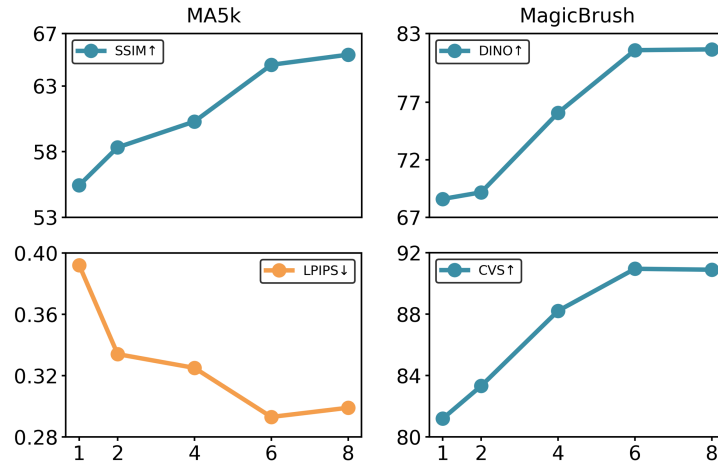
Apart from the used “*What*” prompt, we also investigate a “*How*” prompt as “*how to edit this image and [ins]*” for expressive instructions. Fig. 11 shows that our “*What*” is more aligned, which can guide image editing with more relevant visual implications, such as “*painted in hues of red, orange, and yellow*” for Autumn or “*famous landmarks as Kremlin*” for Russia. “*How*” miscomprehends the instruction as “*replace the whole garden with a beach*”. However, it should only manipulate the end of the stairs yet remain “*the stairway surrounded by lush greenery*”.



**How Many Visual Tokens do We Need?** Our editing head projects the guidance modality from the MLLM to the diffusion model. We follow GILL (Koh et al., 2023) and apply  $N=8$  visual tokens by default. Here we investigate the effectiveness of different numbers of [IMG]. The results indicate that less [IMG] makes the extracted visual imagination insufficient for effective guidance, resulting

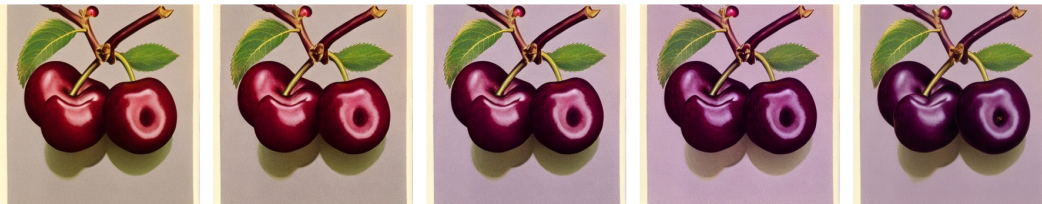


in a significant performance drop. While more [IMG] can bring further enhancements, we also find that the performance gets similar when using more than 4 [IMG].



**Qualitative Results of Different  $\alpha_\gamma$ .** MGIE adopts the weight  $\alpha_\gamma$  to adjust the level of editing. A higher  $\alpha_\gamma$  makes the editing result more similar to the input, while a lower  $\alpha_\gamma$  leads to more editing applied onto the image. Hence we can control the extent of visual transformation for both local (*e.g.*, the color of cherries) and global editing (*e.g.*, the style of the painting).

make the cherry ripe purple



the forest path to a beach



much more abstract



Input Image

$\alpha_\gamma = 2.2$

1.8

1.4

1.0

**Comparison to Description-based Baselines.** In addition to instruction-based baselines, we also consider description-based editing models. We leverage GIT (Wang et al., 2022) to caption the input image as its input description and ChatGPT to merge the edit instruction as the goal description via the prompt “Combine two sentences A: [description] and B: [instruction] into a single sentence. The output should be at most similar to sentence A”. For instance, “a girl is walking at the beach” and “give her a hat” will be transformed into “a girl with a hat is walking at the beach”. For

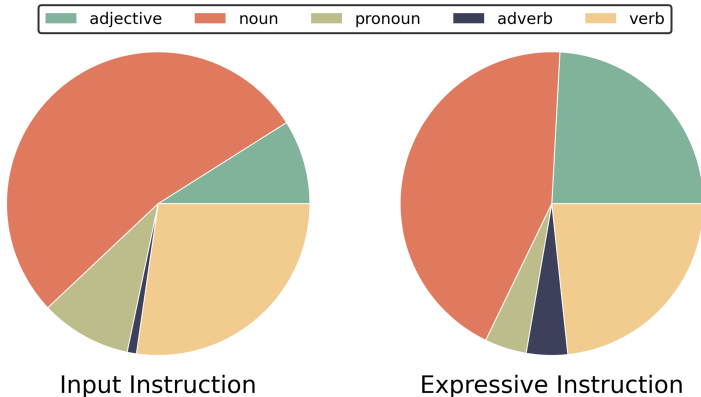
MagicBrush, we directly apply their released descriptions instead. Text2LIVE (Bar-Tal et al., 2022) and Null-Inv (Mokady et al., 2022) only yield feasible results on the traditional L1 distance but are obviously inferior to our MGIE on semantic-level evaluations (*e.g.*, lower CVS), which supports that they cannot present concrete editing results and carry out goal descriptions well. On the other hand, both count on inference optimization (CLIP alignment and DDIM inversion), which takes more than 200 seconds (*vs.* ours 9.2 seconds) for each editing task.

Method	EVR			GIER			MA5k			MagicBrush			
	L1↓	DINO↑	CVS↑	L1↓	SSIM↑	CVS↑	L1↓	SSIM↑	LPIPS↓	L1↓	DINO↑	CVS↑	CTS↑
Text2LIVE	0.169	66.19	78.22	<b>0.126</b>	<u>58.32</u>	79.32	0.165	57.62	0.342	<b>0.071</b>	<b>83.35</b>	<u>89.71</u>	23.59
Null-Inv	0.174	<u>69.24</u>	78.35	0.149	58.24	82.33	0.179	<u>61.36</u>	<u>0.335</u>	<u>0.073</u>	81.72	87.24	27.62
InsPix2Pix	0.189	<u>67.82</u>	<u>81.38</u>	0.144	57.51	<u>86.63</u>	0.176	58.92	0.359	0.101	71.46	85.22	<u>29.34</u>
MGIE	<b>0.163</b>	<b>71.49</b>	<b>81.73</b>	<u>0.135</u>	<b>59.24</b>	<b>88.59</b>	<b>0.133</b>	<b>66.25</b>	<b>0.298</b>	0.082	<u>82.22</u>	<b>91.14</b>	<b>30.40</b>

**Evaluating Image Editing via FID.** As ground-truth goal images are available, we also calculate the Fréchet inception distance (FID) for editing results under the zero-shot or fine-tuned evaluation. However, the differences are all pretty limited. Since most editing results still resemble the original input images, it is difficult for FID to determine their authenticity. These results indicate that FID is insufficient to compare the quality of image editing.

Method	Zero-shot				Fine-tuned			
	EVR	GIER	MA5k	MagicBrush	EVR	GIER	MA5k	MagicBrush
InsPix2Pix	<b>6.19</b>	<b>5.61</b>	5.91	5.69	<b>5.31</b>	<b>5.31</b>	<b>5.30</b>	5.64
LGIE	6.67	5.69	<u>5.80</u>	<b>5.31</b>	<u>5.32</u>	<u>5.42</u>	5.59	<u>5.48</u>
MGIE	<u>6.45</u>	<u>5.64</u>	<b>5.48</b>	<u>5.61</u>	5.53	5.59	<u>5.41</u>	<b>5.42</b>

**Part-of-Speech Distribution.** We investigate part-of-speech (POS) distributions<sup>5</sup> of input instructions and our derived expressive instructions. In general, input instructions involve more nouns but fewer adjectives. In contrast, our expressive instructions can portray concrete edited scenes in detail via more adjectives. The original instructions are also dominated by verbs, which are challenging to perceive. The derivation helps them to be more understandable as adverbs. Moreover, we effectively decrease the number of ambiguous pronouns. More than 68% pronouns (only 13% in our expressive instructions) are unresolvable in input instructions<sup>6</sup>, where the model can not have explicit goals.



**Unseen Editing Operation.** Since there is no removal or photo optimization in IPr2Pr, InsPix2Pix has failed due to the shortage of training examples. Our MGIE is able to handle such editing via the visual-aware derivation of MLLM. We can accurately remove “*the boy in red shirt*” or “*lighten out the yellow tone*”, which demonstrates better generalizability for unseen operations. More qualitative comparisons can be found on our project website<sup>4</sup>.

<sup>5</sup>We adopt flairNLP (<https://github.com/flairNLP/flair>) as the part-of-speech tagger.

<sup>6</sup>We apply AllenNLP (<https://github.com/allenai/allennlp>) for coreference resolution.

remove text



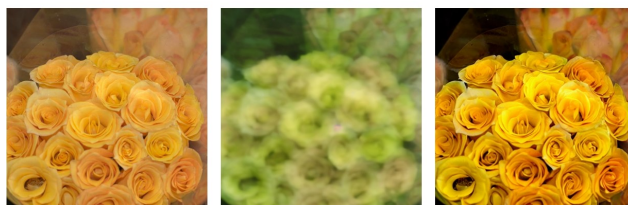
remove boy with red shirt from picture



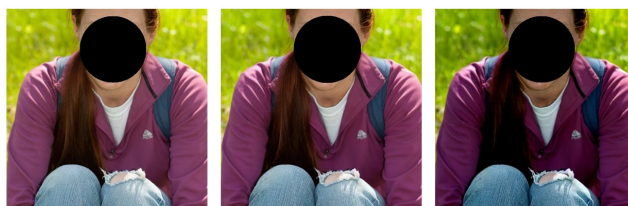
remove hot air balloon



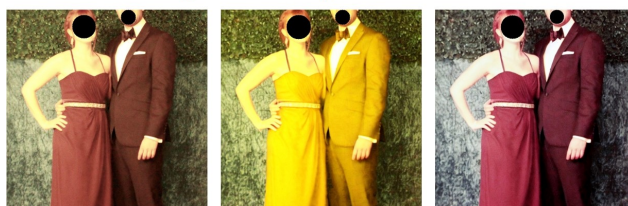
need to clarified, more focus



please reduce the brightness of the image



lighten out yellow tone



Input Image

InsPix2Pix

MGIE

**Ablation Study of Training Loss.** There are two training losses, instruction loss ( $\mathcal{L}_{ins}$ ) and editing loss ( $\mathcal{L}_{edit}$ ), in our MGIE.  $\mathcal{L}_{edit}$  is necessary for training to produce the editing result. Without  $\mathcal{L}_{ins}$ , it will derive full but lengthy guidance to lead  $\mathcal{L}_{edit}$ . However, both LGIE and MGIE drop significantly; LGIE even performs worse than the baseline. This underscores the prominence of learning concise expressive instructions, which offer succinct and relevant guidance. Besides, lengthy instructions via the MLLM will incur additional overhead (29.4 vs. ours 9.2), resulting in an inefficient inference.

Method Setting	MA5k		MagicBrush			
	SSIM $\uparrow$	LPIPS $\downarrow$	DINO $\uparrow$	CVS $\uparrow$	CTS $\uparrow$	
InsPix2Pix	58.92	0.359	71.46	85.22	29.34	
LGIE	- $\mathcal{L}_{ins}$	57.59	0.386	70.79	83.21	28.66
	+ $\mathcal{L}_{ins}$	<b>64.60</b>	<b>0.327</b>	<b>80.90</b>	<b>88.87</b>	<b>30.10</b>
MGIE	- $\mathcal{L}_{ins}$	58.18	0.365	71.50	85.19	29.11
	+ $\mathcal{L}_{ins}$	<b>66.25</b>	<b>0.298</b>	<b>82.22</b>	<b>91.14</b>	<b>30.40</b>

**Adding New Object.** MGIE also supports adding new objects that are not present in the input and placing them in reasonable positions. For instance, the “*hat*” is put on the girl’s head, and the “*river*” is added along with the grass. More surprisingly, the appended “*fireworks*” further makes the beach colorful, which drives the night scene coherent and visually appealing.



**Transferring Image Texture/Color/Emotion.** We attempt transferring visual patterns of images, also controlled through human instructions. For texture, we follow CLVA (Fu et al., 2022) and adopt the style prompt “*make the whole image as texture [ins]*”. InsPix2Pix can only do limited transfer, but MGIE shows clear visual attributes (e.g., “*orange*” or “*pinkish*”) as well as the complex “*colorful circular round*”. We perform fine-grained color manipulation, including “*glasses frame*” or “*hair*”. However, the baseline even alters the whole color. For global colorization (Chang et al., 2023), both InsPix2Pix and our MGIE cannot present appealing results, which indicates the need for fine-tuning. Transferring the emotion is more challenging as the model has to perceive the latent semantics. We are able to illustrate the visual concept of “*bright day*” or “*chaotic and confused*” as the beach in the early morning or the gloomy street at night. MGIE can also transform from the cozy snowy day into suspenseful and thrilling through “*nightmare and scared*”. Although exhibiting promising potential, it still requests more profound texture/emotion perception for each specific goal. We leave them as future research for creative visual editing (Weng et al., 2023).

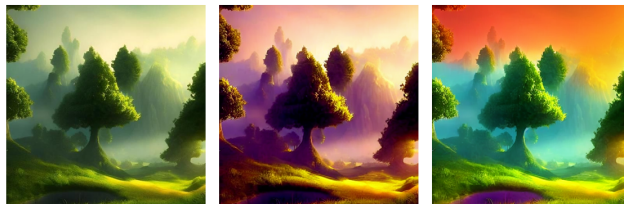
hexagonal, orange, blue, smooth white



pinkish, interlaced, cloth, like pillow cover



colorful smooth pretty circular round



Input Image

InsPix2Pix

MGIE

*color/emotion results on the next page*

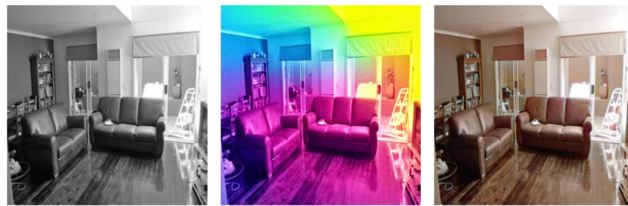
make the frame red



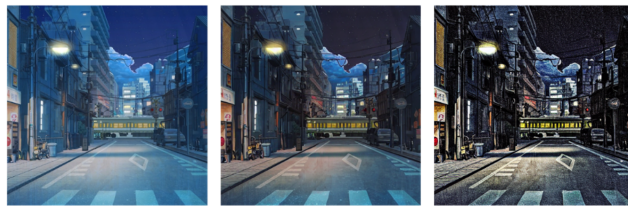
change the hair to green color



as a colorful image



feel chaotic and confused due to the tone



charmed by the beautiful bright day



out of nightmare, utterly scared and shaken



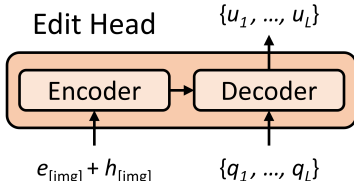
Input Image

InsPix2Pix

MGIE

## B DETAILED EXPERIMENTAL SETUP

**Edit Head to Joint the MLLM and the Diffusion Model.** These appended visual tokens [IMG] are treated as the latent imagination of the editing goal from the MLLM but in the language modality. Inspired by GILL (Koh et al., 2023), we consider an edit head  $\mathcal{T}$  to transform them into actual visual guidance.  $\mathcal{T}$  is a lightweight 4-layer Transformer, which takes word embeddings  $e$  and hidden states  $h$  of [IMG] as the input and generates the visual imagination  $\{u_1, \dots, u_L\}$ , conditioned on learnable query embeddings  $\{q_1, \dots, q_L\}$ . As our diffusion model is inherited from StableDiffusion (Rombach et al., 2022), we apply the same  $L = 77$ , and the dimension of  $u$  is 768.



**Editing Loss of the Diffusion Model.** Our diffusion model is built upon latent diffusion  $\mathcal{F}$  (Rombach et al., 2022), which operates the latent space of the variational autoencoder (VAE). For the goal image  $\mathcal{O}$ , the diffusion process keeps adding noises to the encoded  $o = \text{Enc}_{\text{VAE}}(\mathcal{O})$  and produces a noisy latent  $z_t$ . Our target is to learn the UNet  $\epsilon_\theta$  that predicts the added noise according to the input image  $v = \text{Enc}_{\text{VAE}}(\mathcal{V})$  and the visual imagination  $\{u\}$  from the MLLM. The learning objective is:

$$\mathcal{L}_{\text{edit}} = \mathbb{E}_{o,v,\{u\},\epsilon \sim \mathcal{N}(0,1),t} [\|\epsilon - \epsilon_\theta(z_t, t, v, \{u\})\|_2^2].$$

Following InsPix2Pix (Brooks et al., 2023), we leverage the classifier-free guidance (Ho & Salimans, 2021), which combines both conditional and unconditional (a fixed null value  $\emptyset$ ) denoising. During inference, we let the score estimation  $s_\theta$  extrapolate toward the conditional yet keep away from the unconditional guidance. Since there are two conditionings ( $v$  for image and  $\{u\}$  for instruction), our modified  $s_\theta$  should be:

$$\begin{aligned} s_\theta(z_t, v, \{u\}) &= s_\theta(z_t, \emptyset, \emptyset) \\ &+ \alpha_\gamma \cdot (s_\theta(z_t, v, \emptyset) - s_\theta(z_t, \emptyset, \emptyset)) \\ &+ \alpha_\chi \cdot (s_\theta(z_t, v, \{u\}) - s_\theta(z_t, v, \emptyset)), \end{aligned}$$

where we randomly set  $v = \emptyset$ ,  $\{u\} = \emptyset$ , or both  $= \emptyset$  for 5% of data during training.  $\alpha_\gamma$  and  $\alpha_\chi$  are guidance scales to control the trade-off between input image similarity and instruction alignment. By default, we use  $\alpha_\gamma = 1.5$  and  $\alpha_\chi = 7.5$ .

**Training Cost.** Our MGIE training requires 26 epochs to converge, and InsPix2Pix has 20 epochs (from their released checkpoint). Both MGIE and InsPix2Pix take a similar 1.6 hours per epoch on our node (8 NVIDIA A100 GPUs), where the overall training can be done in two days.

**Human Evaluation.** We sample 100 examples (25 for each dataset) to conduct our human evaluation. Each task is assigned 3 annotators, who rank across baselines and our MGIE, to avoid potential bias. We require workers to have a 97% approval rate and over 500 approved tasks to ensure quality. The worker is awarded \$5 for each task (5 examples) and takes 21 minutes on average to complete.

## C ETHICS DISCUSSION AND LIMITATION

In this paper, we leverage multimodal large language models (MLLMs) with the diffusion model to enhance instruction-based image editing. Even though our work benefits creative visual applications, there are still limitations that should be taken into consideration when interpreting the results. Since our MGIE is built upon pre-trained foundation models, it is possible to inherit bias from LLaVA and StableDiffusion. To mitigate this issue, we make the derived expressive instruction concise through summarization and update the MLLM together with the diffusion model. This end-to-end learning can also reduce the potential harmfulness since the hallucination from the LM will not be expressed over the editing. We can incorporate the safety checker (Rombach et al., 2022) to filter out offensive results during post-processing as the final line of defense. From the perspective of editing, there are

some challenging cases. Compositional command is hard to accomplish in a single step. Our MGIE can successfully remove the left sign but not the subsequent manipulation. In addition, the ability of language grounding (*e.g.*, only the potato should be replaced), as well as numerical perception (*e.g.*, just add to one cupcake), can be improved for more accurate targeting. We leave these directions as future research to achieve more practical and powerful instruction-based image editing.

