GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, Zhe Gan
UC Santa Barbara, Apple

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.
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 IP2Pr 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 IP2Pr 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 \(^1\) respectively.

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; Peng 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 \(W\) to project \(f\) into the language modality. We follow the training of LLaVA (Liu et al., 2023), which is summarized as:

\[
\begin{align*}
C &= \{x_1, x_2, \ldots, x_l\}, \\
f &= \text{Enc}_{\text{vis}}(V), \\
x_t &= \text{MLLM}(\{x_1, \ldots, x_{t-1}\} \mid W(f)),
\end{align*}
\]

\(^1\)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 \( C \). \( C \) 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 \( V \) into a goal image \( O \), by a given instruction \( X \). To handle imprecise instructions, MGIE contains the MLLM and learns to derive explicit yet concise expressive instructions \( E \). To bridge the language and visual modality, we add special image tokens after \( E \) and adopt the edit head \( T \) to transform them. They serve as the latent visual imagination from the MLLM and guide our diffusion model \( F \) 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 \( L \) and make the MLLM learn to generate the summarized outputs. We treat this explicit yet concise guidance as expressive instruction \( E \):

\[
E = \text{Summ}(\text{MLLM}([\text{prompt}, X] \mid \mathcal{W}(f))) = \{w_1, w_2, ..., w_l\},
\]

\[
u_t' = \text{MLLM}(\{w_1, ..., w_{t-1}\} \mid \mathcal{W}(f)),
\]

\[
L_{ins} = \sum_{t=1}^{l} \text{CELoss}(u_t', u_t),
\]

where we apply the cross-entropy loss (CELoss) to train the MLLM via teacher forcing. \( E \) can provide a more concrete idea than \( X \) 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 \( E \) 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 \([\text{IMG}]\) after \( E \), 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 \( E \) and establish a connection between the language and vision modalities.

#### Image Editing via Latent Imagination.

We adopt the edit head \( T \) to transform \([\text{IMG}]\) into actual visual guidance. \( T \) is a sequence-to-sequence model, which maps the sequential visual tokens from the MLLM to the semantically meaningful latent \( \mathcal{U} = \{u_1, u_2, ..., u_L\} \) as the editing guidance:

\[
u_t = T(\{u_1, ..., u_{t-1}\} \mid \{e_{[\text{IMG}]} + h_{[\text{IMG}]}\}),
\]

where \( e \) is the word embedding and \( h \) is the hidden state (from the last layer of MLLM before the LM head) of \([\text{IMG}]\). 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 \( T \) is similar to GILL and BLIP-2 (Li et al., 2023b) for extracting visual features.

To guide image editing with the visual imagination \( \mathcal{U} \), we consider a latent diffusion model \( F \) (Rombach et al., 2022), which includes the variational autoencoder (VAE) and addresses denoising diffusion in the latent space. Our goal of \( F \) is to generate the latent goal \( o = \text{Enc}_{\text{VAE}}(O) \) from preserving the latent input \( v = \text{Enc}_{\text{VAE}}(V) \) 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 Attention(\( Q, K, V \)) = softmax(\( QK^T / \sqrt{\text{dim}} \)) \( V \) with

\[
Q = W_Q^{(i)} \cdot \Phi(z_t), K = W_K^{(i)} \cdot \{u\}, V = W_V^{(i)} \cdot \{u\},
\]

Published as a conference paper at ICLR 2024
### 3.3 Learning of MGIE

Algorithm 1 presents the learning process of the proposed MGIE. The MLLM learns to derive concise \( E \) via the instruction loss \( L_{\text{ins}} \). With the latent imagination from \( [\text{IMG}] \), \( T \) transforms their modality and guides \( F \) to synthesize the resulting image. The editing loss \( L_{\text{edit}} \) 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 \( L_{\text{all}} = L_{\text{ins}} + 0.5 \cdot L_{\text{edit}} \) can be:

\[
\min_{\text{MLLM}, W, T, F} \mathcal{L}_{\text{all}}.
\]

### 4 Experiments

#### 4.1 Experimental Setup

Datasets and Evaluation Metrics. We use IPr2Pr \cite{Brooks2023} as our pre-training data. It contains 1M CLIP-filtered data, where instructions are extracted by GPT-3 \cite{Brown2020}, and images are synthesized by Prompt-to-Prompt \cite{Hertz2023}. For a comprehensive evaluation, we consider various editing aspects. EVR \cite{Tan2019} collects 5.7K triples from PhotoshopRequest. We treat the standard pixel difference (L1) and visual feature similarity from DINO \cite{Caron2021} or the CLIP visual encoder (CVS) between generated images and ground-truth goals as the evaluation metrics. GIER \cite{Shi2020} 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.

Similarity Index (SSIM). **MASk** (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. **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 $E$ from instruction-only inputs but without visual perception.

**Implementation Details.** The MLLM and diffusion model $F$ 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 $T$ 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 $F$ are 5e-4 and 1e-4, respectively. All experiments are conducted in PyTorch (Paszke et al., 2017) on 8 A100 GPUs.

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.

<table>
<thead>
<tr>
<th>Arch. Method</th>
<th>MASk</th>
<th>MagicBrush</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1↓</td>
<td>SSIM↑</td>
</tr>
<tr>
<td>InsPix2Pix</td>
<td>0.176</td>
<td>58.92</td>
</tr>
<tr>
<td>FZ</td>
<td>0.178</td>
<td>57.26</td>
</tr>
<tr>
<td>LGIE</td>
<td>0.163</td>
<td>57.54</td>
</tr>
<tr>
<td>MGIE</td>
<td>0.166</td>
<td>60.11</td>
</tr>
<tr>
<td>FT</td>
<td>0.163</td>
<td>61.38</td>
</tr>
<tr>
<td>E2E</td>
<td>0.144</td>
<td>64.60</td>
</tr>
<tr>
<td>LGIE</td>
<td>0.133</td>
<td>66.25</td>
</tr>
</tbody>
</table>

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 MASk, 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

As there is no object alteration in MASk, feature-based DINO and CVS cannot clearly tell the difference.
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 $\alpha_V$. 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 ($\alpha_V$). We fix $\alpha_X$ as 7.5 and vary $\alpha_V$ in $[0, 2]$. Higher $\alpha_V$ 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 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

\[\text{Figure 4: CLIP-S across images (input / goal) and expressive instructions.}\]

\[\text{Figure 5: Human eval of expressive instructions quality.}\]

\[\text{Figure 6: Human eval of image editing results in terms of instruction following, ground-truth relevance, and overall quality.}\]
Published as a conference paper at ICLR 2024

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.

Input Image | InsPix2Pix | LGIE | MGIE | Ground Truth
--- | --- | --- | --- | ---
EVR | add lightning and make the water reflect the lightning
GIER | remove the christmas tree in the background
MA5k | brighten image a lot, sharpen photo
MagicBrush | let the donuts have strawberry glaze on them

Figure 7: 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.

**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 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.

4Project website: [https://mllm-ie.github.io](https://mllm-ie.github.io)
REFERENCES


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

<table>
<thead>
<tr>
<th>Method</th>
<th>MA5k</th>
<th>MagicBrush</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM↑</td>
<td>LPIPS↓</td>
</tr>
<tr>
<td>InsPix2Pix</td>
<td>58.92</td>
<td>0.359</td>
</tr>
<tr>
<td>+ EncLLaMA</td>
<td>59.08</td>
<td><strong>0.334</strong></td>
</tr>
<tr>
<td>+ EncLLaVA</td>
<td><strong>60.94</strong></td>
<td><strong>0.352</strong></td>
</tr>
<tr>
<td>HIVE</td>
<td>65.17</td>
<td>0.302</td>
</tr>
<tr>
<td>InsEdit</td>
<td>59.59</td>
<td>0.364</td>
</tr>
<tr>
<td>MGIE</td>
<td><strong>66.25</strong></td>
<td><strong>0.298</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method Size</th>
<th>MA5k</th>
<th>MagicBrush</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM↑</td>
<td>LPIPS↓</td>
</tr>
<tr>
<td>InsPix2Pix</td>
<td>58.92</td>
<td>0.359</td>
</tr>
<tr>
<td>LGM</td>
<td>64.40</td>
<td><strong>0.327</strong></td>
</tr>
<tr>
<td>LGM-13B</td>
<td>63.50</td>
<td><strong>0.308</strong></td>
</tr>
<tr>
<td>HIVE</td>
<td>65.17</td>
<td>0.302</td>
</tr>
<tr>
<td>LGM</td>
<td>66.25</td>
<td>0.298</td>
</tr>
<tr>
<td>LGM-13B</td>
<td>65.91</td>
<td><strong>0.279</strong></td>
</tr>
</tbody>
</table>

### 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.

**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 LGM. 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 LGM 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
Published as a conference paper at ICLR 2024

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 $\alpha_V$. MGIE adopts the weight $\alpha_V$ to adjust the level of editing. A higher $\alpha_V$ makes the editing result more similar to the input, while a lower $\alpha_V$ 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).

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 \cite{bar-tal2022text} and Null-Inv \cite{mokady2022null} 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>EVR</th>
<th>GIER</th>
<th>MA5k</th>
<th>MagicBrush</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1↓</td>
<td>DINOT CVST</td>
<td>L1↑</td>
<td>SSIM↑ CVST</td>
</tr>
<tr>
<td>Text2LIVE</td>
<td>0.169</td>
<td>66.19</td>
<td>78.22</td>
<td>0.126</td>
</tr>
<tr>
<td>Null-Inv</td>
<td>0.174</td>
<td>69.24</td>
<td>78.35</td>
<td>0.149</td>
</tr>
<tr>
<td>InsPix2Pix</td>
<td>0.189</td>
<td>67.82</td>
<td>81.38</td>
<td>0.144</td>
</tr>
<tr>
<td>MGIE</td>
<td><strong>0.163</strong></td>
<td><strong>71.49</strong></td>
<td><strong>81.73</strong></td>
<td>0.135</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Zero-shot</th>
<th>Fine-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EVR</td>
<td>GIER</td>
</tr>
<tr>
<td>InsPix2Pix</td>
<td>6.19</td>
<td>5.61</td>
</tr>
<tr>
<td>LGIE</td>
<td>6.67</td>
<td>5.69</td>
</tr>
<tr>
<td>MGIE</td>
<td>6.45</td>
<td>5.64</td>
</tr>
</tbody>
</table>

Part-of-Speech Distribution. We investigate part-of-speech (POS) distributions\(^5\) 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\(^6\), 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\(^4\).

\(^5\)We adopt flairNLP \(\text{https://github.com/flairNLP/flair}\) as the part-of-speech tagger.

\(^6\)We apply AllenNLP \(\text{https://github.com/allenai/allennlp}\) 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

<table>
<thead>
<tr>
<th>Input Image</th>
<th>InsPix2Pix</th>
<th>MGIE</th>
</tr>
</thead>
</table>
Ablation Study of Training Loss. There are two training losses, instruction loss ($L_{\text{ins}}$) and editing loss ($L_{\text{edit}}$), in our MGIE. $L_{\text{edit}}$ is necessary for training to produce the editing result. Without $L_{\text{ins}}$, it will derive full but lengthy guidance to lead $L_{\text{edit}}$. 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.

<table>
<thead>
<tr>
<th>Method Setting</th>
<th>MA5k</th>
<th>MagicBrush</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM↑ LPIPS↓</td>
<td>DINO↑ CVS↑ CTS↑</td>
</tr>
<tr>
<td>InsPix2Pix</td>
<td>58.92 0.359</td>
<td>71.46 85.22 29.34</td>
</tr>
<tr>
<td>LGIE - $L_{\text{ins}}$</td>
<td>57.59 0.386</td>
<td>70.79 83.21 28.66</td>
</tr>
<tr>
<td>LGIE + $L_{\text{ins}}$</td>
<td><strong>64.60</strong> 0.327</td>
<td><strong>80.90</strong> <strong>88.87</strong> <strong>30.10</strong></td>
</tr>
<tr>
<td>MGIE - $L_{\text{ins}}$</td>
<td>58.18 0.365</td>
<td>71.50 85.19 29.11</td>
</tr>
<tr>
<td>MGIE + $L_{\text{ins}}$</td>
<td><strong>66.25</strong> <strong>0.298</strong></td>
<td><strong>82.22</strong> <strong>91.14</strong> <strong>30.40</strong></td>
</tr>
</tbody>
</table>

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**

![hexagonal, orange, blue, smooth white]

**pinkish, interlaced, cloth, like pillow cover**

![pinkish, interlaced, cloth, like pillow cover]

**colorful smooth pretty circular round**

![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 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 \( \{u_1, \ldots, u_L\} \), conditioned on learnable query embeddings \( \{q_1, \ldots, q_L\} \). 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.

![Diagram of Edit Head](image)

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 \( O \), the diffusion process keeps adding noises to the encoded \( o = \text{Enc}_{\text{VAE}}(O) \) 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 \( v = \text{Enc}_{\text{VAE}}(V) \) and the visual imagination \( \{u\} \) from the MLLM. The learning objective is:

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

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 \( \forall \)) 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:

\[
s_\theta(z_t, v, \{u\}) = s_\theta(z_t, \forall, \forall) + \alpha_\mathcal{V} \cdot (s_\theta(z_t, v, \forall) - s_\theta(z_t, \forall, \forall)) + \alpha_\mathcal{X} \cdot (s_\theta(z_t, v, \{u\}) - s_\theta(z_t, v, \forall)),
\]

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

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.