Instruction-based Image Editing with Planning, Reasoning, and Generation

Anonymous ICCV submission

Paper ID 4614

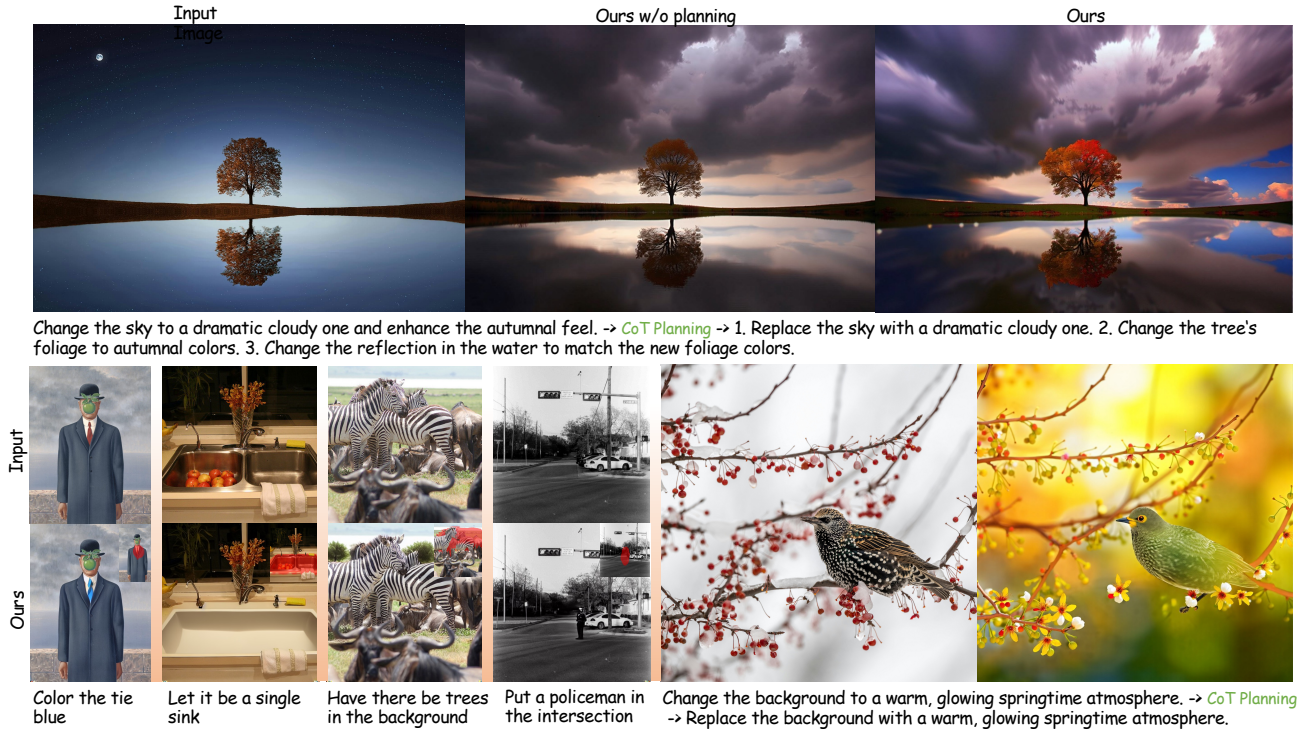

Figure 1. We propose an instruction-based editing method with Planning, Reasoning, and Generation framework, that can edit the image with human language empowered by the (multi-modal) large language model. Row 1 and Row 2 right: Our model could generate more fulfilling contents using instructions obtained by chain-of-thought; Row 2 left: Ours can further reason for the accurate editing region (shown at top right of sub-figures) based on the provided instructions.

Abstract

Editing images via instruction provides a natural way to generate interactive content, but it is a big challenge due to the higher requirement of scene understanding and generation. Prior work utilizes a chain of large language models, object segmentation models, and editing models for this task. However, the understanding models provide only single modality ability, restricting the editing quality. We aim to bridge understanding and generation via a new multi-modality model that provides the intelligent abilities to instruction-based image editing models for more complex cases. To achieve this goal, we separate the instruction editing task with the multi-modality chain of thought prompts, i.e.*, Chain-of-Thought (CoT) planning, editing region reasoning, and editing, individually. For Chain-of-Thought planning, the large language model could reason the appropriate sub-prompts considering the instruction provided and the ability of the editing network. For editing region reasoning, we train an instruction-based editing region generation network with a multi-modal large language model. Finally, for editing image generations, a hint-guided instruction-based editing network is proposed based on the sizeable text-to-image diffusion model to accept the hints for generation. Extensive experiments demonstrate that our method has competitive editing abilities on complex real-world images. Source codes will be publicly available.*

1. Introduction

Humans are familiar with guiding how to perform a task via instructions since instructions effectively encompass actions and the object that needs to be modified. Unlike other settings of language-guided image editing, such as text labels or descriptions of target images, image editing via instructions [1] allows a more user-friendly interaction with concise and accurate action guidance. This interactive approach expands our imaginations to a world in which humans use their language naturally to easily change multimedia resources or artificially generated content. In addition, instruction-based image editing can be extended with voice control in human-computer interaction scenarios and enhance the user experience in commercial products.

Previous methods [1, 14, 37] tune the text-to-image diffusion model for instruction-based editing in an end-to-end fusion. Several works [5, 13, 33] increase the editing ability of diffusion models [1] with the help of Multi-modality Large Language Models (LLMs), where they [5, 13] replace the text embedding with MLLMs directly. These methods have two drawbacks. Firstly, it increases the workload and requirements of the generation network without any additional human prior knowledge, like splitting one complex problem into several simple tasks. Secondly, the whole framework is less interpretable then the editing hints, such as sub-prompts or the editing regions, which can also be modified by users easily.

In the real world, image editing via instructions challenges us with higher requirements of understanding and reasoning ability due to their complexity. Some instructions contain abstract concepts, such as “dramatic” or “beautiful”, which can not be understood well by the text encoder only. Besides, the longer instructions contain multiple actions together. Thus, inspired by Chain-of-Thought [31], we utilize the power of LLM to make the detailed and interoperability prompt so that we can increase the generation ability. Different from the original Chain-of-Thought, which only contains the text prompt, we are focusing on the image editing tasks, we thus consider the multi-modality prompts which contain the prompt planning, the editing region generation, and finally the prompt for instruction-based editing. These multi-modal thought chains provide detailed and explainable intermediate results to divide the complex editing task into multi-run editing tasks.

Based on previous motivation, we propose a novel framework, Multimodal Chain-of-Thought Editing, consisting of an MLLM CoT Planner that generates multimodality hints for editing and a hint-guided editing network that generates the final editing results. The multimodality hints contain two parts: the specific sub-prompts and the corresponding editing regions. We use DeepSeek [4] Reasoning Model with proper promptings, such as ”Let us think step by step,” to trigger the chain of thought prompts. We aim to instantiate abstract concepts, understand the concept in the context of specific situations, or split complex tasks into simple sub-tasks. In details, we not only use the common way “Let us think step by step,” but consider the ability of editing the network in the prompt and enable the planner double-check the answers. Secondly, inspired by LISA [18], we would like the Multi-modal Language Model to reason the edited region directly, given the input image and the editing instructions, leading to a more stable and fine-grained editing quality. The edited mask, reasoned by the Multimodal Language Model, is an excellent external resource for controlling the generated results spatially. In addition, we propose a simple but effective hints-guided network by adding the latent spaces of foreground and background images spatially to the noised states in each step of the denoising process. We found out that the foreground and background images as the conditions of the diffusion models could bring effective hint control to the generated results. We also extend the framework to support classifier-free guidance on three conditions, which, according to the experiments, leads to a slight improvement.

We conduct extensive experiments and achieve state-of-the-art performance on the MagicBrush [37] dataset and the HQEdit-Abstract dataset with abstract concepts instructions, extracted from HQEdit [14]. We also apply our models to real-world cases in the open domain. Our contributions can be summarized as:

- We propose a novel framework, Multimodal Chain-of-Thought Editing, consisting of an MLLM CoT Planner that generates multimodality hints for editing and a hint-guided editing network that generates the final results.
- We propose an effective hint-guided editing framework by adding the foreground and background images as the conditions of generation models.
- We create an instruction-based image editing CoT dataset based on MagicBrush. We also conduct extensive experiments on the MagicBrush dataset and the HQEdit-Abstract dataset with state-of-the-art performance and apply our method to the real-world open-domain cases.

2. Related Work

2.1. Instruction-based Image Editing

Instruction-based image editing [1, 3, 5, 8, 10, 13, 22, 24, 30, 35, 37, 39] provide a more straightforward way for human-like image editing. InstructPix2Pix [1] is the first work to propose this setting and generate a large instruction-based dataset using the Prompt-to-prompt [11] techniques to control the consistency of the spatial structure. MagicBrush [37] proposed the instruction-based fine-grained image editing dataset, and they fine-tuned InstructPix2Pix on their dataset, leading to an improvement in editing quality. Firstly, it is helpful to utilize the masks reasoned by

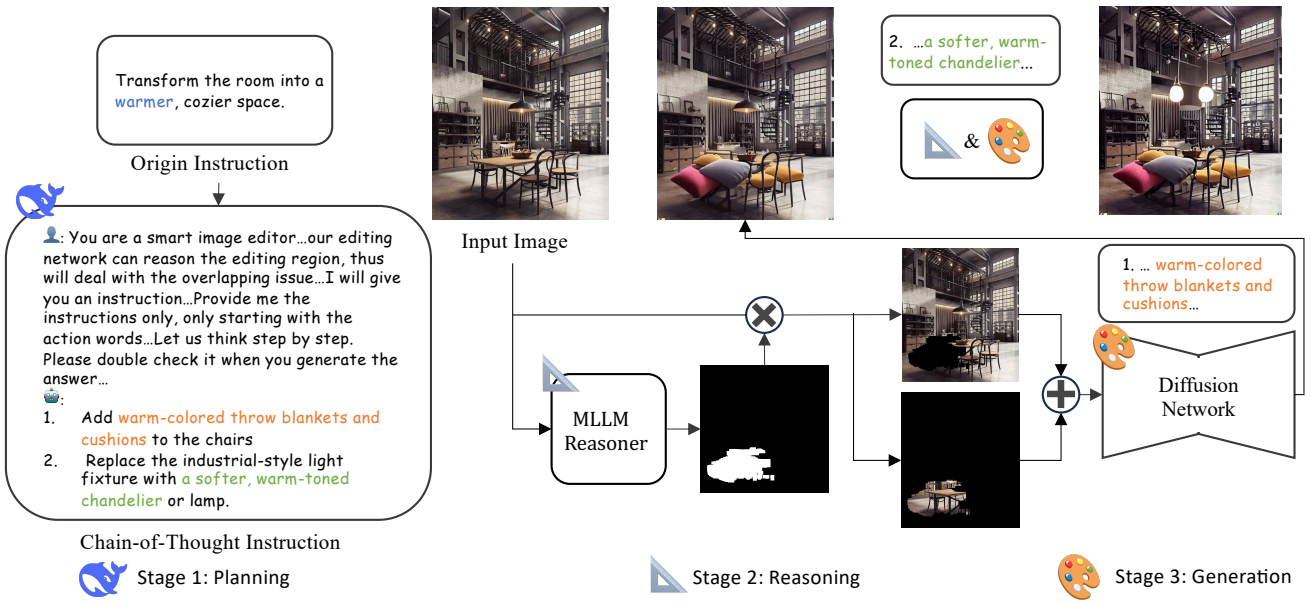

Figure 2. Our Multi-modal Chain-of-thoughts Editing framework executes image editing through three iterative stages, including planning, reasoning, and generation. In stage 1, a **C**hain-of-**T**hought Planner decomposes the user prompts to chain-structured refined editing sub-instructions; For each sub-instruction, an **M**LLM localizes target editing regions (stage 2) via cross-modal reasoning; Then, the conditional Diffusion model **E**dits the latest image (stage 3) while preserving non-target areas. The system cyclically refines outputs through location reasoning by MLLM and image generation by Diffusion model until the original plan in stage 1 is completed.

the model as additional information for instruction-based image editing. Chakrabarty *et al*. [3] tries to use ChatGPT [25] and GroundingDINO [21] to generate the mask to filter out a more high-quality dataset on InstructPix2Pix [1]. InstructEdit [30] utilizes the mask provided by chaining ChatGPT with an object-level segmentation model to obtain the editing results. Qin *et al*. [10] extracts the mask from the cross-attention map and provides an attention-guided editing framework. All of the works [3, 10, 30] fail the scenarios in which we need the editing regions instead of object-level segmentation. In addition, there are several works [5, 13] that take advantage of the strong ability of M-LLMs to solve the open-domain challenges in instruction-based image editing. To the best of our knowledge, we are the first to utilize the Multi-Modality Large Language network to reason the edited regions as a bridge of understanding and generation and thus relieve the workload of diffusion-based editing models in the instruction-based image editing problem.

2.2. Multi-Modality LLMs for Vision Tasks

Multi-modality LLMs are first inspired by GPT-4 [25], which accepts the image input to the large language models and produces corresponding text output. Based on this intuition, several methods have been proposed for understanding the scenes through neural languages [9, 19, 20, 41]. Besides language tasks, research also involves the computer vision understanding tasks via M-LLM, including the grounding information generation [32], semantic generation [18, 42], planing [6]. Our task is more related to the MLLM-based semantic mask generation. Differently, we aim to train a network that is specifically for editing region generation. Then, we can use this hint for the proposed instruction-based image editing network. Recently, several editing works have been proposed by utilizing M-LLMs [7, 13, 33], which are different from ours.

2.3. Controllable Generation in Diffusion Models

Recent works mainly utilize the priors of the diffusion model for conditional generation. ControlNet [38] and T2I-Adapter [23] learn to add additional control signal abilities (*e.g*., the human pose, the canny edge, and depth) to the stable diffusion. Despite the spatial control, another interesting domain of controllable generation is video personalization. Where we can learn to control the identity of the objects or human [36] via the plugins, such as Dreambooth [29], Custom Diffusion [16] and IP-Adapter [34], however, they only work on the pre-trained text-to-image models, which is different from our task, which needs to add controls to the instruction-based editing models.

3. Method

We aim to perform general editing on images following complex natural language instructions. Given an image x_0 and an editing instruction p, our model generates the edited image y. Different from previous globally single-stage end-

to-end framework [1] or directly fine-tune [14, 37], we introduce the MLLM Chain-of-Thought planning framework as the bridge of understanding and generation so that we can utilize the powerful reasoning ability of Multi-modality LLMs [33, 40].

3.1. Image Editing with Multi-modal Chain-of-Thought Prompts

As shown in Fig. 2, we propose a novel framework for general image editing with MLLM CoT and Conditional Editing, where the MLLM CoT Planner contains multi-modality LLMs that can parse the given instruction and the reference image to produce several sub-prompts, these sub-prompts gives detailed thought chain knowledge including the text and editing mask respect to the original image and the given prompt. Then, the framework utilizes these multi-domain prompts for generation.

Specifically, suppose we have a CoT planner, $\mathcal{C}_p(\cdot)$ for generating sub-prompts and a MLLM reasoner $\mathcal{C}_m(\cdot)$ for reasoning the editing regions, these editing hints can be obtained by:

$$\begin{aligned} C_h =& \{(\mathcal{C}_m(x_0, p_i), p_i) : p_i \in \mathcal{C}_p(p_{x_0}, p, K)\} \\ =& \{(m_i, p_i)\}_{1 \le i \le k \le K}, \end{aligned} \tag{1}$$

where p_{x_0} is the global description of the input image x_0 and $m_i = \mathcal{C}_m(x_0, p_i)$. We denote k as the number of sub-prompts decided by $\mathcal{C}_p(\cdot)$ and K as the pre-defined threshold.

After generating the multi-modality prompts from the given image, the conditional generation module is used for instruction-based editing. Suppose we have a generative model $\mathcal{G}(\cdot)$ conditioning on reasoned hints C_h and the input image x_0. The edited results y_0 is obtained iteratively:

$$y_{i+1} = \mathcal{G}(y_i, m_i, p_i);\ for\ 1 \le i \le k, \tag{2}$$

where the edited image $y = y_{k+1}$ and the starting input $y_0 = x_0$. If we assume that the image quality of y_{i+1} is not lower than that of y_i after the operation of conditional generation, the edited quality will remain the same with the input image x_0 after several iterations. However, the assumption is not valid due to the limitation of current state-of-art generation models. Therefore, we set an appropriate small value for K to limit the number of sub-prompts practically.

In detail, we use DeepSeek Reasoning Model [4] as $C_p(\cdot)$ with proper prompting to trigger the Chain-of-Thought ability. The prompting details can be found in Figure 2. We find providing the planner with the editing network ability as a prior could remove some unnecessary instructions. For example, since our editing network can reason about the editing region. Proving this information in the prompt could avoid some position adjustment instructions. At the same time, letting the planner double check the answer via adding "Please double check it when you generate the answer," can make the prompt more accurate and stable, especially for dealing with numbering cases. In addition, we train the M-LLM $C_m(\cdot)$ for reasoning the editing regions(Sec. 3.2). We also perform the learning-based editing $\mathcal{G}(\cdot)$ given the condition set (y_i, m_i, p_i) via a tuned diffusion model (Sec. 3.3).

3.2. Editing Region Reasoning

The editing region is a specific kind of mask with high correlations between the input image and the editing instructions under human opinion. We argue this region is different from the object-level segmentation and might be in detailed illustration than the object level for specific objects or a meaningless region to put something on. Thus, the current universal reasoning segmentation model, *i.e.*, LISA [18] and SEEM [42] might not work well in these cases. As shown in Figure 3, if we want to edit an image with the instruction of "Have the person jump over a tennis ball." the segmented region is an area below the legs of the person, instead of the person itself. In addition, object-level segmentation needs to segment the object precisely, while we only aim to segment an approximate editing region given an input image to accept more possibilities. Since this edited region is not straightforward, it requires the model to be more able to reason about the input instruction. Therefore, we need a stronger model that can recognize the image, accept natural language as a prompt, and, most importantly, have the reasoning ability to generate a correct editing area m_i.

Thus, inspired by the recent advantages of multi-modal language model-based image segmentation [18, 42], we repurpose the reasoning image segmentation network for our editing region generation task. In detail, we fix the parameters of the original Multi-Modal LLM [20] and train a LoRA [12] to generate the reasoning tokens for segmentation. Then, a pre-trained segmentation anything model (SAM [15]) is used to extract the visual feature and generate the reasoning mask with the help of the LLM's output tokens. In this stage, we only train the parameters in LoRA [12] and the decoder of SAM [18] inspired by LISA [18]. We utilize the standard BCE loss to train the network to predict the edit region on the training dataset of MagicBrush [37]. After training, this network can be used to infer the editing region from the image and the instruction.

3.3. Hint-guided Editing Network

After getting the condition set (y_i, m_i, p_i), we propose an efficient network structure to perform the hint-guided image editing. In detail, we utilize the structure of Stable Diffusion [28] as the network structure, where a denoising U-Net $\epsilon_\theta(\cdot)$ is trained for image editing via the paired dataset us-

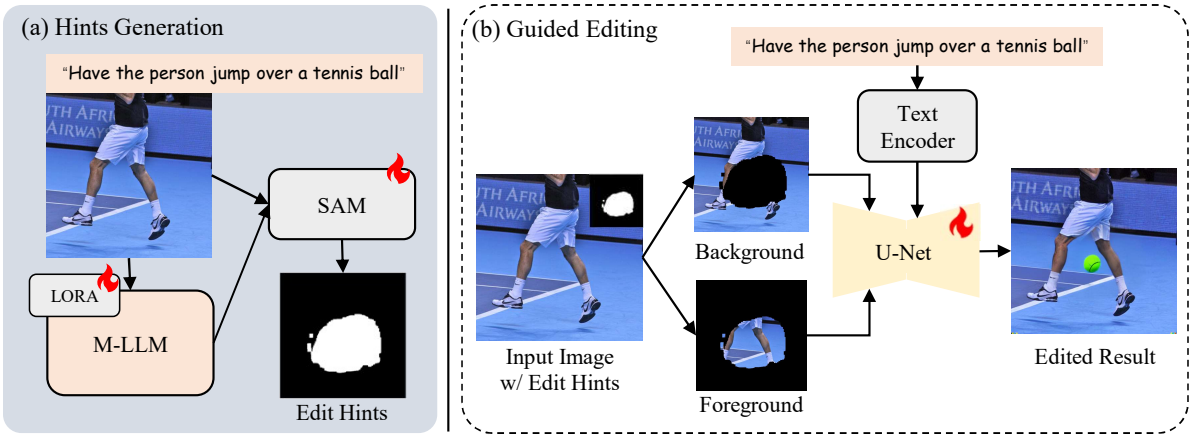

Figure 3. (a) We trained a Multi-modality LLM that generates an editing region and enables better localization given the input image and sub-prompt. (b) Given the editing region and sub-prompt reasoned by M-LLM, we further train a conditional generative diffusion model to edit the image with better locality.

ing Prompt-to-Prompt [11]. The CLIP text encoder $\varepsilon(\cdot)$ is frozen to accept the instruction-level guidance. Aiming to get better control based on the editing hints, we first compute the foreground image x_f and background image x_b with the edit region m_i by:

$$x_f = y_i \odot m_i, \tag{3}$$

$$x_b = y_i \odot (1 - m_i). \tag{4}$$

Then we concatenate the foreground image and background image as an additional spatial condition as the input of Diffusion U-Net $\epsilon_\theta(\cdot)$. We also encode x_f and x_b into the latent space using the latent encoder $\xi(\cdot)$ before feeding into the denoising network. We modify the standard diffusion loss [28] to optimize our network:

$$E_{\xi(y_i),\varepsilon(p_i),\xi(x_f),\xi(x_b),t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[||\epsilon - \epsilon_\theta(z_t, t, \varepsilon(p_i), \xi(x_f), \xi(x_b)||_2^2], \tag{5}$$

So we only need to modify the weights of the first convolution layer to fit this difference. In the training process, we use the ground truth mask as the input since it can provide more reliable guidance. In testing, we perform the editing based on the reasoning results from the hints generation network.

3.4. Classifier-free Guidance for Three Conditions

The denoising diffusion model with conditions will decrease the diversity of the generated results. Adding classifier-free guidance aims to let the model maintain the original generation ability by randomly dropping some conditions during the training process. It is even more important with the increase in the number of conditions. In our hints-guided network, the denoising network $\epsilon_\theta(z_t, x_f, x_b, p_i)$ has three conditions, which denote the foreground image x_f, the background image x_b, and the instruction p_i separately. We extend two conditions [1] to three conditions as below:

$$\begin{aligned}\epsilon_\theta(z_t, x_b, p_i, x_f) &= \epsilon_\theta(z_t, \phi, \phi, \phi) \\ &+ s_f(\epsilon_\theta(z_t, x_f, \phi, \phi) - \epsilon_\theta(z_t, \phi, \phi, \phi)) \\ &+ s_p(\epsilon_\theta(z_t, x_f, x_b, \phi) - \epsilon_\theta(z_t, x_f, \phi, \phi)) \\ &+ s_b(\epsilon_\theta(z_t, x_f, x_b, p_i) - \epsilon_\theta(z_t, x_f, x_b, \phi)),\end{aligned} \tag{6}$$

where s_f, s_b, s_p denote the guidance scales for foreground image condition, background image condition, and text condition accordingly. From Equation 6, we can see that there are four situations. Practically, during the training process, we randomly drop the instruction condition at 5%, drop both the background image and instruction at 5%, and drop all three conditions at 5%.

4. Experiments

4.1. Datasets and Pretrained Models

We train our methods on MagicBrush [37], an instruction-based dataset with high quality for local image editing dataset, which also provides the edited mask compared to other instruction-based datasets. We trained our MLLM reasoner and hints-guided editing network on the released train dataset, which contains 4,600 input images. In addition, although this dataset is high-quality, the number of instances is limited due to the high cost of labor. We augment this dataset five times at last, as introduced in our method, and extend the training dataset to contain 78,000 input image editing pairs, which is a significant number for training an editing model. The details of augmenting data can be found in the Appendix.

We evaluate the performance of our model on two datasets. One is the released test dataset of MagicBrush. In addition, we have built a small dataset with 100 sam-

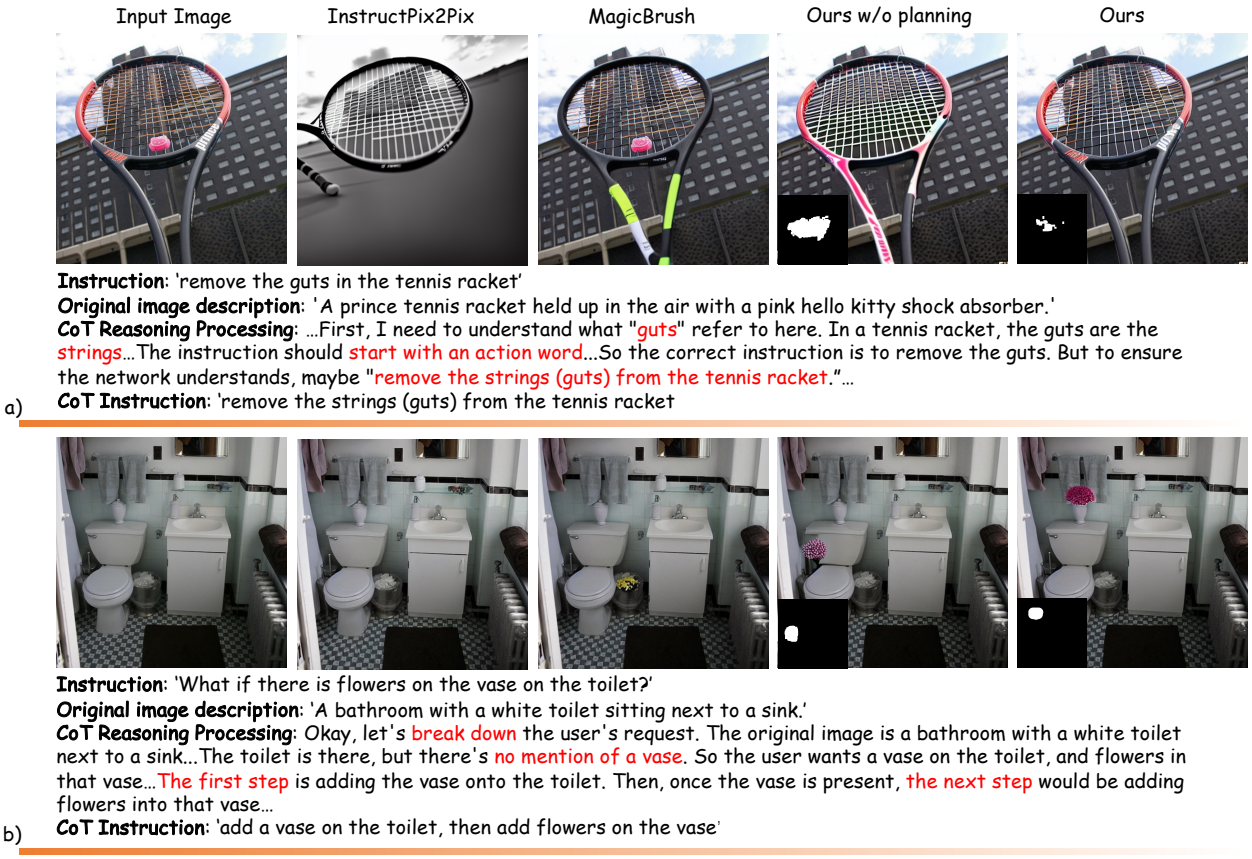

Figure 4. **Examples of our method of the instruction-based image editing on MagicBrush [37]**. Editing regions reasoned by M-LLM are shown in the bottom left corner of our editing results. We show Chain-of-thoughts (CoT) planning under the examples, helping to understand some concept or break down the tasks.

ples, containing abstract concepts, such as "warm", "dramatic", and "playful", extracted from HQEdit [14]. We evaluate the effectiveness of our whole framework, denoted as "ours,", consisting of planning, reasoning, and generation on the HQEdit-abstract dataset. In addition, for the description of the input image in the CoT planning phrase. We use the global description provided by the MagicBrush dataset and utilize gpt-4o [25] to generate the descriptions for HQEdit [14] or other open-domain cases.

We utilize three pretrained models and finetune our method on that. For the hints-generation network, we utilize SAM [15] for segmentation and LLava-7b [19, 20] for multi-modal LLMs. For the hints-guided network, we utilize instructPix2Pix [1] as the initialization for the weights of denoising U-Net parts.

4.2. Baselines and Metrics

We have picked InstructPix2Pix [1], InstructDiffusion [8], MagicBrush [37], and HIVE [39], HQEdit [14] as our baselines. We run the checkpoints provided publicly on MagicBrush's dev dataset, keeping the hyper-parameters the same as those in the released codes. Those works do not need users to provide a mask at the test time.

Following previous image editing methods [37], we use the embeddings of the CLIP [27] and DINO [2] to calculate the cosine similarity between the generated output and the ground-truth output provided by the dataset, which are denoted as CLIP-I and DINO-I separately. In addition, we utilize the global and local descriptions of the ground-truth output. We also use CLIP to calculate the similarity between the generated output and the description, which are denoted as CLIP-T (Global) and CLIP-T (Local). As for the HQEdit-abstract dataset, we conduct a user study of around 23 workers. Each worker needs to answer two questions for each pair. One is for the correctness of editing quality, another one is for the subject rating about the consistency with the corresponding abstract concepts, denoted as abstract-score.

4.3. Implementation details

We train each model individually. For the hints-guided network, we trained our model from the pretrained InstructPix2Pix [1] at the resolution 256 $\times$ 256 with epoch 200 on the training set of the MagicBrush [37]. The batch size

Table 1. **Quantitative results of our instruction-based editing model on MagicBrush [37] test datasets**. We calculate the CLIP [27] similarity using global and local descriptions separately. The total score is the average score among all metrics.

Methods	Total Score	CLIP-I↑	DINO-I↑	CLIP-T↑ (Global)	CLIP-T ↑ (Local)
InstructPix2Pix [1]	0.5457	0.8595	0.7501	0.2942	0.2791
InstructDiffusion [8]	0.5754	0.8980	0.8226	0.2997	0.2814
MagicBrush [37]	0.5853	0.9080	0.8443	**0.3035**	**0.2855**
HIVE [39]	0.5493	0.8599	0.7681	0.2928	0.2762
Ours w/o planning	0.5881	0.9117	0.8554	0.3026	0.2826
Ours	**0.5904**	**0.9172**	**0.8658**	0.2995	0.2789

Table 2. **Quantitative results of our multi-modal chain-of-thoughts editing framework on HQEdit-Abstract Dataset.** We show the voting ratio regarding the correctness of editing quality and the subject rating of abstract concepts.

Methods	Editing ↑ Quality	Abstract↑ Score
MagicBrush [37]	20.71%	22.76%
HQEdit [14]	23.53%	23.01%
Ours w/o planning	**28.64%**	24.80%
Ours	27.10%	**29.41%**

is 2 for 8 V100 GPUs. We set the learning rate to $1e^{-4}$. We use the SD-XL [26] as the generative fill model and set the probability threshold γ to pick the augmented images as 50%. We use the ground-truth mask for training and the mask predicted by the hints-generation network for inference. The DDIM steps for inference are set to 100 as the original instructPix2pix [1]. For the hints-generation network, the total training step is 2500 with a learning rate of $1e^{-4}$ with the ground truth mask from MagicBrush [37]. The threshold K in CoT planning is set to 3.

4.4. Experimental Results

4.4.1. Experimental Results on MagicBrush

Table 1 shows the quantitative results of our methods, where the proposed method shows the state-of-the-art performances compared with baselines. We also give some visual results in Figure 4 to prove the efficiency of the proposed method. Since we infer the edit hints from multi-modality LLM, the mask gives accurate hints of the edited region, which performs better than previous methods.

Based on the proposed editing region-based finetuning, the proposed method can accurately reason the location. Another example is that the original InstructPix2pix [1] and MagicBrush [37] can not perform local editing well since they do not have an explicit mask. In addition, for some difficult examples, instead of changing too much, some baselines, lacking of powerful editing ability, choose not to change the input image, leading to a low CLIP-T(Local) score. In conclusion, our method could edit the image at an appropriate level with the help of reasoning hints.

4.4.2. Experimental Results on HQEdit-Abstract

Table 2 shows the user study results of our methods, compared with HQEdit [14], MagicBrush [37], and our method without prompt planning. The table shows that our method produces better results than the framework without the help of M-LLMs. However, the score of editing quality has slightly decreased due to the decrease in image quality after the operation of conditional generation, illustrated in Sec 3.1. However, due to the powerful knowledge brought by M-LLMs, the edited results with our framework could bring more plentiful results to building an atmosphere regarding abstract concept topics.

Figure 5 shows two examples of our framework. The abstract topic is "warm" for the first line and "dramatic" for the second line. Our CoT planning with multimodal LLMs could instantiate the abstract instruction into more specific details. For example, if we want to create a dramatic nighttime tempest, we should first add turbulent waves and then dark storm clouds and lightning. This information could not be encoded only from the abstract instructions.

Ablation	Methods	CLIP-I↑	CLIP-T↑
Hints method	Pretrained LISA	0.9081	0.2766
	Ground Truth	**0.9277**	**0.2838**
	Ours	0.9219	0.2835
The ratio of augmented data	0	**0.9314**	0.2771
	0.25	0.9290	0.2796
	0.5	0.9219	**0.2835**
	0.75	0.9206	0.2796

Table 3. **Ablation study on the hints generation method and the ratio of the utility of augmented data in the training process.** We evaluate the performance on MagicBrush dev dataset with the original instruction. CLIP-T is calculated via the local description. Both the hint generation and the augmented dataset benefit the quality of generation results.

4.5. Ablation Study

Firstly, we conduct the ablation study on the classifier-free guidance. We added extra conditions for classifier-free guidance, and thus, we varied the values of the Classifier-free Guidance (CFG) foreground image and background

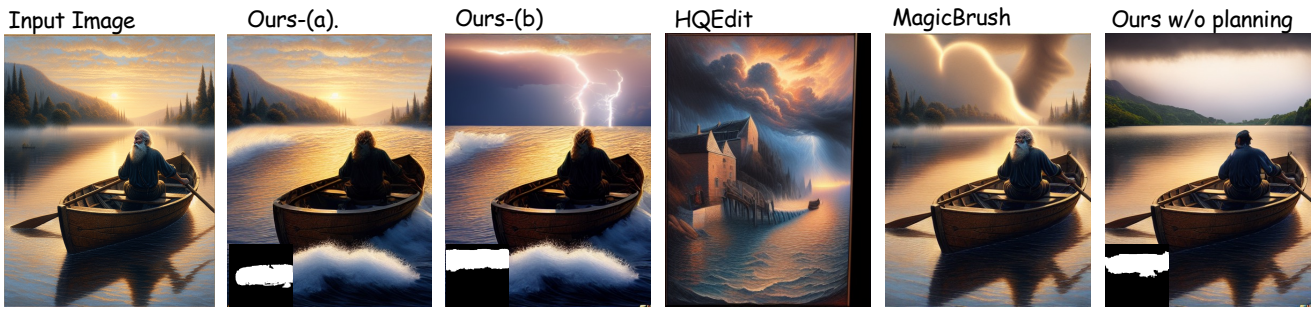

Figure 5. **Examples of our Multimodal Chain-of-Thoughts Editing Framework on HQ-Abstract.** The abstract topic is 'dramatic". Our CoT planning with multimodal LLMs could instantiate the abstract instruction into more specific details. The editing area is shown in the bottom left of each image.

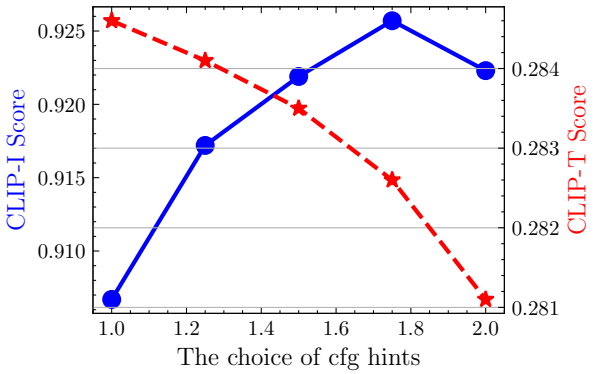

Figure 6. **The influence of editing hints-related CFG.** We evaluate the model performance on MagicBrush dev dataset with the original instruction. We have set the text CFG to a fixed value 7.5.

image simultaneously from 1.0 to 2.0 with a fixed value for the CFG text of 7.5. Figure 6 shows the CLIP-I and CLIP-T for local description with different hints-related CFG at the test time. We can see that more control of the hints-related conditions will increase the CLIP-I scores since we can better maintain the unmasked area of the input images. However, the larger control of hints-related conditions will also hurt the generation ability of our editing model and the diversity of our edited results with the decrease in CLIP-T scores.

In addition, we conduct the ablation study on the mask-generated model. We compare our method with the pre-trained LISA model and the ground truth of the mask provided by the MagicBrush dataset. Finally, we also conducted an ablation study on the random probability γ, indicating how to use our generated data in our training process. The details can be found in Table 3.

4.6. Flux Editing Models with CoT Planning

We extend the Chain-of-Thought Planning on Flux-based Editing Models. Based on Flux [17] text-to-image generation model, we train the flux editing model with Control-net [38] framework. Figure 7 shows an example of flux editing models with CoT planning. We find that with the help of CoT planning, the edited results have more alignment with the input image and more fulfilling contents considering the instruction.

Figure 7. Examples of Flux editing network with Chain-of-Thoughts planning.

5. Conclusion

In this paper, we propose a novel framework, called Multimodal Chain-of-Thought Editing, as the bridge between scene understanding and scene editing. This framework consists of three parts: a Chain-of-Thought planner, an MLLM reasoner, and a hints-guided editing network. The experiments show the advantage of the proposed method over the state-of-the-art method on the benchmark of *i.e.*, MagicBrush [37] and a small dataset with abstract instructions, extracted from HQEdit [14]. In the future, the current MLLM reasoner still faces the issue of inaccurate editing regions. One possible direction is utilizing the current Chain-of-Thought LLMs to improve the reasoning ability. As for the base generation model, it is promising that extending our whole framework on the Flux model comprehensively.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 3, 4, 5, 6, 7

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[3] Tuhin Chakrabarty, Kanishk Singh, Arkadiy Saakyan, and Smaranda Muresan. Learning to follow object-centric image editing instructions faithfully. *arXiv preprint arXiv:2310.19145*, 2023. 2, 3

[4] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 2, 4

[5] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 2, 3

[6] DiFei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. 2023. 3

[7] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 3

[8] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. Instructdiffusion: A generalist modeling interface for vision tasks. *CoRR*, abs/2309.03895, 2023. 2, 6, 7

[9] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023. 3

[10] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. *arXiv preprint arXiv:2312.10113*, 2023. 2, 3

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 5

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[13] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *arXiv preprint arXiv:2312.06739*, 2023. 2, 3

[14] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024. 2, 4, 6, 7, 8

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4, 6

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023. 3

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 8

[18] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 3, 4

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3, 6

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3, 4, 6

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[22] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. Watch your steps: Local image and scene editing by text instructions. *arXiv preprint arXiv:2308.08947*, 2023. 2

[23] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[24] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via visual prompting. *arXiv preprint arXiv:2307.14331*, 2023. 2

[25] OpenAI. Gpt-4 technical report, 2023. 3, 6

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 7

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 4, 5

[29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 3

[30] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*, 2023. 2, 3

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[32] Jinheng Xie, Kai Ye, Yudong Li, Yuexiang Li, Kevin Qinghong Lin, Yefeng Zheng, Linlin Shen, and Mike Zheng Shou. Visorgpt: Learning visual prior via generative pre-training. *arXiv preprint arXiv:2305.13777*, 2023. 3

[33] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*, 2024. 2, 3, 4

[34] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 3

[35] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 2

[36] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023. 3

[37] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023. 2, 4, 5, 6, 7, 8

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 8

[39] 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. *arXiv preprint arXiv:2303.09618*, 2023. 2, 6, 7

[40] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 4

[41] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[42] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 3, 4