

InstructEdit: Improving Automatic Masks for Diffusion-based Image Editing With User Instructions

Qian Wang
KAUST

qian.wang@kaust.edu.sa

Biao Zhang
KAUST

biao.zhang@kaust.edu.sa

Michael Birsak
KAUST

michael.birsak@kaust.edu.sa

Peter Wonka
KAUST

peter.wonka@kaust.edu.sa

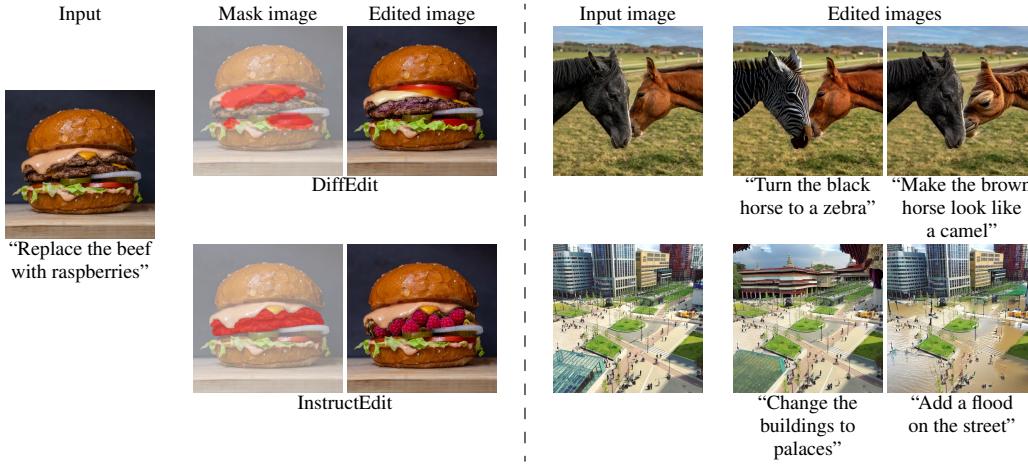


Figure 1: Left: a comparison between DiffEdit and InstructEdit. Right: examples of editing using InstructEdit. Note that InstructEdit only requires user instructions as input, while DiffEdit needs an input caption and an edited caption instead.

Abstract

Recent works have explored text-guided image editing using diffusion models and generated edited images based on text prompts. However, the models struggle to accurately locate the regions to be edited and faithfully perform precise edits. In this work, we propose a framework termed **InstructEdit** that can do fine-grained editing based on user instructions. Our proposed framework has three components: language processor, segmenter, and image editor. The first component, the language processor, processes the user instruction using a large language model. The goal of this processing is to parse the user instruction and output prompts for the segmenter and captions for the image editor. We adopt ChatGPT and optionally BLIP2 for this step. The second component, the segmenter, uses the segmentation prompt provided by the language processor. We employ a state-of-the-art segmentation framework Grounded Segment Anything to automatically generate a high-quality mask based on the segmentation prompt. The third component, the image editor, uses the captions from the language processor and the masks from the segmenter to compute the edited image. We adopt Stable Diffusion and the mask-guided

generation from DiffEdit for this purpose. Experiments show that our method outperforms previous editing methods in fine-grained editing applications where the input image contains a complex object or multiple objects. We improve the mask quality over DiffEdit and thus improve the quality of edited images. We also show that our framework can accept multiple forms of user instructions as input. We provide the code at <https://github.com/QianWangX/InstructEdit> and project page at <https://qianwangx.github.io/InstructEdit/>.

1 Introduction

Generative diffusion models are a versatile tool to generate images [45, 43, 4, 44, 27], videos [19, 13, 18, 5, 51], and 3D shapes [22, 53, 15, 56]. In addition, the powerful representation learned by generative diffusion models makes them a great basis for downstream editing operations. In this paper, we are particularly interested in image editing operations.

Training-free and tuning-free text-guided image editing using diffusion models [16, 10, 48, 28] usually relies on descriptive text captions for both input image and edited image. However, one line of the work focuses on accepting human instructions as input to edit the images, as human instructions are more intuitive for a user to provide, and can be free from specific prompting structures. InstructPix2Pix [6] constructs an “input caption, edited caption, user instruction” triplet dataset by fine-tuning GPT-3 [7] and generates pairs of input image and corresponding edited image using Prompt-to-Prompt [16]. Here we also try to accept human instructions as input. Instead of fine-tuning a large language model, we utilize the in-context learning ability of it to parse user instructions on-the-fly. One challenge of this tuning-free approach is that the user instructions can be as unclear as “Add glasses” and “Turn him into a bearded man”, where no information about the referred object is given. To tackle this problem, we utilize a large language model to understand the user instructions along with a multi-modal model to improve the comprehension.

Using a mask to provide guidance to the models about the specific areas to edit is a logical and intuitive approach. There already exist several editing models [6, 16, 28, 48] that do not require a mask as input, but they solely exploit the input caption and edited caption to modify the text-conditioned diffusion process. This approach generally works well when there is a single object in the image or if there is no object of the same type than the object the user wants to edit. However, with multiple objects in the input image, diffusion models struggle to correctly identify which object the user intends to edit. For example, if the user wants to edit a specific object like “A yellow chair” in an image with multiple chairs present or “The cat on the left” in a group of cats. In such cases, an input mask can help the model to correctly locate the object of interest. Thus, image editing with mask guidance is favorable for local editing, especially in multi-object scenarios, as a mask can explicitly control which regions to edit and which regions to leave untouched. As InstructPix2Pix utilizes the mask-free image editing method Prompt-to-Prompt to generate the paired input-edited images, its ability to accurately locate objects is not sufficient in complex cases. One possible solution is to create more paired images and paired captions to fine-tune InstructPix2Pix. Here we provide another solution by adopting a pre-trained image segmentation model for generating high-quality masks to guide the editing. This does not require any training or dataset collection.

While it is possible to ask a user to paint a mask, this also has disadvantages. Mainly, it requires time-consuming and detailed user interaction. We therefore follow the previous work DiffEdit that employs automatically generated masks. DiffEdit estimates a mask by subtracting the predicted noise guided by the input caption and the edited caption, respectively. In general, the quality of the mask has a significant impact on the visual quality of the resulting edit. While DiffEdit obtains very good masks in many cases, there are several instances where DiffEdit fails to produce high-quality masks: 1) the descriptive captions are not informative enough; 2) the threshold of the mask filtering is not correctly set; 3) there is more than one object of the same or different type; and 4) there are many parts of one object in the image that may cause ambiguity. In our work, we set out to tackle the challenge of computing improved masks in such challenging cases. We adopt a powerful grounding segmentation module called Grounded Segment Anything (Grounded SAM), which combines a segmentation network Segment Anything [24] with an open-set object detector Grounded DINO [30]. By simply providing the segmentation prompt to Grounded SAM, a mask that exactly matches the shape of the object can be generated. This provides extra grounding and segmentation ability to the framework, which helps the model to locate and extract the object(s) to be edited.

In this work, we propose a framework to use large pre-trained models to edit images following user instructions based on DiffEdit. We call our method **InstructEdit** as our method is a natural extension of the original DiffEdit that accepts user instructions as input instead of pairs of captions for input image and edited image. We automatically extract higher-quality masks compared to DiffEdit, and therefore achieve more preferable and stable editing results in more complex multi-object image editing scenarios. Specifically, our method has three components: a language processor, a segmenter and an image editor. We first use the language processor to understand the user instruction by identifying which object(s) should be edited and how. Then it provides the segmentation prompt for the segmenter and captions for the image editor. We use ChatGPT to parse the user instruction and optionally adopt BLIP2 [26] when the user instruction is unclear. The segmenter then accepts the segmentation prompt and generates a mask that outlines the region according to the segmentation prompt. We utilize Grounded Segment Anything as the segmenter which combines both high grounding and segmentation abilities. Finally, the image editor performs image editing using the captions along with the generated mask. We adopt Stable Diffusion along with the mask-guided image editing process to do the editing.

We show that our method outperforms previous editing methods in fine-grained editing applications. Specifically, we are interested in input images that contain 1) one object and we want to edit one part of the object. 2) multiple objects and we want to edit one or multiple objects. We show that we improve the quality of edited images by improving the mask quality over DiffEdit. We also show that our framework can accept multiple forms of user instructions as input. We summarize our contributions as below:

- We propose a diffusion-based text-guided image editing framework that accepts user instruction as input instead of input caption and edited caption.
- We outperform baseline methods in fine-grained editing when we want to edit one part of the object in a single-object image or one or multiple objects in a multi-object image.
- We improve the mask quality over the original DiffEdit and thus improve the image quality.
- We accept various kinds of user instructions as input.

2 Related work

2.1 Image editing using diffusion models

Pre-trained diffusion models [43, 45, 44, 20] can be used to do various image editing tasks. Several works [47, 23, 25] fine-tune the diffusion models weight or optimize a loss function to perform image editing. However, in these works fine-tuning requires a relatively long time to obtain a single edited image, and each editing prompt requires its own fine-tuning process.

Many works [33, 16, 37, 48] achieve good image editing results using a tuning-free approach. Prompt-to-Prompt [16] proposes to edit the cross-attention maps by comparing the input caption and the edited caption. MDP [48] proposes to manipulate the diffusion path by analyzing the sampling formula. In this line of work, a mask is not required during the editing process, thereby making it easier to use than systems that rely on masking. These kinds of methods usually perform well when there is a single foreground object in the image, but fail when more fine-grained control is needed.

By providing a manually designed mask as input, a user can explicitly control which region to edit and which region to preserve. Blended Diffusion [3] and Blended Latent Diffusion [2] utilize a mask to perform text-guided image editing by operating either in pixel space or in latent space. Shape-guided Diffusion [36] proposes to use a mask with inside-outside attention to preserve the shape of the object to be edited. MasaCtrl [8] focuses on performing non-rigid editing while using a mask to alleviate the confusion of foreground with background objects.

Although a mask provides additional control in the editing process, generating a manual mask can be a burden for the user if the mask is not automatically estimated. [36, 8] can also replace the manual mask with an automatic mask generated by cross-attention maps. Another prominent work DiffEdit [10] proposes a way to automatically predict a mask by contrasting the predicted noise conditioned on the input caption and the edited caption. In this work, we intend to exploit the benefit of using a mask without bringing the burden of manually labeling a mask to the user. We adopt large pre-trained models to help us automatically infer a mask.

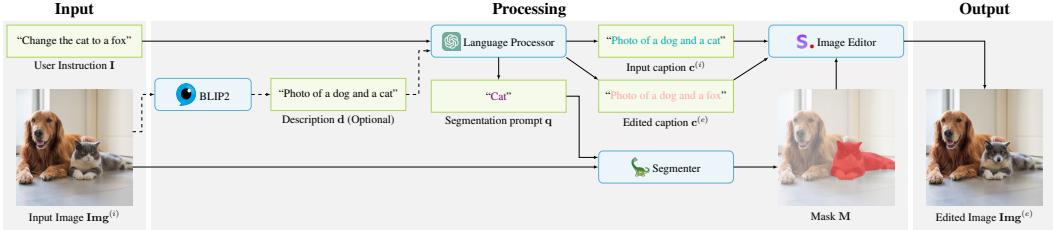


Figure 2: Pipeline: given a user instruction, a language processor first parses the instruction into a **segmentation prompt**, an **input caption**, and an **edited caption**. A segmenter then generates a mask based on the segmentation prompt. The mask along with the input and edited captions are then going to an image editor to produce the final output.

Several recent video editing papers that extend the concepts from image editing to videos [29, 49, 38, 31, 34]. While video editing is beyond the scope of our work, it is very interesting for future research, as our proposed editing capabilities are not yet available in video editing frameworks.

2.2 Foundational models

Large language models. The development of large language models has been a rapidly evolving field in recent years [40, 41, 7, 11, 42, 9]. One important contribution is the GPT series of models [40, 41, 7]. These models are pre-trained on massive amounts of text data and then fine-tuned for specific tasks. Most notably, ChatGPT [35] is a variant of the GPT series of models that are designed specifically for generating human-like responses in conversational settings. We make use of ChatGPT in our work for the purpose of information extraction from user instruction.

Segmentation model and grounding detector. Segment Anything [24] is a segmentation model which uses a combination of different input prompts and enables zero-shot generalization to unfamiliar objects and images without requiring additional training. Grounding DINO [30] is an open-set object detector which combines the Transformer-based detector DINO [54] with grounded pre-training to detect arbitrary objects with human inputs such as category names or referential expressions. Their system outputs multiple pairs of object boxes and noun phrases give a prompt.

Images and language. Vision language models (VLMs) [39, 1, 32, 21, 26] are a powerful class of models that combine computer vision and natural language processing. VLMs have gained significant attention in recent years due to their ability to bridge the gap between visual and textual information, enabling a range of applications such as image captioning, visual question answering, and image retrieval. Especially, BLIP-2 [26] unlocks the capability of zero-shot instructed image-to-text generation. Given an input image, BLIP-2 can generate various natural language responses according to the user’s instruction.

3 Method

3.1 Preliminaries

During the training process of diffusion models, we have the objective function

$$\min_{\theta} \mathbb{E}_{\mathbf{x}_0, \epsilon, t} \|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, \mathbf{c}, t)\|^2, \quad (1)$$

where \mathbf{x}_0 is the input image, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ is Gaussian noise that is added to the input image and ϵ_{θ} is the noise estimator that is used to predict the added noise. t is the denoising timestep while \mathbf{c} is the condition of the diffusion model. In this work we only consider \mathbf{c} to be a text prompt of a text-guided diffusion model. After training, the noise estimator ϵ_{θ} can be used to generate new samples. We use DDIM [46], which is a deterministic sampler with denoising steps

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \cdot f_{\theta}(\mathbf{x}_t, \mathbf{c}, t) + \sqrt{1 - \alpha_{t-1}} \cdot \epsilon_{\theta}(\mathbf{x}_t, \mathbf{c}, t), \quad (2)$$

where $f_{\theta}(\mathbf{x}_t, \mathbf{c}, t) = \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \cdot \epsilon_{\theta}(\mathbf{x}_t, \mathbf{c}, t)}{\sqrt{\alpha_t}}$, and α_t is the noise schedule factor in DDIM. We denote $\epsilon_t = \epsilon_{\theta}(\mathbf{x}_t, \mathbf{c}, t)$. Given an input image, we can use DDIM inversion to invert it into an initial noise

tensor \mathbf{x}_T . Each inversion step is calculated as

$$\mathbf{x}_{t+1} = \sqrt{\alpha_{t+1}} \cdot f_\theta(\mathbf{x}_t, \mathbf{c}, t) + \sqrt{1 - \alpha_{t+1}} \cdot \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t). \quad (3)$$

We iteratively apply this formula until obtaining \mathbf{x}_T . However, if we stop the inversion step at timestep $r \leq T$, we encode \mathbf{x}_0 into a less noised version \mathbf{x}_r . r is called the encoding ratio as in DiffEdit. A larger value r indicates a stronger editing effect, making the edited image guided more by the edited caption but look less like the input image.

3.2 Language processor

Given a user instruction \mathbf{I} as input, we use a large language model ChatGPT [35] to extract segmentation prompt \mathbf{q} for the segmenter, and an input caption $\mathbf{c}^{(i)}$ and an edited caption $\mathbf{c}^{(e)}$ for the image editor. Here, we utilize the in-context learning ability of large language models [12] to achieve zero-shot task learning. In-context learning does not require any tuning of the parameters of the language model. Instead, it learns the pattern from the task examples and makes predictions when it sees a new example. In this work, by giving a few examples, ChatGPT is able to learn the task and follow the instructions. The prompting system for ChatGPT to learn the task in our work is therefore consisting of two parts: a context description and a few task examples. The context description is to show the context that ChatGPT needs to provide a segmentation prompt for the segmenter, and an input caption and an edited caption for the image editor. A task example shows how ChatGPT should manipulate a user instruction that is provided as input.

When the user instruction \mathbf{I} or the description of the object to be edited is unclear, it is difficult for ChatGPT to correctly provide the prompts as it has no access to the content of the input image. The vision-language model BLIP2 can process the image and is able to answer questions about its content. BLIP2 can provide a short description of the original image, which can assist ChatGPT to provide prompts for the segmenter and captions for the image editor.

In this work, we optionally query BLIP2 to obtain a description \mathbf{d} of the image. Given an input image, we first ask BLIP2 “*Is this a photo, a painting or another kind of art?*”. We denote the answer as ρ and reuse it in another query to BLIP2 composed as “ ρ of” to obtain a completed sentence describing the image as input prompt to ChatGPT. ChatGPT in turn can refine the prompt by identifying which object to edit and provide more details even when the user instruction does not specify the content to be edited or the description is incomplete to unambiguously refer to the intended objects in the image.

3.3 Segmenter

We use Grounded Segment Anything as our segmenter to locate the object(s) to be edited and to compute a corresponding mask. Grounded Segment Anything is a framework which combines Grounding DINO [30] and Segment Anything [24]. Grounding DINO is an open-set object detector, which can accept a given text and output one or multiple detected bounding boxes and a text similarity score per bounding box. Segment Anything (SAM) is a powerful segmentation model. It can accept the bounding box output by Grounding DINO and produce high-quality binary masks for the downstream tasks.

Grounding DINO is first applied to get a bounding box for a given segmentation prompt \mathbf{q} by

$$\text{DINO}(\mathbf{x}_0, \mathbf{q}) = \mathbf{b} = [h, w, \Delta h, \Delta w],$$

where $[h, w]$ is the top-left corner coordinate of the detected bounding box in pixel space, and $[\Delta h, \Delta w]$ is the size of the bounding box. Then, the bounding box is refined to a per-pixel binary mask \mathbf{M} by

$$\text{SAM}(\mathbf{x}_0, \mathbf{b}) = \mathbf{M}.$$

3.4 Image editor

We adopt the mask-guided image editing as in DiffEdit [10]. Given an input image $\mathbf{Img}^{(i)}$ (which is also denoted as \mathbf{x}_0 in 3.1), we want to edit it to get the edited image $\mathbf{Img}^{(e)}$. With the automatically generated mask \mathbf{M} and an encoded noise \mathbf{x}_r , the mask-guided DDIM denoising step is formulated as

$$\tilde{\mathbf{y}}_t = \mathbf{M}\mathbf{y}_t + (1 - \mathbf{M})\mathbf{x}_t, \quad (4)$$

where $\mathbf{y}_t = \begin{cases} \mathbf{x}_r & \text{if } t = r, \\ \epsilon_\theta(\mathbf{y}_{t-1}, \mathbf{c}, t) & \text{otherwise.} \end{cases}$. The region within the mask will have the changes guided by the edited caption, while the region outside the mask will be mapped back to the original pixels. We obtain $\text{Img}^{(e)} = \tilde{\mathbf{y}}_0$ after iteratively applying Eq. 4.

4 Experiments

4.1 Settings

Editing types. In this work we focus on image edits restricted to certain object categories or number of occurrence in the input image. We consider replacing objects with other objects and changing attributes of objects. The edited image should faithfully follow the user instructions while also preserving the layout and the appearance of the other parts in the original image.

- Single-object: there is a foreground object consisting of multiple parts and we want to edit one part of that object.
- Multi-object of the same type: there are several objects of the same type and we want to edit one or more of them.
- Multi-object of different types: there are several objects of different types and we want to edit one of them.

Baselines. We choose three text-guided image editing methods that do not require a manual mask as input as our baselines. MDP- ϵ_t [48] is a mask-free editing method that manipulates the diffusion paths by mixing the predicted noise guided by the input caption and the edited caption according to a defined mixing schedule. InstructPix2Pix [6] is also a mask-free editing method. It trains a conditional diffusion model that accepts both image and a user instruction as input to edit the input image. It highlights the user instruction as input without edited captions by exploiting a large language model. DiffEdit [10] is a mask-based editing method that produces an automatically computed mask by subtracting the predicted noises guided by the input caption and the edited caption. InstructEdit adopts the same mask guidance as in DiffEdit to generate the edited image. However, InstructEdit uses a pre-trained segmentation model to automatically compute a new mask.

All the experiments are performed on a single NVIDIA A100. We use Stable Diffusion v1.5 as the backbone of the image editor. We use the model’s weights and implementation of Grounded Segment Anything from <https://github.com/IDEA-Research/Grounded-Segment-Anything>. More details can be found in the Supplementary Materials.

Evaluation. We provide both qualitative and quantitative results for our method. For quantitative metrics, we use LPIPS [55] to measure the similarity between the input image and the edited image, CLIP score [17] to measure the instruction-image compatibility and CLIP directional similarity [14] to see if the change in images is consistent with the change in captions. As quantitative metrics alone usually cannot align with human judgment, we additionally conduct a user study for a better evaluation.

4.2 Baseline comparisons

We provide selected qualitative comparisons between our method and other baselines in Fig. 4. It can be seen that InstructEdit can accurately locate either an object or a part of an object to be edited according to the user instruction in a fine-grained manner. In addition, the edits are faithfully performed within the regions InstructEdit identifies. The three baselines have difficulty to accurately locate the object of interest. We observed two types of issues: the edited region may overshoot the object specified in the user instruction (e.g. in the garage door and main door example a larger area is edited, and in the green buses example more buses that are not green are edited), or the object is not correctly located (e.g. in the windows example all the three methods edit the painting instead of the windows). For mask-free baselines MDP- ϵ_t and InstructPix2Pix, more difficulties are found to perform a correct degree of editing. If we want to preserve the regions that should not be edited, the region that should be edited is not changed significantly enough (e.g. in the middle bus example and the laptops example those edited images are almost the same as the input image). If we want to do a

Table 1: Quantitative comparisons between our method and baselines.

	MDP- ϵ_t	InstructPix2Pix	DiffEdit	InstructEdit
LPIPS \downarrow	0.214	0.290	0.167	0.121
CLIP score \uparrow	26.414	25.844	26.847	27.404
CLIP directional similarity \uparrow	0.079	0.114	0.106	0.082

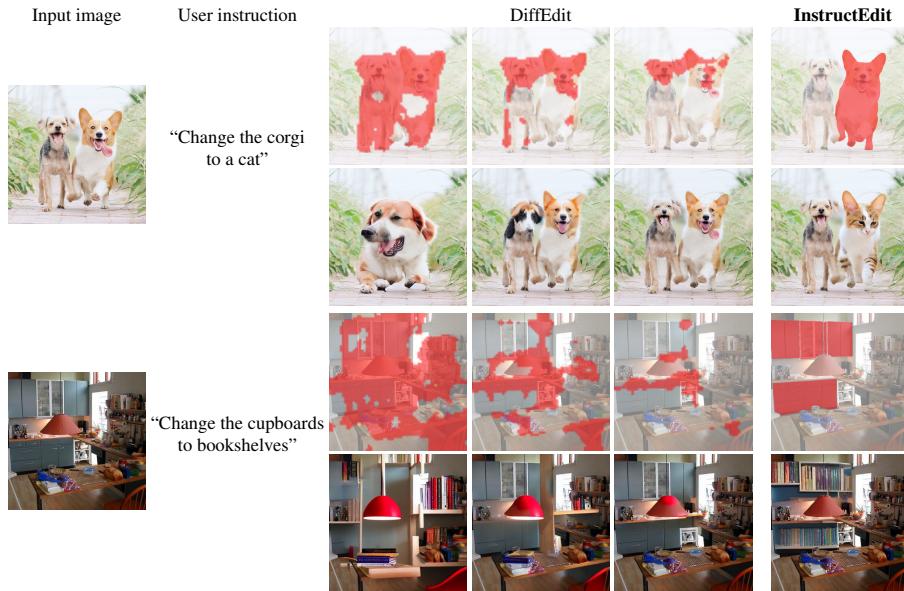


Figure 3: Comparison of the masks (colored in red and blended with the input image) and the corresponding edited image (below each mask) generated by DiffEdit and InstructEdit.

clearly noticeable edit, more regions are changed unexpectedly (e.g. in the cat examples all the cats are edited). These results show that masks and their quality are crucial in fine-grained editing. More qualitative results can be found in the Supplementary Materials. We show more examples in Sec. 4.3 to show that InstructPix2Pix outperforms DiffEdit by improving the quality of masks.

We show the results of the quantitative evaluation in Tab. 1 for the 10 editing examples shown in Fig. 4. Results show that our method has the best semantic preservation of the edited image compared to the input image and the best alignment of the instruction and edited image pair. The metrics do not capture the large improvements due to our method, because CLIP itself does not capture the fine-grained spatial localization required to judge our complex edits. We therefore perform a user study for these 10 editing examples by comparing our method with the other baseline methods. We could acquire 26 workers on MTurk who were presented triplets of images, each triplet consisting of one input image and two edited images. One of the edited images always was the output from our method, the second one was a randomly picked edited image from one of the baselines methods. The results show that InstructEdit was preferred over MDP- ϵ_t , InstructPix2Pix and DiffEdit in 83.0%, 83.0%, and 84.5% of the cases, respectively.

4.3 Mask improvement

We show examples of how InstructEdit improves the mask quality over the original DiffEdit in Fig. 3. For each example, we use the user instruction as input to InstructEdit and design specific input captions and edited captions for DiffEdit. We also select three different mask thresholds θ for DiffEdit to show how the mask threshold influences the mask quality and therefore the image quality. We show that for DiffEdit the generated mask cannot accurately outline the intended region as specified in the user instruction for all different θ . Therefore, the generated images either have too many or too few changes. On the contrary, for InstructEdit there is no such mask threshold and the generated masks can exactly localize the intended region to be edited. By improving the mask quality, InstructEdit can faithfully do the edit and outperform DiffEdit. We show that without the

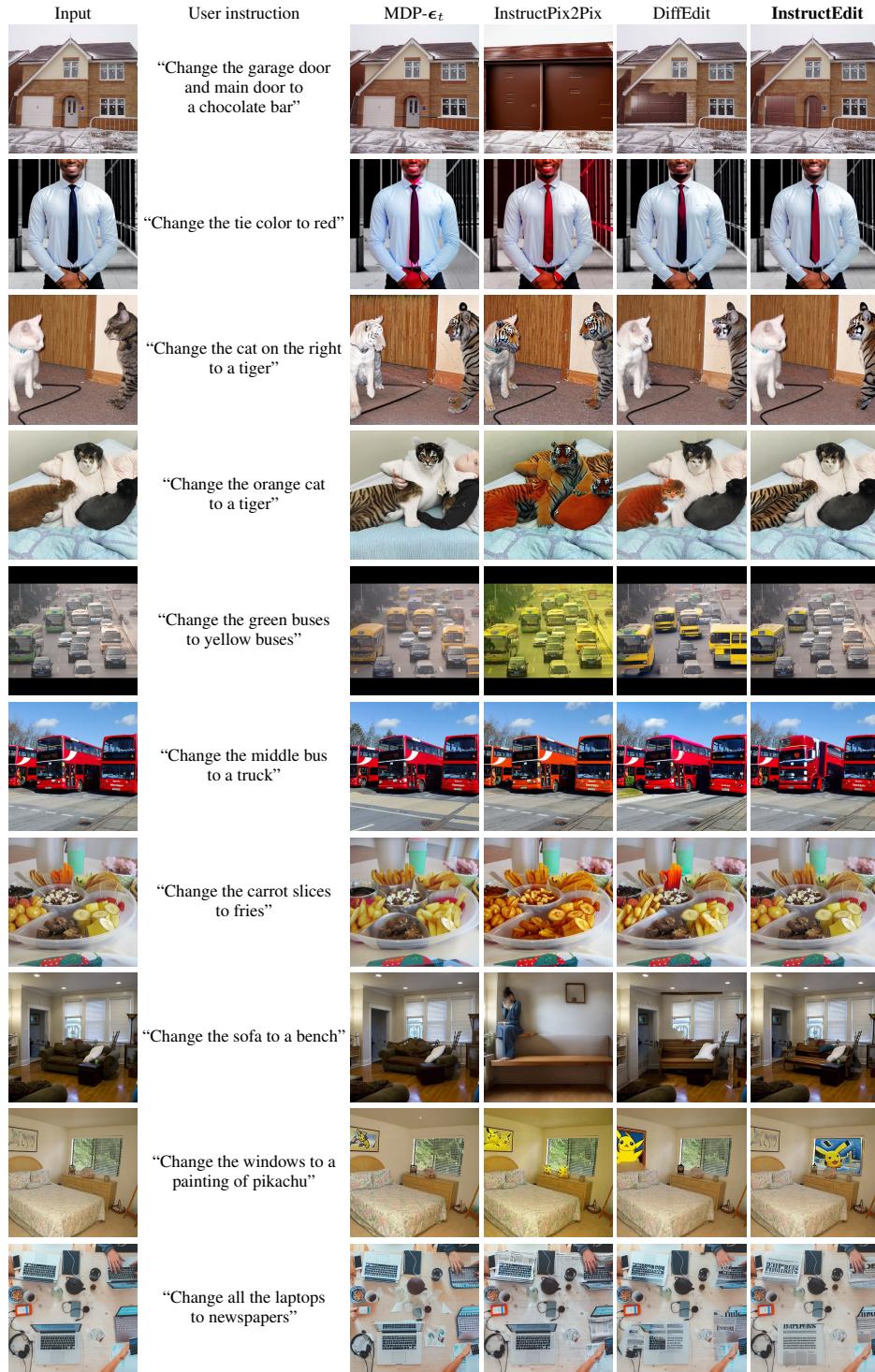


Figure 4: Qualitative results of baselines and our method. Here we use the same form of instruction “Change ... to ...” as an example.



Figure 5: Examples of how BLIP2 improves the quality of edited images by improving the generated prompts. The three generated prompts are **segmentation prompt**, **input caption** and **edited caption**, respectively. For the first example the BLIP2 description of the image is “Photo of a dog and a cat”. We show three examples for w./wo. BLIP2 with increasing encoding ratio r from left to right. In the second example the BLIP2 description is “Painting of a girl with a pearl earring by Jan Van Gogh”. (BLIP2 gets the name of the painting correct but the name of the artist wrong.)

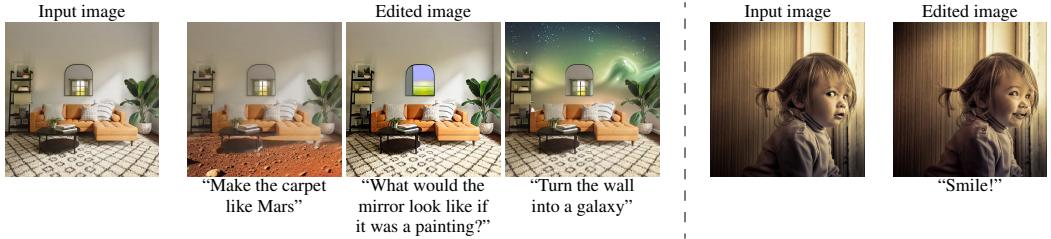


Figure 6: Examples of the variety of the input user instructions. Under each edited image is the input instruction.

help of the grounding pre-trained network, it is very hard for the diffusion model itself to accurately outline the region and do a fine-grained editing at this scale.

4.4 Instruction processing

In Fig. 5 we show the results of an ablation study and how BLIP2 helps to improve the quality of the edited images. When the instruction does not clearly refer to all the prominent objects in the image, e.g. not mentioning a visible “dog” like in the first example, the input and edited captions provided by ChatGPT do not contain “dog” as well. This leads to difficulties when the encoding ratio r is increasing. In the second example, where an unclear “it” is used to refer to the object to be edited, without BLIP2 ChatGPT fails to correctly generate a segmentation prompt and the captions. In such a case, BLIP2 provides an extra description of the image such that ChatGPT can understand which object should be edited and provides improved captions.

We also show in Fig. 6 that InstructEdit is very robust and can correctly understand differently phrased instructions as input. This demonstrates the benefit of adopting a large language model to process the instruction rather than hard-coded parsing.

5 Discussion and Conclusions

There are several limitations in our work. As ChatGPT and BLIP2 are probabilistic models, their outputs are not always optimal and thus may fail to correctly parse the user instruction. Also, as the editing is performed within the generated mask, which has the same shape as the object, it is difficult to conduct deformations. Concurrent work InpaintAnything [52] and EditAnything [50] also utilize Segment Anything to extract a mask and Stable Diffusion to create an image editing framework. Similar to our method they can achieve high-quality edits without a manual mask as input. One feature that is unique to our framework is the ability to accept user instructions as input. For future work, we are interested in incorporating other editing models which specialize in different editing types and extending our work to video editing.

In this work we proposed a framework termed InstructEdit that can do fine-grained editing and directly accept user instructions as input. We use a language processor to process user instructions, a segmenter to generate high-quality masks and an image editor to do the mask-guided editing. Experiments show that our method outperforms previous editing methods when doing fine-grained edits. We improve the mask quality over DiffEdit by incorporating the grounding pre-trained models and thus improve the quality of edited images. Our framework can also accept multiple forms of user instructions as input.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022.
- [3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.
- [4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [8] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing, 2023.
- [9] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [10] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.
- [13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023.
- [14] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021.
- [15] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023.
- [16] 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.
- [17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

- [18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.
- [19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.
- [20] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions, 2023.
- [21] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.
- [22] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.
- [23] 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. *arXiv preprint arXiv:2210.09276*, 2022.
- [24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
- [25] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022.
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
- [27] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Glichen: Open-set grounded text-to-image generation. 2023.
- [28] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022.
- [29] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control, 2023.
- [30] 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.
- [31] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos, 2023.
- [32] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179*, 2022.
- [33] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.
- [34] Eyal Molad, Eliah Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors, 2023.
- [35] OpenAI. Introducing chatgpt, 2023.
- [36] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention. *arXiv*, 2022.
- [37] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023.
- [38] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing, 2023.
- [39] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
- [40] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [41] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

- [42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.
- [43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [45] Chitwan Saharia, 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, 2022.
- [46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [47] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image, 2022.
- [48] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path, 2023.
- [49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, 2023.
- [50] Defeng Xie, Ruichen Wang, Jian Ma, Chen Chen, Haonan Lu, Dong Yang, Fobo Shi, and Xiaodong Lin. Edit everything: A text-guided generative system for images editing, 2023.
- [51] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation, 2022.
- [52] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting, 2023.
- [53] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445*, 2023.
- [54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.
- [55] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.
- [56] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *arXiv preprint arXiv:2305.04461*, 2023.

A Broader Impacts

Our framework enables easier user interaction for image editing. Users do not need to provide two captions for images but only an instruction. This simplicity allows a larger community to use the proposed framework for image editing, while also possibly increasing the potential of producing malicious content. A misuse of the method also allows for the creation of deep fakes and offensive content targeted at marginalized communities. The misuse of such techniques raises concerns such as reputational damage, privacy invasion, the spread of misinformation, perpetuating harmful stereotypes, inciting violence, and reinforcing biases. Responsible research, development, and deployment practices are necessary to mitigate harm. This includes developing robust detection mechanisms for deep fakes, implementing ethical guidelines, conducting impact assessments, collaborating with ethics researchers in AI and affected communities, and promoting education and awareness to foster responsible use. By prioritizing ethical considerations, we can harness the potential of image editing technologies while protecting individuals' rights and ensuring their positive impact on society. For the language processor, a safeguard for large language models can help to filter negative content. For the image editor, a negative content detector can also be adopted for the edited image.

B DiffEdit

DiffEdit is a mask-based text-guided image editing model that can automatically generate masks. During the denoising process, different text prompts will guide the diffusion model to yield different predictions. By contrasting two predictions, the difference can give information for which image regions the input caption and edited caption have different estimates. We refer to the text prompt that is used to generate or inverse the input image as input caption $\mathbf{c}^{(i)}$, and to the text prompt that is used to describe the edited image as edited caption $\mathbf{c}^{(e)}$. The predicted noise at each timestep t guided by $\mathbf{c}^{(i)}$ and $\mathbf{c}^{(e)}$ are $\epsilon_t^{(i)} = \epsilon_\theta(\mathbf{x}_t, \mathbf{c}^{(i)}, t)$ and $\epsilon_t^{(e)} = \epsilon_\theta(\mathbf{x}_t, \mathbf{c}^{(e)}, t)$, respectively. The difference of the prediction $\epsilon_t^{(d)}$ at timestep t is then calculated as:

$$\epsilon_t^{(d)} = |\epsilon_t^{(i)} - \epsilon_t^{(e)}|, \quad (5)$$

$\tilde{\epsilon}_t^{(d)}$ is finally decoded from the latent to a binary mask image \mathbf{M} with a threshold θ .

DiffEdit uses DDIM inversion to invert the input image \mathbf{x}_0 into the initial noise \mathbf{x}_T which can generate \mathbf{x}_0 . Each inversion step is calculated as:

$$\mathbf{x}_{t+1} = \sqrt{\alpha_{t+1}} \cdot f_\theta(\mathbf{x}_t, \mathbf{c}, t) + \sqrt{1 - \alpha_{t+1}} \cdot \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t). \quad (6)$$

We iteratively apply this formula until we obtain \mathbf{x}_T . However, if we stop the inversion step at timestep $r \leq T$, we encode \mathbf{x}_0 into a less noised version \mathbf{x}_r . r is called the encoding ratio as in DiffEdit. A larger value of r indicates a stronger edit effect, making the edited image guided more by the edited caption but less similar to the input image.

C Implementation details

C.1 Example of task template

We show one task template we provide to ChatGPT:

For example, if the user says “Change the dog to a cat”, you need to give the segmentation model only the keyword “Dog”. You also need to give the image editing model two text prompts: “Photo of a dog”, and “Photo of a cat”. Your answer should be in the form of: Segmentation prompt: Dog. Editing prompt 1: “Photo of a dog”. Editing prompt 2: “Photo of a cat”.

Here, *Editing prompt 1* is the input caption $\mathbf{c}^{(i)}$ and *Editing prompt 2* is the edited caption $\mathbf{c}^{(e)}$.

C.2 Baselines

For MDP- ϵ_t , we fix the interpolation factor as 1 as default and vary the editing starting timestep and ending timestep; For InstructPix2Pix, we only tune the classifier-free guidance factor for text

condition; For DiffEdit and InstructEdit, as they both share the same generation process after obtaining a mask, we tune the encoding ratio r . For DiffEdit we also tune the additional parameter θ which controls the threshold when computing a binary mask. For InstructPix2Pix and InstructEdit, we use the same user instruction as input, while for MDP- ϵ_t and DiffEdit we manually design the input caption and edited caption based on the user instruction.

For MDP- ϵ_t , we use the official implementation from <https://github.com/QianWangX/MDP-Diffusion>. For InstructPix2Pix, we use the official implementation from <https://github.com/timothybrooks/instruct-pix2pix>. For DiffEdit, as there is no official implementation available, we refer to https://github.com/johnrobinson/diffusion_experiments/blob/main/DiffEdit.ipynb and modify the parts that are not consistent with the paper [10].

D More results

D.1 Qualitative and quantitative comparison

We show more qualitative results in Figs. 12 and 13. We show that InstructEdit performs better than the baseline methods in the majority of cases.

We also provide a additional user study for the 20 examples shown in Figs. 12 and 13. We presented triplets of images as done in the main paper, each triplet consisting of one input image and two edited images. One of the edited images was the output from our method, while the second one was a randomly picked edited image from one of the baselines methods. For every example, we ask the participant a question: “*Which image applied the instruction more appropriately?*”. Results show that InstructEdit was preferred over MDP- ϵ_t , InstructPix2Pix and DiffEdit in 67.5%, 61.0%, and 57.0% of the cases, respectively. We show a screenshot of the user study interface in Fig. 11.

D.2 Mask improvement

We show more results for mask improvement of InstructEdit over DiffEdit in Fig. 14. The quality of the automatic masks generated by DiffEdit is highly affected by the mask threshold θ . Nevertheless, the mask area generated by DiffEdit can be completely off the region that should be edited under different θ . With the help of the grounding model and segmentation model, however, the generated mask can accurately outline the region to be edited without tuning θ .

D.3 Encoding ratio

We show the effect of increasing the encoding ratio and compare the results with an inpainting baseline in Fig. 7. The inpainting baseline from <https://github.com/IDEA-Research/Grounded-Segment-Anything/tree/main> is also an image editing framework, which also adopts the Grounded Segment Anything to extract a high-quality mask for the input image, but uses the Stable Diffusion Inpainting model <https://huggingface.co/runwayml/stable-diffusion-inpainting> as the image editing model.

In general, increasing the encoding ratio can lead to a larger change in the edited image compared to the original image. The edited image will follow the user instruction more and look less like the input image. Compared to using an inpainting model as an image editing model, using the mask-guided generation enables more flexibility in choosing either a larger change or a smaller change. An inpainting model however will completely ignore the original pixel information inside the mask region and only follow the inpainting prompt.

E Failure cases

We identify one common failure case for each component in our framework and show those cases in Figs. 8 to 10. For the language processor, BLIP2 sometimes may provide a description that is not helpful to the editing task. For example in Fig. 8, describing the brand of the car does not help change the color of the target car. For the segmenter, Grounded DINO could fail to correctly locate the object given a certain direction. In Fig. 9, the object to be edited should be the dog on the right, but Grounded DINO gives a larger probability score to the dog on the left and therefore segmenter

produces the wrong mask. For the image editor, as all the editing are performed within the mask, the editor is not good at deforming the object as shown in Fig. 10.

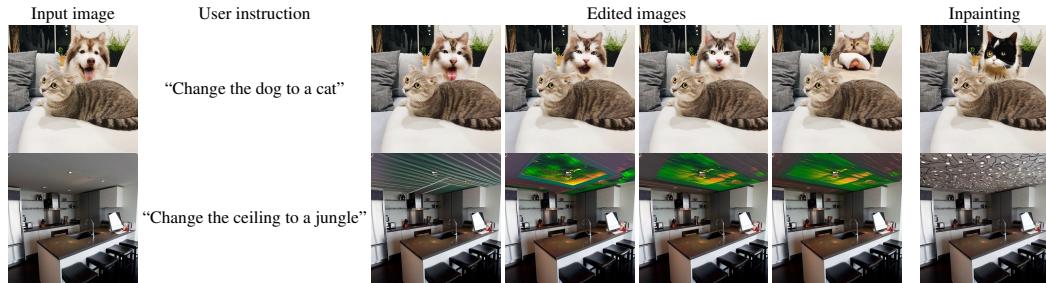


Figure 7: Effect of increasing encoding ratio and comparison with an inpainting baseline. From left to right the edited images are edited by increasing encoding ratios. Note that the inpainting baseline does not accept user instruction as input, but only inpainting prompt.

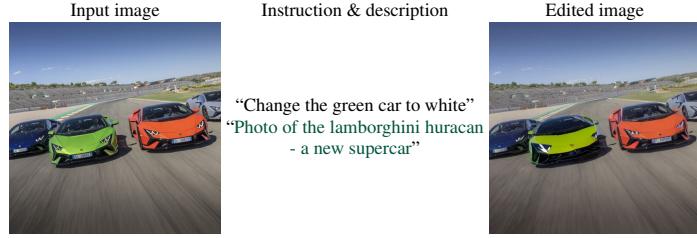


Figure 8: Failure case of the language processor. Note that we show the **BLIP2** description below the user instruction. The generated BLIP2 description is irrelevant to the editing task.

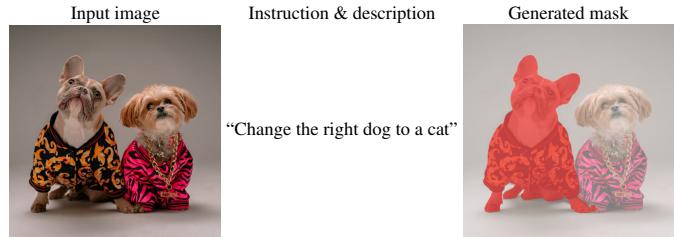


Figure 9: Failure case of the segmenter. The segmenter here fails to mask the dog on the right side.

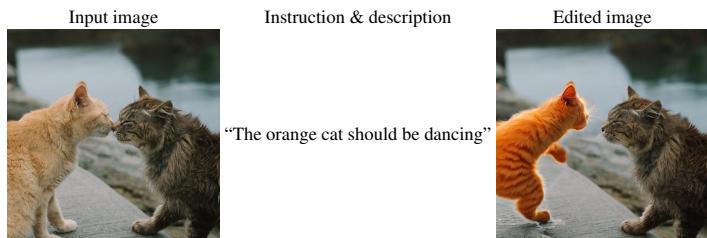


Figure 10: Failure case of the image editor. The edited cat fails to preserve the identity of the cat in the original image.



Figure 11: Screenshot of the user study interface.

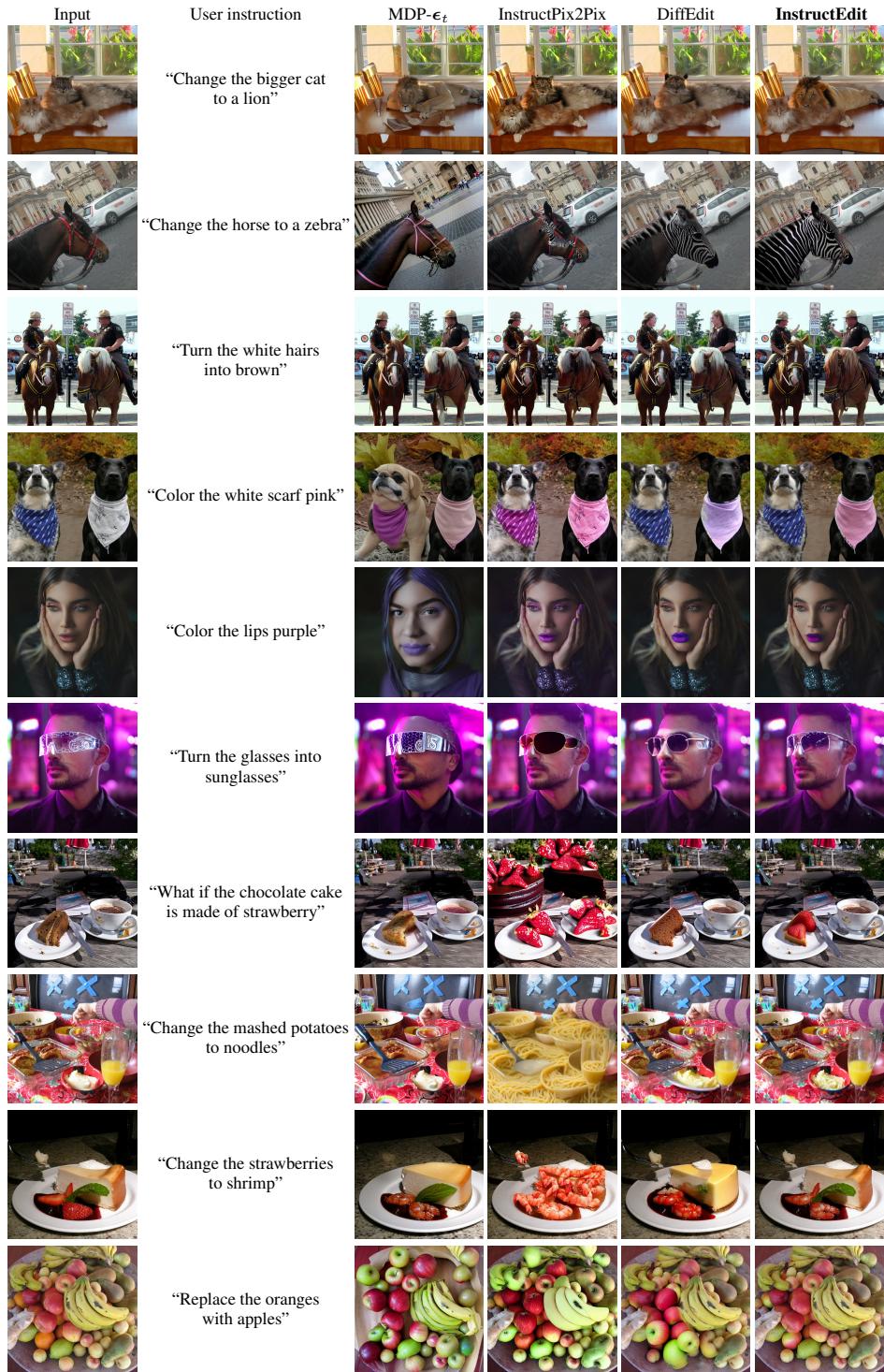


Figure 12: Qualitative results of baselines and our method.

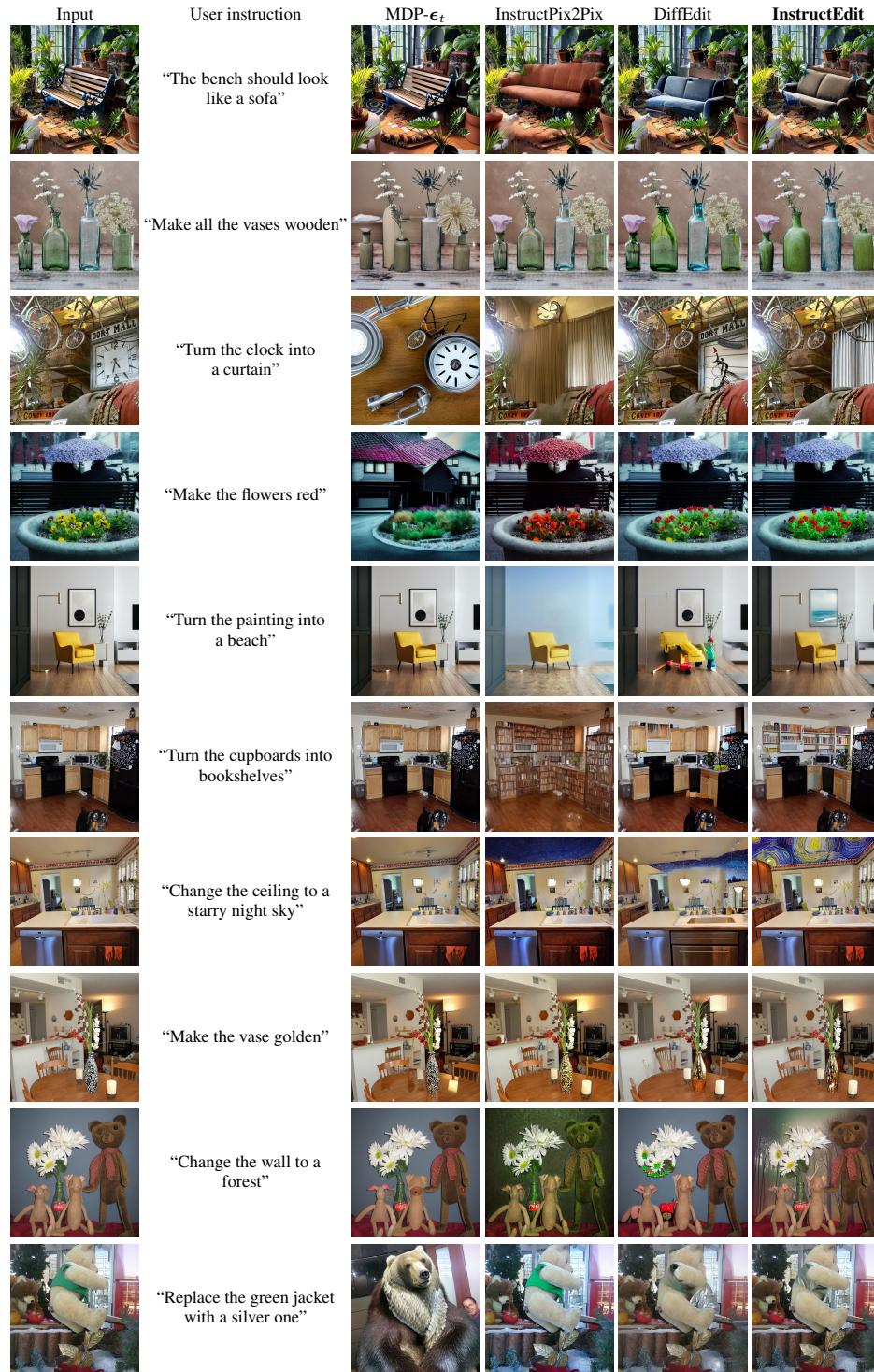


Figure 13: Qualitative results of baselines and our method.

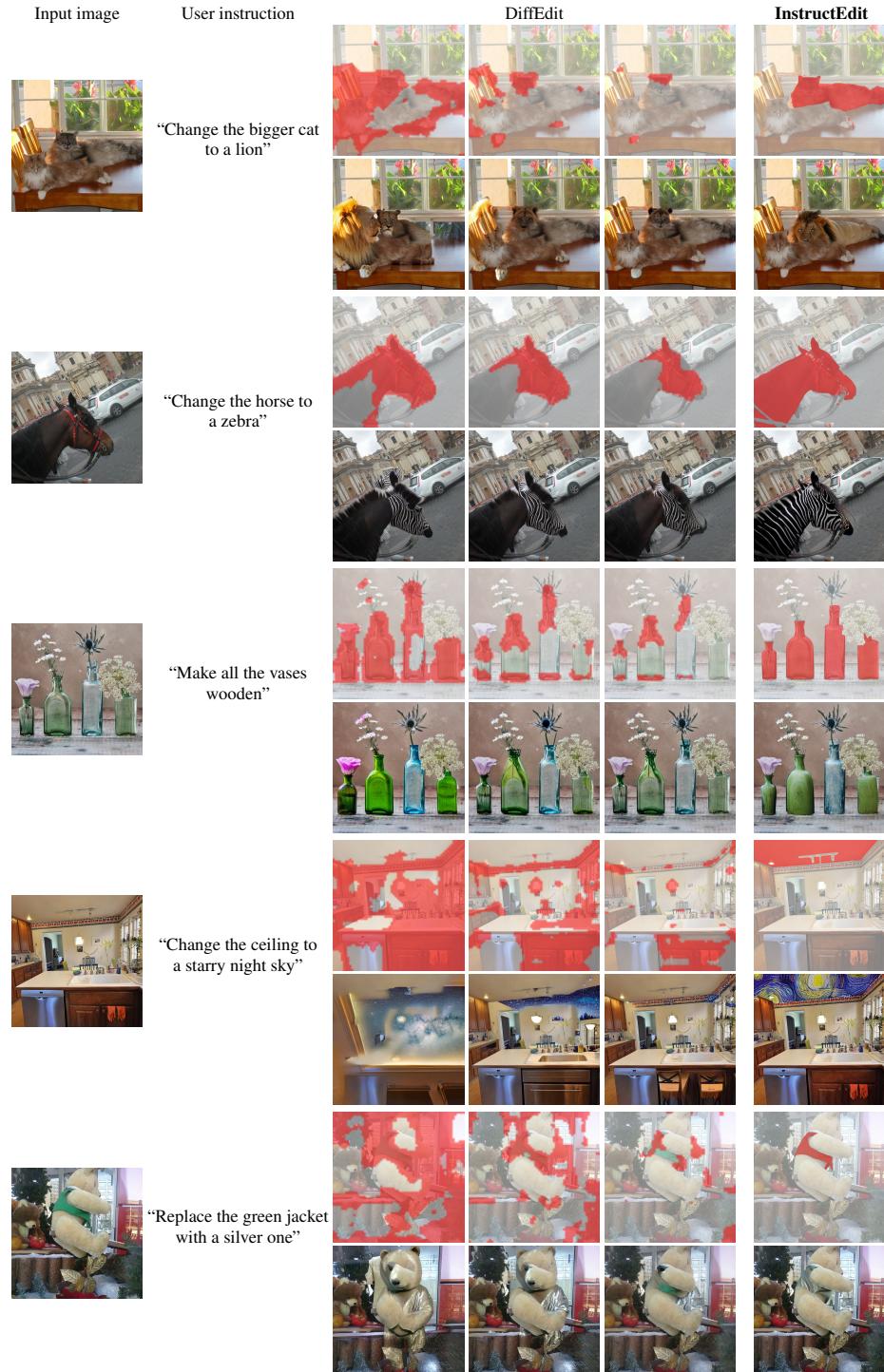


Figure 14: Comparison of the masks (colored in red and blended with the input image) and the corresponding edited image (below each mask) generated by DiffEdit and InstructEdit.