# Image Inpainting Models are Effective Tools for Instruction-guided Image Editing

**Solution for GenAI Media Generation Challenge Workshop @ CVPR**

Xuan Ju[1,2], Junhao Zhuang[1], Zhaoyang Zhang[1], Yuxuan Bian[1,2], Qiang Xu[2]
[1]ARC Lab, Tencent PCG, [2]The Chinese University of Hong Kong

https://github.com/TencentARC/BrushNet/tree/main/InstructionGuidedEditing

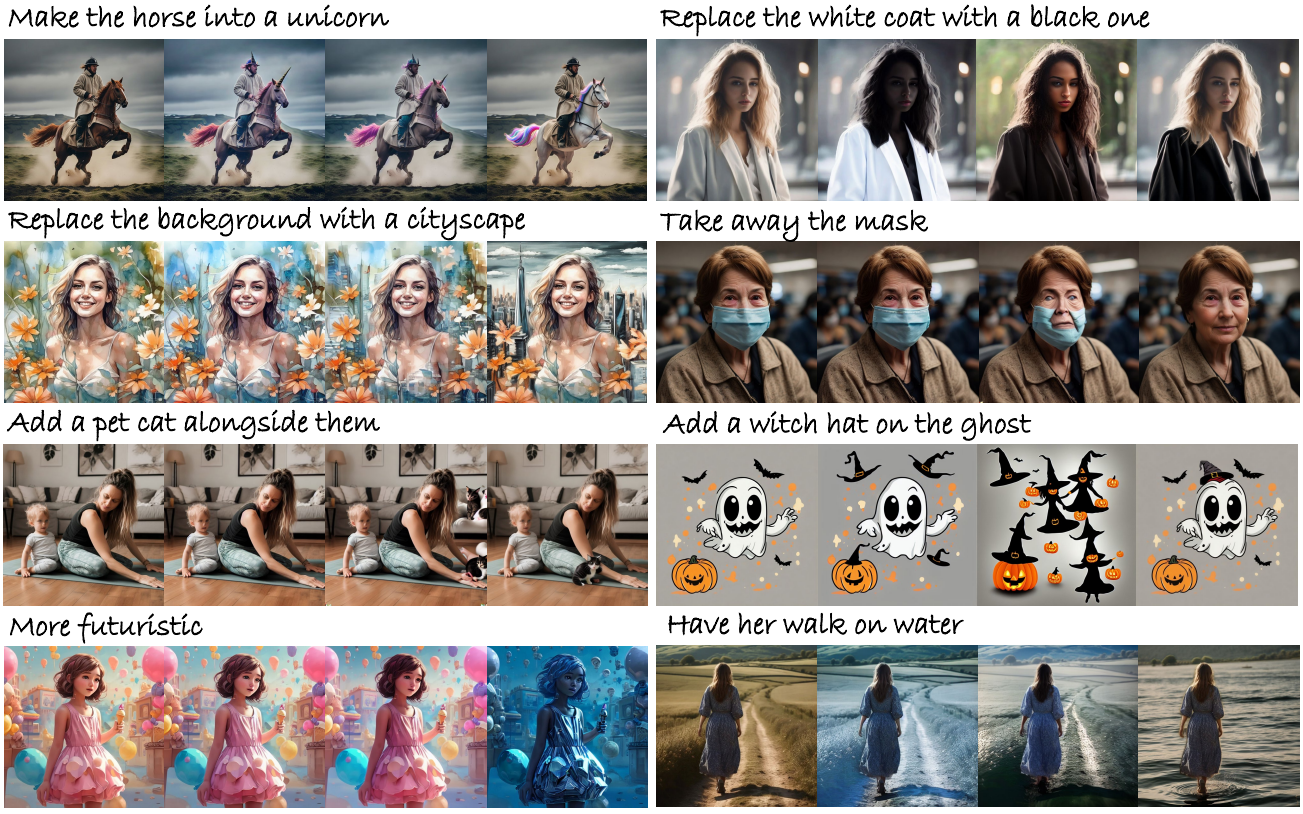


Figure 1: **The comparison of Previous Text-Guided Image Editing Methods and Ours.** Text prompt is shown at the top of each group of images. Images include editing category of local edit, background edit, global edit, addition, and remove.

## Abstract

*This is the technique report for the winning solution of the CVPR2024 GenAI Media Generation Challenge Workshop's Instruction-guided Image Editing track. Instruction-guided image editing has been largely studied in recent years. The most advanced methods, such as SmartEdit and MGIE, usually combine large language models with diffusion models through joint training, where the former provides text understanding ability, and the latter provides image generation ability. However, in our experiments, we find that combining large language models and image generation models through intermediary mask guidance instead of fine-tuning leads to a better editing performance and success rate than previous methods. We use a 4-step process, **IIIE** (**I**npainting-based **I**nstruction-guided **I**mage **E**diting): editing category classification, main editing object identification, editing mask acquisition, and image inpainting. Results show that through proper combinations of language models and image inpainting models, our pipeline can reach a high success rate with satisfying visual quality.*

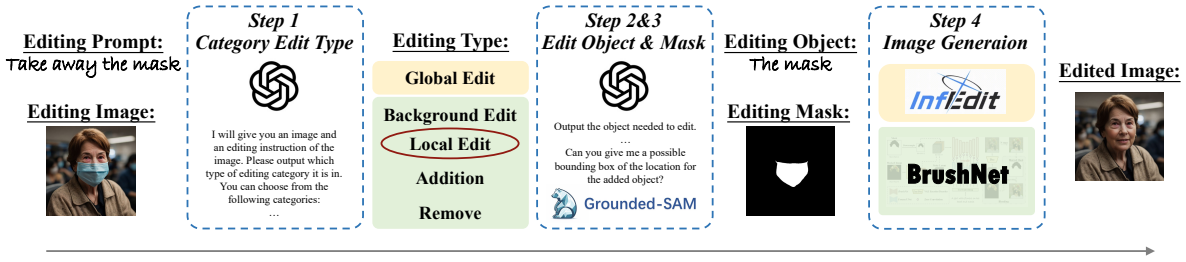


Figure 2: **The pipeline of our proposed 4-step image editing process ITIE.**

## 1. Solution

In the wake of the rapid advancements in generative model capabilities, the field of text-guided image generation and editing has seen unprecedented progress. This has resulted in creating images with unparalleled quality, aesthetics, and adherence to text guidance. Yet, a significant challenge remains the absence of a universally accepted, easily accessible benchmark for evaluating the in-depth capabilities of the generated models. This issue stems from the lack of a comprehensive, large-scale evaluation dataset, standardized evaluation protocols, and the insufficiency of current automatic metrics.

This technique report is the winning solution for The GenAI Media Generation Challenge (MAGIC), which moves us towards addressing these issues. The MAGIC hosts two challenge tracks: (1) text-to-image generation and (2) text-guided image editing. This technique report presents the solution for the second track, text-guided image editing. We list the instructions for this track here:

**Guidelines** For text-guided image editing, we test the capacity of the model to change a given image's contents based on some text instructions. The specific type of instructions that we test for are the following:

- Addition: Adding new objects within the images.
- Remove: Removing objects
- Local: Replace local parts of an object and later the object's attributes, i.e., make it smile
- Global: Edit the entire image, i.e., let's see it in winter
- Inpaint: Alter an object's visual appearance without affecting its structure
- Background: Change the scene's background

**Evaluation Protocol** We leverage both human and automated evaluations. In human-based evaluations, we use human annotators. We mainly evaluate the following aspects:

- Edit faithfulness - whether the edited image follows the editing instruction
- Content preservation - whether the edited image preserves the regions of the original image that should not be changed
- Visual quality - whether the edited image is artifact-free, keeping the core visual features of the original image, etc

On automatic evaluation, similar to the text to image track, we will leverage existing methods that have developed automatic metrics to help in assessing the outputs of the image based on the prompt and instruction.
To determine winners, we use automatic evaluation to help prune the total number of entries to 10 finalists. At 10, we would use human annotation and evaluation to determine the final winners.

Previous instruction-guided image editing methods include two categories: (1) diffusion model finetuning (*e.g.*, InstructPix2Pix [1] and InstructDiffusion [3]), and (2) large language model (LLM)+diffusion model finetuning (*e.g.*, SmartEdit [4] and MGIE [2]) [1]. These methods either not include LLMs in the model, or use joint fine-tuning of LLMs and diffusion models with the LLMs' capabilities not being fully unleashed, leading to a weak understanding of instructions. Consequently, these methods show low success rates and unsatisfying results. Contrary to these methods, we find that a simply tool-based combination of LLMs and diffusion models can lead to a much better visual results. This comes from a full utilization of the language understanding ability of LLMs.

In this competition, we propose a 4-step process for instruction-based image editing, IIIE (Inpainting-based

[1]Noted that instruction-guided image editing is a subclass of text-guided image editing [5].

Instruction-guided Image Editing). Firstly, we use GPT4-o to categorize the current editing instructions into one of the editing categories: Local Edit, Background Edit, Global Edit, Addition, and Remove. Then, we find the main editing object by making further conversations with GPT4-o based on the editing category. After that, we use Grounded-SAM combined with GPT4-o to obtain the editing mask. Finally, we use image generation model to perform image editing based on a target prompt generated by GPT4-o. The pipeline of our solution is shown in Fig. 2. All code and generated results are released in our GitHub code. Visualization comparison in Fig. 1 show a higher success rate and visual quality of IIIE compared to previous methods. We hope our findings can offer some insights for relevant field.

## 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

[2] 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] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12709–12720, 2024. 2

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

[5] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 2