EditAnything: Empowering Unparalleled Flexibility in Image Editing and Generation

Shanghua Gao
Nankai University, Sea AI Lab
Tianjin, China
shgao@mail.nankai.edu.cn

Zhijie Lin
Sea AI Lab
Singapore
linzhijie@zju.edu.cn

Xingyu Xie
Peking University, Sea AI Lab
Beijing, China
xyxie@pku.edu.cn

Pan Zhou
Sea AI Lab
Singapore
zhoupan@sea.com

Ming-Ming Cheng
Nankai University
Tianjin, China
cmm@nankai.edu.cn

Shuicheng Yan
BAAI, Skywork AI
Beijing, China
shuicheng.yan@gmail.com

ABSTRACT

Image editing plays a vital role in computer vision field, aiming to realistically manipulate images while ensuring seamless integration. It finds numerous applications across various fields. In this work, we present EditAnything, a novel approach that empowers users with unparalleled flexibility in editing and generating image content. EditAnything introduces an array of advanced features, including cross-image dragging (e.g., try-on), region-interactive editing, controllable layout generation, and virtual character replacement. By harnessing these capabilities, users can engage in interactive and flexible editing, giving captivating outcomes that uphold the integrity of the original image. With its diverse range of tools, EditAnything caters to a wide spectrum of editing needs, pushing the boundaries of image editing and unlocking exciting new possibilities. The source code is released at https://github.com/sail-sg/EditAnything.

ACM Reference Format:
Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. EditAnything: Empowering Unparalleled Flexibility in Image Editing and Generation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3581783.3612680

1 INTRODUCTION

Image editing is a computer vision fundamental task, playing a crucial role in various applications, e.g. photography and graphic design [3, 5, 7, 8, 13]. The ability to manipulate and enhance images has revolutionized the way we perceive and interact with visual content. With the advancement of generative models, e.g. GANs and diffusion models [1, 2, 9–12], image editing has become more accessible to both professionals and amateurs alike. However, despite the progress made in this field, there are still significant challenges that need to be addressed.

One of the primary challenges in image editing is achieving realistic and visually pleasing results while maintaining the integrity and authenticity of the image [3, 5, 7, 8, 13]. For instance, when applying modifications like object replacement or image retouching, it is essential to ensure that the edited image seamlessly blends with the surrounding content, without leaving significant traces of manipulation. Achieving this level of realism often requires sophisticated algorithms that can handle complex image structures, lighting conditions, and diverse visual content. Additionally, the increasing demand for real-time and interactive image editing further adds to the challenge, as it requires efficient algorithms capable of processing images quickly and accurately. Addressing these challenges is crucial to advancing the field of image editing and enabling a wide range of applications in visual media production and creative industries.

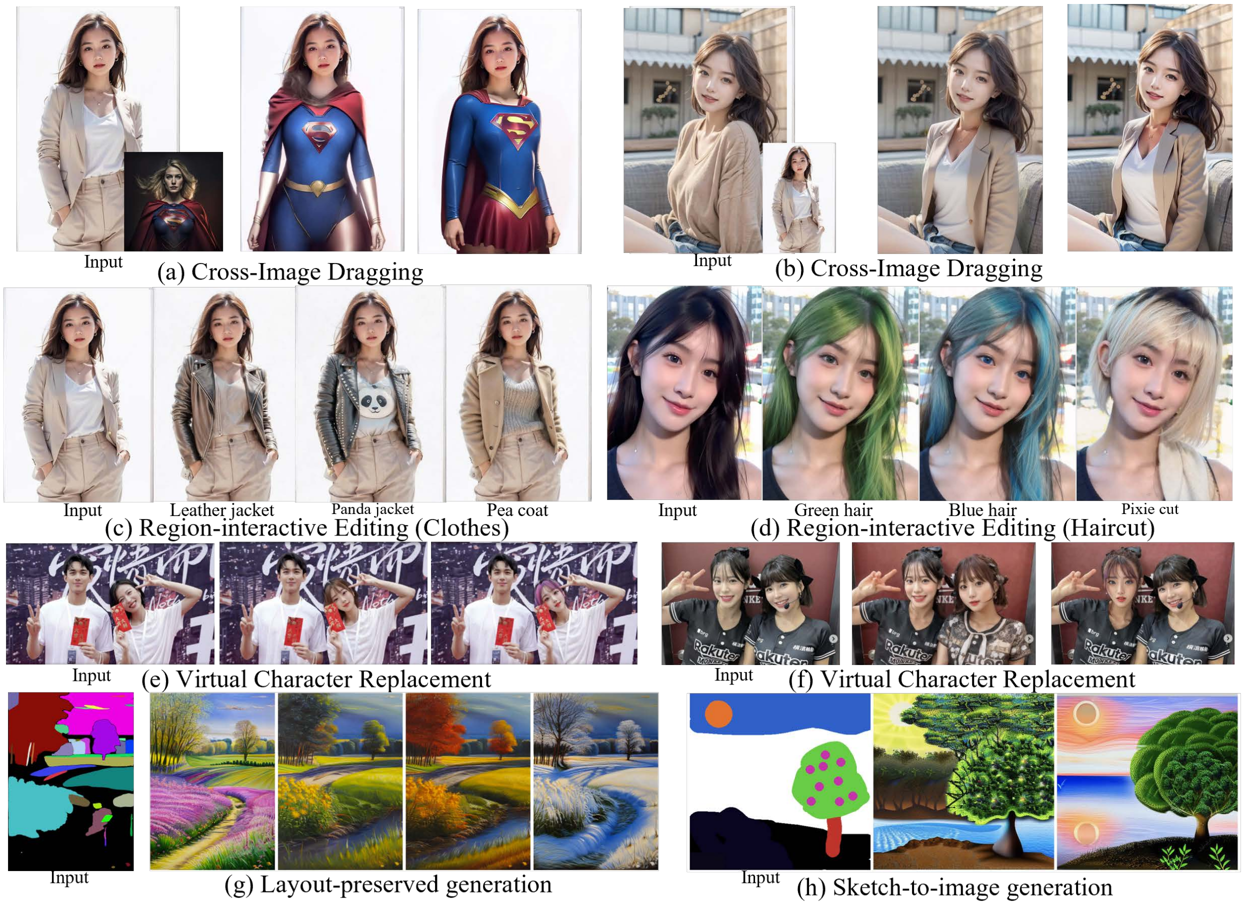
EditAnything is specifically designed to empower users with unparalleled freedom in editing and generating image content. This remarkable approach incorporates a myriad of powerful features, including cross-image dragging (e.g., try-on), region-interactive editing via text commands and mouse-clicking, controllable layout generation, and virtual character replacement for specific image object characters. These functionalities enable users to engage in image editing with unparalleled creativity and flexibility, giving captivating effects and innovations while preserving the integrity and authenticity of the image. EditAnything provides an extensive range of tools and features, allowing users to effortlessly adjust image elements or create entirely new content. With EditAnything, users can achieve highly realistic and visually pleasing results, catering to diverse editing needs and pushing the boundaries of image editing possibilities.

2 METHOD

The large-scale text-to-image diffusion models [6–8] have taken an enormous step of image generation in both quality and flexibility. To further enable a fine-grained layout control, we build the novel EditAnything upon the pre-trained Stable Diffusion [7], and add the segmentation mask generated by Segment Anything Model (SAM) [4] as the extra condition using the scheme of ControlNet [13]. Therefore, based on EditAnything, we can generate high-quality images or make specific regional editing while maintaining the given layout. Based on the inpainting model, we keeps the original image feature of unchanged region and applies Tile refinement [13] to ensure the realistic and quality of edited image. For the training-free image dragging and merging, we then utilize the

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612680



(a) Cross-Image Dragging

(b) Cross-Image Dragging

(c) Region-interactive Editing (Clothes)

(d) Region-interactive Editing (Haircut)

(e) Virtual Character Replacement

(f) Virtual Character Replacement

(g) Layout-preserved generation

(h) Sketch-to-image generation

Figure 1: EditAnything (controllable layout generation includes layout-preserved generation and sketch-to-image generation)

reference image as another control and employ the cross-attention mechanism and the cross-region adaptive normalization to aggregate semantic information to edited regions, where the proposed layout control is a helpful constraint to keep the generated images consistent and harmonious.

3 EXPERIMENTAL RESULTS

Cross-Image Dragging. EditAnything redefines image editing with its groundbreaking cross-image dragging functionality, enabling seamless merging of regions from different images without training. Users can effortlessly blend objects or areas from one image into another, expanding their creative possibilities. Fig.1 (a) showcases EditAnything's ability to smoothly drag Superman's costume onto another person. In Fig.1 (b), EditAnything transcends pose differences, transferring a front-standing shirt from a reference image to a side-sitting person in the target image. This flexibility frees users to explore limitless potential and create stunning images beyond the constraints of individual photographs.

Region-Interactive Editing. EditAnything revolutionizes region-interactive editing. Leveraging image layout maps, EditAnything empowers users to seamlessly edit and generate image regions. Using class-agnostic segmentation masks generated by the Segment Anything model, users can effortlessly manipulate image regions through text inputs or mouse-clicking. This interactive capability unleashes users' imagination, enabling easy editing of clothing, hairstyles, and more, as shown in Fig.1 (c) & (d).

Virtual Character Replacement. EditAnything, in conjunction with a specific LORA model, offers the capability of performing real-to-virtual character replacements, prioritizing privacy protection. This unique functionality caters to a wide range of image editing requirements while preserving personal privacy. With EditAnything's controllable image layout generation, users can seamlessly transfer facial features, body postures, and expressions to virtual characters while maintaining privacy. The Tile model enhances image details, consistency, and resolution, preserving the original style during the editing process, as shown in Fig.1 (e) & (f).

Controllable Layout Generation. EditAnything's controllable image generation is based on layout, leveraging the Segment Anything Model (SAM) to generate class-agnostic segmentation masks. By training a ControlNet using SAM, EditAnything gains the ability to control generative models like StableDiffusion, allowing for controllable layout-based image editing and generation. Edited images maintain structural consistency while adhering to the original layout, as illustrated in Fig.1 (g). Users can modify the seasonal style of an image while preserving the layout, such as plants. In Fig.1 (h), EditAnything also enables the generation of corresponding images based on user-provided layout sketches.

ACKNOWLEDGEMENT.

This research was supported by the NSFC (NO. 62225604) and the Fundamental Research Funds for the Central Universities (Nankai University, 070-63233089). The Supercomputing Center of Nankai University supports computation.

REFERENCES

[1] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. Masked diffusion transformer is a strong image synthesizer. In *International Conference on Computer Vision*.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[3] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

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

[5] Lorenzo Luzi, Ali Siahkoohi, Paul M Mayer, Josue Casco-Rodriguez, and Richard Baraniuk. 2022. Boomerang: Local sampling on image manifolds using diffusion models. *arXiv preprint arXiv:2210.12100* (2022).

[6] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[9] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

[10] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

[11] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

[12] Zike Wu, Pan Zhou, Kenji Kawaguchi, and Hanwang Zhang. 2023. Fast Diffusion Model. arXiv:2306.06991 [cs.CV]

[13] Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).