Edit Everything: A Text-Guided Generative System for Images Editing

Defeng Xie[*]　　　　**Ruichen Wang**　　　　**Jian Ma**

Chen Chen [†]　　　　**Haonan Lu** [†]　　　　**Dong Yang** [†]

OPPO Research Institute

Fobo Shi [*]　　　　　　　　　　**Xiaodong Lin**
Central China Normal University　　　　Rugster Unversity

Abstract

We introduce a new generative system called Edit Everything, which can take image and text inputs and produce image outputs. Edit Everything allows users to edit images using simple text instructions. Our system designs prompts to guide the visual module in generating requested images. Experiments demonstrate that Edit Everything facilitates the implementation of the visual aspects of Stable Diffusion with the use of Segment Anything model and CLIP. Our system is publicly available at https://github.com/DefengXie/Edit_Everything.

1 Introduction

"While drawing I discover what I really want to say."

— Dario Fo

Visualization, including images, paintings, shots, illustrations, and photographs, can usually be described with text. However, creating these images often requires specialized skills and a significant time investment [18]. Consequently, a generative system has the potential to produce realistic images based on natural language, allowing humans to efficiently generate a wide range of visual content. Additionally, this system offers an unprecedented opportunity for continuous improvement and precise control over image editing, making it a crucial tool for real-world applications.

Recently, diffusion models have shown promising performances in generating high-quality realistic images [17, 14, 15, 12]. In particular, a text-guided method significantly improves the diversity and fidelity [10]. To address photorealism in the conditional setting, [2] proposes a classifier to guide diffusion models, allowing them to generate realistic images toward a classifier's label. [5] shows that guidance can be indeed performed by a pure generative model without a classifier, named classifier-free guidance. Classifier-free guidance combines conditional and unconditional score estimates to attain a trade-off between sample quality and diversity similar to that with classifier guidance. Inspired by the ability of guided diffusion models with text, [10] first implements CLIP to guide diffusion models towards text prompt [11]. However, compared to the quality of generated images,

---

[*]Equal Contributions.

[†]Correspondence to: Chen Chen <chenchen4@oppo.com>, Haonan Lu <luhaonan@oppo.com>, Dong Yang <yangdong1@oppo.com or dongyang3-c@my.cityu.edu.hk>.
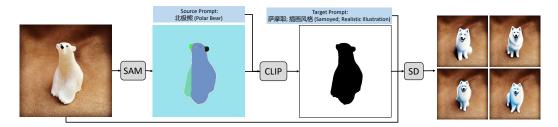
Figure 1: The network architecture of Edit Everything. The original image is separated into several segments with the help of Segment Anything model (SAM). Next, These segments are ranked based on the source prompt, and the target segment is chosen based on the highest score calculated by our trained CLIP model. The source prompt is a text that describes the target object and editing styles. Finally, guided by the target prompt, Stable Diffusion (SD) generates the replacement object for the mask segment. This process is seamless and efficient, resulting in high-quality image editing.

CLIP guidance cannot compete with classifier-free guidance. Another promising solution is that [20] designs a neural network structure (ControlNet) to control diffusion models and support conditional inputs. Diffusion models can be augmented by ControlNets to generate edge maps, segmentation maps, key points, etc.

Inspired by the remarkable performance of ControlNet and CLIP guidance in significantly enhancing the image quality, we leverage Segment Anything model (SAM) and CLIP to guide diffusion models [11, 8]. In our work, we create a text-guided generative system called Editing Everything, combining SAM, CLIP and Stable Diffusion (SD) [14]. With this framework, we aim to improve the quality of generated images, and provide an efficient and accurate tool for researchers and practitioners in various domains. First, we train a CLIP with 400 million parameters and a SD with 1 billion parameters for Chinese scenarios. Our trained models power Editing Everything, equipping it with the ability to understand Chinese text prompts and guide diffusion models towards text prompts. Second, Editing Everything can readily utilize text prompts for zero-shot generation, while it has difficulty creating realistic images for complex prompts. To address this problem, we propose a solution that this system breaks down complex prompts into smaller entities, which are then replaced in a sequential manner. The system also facilitates iterative sample generation until it matches complex prompts, proving advantageous for users. Third, by leveraging Editing Everything, users can efficiently modify images with different styles and objects.

2 Methods

2.1 Architecture

The text-guided generative system, Editing Everything, is composed of three main components: Segment Anything Model (SAM), CLIP, and Stable Diffusion (SD). SAM is used to extract all segments of an image, while CLIP is trained to rank these segments based on a given source prompt. The source prompt describes the interested object. The highest scoring segment is then selected as the target segment. Finally, SD is guided by a target prompt to generate a new object to replace the selected segment. This allows for a precise and personalized method of image editing.

For Chinese scenarios, our CLIP is pre-trained on Chinese corpus and image pairs. And Stable Diffusion is also pre-trained on Chinese corpus. Our generative system has successfully generated realistic images on Chinese scenarios. However, for English scenarios, we implement available open-source CLIP and Stable Diffusion.

2.2 Pre-training Data

Table 1 shows that our pre-training dataset consists of several open sources. Our trained models only use the Chinese corpus and related images. Here are details of these datasets:

Wukong. Wukong is the largest Chinese cross-modal dataset, an enormous collection of 100 million Chinese image-text pairs from the web [4]. This dataset is designed for large-scale experiments and
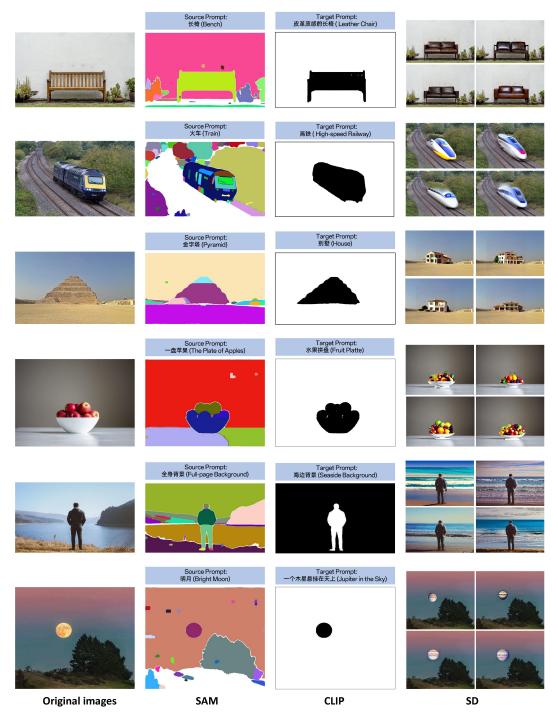
Figure 2: Text-guided image editing examples created by Edit Everything. Our advanced system detects the dark region, and erases them by the source target. And then we apply SD to fill it based on the target prompt. Our system is able to produce various styles and seamlessly match the surrounding context.

serves as a valuable resource for researchers in language-vision fields. In our training, we use 100 million image-text pairs.

Zero and R2D2. Zero and R2D2 is a large-scale cross-modal dataset, which contains the public dataset ZERO-Corpus and five human-annotated datasets designed for downstream tasks [19]. It has
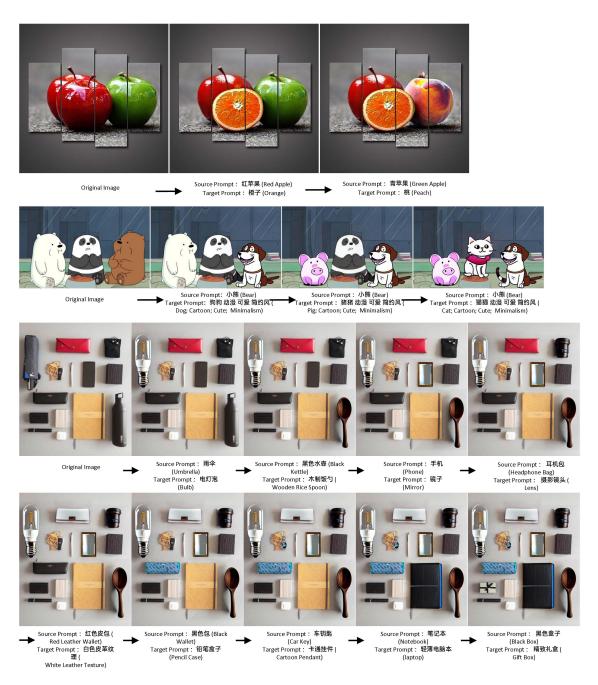
Figure 3: Iteratively replacing objects of an image step by step using Editing Everything.

250 million images paired with 750 million text. Based on Zero and R2D2, we can extract 23 million high-quality Chinese image-text pairs for our training.

Laion-5b. Laion-5b is a large-scale multi-modal dataset consisting of 5.85 billion CLIP-filtered image-text pairs, of which 2.32B are English language [16]. We have processed 80 million Chinese image-text pairs by filtering and pre-processing the data.

Crawled Data. We have crawled an extensive collection of 100 million image-text pairs from the web. This dataset covers various domains, such as shopping, common knowledge, tools, etc. However, we regret to inform that due to commercial restrictions, we cannot release this data to the public.

4

Table 1: Statistics of pre-training datasets.

Dataset	Proportion (%)	Number of image-text pairs (million)
Wukong	33.0	100
Zero and R2D2	7.6	23
Laion-5b	26.4	80
Crawled Data	33.0	100

Figure 4: Comparisons of images generated by open-source models and our trained models. Our models could support Chinese inputs.

2.3 Implementation

CLIP is composed of a text encode and an image encode. The text encode is a Chinese BERT with the WordPiece [1]. The image encode utilizes a ViT-L/14 trained by OpenAI's CLIP [3, 11]. Stable Diffusion is a powerful model that combines VAE [7, 13], UNet [2], and text encoder to achieve exceptional performance. The VAE parameters are kept frozen during the training process. The UNet is designed to learn a data distribution, $p(x)$, by gradually denoising a normally distributed variable, allowing it to learn the reverse process of a fixed Markov Chain of length T.

Our trained models use the AdamW optimizer [9], with the following hyper parameters: $\beta_1 = 0.9, \beta_2 = 0.99$. To optimize our training process, we implement a linear learning rate schedule, and apply a weight decay of 0.1 and gradient clipping at 1.0. In addition, warm-up steps are set to 2000 and the batch size is 180. To reduce computation costs, we run them on 40 Tesla V100s.

3 Main Results

3.1 Simple Prompts

In Figure 2, we observe that Editing Everything with text guidance is capable of editing any object within an image and generating a diverse range of realistic images. It can also seamlessly match different styles of illustrations, such as realistic or painted styles. In editing tasks, Editing Everything can effectively modify existing images using source prompts, and insert new objects, shadows, and reflections based on target prompts. The results are highly realistic.

3.2 Complicated Prompts

In Figure 3. we demonstrate the iterative creation of images using Editing Everything based on complex prompts. This process involves the step-by-step replacement of source objects with target objects. While this pipeline may not be efficient, it produces highly accurate control.

3.3 Further Comparisons

To further demonstrate the performance of our system, we present a comparison between our trained models and open-source models in Figure 4. Our trained models consist of open-source SAM, trained CLIP, and trained Stable Diffusion. In contrast, open-source models contain open-source SAM, Taiyi-CLIP [6], and trained stable diffusion [14]. Notably, our trained models are capable of accepting Chinese text inputs. Moreover, due to crawled high-quality image-text datasets, our models outperform open-source models.

4 Limitations

Our generative system, Editing Everything, consists of SAM, CLIP, and SD models. Their architecture is not modified in any way. To enhance their performance in Chinese scenarios, we trained these models on our own crawled dataset.

5 Conclusion

In this paper, we propose a new generative system, namely Editing Everything, to design text prompts to help Stable Diffusion to generate images toward target prompts. This generative system implements Segment Anything Model, CLIP, Stable Diffusion. Based on our trained models, this work firstly provides an efficient solution for Chinese scenarios. Compared to open-source models, our generative system achieve a great performance for image editing.

6 Acknowledgements

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35: 26418–26431, 2022.

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[6] IDEA-CCNL. Fengshenbang-lm. `https://github.com/IDEA-CCNL/Fengshenbang-LM`, 2021.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[8] 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.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[10] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[11] 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.

[12] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[13] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

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

[15] 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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

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

[18] Sudheendra Vijayanarasimhan and Kristen Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2262–2269. IEEE, 2009.

[19] Chunyu Xie, Heng Cai, Jianfei Song, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, et al. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework. *arXiv preprint arXiv:2205.03860*, 2022.

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