SEEDEDIT: ALIGN IMAGE RE-GENERATION TO IMAGE EDITING

Yichun Shi[*], **Peng Wang**[*], **Weilin Huang**
Seed Team, ByteDance

ABSTRACT

We introduce SeedEdit, a diffusion model that is able to revise a given image with any text prompts. In our perspective, the key to such a task is to obtain an optimal balance between maintaining the original image, i.e. *image reconstruction*, and generating a new image, i.e. *image re-generation*. To this end, we start from a weak generator (text-to-image model) that creates diverse pairs between such two directions and gradually align it into a strong image editor that well balances between the two tasks. SeedEdit can achieve more diverse and stable editing capability over prior image editing methods, enabling sequential revision over images generated by diffusion models. Our website is `https://team.doubao.com/seededit`.

1 INTRODUCTION

Today's diffusion models can create realistic and diverse images from only text descriptions. However, these generated images are usually quite uncontrollable and to some extent, the generation process is like throwing a dice until one sees a good output. To obtain more controllability over the generated content, a desired feature is instructional image editing, *i.e.* revising an input image with text descriptions. This can be regarded as a intersection between image generation and image understanding, both of which are quite mature today. Yet to this date, the technology of image editing itself still falls far behind both generation and understanding.

Existing image editing for diffusion models can be roughly categorized into two types. Firstly, the training-free methods combine specific techniques such as DDIM Inversion (Nichol et al., 2021; Mokady et al., 2023), test-time fine-tuning (Ruiz et al., 2023; Kawar et al., 2023), attention control (Cao et al., 2023; Hertz et al., 2022) to reconstruct an input image and generate a new one with the new text guidance. But since both the reconstruction and the re-generation process suffer from instability, the combination of these two accumulates into more error into the edited image, which could be inconsistent with either the input image or the target description.

The second type of methods are data driven approaches, where a large-scale pairwise editing dataset is prepared to train a instructional diffusion model (Brooks et al., 2023; Zhang et al., 2024; Hui et al., 2024; Wasserman et al., 2024; Zhao et al., 2024). The main difficulty here, however, is to prepare a diverse and high-quality editing dataset. Unlike image datasets that can be massively collected from the Internet, image editing pairs are very rare and it is almost impossible to collect a high-quality dataset that covers all types of editing pairs. So existing works attempt to use certain tools, such as Prompt-to-Prompt (Hertz et al., 2022) or in-painting to create such a dataset. But consequently, their performance is limited by these data creation tools, who themselves are not satisfying either.

To overcome the above mentioned difficulties, we introduce a new framework to convert an image generation diffusion model to one that edits images. We recognize that image editing is essentially a balance between image reconstruction and re-generation, and hence we develop a pipeline that first generates diverse pairwise data that scatters into these two directions, and then gradually align a image-conditioned diffusion model to arrive at an optimal balance between these two tasks. Overall, it leads to a model that is capable of revising images with either instructions or descriptions, which we call SeedEdit, and yields superior performance compared to prior studies.

[*]Equal Contribution

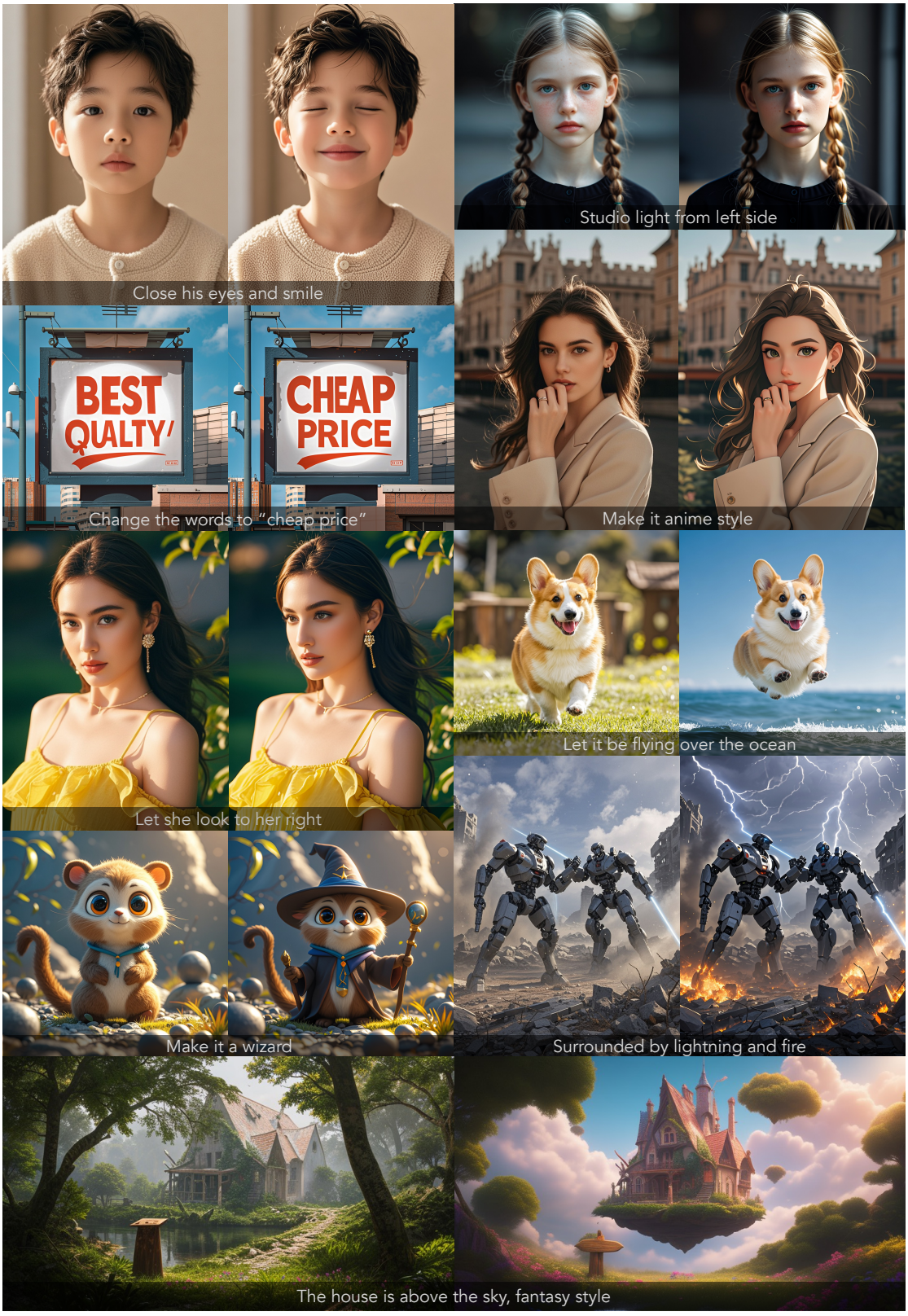
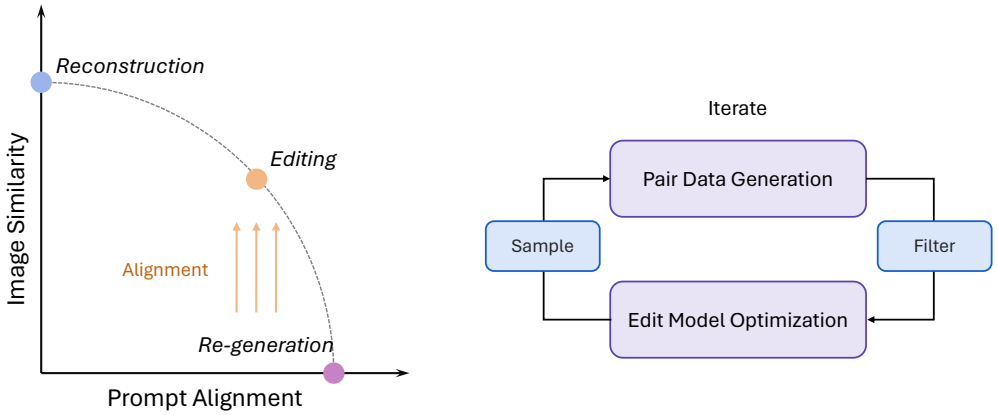


Figure 1: Example images edited by our method with one unified model and instructions only.



(a) Objective Illustration (b) Optimization

Figure 2: Overview of SeedEdit framework. We align a T2I model as lower bound for editing by improving image consistency. **Right**: Our optimization pipeline, we have an init edit model based on T2I and then iteratively conduct data sampling and model optimization to reach the optimal balance.

2 SeedEdit

The core difficulty of the image editing problem is the scarcity of pairwise image data. We address this problem from an alignment perspective. In particular, we regard text-to-image (T2I) model as a weak editing model, which achieves "editing" by generating a new image with a new prompt. We then distill and align such a weak editing model into a strong one by maximally inherit the re-generation capability while improving image consistency, as shown in Figure 2.

2.1 T2I model for Editing Data Generation

Our initial editing data are generated using a pre-trained T2I model as an editing model, where a pair of images before and after editing can be generated with corresponding text descriptions, similar to InstuctPix2Pix Brooks et al. (2023). With such data, we could distill a T2I model into an image-conditioned editing model. However, such naive re-generation could lead to inconsistency between the two images. To improve consistency, there exist various approaches, such as prompt-to-prompt (Hertz et al., 2022; Brooks et al., 2023) and attention control Cao et al. (2023). However, these techniques can generate very limited types of pair data and can hardly cover all types of image editing. Therefore, we combine different re-generation techniques and parameters to create a much more diverse dataset. In particular, we generate a large-scale pairwise dataset with more randomness to ensure diversity, and then we apply filters to choose good examples for model training and alignment. Fig 3 illustrates that our aligned model performs much better than naive re-generation based on the CLIP metrics.

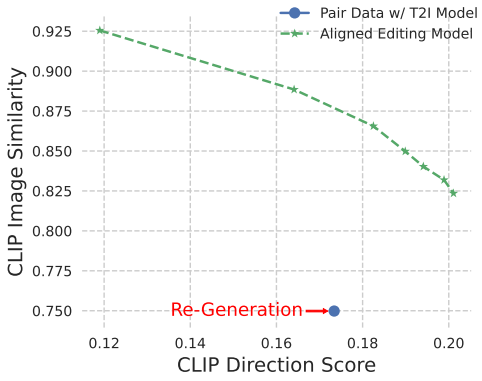
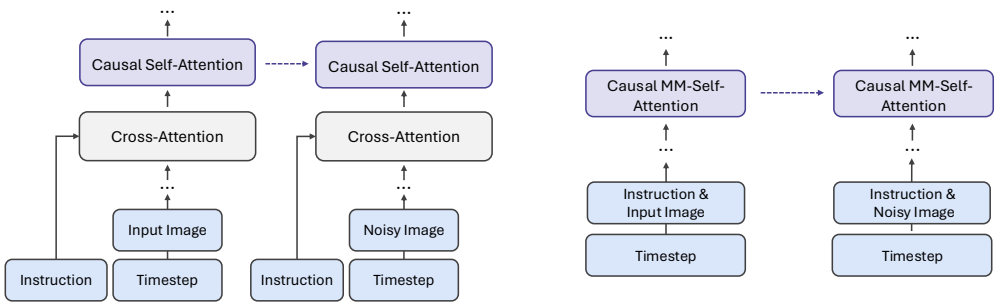


Figure 3: Our aligned editing model can achieve a similar or higher direction score (prompt alignment) with much higher image similarity compared to re-generation. The green curve is drawn by sampling different CFG for editing model.

2.2 Causal Diffusion Model with Image Input

The model architecture of our image-conditioned diffusion model is shown in Fig. 4. Unlike previous studies that add additional input channels for image conditioning (Brooks et al., 2023), we reuse self-attention for this purpose, where two branches of the diffusion model (shared parame-



(a) Causal UNet (b) Causal MM-DiT

Figure 4: Architecture of causal diffusion model for image conditioning. Two branches with shared parameters are applied to the input image and instructions, respectively.

ters) are applied to the input and output image, respectively. This is inspired by prior training-free methods (Cao et al., 2023) and we empirically found that such an architecture performs better on geometric deformation tasks and introduces fewer new parameters. Specifically, a causal self-attention structure is introduced such that two networks can build communications based on intermediate features. If we drop the input branch, it leads to the original T2I diffusion model, allowing for a mixed training on editing and T2I data.

2.3 Iterative Alignment

Because of the noisy dataset, the initial editing model trained on the pair of examples may not be sufficiently robust for applications. That is, like the dataset itself, the model is able to cover diverse editing tasks, but only with a limited success rate. To further ensure the robustness of the model, we propose progressively aligning the editing model by adding additional rounds of fine-tuning. In particular, since we already have an editing model at this stage, we may prepare a new set of data based on the current model following a similar pipeline for data generation. The results are then labeled and filtered again to fine-tune the editing model as in Sec. 2.2. We repeat this process for multiple rounds until the model converges, i.e. no more improvements over our metrics.

3 Experiments

3.1 Benchmark and Metrics

Two base models are evaluated for our experiments, namely SDXL (Podell et al., 2023) and an in-house T2I model based on DiT architecture (Peebles & Xie, 2023; Esser et al., 2024).

We use two public datasets to evaluate image editing performance. The HQ-Edit dataset proposed in (Hui et al., 2024) and Emu Edit dataset from (Sheynin et al., 2024). The former is composed of 293 Dalle3 generated images and the latter is composed of 535 real in-the-wild image inputs. We note that our method is mainly focused on the application scenarios in the HQ-Edit benchmark, where we want to revise T2I generated images with arbitrary instructions. Emu Edit is rather different from our training data, which mostly includes local editing on real-scene images. Therefore, we consider Emu Edit as an Out-of-Domain (OOD) test but mainly rely on HQ-Edit to evaluate the application potential of our method.

We adopt two metrics to evaluate the editing performance. The first is CLIP-based (Brooks et al., 2023), where CLIP Direction Score is used to evaluate the alignment of the editing prompt and the CLIP image similarity is used to measure consistency. The second is LLM-as-evaluator, where GPT is used to replace the CLIP Direction score to measure the success of the editing.

3.2 Image Editing Comparison

We compare our method with several state-of-the-art image editing methods, including a training-free method Prompt-to-Prompt (Null-text Inversion) (Hertz et al., 2022; Mokady et al., 2023), and

Model	HQ-Edit			Emu Edit		
	GPT↑	CLIP_{dir}↑	CLIP_{img}↑	GPT↑	CLIP_{dir}↑	CLIP_{img}↑
Prompt-to-Prompt	26.93	0.0811	0.7462	12.69	0.0488	0.6568
Instruct-Pix2Pix	47.50	0.1224	0.8390	31.39	0.0726	0.8092
MagicBrush	47.51	0.1287	0.8008	44.25	0.0856	0.7930
Emu Edit	N/A	N/A	N/A	64.51	0.1094	**0.8206**
UltraEdit	54.17	0.1473	0.8281	46.95	0.0933	0.8072
SeedEdit (SDXL)	71.24	0.1656	**0.8698**	66.48	**0.1162**	0.8025
SeedEdit (in-house T2I)	**78.54**	**0.1766**	0.8524	**75.03**	0.1137	0.7875

Table 1: Quantitative evaluation on image editing benchmarks.

data-driven methods Instruct-Pix2Pix (Brooks et al., 2023), MagicBrush Zhang et al. (2024), Emu Edit Sheynin et al. (2024) and UltraEdit Zhao et al. (2024). Since Emu Edit is not open-sourced, we only compare them on their own test set. For the other methods, we used their model released with default parameters for comparison. Table 1 shows the quantitative results of the baselines and our method. Overall, our method shows a significantly higher editing score on both benchmarks than open-source baselines. Meanwhile, we also observe a higher CLIP image similarity on the HQ-Edit dataset, which indicates a better preservation of the content in the original image.

Although we mainly focus on the application scenario for revising T2I images as in HQ-Edit, our method also achieves descent quantitative scores on the Emu Edit benchmark, which is comparable/better to the original Emu Edit method. However, in general, we observe that the quality of the generated images of all methods (including ours) is not so satisfying on the Emu Edit benchmark, which proves our belief that the revisement of T2I images could be a first step to be solved before editing on arbitrary in-the-wild images.

Fig 5 shows some qualitative examples of our method and baselines on the HQ-Edit benchmark. A major difference between our method is that our method could understand rather ambiguous instructions and when performing fine-grained editing with a higher success rate.

Lastly, we compare the image editing capabilities of SeedEdit (in-house T2I model) with other commercial SoTA tools, such as DALLE3 Edit[1] and Midjourney[2], which allow the editing of self-generated images. Fig. 7 presents a qualitative comparison of the results. In general, both DALLE3 and Midjourney tend to introduce more unintended content changes beyond the specified editing prompt. Between the two, Midjourney produces more aesthetically pleasing images, while DALLE3 demonstrates superior adherence to the prompt instructions. In contrast, as shown in the last column, SeedEdit strikes a better balance, offering more precise editing that closely follows the given instructions. Furthermore, we conducted an internal user study that indicated a strong preference for the results generated by our method.

4 Conclusion

In this work, we introduced SeedEdit, a progressive alignment framework to adapt a pre-trained T2I diffusion model to image editing model, which maximizes both prompt alignment and image consistency. An causal diffusion model is proposed to take both images and texts as conditions for image generation. An iterative data generation and fine-tuning framework is proposed to align the diffusion towards precise image editing. Experimental results demonstrate that our method yields superior results compared to existing methods by a large margin.

[1] https://openai.com/index/hello-gpt-4o/

[2] https://docs.midjourney.com/docs/the-web-editor

(a) Get rid of the traffic.

(b) Let the subject raise up hands.

Figure 5: Example Results of different methods on the HQ-Edit benchmark.

(a) Get rid of the cat peering out from the suitcase.

(b) Change the time of the day to night

Figure 6: Example results of different methods on the Emu Edit benchmark.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

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. In *ICCV*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

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

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:2404.09990*, 2024.

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. In *CVPR*, 2023.

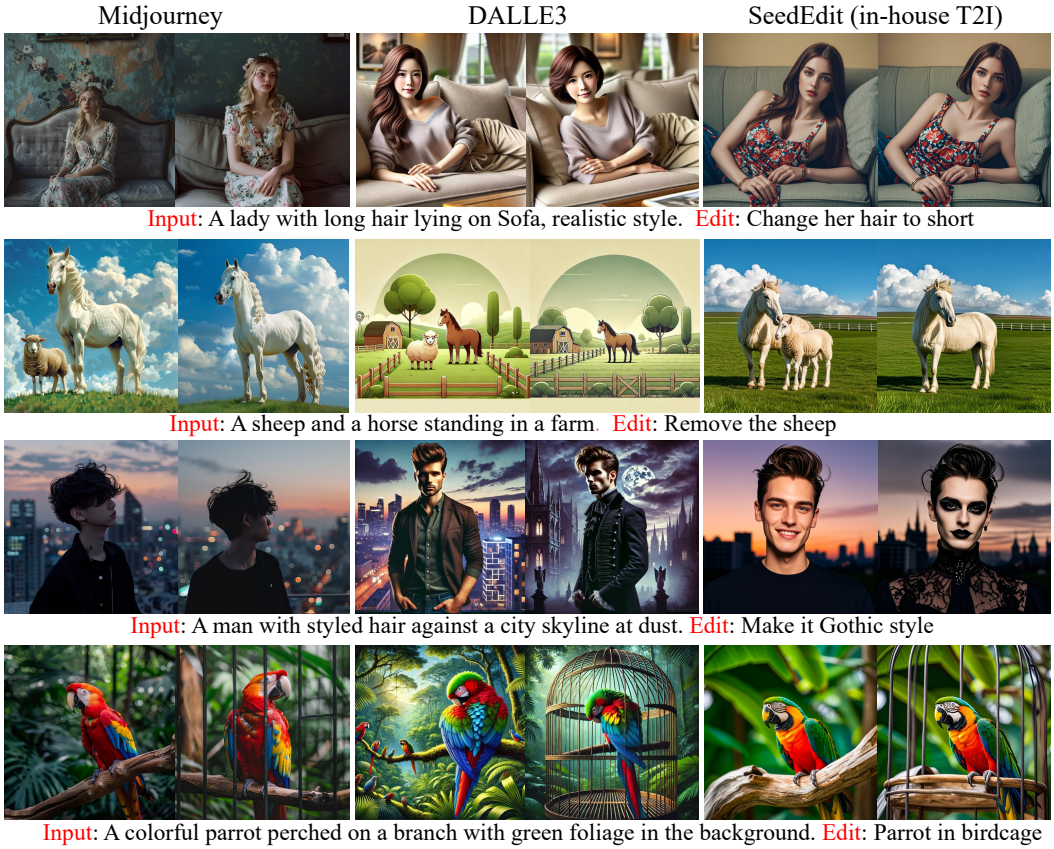


Figure 7: Example results from different products for editing based image generation.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023.

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:2112.10741*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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:2307.01952*, 2023.

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. In *CVPR*, 2023.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *CVPR*, 2024.

Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. *arXiv:2404.18212*, 2024.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS*, 2024.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *arXiv:2407.05282*, 2024.