Supplementary-DiffEditor: Boosting Accuracy and Flexibility on Diffusion-based Image Editing

Chong Mou[1,3] Xintao Wang[2] Jiechong Song[1] Ying Shan[2] Jian Zhang[1,3]*

[1]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University

[2]ARC Lab, Tencent PCG [3]Peking University Shenzhen Graduate School-Rabbitpre AIGC Joint Research Laboratory

In this supplementary material, we provide the following additional details to facilitate the understanding of our paper:

(1) In Sec. 1, we provide more ablation studies of our proposed DiffEditor. It includes regional SDE, guidance time steps, image prompt, and text input.

(2) In Sec. 2, we provide more comparison results between our DiffEditor and other methods on content dragging, object moving, resizing, pasting, and appearance replacing tasks.

(3) In Sec. 3, we present more comparisons between our method and other text-guided methods.

(4) In Sec. 4, we provide a user study of different methods on different editing tasks.

(5) In Sec. 5, we provide a further discussion of our limitations.

1. More Ablation Studies

In this section, we conduct ablation studies on more detailed components of our proposed DiffEditor.

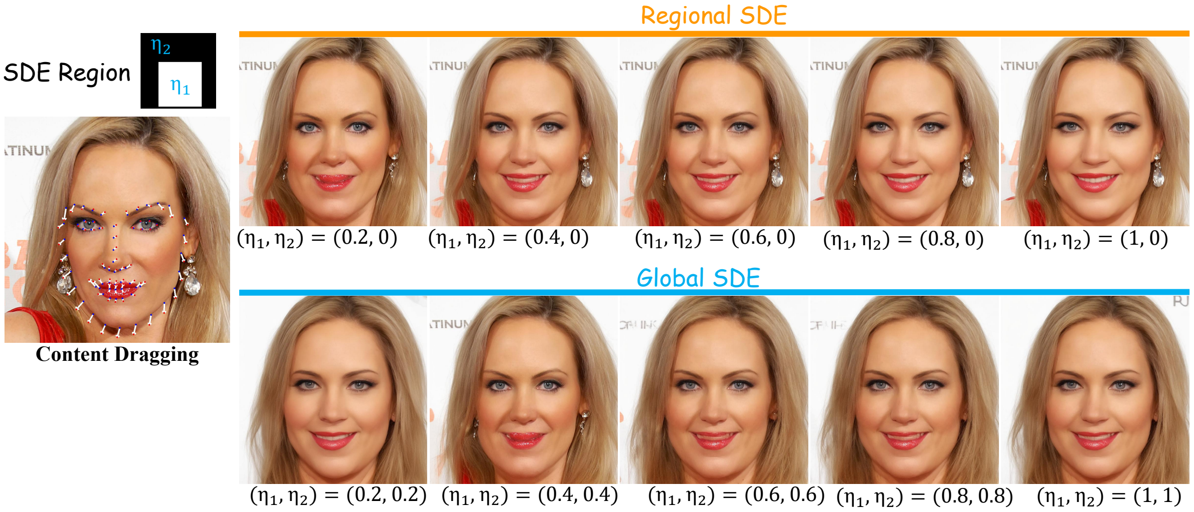


Figure 1. The impact of using regional SDE and the intensity of SDE randomness on the editing results. The comparison between the first and the second rows shows that regional SDE can improve editing flexibility while maintaining good content consistency. The comparison within each row shows that the greater the randomness intensity, the higher the editing flexibility.

Regional SDE. In our image editing process, the diffusion sampling is defined as:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\frac{\mathbf{z}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta^t(\mathbf{z}_t)}{\sqrt{\alpha_t}}}_{\text{"predicted } \mathbf{z}_0 \text{ "}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\boldsymbol{\epsilon}_\theta^t(\mathbf{z}_t)}_{\text{"direction pointing to } \mathbf{z}_t \text{ "}} + \underbrace{\sigma_t\boldsymbol{\epsilon}}_{\text{"noise"}}, \tag{1}$$

*Corresponding author

This work was supported by National Natural Science Foundation of China under Grant 62372016.

where $\sigma_t = \eta\sqrt{(1-\alpha_{t-1})\,/\,(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$. When $\eta = 1$ for all t, it becomes DDPM [3], *i.e.*, a stochastic differential equation (SDE). As $\eta = 0$, the sampling process becomes deterministic, *i.e.*, an ordinary differential equation (ODE). Most diffusion-based image editing works (*e.g.*, DragonDiff [8] and DragDiff [5]) rely on ODE to achieve better content consistency. In this paper, we explore utilizing the randomness in SDE to improve editing flexibility while preserving content consistency. Specifically, we design a regional SDE to locally inject the randomness:

$$\begin{aligned} &\mathbf{z}_{t-1} = \mathbf{m}_{edit} \cdot \mathcal{F}(\mathbf{z}_t; \eta_1(t)) + (1-\mathbf{m}_{edit}) \cdot \mathcal{F}(\mathbf{z}_t; \eta_2(t)), \\ &(\eta_1(t), \eta_2(t)) = \begin{cases} (0.4, 0.2), & t \in \tau_{SDE} \\ (0, 0), & t \notin \tau_{SDE} \end{cases} \qquad , \end{aligned} \tag{2}$$

where $\mathcal{F}(\mathbf{z}_t; \eta(t))$ is used to simplify Eq. 1. $\mathbf{m}_{edit}$ locates the editing region. In Fig. 1, we conducted an ablation study on the details of our proposed regional SDE. The comparison between the first and the second rows shows that regional SDE can improve editing flexibility (*e.g.*, open the mouth) while controlling the flexibility in a local region to maintain good content consistency in unedited regions (*e.g.*, the background content). The comparison within each row shows that the greater the randomness intensity, the higher the editing flexibility. After balancing the editing flexibility and accuracy, we design the regional SDE as Eq. 2.

Guidance time steps. Our editing process is built on 50-step diffusion sampling. We incorporate editing guidance every k steps (2 in our method) in a time interval $n = [t_1, t_2]$ ($[0, 30]$ in our method). The selection of n and k impacts the complexity and performance of the editing process. In Fig. 2, we demonstrate the impact of the k on editing performance and complexity. It can be seen that having guidance-skip (*i.e.*, $k > 1$) can reduce the time complexity. Due to our accurate guidance design, we can also achieve promising editing results with large k. After balancing between complexity and performance, we choose $k = 2$. Fig 3 demonstrates the impact of the guidance time interval n on editing performance and complexity. It shows the importance of adding editing guidance in the early sampling time steps. The later the sampling, the weaker the role of editing guidance. In addition, adding guidance information in more sampling steps helps improve the quality of texture details. After balancing the performance and computational complexity, we add score-based gradient guidance in the first 30 steps.

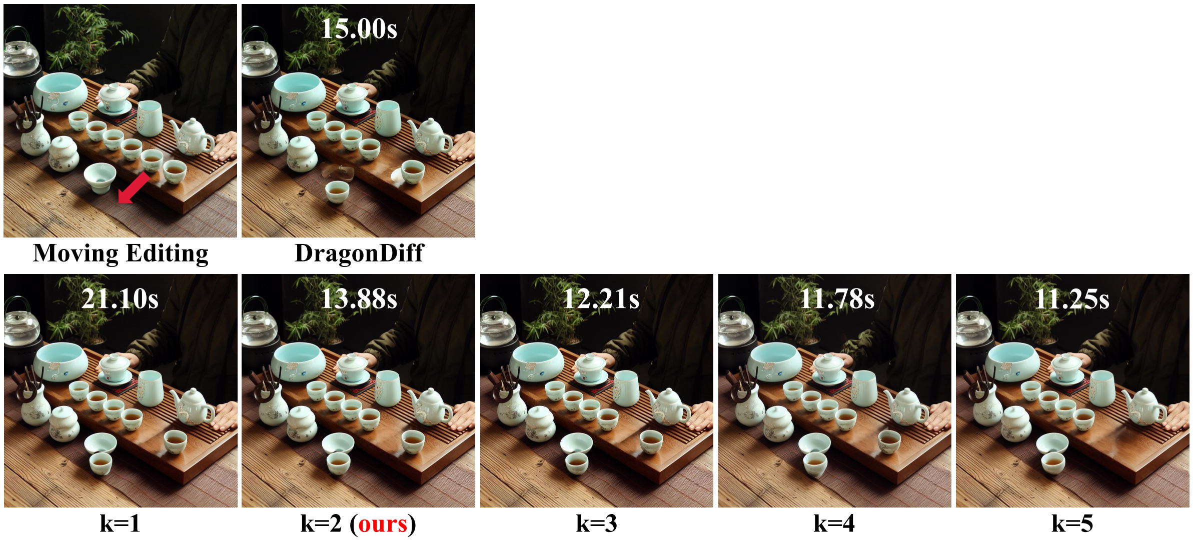


Figure 2. The influence of providing editing guidance every k steps on performance and time complexity. The results show the trade-off that applying guidance-skip (*i.e.*, $k > 1$) can reduce the time complexity but weaken the editing quality.

Image prompt. In this paper, we design image prompt to enhance the description of the editing content. Mathematically, it is defined as:

$$\text{Att}(\mathbf{Q}, \mathbf{K}^{'}, \mathbf{V}^{'}, \mathbf{K}^{''}, \mathbf{V}^{''}) = \mathcal{S}(\frac{\mathbf{Q}(\mathbf{K}^{'})^T}{\sqrt{d}})\mathbf{V}^{'} + \gamma \cdot \mathcal{S}(\frac{\mathbf{Q}(\mathbf{K}^{''})^T}{\sqrt{d}})\mathbf{V}^{''}, \tag{3}$$

where $(\mathbf{K}^{'}, \mathbf{V}^{'})$ and $(\mathbf{K}^{''}, \mathbf{V}^{''})$ refer to the keys and values from the text and image prompt, respectively. γ is a weight to balance these two terms. $\mathcal{S}$ is the function of Softmax. In Fig. 4, we conduct the ablation study of γ and the length of the image prompt (*i.e.*, number of tokens). One can see that increasing the number of tokens and raising the weight of the image

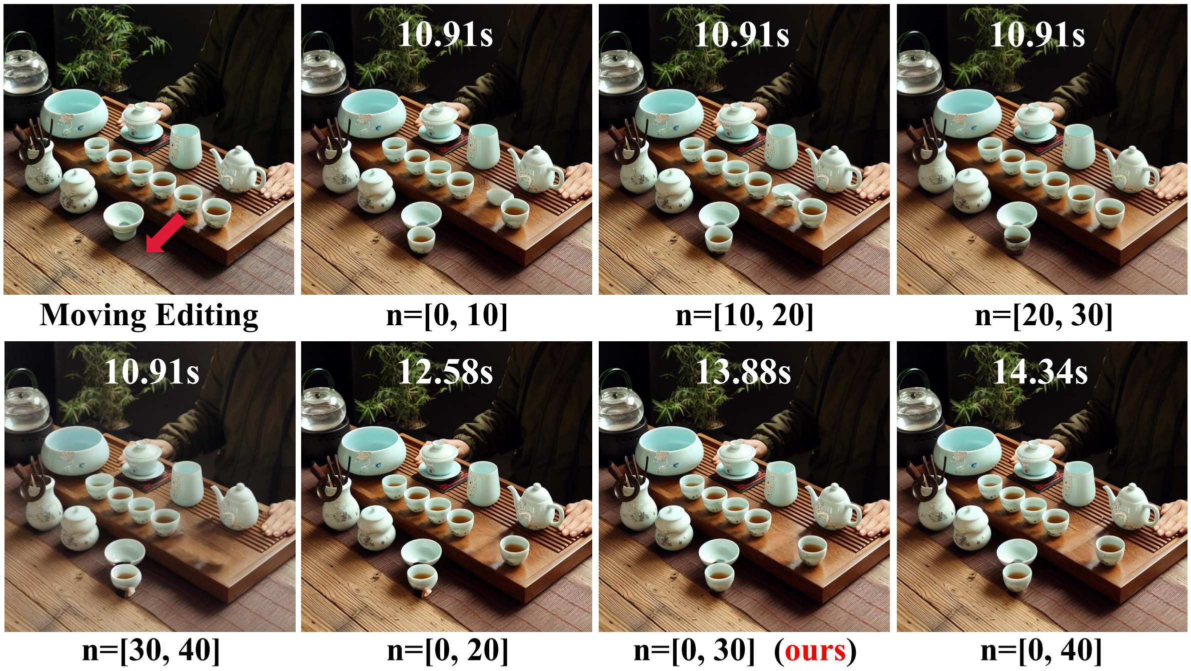


Figure 3. The editing results generated by introducing editing guidance at different time intervals. The results show the importance of adding guidance information in the early sampling stage, and introducing editing guidance in more sampling steps helps refine the details.

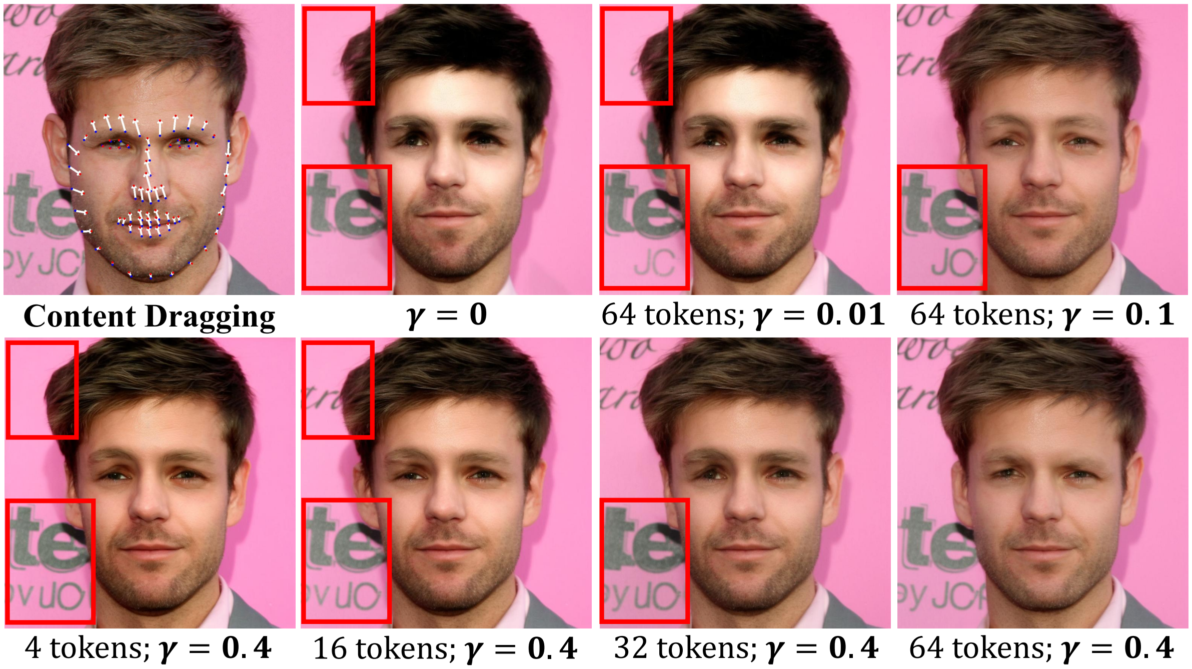
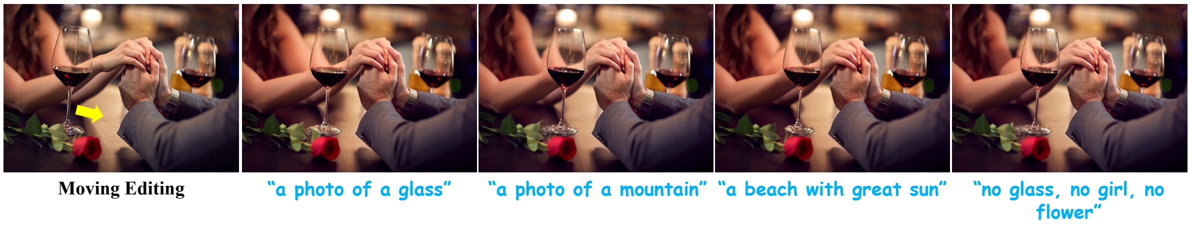


Figure 4. The influence of the token numbers and weight of the image prompt on the editing results. The results demonstrate that an increased number of tokens and higher weights have a positive effect on the representation of the image content.

prompt can both enhance the description of the source image content. It is worth noting that if γ is too large, it will make the editing result being overly consistent with the original image, thereby affecting the editing effect. Finally, we choose $\gamma = 0.5$ and the number of tokens to be 64.



Moving Editing

"a photo of a glass"

"a photo of a mountain"

"a beach with great sun"

"no glass, no girl, no flower"

Figure 5. The influence of text prompt on the editing result. This figure shows that the editing results are similar and meet the editing requirements under different text inputs, *i.e.*, our method is insensitive to text prompts.

Table 1. User study on object pasting, object moving, object resizing, appearance replacing, and content dragging. The evaluation includes editing accuracy (*i.e.*, the consistency of editing results with editing goals and original content) and editing quality (*i.e.*, the image quality of editing results). For each application, we select 8 samples and allow 16 volunteers to rate (0-5) them on these two aspects.

	Object Pasting		Object Moving & Resizing		Appearance Replacing		Content Dragging	
	Accuracy	Quality	Accuracy	Quality	Accuracy	Quality	Accuracy	Quality
DragGAN [6]	-	-	-	-	-	-	4.32	3.27
Paint-by-examplt [10]	3.84	**4.43**	-	-	-	-	-	-
Self-Guidance [2]	-	-	2.24	2.82	2.43	2.78	-	-
DragDiff [8]	-	-	-	-	-	-	4.10	4.21
DragonDiff [5]	4.24	4.15	4.42	4.21	3.96	4.17	4.29	4.33
DiffEditor (Ours)	**4.38**	4.29	**4.57**	**4.37**	**4.48**	**4.52**	**4.45**	**4.53**

Text input. Although our method is based on the image feature correspondence [9] to achieve fine-grained image editing, our approach is built on pre-trained StableDiffusion [7], so there is also text input. In Fig. 5, we demonstrate the impact of text on the editing results. It can be seen that under different text inputs (relevant, irrelevant, or even contradictory to the image content), our method can produce stable and satisfactory editing results that meet the requirements. Therefore, our method is insensitive to text prompts.

2. More Comparison between Our Method and other Methods

In this section, we present more comparisons with other methods. Fig. 8 shows more comparisons of our method with DragGAN [6], DragDiff [8], and DragonDiff [5] in the content dragging task. It can be seen that DragGAN has degradation on some editing content that deviates from the training distribution (e.g., the hat, a dog in a barrel). DragDiff and DragonDiff have limitations in flexibility, such as making the dog open its mouth. In comparison, our method performs better in editing flexibility and accuracy. Fig. 9 and Fig. 10 show comparisons of our method with Self-Guidance [2] and DragonDiff in tasks of appearance replacing, object moving, and object resizing. Due to the lack of consistency constraints in Self-Guidance, the editing results deviate from the original image. In addition, it is difficult for the text to accurately correspond to the image content in complex scenes, resulting in a lack of editing accuracy of Self-Guidance. Due to the accurate and refined guidance design, our DiffEditor outperforms DragonDiff in editing accuracy. Fig. 11 shows more comparisons of our method with Point-by-example [10] and DragonDiff in the object pasting task. Although the trained Point-by-example can achieve more natural pasting effects, it is difficult to maintain the object identity. The pasting results of our DiffEditor are more consistent with the original object.

3. More Comparison between Our Method and Text-guided Methods

In this section, we compare our approach with more text-guided image editing methods, *e.g.*, Null-text inversion [4] and InstructPix2Pix [1], as shown in Fig. 6. It can be seen that text guidance struggles to establish an accurate correspondence with the editing content, making it difficult to achieve accurate editing. For instance, it is challenging to edit a single object through text guidance in multi-object scenarios. Moreover, most text-guided image editing methods rely on text to specify object appearance, resulting in uncontrollable randomness. In contrast, our method enables accurate specification of the editing region and additional reference appearance.

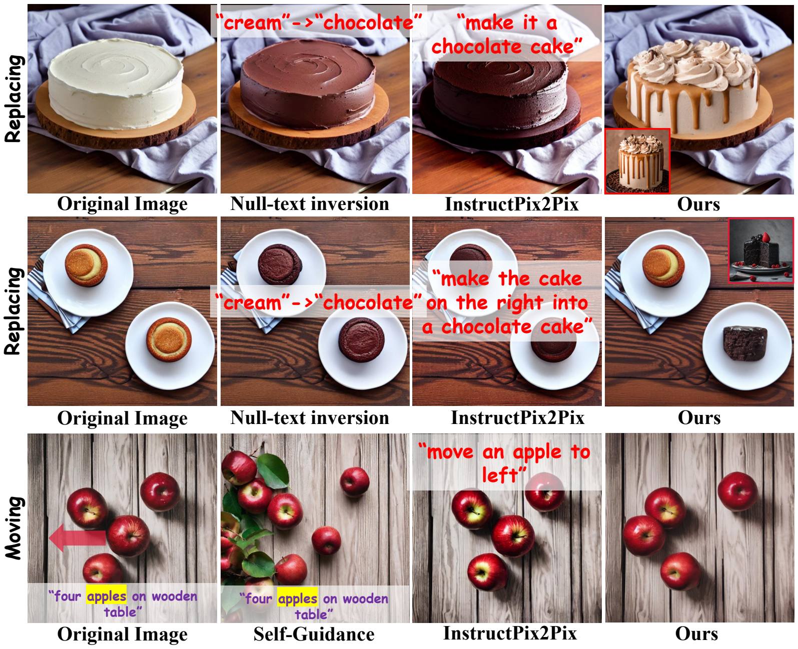


Figure 6. More visual comparison between our DiffEditor and other text-guided methods, *i.e.*, Null-text inversion [4], InstructPix2Pix [1], and Self-Guidance [2]. The results demonstrate that text-guided image editing methods struggle to achieve accurate image editing, especially in multi-object scenarios.

4. User Study on Different Tasks

In Tab. 1, we conduct a user study of different methods on different editing tasks. The evaluation includes editing accuracy (*i.e.*, the consistency of editing results with editing goals and original content) and editing quality (*i.e.*, the image quality and harmony of editing results). For each application, we select 8 samples and allow 16 volunteers to rate (0-5) them on these two aspects. The results demonstrate the promising performance of our method.

5. Discussion of Limitations and Future Works

As described in Sec. 5 in our main paper, our method has limitations in editing scenarios that require a large amount of content imagination while maintaining consistency. Here, we present a failure case in Fig. 7. Our editing goal is to change the direction of the car by dragging the car's front and rear in the opposite direction. However, the editing result slightly changes the direction of the car and shortens the length of the car. The result mainly reflects two limitations: (1) In some scenarios, the model cannot understand our editing goals well by only specifying the start and target points of the edit; (2) The model still has limitations in large-scale content imagination. In future work, we will focus on addressing these limitations.

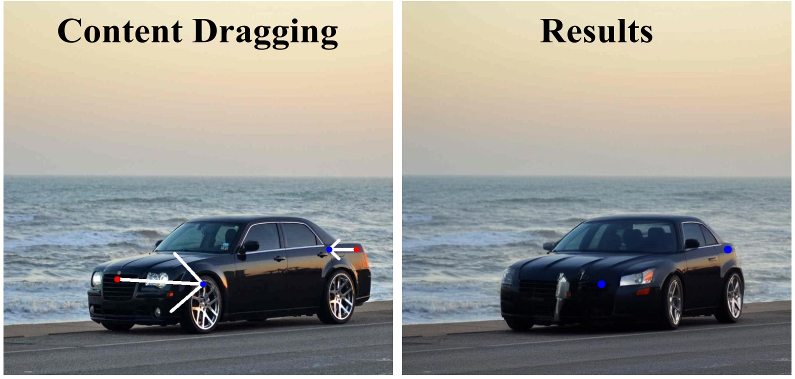


Figure 7. The visualization of a failure case of our method. Our editing goal is to change the direction of the car, but the model does not adequately understand our editing goal.

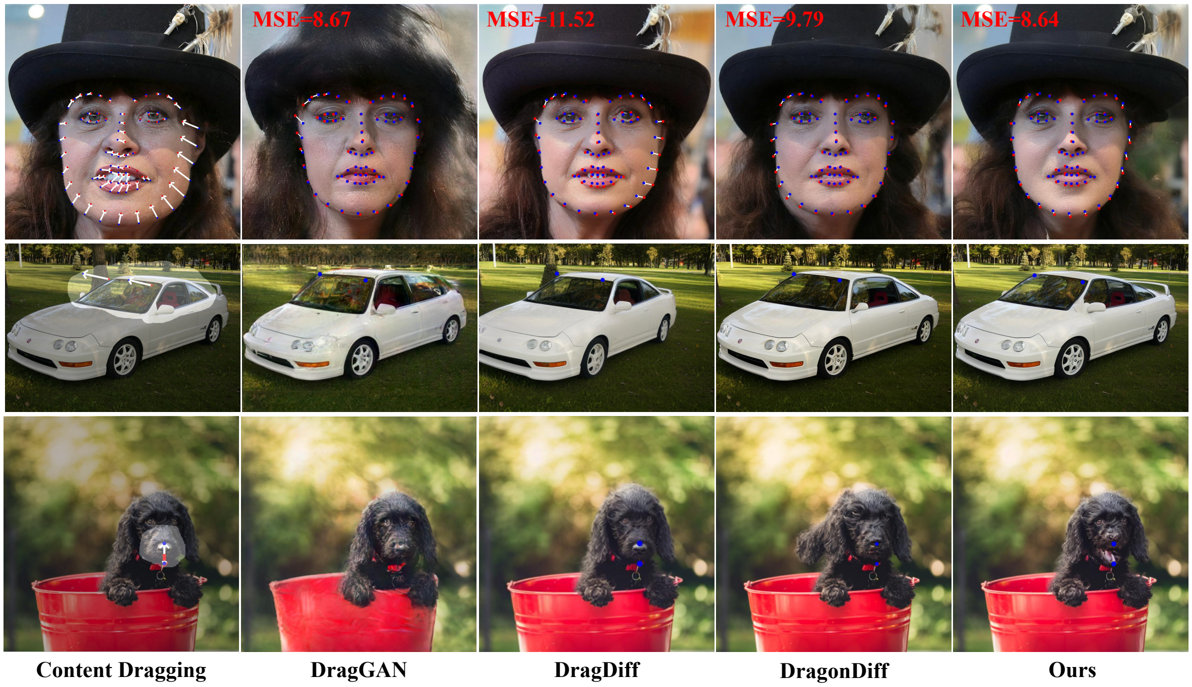


Figure 8. More visual comparison between our DiffEditor and other methods (*i.e.*, DragGAN [6], DragDiff [8], and DragonDiff [5]) on content dragging.

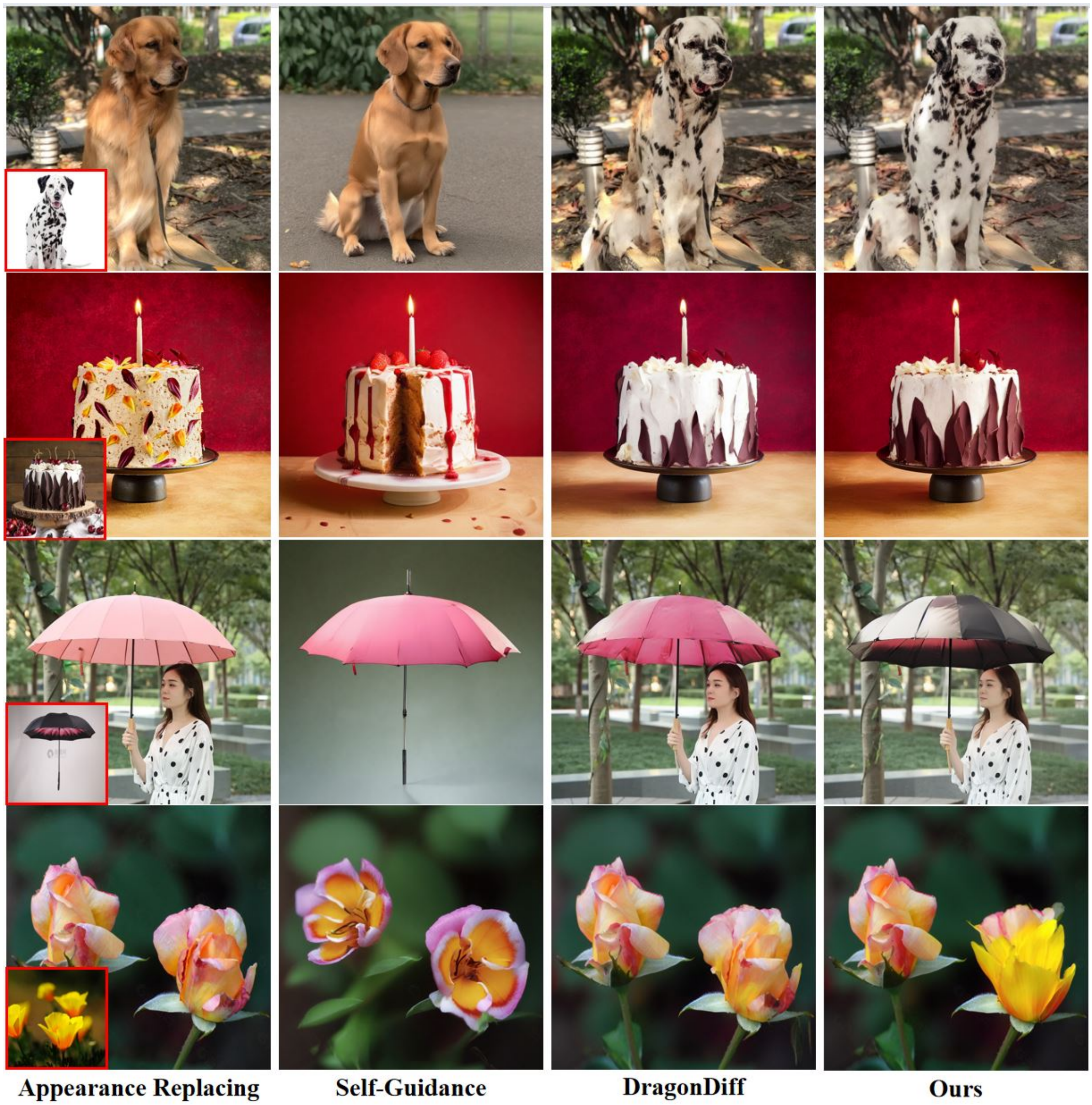


Figure 9. More visual comparison between our DiffEditor and other methods (*i.e.*, Self-Guidance [2], DragonDiff [5]) on appearance replacing.

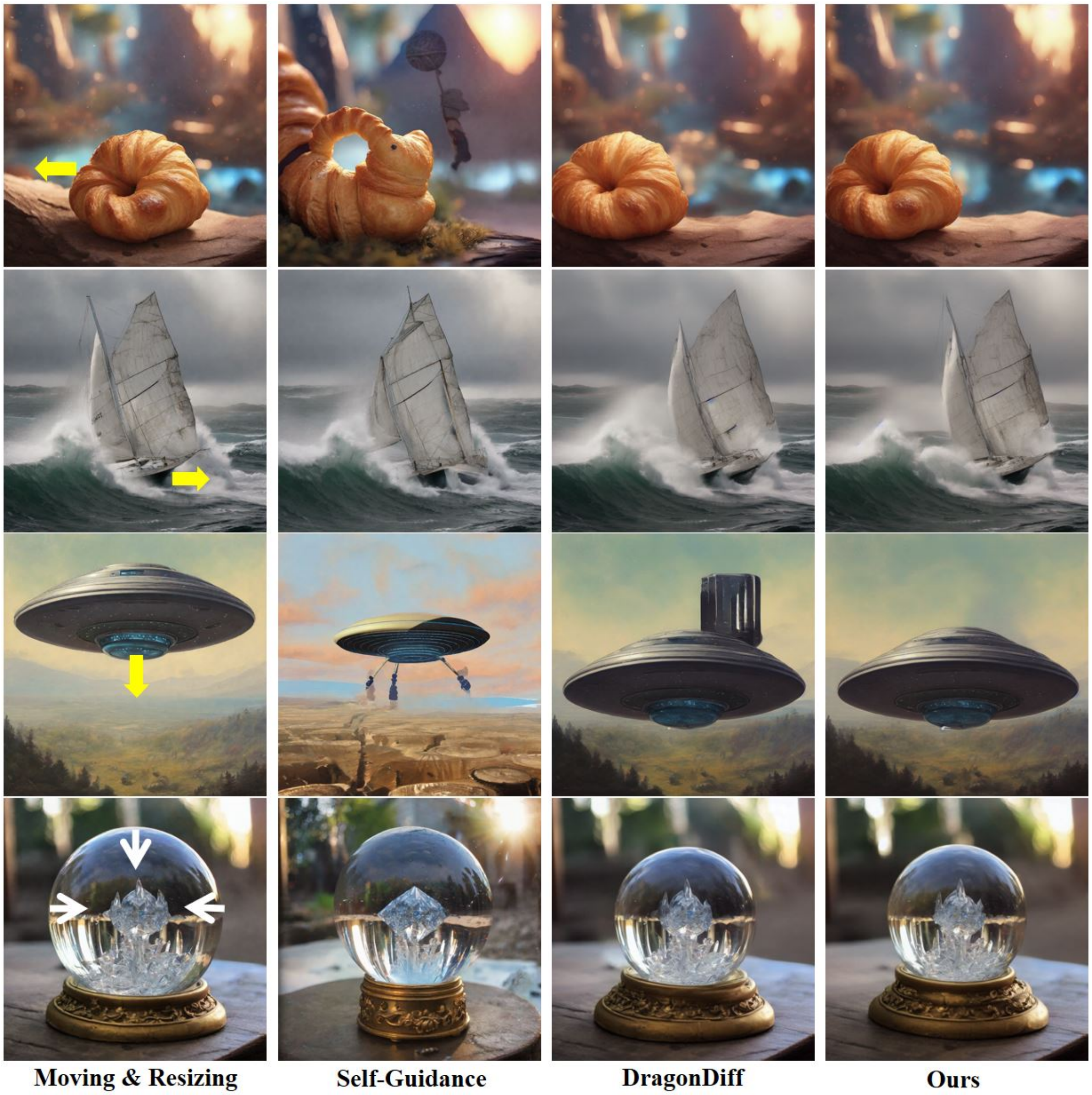


Figure 10. More visual comparison between our DiffEditor and other methods (*i.e.*, Self-Guidance [2], DragonDiff [5]) on object moving and resizing.

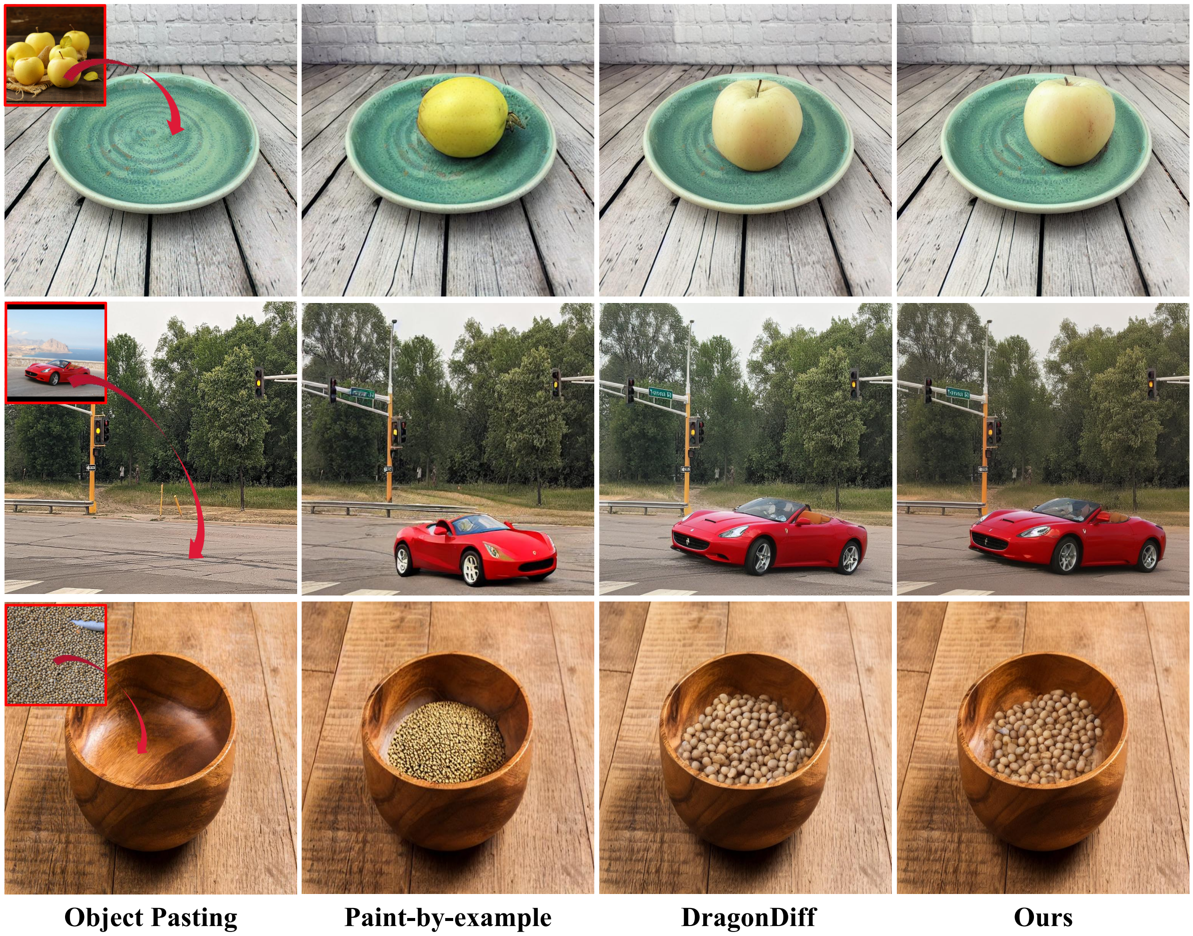


Figure 11. More visual comparison between our DiffEditor and other methods (*i.e.*, Paint-by-example [10], DragonDiff [5]) on object pasting.

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. 4, 5

[2] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023. 4, 5, 7, 8

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[4] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 4, 5

[5] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 2, 4, 6, 7, 8, 9

[6] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. *arXiv preprint arXiv:2305.10973*, 2023. 4, 6

[7] 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. 4

[8] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 2, 4, 6

[9] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 4

[10] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 4, 9