

Inversion-Free Image Editing with Language-Guided Diffusion Models

Supplementary Material

1. DDCM v.s. Inversion-based Editing

Existing inversion-based editing methods are limited for real-time and real-world language-driven image editing applications. First, most of them still depend on a time-consuming inversion process to obtain the inversion branch as a set of anchors. Second, consistency remains a bottleneck given the efforts from optimization and calibration. Recall that dual-branch inversion methods perform editing on the target branch by iteratively calibrating the z_t^{tgt} with the actual distance between the source branch and the inversion branch at t , as is boxed in Figure 7a. While they ensure faithful reconstruction by leaving the source branch untouched from the target branch, the calibrated z_t^{tgt} does guarantee consistency from z_t^{src} in the source branch, as can be seen from the visible difference between z_0^{src} and z_0^{tgt} in Figure 7a. Third, all current inversion-based methods rely on variations of diffusion sampling, which are incompatible with efficient Consistency Sampling using LCMs.

DDCM offers an alternative to address these limitations, introducing an Inversion-Free Image Editing (InfEdit) framework. While also adopting a dual-branch paradigm, the key of our InfEdit method is to directly calibrate the initial z_0^{tgt} rather than the z_t^{tgt} along the branch, as is boxed in Figure 7b. InfEdit starts from a random terminal noise $z_{\tau_1}^{\text{src}} = z_{\tau_1}^{\text{tgt}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As shown in Figure 7b, the source branch follows the DDCM sampling process without explicit inversion, and we directly compute the distance $\Delta \varepsilon^{\text{cons}}$ between $\varepsilon^{\text{cons}}$ the $\varepsilon_{\theta}^{\text{src}}$ (the predicted noise to reconstruct a \bar{z}_0^{src}). For the target branch, we first compute the $\varepsilon_{\theta}^{\text{tgt}}$ to predict \bar{z}_0^{tgt} , and then calibrate the predicted target initial with the same $\Delta \varepsilon^{\text{cons}}$. Algorithm 1 outlines the mathematical details of this process, in which we slightly abuse the notation to define $f_{\theta}(z_t, t, \varepsilon) = (z_t - \sqrt{1 - \alpha_t} \varepsilon) / \sqrt{\alpha_t}$.

2. Additional Experiments and Discussions

2.1. Image-to-Image Translation Tasks

We further evaluate InfEdit (VI+UAC) in scene-level and object-level I2I translation tasks for more general comparisons. The baselines we considered include Text2LIVE [3], SDEdit [21], CycleD [36], NT [23], MasaCtrl [4], as well as the training-based state-of-the-art CycleNet [37]. As shown in Table 3, InfEdit strikes an effective balance between translation effects and consistency. Qualitative examples are shown in Figure 19 and 20.

2.2. Output Diversity

InfEdit allows for diverse outputs from the same language prompt. We present qualitative examples in Figure 6.



Figure 6. Diverse output from “a painting of a waterfall (+and angels) in the mountains” with different seeds.

2.3. Connecting InfEdit to LLMs

InfEdit supports editing with language descriptions but does not follow instructions. We can further leverage the instruction-following capability of large language models (LLMs) to follow the image editing instructions, which improves the user experience. Motivated by recent work [18, 19], we validate the feasibility of prompting GPT-4 [27] to break down editing instructions into adequate source and target prompts for InfEdit. We release a Gradio demo.¹

3. Ethics Statement

While InfEdit offers promising advancements in image editing, it is crucial to consider its broader ethical, legal, and societal implications.

Copyright Infringement. As an advanced image editing tool, InfEdit could be used to modify and repurpose artists’ original works, raising concerns over copyright violations. It’s vital for practitioners to respect the rights of creators and maintain the integrity of the creative economy, ensuring adherence to licensing and copyright laws.

Deceptive Misuse. If exploited by nefarious entities, InfEdit’s capability to generate convincing image alterations could be used for misinformation, fraud, or identity theft. This necessitates responsible user guidelines and strong security protocols to prevent such misuse and safeguard against security threats.

Bias and Fairness. Furthermore, InfEdit builds upon pre-trained latent diffusion models and latent consistency models, which might carry inherent biases, leading to potential fairness issues. While the method is algorithmic and not pre-trained on large web-scale datasets, it’s important

¹<https://huggingface.co/spaces/sled-umich/InfEdit>

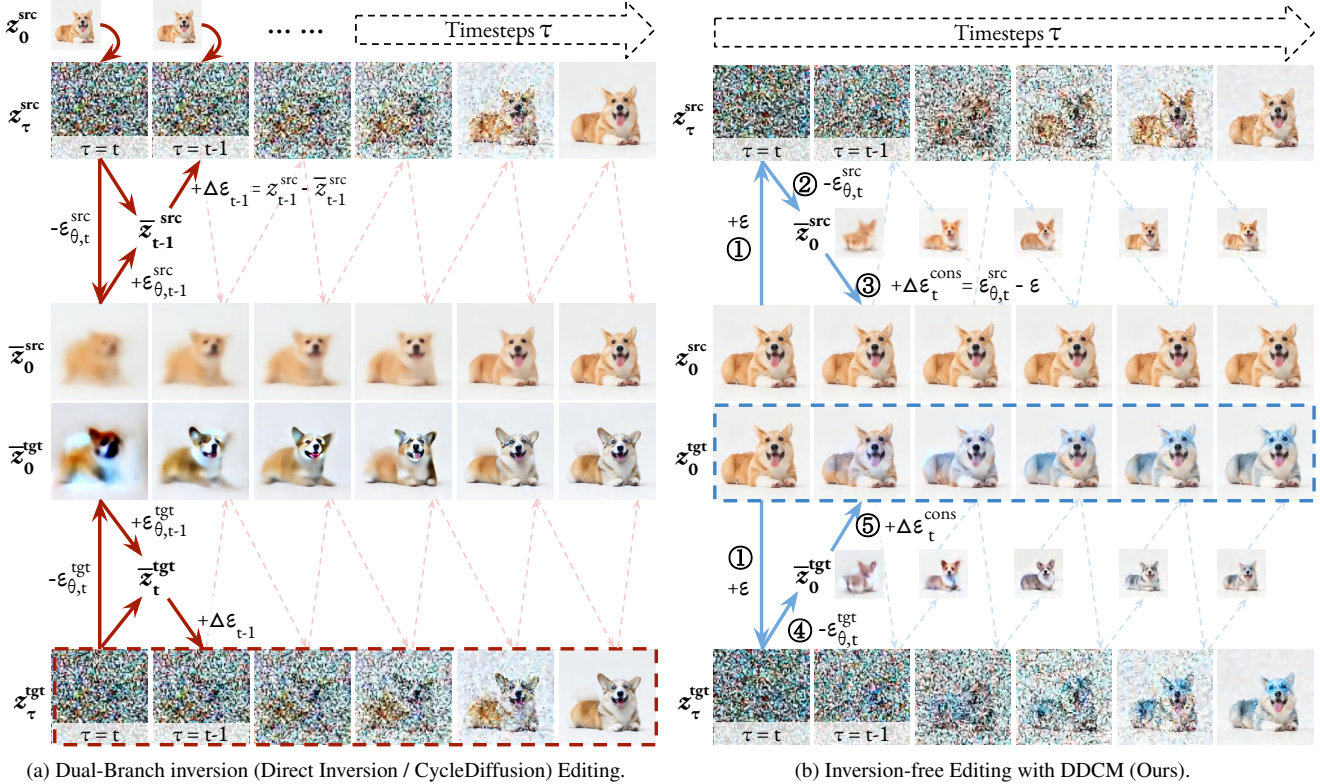


Figure 7. A comparative overview of the dual-branch inversion editing and inversion-free editing enabled by DDCM. While the former iteratively calibrates z_τ^{tgt} in the target branch, inversion-free editing iteratively polishes the target branch initial z_0^{tgt} . In (b), we initialize z_0^{tgt} with z_0^{src} for visualization purposes, while in principle it can start from any random noise. The circled numbers correspond to Algorithm 1.

Task	Summer↔Winter (512 × 512)					Horse↔Zebra (512 × 512)				
	FID ↓	CLIP Sim ↑	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	CLIP Sim ↑	LPIPS ↓	PSNR ↑	SSIM ↑
CycleNet	79.79	24.12	0.15	25.88	0.69	76.83	25.27	0.08	26.21	0.74
Text2LIVE	86.12	25.98	0.27	16.83	0.68	103.14	31.55	0.16	20.98	0.81
SDEdit	90.51	23.26	0.30	18.59	0.43	63.04	27.97	0.33	18.49	0.44
CycleD	84.52	24.40	0.24	21.66	0.68	41.17	29.09	0.29	19.41	0.61
NT+P2P	92.65	24.82	0.24	20.19	0.66	106.83	26.57	0.21	21.45	0.66
MasaCtrl	114.83	17.11	0.37	14.66	0.43	239.61	21.15	0.41	16.31	0.37
InfEdit	75.63	23.07	0.18	21.99	0.68	61.81	28.16	0.16	21.80	0.72

Table 3. Image2Image translation comparison. InfEdit methods achieve a favorable balance between consistency and translation quality.

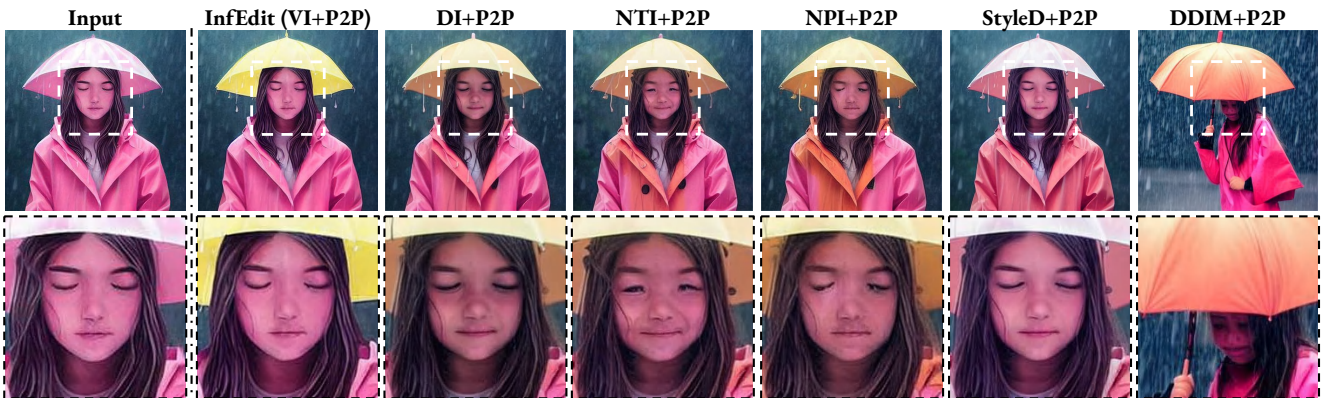
to recognize and mitigate any encoded biases in these pre-trained backbones to ensure fairness and ethical use.

By proactively addressing these concerns, we can leverage InfEdit’s capabilities responsibly, prioritizing ethical considerations, legal compliance, and the welfare of society. This approach is essential for advancing technology while safeguarding our community’s values and trust.

4. Additional Qualitative Examples

We present a qualitative example in Figure 8, showing that with the same P2P attention control, VI allows faithful se-

mantic changes and better consistency. We present overall qualitative comparisons in Figure 9 and the following.



A girl with a ~~pink~~ yellow umbrella in the rain.

Figure 8. A qualitative example for ablation over inversion methods. With the same P2P attention control, InfEdit (VI+P2P) allows faithful semantic changes as well as better consistency.

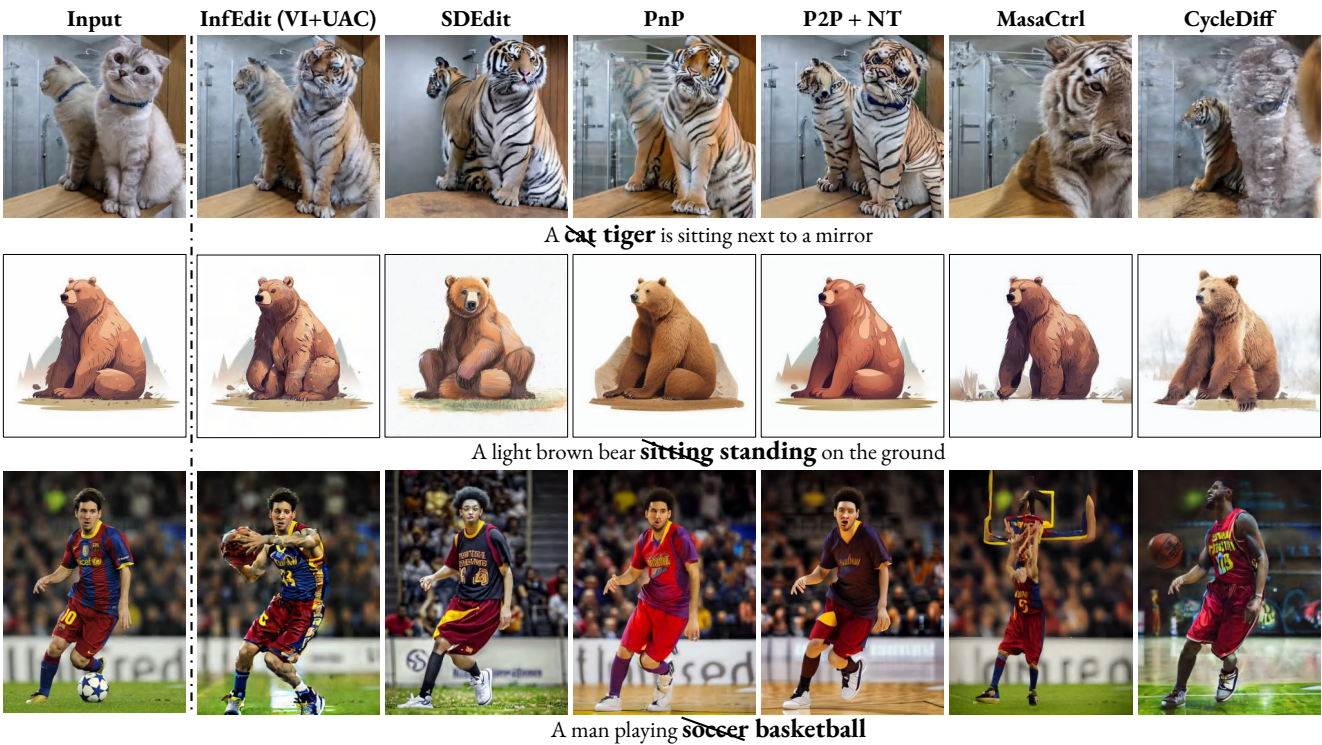


Figure 9. Qualitative comparisons of InfEdit (VI+UAC) against baselines. InfEdit attains editing goals with the best consistency.



Figure 10. Additional comparison on changing object tasks.

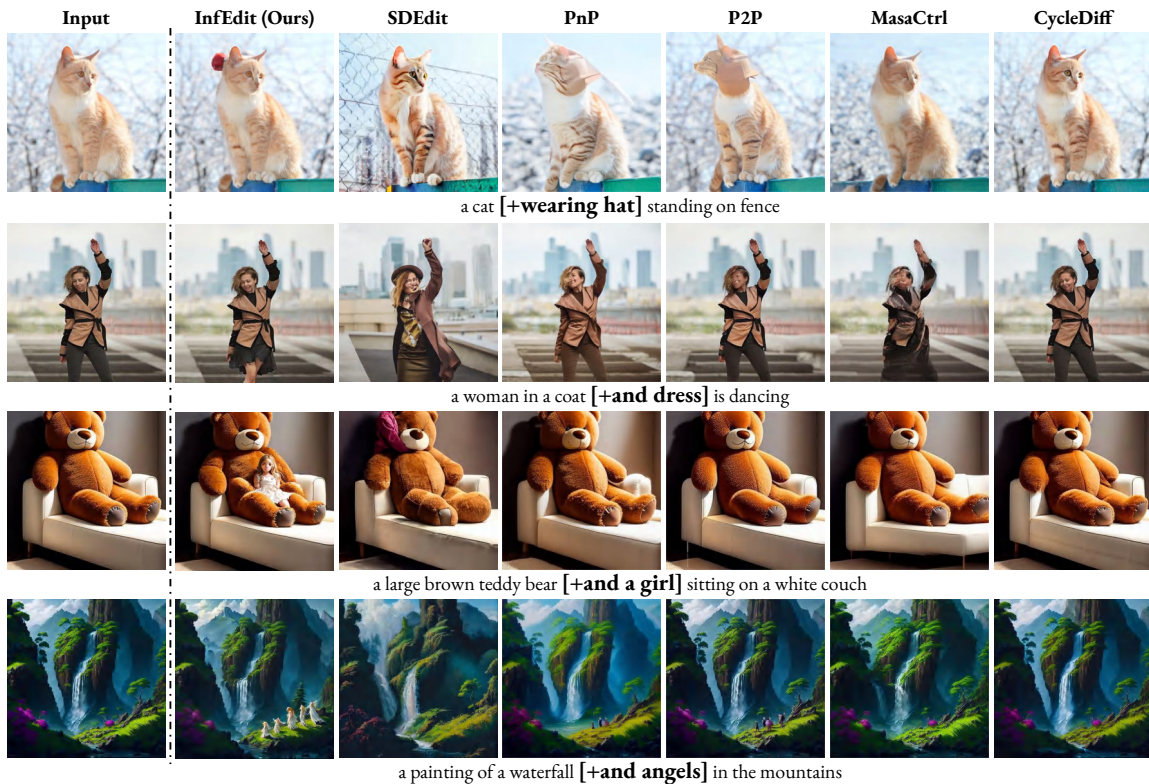


Figure 11. Additional comparison on adding object tasks.



Figure 12. Additional comparison on deleting object tasks.

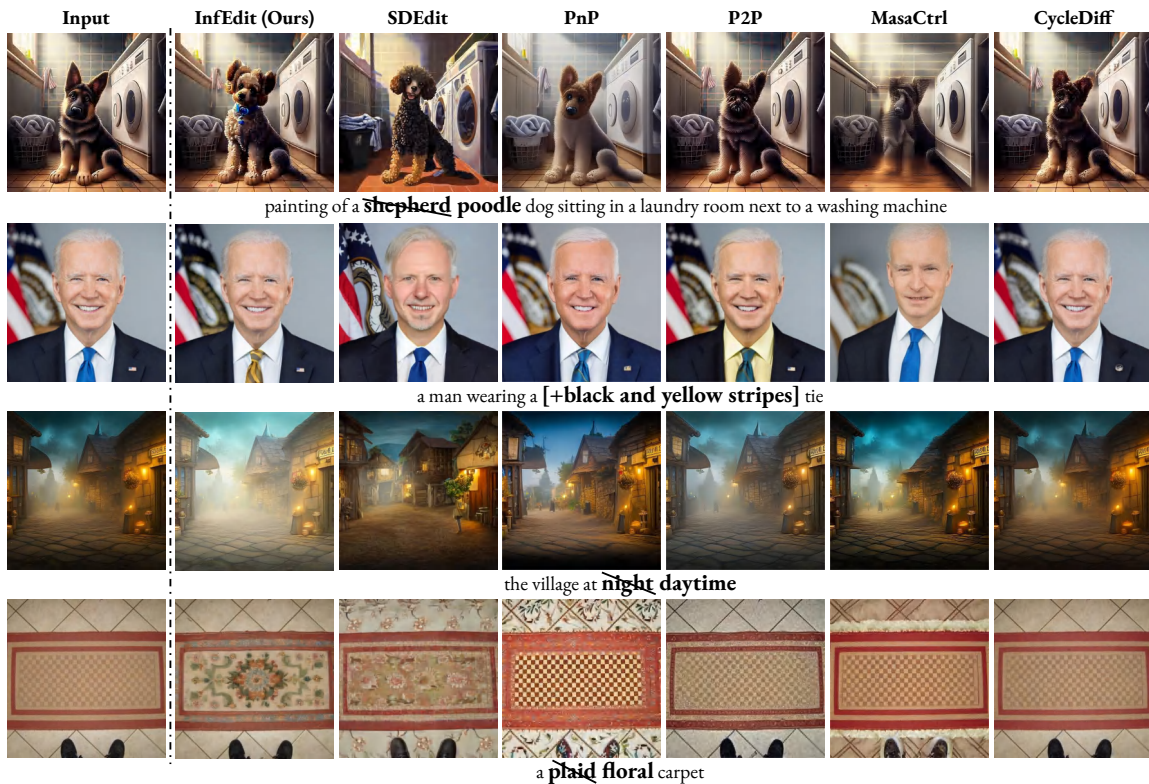


Figure 13. Additional comparison on changing content tasks.

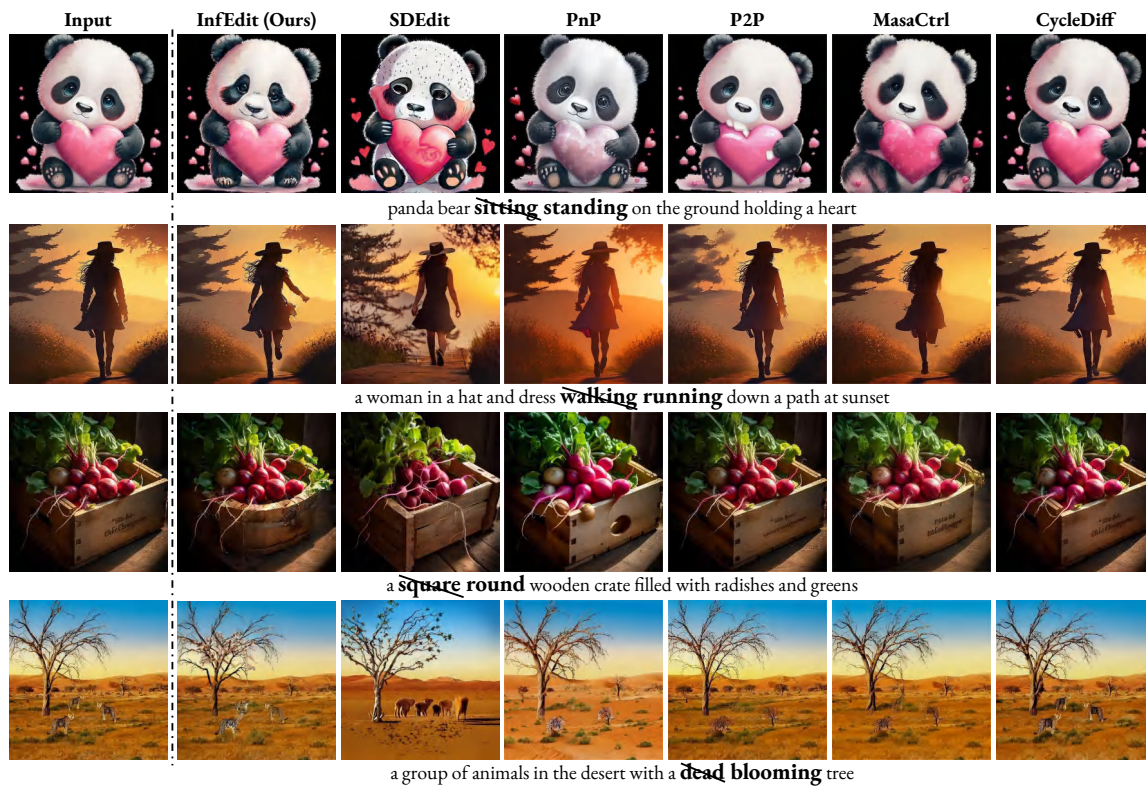


Figure 14. Additional comparison on changing pose tasks.

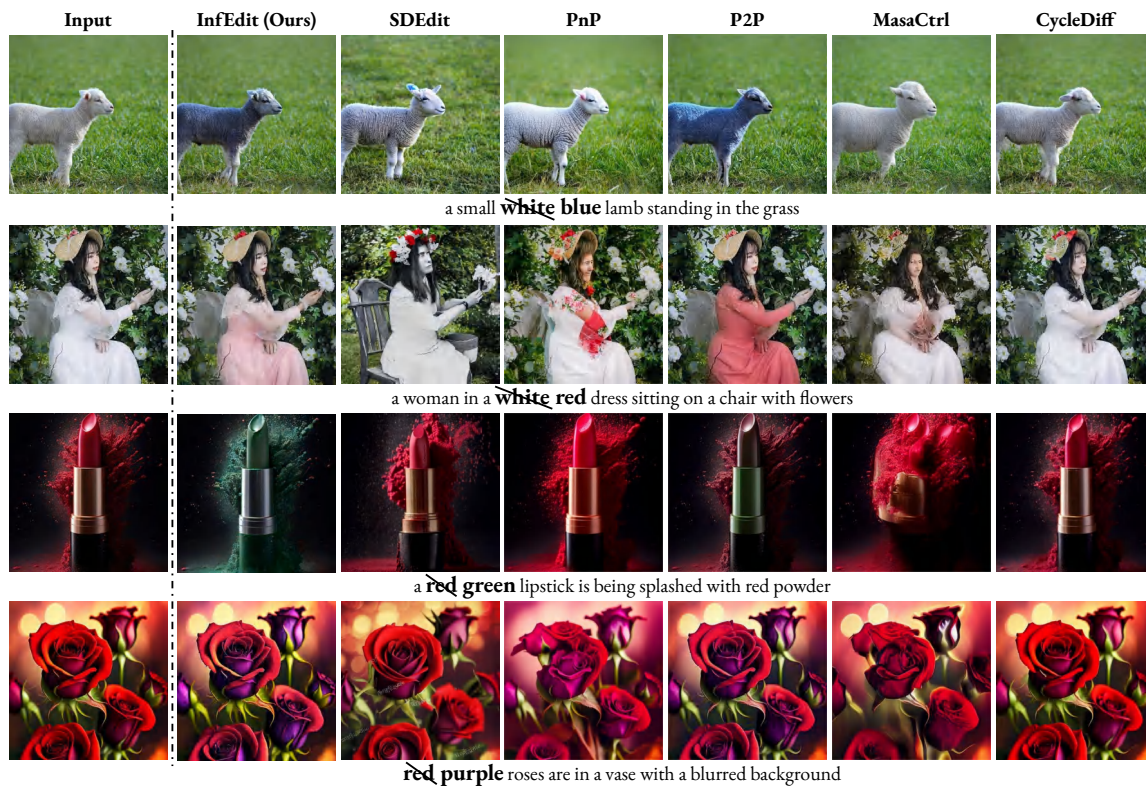


Figure 15. Additional comparison on changing color tasks.

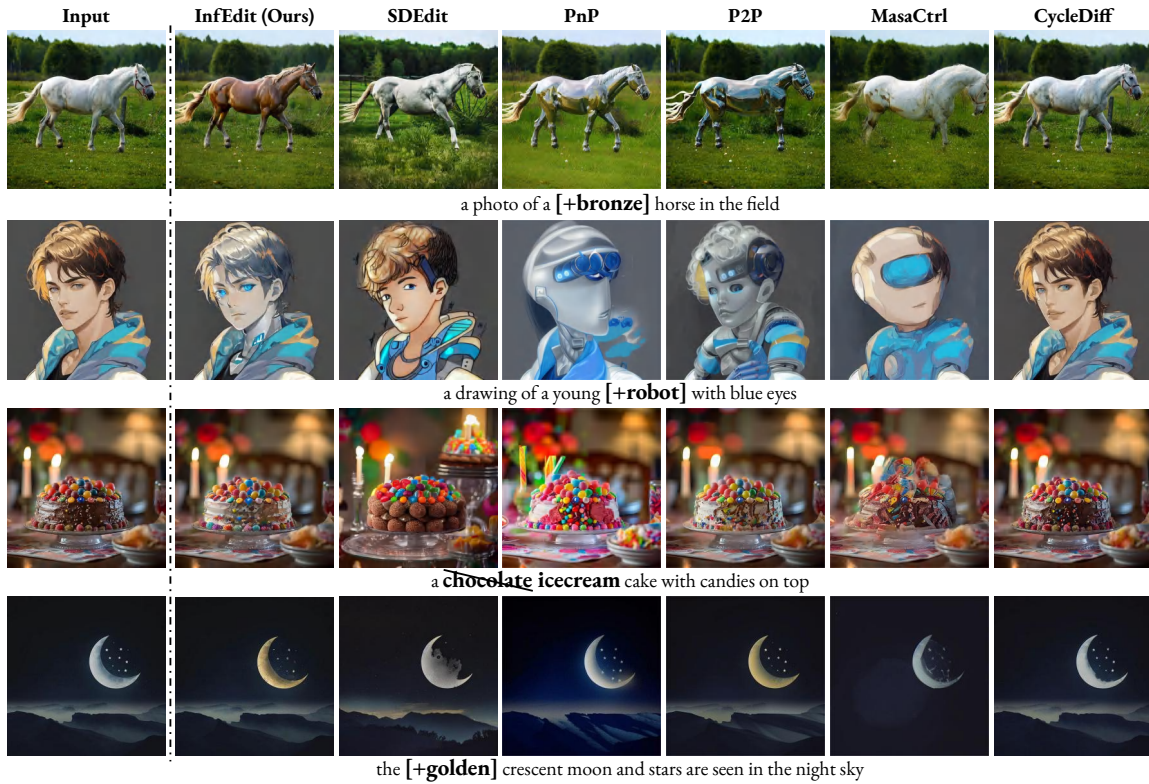


Figure 16. Additional comparison on changing material tasks.

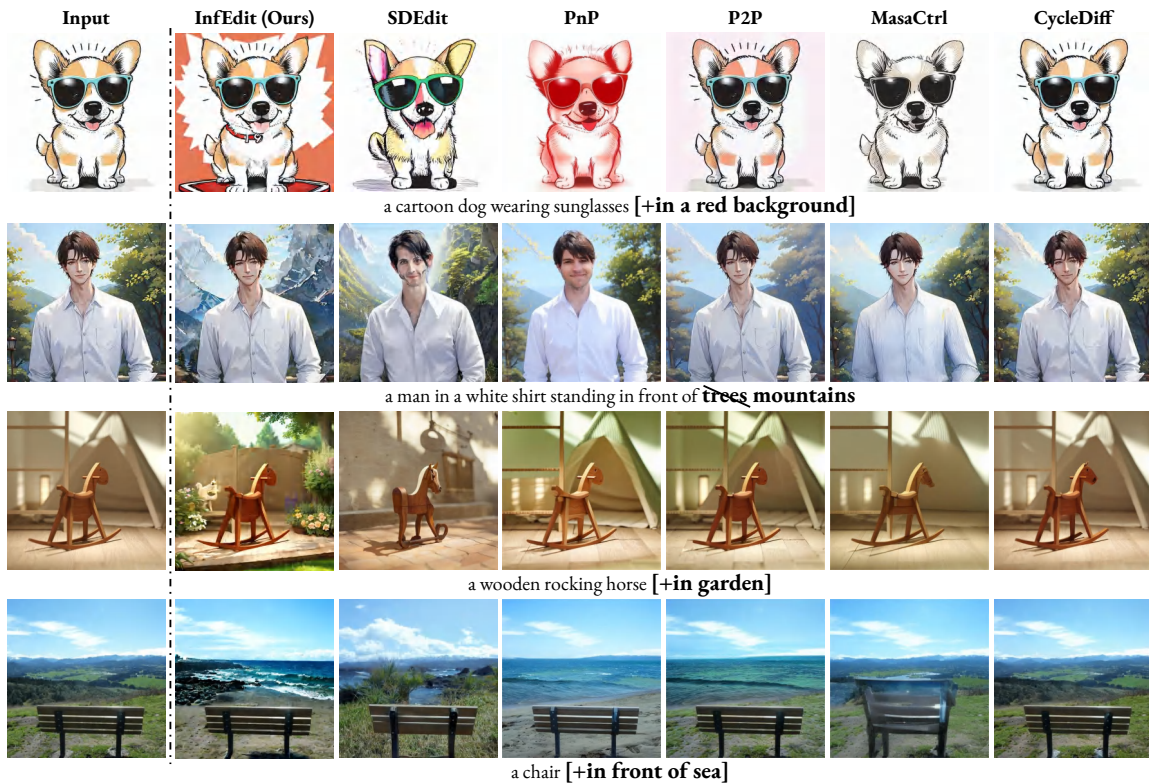


Figure 17. Additional comparison on changing background tasks.

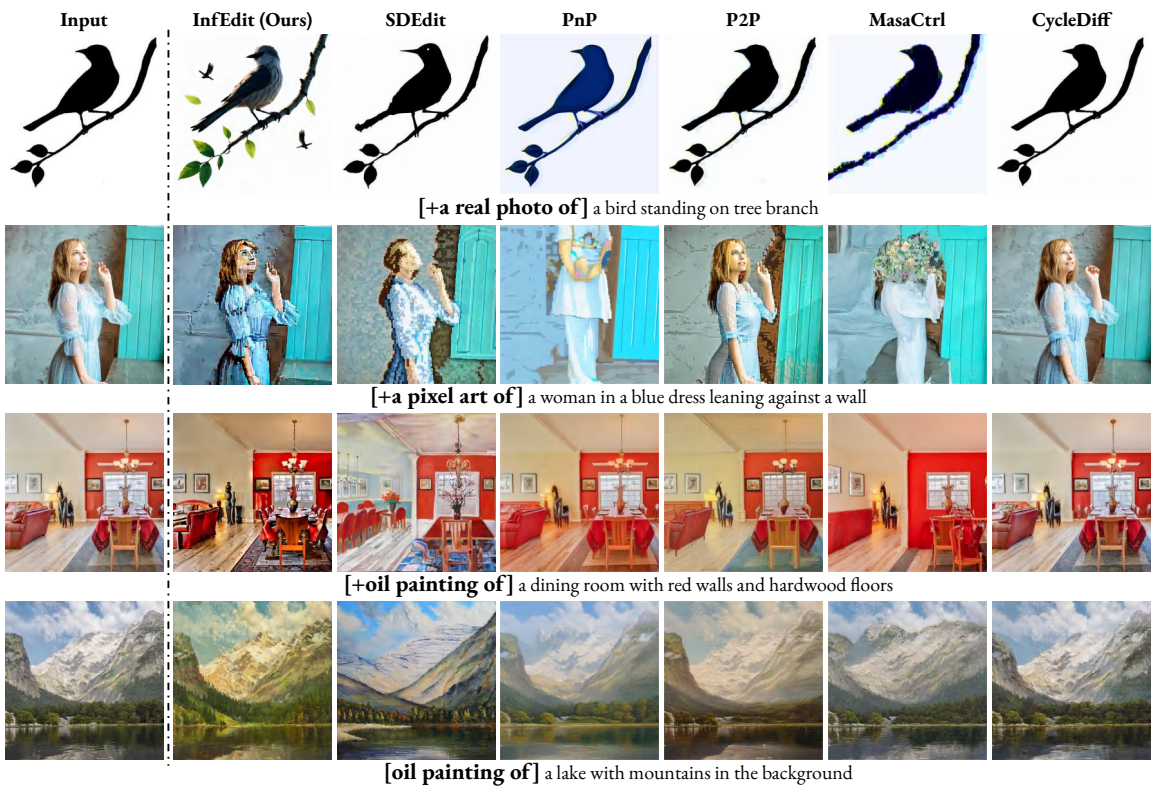


Figure 18. Additional comparison on changing style tasks.

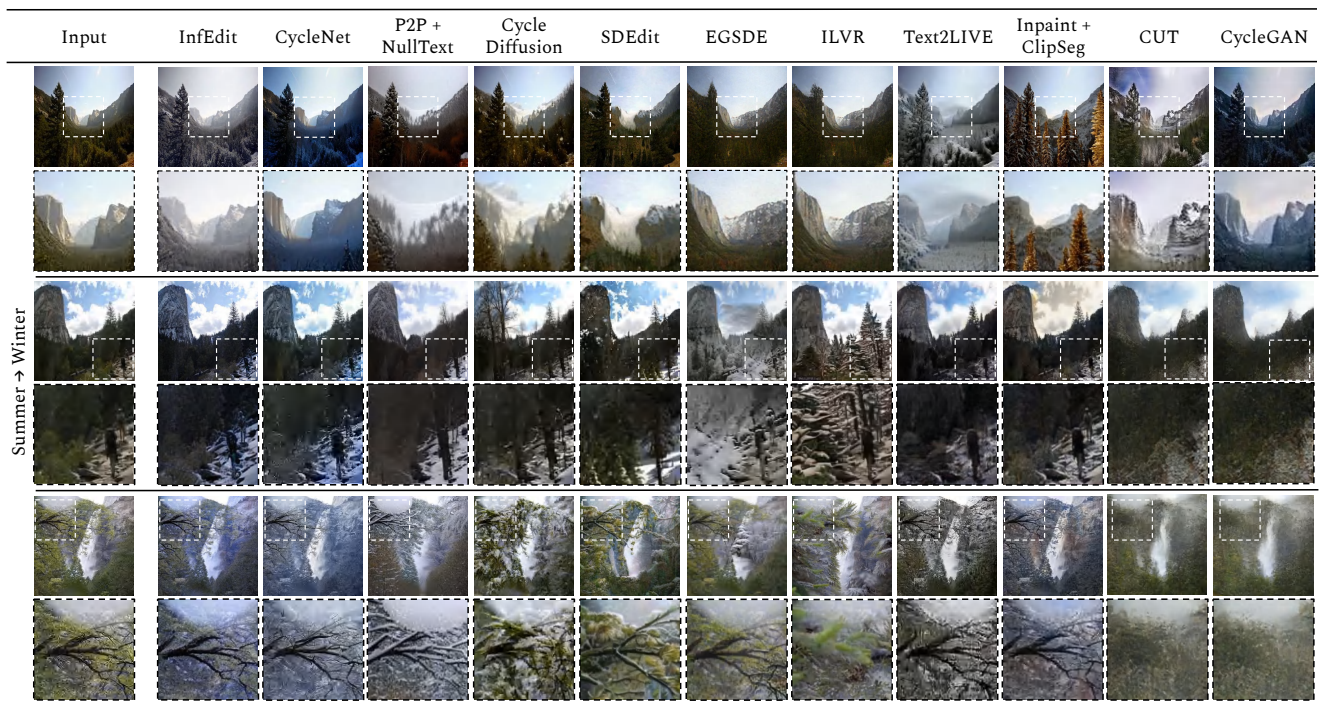


Figure 19. Qualitative comparison of InfEdit on Summer2Winter task with other baselines.

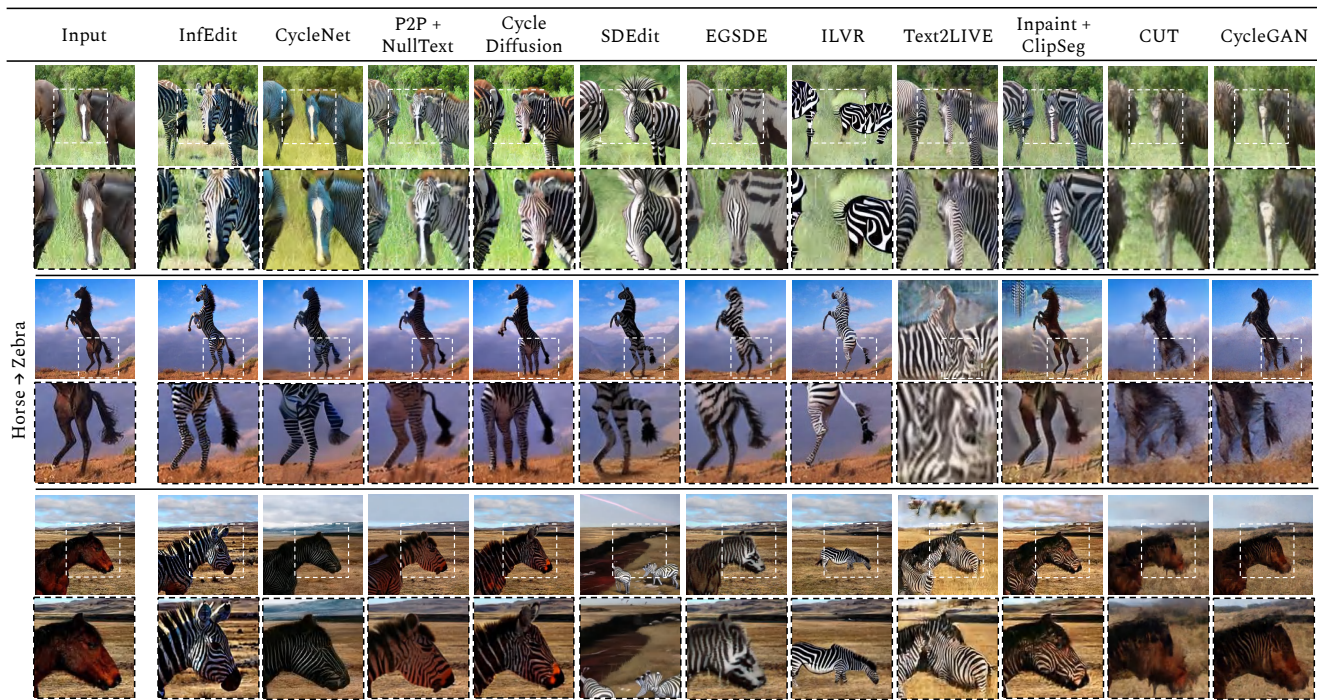


Figure 20. Qualitative comparison of InfEdit on Horse2Zebra task with other baselines.

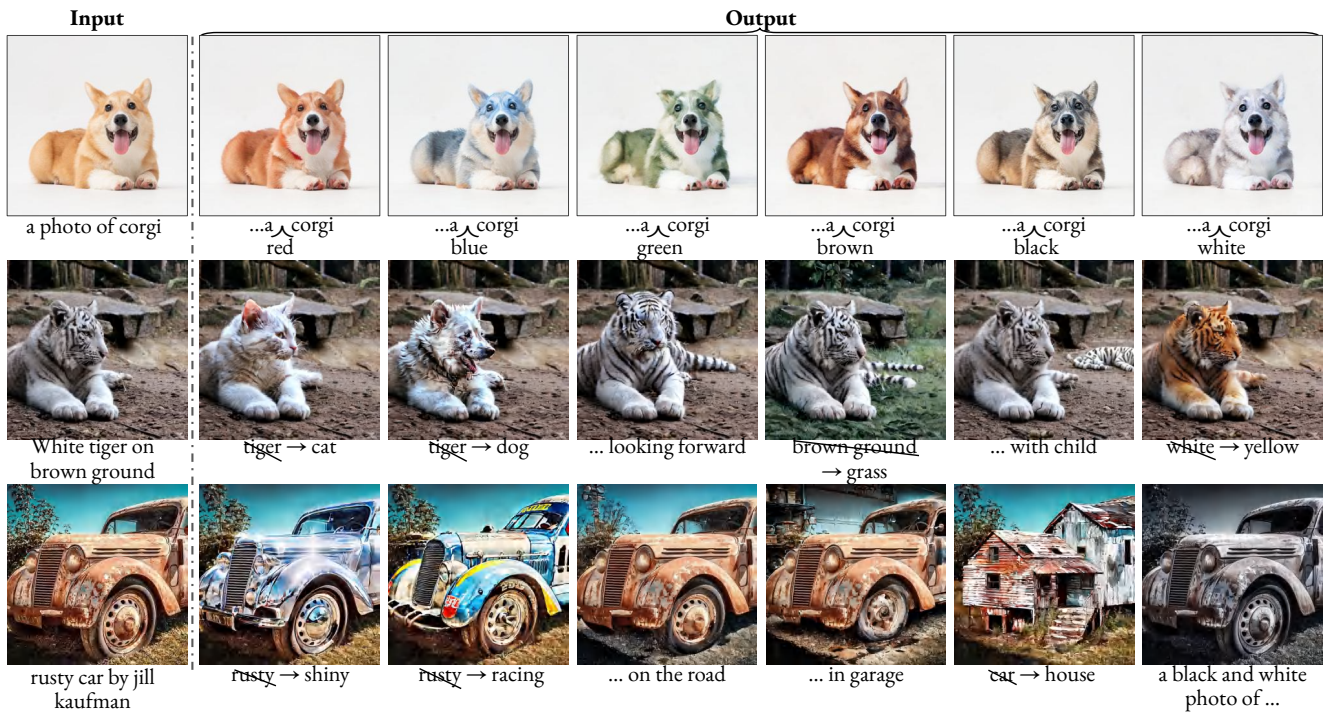


Figure 21. Additional results of InfEdit in various complex image editing tasks.

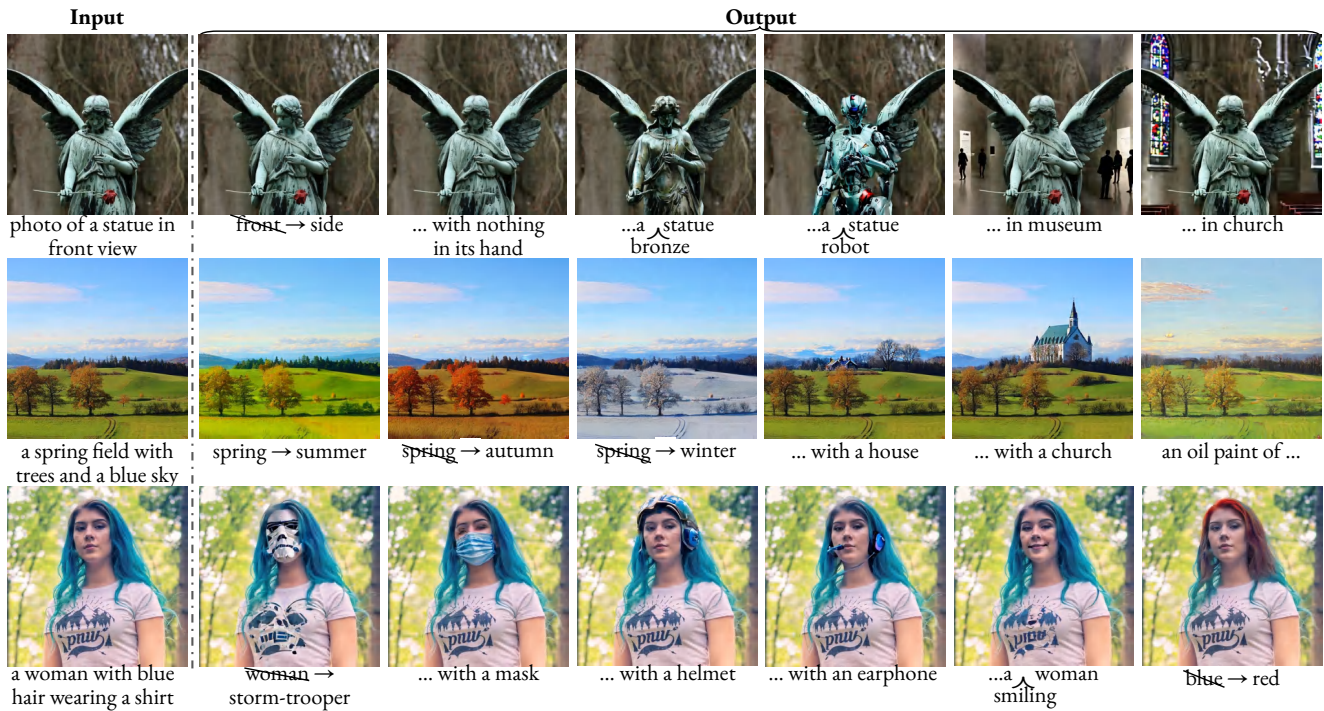


Figure 22. Additional results of InfEdit in various complex image editing tasks.

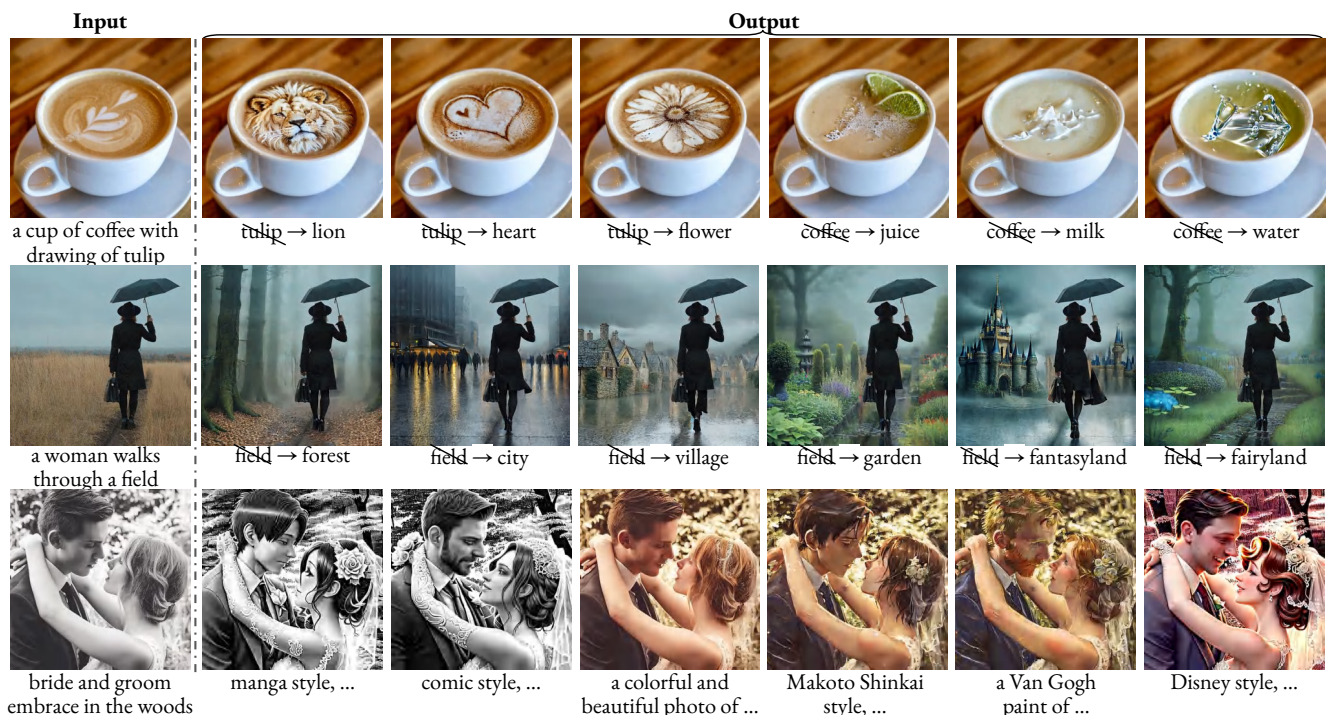


Figure 23. Additional results of InfEdit in various complex image editing tasks.



Figure 24. Additional results of InfEdit compared with other method.

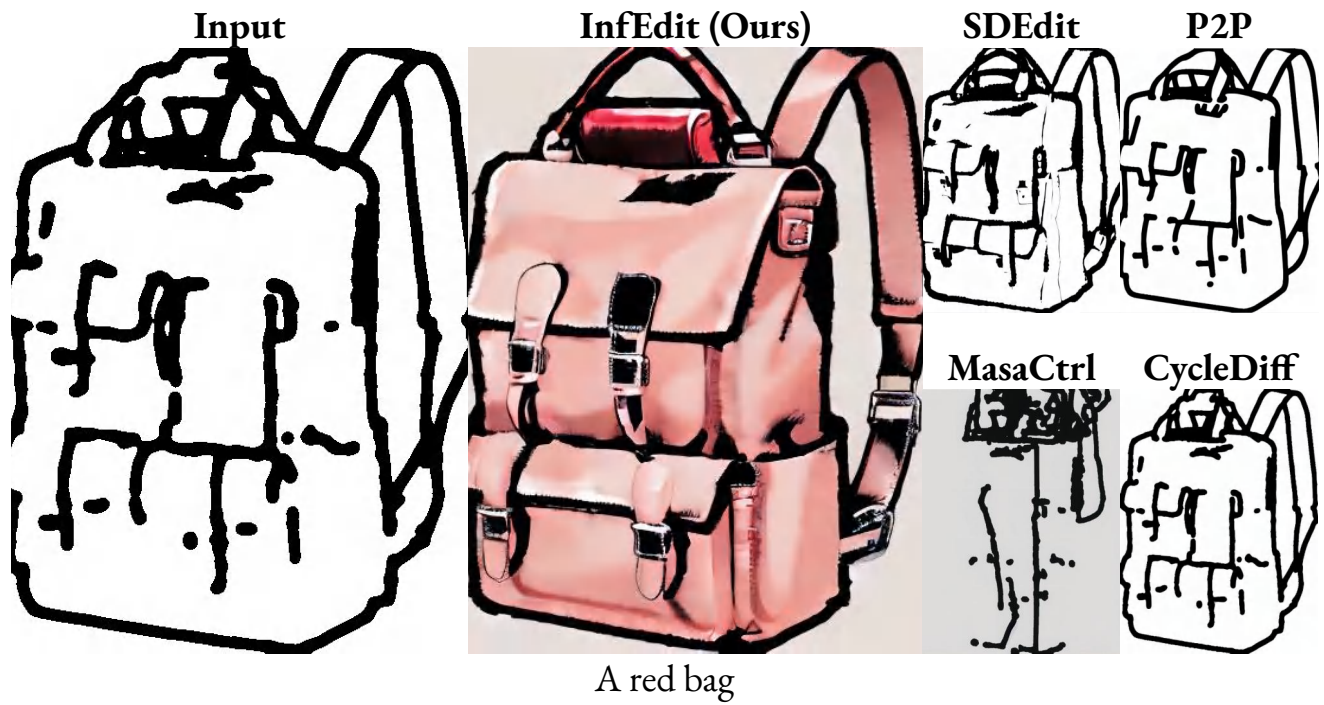


Figure 25. Additional results of InfEdit compared with other method.



Figure 26. Additional results of InfEdit in multi-modal editing.

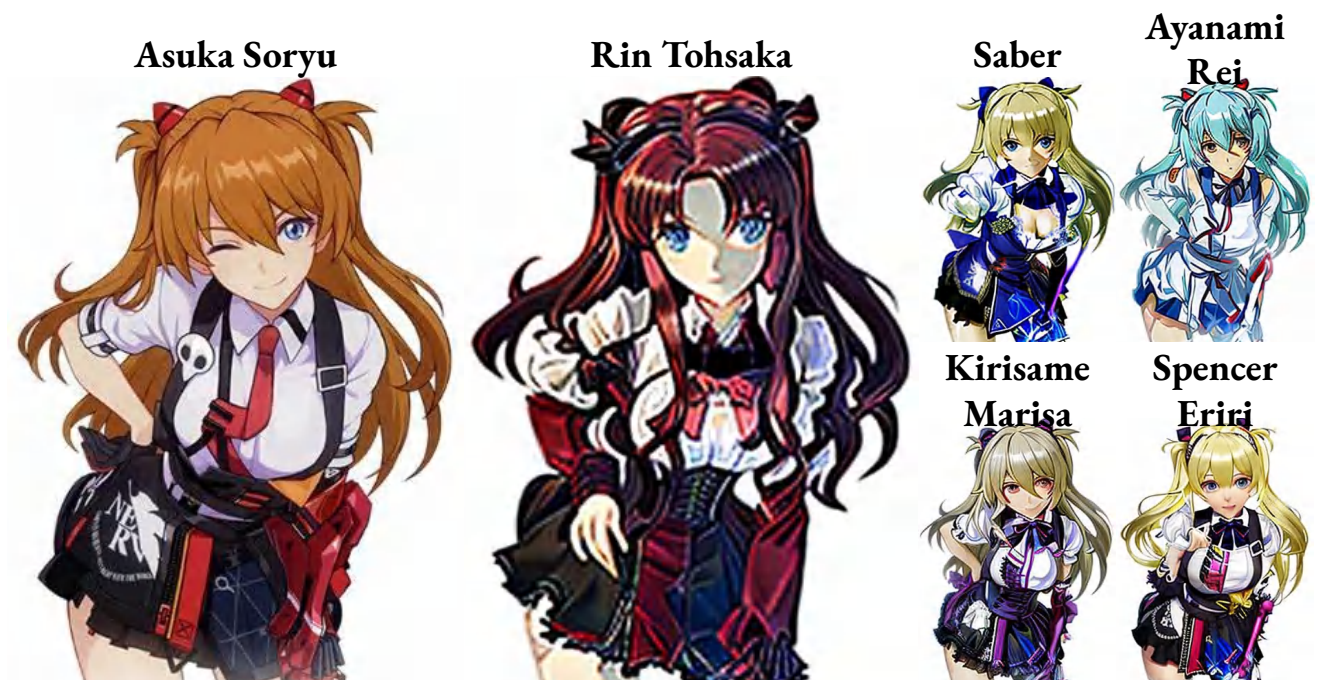



Figure 27. Additional results of InfEdit in multi-modal editing.



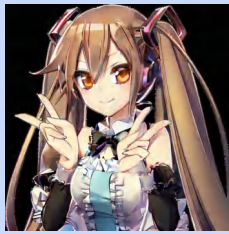
 a anime girl with ~~green~~ orange hair.



 a anime girl with ~~shirt~~ lace skirt.



 a anime girl with ~~green~~ red eyes.



 a anime girl with ~~smile~~ angry face.

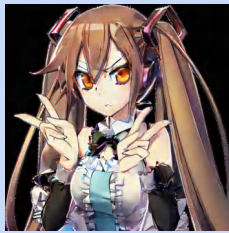


Figure 28. Multi-turn editing via InfEdit.