DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing (Appendix)

A. Details About DRAGBENCH Dataset

We have collected 205 images and provided 349 pairs of handle and target points in total. Images in our DRAGBENCH are classified into the following 10 categories: animals, art works, buildings (city view), buildings (countryside view), human (head), human (upper body), human (full body), interior design, landscape, other objects. All human-related images are selected from Midjourney generation results to avoid potential legal concerns. All the other images are real images downloaded from unsplash (https://unsplash.com/), pexels (https://www.pexels.com/zh-cn/), and pixabay (https://pixabay.com/). Some examples of our dataset is given in Fig. 1.

B. Links to the Stable Diffusion's Finetuned Variants Used by Us

Here, we provide links to the fine-tuned variants of Stable Diffusion used by us:
Counterfeit-V2.5 (https://huggingface.co/gsdf/Counterfeit-V2.5),
Majixmix Realistic (https://huggingface.co/emilianJR/majicMIX_realistic),
Realistic Vision (https://huggingface.co/SG161222/Realistic_Vision_V2.0),
Interior Design Supermix (https://huggingface.co/stablediffusionapi/interiordesignsuperm),
DVarch (https://huggingface.co/stablediffusionapi/dvarch).

C. More Details on Editing Diffusion-Generated Images

Here we introduce more details about editing diffusion-generated images. Firstly, different from editing real images, we **do not** need to conduct LoRA fine-tuning before latent optimization. This is because the purpose of LoRA fine-tuning is to help better encode the features of the original image into the diffusion UNet. However, for diffusion-generated images, the image features are already well-encoded as the diffusion model itself can generate these images. In addition, during the latent optimization stage, we do not have to perform DDIM inversion as the diffusion latents are readily available from the generation process of the diffusion models.

Another details we need to attend to is the presence of classifier-free guidance (CFG) when editing generated images. As described in the main text, when editing real images, we turn off the CFG as it pose challenges to DDIM inversion. However, when editing generated images, we inevitably have to deal with CFG, as it is one of the key component in diffusion-based image generation. CFG introduces another forward propagation pass of the UNet during the denoising process with a negative text embedding from null prompt or negative prompt. This makes a difference during our latent optimization stage, as now we have two UNet feature maps (one from the forward propagation with positive text embedding and the other one from the negative text embedding) instead of only one. To deal with this, we concatenate these two feature maps along the channel dimension and then use the combined feature maps to supervise latent optimization. This simple strategy have been proven to be effective as shown in our empirical results.

D. Execution Time

Given a real image with the resolution of 512×512, the execution time of different stages in DRAGDIFFUSION on a A100 GPU is as follows: LoRA fine-tuning is around 25 seconds, latent optimization is around 10 to 30 seconds depending on the magnitude of the drag-instruction, the final Latent-MasaCtrl guided denoising is negligible comparing to previous steps (about 1 to 2 seconds)

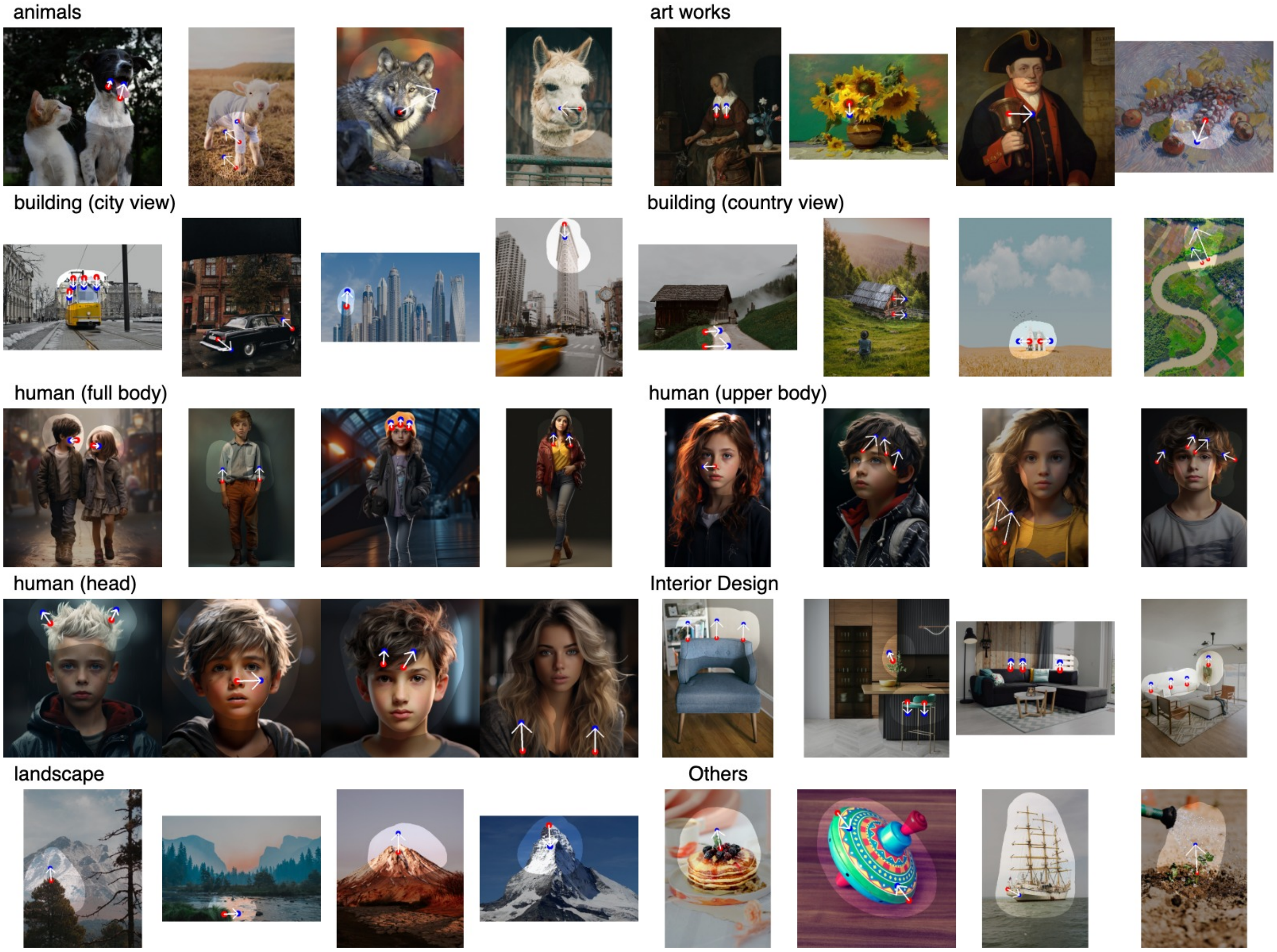

Figure 1. Examples of our DRAGBENCH dataset. Each image is accompanied by a set of drag-based editing instruction.

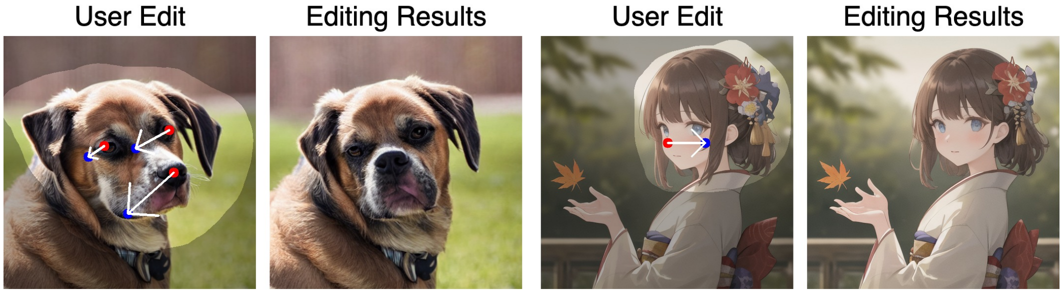

Figure 2. Limitation of DRAGDIFFUSION. Occasionally, some of the handle points cannot precisely reach the desired target.

E. Limitations

As shown in Fig. 2, the limitation of our DRAGDIFFUSION is that, occasionally, some of the handle points cannot precisely reach the desired target. This is potentially due to inaccurate point-tracking or difficulties in latent optimization when multiple pairs of handle and target points are given. We leave making the drag-based editing on diffusion models more robust and reliable as our future work.

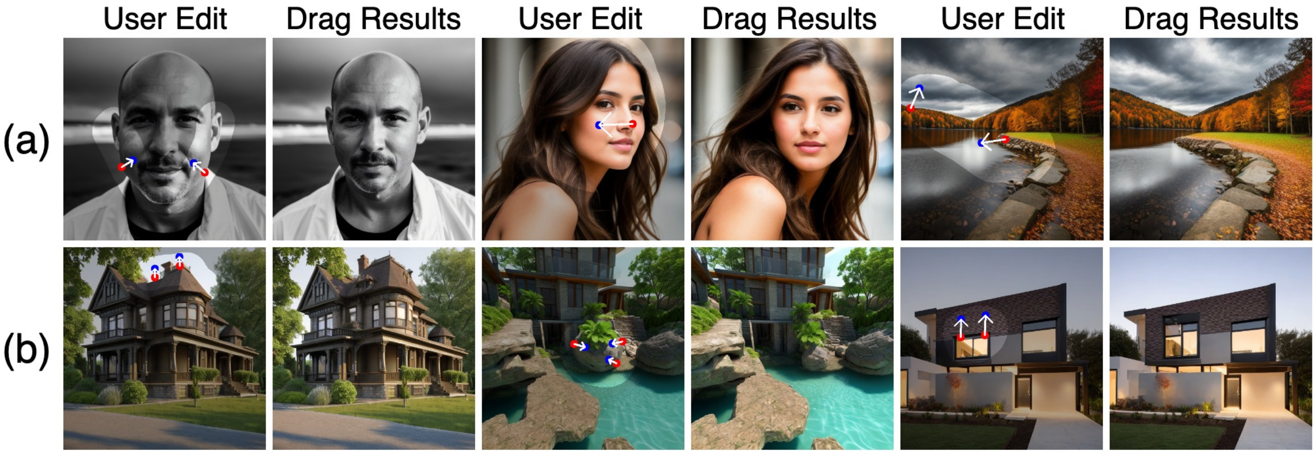

Figure 3. Editing results on diffusion-generated images with **(a)** Realistic Vision; **(b)** DVarch.

F. More Qualitative Results

To start with, we provide more qualitative comparisons between DRAGDIFFUSION and DRAGGAN in Fig. 4. These results consistently showing that our approach demonstrate much better versatility than DRAGGAN. In addition, we provide comparisons with DRAGGAN on samples from StyleGAN's training dataset in Fig. 6. Our results show that even in StyleGAN's training domain, our approach outperforms DRAGGAN. We posit this is primarily because diffusion models behave in a more robust manner when dealing with real input images.

Next, we demonstrate results of applying DRAGDIFFUSION on images generated by two more fine-tuned variants of Stable-Diffusion-1.5, namely Realistic-Vision and DVarch. Results are shown in Fig. 3. These results along with the results in the main text corroborate the generality of our approach on different diffusion models.

Also, we provide more results on generated images beyond the 512×512 resolution as in the main text. These results are shown in Fig. 5, which further demonstrate the versatility of DRAGDIFFUSION.

Finally, examples of dragging across the diagonal of an image and examples with SDXL are provided in Fig. 7 and Fig. 8, respectively.

G. Visual Ablation on the Number of Identity-preserving fine-tuning steps

In the main paper, we ablate on the effect of the number of identity-preserving fine-tuning steps (denoted by n in this section). We show through numerical experiments that $n \geq 80$ produce ideal results in Fig. 7 (b) of the main text. In this section, we provide visualization that corroborate with conclusions in the main text, showing setting $n \geq 80$ produces editing results that are free from artifacts such as distorted faces and scenes, unexpected hands, *etc*. Results are shown in Fig. 9.

H. Visual Ablation on the UNet Feature Maps

In the main paper, we have studied the effect of using UNet feature maps produced by different blocks of UNet *decoder* for our approach. In this section, we provide visualization that corroborates with conclusions in the main text. Results are shown in Fig. 10. According to the results, using the 1-st block of feature maps will lead to unfavorable preservation of local details due to lack of fine-grained information. This corresponds to the low Image Fidelity (IF) and high Mean Distance (MD) as in main text Fig. 7 (c) when Block number is 1.

On the other hand, as the 4-th block of UNet feature maps only contains low-level information, the editing results is almost the same as the original real image, indicating ineffective editing. This corresponds to the high IF and high MD as in main text Fig. 7 (c) when Block number is 4.

Finally, using the 2-nd or the 3-rd block of UNet feature maps can can yield reasonable editing. However, if observing more closely, we can see that using the 3-rd block of features yields slightly better preservation of local details (*e.g*. more reasonable headwrap in Fig. 10 (a) and better details of buildings by the river in the Fig. 10 (b)). Correspondingly, in main text Fig. 7 (c), we also show using UNet feature maps output by the 3-rd block can yield better results (lower MD and higher IF).

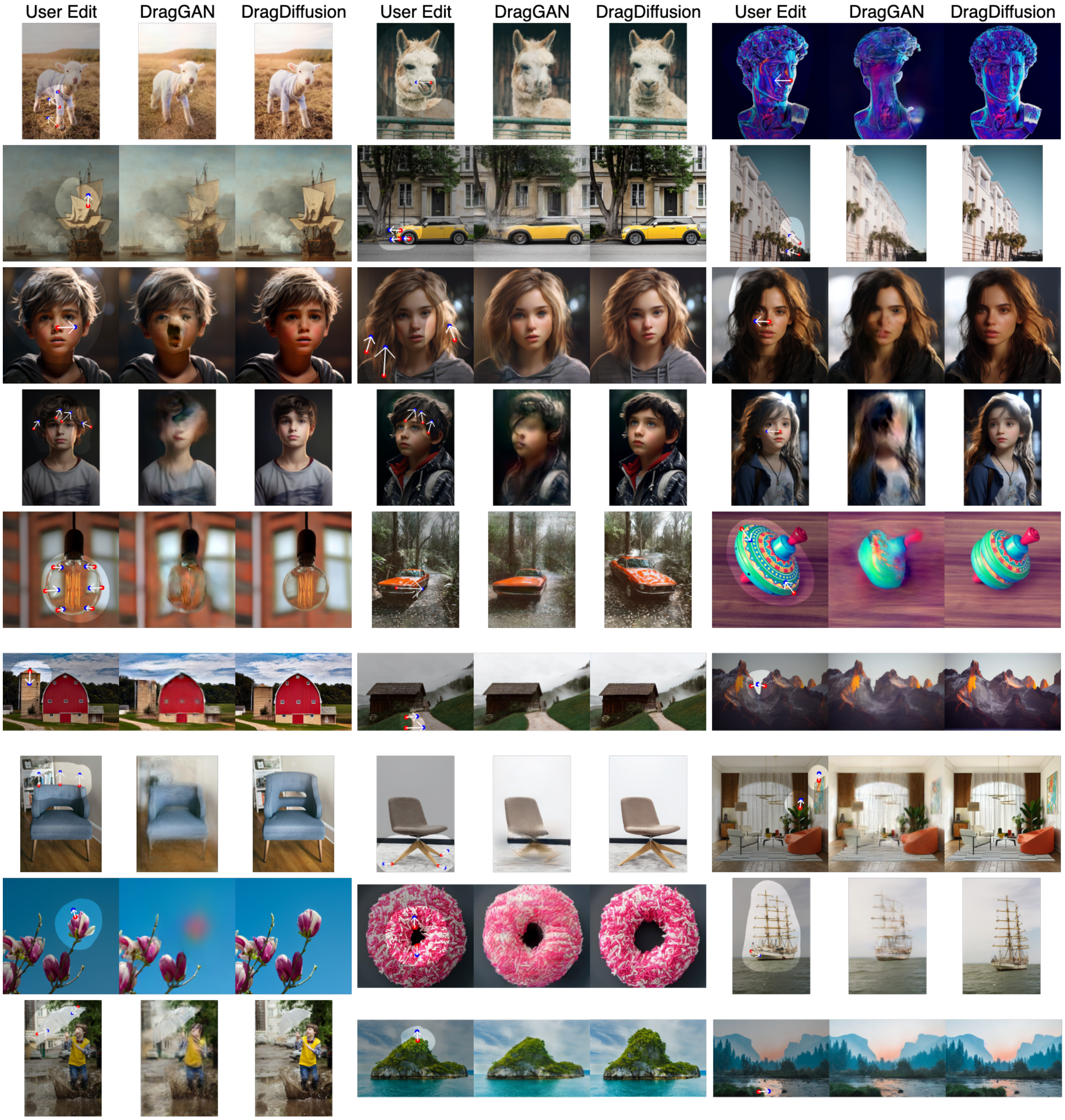

Figure 4. Additional comparisons between DRAGGAN and DRAGDIFFUSION. All images are from our DRAGBENCH dataset. Both approaches are executed under the same drag-based editing instruction. **Zoom in to check the details.**

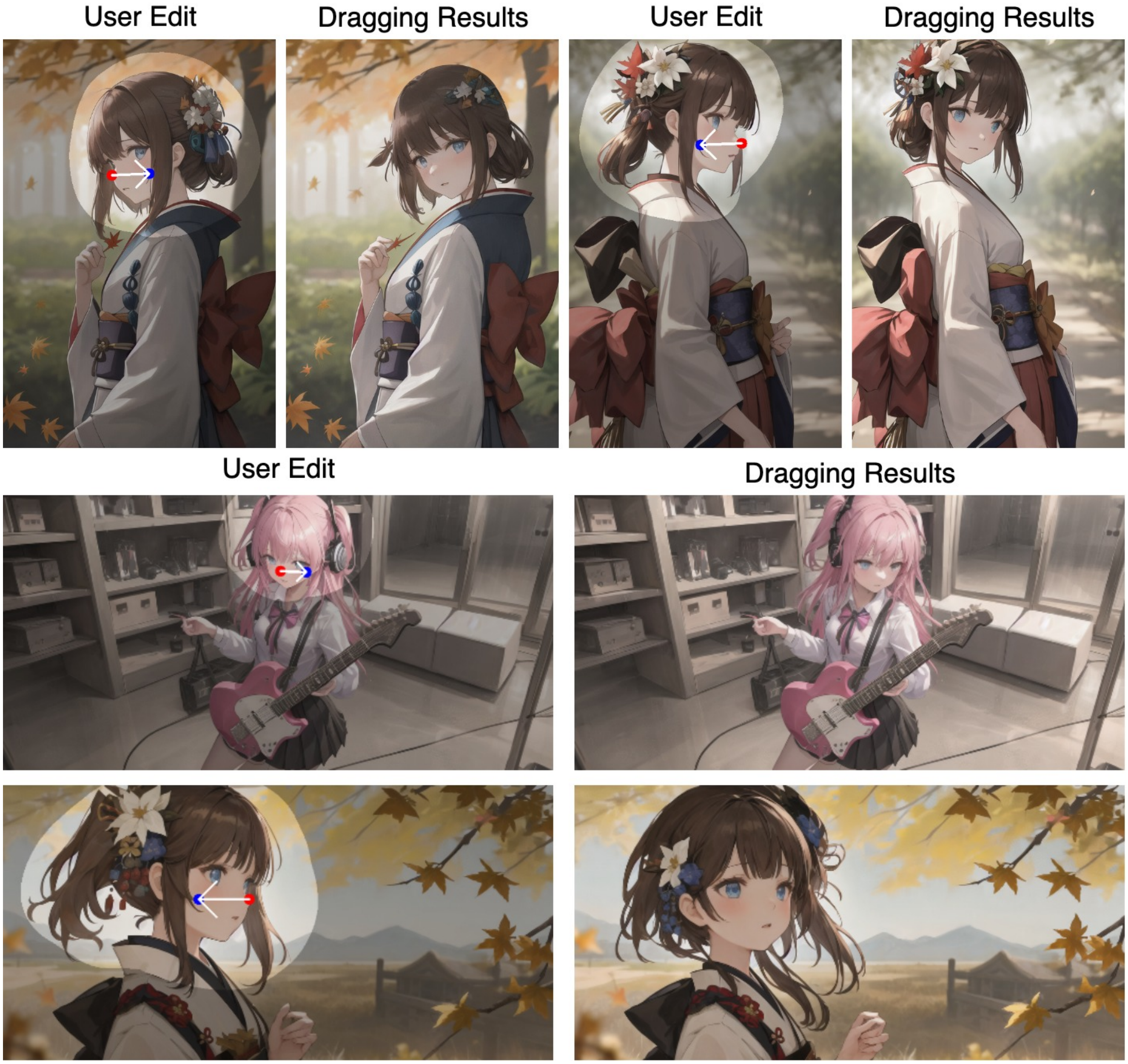

Figure 5. Editing results from DRAGDIFFUSION beyond 512 × 512 resolution. Results are produced by perform drag-based edits on images generated by Counterfeit-V2.5. The resolution of images in the first row are 768 × 512, while the images in the second row are 512 × 1024.

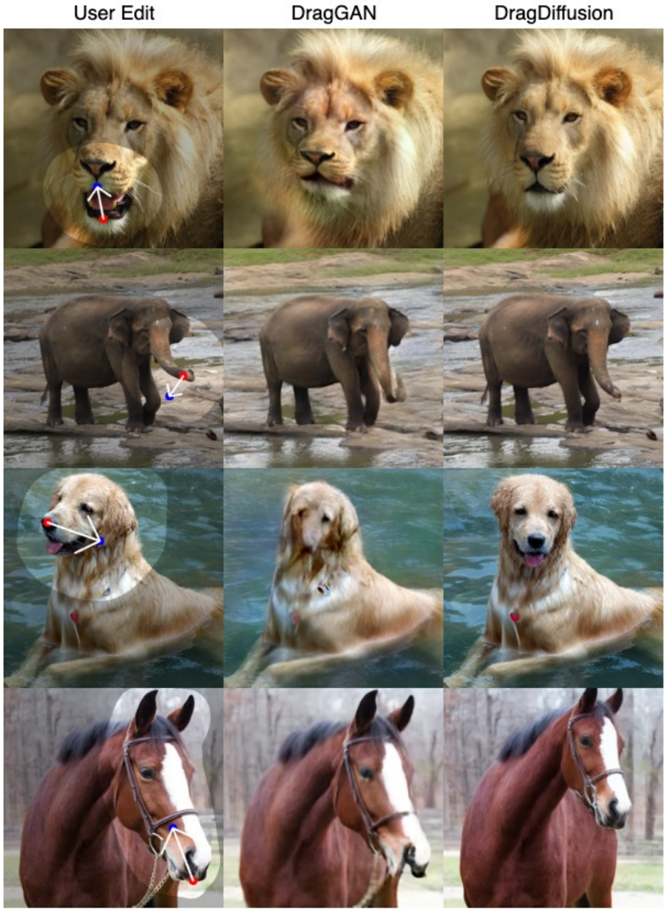

Figure 6. **"In domain comparisons" with DragGAN.**

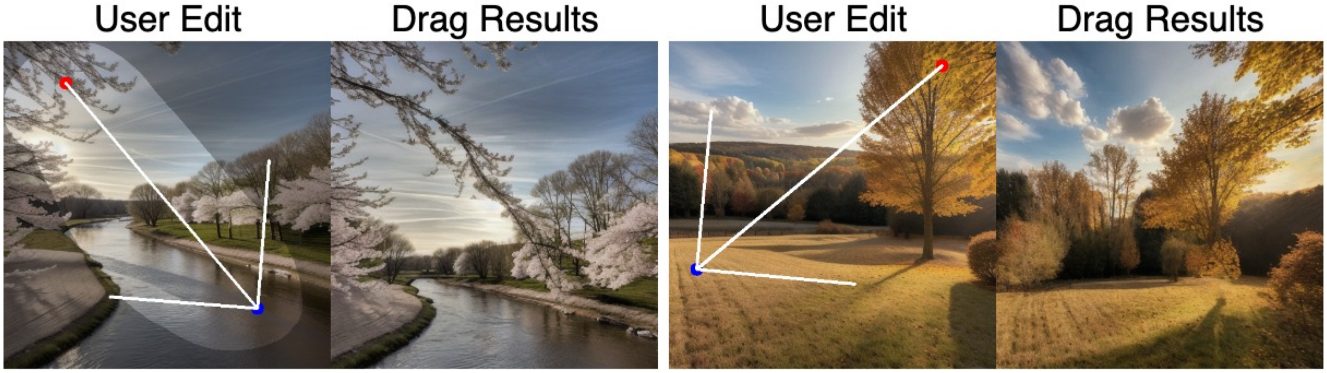

Figure 7. **Dragging across the diagonal.**

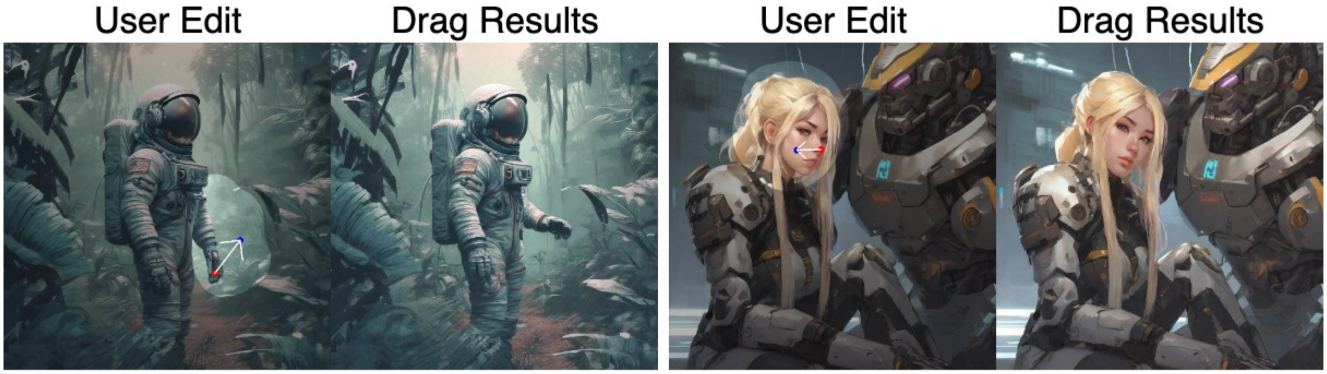

Figure 8. **Results on SDXL (resolution:** 1024×1024**).**

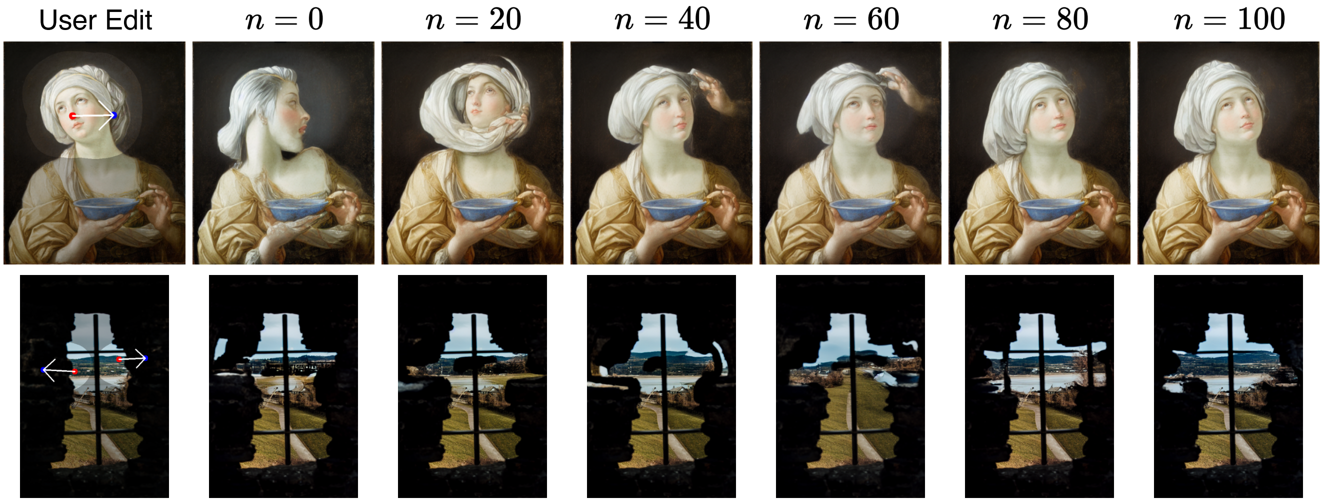

Figure 9. Visual ablation study on the number of **identity-preserving fine-tuning steps (denoted as** n**). Zoom in to view details.** From the visualization, we see that setting $n < 80$ can produce undesired artifacts in the dragging results (*e.g*., distorted faces and scenes, unexpected hands, *etc*.). On the other hands, $n \geq 80$ normally produces reasonable results without artifacts.

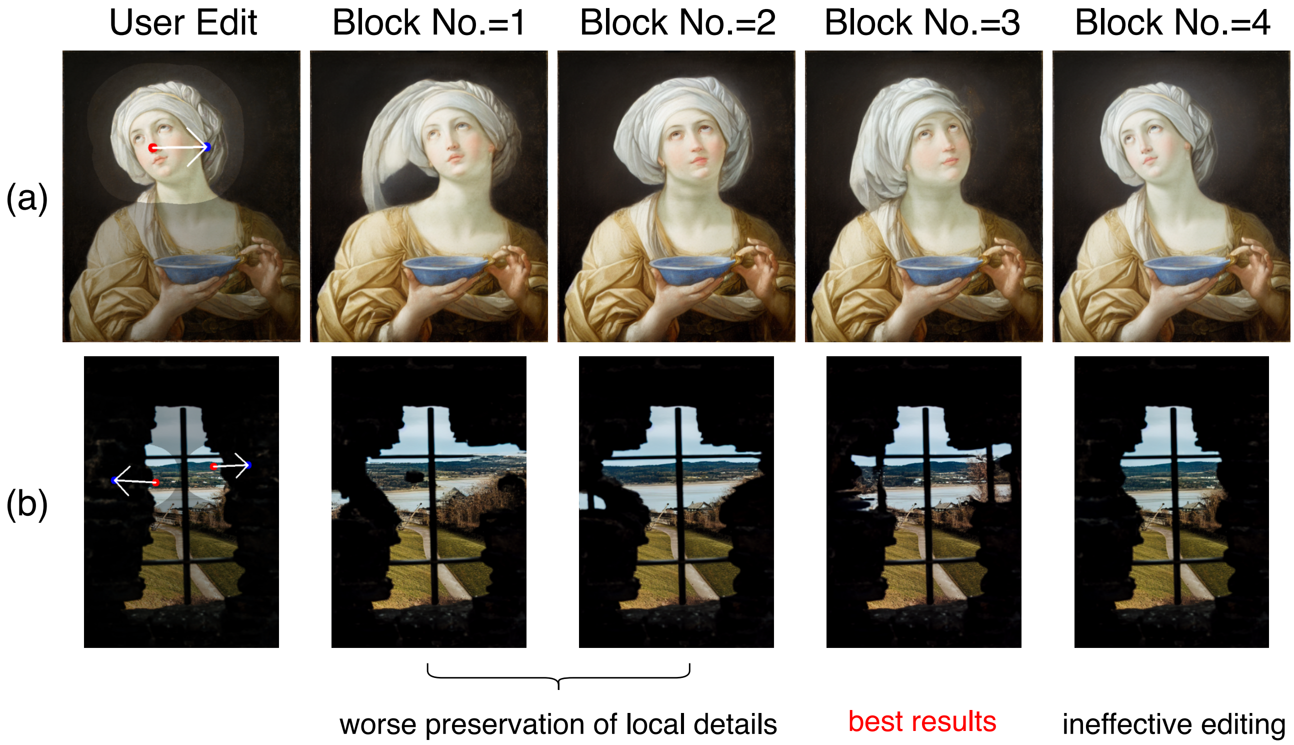

Figure 10. Visual ablation study on the block number of the UNet feature map. **Zoom in to view details.** As in the figure, using feature maps of the 2-nd and 3-rd blocks produce reasonable results. However, if observing more closely, we can see that using the 3-rd block of features yields slightly better preservation of local details (*e.g*. more reasonable headwrap in **(a)** and better details of buildings by the river in **(b)**).

	art works	landscape	city	countryside	animals	head	upper body	full body	interior design	other objects
DRAGGAN	0.71	0.84	0.74	0.79	0.72	0.91	0.33	0.31	0.57	0.71
DRAGDIFFUSION	0.88	0.88	0.89	0.88	0.87	0.85	0.89	0.95	0.90	0.87

Table 1. **Comparisons of Image Fidelity (1-LPIPS) on DRAGBENCH on each category (↑).**

	art works	landscape	city	countryside	animals	head	upper body	full body	interior design	other objects
DRAGGAN	59.51	47.60	41.94	46.96	60.12	65.14	82.98	37.01	75.65	58.25
DRAGDIFFUSION	30.74	36.55	26.18	43.21	39.22	36.43	39.75	20.56	24.83	39.52

Table 2. **Comparisons of Mean Distance on DRAGBENCH on each category (↓).**

I. Detailed Comparisons on DRAGBENCH by Category

In the main paper Fig. 8, we report the Mean Distance (MD) and Image Fidelity (IF) averaging over all samples in DRAGBENCH. In this section, we provide detailed comparisons between DRAGGAN and DRAGDIFFUSION on each category in DRAGBENCH. Comparisons in terms of IF (*i.e.*, 1-LPIPS) and MD are given in Tab. 1 and Tab. 2, respectively. According to our results, DRAGDIFFUSION significantly outperforms DRAGGAN in every categories of DRAGBENCH.