[27] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2, 4, 6

[28] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22873–22882, 2023. 2, 3

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 3

Contrastive Denoising Score for Text-guided Latent Diffusion Image Editing

Supplementary Material

Method	Metric		
	CLIP Acc (↑)	Dist (↓)	LPIPS (↓)
SDEdit + word swap	99.2%	0.066	0.126
DiffuseIT	99.2%	0.066	0.255
DDPM inv + P2P	85.7%	0.073	0.147
DDPM inv + PnP	86.1%	0.078	0.166
DDS	**100**%	0.031	0.091
Ours	**100**%	**0.027**	**0.088**

Table 3. Quantitative evaluation for the cat → pig task. 'Dist' denots DINO-ViT structure distance.

A. Implementation details

For implementation, we referenced the official source code of Delta Denoising Score[2] by using Stable Diffusion v1.4. We extracted intermediate features from self-attention layers and apply PatchNCE loss similar to CUT. Inspired by the analysis of CUT, we applied PatchNCE to all the upsampling self-attention layers but excluded features from the U-Net bottleneck layer, as it is related to the overall semantics of the images. For hyperparameters, we utilizes patch sizes of 1×1 or 2×2 with 256 patches and set the weight of ℓ_{con} to 3.0. Other settings regarding to DDS, including the number of optimization steps, optimizer, and learning rate, adhere to the default configurations provided in the official code. The code is available to public on https://github.com/HyelinNAM/CDS. All image manipulations were conducted using an NVIDIA RTX 6000, and the processing time for editing each image was approximately 2 minutes and 50 seconds.

B. Quantitative Results

In addition to the quantitative results presented in the main paper, Tab. 3 also demonstrates the effectiveness of our proposed method in achieving optimal editing outcomes while maintaining the structural elements of the source image. In contrast, other baseline models achieve low structure distance. This implies that despite achieving high CLIP accuracy, the existing methods edited the image without considering the source structure. The visual comparison for both tasks(see Fig. 10) also confirms the aforementioned quantitative evaluation.

[2] https://github.com/google/prompt-to-prompt/blob/main/DDS_zeroshot.ipynb

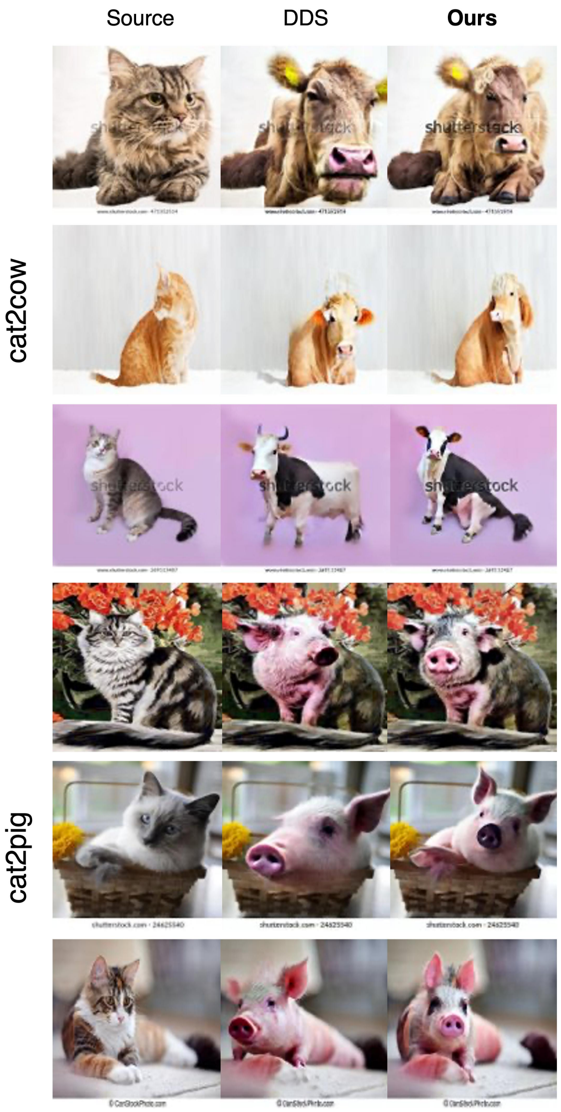

Figure 10. Sample results of the cat2dog task from DDS and CDS.

C. Additional Ablation studies

C.1. Patch size

First, we evaluate the impact of patch size by varying its size. As shown in Fig. 11a, we observe that the patch size has an effect on the extent of content preservation. As we are regulating the latent, which is more compact than image pixels, utilizing small patch size shows better impact on preserving structural elements and background details. Therefore, we decide to use a small patch, specifically 1×1 and 2×2, to align with our objectives.

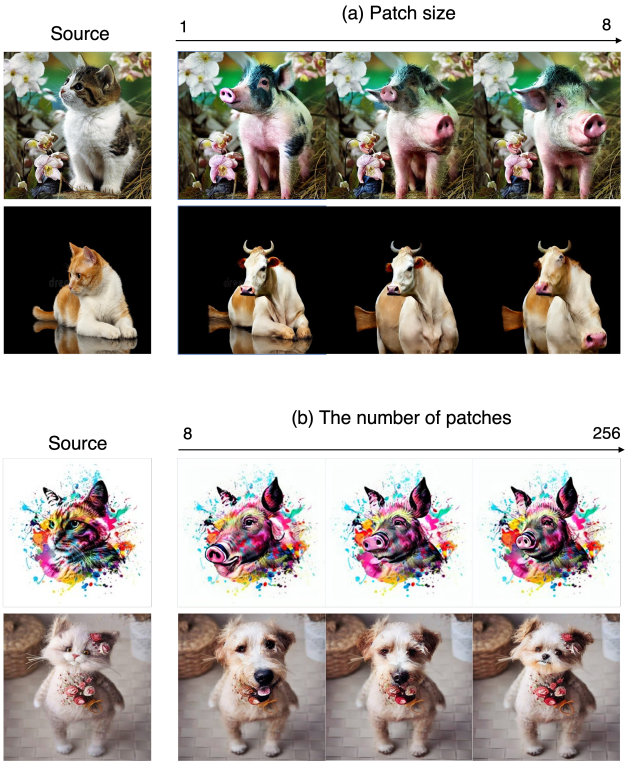

Figure 11. Ablation study on (a) patch size and (b) the number of patches. The given prompt is “cat → pig,” “cat → cow,”, “cat → pig” and ”cat → dog” from top to bottom.

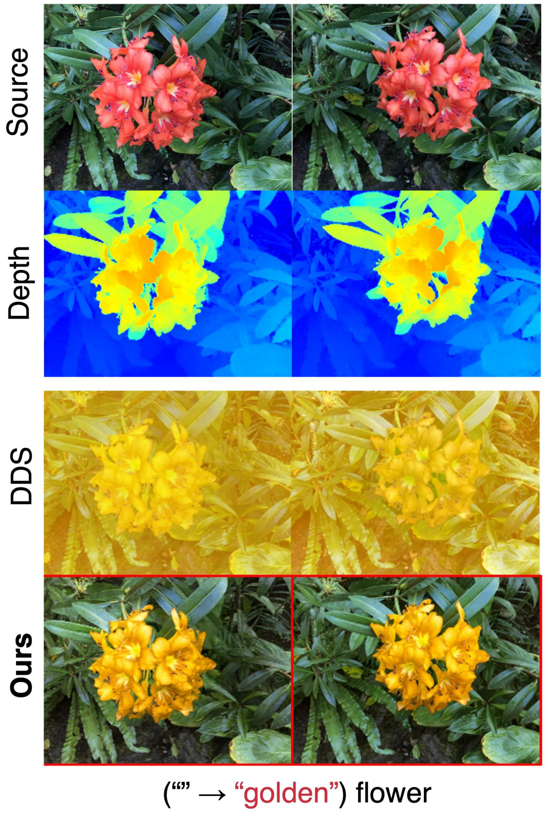

Figure 12. Additional results on NeRF editing.

C.2. The number of patches

We also ablate the impact of the number of patches and found that it also determines the extent of the regulation. As the number of patches increases, we observe a better preservation of structural aspects of the original image, such as facial structure and head angles (see Fig. 11b). Therefore, we chose 256 number of patches.

D. Additional results

D.1. Qualitative results

In Figs. 13 and 14, we show our edited outputs with various images and prompts. The results clearly demonstrate that our method can be applied not only to changing objects but also to diverse cases, such as adding a smile or altering gender. The proposed framework is capable of performing the edits while still retaining the other details, such as background details.

D.2. NeRF editing

We further provide additional results on NeRF editing. To fine-tune the NeRF model, we utilized the recent pre-trained model, TensoRF [4]. For efficiency, we downsampled the images to a resolution of 504x378 in pre-training stage. For fine-tuning the pre-trained model, we further downsampled the resolution to 252x189 due to resource constraints.

For fine-tuning, we rendered source and target images from the pre-trained source NeRF model ϕ and fine-tuned model θ, respectively. With the same view direction d, we obtain two rendered view $\hat{x}, x$, representing rendered 2D images from source and target model, respectively. By embedding these two images into the encoder of the Stable Diffusion model, we obtain source and target latent $\hat{\boldsymbol{z}}, \boldsymbol{z}$. With the prepared latents, we can calculate DDS gradient as follows:

$$\mathcal{L}_{\text{DDS}}(\theta, y_{trg}) = \nabla_\theta \mathcal{L}_{\text{SDS}}(\boldsymbol{z}, y_{trg}) - \nabla_\theta \mathcal{L}_{\text{SDS}}(\hat{\boldsymbol{z}}, y_{src}). \tag{7}$$

We also utilize our proposed contrastive loss along with the aforementioned DDS gradient to update the NeRF parameter θ. During training, we used the Adam optimizer with a learning rate of 0.01, and conducted 400 iterations for fine-tuning. The overall process takes about 8 minutes per sample.

In Fig. 12, we show the comparison results between the baseline DDS and CDS. We observed that basic DDS model struggles to accurately capture the shape of the original 3D object, often resulting in changes to the entire color tones. In contrast, our method enables localized object editing without compomising the shape of the original object.

E. Limitations and Negative social impact

Given that our framework manipulates images based on user intentions, there is a potential for misuse, including the creation of deepfakes or other forms of disinformation. Additionally, our method has dependency on generative priors of a large text-to-image diffusion models, which may contain undesired biases. Therefore, ensuring ethical implementation and appropriate regulation are imperative for these methods.

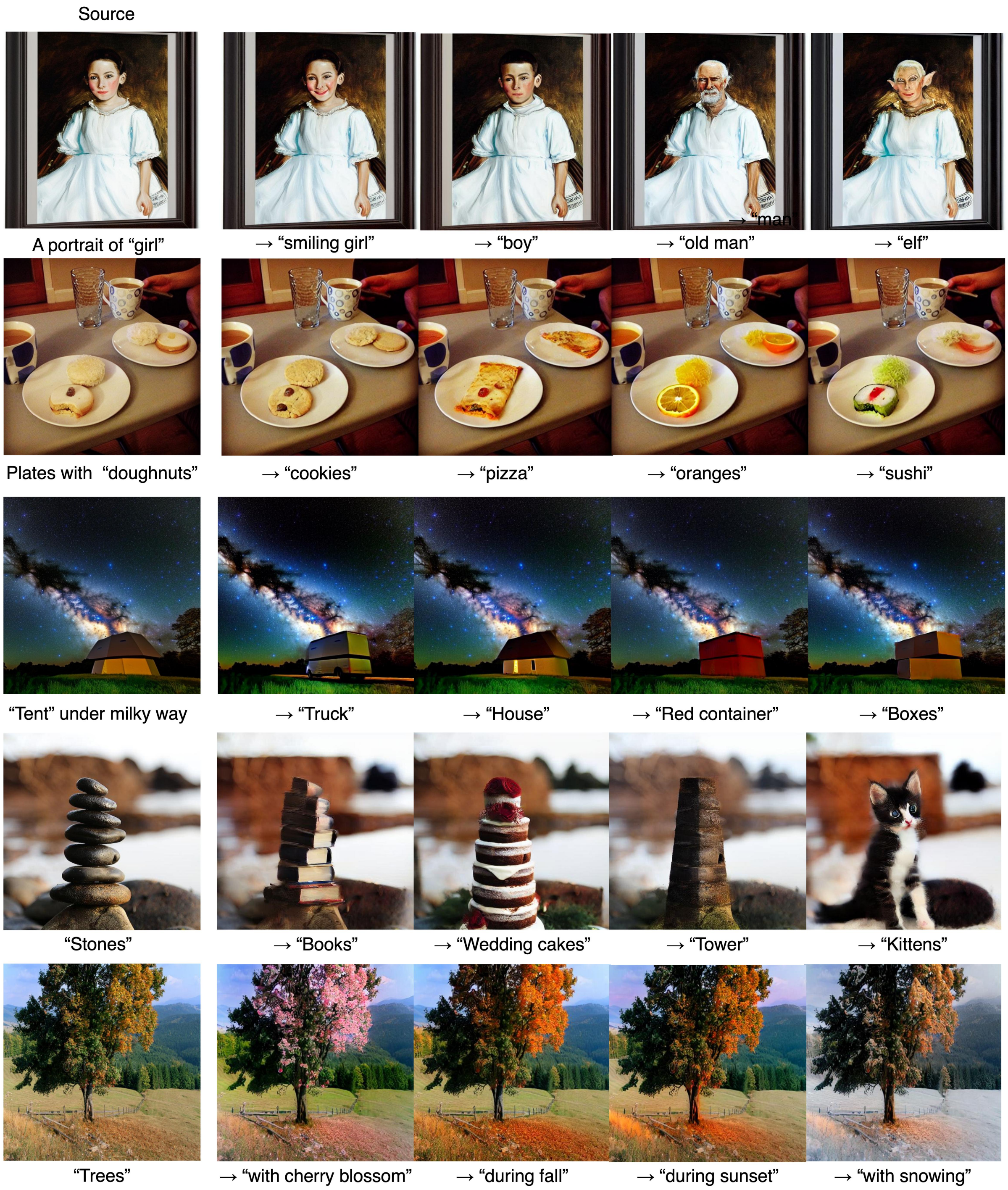

Figure 13. Additional qualitative results with various images and prompts.

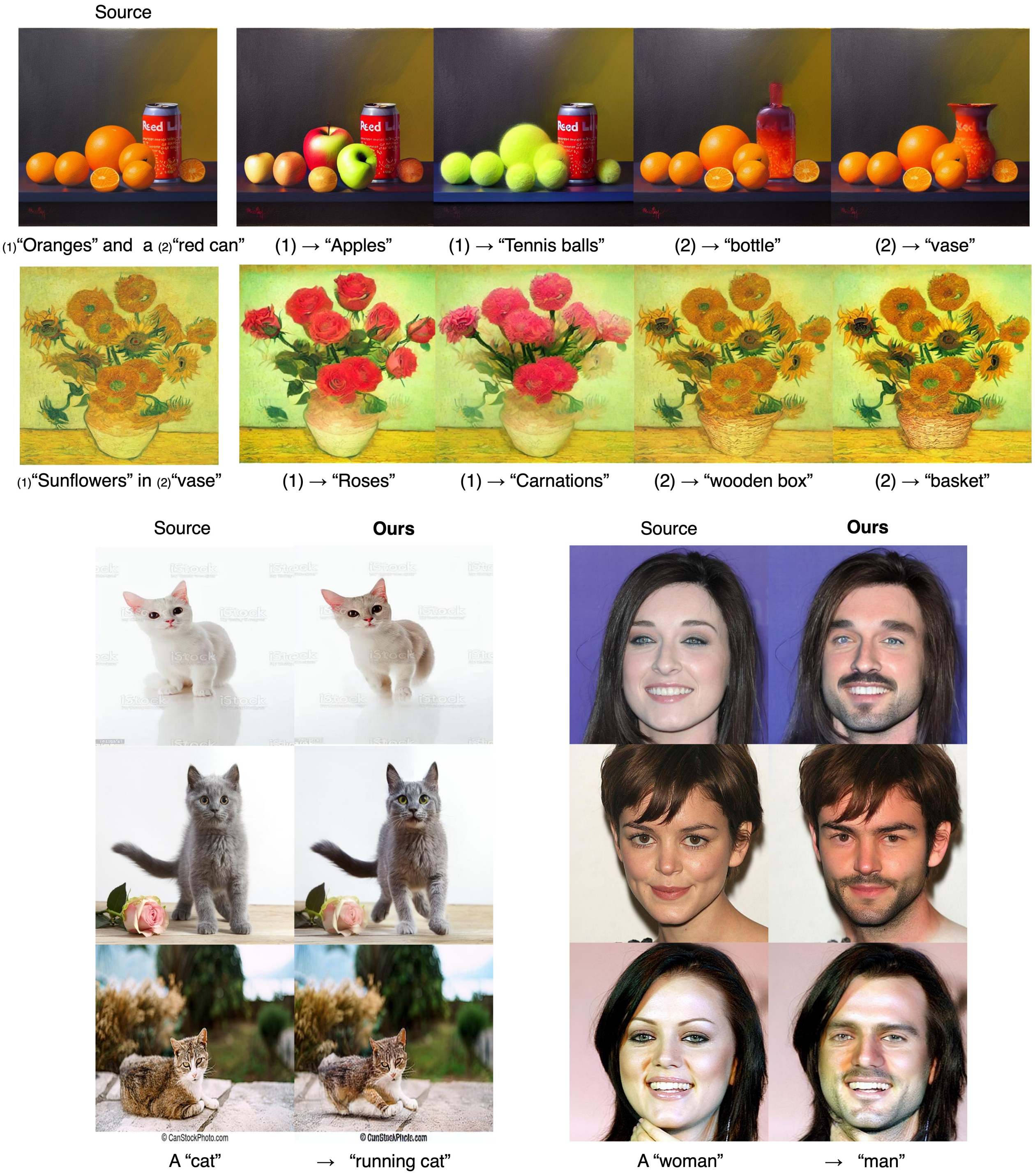

Figure 14. Additional qualitative results with various images and prompts.