arXiv:2311.18608v1 [cs.CV] 30 Nov 2023

Contrastive Denoising Score for Text-guided Latent Diffusion Image Editing

Hyelin Nam[1], Gihyun Kwon[2], Geon Yeong Park[2], Jong Chul Ye[1]
Kim Jae Chul Graduate School of AI[1], Dept. of Bio and Brain Engineering[2], KAIST
{hyelin.nam, pky3436, cyclomon, jong.ye}@kaist.ac.kr

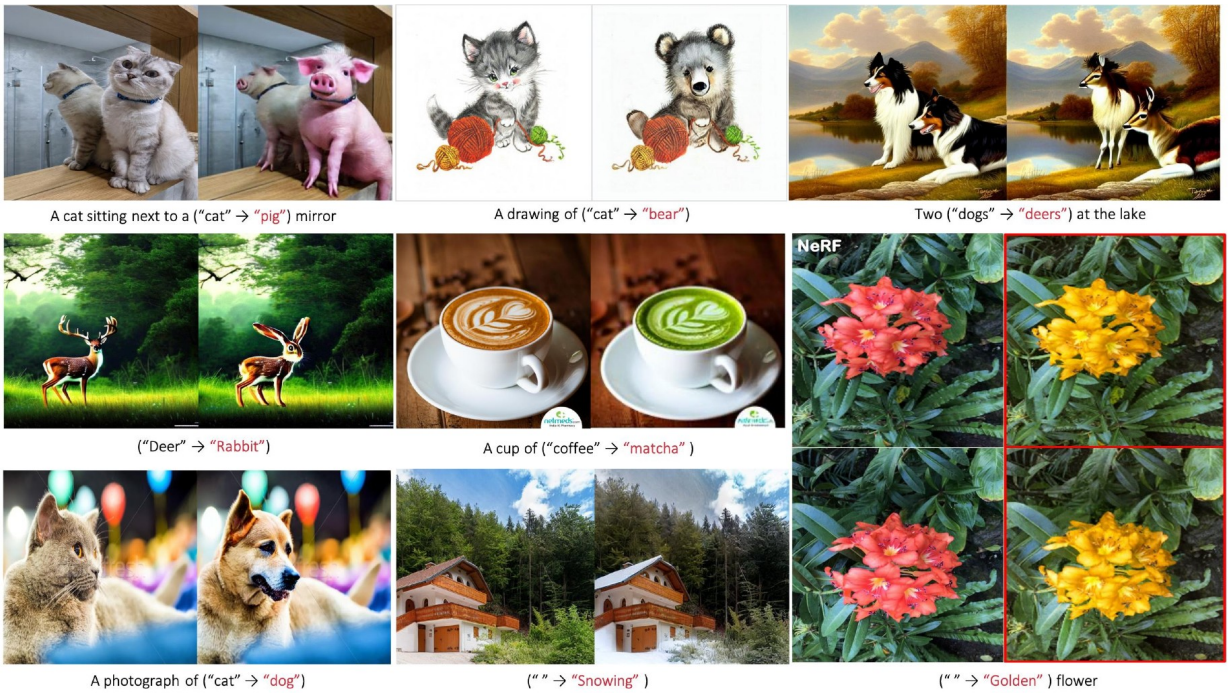

Figure 1. Text-guided Image Editing results. Our method successfully translates source images with a well-balanced interplay between maintaining the structural elements of the source image and transforming the content in alignment with the target text prompt.

Abstract

With the remarkable advent of text-to-image diffusion models, image editing methods have become more diverse and continue to evolve. A promising recent approach in this realm is Delta Denoising Score (DDS) - an image editing technique based on Score Distillation Sampling (SDS) framework that leverages the rich generative prior of text-to-image diffusion models. However, relying solely on the difference between scoring functions is insufficient for preserving specific structural elements from the original image, a crucial aspect of image editing. Inspired by the similarity and importance differences between DDS and the contrastive learning for unpaired image-to-image translation (CUT), here we present an embarrassingly simple yet very powerful modification of DDS, called Contrastive Denoising Score (CDS), for latent diffusion models (LDM). Specifically, to enforce structural correspondence between the input and output while maintaining the controllability of contents, we introduce a straightforward approach to regulate structural consistency using CUT loss within the DDS framework. To calculate this loss, instead of employing auxiliary networks, we utilize the intermediate features of LDM, in particular, those from the self-attention layers, which possesses rich spatial information. Our approach enables zero-shot image-to-image translation and neural radiance field (NeRF) editing, achieving a well-balanced interplay between maintaining the structural details and transforming content. Qualitative results and comparisons demonstrates the effectiveness of our proposed method. Project page with code is available at this link.

1. Introduction

Diffusion models (DMs) have made significant strides in controllable multi-modal generation tasks, particularly in the domain of text-to-image (T2I) synthesis. Evolving from basic models, recent Latent Diffusion Model (LDM) showed notable efficacy in T2I task [21, 23, 24]. Building on the progress of T2I models, various approaches have been explored to adapt these models for text-conditioned image editing tasks [8, 19, 27]. In the realm of text-guided image editing, initial work primarily focused on incorporating source image conditions into the sampling or reverse process, such as guiding the generation process through gradient-based sampling [13, 28], or directly training models that takes the source image as conditional input [2]. The progress of T2I models has led to significant advancements in the field of image editing, with researchers directly leveraging the properties of T2I models [8, 27].

In this context, one promising recent approach in this realm is Delta Denoising Score (DDS) [7] - an image editing technique builds upon the Score Distillation Sampling (SDS) [20] framework. SDS allows the optimization of a parametric image generator such as 3D NeRF [15], capitalizing on the rich generative prior of the diffusion model from which it samples [23]. To adapt the SDS framework for editing real images, DDS introduced an additional reference branch with a matching text prompt to refine the noisy gradient of SDS. Then, DDS queries the generative model with two pairs of images and texts. By utilizing the difference between the outcomes of the two queries, which provides a cleaner gradient direction, the target image is updated incrementally. Unfortunately, in DDS the structural details of the source image are often neglected, as shown in Fig. 3.

As preserving structural consistency has been deemed a crucial element in image manipulation, mechanisms to enforce structural consistency in the editing process have continued to evolve. In classical CycleGAN [30], target appearance is enforced using an adversarial loss, while content is preserved using cycle-consistency. However, cycle-consistency assumes the bijection between two domain and should train two generators, which is often too restrictive. On the other hand, the Contrastive Unpaired Translation (CUT) [18] proposed an alternative, rather straightforward way of maintaining correspondence in content by maximizing the mutual information between corresponding input and output patches in the latent domains. Similar idea has been employed within diffusion models, either by using features from ViT [13] or attention layer of the score network [28]. Unfortunately, the existing works entails additional encoder training, which is inefficient. Furthermore, these approaches have been used only for pixel-domain diffusion models, and we are not aware of any prior work using pretrained latent diffusion models.

To address these challenges, here we propose an embarrassingly simple yet amazingly effective zero-shot training-free method for applying CUT loss to DDS, which can be used for pretrained latent diffusion models such as Stable Diffusion [22]. Specifically, we demonstrate that the intermediate features of LDM, particularly those from self-attention, contain rich spatial information, allowing them to be directly utilized for applying CUT loss. We are aware that many recent LDM-based image manipulation methods have also focused on manipulating the attention layers [8, 27]. However, the spatial information in the attention layers are usually exploited either in somewhat supervised manner by identifying the semantically important attention map [19]. On the other hand, our way of utilizing the attention is in fully unsupervised manner by randomly selecting set of patches from the same spatial locations. Therefore, we believe that this is the first work that applies the CUT loss to the LDM model for unsupervised image translation. We validate the effectiveness of our proposed loss on various text-driven image editing tasks. Furthermore, as our work is rooted in the score distillation method, it can be applied to various domains, such as Neural Radiance Fields (NeRF)[15], as demonstrated in our experiments. To summarize, we make the following key contributions:

- For the purpose of suitable structural consistency, we incorporate the CUT loss within the DDS framework and present a method for applying CUT loss using latent representations. To the best of our knowledge, this is the first work that applies the CUT loss to the LDM model and demonstrates its power in zero-shot image editing.
- We show that the intermediate latent features from the self-attention layers can be employed for applying CUT loss without the need for an additional network.
- We show that our method outperforms existing state-of-the-art baselines, achieving a significantly better balance between preserving the structural details of the original image and transforming the content in alignment with a target text prompt.
- This method, being grounded in the score distillation framework, is extendable to multiple domains including NeRF.

2. Related works

Image editing with Text-to-image Diffusion Models Recently, various Text-to-image Diffusion models [21, 23, 24, 29] have been developed, marking substantial progress in the field of image generation. The advancements have reached a stage where the generated images are closely resemble real-world visuals, often indistinguishable to the human eyes. The popularity of open-source generative models, particularly models like LDM [23], has led to extensive exploration of various applications. The incorporation of a text-conditioned injection framework through the cross-

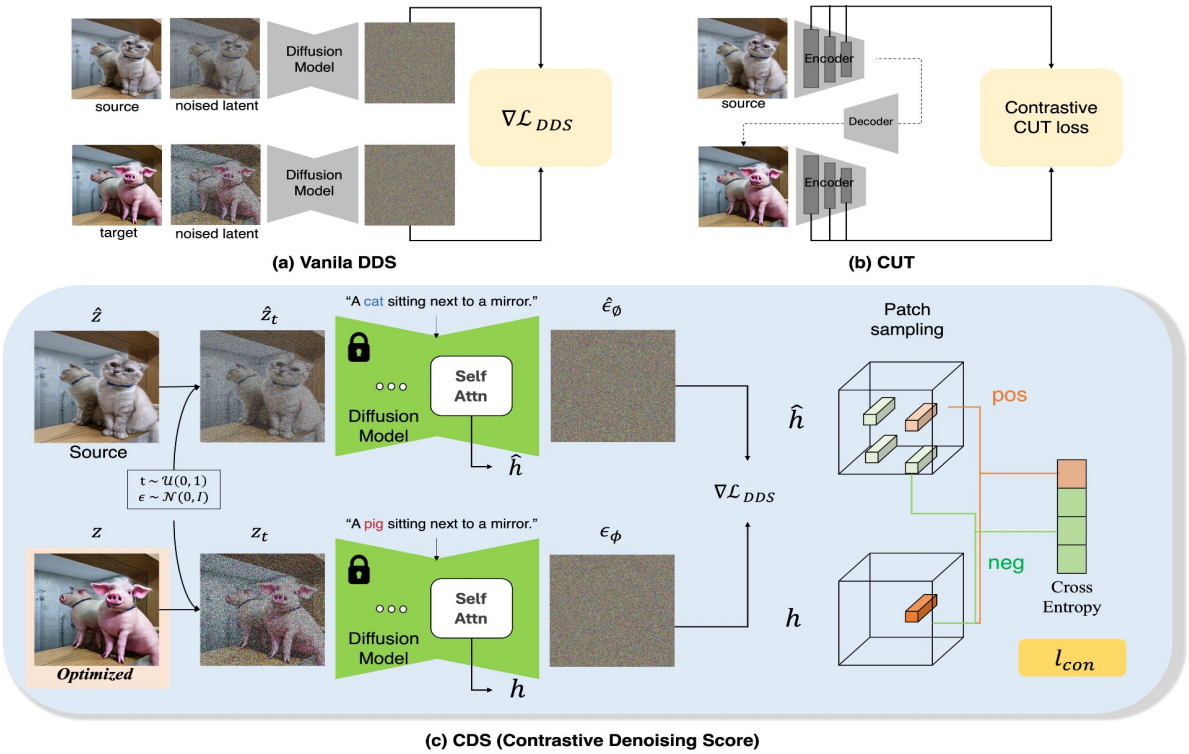

Figure 2. Overall pipeline of CDS. During inference for DDS, we extract the intermediate features of the self-attention layers and calculate $\mathcal{L}con$. This loss enable us to regulate structural consistency and generate a reliable image.

attention layer of the model has enabled a diverse image editing [8] and translation [14, 27] tasks. Recent methods have introduced innovative approaches such as re-weighting for editing specific components [8], the injection of self-attention features for image translation [27], and combinations with methods for inverting real images [6, 16]. These approaches demonstrate improved editing performance.

While most of these methods are applied during the reverse process in pre-trained models, an extended method of Score distillation Sampling (SDS) [20] method has shown promising performance not only in 2D but also in 3D object generation tasks. However, the sampling method utilized by SDS often leads to blurry outputs due to its reliance on the gradient of the difference between pure noise and the target text score prediction. To address this challenge, Delta Denoising Score (DDS) [7] introduces an alternative editing approach using the gradient between source text score and target text score direction. Despite this improvement, the existing framework overlooks a critical aspect of editing: maintaining structural consistency between the source and output images, thereby limiting its overall editing performance. To address this, we propose a new framework that enhances the performance of DDS by introducing appropriate contrastive loss to maintain the structural similarity.

Consistency Regularization for Image Manipulation In image editing and translation, preserving the structural components of the source image while transforming its semantics is crucial. The initial work of CycleGAN [30] introduced cycle consistency, translating output images back to the source domain. Building on this, subsequent studies [1, 5, 10] proposed various consistency regularization techniques to enhance Image-to-Image (I2I) performance. On the other hand, Contrastive Unpaired Translation (CUT) [18] introduced a method of applying contrastive learning to patch-wise representations, effectively preserving structural information between source and output. Inspired by this work, the applications of the CUT Loss to StyleGAN [12] and Diffusion models [13, 28] in subsequent studies for image translation and style transfer yielded promising results. While various techniques have been proposed, applying these methods to the pretrained text-to-image latent diffusion models like StableDiffusion still remains an open problem, as finetuing the off-the-self LDM is compute heavy. Leveraging the simultaneous network prediction for source and output in the DDS framework, we discovered a natural integration of CUT loss to DDS using an off-the-self LDM without finetuing. This zero-shot integration significantly improves image editing output compared to traditional DDS.

3. Main Contribution: The CDS

3.1. Key Observation

DDS. We begin with a concise overview of DDS and then compare its similarity and difference from CUT. This comparative explanation clearly illustrates the missing component in DDS and inspire us how to improve DDS.

The noise that text conditioned diffusion models using classifier-free guidance (CFG) [9] predicted can be expressed as:

$$\boldsymbol{\epsilon}_\phi^\omega(\boldsymbol{z}_t, y, t) = (1+\omega)\boldsymbol{\epsilon}_\phi(\boldsymbol{z}_t, y, t) - \omega\boldsymbol{\epsilon}_\phi(\boldsymbol{z}_t, \emptyset, t) \quad (1)$$

where ω denotes the guidance parameter, and $\boldsymbol{\epsilon}_\phi$ denote a noise prediction network with parameters set ϕ. Additionally, y and $\emptyset$ represent the text and null-text prompt, respectively, and $\boldsymbol{z}_t$ represent a noisy latent from the clean latent $\boldsymbol{z}_0$ at the noise timestep $t{\sim}\mathcal{U}(0,1)$.

SDS leverages the gradient of the diffusion loss function:

$$\mathcal{L}_{\text{SDS}}(\theta; y_{tgt}) = \|\boldsymbol{\epsilon}_\phi^\omega(\boldsymbol{z}_t(\theta), y, t) - \boldsymbol{\epsilon}\|^2 \quad (2)$$

where y refers to the target prompt, $\boldsymbol{\epsilon}{\sim}\mathcal{N}(0, \mathbf{I})$ and

$$\boldsymbol{z}_t(\theta) = a_t\boldsymbol{z}_0(\theta) + b_t \quad (3)$$

for some DDPM noise schedule (a_t, b_t) with $a_t^2 + b_t^2 = 1$ and with $\boldsymbol{z}_0(\theta)$ denotes the target latent parameterized by θ that should be optimized to follow the target prompt. It has been demonstrated that $\nabla_\theta\mathcal{L}_{\text{SDS}}$ is efficient gradient term for generating images that exhibits a heightened level of fidelity to the given prompt. However, SDS produces blurry outputs that primarily emphasize objects mentioned in the prompt, rendering it insufficient for practical image editing purposes.

DDS expands SDS framework for the domain of image editing, by utilizing not only target text prompt y but also a reference pair of image $\hat{z}_0$ and text $\hat{y}$. Specifically, the DDS loss is given by

$$\mathcal{L}_{\text{DDS}}(\theta; y_{tgt}) = \|\boldsymbol{\epsilon}_\phi^\omega(\boldsymbol{z}_t(\theta), y, t) - \boldsymbol{\epsilon}_\phi^\omega(\hat{\boldsymbol{z}}_t, \hat{y}, t)\|^2 \quad (4)$$

where $\hat{\boldsymbol{z}}_t = a_t\hat{\boldsymbol{z}}_0 + b_t$. In a same manner as SDS, $\boldsymbol{z}_0(\theta)$ is updated incrementally in the direction of the $\nabla_\theta\mathcal{L}_{\text{DDS}}$. As shown in Fig. 2a, the score supplied by the reference branch aligns with the score from the output branch by minimizing (4). Given that the score $\boldsymbol{\epsilon}_\phi^\omega$ can be understood as feature vector, DDS attempts to minimize the differences in the feature domain.

CUT. The basic idea of CUT is to exploit patch-wise contrastive learning in the feature domain for one-sided translation. Specifically, a generated output should produce feature whose patch appear closer to its corresponding patches from the input image, in comparison to other random patches. As shown in Fig. 2b, CUT use a multilayer, patchwise contrastive loss, which maximizes mutual information between corresponding input and output patches. This enables one-sided translation in the unpaired setting without imposing the cycle consistency. CUT has been demonstrated that the CUT loss is effective in maintaining correspondence in content and structure, so it has been widely utilized in image editing.

3.2. Contrastive Denoising Score

By inspection of Fig. 2a and 2b, we can see the striking similarity between the two. Aside from the actual image generation processes (i.e. by trained decoder network in CUT, and image optimization in DDS), both approaches attempt to align the features from the reference and reconstruction branches.

While DDS offers a denoised editing direction that focuses on editing the pertinent part of the image, such that it matches the target text, empirical observations reveal instances of failure cases. For example, as shown in Fig. 3, the pose or structural details of the content in the source image are not preserved. Recall that the main objective of text-driven image editing is not only aligning to the content specified in a target prompt, but also incorporating the structure and details of an input source image.

With the aim of regulating the excessive structural changes, we are therefore interested in the key idea from CUT, which is shown effective in maintaining input structure. However, the original CUT algorithm requires training an encoder to extract spatial information from the input image, which is inefficient. We therefore aim to calculate CUT loss without introducing auxiliary encoder, by fully leveraging the information of the latent representation of LDM.

One potential approach is to compute CUT loss by directly leveraging the score $\boldsymbol{\epsilon}_\phi^\omega(\boldsymbol{z}_t(\theta), y, t)$ and $\boldsymbol{\epsilon}_\phi^\omega(\hat{\boldsymbol{z}}_t, \hat{y}, t)$. It shows an effect; however, in certain cases, we have observed that semantic changes to align with the content specified in a target text were also suppressed due to information entanglement. On the other hand, intermediate representations from self-attention layers have been shown to contain rich spatial information, disentangled from semantic information [27]. Even considering how it operates, self-attention layers contain similarity information between spatial patches in the given representations, which is exactly CUT loss requires. Therefore, we calculate CUT loss utilizing the latent representation of self-attention layers. For visualization results of CUT loss in other feature extraction spaces, please refer to Fig. 6.

Specifically, we begin with briefly describing the self-attention layers that compose the denoising U-Net $\boldsymbol{\epsilon}_\theta$ in the Stable Diffusion (SD) model. During each timestep t of the denoising process, the noisy latent representation $\boldsymbol{z}_t$ is fed as input to the denoising network. For self-attention layer l, $\hat{\boldsymbol{h}}_l$ and $\boldsymbol{h}_l$ represent the intermediate features passed through the residual block and self-attention block condi-

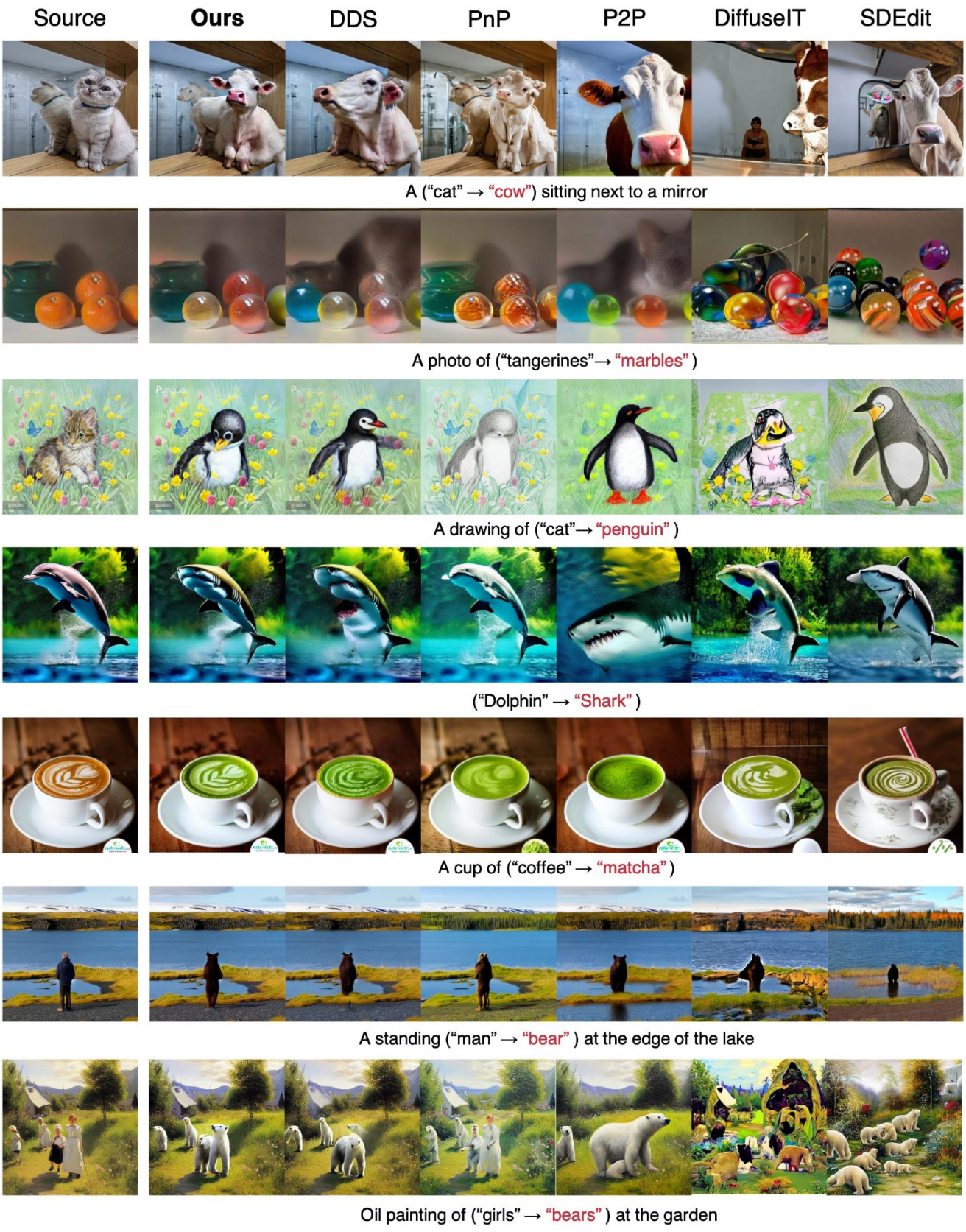

Figure 3. Comparison with baseline models. Our proposed method demonstrates outstanding performance in effectively regulating structural consistency.

Table 1. Quantitative evaluation to the state-of-the-art editing methods. Baseline results are from [19].

Method	Metric	
	CLIP Acc (↑)	Dist (↓)
SDEdit + word swap	71.2%	0.081
DDIM + word swap	72.0%	0.087
P2P	66.0%	0.080
Pix2Pix-zero	92.4%	0.044
DDS	**97.9**%	0.0226
Ours	97.5%	**0.0203**

tioned on $\hat{y}$ and y, respectively.

During the denoising process of each branch, which is the part of the DDS gradient computation, we obtain $\hat{\boldsymbol{h}}_l$ and $\boldsymbol{h}_l$. Then, we randomly selecting patches from the feature map h_l. Initially, a "query" patch is sampled from the feature map $\boldsymbol{h}_l$. We denote $s \in \{1, ..., S_l\}$, where S_l is the number of query patches. Then, for each query, the patch at the corresponding location on the feature map $\hat{\boldsymbol{h}}_l$ is designated as the "positive", while the non-corresponding patches within the feature map serve as "negatives". We refer to the positive patch as $\hat{\boldsymbol{h}}_l^s$ and the other patches as $\hat{\boldsymbol{h}}_l^{S\setminus s}$. The objective of the CUT loss is to maximize the mutual information between "positive" while simultaneously minimize the mutual information between "negatives." This process can be formulated as the patchNCE loss:

$$\ell_{con}(\boldsymbol{z}, \hat{\boldsymbol{z}}) = \mathbb{E}_{\boldsymbol{h}} \left[\sum_l \sum_s \ell(\boldsymbol{h}_l^s, \hat{\boldsymbol{h}}_l^s, \hat{\boldsymbol{h}}_l^{S\setminus s}) \right] \tag{5}$$

where, $\ell(\cdot)$ denotes cross-entropy loss:

$$\ell(\boldsymbol{h}, \boldsymbol{h}^+, \boldsymbol{h}^-) = -\log \left(\frac{\exp(\boldsymbol{h} \cdot \boldsymbol{h}^+/\tau)}{\exp(\boldsymbol{h} \cdot \boldsymbol{h}^+/\tau) + \sum \exp(\boldsymbol{h} \cdot \boldsymbol{h}^-/\tau)} \right) \tag{6}$$

for some parameter $\tau > 0$. By using this simple ℓ_{con} loss additionally, we can regularize DDS to maintain structural consistency between the $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$.

Table 2. User study results. Our CDS shows the best results.

Method	Metric		
	Text (↑)	Structure (↑)	Quality (↑)
SDEdit	3.77	2.90	3.43
DiffuseIT	3.17	2.94	2.83
P2P	3.89	2.69	3.61
PnP	3.36	3.70	3.22
DDS	4.06	4.05	3.64
Ours	**4.43**	**4.65**	**4.20**

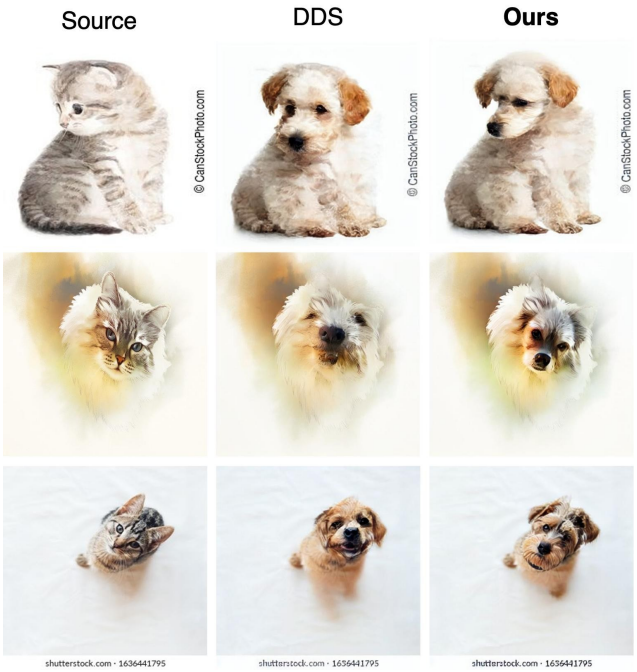

Figure 4. Sample results of DDS and CDS on image from cat2dog dataset retrieved from the LAION 5B dataset.

4. Experiments

4.1. Experimental setting

Implementation. For implementation, we referenced the official source code of Delta Denoising score[1]. We modified the source code to extract intermediate features from attention layers and apply PatchNCE loss similar to CUT. Inspired by the analysis of CUT, we applied PatchNCE to all of the up-sampling self attention layers. Different from the default CUT, we did not use additional projection layer. We did not included PatchNCE loss to U-Net bottleneck layer, as bottleneck is related to overall semantics of the images. Additional experimental details are provided in our Supplementary Materials.

Baseline methods. To comprehensively evaluate the performance of our method, we conduct comparative experiments, comparing it to several state-of-the-art methods. Our method is compared against five baselines including vanilla DDS. For implementation, we referred the official source code of each methods, except for SDEdit which we used the implementation of Stable Diffusion. Since P2P implementation requires additional inversion process, we used DDPM inversion [11] in this step.

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

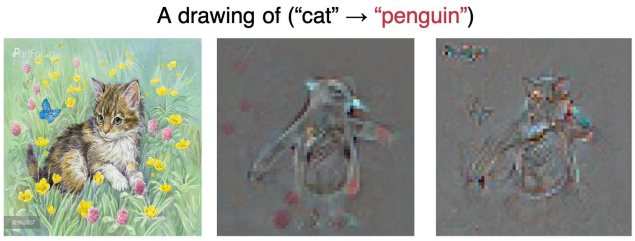

Figure 5. Gradient visualization on DDS and our proposed method.

4.2. Experimental Results

Qualitative results To compare the qualitative results, we show the edited outputs in Fig. 3. For the sampling-based methods of DiffuseIT and SDEdit, the results changes the source image attribute to target text conditions, but the outputs structures are severely deformed from the original source image structures, and some cases have unwanted artifacts. For the attention-based method of P2P, the results also follow the target text conditions. However, the method suffer from severe structure inconsistency because the quality is largely effected by the performance of inversion method and timesteps for attention map modulation. For the another baseline of PnP diffusion, the overall structure is well maintained from source images, and the output semantic reflects the text conditions. However, the method still does not fully keep the consistency between source and output, and for difficult cases such as cat→cow, the output shows unrealistic result with artifacts. DDS shows decent performance in text-guided editing, but still the method fails in maintaining the structural consistency between the source images. On the other hand, our method CDS can successfully edit the source images with their original structural information and our results do not modify the unrelated regions (e.g. background) in which most of the baseline methods shows degraded performance.

Quantitative results In order to further measure the generation quality of our proposed method, we conducted quantitative experiments. We focus on image-to-image translation tasks, such as cat → dog, which can evaluate both structural consistency and semantic changes. Source images are retrieved from the LAION 5B dataset [25]. To evaluate whether our results reflect the editing direction (source → target), we measure CLIP Accuracy and DINO-ViT structure distance.

Table 1 shows that our proposed method gets a high CLIP-Acc while having low structure distance, which indicates that our output can achieve optimal editing results while preserving the structural elements of the original input image. Visualizations are provided in the Fig. 4.

Moreover, in order to evaluate the perceptual quality of translated image, we conducted an additional user study. We gathered feedback from 20 subjects and asked the users in three different parts: 1) Text-match, 2) Structure consistency, 3) Overall quality of generated image. In Table 2, our model showed the best performance.

4.3. Ablation studies

In order to evaluate the proposed loss, we conducted various ablation study.

CUT loss. First, we ablate our loss ℓ_{con} and demonstrate their effectiveness in Fig. 4. We observed that excluding ℓ_{con}, which is vanilla DDS, resulted in a loss of structural details even though the overall contents are changed. When we apply ℓ_{con}, structural attributes such as leg and pose are preserved. This indicates that ℓ_{con} is beneficial for preserving the overall structure of source image. Overall, when we apply ℓ_{con} utilizing features from the self-attention layers, we can reliably edit images with both reflecting the source image structure and the target text semantics.

Furthermore, to investigate the effect of the proposed loss on DDS gradients, we also present a study on the visualization of gradients for both vanilla DDS and improved DDS with our proposed loss. In Fig. 5, we show the visualization results of gradients. For the gradients of vanilla DDS, the spatial information does not accurately follow the original structure of the source images. However, in our proposed DDS, the gradient has much more detailed structural information such as cat's ears and pose. This shows that our proposed CUT loss framework enhances the spatial details of DDS gradient, leading to further improvement in the final edited output.

CUT loss location. First, as shown in Fig. 6, we evaluated the effectiveness of feature extraction layer for CUT loss calculation. We show the generated outputs with applying the additional CUT loss on the the direct score network output, on the hidden state of cross attention layer, and on the hidden state of self-attention layers, respectively. The results in Fig. 6 illustrate that the direct application of the loss to the score output excessively constrains information. When apply loss on cross attention, we can see that the outputs does not correctly preserve the original structures. Our results show that proposed self-attention layer has best performance in editing aspect. This demonstrates that the features extracted from self-attention layers possess disentangled spatial information, which aligns well with the requirements of CUT loss.

CUT loss weight. For further analysis on our proposed consistency loss on DDS framework, we conducted addi-

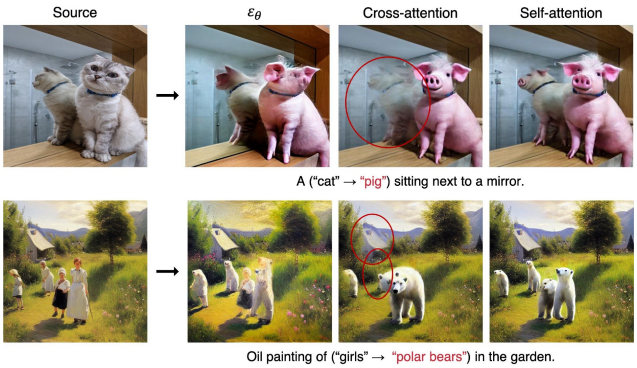

Figure 6. Qualitative results for ablation study on feature extraction location for contrastive loss.

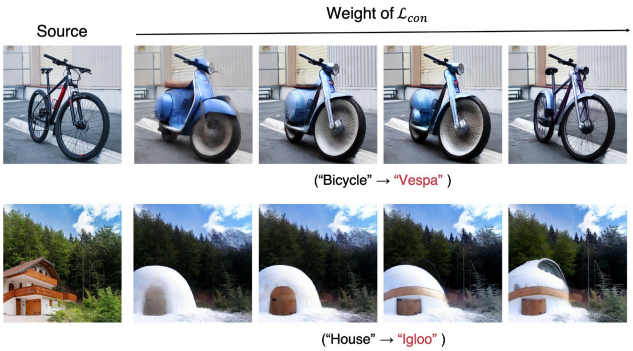

Figure 7. Ablation study on weights of contrastive loss.

tional experiments with changing the loss weights. Intuitively, we expect to have better structural consistency with higher contrastive loss weights, and better semantic change with lower weights. In Fig. 7, we can observe that we can control the editibility and consistency with varying the weight of contrastive loss. Not only for the target object structure, we can see that using stronger loss can affect the preservation of background area.

4.4. Extension to NeRF

Since our propose method is an improved version of score distillation framework, we can transfer the distilled score gradient to other generator network, such as NeRF. Inspired by the recent ED-NeRF [17], which aimed to leverage DDS for text-guided 3D object editing, we applied our proposed method to pre-trained NeRF. Beginning with original NeRF model, we rendered the 2D images for reference and target NeRF models with same view directions. Subsequently, we applied our proposed framework to NeRF fine-tuning task. For comparison, we also conducted experiment with our baseline of DDS. Additional experimental details are provided in our Supplementary Materials.

In Fig. 8, we show the comparison between applying vanilla DDS and our proposed framework to pre-trained

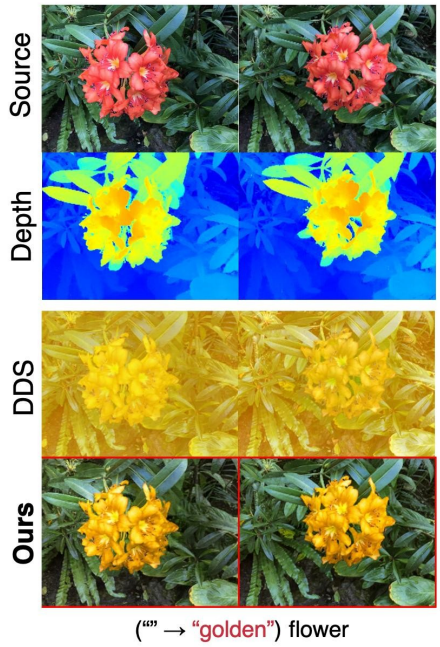

Figure 8. Results on NeRF editing. As an extension of our proposed framework, we applied our method to the NeRF 3D object editing task. Compared to baseline DDS method, our approach can captures the structure of the original source object.

NeRF. In the case of the vanilla DDS, the outputs failed to capture the structural information of source images, resulting in the model changing the entire pixel colors. However, with our proposed method, we were able to accurately edit the target object (e.g. flower) to match the desired text conditions. The results demonstrate the effectiveness of our proposed scheme in the 3D NeRF space, highlighting that our method is extendable to multiple domains based on the score distillation framework.

5. Conclusions

In this paper, we introduced the CUT loss within DDS framework to preserve structural consistency. Unlike the original CUT algorithm, which required additional network inference, we leverage the rich spatial information inherent in latent representation extracted from LDM, especially the self-attention layer. These loss allows us to successfully generate image with a better balance between preserving the structural details of the original image and transforming the content in alignment with a target text prompt. Qualitative and quantitative experiments demonstrate that the effectiveness of our proposed method and its scalability to multiple domains.

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 11

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 12

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[6] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, Di Liu, Qilong Zhangli, Jindong Jiang, Zhaoyang Xia, Akash Srivastava, and Dimitris Metaxas. Improving tuning-free real image editing with proximal guidance, 2023. 3

[7] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score, 2023. 2, 3

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[10] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[11] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. *arXiv preprint arXiv:2304.06140*, 2023. 6

[12] Gihyun Kwon and Jong Chul Ye. One-shot adaptation of gan in just one clip. *arXiv preprint arXiv:2203.09301*, 2022. 3

[13] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[14] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3

[15] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[16] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. 3

[17] Jangho Park, Gihyun Kwon, and Jong Chul Ye. Ed-nerf: Efficient text-guided editing of 3d scene using latent space nerf, 2023. 8

[18] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision – ECCV 2020*, pages 319–345, Cham, 2020. Springer International Publishing. 2, 3

[19] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023. 2, 6, 11

[20] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 2, 3

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[23] 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. 2

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 7, 11

[26] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 11

[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, 3, 4, 11

[28] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer, 2023. 2, 3

[29] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian

Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 2

[30] 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 (ICCV)*, 2017. 2, 3

Contrastive Denoising Score for Text-guided Latent Diffusion Image Editing

Supplementary Material

A. Implementation details

For implementation, we referenced the official source code of Delta Denoising Score[2] by using Stable Diffusion v1.4. We modified the source code to extract intermediate features from self-attention layers and apply the CUT loss. For hyperparameters, we use the patch sizes of 1×1 and 2×2 with 256 patches and 3.0 for ℓ_{con} weight. 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. 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

This section provides the experimental setup and presents additional results for quantitative comparison. To guarantee fair evaluations, all methods utilize the pre-trained Stable-Diffusion v1.4, and adopt the same classifier-free guidance scale.

B.1. Dataset

For the animal transition task, we conduct three tasks: (1) cat $\rightarrow$ dog, (2) cat $\rightarrow$ cow and (3) cat $\rightarrow$ pig. Adhering to the data collection protocol outlined in [19], we gather 250 images relevant to cats from the LAION 5B dataset [25]. Images are selected based on high CLIP similarity to the source word.

B.2. Metrics

Motivated by [19, 27], we measure the CLIP accuracy (CLIP Acc) and the DINO-ViT structure distance. CLIP Acc represents the degree to which the targeted semantic contents are accurately reflected in the generated images. On the other hand, the structure distance [26, 27] measures the extent to which the overall structure of the input image is well-preserved. It is defined as the difference in self-similarity among the keys obtained from the attention module at the deepest layer of DINO-ViT [3].

B.3. Additional Results

In addition to quantitative results in main paper, Table 3 and 4 also demonstrate the effectiveness of our proposed method in achieving optimal editing outcomes while maintaining the structural element of the source image. In contrast, other baseline models achieve low structure distance.

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

Table 3. Quantitative evaluation for the cat $\rightarrow$ cow task.

Method	Metric	
	CLIP Acc (↑)	Dist (↓)
SDEdit + word swap	99.6%	0.070
DDIM + word swap	**100**%	0.136
P2P	86.1%	0.078
Pix2Pix-zero	86.6%	0.908
DDS	99.6%	0.040
Ours	97.9%	**0.033**

Table 4. Quantitative evaluation for the cat $\rightarrow$ pig task.

Method	Metric	
	CLIP Acc (↑)	Dist (↓)
SDEdit + word swap	99.2%	0.066
DDIM + word swap	**100**%	0.116
P2P	85.7%	0.073
Pix2Pix-zero	55.0%	0.106
DDS	**100**%	0.031
Ours	**100**%	**0.027**

This implies that despite achieving high CLIP accuracy, the existing methods edited the image without considering the source structure. The visualizations comparison for two tasks (Fig. 4) also confirm the aforementioned quantitative evaluation.

B.4. User study Details

For the user study in Table 2 of the main text, we presented six comparison results and collected feedback from 20 participants, ranging in age from their 20s to 50s. After each participant viewed images generated by our model and the baselines, they provided feedback through a scoring survey. We set the minimum score as 1 and the maximum score as 5, and user choose the score among 5 options. To measure the performance of editing, we asked three questions for each sample: 1) (Text-match) Does the image reflect the target text condition?, 2) (Structural consistency) Does the image contain the content and structure information of source images?, 3) (Overall quality) Are the generate images realistic?

C. Additional Ablation studies

C.1. Patch size

First, we evaluate the impact of patch size by varying its size. In Fig. 9a, we can observe that the patch size has an

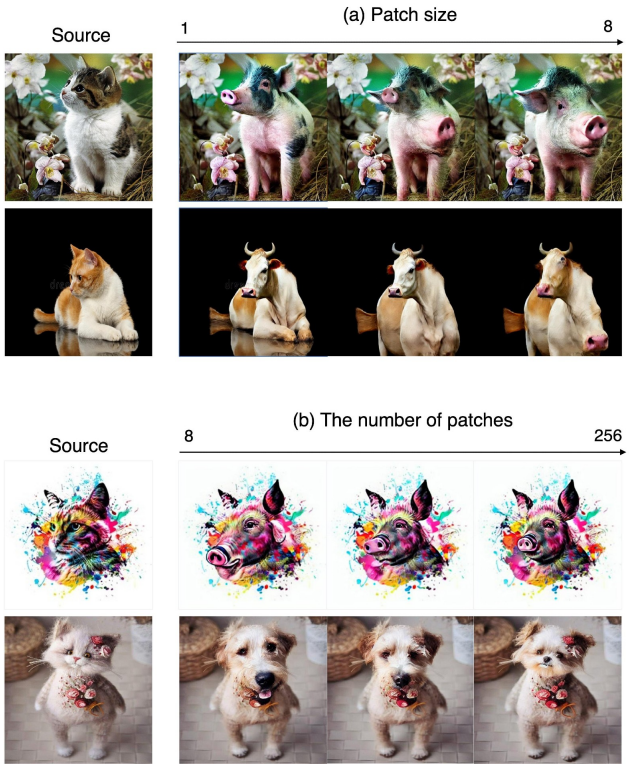

Figure 9. Ablation study on (a) patch size and (b) the number of patches. The given prompt is "cat → pig," "cat → cow,", "cat → pig" and "cat → dog" from top to bottom.

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.

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. 9b). Therefore, we chose 256 number of patches.

D. Additional results

D.1. Qualitative results

In Figs. 12 and 13, 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.

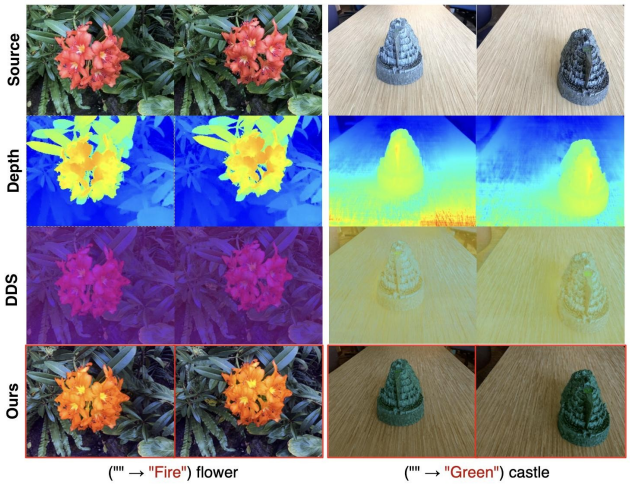

Figure 10. Additional results on NeRF editing.

D.2. NeRF editing

We also provide additional results on NeRF editing. For NeRF model fine-tuning, we prepare the pre-trained NeRF. In this part, we leveraged recent model of TensoRF[4]. For efficiency, we used downsampled images which has resolution of 504x378 in pre-training stage. For fine-tuning the pre-trained model, we downsampled the resolution to 252x189, due to the limited resources.

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

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

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

In Fig. 10, we show the comparison results between the baseline DDS and our proposed DDS with contrastive loss. We found that basic DDS model shows difficulty in capturing the shape of the original 3D object, with changing the entire color tones. For our method, we can edit the localized object without damaging the shape of original object.

E. Limitations and Negative social impact

Given that our framework manipulates images based on user intentions, there exists a potential for misuse, including the creation of deepfakes or other forms of disinformation. Moreover, 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.

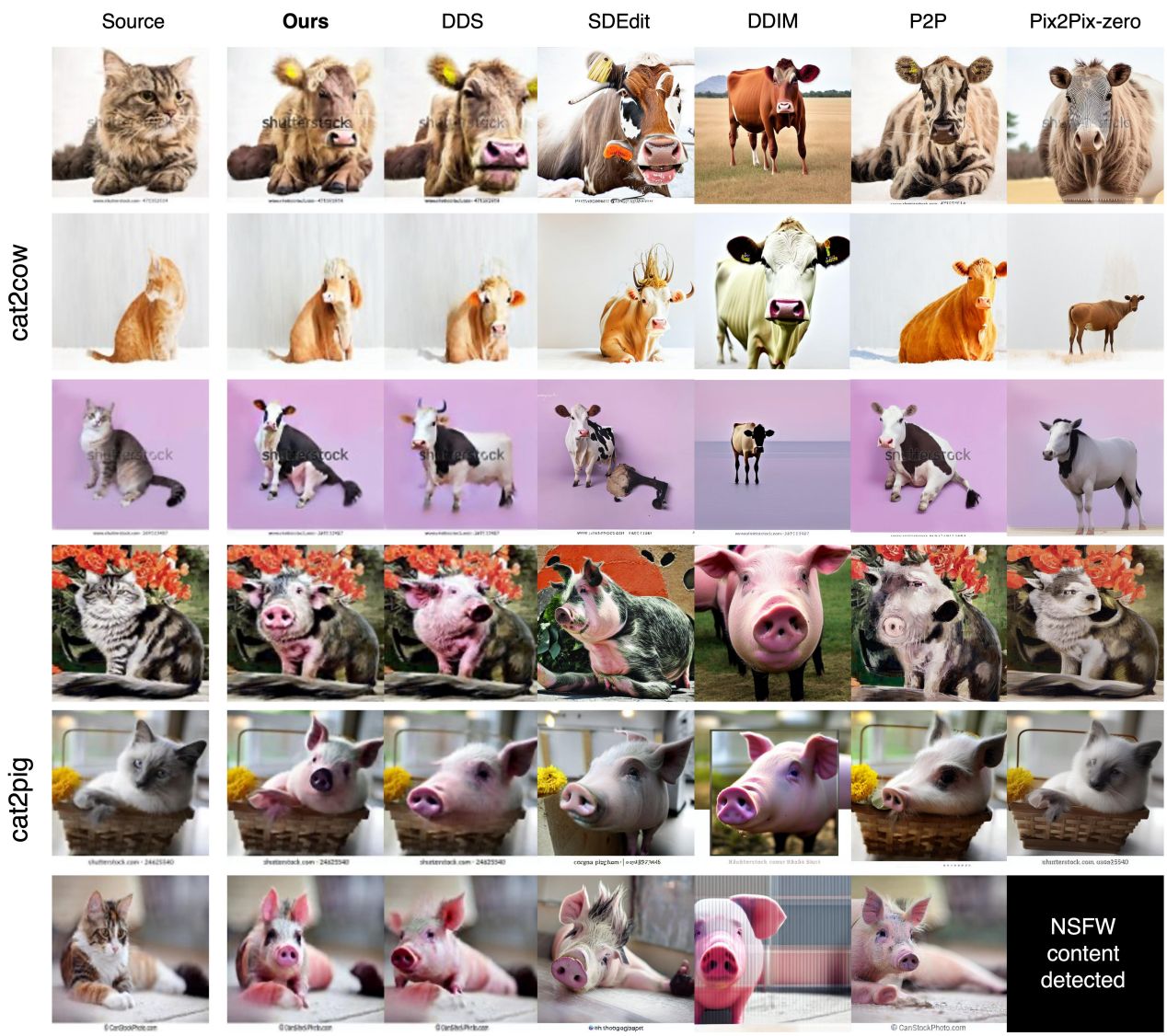

Figure 11. Additional comparison results with baseline models. In the last row of the cat2pig task, Pix2Pix-zero generates an NSFW(Not Safe For Work) image, so we are unable to include it.

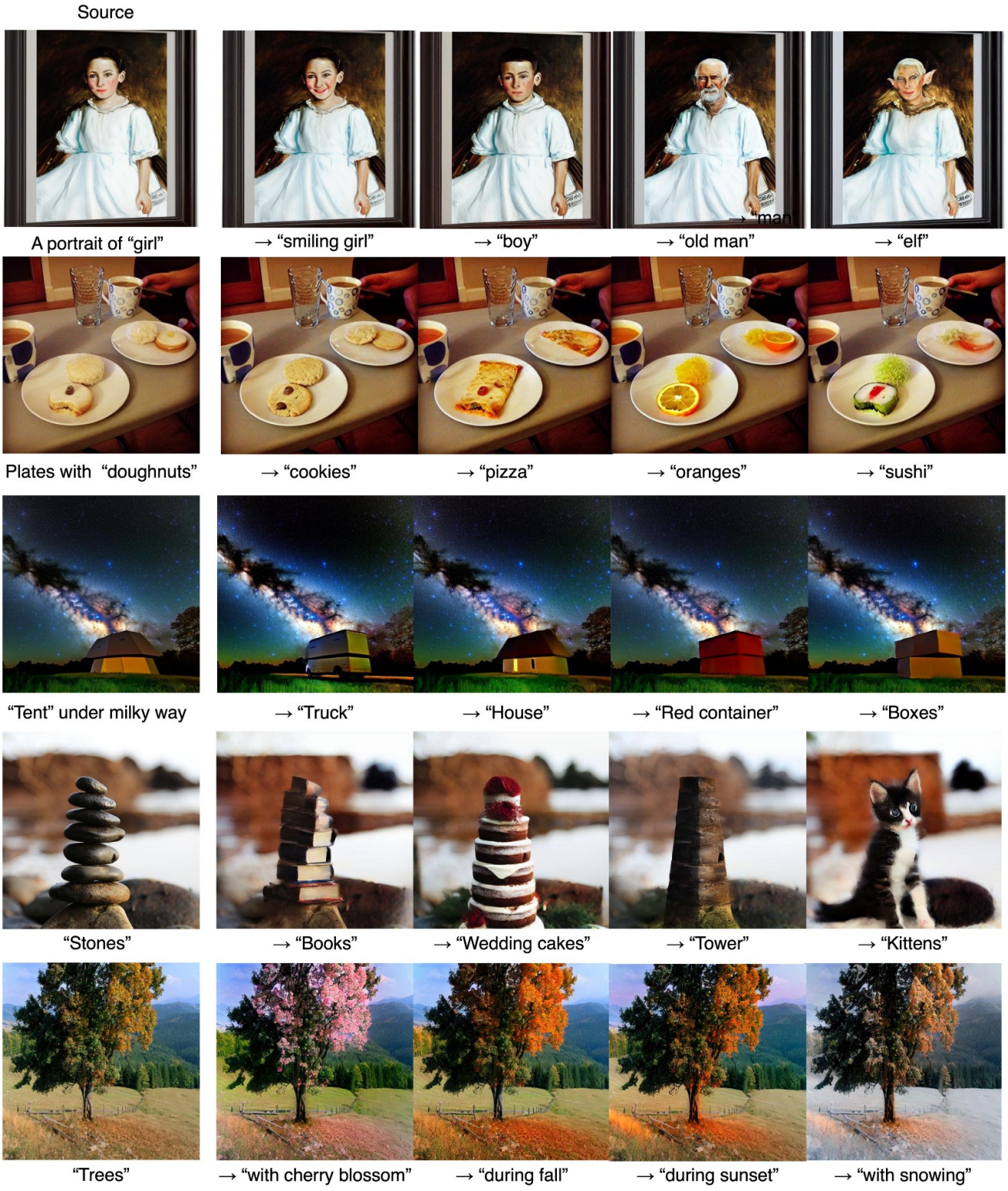

Figure 12. Additional qualitative results with various images and prompts.

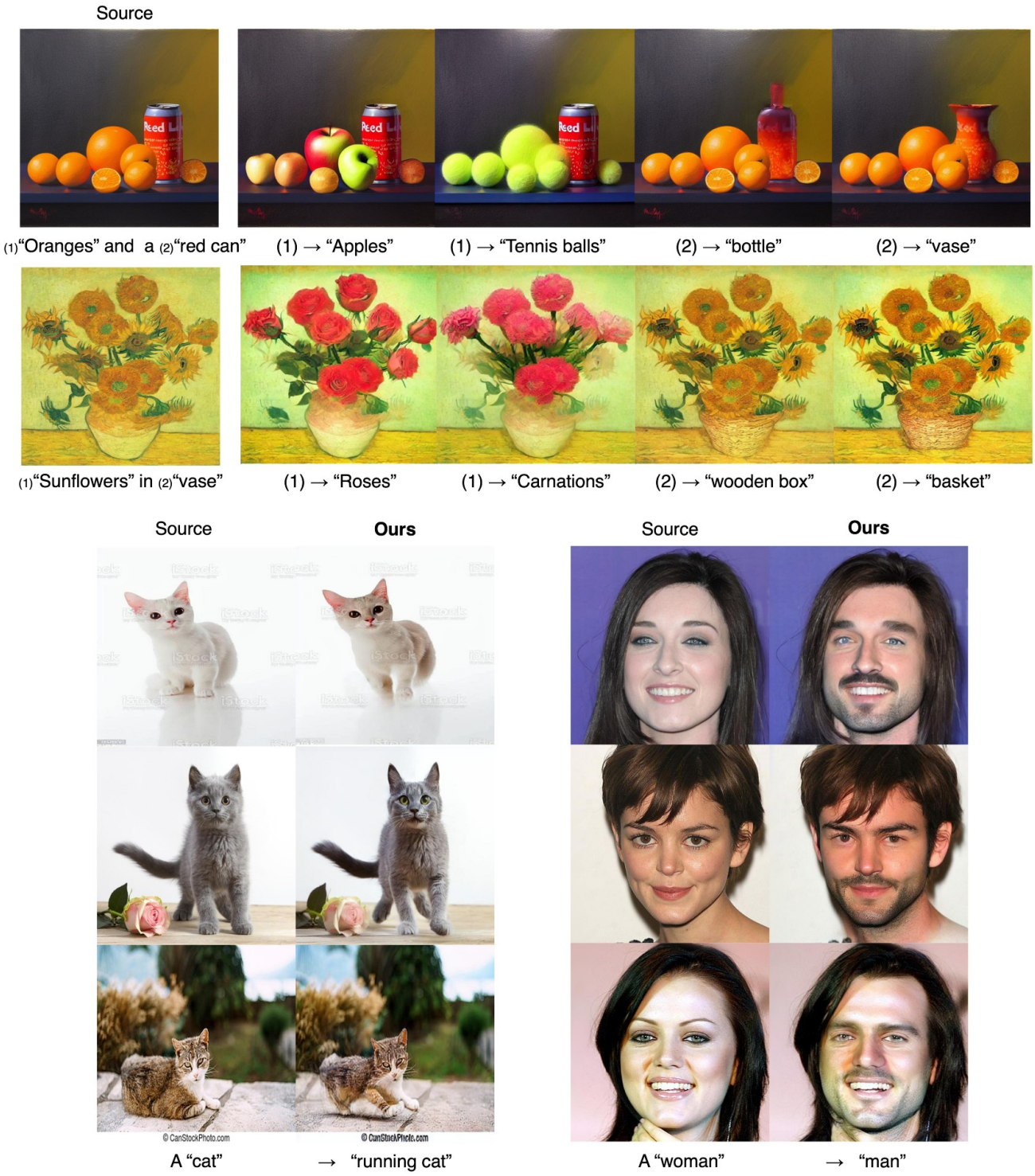

Figure 13. Additional qualitative results with various images and prompts.