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

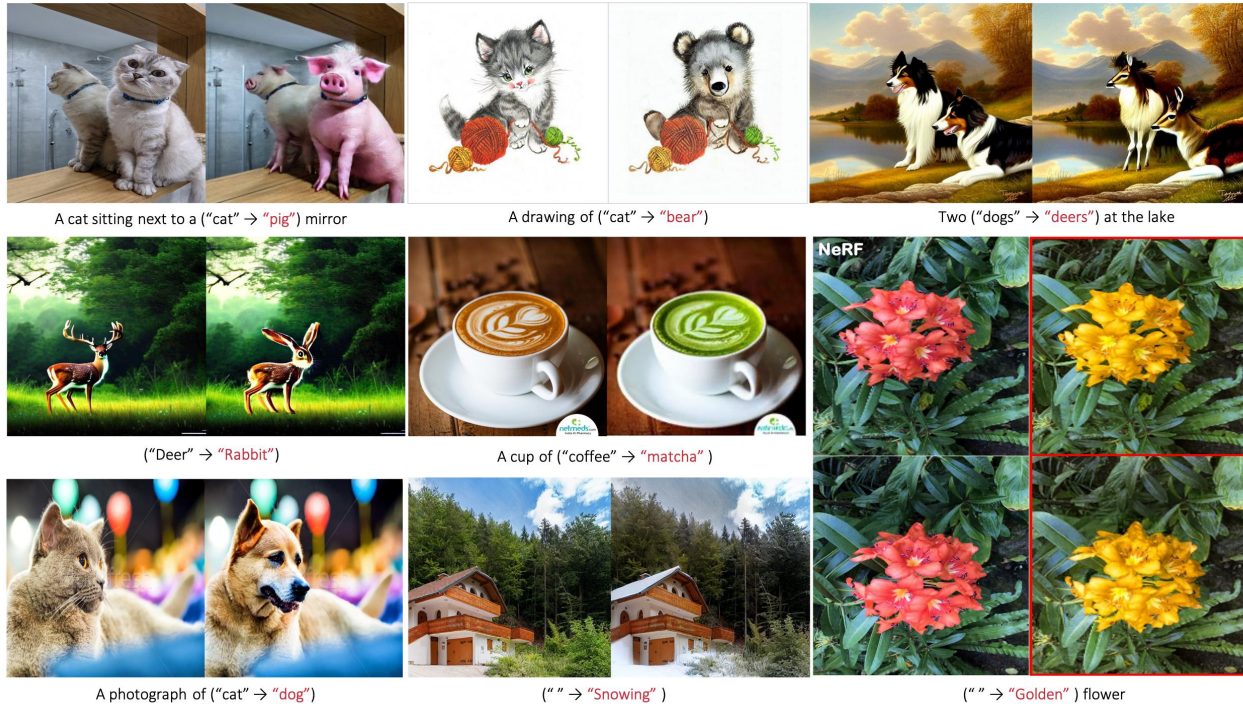{hyelin.nam, cyclomon, pky3436, jong.ye}@kaist.ac.kr

Figure 1. Text-guided Image Editing results via **C**ontrastive **D**enoising **S**core. **CDS** 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. To address this, here we present an embarrassingly simple yet very powerful modification of DDS, called Contrastive Denoising Score (CDS), for latent diffusion models (LDM). Inspired by the similarities and differences between DDS and the contrastive learning for unpaired image-to-image translation(CUT), we introduce a straightforward approach using CUT loss within the DDS framework. Rather than employing auxiliary networks as in the original CUT approach, we leverage the intermediate features of LDM, specifically 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 structural correspondence between the input and output while maintaining content controllability. Qualitative results and comparisons demonstrates the effectiveness of our proposed method. Project page: https://hyelinnam.github.io/CDS/

1. Introduction

Diffusion models(DMs) have made significant strides in controllable multi-modal generation tasks, particularly in text-to-image(T2I) synthesis. Evolving from basic models, recent Latent Diffusion Model(LDM) showed notable efficacy in T2I task [23–25]. Building on the progress of T2I models, various approaches have been explored to adapt these models for text-conditioned image editing tasks. Initial work primarily focused on incorporating source image conditions into the sampling or reverse process. This involved guiding the generation process through gradient-based sampling [13, 28], or directly training models that take the source image as conditional input [2]. With the advent of LDM, researchers have begun directly leveraging the properties of T2I models [8, 27].

In this context, one promising recent approach is Delta Denoising Score(DDS) [7] - an image editing technique builds upon the Score Distillation Sampling(SDS) [21] framework. SDS allows the optimization of a parametric image generator such as 3D NeRF [16], capitalizing on the rich generative prior of the diffusion model from which it samples. However, SDS is recognized for prominent concerns such as over-saturation and over-smoothing. To address these issue, 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 is reconized as crucial in image manipulation, mechanisms to enforce structural consistency in the editing process have continued to evolve. In classical CycleGAN [29], target appearance is enforced using an adversarial loss, while content is preserved using cycle-consistency. However, cycle-consistency assumes the bijection between two domains, necessitating the training of two generator, which can be overly restrictive. Contrastive Unpaired Translation (CUT) [19] proposed an alternative approach by maximizing the mutual information between corresponding input and output patches in the latent domains. Building on the success of these approaches, similar ideas have incorporated into diffusion models, utilzing features from ViT [13] or the attention layer of the score network [28]. However, these approaches have only been applied to pixel domain diffusion models, not latent diffusion models, and also require additional encoder training, which is inefficient.

To address these challenges, here we propose an embarrassingly simple yet amazingly effective zero-shot training-free method called Contrastive Denoising Score(**CDS**). This method integrates the application of CUT loss to pretrained latent diffusion models such as Stable Diffusion [24], within DDS framework. We utilize the rich spatial information in the self-attention features of LDM's to compute the CUT loss.

Importantly, the application of CUT loss has been limited to pixel-domain diffusion models, and we are not aware of any prior work using latent diffusion models. Moreover, while previous image manipulation methods often use attention layers in a supervised manner [8, 20, 27], our approach utilizes attention in a fully unsupervised manner, randomly selecting sets of patches from the same spatial locations. Therefore, we believe that this is the first work to apply CUT loss to LDM 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) [16], as demonstrated in our experiments.

To summarize, we make the following key contributions:

- For the purpose of suitable structural consistency, we integrate CUT loss into the DDS framework using latent latent representations. To the best of our knowledge, this is the first work to apply CUT loss LDM for zero-shot image editing.
- We demonstrate that utilizing intermediate latent features from self-attention layers enables the application of CUT loss without requiring 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, the advancements in Text-to-image Diffusion models have led to generated images that closely resemble real-world visuals, often indistinguishable to the human eye [23–25]. The popularity of open-source generative models, particularly models like LDM [24], has led to extensive exploration in image editing. The incorporation of a text-conditioned injection framework through the cross-attention layer of the model has enabled a diverse image editing [8] and translation [15, 27] tasks. Recent methods also 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, 17]. These approaches demonstrate improved editing performance.
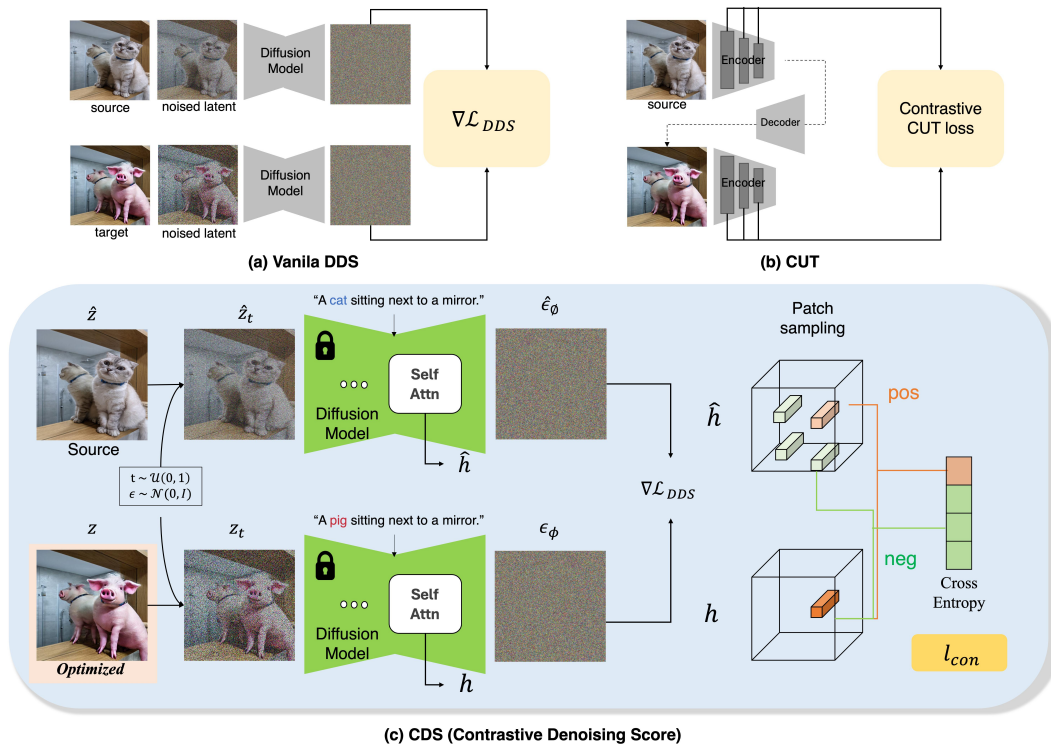
(a) Vanila DDS

(b) CUT

(c) CDS (Contrastive Denoising Score)

Figure 2. Overall pipeline of **CDS**. We extract the intermediate features of the self-attention layers and calculate ℓ_{con}. This loss enables us to regulate structural consistency and generate reliable images.

While most of these methods are typically applied during the reverse process in pre-trained models, an extended method known as Score Distillation Sampling (SDS) [21] method has shown promising performance, both in 3D object generation and 2D image generation tasks. However, 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, Delta Denoising Score (DDS) [7] introduces an alternative editing approach using the gradient between the source text score and the direction of the target text score. Despite this improvement, DDS overlooks the critical aspect of editing: maintaining structural consistency between the source and output images, thus limiting its overall editing performance. To address this, we propose a new framework that enhances the performance of DDS by introducing an appropriate contrastive loss to maintain 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 [29] introduced cycle consistency, translating output images back to the source domain. Building on this, subsequent studies [1, 5, 11] proposed various consistency regularization techniques to enhance Image-to-Image (I2I) performance. On the other hand, Contrastive Unpaired Trans-

lation (CUT) [19] introduced a method of applying contrastive learning to patch-wise representations, effectively preserving structural information between source and output. Inspired by this work, subsequent studies have explored the application of the CUT loss to StyleGAN [14] and Diffusion models [13, 28], yielding promising results. While various techniques have been proposed, applying these methods to pretrained text-to-image latent diffusion models like StableDiffusion still remains an open problem, as fine-tuning the off-the-shelf LDM is computationally intensive. We discovered a natural integration of CUT loss into the off-the-shelf LDM within the DDS framework without fine-tuning. This zero-shot integration significantly improves editing output compared to traditional DDS approaches.

3. Main Contribution: The CDS

3.1. Key Observation

DDS. We begin with a concise overview of DDS and then compare its similarities and differences with CUT. This comparative explanation clearly illustrates the missing component in DDS and inspires us on how to improve DDS.

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

pressed 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 while $\boldsymbol{z}_t$ 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:

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

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 [10] 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 $\nabla_\theta \mathcal{L}_{\mathrm{SDS}}$ is an efficient gradient term for generating images that exhibit 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 the SDS framework for image editing, utilizing not only the target text prompt y but also a reference pair of image $\hat{\boldsymbol{z}}_0$ and text $\hat{y}$. Specifically, the DDS loss is given by

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

where $\hat{\boldsymbol{z}}_t = a_t \hat{\boldsymbol{z}}_0 + b_t$. In the same manner as SDS, $\boldsymbol{z}_0(\theta)$ is updated incrementally in the direction of $\nabla_\theta \mathcal{L}_{\mathrm{DDS}}$. 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 a feature vector, DDS aims to minimize the differences in the feature domain.

CUT. The fundamental idea of CUT is to exploit patch-wise contrastive learning in the feature domain for one-sided translation. Specifically, the generated output should produce features whose patches resemble corresponding patches from the input image, in contrast to other random patches. As shown in Fig. 2b, CUT utilizes a multilayer, patch-wise contrastive loss, which maximizes mutual information between corresponding input and output patches. This facilitates one-sided translation in the unpaired setting without imposing the need for enforcing the cycle consistency. CUT has been demonstrated that the CUT loss is effective in maintaining correspondence in content and structure, making it widely utilized technique in image editing.

3.2. Contrastive Denoising Score

By inspection of Fig. 2a and Fig. 2b, we observe a striking similarity between the two approaches. Aside from the image generation processes (i.e., via a trained decoder network in CUT and image optimization in DDS), both approaches aim 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 to match 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 excessive structural changes, we are interested in the key idea from CUT, which has been shown to be effective in maintaining the input structure. However, the original CUT algorithm requires training an encoder to extract spatial information from the input image, which is inefficient. Therefore, we aim to calculate CUT loss without introducing an auxiliary encoder, fully leveraging the information of the latent representation of LDM.

One potential approach is to compute CUT loss by directly leveraging the scores $\boldsymbol{\epsilon}_\phi^\omega(\boldsymbol{z}_t(\theta), y, t)$ and $\boldsymbol{\epsilon}_\phi^\omega(\hat{\boldsymbol{z}}_t, \hat{y}, t)$. While this approach demonstrates effectiveness, 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 by 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 conditioned 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 select patches from the feature map $\boldsymbol{h}_l$. 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 cor-
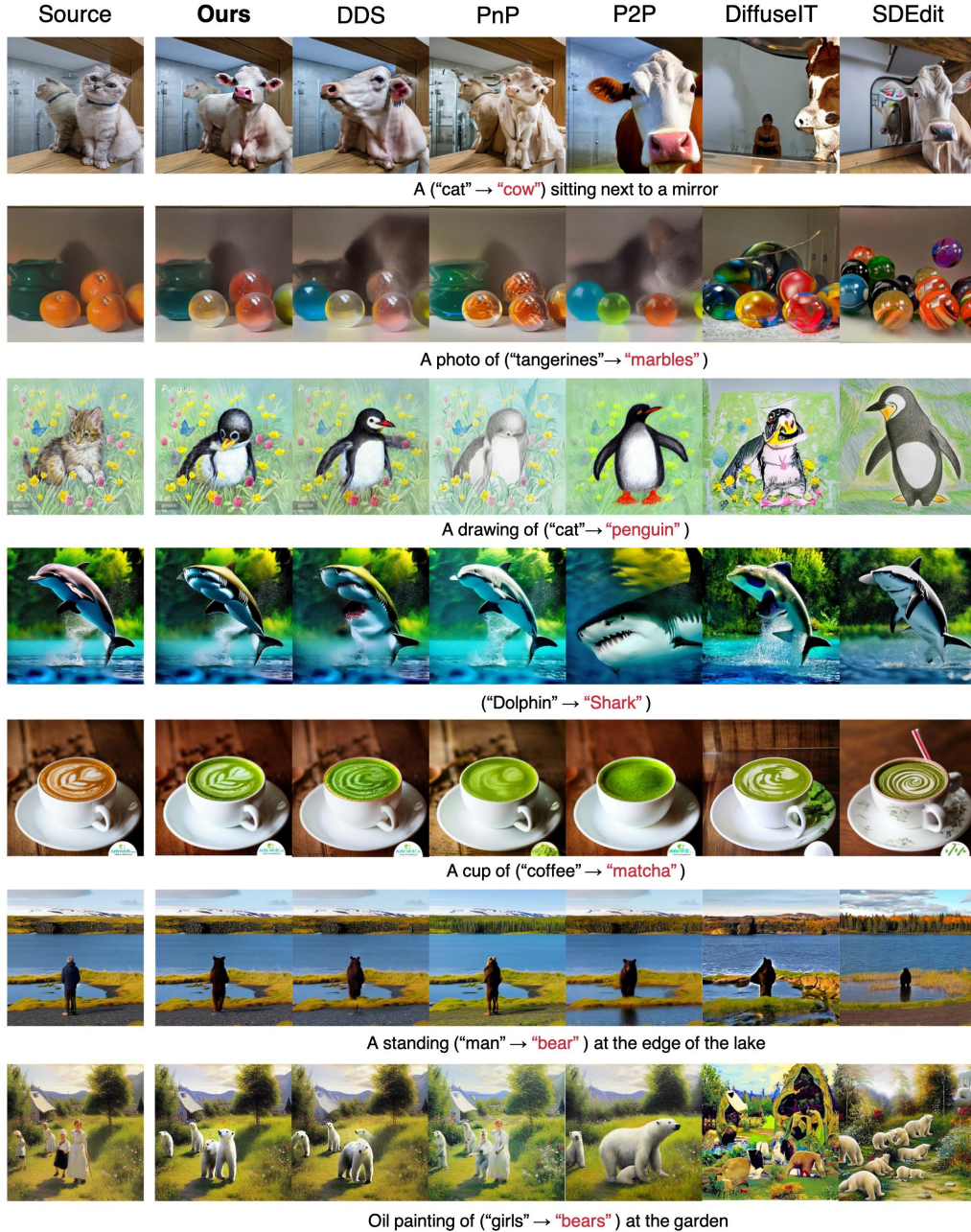
Figure 3. Comparison with baseline models. CDS demonstrates outstanding performance in effectively regulating structural consistency.

responding 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" patches while simultaneously minimizing the mutual information between "negatives." This pro-

cess can be formulated as the patchNCE loss:

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

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

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

Method	cat2dog			cat2cow		
	CLIP Acc (↑)	Dist (↓)	LPIPS (↓)	CLIP Acc (↑)	Dist (↓)	LPIPS (↓)
SDEdit + word swap	**97.9**%	0.052	0.138	97.9%	0.052	0.138
DiffuseIT	**97.9**%	0.174	0.277	93.3%	0.175	0.302
DDPM inv + P2P	86.1%	0.072	0.150	86.1%	0.078	0.158
DDPM inv + PnP	91.6%	0.073	0.156	95.0%	0.079	0.179
DDS	**97.9**%	0.023	0.080	**99.6**%	0.040	0.116
Ours	97.5%	**0.020**	**0.079**	97.9%	**0.033**	**0.112**

Table 1. Quantitative evaluation for the cat → dog and cat → cow tasks. 'Dist' denotes DINO-ViT structure distance. Results for the cat → pig are provided in our Supplementary Materials.

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

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

Table 2. User study results.

4. Experiments

4.1. Experimental setting

Implementation. For implementation, we referenced the official source code of DDS[1] by using Stable Diffusion v1.4. More 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 [15] which we used the implementation of Stable Diffusion. For methods requiring additional inversion process, we employed DDPM inversion [12].

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 change the source image attribute to target text conditions. However, the output structures are severely deformed compared to the original source image structures, and in some cases, unwanted artifacts are present. For the attention-based method of Prompt-to-Prompt(P2P) [8], the results also follow the target text conditions. However, the method suffers

---

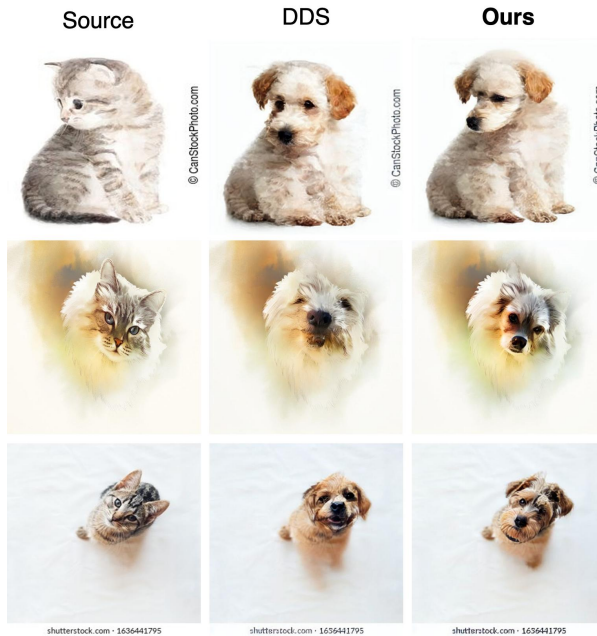[1] https://github.com/google/prompt-to-prompt/blob/main/DDS_zeroshot.ipynb

Figure 4. Sample results of the cat2dog task from DDS and CDS.

from severe structure inconsistencies because the quality is largely affected by the performance of the inversion method and the timesteps for attention map modulation. For the Plug-and-Play(PnP) [27] diffusion baseline, and the overall structure is well maintained from source images, and the output semantics reflect the text conditions. However, the method still does not fully maintain 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 it fails to maintain structural consistency between the source images. On the other hand, our method, CDS successfully edits the source images while preserving their original structural information and does not modify the unrelated regions (e.g. background).

Quantitative results In order to further measure the generation quality of our proposed method, we conduct three tasks: (1) cat → dog, (2) cat → cow and (3) cat → pig. Adhering to the data collection protocol outlined in [20], we gather 250 images relevant to cats from the LAION 5B dataset [26]. In order to assess whether targeted semantic contents are accurately reflected in the generated images, we measure CLIP [22] accuracy(CLIP Acc). Additionally, to measure the structural consistency, we calculated DINO-ViT [3] structure distance and LPIPS distance between the given source image and the outputs. The DINO-ViT structure distance is defined as the difference in self-similarity among the keys obtained from the attention module at the deepest layer of DINO-ViT. Tab. 1 shows that our pro-

A drawing of ("cat" → "penguin")

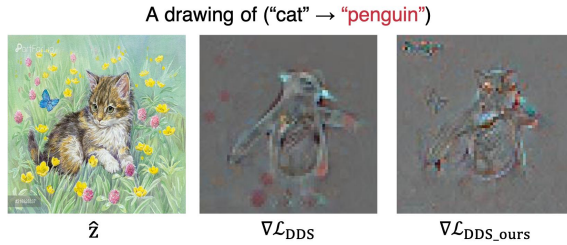$\hat{z}$ $\nabla\mathcal{L}_{\text{DDS}}$ $\nabla\mathcal{L}_{\text{DDS\_ours}}$

Figure 5. Gradient visualization on DDS and CDS.

posed method achieves high CLIP-Acc while maintaining low structure distance, indicating optimal editing results while preserving the structural elements of the original input image. Visualizations are provided in the Figure 4.

For human evaluation, we presented six comparison results and collected feedback from 20 participants. 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. The following are the provided questions for evaluation: (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?* In Tab. 2, our model showed the best performance.

5. Analysis

5.1. Ablation Study

CUT loss. First, we ablate our loss ℓ_{con} and demonstrate its 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} using features from the self-attention layers, we can reliably edit images while reflecting both the source image structure and the target text semantics.

Furthermore, we also conducted a study on the visualization of gradients for both vanilla DDS and CDS with the proposed loss. In Fig. 5, we present 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 CDS, the gradient contains 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 the DDS gradient, leading to further improvement in the final edited output.

CUT loss location. First, we evaluated the effectiveness of feature extraction layer for CUT loss calculation, as
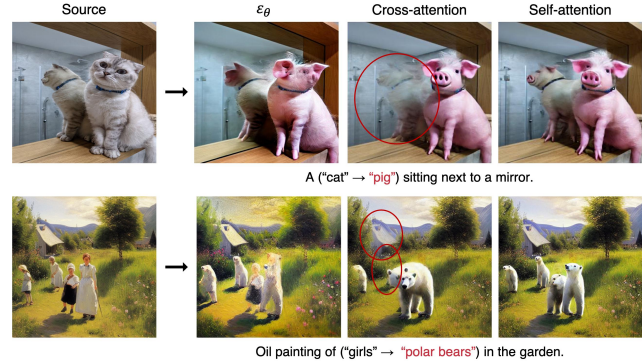

Figure 6. Qualitative results for ablation study on feature extraction location for contrastive loss.
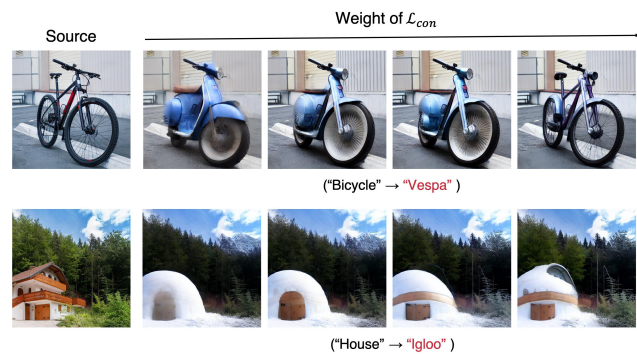

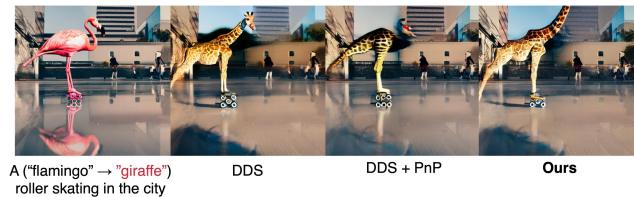Figure 7. Ablation study on weights of contrastive loss.

Figure 8. Comparison with feature injection method.

shown in Fig. 6. We show the generated outputs with the additional CUT loss applied on the direct score network output, the hidden state of cross attention layer, and the hidden state of self-attention layers, respectively. Additionally, the results in Fig. 6 illustrate that directly applying the losses to the score output excessively constrains information. When applying the loss on cross attention features, we can see that the outputs do not correctly preserve the original structures. Our results show that applying proposed method has the best performance. This demonstrates that the features extracted from self-attention layers possess disentangled spatial information, which aligns well with the requirements of CUT loss.
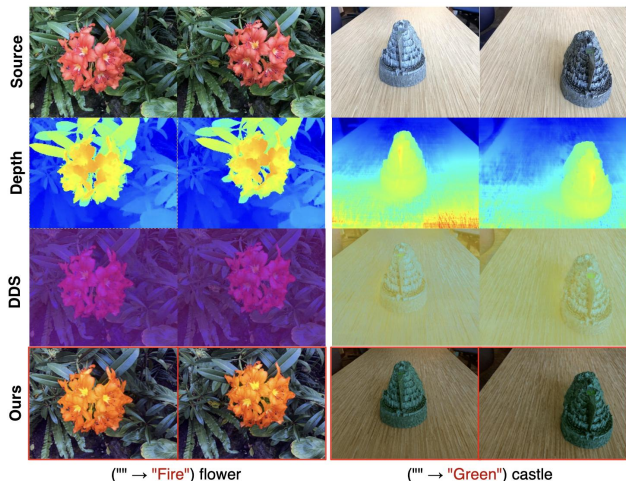
Figure 9. Results on NeRF editing. As an extension of our proposed framework, we applied our method to the NeRF 3D object editing task.

CUT loss weight. For further analysis of our proposed loss, we conducted additional experiments by varying the loss weights. Intuitively, we expect better structural consistency with higher contrastive loss weights and better semantic changes with lower weights. In Fig. 7, we observe that we can control the editability and consistency by varying the weight of contrastive loss. We can see that using a stronger loss not only affects the preservation of the target object structure but also impacts the preservation of the background area.

5.2. Comparison with feature injection method

We noted that recent works, including PnP, leverage feature injection for the purpose of structure preservation. Therefore, we compared the outputs of DDS, DDS with feature injection following the PnP approach, and CDS. In Fig. 8, we can see that while feature injection aids in structural preservation, its forced application without considering semantic information can result in over-constraint during extensive edits. In contrast, CDS employs regularization based on semantic similarity between patches, determined through the inner product in a well-pretrained feature space. This enables CDS to adeptly balance text-aligned editing with the preservation of structural elements.

5.3. Extension to NeRF

Since our propose method is an improved version of the score distillation framework, we can transfer the distilled score gradient to other generator networks, such as NeRF. Inspired by the recent work ED-NeRF [18], which aimed to leverage DDS for text-guided 3D object editing, we applied our proposed method to a pre-trained NeRF. Beginning with original NeRF model, we rendered the 2D images for refer-

ence and target NeRF models with the same view directions. Subsequently, we applied our proposed framework to the NeRF fine-tuning task. For comparison, we also conducted experiments with our baseline DDS. Additional experimental details are provided in our Supplementary Materials.

In Fig. 9, we compare the results of applying vanilla DDS and our proposed framework to pre-trained NeRF model. With vaniila DDS, the outputs fail to capture the structural information of source images, resulting in changes to entire pixel colors. However, our proposed method accurately edits the target object(e.g.a flower) to match the desired text conditions. These 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.

6. Conclusion

In this paper, we proposed CDS, which incorporates 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 representations extracted from LDM, particularly the self-attention layer. This loss allows us to successfully generate images 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. Regarding limitations, failure cases can arise from unfavorable random patch selections or when the source object has unconventional poses, such as looking back.

Acknowledgements

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 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. 6

[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*, pages 2427–2436, 2019. 3

[6] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4291–4301, 2024. 2

[7] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 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, 6

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[11] 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)*, pages 172–189, 2018. 3

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

[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] Gihyun Kwon and Jong Chul Ye. One-shot adaptation of gan in just one clip. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[15] 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. 2, 6

[16] 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

[17] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. 2

[18] JangHo Park, Gihyun Kwon, and Jong Chul Ye. ED-neRF: Efficient text-guided editing of 3d scene with latent space neRF. In *The Twelfth International Conference on Learning Representations*, 2024. 8

[19] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. 2, 3

[20] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 6

[21] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[24] 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

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. 2

[26] 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. 6

[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