Identity-preserving Distillation Sampling by Fixed-Point Iterator

Supplementary Material

Figure S1. **Posterior mean with/without FPR.** When the prompt y is given by "portrait of a worried-looking woman in a dress", the posterior mean $\mathbf{z}_{0|t}$ is obtained $(first\ row)$ without FPR, $(second\ row)$ with FPR w.r.t $\mathbf{z}_t$, and $(third\ row)$ with FPR w.r.t ϵ.

A. Posterior mean analysis

To investigate how much identity of the original image $\mathbf{z}$ is contained in the text-conditioned score $\epsilon_\phi(\mathbf{z}, y, t)$, we conduct the experiment in which the posterior mean is obtained from various timesteps. As shown in the first row of Fig. S1, more primary information is damaged as the timestep t increases. On the other hand, when using FPR, since the score $\epsilon_\phi(\mathbf{z}, y, t)$ is modified to preserve the identity of $\mathbf{z}$, we can see that it has more information than before, even at large timestep, as described in the second and third row of Fig. S1. Note that the score $\epsilon_\phi(\mathbf{z}, y, t)$ can be controlled by updating the injection noise ϵ or the noisy latent $\mathbf{z}_t$. Of the two options, it has been updated for $\mathbf{z}_t$ because it contains more content details.

B. Metrics for FPR

As defined in Eq. (6), $d(\mathbf{x}_1, \mathbf{x}_2)$ can be any metric to calculate the difference between two inputs. For comparison, we consider three different strategies: (1) Euclidean loss, (2) L1 loss, and (3) SSIM loss. As demonstrated in Fig. S2, all metrics can be applied to our method for image editing according to text prompts. Among these, the use of Euclidean loss is particularly notable, as it effectively preserved the original information while producing visually superior results.

Figure S2. **Ablation study for loss function.** Edited results of *(first)* the source image from prompt *"a drawing of a cat"* to *"a drawing of a dog"* using *(second)* Euclidean, *(third)* L1, and *(fourth)* SSIM loss function for FPR.

C. Implementation details

For experiments, we implement our method based on the official code of CDS [1] by using Stable Diffusion v1.4. All baselines are implemented based on the official code and setting for each method. For the proposed FPR, we set the scale λ to 1.0 and iteration N to 3. The range of timesteps, optimization, learning rate, and number of optimization steps correspond to the default settings employed in DDS and CDS. All experiments are conducted on a single NVIDIA RTX 3090.

D. Evaluation metrics

Our purpose is to preserve the source information by optimizing the score $\epsilon_\phi^{\text{src}}$. Thus, in addition to the LPIPS, we newly utilize IoU and background PSNR as our metrics to measure the structural similarity between the source and edited image.

IoU. The aim of *Cat-to-Others* task is to translate the cat into another animal. Thus, the segmentation mask of the cat and translated animal can be obtained using the language Segment-Anything model (lang-SAM)[2], which is an open-source project to segment some objects from the text prompt. IoU of the source and target mask represents how much the area of the cat changes after image editing. The lower the IoU, the more similar the region of the cat and the region of the translated animal, meaning the overall shape is preserved. To this end, first, the mask about the prompt is obtained from an image using lang-SAM. For example, 'cat' is segmented from the source image to get the mask M_{src}, while 'dog' is segmented from the edited image to obtain the mask M_{trg}, as shown in Fig. S3 (a). After getting masks, we calculate IoU from the masks that are given by:

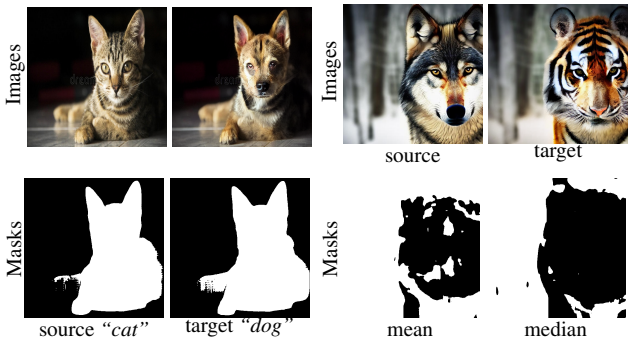
$$\text{IoU} = \frac{(M_{\text{src}} \cap M_{\text{trg}})}{(M_{\text{src}} \cup M_{\text{trg}})}$$

[1] https://hyelinnam.github.io/CDS/
[2] https://github.com/paulguerrero/lang-sam

Background PSNR. Since the editing prompts of IP2P dataset [1] is complex than *Cat-to-Others* dataset [7, 8], it is hard to get mask by lang-SAM. Therefore, we use background PSNR to evaluate how much the original information is preserved. The residual of the source and target images is calculated, and the standard deviation σ of each pixel of the residual image is computed with window size 30. Then, the mask M_{PSNR} is acquired by thresholding the σ. Since the range of σ varies according to the edited results for each method, we use the mean or median values of σ to set an appropriate threshold (see Fig. S3 (b)). For the background PSNR of Tab. 1, we use mean threshold. Finally, we calculate PSNR values from masked source and target images:

$$\text{PSNR}_{\text{back}} = \text{PSNR}(M_{\text{PSNR}} \odot \mathbf{z}_{\text{src}}, M_{\text{PSNR}} \odot \mathbf{z}_{\text{trg}})$$

where $\odot$ is pixel-wise multiplication.



(a) IoU mask (b) Background PSNR mask

Figure S3. **Calculated masks** for IoU and background PSNR. In (a), *(second row)* each mask for *(first row)* the source and target image is obtained by using lang-SAM for IoU. In (b), *(second row)* a mask is calculated for *(first row)* the source and target image to measure background PSNR between the masked source and target image. The mask can be generated by thresholding method, mean and median

E. Extension to other methods

Since our method optimizes the source latent to estimate a more accurate score, it can be applied to other methods that are based on SDS despite that we report the results using our method to DDS.

During SDS optimization, FPR can be used to preserve the original content and reduce the blurry effect. As shown in Fig. S4, the conserved rate of the information of the source image is controllable by the number of FPR iteration.

When the proposed FPR is integrated into CDS, the texture of the source image is further maintained, as illustrated in Fig. S5. In addition, FPR promoted reducing the over-boosting of color often found in the translated images of

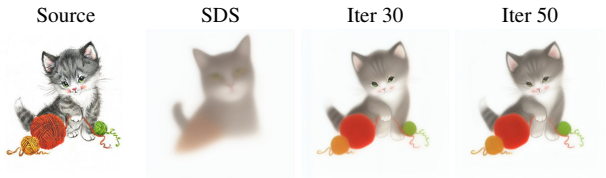


Figure S4. **SDS with FPR.** Given *(first)* source image and prompt *"a drawing of a cat"*, *(second)* SDS optimization, *(third, fourth)* SDS optimization with FPR for 30 and 50 iterations are applied. Each result uses 200 steps for optimization.

Figure S5. **CDS with FPR.** Given *(first)* source image, source prompt *"a drawing of a cat"*, and target prompt *"a drawing of a pig"*, *(second)* CDS translation, *(third)* CDS optimization with FPR for $N = 3$ and $\lambda = 1.0$.

CDS. This confirms that the proposed FPR can be a universal regularization to preserve the identity of the source image for text-guided image editing.

Furthermore, FPR can help optimize not only pixel space but also the parametric editor such as PDS [6]. As demonstrated in Fig. S6, Fig. S7, and Tab. S1, the edited results with our method show that FPR assists in maintaining the original contents. By comparing the first and second rows of Fig. S6, the use of FPR results in the preservation of source components more effectively compared to PDS. Similarly, in the third and fourth rows, the results obtained using FPR retain key original features, such as the shape and color of the face as well as the color of the clothing. Furthermore, the gradient weights, FPR assigns minimal weight to the structure of the source image, such as the background, while primarily focusing the weights on the editing points. For 3D and 2D editing, we implement the experiments based on official code of PDS[3]. We use the subset of Instruct-NeRF2NeRF [2] for 3D editing and Scalable Vector Graphics (SVGs) with their text description used in [5] for 2D editing.

	NeRF		SVG	
Metric	CLIP(↑)	LPIPS(↓)	CLIP(↑)	LPIPS(↓)
SDS	0.305	0.814	**0.346**	0.552
DDS	**0.306**	0.875	0.344	0.557
PDS	0.292	0.662	0.324	0.326
Ours+PDS	0.295	**0.587**	0.327	**0.274**

Table S1. **Quantitative results** for PDS.

[3] https://github.com/KAIST-Visual-AI-Group/PDS

Figure S6. **3D Qualitative results for PDS** on subset of Instruct-NeRf2NeRF [2]. From left to right, each column represents the source image, the edited image, and the gradient weight. The gradient weight indicates which regions the model primarily references during the editing process. The results demonstrate that FPR operates effectively in End-to-End NeRF while preserving the structure and identity of the source image.

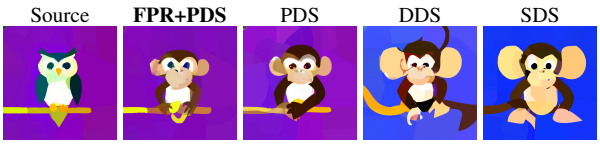


Figure S7. **2D Qualitative results for PDS on VectorFusion [5]**. In SVG editing, our method can be utilized with PDS and help the source identity maintained.

F. Additional results

We also provide qualitative results for *Cat-to-Others* task, as demonstrated in Fig. S8. With DDS and CDS, the direction of the gaze changes when translated from the cat to the squirrel, while it remains the same with IDS. Note that the proposed IDS can also retain the hue of the source image without overemphasizing the colors, as demonstrated in *Cat-to-Tiger* task. This confirms that the proposed IDS consistently offers suitable editing of cat images into the diverse animals, while conserving the identity of the source against other algorithms.

The trends in the quantitative results are also consistent with the qualitative result, as represented in Tab. S2. Our method provides the best performance for LPIPS and IoU in most *Cat-to-Others* tasks. This shows again that the self-

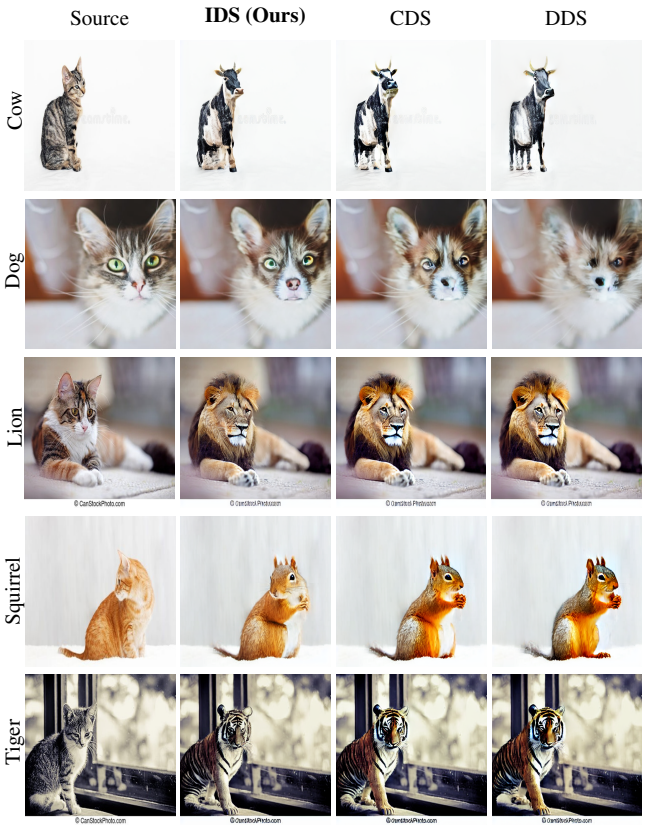


Figure S8. **Qualitative results** of *Cat-to-Others* task. The leftmost text means each target prompt, and each row shows the editing results from 'Cat" to the target prompt.

correction of the score using the proposed algorithm is crucial for maintaining the identity.

G. Limitations

Success rate. As discussed in Sec. 7, our method optimizes the latents only for source information, resulting in low CLIP scores. To demonstrate it does not mean *"IDS fails to translate the source image"*, we measure the success rate. To calculate the success rate, we classify the transformed images with the pre-trained ResNet classifier on the *Cat-to-dog* task. We treat the results of classified top 1 as the success sample if they belonged to the class of *dog*. The success rate in Tab. S3 claims that the low CLIP score of IDS did not fall to convert, but occurred in the process of maintaining the source identity.

Failure case for complex prompt. Because our method only considers the source information, it struggles with translating the given image for complex text prompts. Although we tried to modify the image with more complex prompts, it failed not only in our method but also in all SDS-based translation methods, as shown in Fig. 8.

Computational overhead. Our method requires additional

	cat2cow		cat2dog		cat2lion		cat2tiger		cat2penguin	
Metric	LPIPS (↓)	IoU (↑)	LPIPS (↓)	IoU (↑)	LPIPS (↓)	IoU (↑)	LPIPS (↓)	IoU (↑)	LPIPS (↓)	IoU (↑)
P2P [4]	0.43	0.57	0.42	0.51	0.46	0.57	0.47	0.57	0.46	0.54
PnP [9]	0.52	0.55	0.47	0.59	0.51	0.58	0.52	0.58	0.52	0.52
DDS [3]	0.29	0.65	0.22	0.72	0.29	0.69	0.30	0.71	0.28	0.66
CDS [7]	0.25	0.72	0.19	0.74	0.25	**0.74**	0.27	0.75	0.24	**0.72**
IDS (Ours)	**0.21**	**0.74**	**0.17**	**0.75**	**0.21**	0.71	**0.21**	**0.76**	**0.21**	**0.72**

Table S2. **Quantitative results** for *Cat-to-Others* task. LPIPS [10] and IoU are used. Lower LPIPS and higher IoU mean better identity preserving.

computational costs due to repetitive adaptations of FPR for each optimization steps. However, it can be controlled by adjusting hyperparameters such as the number of FPR iterations or the number of optimization steps, as reported in Tab. 4.

	IDS	CDS	DDS
success rate (%)	37.40	34.87	34.03

Table S3. **Success rate** for *Cat-to-dog* task. A higher score means more translated results are classified as *dog*.

H. Social impact

By optimizing for a given image, our method properly mitigates the undesired biases introduced by the generative priors of large text-to-image diffusion models. However, the issue of bias toward target information persists. Furthermore, the method's potential misuse for generating fake content highlights a critical ethical challenge commonly associated with image editing techniques. To mitigate these risks, it is essential to implement robust safeguards, such as stricter content authentication mechanisms, and ethical guidelines for usage.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[2] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 2, 3

[3] 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. 4

[4] 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. 4

[5] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1911–1920, 2023. 2, 3

[6] Juil Koo, Chanho Park, and Minhyuk Sung. Posterior distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13352–13361, 2024. 2

[7] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9192–9201, 2024. 2, 4

[8] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[9] 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. 4

[10] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4