Posterior Distillation Sampling — Supplementary Material

Juil Koo Chanho Park Minhyuk Sung
KAIST
{63days,charlieppark,mhsung}@kaist.ac.kr

In this supplementary material, we first present additional editing results with more diverse representations, including 3D Gaussian Splats (3DGS) [6] and 2D images, in Section S.1. Then, in Section S.2, we provide a more detailed derivation of Posterior Distillation Sampling introduced in Section 4 of the main paper. Following that, Section S.3 presents the implementation details of NeRF editing and SVG editing discussed in Sections 6.1 and 6.2 of the main paper, respectively. Subsequently, Section S.4 provides the details of our user study setups. Lastly, the effect of the refinement stage, discussed in Section 5 of the main paper, is detailed in Section S.5.

S.1. Editing 3D Gaussian Splats [6] and 2D Images

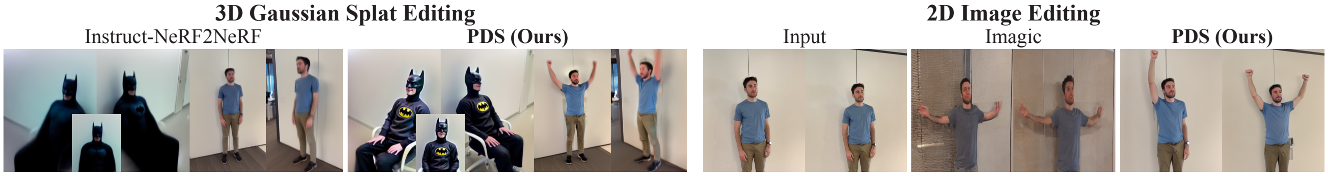


Figure S1. **Editing of more diverse representations, 3D Gaussian Splats [6] and 2D images.** PDS consistently outperforms the baselines. The target attributes are *"Batman"* and *"raising the arms."*

PDS encompasses various editing scenarios, not confined within a specific parameter space. To further assess the versatility and generalizability of PDS in editing tasks, we include both 3D Gaussian Splat (3DGS) [6] editing and 2D image editing. As NeRF editing, Figure S1 shows that PDS outperforms Instruct-NeRF2NeRF [1] in 3DGS representation while uniquely realizing geometric changes. In 2D image editing, PDS demonstrates superior performance compared to Imagic [5], which is introduced for 2D image editing using pre-trained 2D diffusion models. PDS edits the input image while preserving other details with high fidelity. On the other hand, Imagic [5] leaves artifacts, losing the identity of the source content.

S.2. Derivation of Posterior Distillation Sampling

For a comprehensive derivation of Equation 14 in the main paper, we first remind that the objective function of PDS is expressed as:

$$\mathcal{L}_{\tilde{\mathbf{z}}_t}(\mathbf{x}_0^{\text{tgt}}) = \mathbb{E}\left[\|\tilde{\mathbf{z}}_t^{\text{tgt}} - \tilde{\mathbf{z}}_t^{\text{src}}\|_2^2\right] \tag{1}$$

$$= \mathbb{E}\left[\left\|\frac{\mathbf{x}_{t-1}^{\text{tgt}} - \boldsymbol{\mu}_\phi(\mathbf{x}_t^{\text{tgt}}, y^{\text{tgt}}; \boldsymbol{\epsilon}_\phi)}{\sigma_t} - \frac{\mathbf{x}_{t-1}^{\text{src}} - \boldsymbol{\mu}_\phi(\mathbf{x}_t^{\text{src}}, y^{\text{src}}; \boldsymbol{\epsilon}_\phi)}{\sigma_t}\right\|_2^2\right] \tag{2}$$

$$= \mathbb{E}\left[\frac{1}{\sigma_t^2}\left\|(\mathbf{x}_{t-1}^{\text{tgt}} - \mathbf{x}_{t-1}^{\text{src}}) - \left(\boldsymbol{\mu}_\phi(\mathbf{x}_t^{\text{tgt}}, y^{\text{tgt}}; \boldsymbol{\epsilon}_\phi) - \boldsymbol{\mu}_\phi(\mathbf{x}_t^{\text{src}}, y^{\text{src}}; \boldsymbol{\epsilon}_\phi)\right)\right\|_2^2\right]. \tag{3}$$

Given that $\tilde{\mathbf{z}}_t^{\text{src}}$ and $\tilde{\mathbf{z}}_t^{\text{tgt}}$ share the same noises $\boldsymbol{\epsilon}_{t-1}$ and $\boldsymbol{\epsilon}_t$ for their respective $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, the difference between $\mathbf{x}_{t-1}^{\text{tgt}}$ and $\mathbf{x}_{t-1}^{\text{src}}$ results in a constant multiple of the difference between $\mathbf{x}_0^{\text{tgt}}$ and $\mathbf{x}_0^{\text{src}}$:

$$\mathbf{x}_{t-1}^{\text{tgt}} - \mathbf{x}_{t-1}^{\text{src}} = \sqrt{\bar{\alpha}_{t-1}}(\mathbf{x}_0^{\text{tgt}} - \mathbf{x}_0^{\text{src}}). \tag{4}$$

Following our notation $\hat{\boldsymbol{\epsilon}}_t^{\text{src}} := \boldsymbol{\epsilon}_\phi(\mathbf{x}_t^{\text{src}}, y^{\text{src}}, t)$ and $\hat{\boldsymbol{\epsilon}}_t^{\text{tgt}} := \boldsymbol{\epsilon}_\phi(\mathbf{x}_t^{\text{tgt}}, y^{\text{tgt}}, t)$ introduced in Section 4 of the main paper, the difference between the approximated posterior means is also expressed as follows:

$$\boldsymbol{\mu}_\phi(\mathbf{x}_t^{\text{tgt}}, y^{\text{tgt}}; \boldsymbol{\epsilon}_\phi) - \boldsymbol{\mu}_\phi(\mathbf{x}_t^{\text{src}}, y^{\text{src}}, \boldsymbol{\epsilon}_\phi) = (\gamma_t + \delta_t\sqrt{\bar{\alpha}_t})(\mathbf{x}_0^{\text{tgt}} - \mathbf{x}_0^{\text{src}}) - \gamma_t\sqrt{\frac{1}{\bar{\alpha}_t} - 1}(\hat{\boldsymbol{\epsilon}}_t^{\text{tgt}} - \hat{\boldsymbol{\epsilon}}_t^{\text{src}}), \tag{5}$$

where $\boldsymbol{\mu}_\phi(\mathbf{x}_t, y; \boldsymbol{\epsilon}_\phi)$ can be expanded as shown in the following equation:

$$\begin{aligned}
\boldsymbol{\mu}_\phi(\mathbf{x}_t, y; \boldsymbol{\epsilon}_\phi) &= \gamma_t \tilde{\mathbf{x}}_0(\mathbf{x}_t, y; \boldsymbol{\epsilon}_\phi) + \delta_t \mathbf{x}_t & (6)\\
&= \gamma_t \left(\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, y, t) \right) + \delta_t \mathbf{x}_t & (7)\\
&= (\frac{\gamma_t}{\sqrt{\bar{\alpha}_t}} + \delta_t)\mathbf{x}_t - \gamma_t\sqrt{\frac{1}{\bar{\alpha}_t} - 1}\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, y, t) & (8)\\
&= (\gamma_t + \delta_t\sqrt{\bar{\alpha}_t})\mathbf{x}_0 + \sqrt{\frac{1}{\bar{\alpha}_t} - 1}(\gamma_t + \delta_t\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_t - \gamma_t\sqrt{\frac{1}{\bar{\alpha}_t} - 1}\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, y, t). & (9)
\end{aligned}$$

Incorporating Equation 4 and Equation 5 into Equation 3, we can reformulate the objective function of PDS as follows:

$$\begin{aligned}
\mathcal{L}_{\tilde{\mathbf{z}}_t}(\mathbf{x}_0^{\text{tgt}}) &= \mathbb{E}\left[\frac{1}{\sigma_t^2} \left\| (\sqrt{\bar{\alpha}_{t-1}} - \gamma_t - \delta_t\sqrt{\bar{\alpha}_t})(\mathbf{x}_0^{\text{tgt}} - \mathbf{x}_0^{\text{src}}) + \gamma_t\sqrt{\frac{1}{\bar{\alpha}_t} - 1}(\hat{\boldsymbol{\epsilon}}_t^{\text{tgt}} - \hat{\boldsymbol{\epsilon}}_t^{\text{src}}) \right\|_2^2 \right] & (10)\\
&= \mathbb{E}\left[\frac{1}{\sigma_t^2} \left((\sqrt{\bar{\alpha}_{t-1}} - \gamma_t - \delta_t\sqrt{\bar{\alpha}_t})^2(\mathbf{x}_0^{\text{tgt}} - \mathbf{x}_0^{\text{src}})^2 \right.\right. & (11)\\
&\qquad + 2(\sqrt{\bar{\alpha}_{t-1}} - \gamma_t - \delta_t\sqrt{\bar{\alpha}_t})\gamma_t\sqrt{\frac{1}{\bar{\alpha}_t} - 1}(\mathbf{x}_0^{\text{tgt}} - \mathbf{x}_0^{\text{src}})(\hat{\boldsymbol{\epsilon}}_t^{\text{tgt}} - \hat{\boldsymbol{\epsilon}}_t^{\text{src}}) \\
&\qquad \left.\left. + \gamma_t^2(\frac{1}{\bar{\alpha}_t} - 1)(\hat{\boldsymbol{\epsilon}}_t^{\text{tgt}} - \hat{\boldsymbol{\epsilon}}_t^{\text{src}})^2 \right) \right].
\end{aligned}$$

By taking the gradient of $\mathcal{L}_{\tilde{\mathbf{z}}_t}$ with respect to θ while ignoring the U-Net jacobian term, $\frac{\partial \hat{\boldsymbol{\epsilon}}_\phi^{\text{tgt}}}{\partial \mathbf{x}_0^{\text{tgt}}} = \mathbf{I}$, one can obtain PDS as follows:

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\text{PDS}} &= \frac{\partial \mathcal{L}_{\tilde{\mathbf{z}}_t}(\mathbf{x}_0^{\text{tgt}})}{\partial \mathbf{x}_0^{\text{tgt}}} \cdot \frac{\partial \mathbf{x}_0^{\text{tgt}}}{\partial \theta} & (12)\\
&= \mathbb{E}\left[\frac{2}{\sigma_t^2} \left((\sqrt{\bar{\alpha}_{t-1}} - \gamma_t - \delta_t\sqrt{\bar{\alpha}_t})^2(\mathbf{x}_0^{\text{tgt}} - \mathbf{x}_0^{\text{src}}) + (\sqrt{\bar{\alpha}_{t-1}} - \gamma_t - \delta_t\sqrt{\bar{\alpha}_t})\gamma_t\sqrt{\frac{1}{\bar{\alpha}_t} - 1}(\hat{\boldsymbol{\epsilon}}_t^{\text{tgt}} - \hat{\boldsymbol{\epsilon}}_t^{\text{src}}) \right) \frac{\partial \mathbf{x}_0^{\text{tgt}}}{\partial \theta} \right]. & (13)
\end{aligned}$$

Thus, the coefficients $\psi(t)$ and $\chi(t)$ in Equation 14 of the main paper are as follows:

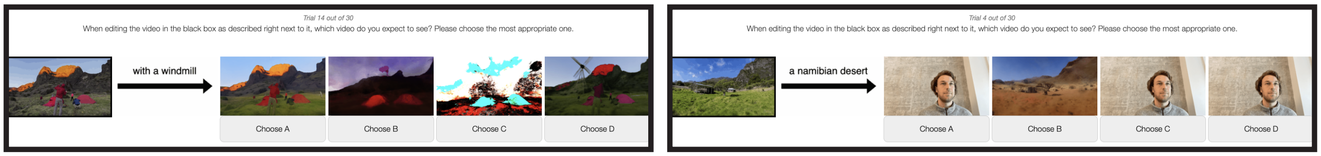
$$\begin{aligned}
\psi(t) &= \frac{2(\sqrt{\bar{\alpha}_{t-1}} - \gamma_t - \delta_t\sqrt{\bar{\alpha}_t})^2}{\sigma_t^2}, & (14)\\
\chi(t) &= \frac{2(\sqrt{\bar{\alpha}_{t-1}} - \gamma_t - \delta_t\sqrt{\bar{\alpha}_t})}{\sigma_t^2}\gamma_t\sqrt{\frac{1}{\bar{\alpha}_t} - 1}. & (15)
\end{aligned}$$

In practice, we sample non-consecutive timesteps for $t-1$ and t as in DDIM [9] since the coefficients become 0 when they are consecutive. Given a sequence of non-consecutive timesteps $[\tau_i]_{i=1}^S$, a more generalized form of PDS is represented as follows:

$$\nabla_\theta \mathcal{L}_{\text{PDS}} = \mathbb{E}_{i, \boldsymbol{\epsilon}_{\tau_i}, \boldsymbol{\epsilon}_{\tau_{i-1}}}\left[\psi(i)(\mathbf{x}_0^{\text{tgt}} - \mathbf{x}_0^{\text{src}}) + \chi(i)(\hat{\boldsymbol{\epsilon}}_{\tau_i}^{\text{tgt}} - \hat{\boldsymbol{\epsilon}}_{\tau_i}^{\text{src}})\frac{\partial \mathbf{x}_0^{\text{tgt}}}{\partial \theta} \right], \tag{16}$$

where

$$\begin{aligned}
\psi(i) &= \frac{2(\sqrt{\bar{\alpha}_{\tau_{i-1}}} - \gamma_{\tau_i} - \delta_{\tau_i}\sqrt{\bar{\alpha}_{\tau_i}})^2}{\sigma_{\tau_i}^2}, & (17)\\
\chi(i) &= \frac{2(\sqrt{\bar{\alpha}_{\tau_{i-1}}} - \gamma_{\tau_i} - \delta_{\tau_i}\sqrt{\bar{\alpha}_{\tau_i}})}{\sigma_{\tau_i}^2}\gamma_{\tau_i}\sqrt{\frac{1}{\bar{\alpha}_{\tau_i}} - 1}. & (18)
\end{aligned}$$



Trial 14 out of 30

When editing the video in the black box as described right next to it, which video do you expect to see? Please choose the most appropriate one.

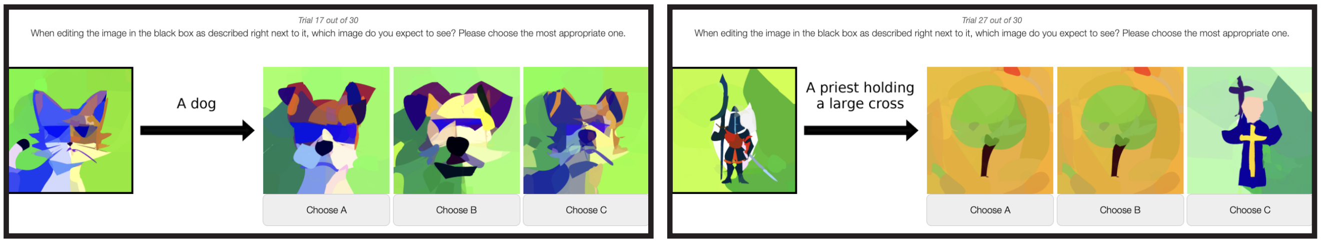
Trial 4 out of 30

When editing the video in the black box as described right next to it, which video do you expect to see? Please choose the most appropriate one.

(a) Main problem

(b) Vigilance task

Figure S2. **NeRF editing user study screenshots.** The participants are presented with NeRF scene videos and editing prompts, and are asked to answer the following question: `When editing the video in the black box as described right next to it, which video do you expect to see? Please choose the most appropriate one.`



Trial 17 out of 30

When editing the image in the black box as described right next to it, which image do you expect to see? Please choose the most appropriate one.

Trial 27 out of 30

When editing the image in the black box as described right next to it, which image do you expect to see? Please choose the most appropriate one.

(a) Main problem

(b) Vigilance task

Figure S3. **SVG editing user study screenshots.** Given SVG images and editing prompts, the participants are asked to answer the following question: `When editing the image in the black box as described right next to it, which image do you expect to see? Please choose the most appropriate one.`

For more details on timestep sampling, refer to the implementation details in the next section.

S.3. Implementation Details

In this section, we provide the implementation details of NeRF and SVG editing presented in Section 6.1 and Section 6.2 of the main paper, respectively.

NeRF Editing. We run the PDS optimization for 30,000 iterations with classifier-free guidance [3] weights within $[30, 100]$ depending on the complexity of editing. As detailed in Section S.2, we sample non-consecutive timesteps τ_{i-1} and τ_i since the coefficients $\psi(\cdot)$ and $\chi(\cdot)$ become zero when the sampled timesteps are consecutive. For this, we define non-consecutive timesteps $[\tau_i]_{i=1}^{S}$, which is a subset sequence of the total forward process timesteps of the diffusion model, $[1, ..., T]$. Specifically, we select these timesteps such that $\tau_i = \lfloor 2i \rfloor$, resulting in a subset sequence length of $S = 500$ out of the total $T = 1000$ timesteps. We then randomly sample the index i within a ratio range of $[0.02, 0.98]$, i.e., $i \sim \mathcal{U}(10, 490)$.

During the refinement stage, we randomly choose and replace $\tilde{I}_v$ every 10 iterations, over total 15,000 iterations. We denote a SDEdit [7] operator by $\mathcal{S}(\mathbf{x}_0; t_0, \boldsymbol{\epsilon}_\phi)$ which samples $\mathbf{x}_{t_0} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t_0}}\mathbf{x}_0, (1-\bar{\alpha}_{t_0})\mathbf{I})$ then starts denoising it from t_0 using $\boldsymbol{\epsilon}_\phi$. For the denoising process, we randomly sample t_0 within a ratio range of $[0, 0.2]$ out of total denoising steps $N = 20$.

SVG Editing. Across all optimizations, SDS [8], DDS [2], and our proposed PDS, we apply the same classifier-free guidance weight of 100. For SDS [8], we sample t within a ratio range of $[0.05, 0.95]$ following VectorFusion [4]. For DDS [2], we follow its original setup, sampling t within $[0.02, 0.98]$. For PDS, we sample i out of a ratio range of $[0.1, 0.98]$.

S.4. Details of User Studies

We conduct user studies for the human evaluation of NeRF and SVG editing through Amazon's Mechanical Turk. We collected survey responses only from those participants who passed our vigilance tasks. To design our vigilance tasks, we create examples where, except for the correct answer choice, all other choices are replaced with ones from different scenes or unrelated SVG examples. Screenshots of our NeRF and SVG editing user studies, including examples of vigilance tasks, are displayed in Figure S2 and Figure S3, respectively. In the NeRF and SVG editing user studies, we received 42 and 17 valid responses, respectively.

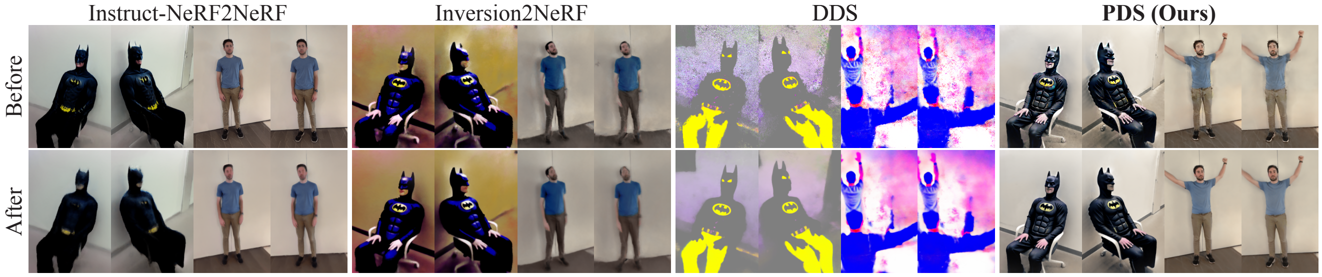


Figure S4. **The effect of the refinement stage.** The overall editing outcomes are determined before the refinement stage, whereas the refinement stage plays the role of removing artifacts. The target attributes are *"Batman"* and *"raising the arms."*

S.5. Effect of the Refinement Stage

Figure S4 illustrates an ablation study of the refinement stage across various editing methods. As depicted, the desired complex edits — making the man raise his arms — are achieved solely through the optimization of PDS. The overall editing outcomes are realized before the refinement stage, and the refinement stage further enhances the fidelity of the outputs.

References

[1] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *ICCV*, 2023. 1

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

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

[4] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *CVPR*, 2023. 3

[5] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 1

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 1

[7] 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 *ICLR*, 2022. 3

[8] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023. 3

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2