Supplementary material of "DreamSampler: Unifying Diffusion Sampling and Score Distillation for Image Manipulation"

Jeongsol Kim[1,*], Geon Yeong Park[1,*], and Jong Chul Ye[2]

[1] Dept. of Bio & Brain Engineering, KAIST
[2] Kim Jae Chul AI graduate school, KAIST
{jeongsol, pky3436, jong.ye}@kaist.ac.kr
* Equal contribution

1 Proof of Theorem 1.

The solution of latent optimization problem

$$\bar{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_\varnothing)\|^2 + \frac{\gamma}{1-\gamma}\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{ref})\|^2 \tag{1}$$

is computed as a closed form of

$$\bar{\boldsymbol{z}} = (1-\gamma)\hat{\boldsymbol{z}}_{0|t}(c_\phi) + \gamma\hat{\boldsymbol{z}}_{0|t}(c_{ref}) \tag{2}$$

because the objective function is convex and

$$2(1-\gamma)(\bar{\boldsymbol{z}} - \hat{\boldsymbol{z}}_{0|t}(c_\phi)) + 2\gamma(\bar{\boldsymbol{z}} - \hat{\boldsymbol{z}}_{0|t}(c_{ref})) = 0. \tag{3}$$

Then,

$$\begin{aligned}
\bar{\boldsymbol{z}} &= (1-\gamma)\hat{\boldsymbol{z}}_{0|t}(c_\varnothing) + \gamma\hat{\boldsymbol{z}}_{0|t}(c_{ref}) && (4)\\
&= \hat{\boldsymbol{z}}_{0|t}(c_\varnothing) - \gamma[\hat{\boldsymbol{z}}_{0|t}(c_\varnothing) - \hat{\boldsymbol{z}}_{0|t}(c_{ref})] && (5)\\
&= \hat{\boldsymbol{z}}_{0|t}(c_\varnothing) - \gamma\left[\frac{\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, c_\varnothing)}{\sqrt{\bar{\alpha}_t}} - \frac{\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, c_{ref})}{\sqrt{\bar{\alpha}_t}}\right] && (6)\\
&= \hat{\boldsymbol{z}}_{0|t}(c_\varnothing) - \gamma\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}[\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, c_\varnothing) - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, c_{ref})] && (7)
\end{aligned}$$

Finally, by substituting Tweedie's formula into $\hat{\boldsymbol{z}}_{0|t}(c_\varnothing)$, we can get

$$\bar{\boldsymbol{z}} = \frac{\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}[\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, c_\varnothing) + \gamma\{\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, c_{ref}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, c_\varnothing)\}]}{\sqrt{\bar{\alpha}_t}} \tag{8}$$

This concludes the proof.

2 Implementation details

In this section, we provide details on the implementation and experimental settings of DreamSampler.

2.1 Image Restoration through Vectorization

For the reverse sampling, we set the CFG scale to 100, and NFE to 1000 for both super-resolution and Gaussian deblurring. For the Lagrangian coefficient, we set $\lambda_{SDS} = 2.4, \lambda_{DC} = 3$ for Guassian deblur and $\lambda_{SDS} = 1, \lambda_{DC} = 4$ for super-resolution. Other optimization configurations, e.g. learning rate, optimization algorithms, paths, etc, follow [8]. Specifically, a self-intersection regularizer $\mathcal{L}_{xing}$ is used with a weight 0.01. The learning rate initiates at 0.02 and linearly escalates to 0.2 across 500 steps, subsequently undergoing a cosine decay to 0.05 upon optimization completion for control point coordinates. Fill colors are subjected to a learning rate that is 20 times lower than that of control points, while the solid background color is allocated a learning rate 200 times lower.

2.2 Real Image Editing

We utilize the DDIM inversion using the null-text to initialize the latent $\boldsymbol{z}_T$. The null-text is defined by empty text "" for both inversion and sampling, which could be leveraged in general. For the noise addition steps at each timestep, we leverage the deterministic sampling by setting $\eta\beta_t = 0$. For the CFG scale γ, we found that time-dependent value between $0.1\bar{\alpha}_t$ and $0.3\bar{\alpha}_t$ can edit image with robustness. As $\bar{\alpha}_t \to 1$ when $t \to 0$, this scale allows early-stage sampling to reconstruct the source image while the later-stage sampling reflects guided direction, according to (17). Here, γ indicates the interpolation coefficient for the latent optimization problem

$$\min_{\boldsymbol{z}} (1-\gamma)\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_\phi)\|^2 + \gamma\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{tgt})\|^2, \tag{9}$$

which is analyzed in section 4.1.

2.3 Text-guided Image Inpainting

For the inapinting task, we utilize the DDIM inversion with null-text where the null-text is "out of focus, depth of field" to adopt concept negation for better generation quality. After solving the latent optimization problem, we add stochastic noise by setting $\eta\beta_t = \sqrt{\bar{\alpha}_t}\sqrt{1-\bar{\alpha}_{t-1}}$. As other tasks, we set NFE to 200.

From the general formulation of DreamSampler, we can derive the latent optimization problem for the inpainting task. Suppose that the generator $g = \mathcal{D}_\varphi$ and $\phi = \boldsymbol{z}$ so the generator maps the latent vector to pixel space. Note that $\mathcal{D}_\varphi$ is a component of autoencoder that satisfies $\boldsymbol{z} = \mathcal{E}_\phi(\mathcal{D}_\varphi(\boldsymbol{z}))$ called perfect reconstruction constraint. Then, by defining the regularization function as

$$\mathcal{R}(\mathcal{D}_\varphi(\boldsymbol{z})) = \|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{tgt})\|^2 + \|\boldsymbol{y} - \mathcal{A}\mathcal{D}_\varphi(\boldsymbol{z})\|^2 \tag{10}$$

we can reach to the latent optimization problem of

$$\min_{\boldsymbol{z},\boldsymbol{x}} \gamma_1 \underbrace{\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{tgt})\|^2}_{\text{Score distillation}} + \gamma_2 \underbrace{\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{\phi})\|^2}_{\text{Proximal term}} + \gamma_3 \underbrace{\left[\|\boldsymbol{y} - \mathcal{A}\mathcal{D}_{\varphi}(\boldsymbol{z})\|^2 + \|\boldsymbol{z} - \mathcal{E}_{\phi}(\boldsymbol{x})\|^2\right]}_{\text{Data consistency}} \tag{11}$$

where the last term comes from the perfect reconstruction constraint, and $\gamma_1 + \gamma_2 + \gamma_3 = 1$. The objective function comprises three components: data consistency, which ensures alignment between the current estimate and the given measurement; score distillation, which serves as textual guidance; and a proximal term, which regulates the solution to remain close to the inverted trajectory. Following TReg [9], we solve the optimization problem sequentially. First, using the approximation $\boldsymbol{x} = \mathcal{D}_{\varphi}(\boldsymbol{z})$, the optimization problem with respect to $\boldsymbol{x}$ becomes

$$\min_{x} \|\boldsymbol{y} - \mathcal{A}\boldsymbol{x}\|^2 + \|\boldsymbol{z} - \mathcal{E}_{\phi}(\boldsymbol{x})\|^2 + \lambda\|\boldsymbol{x} - \mathcal{D}_{\phi}(\boldsymbol{z}) + \boldsymbol{\eta}\|^2 \tag{12}$$

where the dual variable $\boldsymbol{\eta}$ is set to a zero vector for simplicity. Then, by initializing $\boldsymbol{z} = \hat{\boldsymbol{z}}_{0|t}(c_{\phi})$, we have

$$\hat{\boldsymbol{x}}_0(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{x})\|^2 + \lambda\|\boldsymbol{x} - \mathcal{D}_{\varphi}(\hat{\boldsymbol{z}}_{0|t}(c_{\phi}))\|^2. \tag{13}$$

This could be solved by conjugate gradient (CG) method. Subsequently, using the approximation $\boldsymbol{z} = \mathcal{E}_{\phi}(\boldsymbol{x})$ with $\boldsymbol{\eta} = \boldsymbol{0}$, the optimization with respect to $\boldsymbol{z}$ becomes

$$\min_{z} \gamma_1\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{tgt})\|^2 + \gamma_2\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{\phi})\|^2 + \gamma_3\|\boldsymbol{z} - \mathcal{E}_{\phi}(\hat{\boldsymbol{x}}_0(\boldsymbol{y}))\|^2, \tag{14}$$

which leads to a closed-form solution

$$\bar{\boldsymbol{z}} = \bar{\alpha}_t\hat{\boldsymbol{z}}_{0|t}(c_{tgt}) + (1-\bar{\alpha}_t)^2\hat{\boldsymbol{z}}_{0|t}(c_{\phi}) + \bar{\alpha}_t(1-\bar{\alpha}_t)\hat{\boldsymbol{z}}_{0|t}(\boldsymbol{y}) \tag{15}$$

where $\hat{\boldsymbol{z}}_{0|t}(\boldsymbol{y}) := \mathcal{E}_{\phi}(\hat{\boldsymbol{x}}_0(\boldsymbol{y}))$. Specifically, we apply localized distillation gradient by leveraging a mask so the objective function inside/outside of masked region is differ as

$$\min_{z} \gamma_1\|\mathcal{M} \odot (\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{tgt}))\|^2 + \gamma_2\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{0|t}(c_{\phi})\|^2 + \gamma_3\|\boldsymbol{z} - \mathcal{E}_{\phi}(\hat{\boldsymbol{x}}_0(\boldsymbol{y}))\|^2 \tag{16}$$

where $\mathcal{M}$ denotes pixel-wise mask with ones inside the masked region and zeros elsewhere, while $\odot$ denotes element-wise multiplication. Hence, the closed form solution is described as

$$\bar{\boldsymbol{z}} = \begin{cases} \bar{\alpha}_t\hat{\boldsymbol{z}}_{0|t}(c_{tgt}) + (1-\bar{\alpha}_t)^2\hat{\boldsymbol{z}}_{0|t}(c_{\phi}) + \bar{\alpha}_t(1-\bar{\alpha}_t)\hat{\boldsymbol{z}}_{0|t}(\boldsymbol{y}) & \text{inside mask} \\ (1-\bar{\alpha}_t)\hat{\boldsymbol{z}}_{0|t}(c_{\phi}) + \bar{\alpha}_t\hat{\boldsymbol{z}}_{0|t}(\boldsymbol{y}) & \text{otherwise} \end{cases} \tag{17}$$

Also, we use additional DPS step to ensure the consistency between inside and outside of masked region. By following TReg, we compute the (17) when $\Gamma = \{t|t \bmod 3 = 0, t \leq 170\}$ and apply DPS gradie nt otherwise. In summary, the pseudocode of DreamSampler for the inpainting task is described as Algorithm 4

Algorithm 4 DreamSampler for Image Inpainting

Require: measurement $\boldsymbol{y}$, image encoder $\mathcal{E}_\phi$, latent diffusion model $\boldsymbol{\epsilon}_\theta$, null-text embedding $c_\emptyset$, conditioning text embedding c_{tgt}.

$\boldsymbol{z}_0 \leftarrow \mathcal{E}_\phi(\boldsymbol{y})$

$\boldsymbol{z}_T \leftarrow \text{Inversion}(\boldsymbol{z}_0)$

for $t \in [T, 0]$ **do**

$\hat{\boldsymbol{\epsilon}}_\theta(c_\emptyset), \hat{\boldsymbol{\epsilon}}_\theta(c_{tgt}) \leftarrow \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, c_\emptyset), \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, c_{tgt})$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$

$\tilde{\boldsymbol{\epsilon}}_t \leftarrow (\sqrt{1-\bar{\alpha}_{t-1}-\eta^2\beta_t^2}\hat{\boldsymbol{\epsilon}}_\theta(c_\emptyset) + \eta\beta_t\boldsymbol{\epsilon})/\sqrt{1-\bar{\alpha}_{t-1}}$

if $t \in \Gamma$ **then**

$\hat{\boldsymbol{z}}_{0|t}(c_\emptyset) \leftarrow (\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(c_\emptyset))/\sqrt{\bar{\alpha}_t}$

$\hat{\boldsymbol{z}}_{0|t}(c_{tgt}) \leftarrow (\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(c_{tgt}))/\sqrt{\bar{\alpha}_t}$

$\hat{\boldsymbol{x}}_0(\boldsymbol{y}) \leftarrow \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{x})\|^2 + \lambda\|\boldsymbol{x} - \mathcal{D}_\varphi(\hat{\boldsymbol{z}}_{0|t}(c_\emptyset))\|^2$

$\bar{\boldsymbol{z}}_{in} \leftarrow \bar{\alpha}_t\hat{\boldsymbol{z}}_{0|t}(c_{tgt}) + (1-\bar{\alpha}_t)^2\hat{\boldsymbol{z}}_{0|t}(c_\emptyset) + \bar{\alpha}_t(1-\bar{\alpha}_t)\mathcal{E}_\phi(\hat{\boldsymbol{x}}_0(\boldsymbol{y}))$

$\bar{\boldsymbol{z}}_{out} \leftarrow (1-\bar{\alpha}_t)\hat{\boldsymbol{z}}_{0|t}(c_\emptyset) + \bar{\alpha}_t\mathcal{E}_\phi(\hat{\boldsymbol{x}}_0(\boldsymbol{y}))$

$\bar{\boldsymbol{z}} \leftarrow \mathcal{M} \odot \bar{\boldsymbol{z}}_{in} + (1-\mathcal{M}) \odot \bar{\boldsymbol{z}}_{out}$

$\boldsymbol{z}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\bar{\boldsymbol{z}} + \sqrt{1-\bar{\alpha}_{t-1}}\tilde{\boldsymbol{\epsilon}}_t$

else

$\hat{\boldsymbol{z}}_{0|t}(c_\emptyset) = (\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(c_\emptyset))/\sqrt{\bar{\alpha}_t}$

$\boldsymbol{z}'_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{z}}_{0|t}(c_\emptyset) + \sqrt{1-\bar{\alpha}_{t-1}}\tilde{\boldsymbol{\epsilon}}_t$

$\boldsymbol{z}_{t-1} \leftarrow \boldsymbol{z}'_{t-1} - \rho_t\nabla_{\boldsymbol{z}_t}\|\mathcal{A}(\mathcal{D}_\varphi(\hat{\boldsymbol{z}}_{0|t})) - \boldsymbol{y}\|$

end if

end for

3 More qualitative comparison for text-guided inpainting

In this section, we show qualitative comparison of the DreamSampler for text-guided image inpainting tasks, against multiple state-of-the-art diffusion-based algorithms. To ensure the fair comparison, we leverage the StableDiffusion v1.5 checkpoint for all tasks. In parallel with the quantitative results in the main body, DreamSampler achieves better fidelity and data consistency than baselines, as shown in Figure S1 and S2.

3.1 More results for Text-guided Inpainting

Fig. S1: Qualitative comparison for inpainting task with 512x512 FFHQ dataset. Text prompt "A photography of face wearing glasses" is given.

Fig. S2: Qualitative comparison for inpainting task with 512x512 FFHQ dataset. Text prompt "A photography of face with smile" is given.

Fig. S3: Results for text-guided inpainting with 512x512 FFHQ dataset. Text prompt "wearing glasses" is given.

Fig. S4: Results for text-guided inpainting with 512x512 FFHQ dataset. Text prompt "smile" is given.

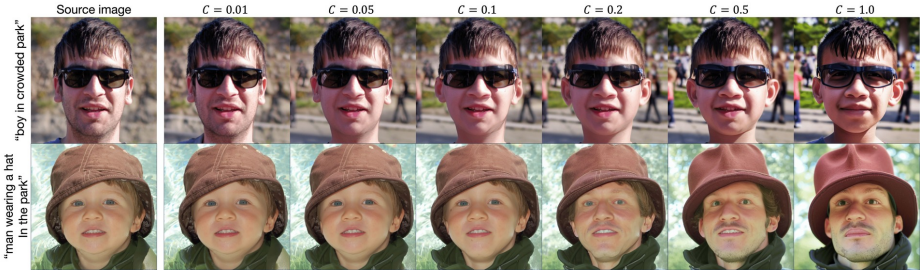

Fig. S5: Ablation study on effects of γ on real image editing task.

4 Ablation study

4.1 Effect of γ in Real Image Editing

From the optimization problem (9), γ could be interpreted as the interpolation weight between $\boldsymbol{z}_{0|t}(c_\phi)$ and $\boldsymbol{z}_{0|t}(c_{tgt})$, where c_ϕ is utilized for DDIM inversion to initialize $\boldsymbol{z}_T$. Thus, setting γ close to 1 steers the solution of the problem towards $\boldsymbol{z}_{0|t}(c_{tgt})$, while setting γ close to 0 directs the solution towards $\boldsymbol{z}_{0|t}(c_\phi)$. Considering the role of γ as a scale for the CFG, this interpretation aligns well. In this study, we use time-dependent $\gamma_t = C\bar{\alpha}_t$ where C denotes a constant. To demonstrate the effect of γ in a real image editing task, we generated multiple images by varying the value of C from 0.01 to 1.0. The results in Figure S5 demonstrate that when C is smaller, the generated image closely resembles the reconstruction, while when C is larger, the generated image adheres closely to the text guidance.

4.2 Effect of Distillation in Image Restoration through Vectorization

To examine the significance of text-guided distillation in the context of restoration, we adjusted the parameter $\lambda_{SDS} = 0$ in Equation (21) of the main manuscript. Figure S6 highlights that relying solely on data consistency regularization falls short of achieving a sufficient level of restoration, manifested through the emergence of scattering artifacts within the SVGs. This suggests that the absence of a text-guided distillation refinement process may contribute to severe degradation.

5 Additional Results

5.1 3D representation learning using degraded view

Since the proposed framework is defined in terms of an arbitrary generator g and generic parameter ψ, we apply Dreamsampler to the 3D NeRF [17] representation learning, specifically targeting scenarios with degraded views. We acknowledge that low-quality, noisy measurements can compromise the detail of 3D modeling,

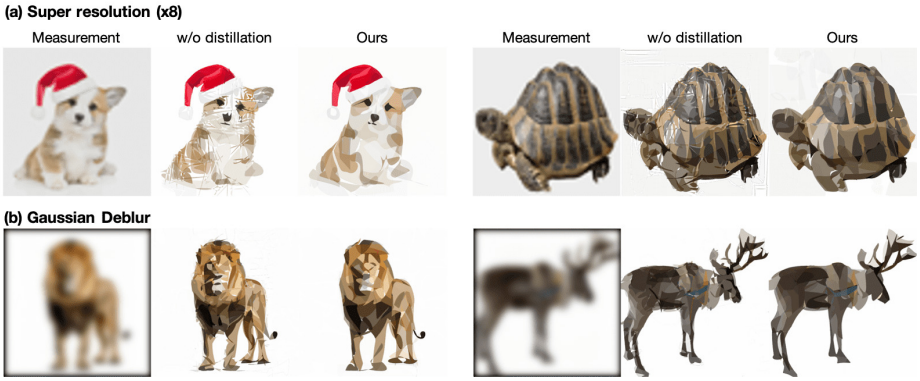

Fig. S6: Ablation study on effects of text-guided distillation on restoration task.

aligning with the motivation of vectorized image restoration task. In this context, the parameters ψ of the generator consist of NeRF MLP, which parameterizes volumetric density and albedo (color). Our goal is to accurately reconstruct NeRF parameters that, when rendered, closely align with the provided degraded view $\boldsymbol{y}$ and text conditions $c_{\boldsymbol{y}}$.

For this NeRF inverse problem, we consider Gaussian blur operator $\mathcal{A}$ with 5×5 kernel and standard deviation of 10 following [12]. For a single input image, we introduce novel 16 degraded view images with SyncDreamer [13] and blurring operator $\mathcal{A}$. Then, we first pre-train NeRF with the data consistency regularization $\ell(\boldsymbol{y}, \mathcal{A}g(\psi))$ for warm-up, where $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ represents a generic loss function. To facilitate effective representation learning, we integrate ℓ_2-loss and perceptual similarity loss such as LPIPS [28] for data consistency. Then the latent optimization framework of the NeRF inverse problem is defined as follows:

$$\min_{\psi}(1-\gamma)\lambda_{SDS}\|\mathcal{E}_\phi(g(\psi)) - \hat{\boldsymbol{z}}_{0|t}(c_{\boldsymbol{y}})\|^2 + \gamma\lambda_{DC}\|\boldsymbol{y} - \mathcal{A}g(\psi)\|^2, \tag{18}$$

which is analogous to the SVG inverse problem. We set $\gamma = \bar{\alpha}_t$ as SVG experiments. Following [29], we additionally adapt pixel-space distillation loss to enhance supervision for high-resolution images such as

$$\|g(\psi) - \mathcal{D}_\varphi(\hat{\boldsymbol{z}}_{0|t}(c_{\boldsymbol{y}}))\|^2, \tag{19}$$

where $\mathcal{D}_\varphi(\hat{\boldsymbol{z}}_{0|t}(c_{\boldsymbol{y}}))$ represents a recovered image via the decoder $\mathcal{D}_\varphi$. We further employ z-coordinates regularization and kernel smoothing techniques to improve sampling in high-density areas and address texture flickering issues.

The findings, as illustrated in S7, reveal that Dreamsampler successfully achieves high-quality 3D representation despite the compromised quality of the degraded views. Thus, Dreamsampler demonstrates its efficacy in learning 3D representations by directly leveraging degraded views and text conditions, effectively facilitating both distillation and reconstruction processes.

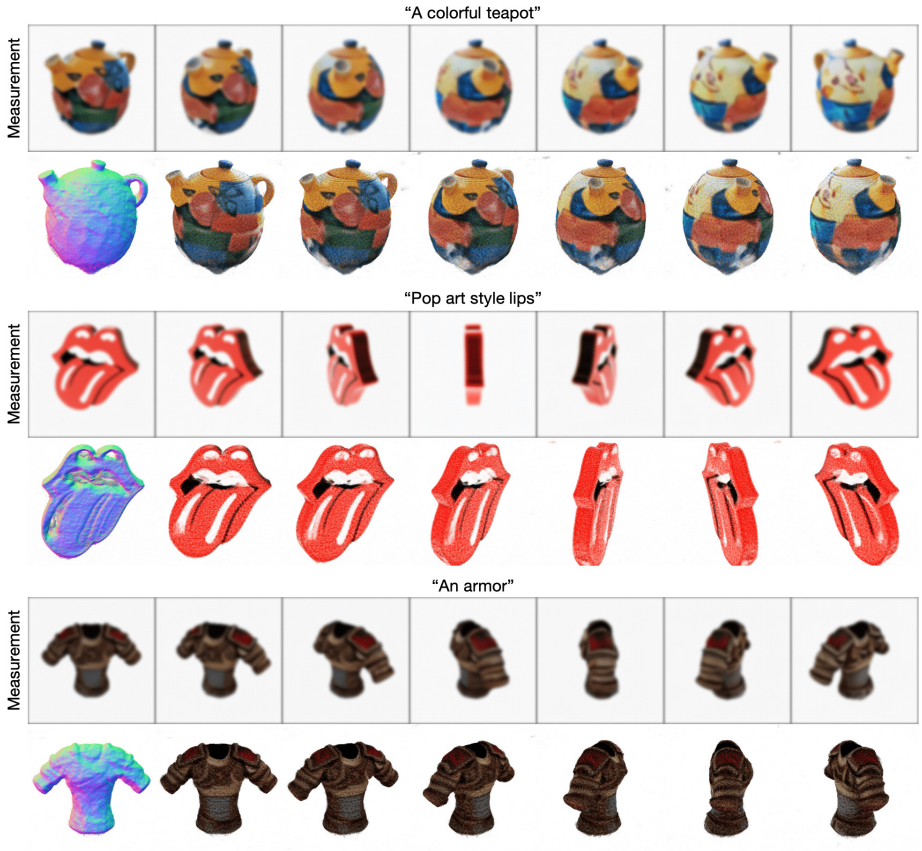

Fig. S7: Novel view synthesis via DreamSampler. Blurry views are given.

5.2 More results for Real Image Editing

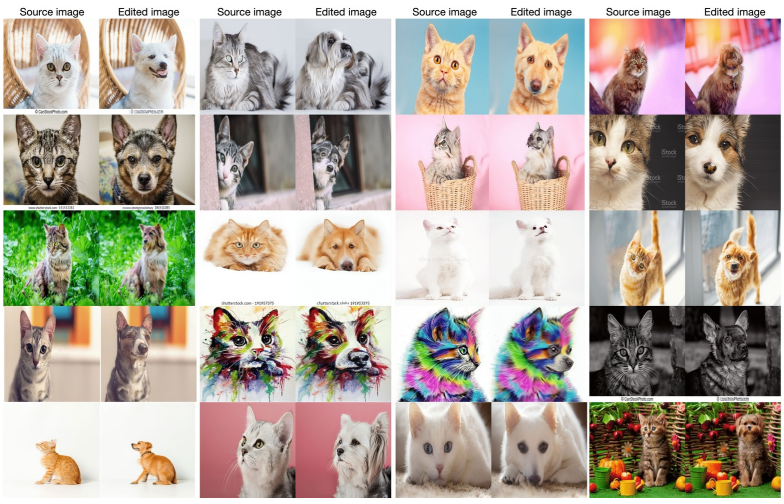

Fig. S8: Results for real image editing. Cat → Dog. Best views are displayed.

Fig. S9: Results for real image editing. Horse → Zebra. Best views are displayed.

5.3 Text to 3D representation learning with DreamSampler

We demonstrate the generation ability of Dreamsampler with Text-to-3D generation task. We set $\mathcal{R}(g(\psi)) = \|g(\psi) - \mathcal{D}(\boldsymbol{z}_{0|t}(c))\|^2$ as our regularization to improve pixel-level details in high-resolution images as similar to [29]. We use Adam optimizer with $lr = 10^{-3}$ for NeRF weights, optimized for 10^4 iterations. Figiure S11 shows that Dreamsampler significantly improves the Text-to-NeRF performance with key differences such as decreasing time-step schedule and pixel-domain regularizer $\mathcal{R}(g(\psi))$.

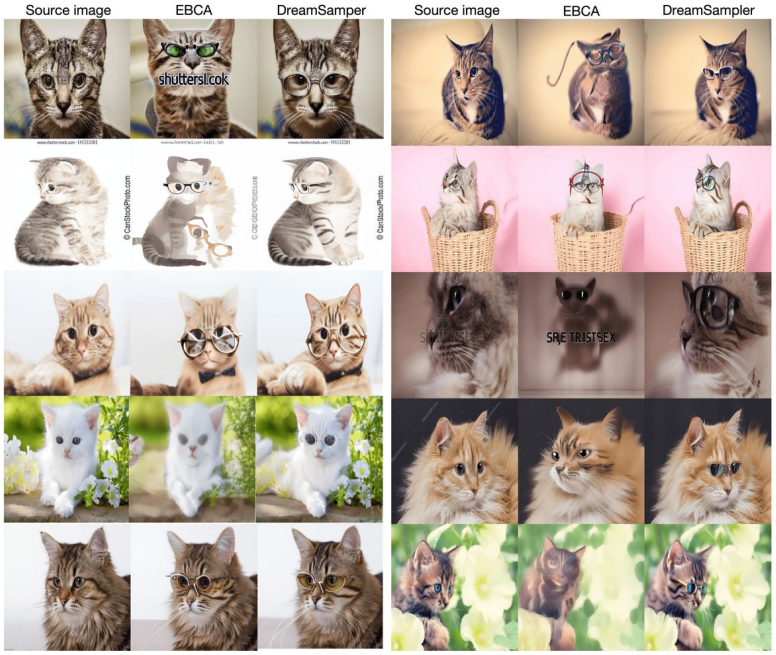

Fig. S10: Results for real image editing. Cat → Cat with Glasses. Best views are displayed.

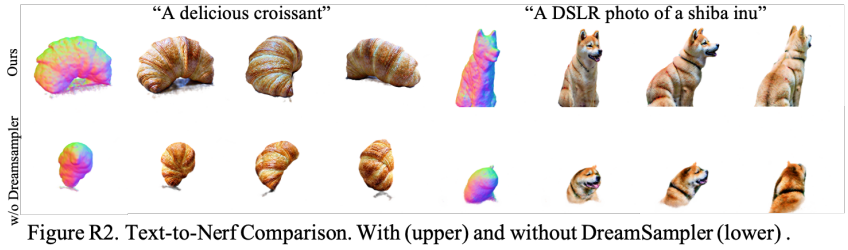

Figure R2. Text-to-Nerf Comparison. With (upper) and without DreamSampler (lower) .

Fig. S11: Text-to-NeRF Comparison. With(upper) and without DreamSampler(lower).