Identity-preserving Distillation Sampling by Fixed-Point Iterator

SeonHwa Kim[1] Jiwon Kim[1] Soobin Park[2] Donghoon Ahn[1] Jiwon Kang[1]
Seungryong Kim[3*] Kyong Hwan Jin[1*] Eunju Cha[2*]
[1]Korea University [2]Sookmyung Women's University [3]KAIST
{sunkim0062, jwonkim, dhahn99, jiwon7258, kyong_jin}@korea.ac.kr,
{psb1219j, eunju.cha}@sookmyung.ac.kr, seungryong.kim@kaist.ac.kr

Abstract

Score distillation sampling (SDS) demonstrates a powerful capability for text-conditioned 2D image and 3D object generation by distilling the knowledge from learned score functions. However, SDS often suffers from blurriness caused by noisy gradients. When SDS meets the image editing, such degradations can be reduced by adjusting bias shifts using reference pairs, but the de-biasing techniques are still corrupted by erroneous gradients. To this end, we introduce Identity-preserving Distillation Sampling (IDS), which compensates for the gradient leading to undesired changes in the results. Based on the analysis that these errors come from the text-conditioned scores, a new regularization technique, called fixed-point iterative regularization (FPR), is proposed to modify the score itself, driving the preservation of the identity even including poses and structures. Thanks to a self-correction by FPR, the proposed method provides clear and unambiguous representations corresponding to the given prompts in image-to-image editing and editable neural radiance field (NeRF). The structural consistency between the source and the edited data is obviously maintained compared to other state-of-the-art methods. Our code is https://github.com/shhh0620/IDS

1. Introduction

Diffusion models [4, 8, 9, 22, 25] have shown powerful representations on text-to-image (T2I) generative tasks. With the advance of classifier guidance (CG) and classifier-free guidance (CFG) paradigms [1, 4, 8, 10], diffusion models improve the quality of generated samples [9, 25]. Such high-quality image generators can be easily extended to image editing by simply modifying forward/reverse iterations [14], applying CFG with a target prompt [2, 7] or interchanging attention layers [26].

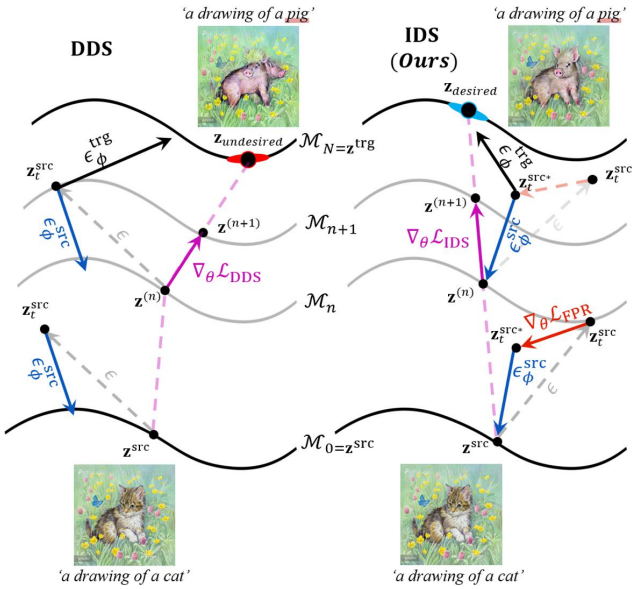


Figure 1. **Trace of guided updating** from source to target images using delta denoising score (DDS) and identity-preserving distillation sampling (IDS). DDS moves a gradient of score function toward $\mathcal{M}_{\mathbf{z}}$ manifold directed by stochastic direction ϵ. In contrast, IDS moves a gradient with a corrected direction by a fixed-point regularization.

Recently, Delta Denoising Score (DDS) [6] is proposed to edit a source image by distilling the rich generative prior of T2I diffusion models. It is based on the analysis of Score Distillation Sampling (SDS) [20], originally developed to optimize a parametric generator such as Neural Radiance Fields (NeRFs) [16] by exploiting the learned score of the diffusion models. Even though SDS offers remarkable performance in synthesizing 3D scenes, noisy gradients from stochastic perturbations lead to significantly over-saturated results that faithfully follow the given text prompts. In the context of image editing, text prompts do not often include information about the identity of the source image, such as the background, the object's pose, or the structure of the

*Corresponding author.

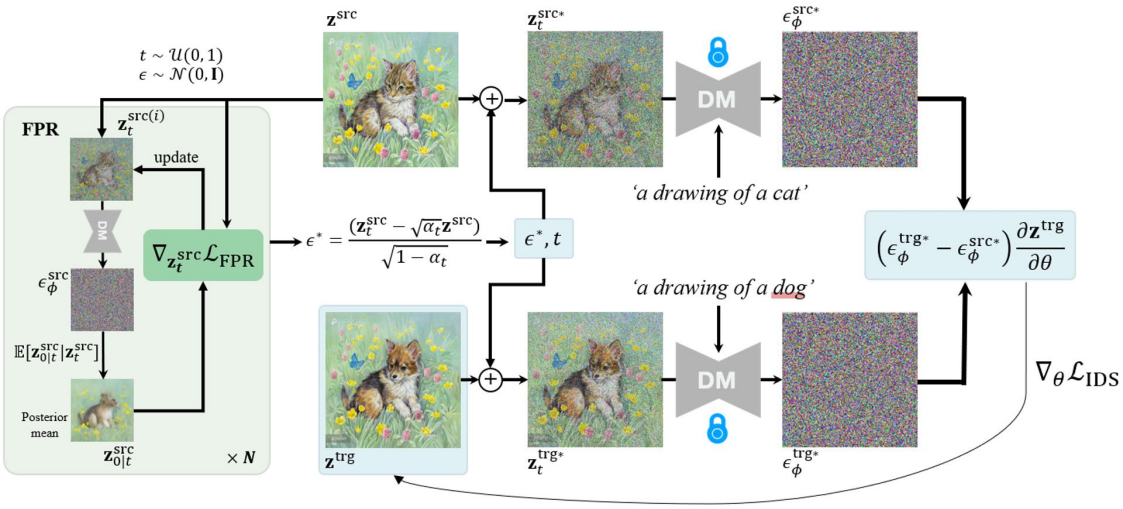


Figure 2. **Flowchart of IDS.** The backbone of our algorithm employs DDS [6] framework to distill score function into a target image. Our fixed-point regularization (FPR) obtains a guided noise, ϵ^*, from iterative updates using posterior mean computed by Tweedie's formula. When distilling the score function to a target image, the guided noise is updated while maintaining the identity of the source.

content, which should be retained during updates. Thus, DDS is designed to resolve such blurriness by erasing gradients of non-text-aligned features from SDS gradients. There is no explicit procedure to preserve the source's identity in DDS updates because the fine gradient may provide the conserved identity. However, this cannot be guaranteed if many variations in the structure are possible, such as editing the image of a cat into a pig, as shown in Fig. 1. To alleviate this problem, Contrastive Denoising Score (CDS) [17] and Posterior Distillation Sampling (PDS) [12] are introduced to maximize the mutual information of the source image and edited image. Although such algorithms rely heavily on text prompts, the algorithms have yet to analyze the inherent error caused by text-conditioned scores.

To this end, we investigate the underlying meaning of text-conditioned score. The gradient maps the stochastic latent, generated by applying the forward diffusion process to the given image, to one of the possible images described by the prompt, including the original image. Simply, the score obtained from the latent of the source image ('source latent') and the source prompt can be a gradient to another image represented by the identical text. Based on this interpretation, the accumulation of misaligned directions causes the loss of the source's identity, leading to structural changes in the result with DDS, as shown in Fig. 1.

To address this issue, we propose a novel score distillation sampling to effectively preserve the identity of the original image by self-correcting the misaligned gradients, called **I**dentity-preserving **D**istillation **S**ampling (**IDS**). The key insight is that if the score is precisely adjusted to the source image, the conditional expectation of the source image given the source latent contains meaningful information that should be preserved during the editing. This conditional expectation corresponds to the posterior mean computed by Tweedie's formula using the learned score [3, 5]. The source latent is iteratively updated to make the posterior mean similar to the source image. This procedure, named a fixed-point iterative regularization (FPR), results in the aligned score with the source that provides reliable gradients for editing, as illustrated in Fig. 1. Following IDS update is performed using guided noise extracted from the refined source latent, rather than random Gaussian noise. This further ensures the identity preservation. Our method demonstrated superior results compared with baselines in two tasks: editing images by prompts and editing NeRF.

In summary, our main contributions are as follows:

- We obtain reliable gradients for the score distillation function by a fixed-point iterator with respect to posterior means. The iterator corrects the text-conditioned score, guiding SDS gradients toward reliable pre-trained manifolds.
- Our fixed-point regularization preserves the identities of sources such as structures and poses in edited targets for 2D and 3D editing. Such preservation is well demonstrated in qualitative and quantitative results.

2. Related works

2.1. Image Editing with Diffusion Models

With the great success of image generation using diffusion models, the pre-trained diffusion models have been recently employed for image editing tasks, demonstrating significant advancements in the quality and flexibility of generated edits [2, 7, 12, 14, 17, 26]. Stochastic Differential

Editing (SDEdit) [14] is a pioneering work in which the source image was modified by adding noise and solving reverse stochastic differential equations. Thanks to the text-conditional Latent Diffusion Model (LDM), a.k.a. Stable Diffusion [22], text-driven editing approaches have been introduced. Specifically, the text embedding was injected through the cross-attention layer of the model for image editing and translation, while retaining the structure of the original image [7, 26]. The editing was further controlled by rescaling the attention of the specific word [7] or by manipulating the self-attention features [26]. These approaches provide greater control by balancing fidelity between the edited prompt and the source image without the need for model training, fine-tuning, additional data, or optimization. However, the current DDIM-based inversion [24] can lead to unsatisfactory reconstructions for real images, and the cross-attention bottleneck limits its effectiveness for broader edits. Crafting suitable prompts also remains challenging for complex compositions.

2.2. Score distillation sampling

Score Distillation Sampling (SDS) [20] enables text-driven 3D synthesis by leveraging probability density distillation loss to distill knowledge from 2D diffusion models, allowing high-quality 3D scene generation based on textual prompts without 3D training data. However, SDS has limitations, often producing oversaturated and overly smooth 3D models, and lacking diversity across initializations.To address these limitations of SDS, various models have been proposed based on exploiting multi-step denoising [30], a variation approach [27], negative conditioning [11], and ordinary differential equation trajectory [28]. To mitigate the limitation of noisy gradients in SDS, which hampers precise image editing, DDS [6] was introduced. By computing the delta between the derived gradient and the target pair, DDS effectively isolates and removes unwanted noise in the gradient direction. Despite these advancements, DDS still faces challenges in preserving the complete structural consistency of the source image's identity.

3. Preliminaries

3.1. Diffusion Model and Sampling Guidance

Text-to-image diffusion models $\epsilon_\phi(\cdot)$ are based on diffusion probabilistic models (DPMs) [9, 22, 25]. The models are trained to estimate the denoising score when the original image $\mathbf{z}_0$ and the text condition y are given:

$$\mathcal{L}(\phi) = \mathbb{E}_{t,\epsilon}[\|\epsilon_\phi(\mathbf{z}_t, y, t) - \epsilon\|_2^2],$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $t \sim \mathcal{U}(0, 1)$. $\mathbf{z}_t$ refers to the stochastic latent of $\mathbf{z}_0$ via the forward diffusion process as follows:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon, \tag{1}$$

where α_t is noise schedule. With the trained $\epsilon_\phi(\cdot)$, high-quality samples can be generated using the classifier-free guidance (CFG) [8] by subtracting unconditioned denoising score from the conditioned score with guidance scale ω:

$$\epsilon_\phi^\omega(\mathbf{z}_t, y, t) = (1+\omega)\epsilon_\phi(\mathbf{z}_t, y, t) - \omega\epsilon_\phi(\mathbf{z}_t, \varnothing, t). \tag{2}$$

3.2. Score Distillation Sampling (SDS)

With pretrained text-to-image diffusion models $\epsilon_\phi(\cdot)$, SDS [20] synthesizes 3D data $\mathbf{z}$ for a given text prompt y by optimizing the differentiable rendering function paremetrized by θ, where $\mathbf{z} = g(\theta)$:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\mathbf{z}, y) = \mathbb{E}_{t,\epsilon}\left[\omega(t)(\epsilon_\phi^\omega(\mathbf{z}_t, y, t) - \epsilon)\frac{\partial \mathbf{z}}{\partial \theta}\right]. \tag{3}$$

The optimized parameters θ^* provide the text-conditioned 3D volume that follows the diffusion prior [20]. However, a single text prompt y can refer to many different 3D volumes, each with diverse backgrounds or structural details of the object. Therefore, an inherent limitation of SDS [20] is that the score conditioned by the prompt y does not always provide the diffusion prior to the identical object during the optimization process, leading to blurry and unclear results.

3.3. Delta Denoising Score (DDS)

DDS [6] is proposed to synthesize the image $\mathbf{z}^{\text{trg}}$ from the given source image $\mathbf{z}^{\text{src}}$ and its corresponding prompt y^{src}, which is aligned to the target prompt y^{trg}. Based on the insight that the gradient should be zero if y^{trg} matches y^{src}, DDS minimizes the identity change of $\mathbf{z}^{\text{src}}$ by simple replacing ϵ in (3) with the score $\epsilon_\phi^\omega(\mathbf{z}_t^{\text{src}}, y^{\text{src}}, t)]$ as follows:

$$\nabla_\theta \mathcal{L}_{\text{DDS}} = \mathbb{E}_{t,\epsilon}\left[(\epsilon_\phi^\omega(\mathbf{z}_t^{\text{trg}}, y^{\text{trg}}, t) - \epsilon_\phi^\omega(\mathbf{z}_t^{\text{src}}, y^{\text{src}}, t))\frac{\partial \mathbf{z}^{\text{trg}}}{\partial \theta}\right]. \tag{4}$$

For simplicity, we denote $\epsilon_\phi^{\text{trg}} = \epsilon_\phi^\omega(\mathbf{z}^{\text{trg}}, y^{\text{trg}}, t)$ and $\epsilon_\phi^{\text{src}} = \epsilon_\phi^\omega(\mathbf{z}^{\text{src}}, y^{\text{src}}, t)$. Here, $\epsilon_\phi^{\text{trg}}$ and $\epsilon_\phi^{\text{src}}$ can be interpreted as the gradients representing the direction from $\mathbf{z}_t^{\text{trg}}$ to $\mathbf{z}^{\text{trg}}$ and the direction from $\mathbf{z}_t^{\text{src}}$ to $\mathbf{z}^{\text{src}}$, respectively. $\theta = \mathbf{z}^{\text{trg}}$ is thus gradually optimized along the direction from $\mathbf{z}^{\text{src}}$ to $\mathbf{z}^{\text{trg}}$, as shown in Fig. 1. It is worth noting that the guidance of the update can be calculated at the same point $\mathbf{z}_t^{\text{trg}}$, thanks to the shared ϵ. However, the slight error in the gradient caused by the score $\epsilon_\phi^{\text{src}}$ still leads to the incorrect direction for the optimization.

3.4. Fixed-point Iteration

In numerical analysis, a fixed-point iteration [18] is an iterative method to find fixed points of a function f, where $f(x) = x$. Given an initial point x_0, the iteration is defined as:

$$x_{n+1} = f(x_n), \quad n = 0, 1, 2, \ldots$$

Under appropriate conditions, this sequence converges to a fixed point. Thanks to its applicability to non-linear problems with low computational costs, fixed-point iteration is widely used in optimization, including applications in the context of diffusion models [13].

4. Method

Given the source pair $\{\mathbf{z}^{\text{src}}, y^{\text{src}}\}$, the aim of our work is to provide an edited result $\mathbf{z}^{\text{trg}}$ that is aligned with y^{trg} while maintaining the source's identity. To this end, we introduce a novel approach called **Identity-preserving Distillation Sampling (IDS)**, which (1) corrects the error of the gradient aligned with the text prompt by the fixed-point iterator and (2) provides the result $\mathbf{z}^{\text{trg}}$ using the guided noise.

4.1. Motivation

Analysis of the text-conditioned score. We first investigated how much identity of the given image $\mathbf{z}^{\text{src}}$ could be contained in the text-conditioned score $\epsilon_\phi^{\text{src}}$. To do this, we conducted the experiment to compare the original image $\mathbf{z}^{\text{src}}$ and the posterior mean $\mathbf{z}_{0|t}^{\text{src}} = \mathbb{E}[\mathbf{z}^{\text{src}}|\mathbf{z}_t^{\text{src}}]$, which is given by:

$$\mathbf{z}_{0|t}^{\text{src}} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t^{\text{src}} - \sqrt{1-\alpha_t}\epsilon_\phi^{\text{src}}\right), \tag{5}$$

where $\mathbf{z}_t^{\text{src}}$ denotes the source latent generated by (1). As shown in the first row of the supplementary Fig. S1, it is difficult to recognize the features of $\mathbf{z}^{\text{src}}$ in $\mathbf{z}_{0|t}^{\text{src}}$, such as hairstyle, details of eyes, and background. This demonstrates that the score $\epsilon_\phi^{\text{src}}$ is not exactly adjusted to the given image $\mathbf{z}^{\text{src}}$. This deformation becomes more pronounced with increasing t. The experiment confirms that $\epsilon_\phi^{\text{src}}$ may not be a precise guidance to the source image $\mathbf{z}^{\text{src}}$. Therefore, the text-conditioned score $\epsilon_\phi^{\text{src}}$ needs to be modified to maintain the identity of the source image $\mathbf{z}^{\text{src}}$ in the edited result $\mathbf{z}^{\text{trg}}$.

Accumulated error in DDS. The transformed image $\mathbf{z}^{\text{trg}}$ can be converted back to the original image $\mathbf{z}^{\text{src}}$ by reversing the set of ϵ used to synthesize $\mathbf{z}^{\text{trg}}$ from $\mathbf{z}^{\text{src}}$ and swapping $\{\mathbf{z}^{\text{src}}, y^{\text{src}}\}$ and $\{\mathbf{z}^{\text{trg}}, y^{\text{trg}}\}$ to calculate the DDS loss in (4). If the guidance from $\mathbf{z}^{\text{src}}$ to $\mathbf{z}^{\text{trg}}$ is computed exactly, the perfect reconstruction can be achieved. Nevertheless, as can be seen from the second row in Fig. 3, DDS [6] fails to restore the original image $\mathbf{z}^{\text{src}}$ from the edited image $\mathbf{z}^{\text{trg}}$, which implies that the direction from $\mathbf{z}^{\text{src}}$ to $\mathbf{z}^{\text{trg}}$ is calculated incorrectly. Based on our analysis, this error is because the text-conditioned score $\epsilon_\phi^{\text{src}}$ do not refer to the source $\mathbf{z}^{\text{src}}$, which can be explictly expressed as the difference between the injected noise ϵ and the score $\epsilon_\phi^{\text{src}}$. While the optimization is being processed, the error inevitably accumulates, leading to the undesirable change to the structure and the pose. To address these issues, we investigated whether the guidance from $\mathbf{z}^{\text{src}}$ to $\mathbf{z}^{\text{trg}}$ can be properly provided while preserving the source's identity, when the timestep t is constrained by $t \sim \mathcal{U}(0, 0.2)$. This is because the posterior mean $\mathbf{z}_{0|t}^{\text{src}}$ and the source image $\mathbf{z}^{\text{src}}$ are similar for small timestep t, as illustrated in the first row of supplementary Fig. S1. However, as depicted in the first row of Fig. 3, DDS yields unrealistic result with this setting, whereby the structure of the given image $\mathbf{z}^{\text{src}}$ is overemphasized. This implies that it is not sufficient to simply limit the timestep t to prevent the score from deviating too far from $\mathbf{z}^{\text{src}}$ to correct the misalignment of the score to $\mathbf{z}^{\text{src}}$. Hence, we propose a fundamental approach to refine the gradient to achieve identity consistency without unwanted overemphasis on details.

Figure 3. **Accumulated error in DDS.** $\mathbf{z}^{\text{trg}}$ is edited image of source image $\mathbf{z}^{\text{src}}$ by prompt $y^{\text{src}} \rightarrow y^{\text{trg}}$. $\mathbf{z}^{\text{src}\dagger}$ is the inverted image of $\mathbf{z}^{\text{trg}}$ by prompt $y^{\text{trg}} \rightarrow y^{\text{src}}$. *(First row)* Inversion result of DDS with timestep $t \sim \mathcal{U}(0, 0.2)$. *(Second row)* Inversion result of DDS with $t \sim \mathcal{U}(0, 1)$. *(Third row)* Inversion result of ours.

4.2. Identity-preserving Distillation Sampling (IDS)

Fixed-point Regularization (FPR). Here, we introduce a **F**ixed-**p**oint **R**egularization (FPR) method that adjusts the text-conditioned score $\epsilon_\phi^{\text{src}}$ to the source image $\mathbf{z}^{\text{src}}$. Our key premise is that if the score $\epsilon_\phi^{\text{src}}$ is rightly estimated as a gradient to $\mathbf{z}^{\text{src}}$, the posterior mean $\mathbf{z}_{0|t}^{\text{src}}$ also contains sufficient information about $\mathbf{z}^{\text{src}}$. Therefore, FPR loss is designed to minimize the difference between $\mathbf{z}^{\text{src}}$ and $\mathbf{z}_{0|t}^{\text{src}}$ as follows:

$$\mathcal{L}_{\text{FPR}} = d(\mathbf{z}^{\text{src}}, \mathbf{z}_{0|t}^{\text{src}}), \tag{6}$$

where $d(\mathbf{x}_1, \mathbf{x}_2)$ can be any metric to compare $\mathbf{x}_1$ and $\mathbf{x}_2$. Here, we employed the Euclidean loss, and further investigations using various metrics are provided in Supplementary Materials.

The score $\epsilon_\phi^{\text{src}}$ needs to be modified to minimize the FPR loss before obtaining the updated direction. There are two ways to control the score $\epsilon_\phi^{\text{src}}$ by altering the injection noise ϵ or the source latent $\mathbf{z}_t^{\text{src}}$. As illustrated in supplementary Fig. S1, the proposed FPR revises the score $\epsilon_\phi^{\text{src}}$ to serve the source's identity for both approaches. Note that the score incorporates the content details, with the updates being performed with respect to the source latent $\mathbf{z}_t^{\text{src}}$ compared to the noise ϵ. Thus, $\mathbf{z}_t^{\text{src}}$ is updated to minimize the FPR loss as follows:

$$\mathbf{z}_t^{\text{src}} \leftarrow \mathbf{z}_t^{\text{src}} - \lambda \nabla_{\mathbf{z}_t^{\text{src}}} \mathcal{L}_{\text{FPR}}, \tag{7}$$

where λ and N denote a regularization scale and the number of iterations, respectively.

Algorithm 1 Fixed-point Regularization (FPR)

Require: $\mathbf{z}^{\text{src}}$, y^{src}, ϵ_ϕ, ω, λ, N

$\epsilon \sim \mathcal{N}(0, \mathbf{I})$
$t \sim \mathcal{U}(0, 1)$
$\mathbf{z}_t^{\text{src}} \leftarrow \sqrt{\alpha_t}\mathbf{z}^{\text{src}} + \sqrt{1-\alpha_t}\epsilon$
for i = 1, . . . , N **do**
  $\epsilon_\phi^{\text{src}} \leftarrow (1+\omega)\epsilon_\phi(\mathbf{z}_t^{\text{src}}, y^{\text{src}}, t) - \omega\epsilon_\phi(\mathbf{z}_t^{\text{src}}, \varnothing, t)$
  $\mathbf{z}_{0|t}^{\text{src}} \leftarrow \frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t^{\text{src}} - \sqrt{1-\alpha_t}\epsilon_\phi^{\text{src}})$
  $\mathcal{L}_{\text{FPR}} \leftarrow d(\mathbf{z}_{0|t}^{\text{src}}, \mathbf{z}^{\text{src}})$
  $\mathbf{z}_t^{\text{src}} \leftarrow \mathbf{z}_t^{\text{src}} - \lambda \nabla_{\mathbf{z}_t^{\text{src}}} \mathcal{L}_{\text{FPR}}$
end for
$\epsilon^* \leftarrow \frac{1}{\sqrt{1-\alpha_t}}(\mathbf{z}_t^{\text{src}} - \sqrt{\alpha_t}\mathbf{z}^{\text{src}})$
return ϵ^*

Editing with guided noise. Thanks to the proposed FPR, the optimized source latent $\mathbf{z}_t^{\text{src}*}$ containing the source's identity can be obtained. Then, the guided noise ϵ^* is extracted as follows:

$$\epsilon^* = \frac{1}{\sqrt{1-\alpha_t}}(\mathbf{z}_t^{\text{src}*} - \sqrt{\alpha_t}\mathbf{z}^{\text{src}}). \tag{8}$$

ϵ^* is utilized to produce the stochastic latent $\mathbf{z}_t^{\text{trg}*}$ by applying the forward diffusion process to the target image $\mathbf{z}^{\text{trg}}$. With $\mathbf{z}_t^{\text{src}*}$ and $\mathbf{z}_t^{\text{trg}*}$, the updated direction is given by:

$$\nabla_\theta \mathcal{L}_{\text{IDS}} = \mathbb{E}_{t,\epsilon}\left[(\epsilon_\phi^\omega(\mathbf{z}_t^{\text{trg}*}, y^{\text{trg}}, t) - \epsilon_\phi^\omega(\mathbf{z}_t^{\text{src}*}, y^{\text{src}}, t))\frac{\partial \mathbf{z}^{\text{trg}}}{\partial \theta}\right]. \tag{9}$$

It is worth noting that ϵ^* guides the appropriate gradients for editing while conserving the source's identity. In contrast to DDS, the proposed IDS perfectly reconstructs the source from the edited result $\mathbf{z}^{\text{trg}}$, as shown in the third row of Fig. 3. This confirms that the correct score and the corresponding injection noise can preserve the identity without further consideration of mutual information. The flowchart of our IDS is illustrated in Fig. 2.

5. Results

We evaluate our method through editing experiments conducted on two experiments. In Sec. 5.1, we perform a comparison on image-to-image editing across several datasets. In Sec. 5.2, we extend our evaluation to editable Neural Radiance Fields (NeRF) [16].

5.1. Text-guided image editing

Baselines. To evaluate our method, we conduct comparative experiments against four state-of-the-art image editing models: Prompt-to-Prompt (P2P) [7], Plug-and-Play (PNP) [26], DDS [6], and CDS [17]. The implementations of the baselines are carried out by referencing the official source code for each method. More details are provided in Supplementary Materials.

Qualitative Results. We present the qualitative results comparing our method with the baselines in Fig. 4. Prompt-to-Prompt (P2P) [7] performs image editing after applying DDIM inversion [4, 24] to the source image, leading to disregarding the structural components of the source image and following the target prompt excessively. Plug-and-Play (PnP) [26] has limitations in object recognition, as seen in the fourth row of Fig. 4. The third row of Fig. 4 demonstrates that DDS [6] and CDS [17] exhibited limitations, particularly in preserving the structural characteristics of the source image. In contrast, our method successfully edits the image while preserving the structural integrity of the source image.

Quantitative Results. To measure the identity-preserving performance, we utilize two datasets. First, we collect 250 cat images from the LAION 5B dataset [23] based on [17] for *Cat-to-Others* task and measure Intersection over Union (IoU). Second, we gather 28 images from the InstructPix2Pix (IP2P) dataset [2], which contains the pairs of source and target images and corresponding prompts and calculate the background Peak-Signal-to-Noise-Ratio (PSNR). Details of the metrics are provided in Supplementary Materials. In addition, we use the LPIPS score [29] for each experiment to quantify the similarity between source and target images. The results are presented in Tab. 1. Our method consistently achieves the lowest LPIPS score across all datasets, indicating that it best preserves the structural semantics of the source images.

For user evaluation, we present 35 comparison sets for four baselines and our method, gathering responses from 47 participants. Participants are asked to choose the most appropriate image for the following three questions: 1. *Which*

	cat2pig		cat2squirrel		Ip2p	
Metric	IoU (↑)	LPIPS (↓)	IoU (↑)	LPIPS (↓)	PSNR (↑)	LPIPS (↓)
P2P [7]	0.58	0.42	0.52	0.46	20.88	0.47
PnP [26]	0.55	0.52	0.53	0.52	23.81	0.39
DDS [6]	0.69	0.28	0.65	0.30	26.02	0.24
CDS [17]	0.72	0.25	**0.71**	0.26	27.35	0.21
IDS (Ours)	**0.74**	**0.22**	**0.71**	**0.24**	**29.25**	**0.19**

Table 1. **Quantitative results** for image editing. LPIPS [29] and IoU was measured on LAION 5B [23], while LPIPS and background PSNR was measured on InstructPix2Pix [2].

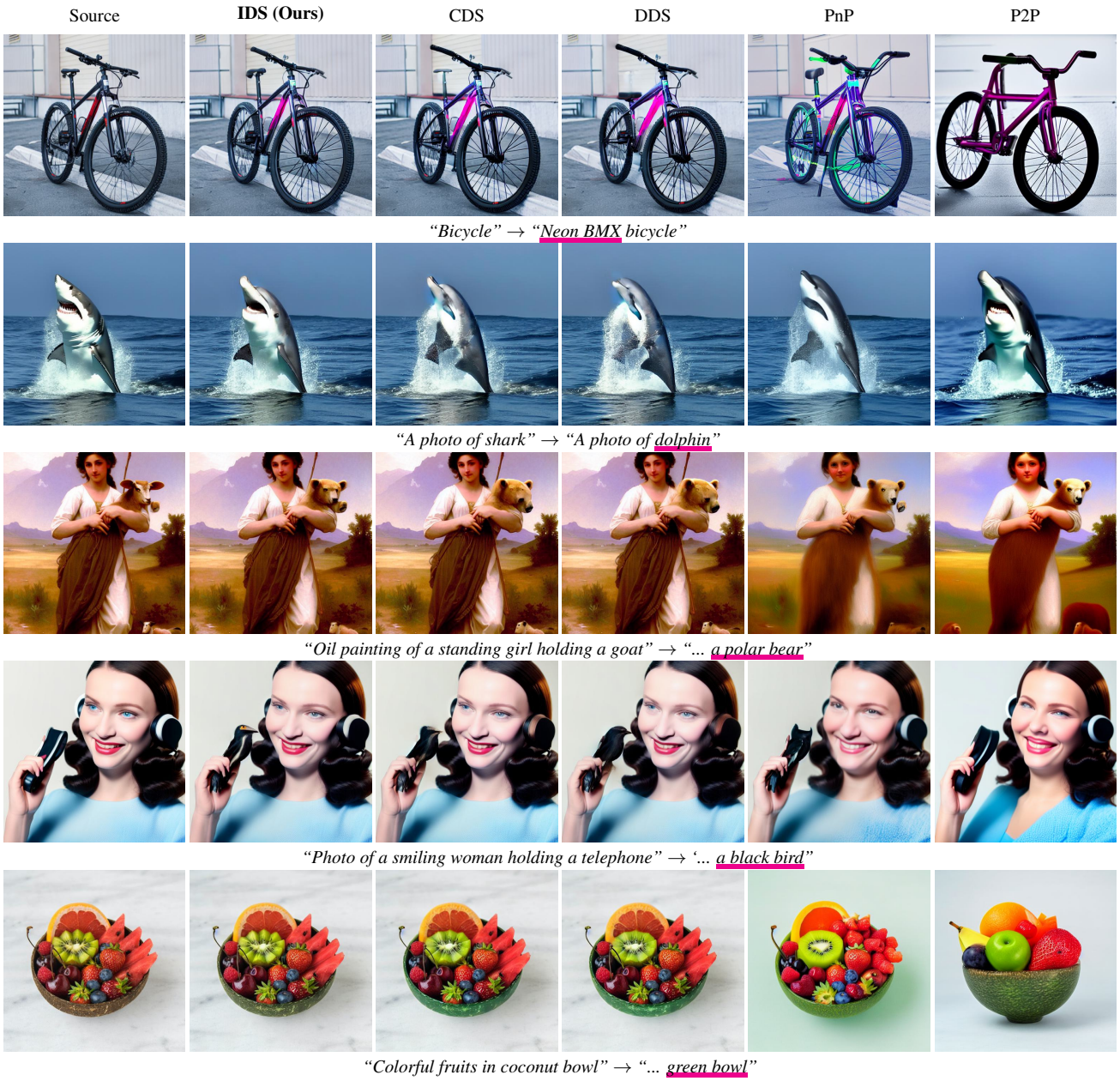


Figure 4. **Qualitative results** of InstructPix2Pix dataset [2]. Our method successfully edits the image aligning with the target text prompt while preserving the structural integrity of the source image.

image best fits the text condition? 2. *Which image best preserves the structural information of the original image?* 3. *Which image has the best quality for text-based image editing?* Additionally, we measure the GPT score using the Dreambench++ [19] method, which generates human-aligned assessments for the same questions by refining the scoring into ten distinct levels. As shown in Tab. 2, our method receives the highest ratings for all questions.

	User Preference Rate (%)			GPT score [19]		
Metric	Text (↑)	Preserving (↑)	Quality (↑)	Text (↑)	Preserving (↑)	Quality (↑)
P2P [7]	11.13	4.80	8.09	5.66	5.37	5.77
PnP [26]	7.72	7.17	6.93	6.54	6.77	6.74
DDS [6]	20.30	10.82	16.23	7.60	7.51	7.37
CDS [17]	17.02	16.72	17.08	8.26	8.00	8.09
IDS (Ours)	**43.83**	**60.49**	**51.67**	**8.97**	**9.00**	**8.80**

Table 2. **User study and GPT scores** [19] show that our method achieved the highest scores across all questions for image editing.

5.2. Editing NeRF

We conduct experiments involving 3D rendering of edited images to demonstrate the effectiveness of our method in

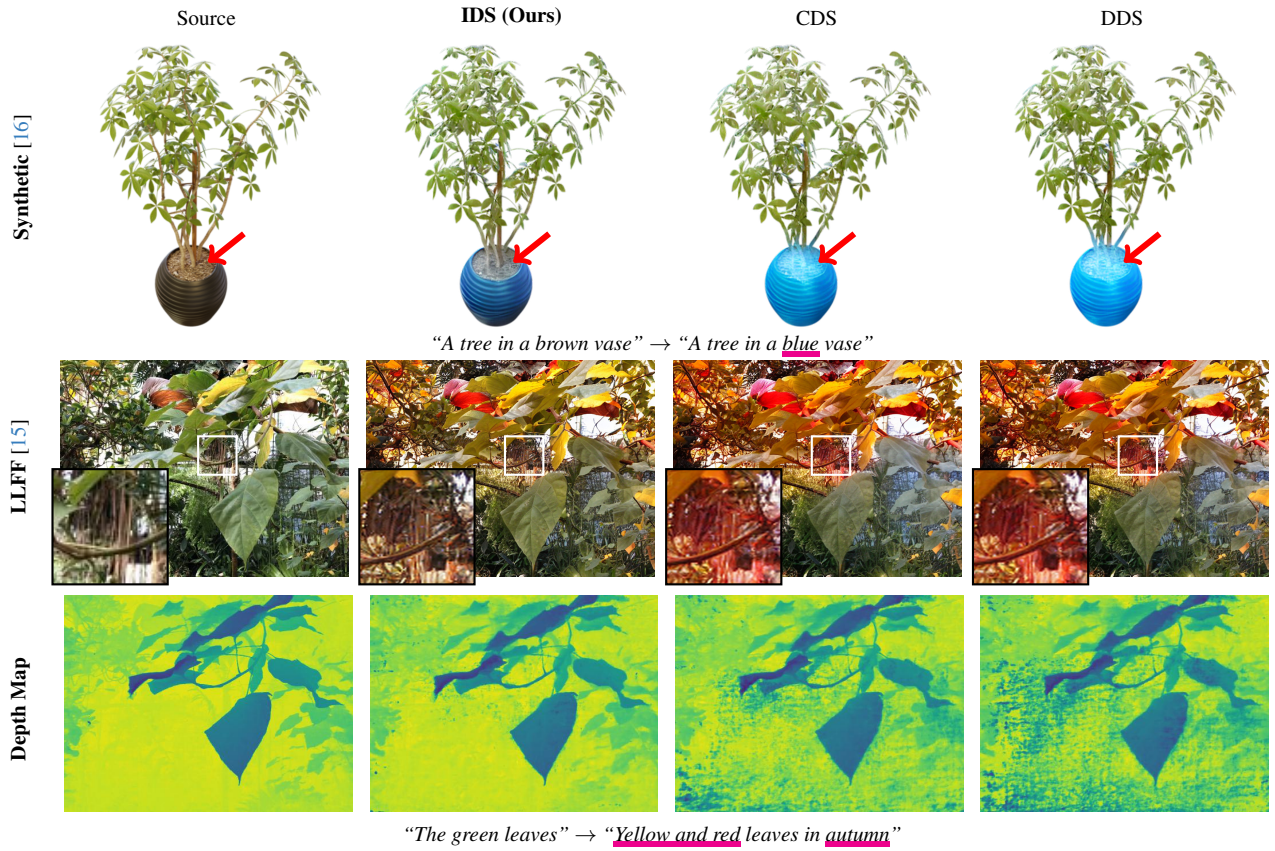


Figure 5. **Qualitative results on Synthetic 360° and LLFF datasets.** IDS outperforms the baselines by preserving the structural consistency of the source image and maintaining the integrity of regions that should remain unchanged, while precisely editing only the areas specified by the target prompt. Furthermore, comparisons of the depth map results also highlight the superior consistency of our method over other baseline models.

maintaining structural consistency. This approach is particularly relevant as consistency has an even greater impact on outcomes in 3D environments.

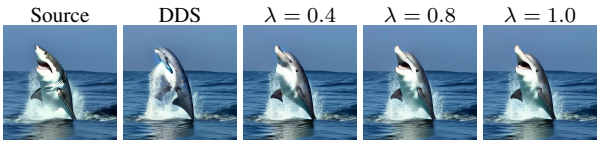
Datasets. We evaluated our method on widely used NeRF datasets: Synthetic NeRF [16] and LLFF [15]. Since NeRF datasets have no given pairs of source and target prompts, we manually composed image descriptions.

Qualitative Results. Fig. 5 illustrates the qualitative results of our method compared with NeRF editing baselines. In the first row, the target prompt specifies a precise part of the image for fine-grained editing. DDS [6] and CDS [17] fail to differentiate and edit the specific area. At the same time, our method accurately identifies the region indicated by the target prompt in the image and performs detailed editing exclusively on that part. The second row demonstrates a scenario in which the target prompt is designed to edit the mood of the image. Our approach adjusts the colors associated with "autumn" and "leaves" throughout the image while maintaining consistency in the "trunk" whereas DDS and CDS also changed the "trunk". In terms of depth maps, our method generates clean depth maps with minimal noise after image editing, whereas DDS and CDS introduce noticeable noise into the depth maps.

Metric	CLIP [21] (↑)	User Preference Rate (%)		
		Text (↑)	Preserving (↑)	Quality (↑)
DDS [6]	0.1596	36.88	28.37	32.62
CDS [17]	0.1597	22.70	23.40	21.28
IDS (Ours)	**0.1626**	**40.42**	**48.23**	**46.10**

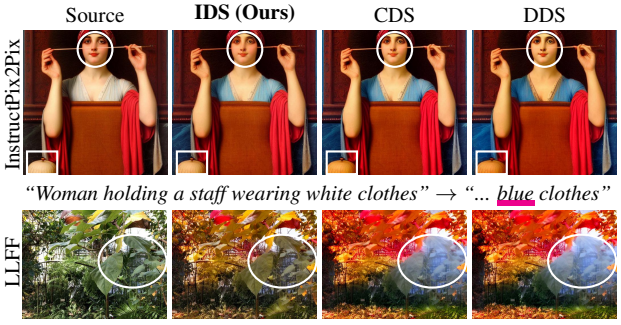
Table 3. **Quantitative results of NeRF editing** with respect to CLIP score and User Preference Rate. IDS demonstrates superior quantitative performance compared to the baselines.

Quantitative Results. Based on edited images, we performed 3D rendering and subsequently conducted quantitative evaluations provided in Tab. 3. To assess whether the edited 3D images are precisely aligned with the target prompts, we measured the CLIP [21] scores at 200k iterations of training on the LLFF dataset. We additionally present a user evaluation conducted under the same setup in Sec. 5.1. Consistent with the trends observed in the qualitative results, our method demonstrates superior performance in the quantitative evaluations compared to other baselines.



"A photo of shark" $\rightarrow$ *"A photo of dolphin"*

Figure 6. **Ablation study** on scale λ. To show the effect of scale more extremely, the number of iterations of FPR is set as 1. The result of DDS is the same as $\lambda = 0$ since it means no update for source identity.



"Woman holding a staff wearing white clothes" $\rightarrow$ *"... blue clothes"*

"The green leaves" $\rightarrow$ *"Yellow and red leaves in autumn"*

Figure 7. **Optimization steps.** Our method effectively preserves the consistency of the source image, even as the number of iteration steps increases up to 400.

	optim step	FPR iter	LPIPS(↓)	CLIP(↑)	time (sec/img)	Memory (GB)
DDS	200	-	0.240	0.293	22.45	6.27
CDS	200	-	0.210	0.287	59.31	8.83
IDS	200	1	0.199	0.285	50.80	8.63
		3	0.190	0.277	107.77	
	100	3	0.165	0.265	54.04	
	150		0.180	0.272	81.25	

Table 4. **Computational complexity** on 28 images of InstructPix2Pix [2] for various settings. Lower LPIPS and higher CLIP scores mean better quality.

6. Discussions

6.1. Ablation studies on FPR

FPR iteration N**.** We conduct experiments on FPR iterations N to evaluate its impact and determine the optimal iteration count. Although performing just one iteration of FPR is sufficient to preserve the source identity, as shown in lower LPIPS score than baselines of Tab. 4, we set $N = 3$ to emphasize the purpose of our method.

Scale λ**.** The scaling factor λ of FPR determines how much information of source latent $\mathbf{z}^{\text{src}}$ is kept. As shown in Fig. 6, increasing the scale preserves the attributes of the source image, resulting in more successful editing when it is hard to translate due to the structural mismatch between the source and the target prompt.

6.2. Optimization steps

To show that our method can prevent error accumulation during translation, we set the experiment to extend the number of optimization steps from 200 to 400. In the results of DDS and CDS, there is color boosting or loss of details due to the cumulated error. In contrast, IDS maintains the characteristics of the original images, such as the color of the pumpkin in the first row of Fig. 7 and the shape of the leaf in the second row of Fig. 7, better than other methods.

7. Limitation

The proposed IDS demonstrates outstanding performance across evaluation metrics assessing consistency between source and target images. However, during FPR process, IDS relies solely on information from the source ($\{\mathbf{z}^{\text{src}}, y^{\text{src}}\}$) without incorporating target-side information. This results in comparatively lower CLIP scores [21] than other baselines (Tab. 5) and failure cases for more complex translations (Fig. 8). In addition, our method requires additional computational overhead (Tab. 4) since FPR is applied to each optimization iteration. Detailed discussion about our limitation is provided in Sec. G of supplementary. Our future direction will explore changing the score conditioned by the target prompt y^{trg}, leading to a better alignment with y^{trg}.

	P2P [7]	PnP [26]	DDS [6]	CDS [17]	**IDS (Ours)**
cat2lion	0.29	0.21	**0.30**	0.29	0.29
cat2dog	**0.27**	0.26	**0.27**	**0.27**	0.26
Ip2p	0.28	**0.30**	0.29	0.29	0.28

Table 5. **Limitation of IDS with respect to CLIP score[21]** for image editing on LAION 5B [23] and InstructPix2Pix [2].

"Photo free night, house, aurora" $\rightarrow$ *"... with two dogs"*

Figure 8. **Failure case** for complex text prompt.

8. Conclusion

We proposed a new distillation sampling method using a fixed-point regularization which aligns the text-conditioned score towards identity-preserved manifolds. The proposed fixed-point regularization preserves the source's identity by re-projecting the intermediate score status on posterior means. In this manner, corrected noises guide a gradient of distilled score toward identity-consistent manifolds. Owing to self-correction by a fixed-point iterator and guided injection noise, the proposed identity-preserving distillation sampling provides clear and unambiguous representations corresponding to the given prompts in text-guided image editing and editable neural radiance field (NeRF). Furthermore, our model can be utilized as a universal module in addition to the existing score-sampling processes.

Acknowledgement This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00335741, RS-2024-00357197).

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. *arXiv preprint arXiv:2403.17377*, 2024. 1

[2] 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. 1, 2, 5, 6, 8

[3] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 5

[5] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 2

[6] 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. 1, 2, 3, 4, 5, 6, 7, 8

[7] 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. 1, 2, 3, 5, 6, 8

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 3

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[10] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023. 1

[11] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 3

[12] 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

[13] Barak Meiri, Dvir Samuel, Nir Darshan, Gal Chechik, Shai Avidan, and Rami Ben-Ari. Fixed-point inversion for text-to-image diffusion models. *arXiv preprint arXiv:2312.12540*, 2023. 4

[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. 1, 2, 3

[15] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 7

[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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 5, 7

[17] 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, 5, 6, 7, 8

[18] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014. 3

[19] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 6

[20] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 3

[21] 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. 7, 8

[22] 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. 1, 3

[23] 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. 5, 8

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5

[25] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 3

[26] 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. 1, 2, 3, 5, 6, 8

[27] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[28] Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9892–9902, 2024. 3

[29] 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. 5

[30] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 3