Stable Score Distillation

Haiming Zhu[1], Yangyang Xu[2]*, Chenshu Xu[1], Tingrui Shen[3],
Wenxi Liu[4], Yong Du[5], Jun Yu[2], Shengfeng He[1]*
[1]Singapore Management University [2]Harbin Institute of Technology (Shenzhen)
[3]South China University of Technology [4]Fuzhou University [5]Ocean University of China

(a) Original Views (b) CSD[52] (c) DDS[10] (d) Ours

Figure 1. We propose Stable Score Distillation (SSD), a method that improves text-guided editing by preserving original content structure and enhancing realism in edited results. The first column shows the original views, while the remaining columns display comparative results with existing score distillation methods on the SD model. SSD demonstrates superior performance, maintaining the integrity of the original content and producing more realistic edits.

Abstract

Text-guided image and 3D editing have advanced with diffusion-based models, yet methods like Delta Denoising Score often struggle with stability, spatial control, and editing strength. These limitations stem from reliance on complex auxiliary structures, which introduce conflicting optimization signals and restrict precise, localized edits. We introduce Stable Score Distillation (SSD), a streamlined framework that enhances stability and alignment in the editing process by anchoring a single classifier to the source prompt. Specifically, SSD utilizes Classifier-Free Guidance (CFG) equation to achieve cross-prompt alignment, and introduces a constant term null-text branch to stabilize the optimization process. This approach preserves the original content's structure and ensures that editing trajectories are closely aligned with the source prompt, enabling smooth, prompt-specific modifications while maintaining coherence in surrounding regions. Additionally, SSD incorporates a prompt enhancement branch to boost editing strength, particularly for style transformations. Our method achieves state-of-the-art results in 2D and 3D editing tasks, including NeRF and text-driven style edits, with faster convergence and reduced complexity, providing a robust and efficient solution for text-guided editing. Code is available at: https://github.com/Alex-Zhu1/SSD.

*Corresponding authors: Yangyang Xu (xuyangyang@hit.edu.cn) and Shengfeng He (shengfenghe@smu.edu.sg).

1. Introduction

Text-based image generation has achieved remarkable progress, particularly with the advent of diffusion models [13, 28, 40, 41, 43, 56]. These models leverage strong priors to produce high-quality images, facilitating significant advances in text-to-3D generation [37, 47, 58]. Moreover, text-guided 3D editing has enabled intricate modifications to shape and texture [26, 31], supporting flexible and precise 3D scene manipulation.

Unlike generation tasks that create new content, editing tasks aim to modify specific elements within an image while preserving surrounding areas [51]. However, directly applying methods like Score Distillation Sampling (SDS) to editing tasks can yield undesired effects, such as blurring across the image. This arises because SDS optimizes globally to the prompt, affecting regions beyond the targeted area [10, 20]. DDS [10] addresses this by introducing a dual-branch architecture, pairing the source image with its description to leverage the model's inherent bias and isolate specific prompt changes. Further, CSD [52] achieved

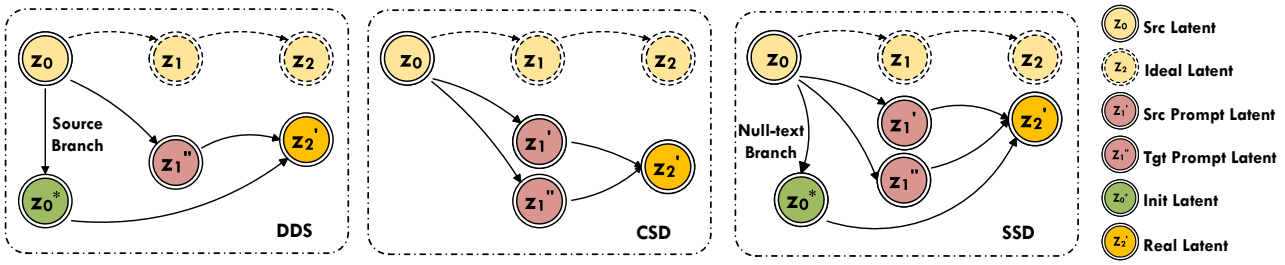

Figure 2. Illustration of three distillation-based approaches. *Note that we assume a 2-step optimization process for illustration, where the subscript t represents the iteration number.* **DDS** utilizes the source branch to obtain initial latent Z_0^*, while **CSD** employs two classifiers to derive Z_1' and Z_1'' for cross-prompt editing. Our **SSD** method designs a CFG classifier to determine the cross-prompt editing, introduces the null-text branch as the initial latent Z_0^*, and further constructs the cross-trajectory term (see Sec. 4.1) for stable optimization.

Figure 3. The optimization process of DDS and our SSD. SSD preserves the source structure effectively during optimization iterations, while DDS cannot preserve it effectively.

scene editing by incorporating a classifier component within Classifier-Free Guidance (CFG) [12] to refine the prediction score by applying the classifier to both the source and target prompts. NFSD [20] further decomposes the CFG score, highlighting the classifier as the primary driver of prompt direction.

Despite their success, we argue that current distillation-based approaches face inherent limitations, such as low editing quality and loss of source content. As shown in Fig. 2, DDS [10] relies on the source branch to remove model bias, but lacks the explicit guidance to preserve the source content [24] during optimization. As shown in Fig. 1c, DDS changes man's clothe during editing his faces. Additionally, although introducing source prompt components is intended to improve prompt specificity [19], it can amplify noise and introduce overlapping objectives that hinder stable convergence. This results in artifacts or unintended variations, especially in the unedited regions. Correspondingly, CSD [52] utilizes dual classifiers to refine the prompt editing direction (see Fig. 2). However, it lacks the explicit source preservation to restrict edits precisely to the target areas. As shown in Fig. 1b, this causes the structure deformation and annoying artifacts in the edited regions.

Our insights into these limitations lead to two key observations: (1) **Cross-prompt**: a single classifier, providing the editing direction from source prompt to target, and (2) **Cross-trajectory**: stability in the editing process can be achieved by aligning the editing direction closely with the structure of the source content.

In this paper, we propose Stable Score Distillation (SSD), a streamlined approach for stable and precise text-guided editing. To achieve a smooth editing direction, we employ the CFG equation for both the source and target prompts, ensuring a gradual transition of the original contextual texture as the model adapts to the specified changes. This approach contrasts with DDS [10], as it eliminates the need for a auxiliary source branch, enabling our method to focus editing gradients precisely within target regions while ensuring a stable transition, as illustrated in Fig. 3. Moreover, for aligning the editing direction with the source prompt, facilitating smoother and more controlled progression toward the target prompt, we design an cross-trajectory strategy to ensure that edits respect the original structure, supporting subtle and stable transformations within designated areas. While NFSD [20] utilizes negative-branch and DDS utilizes source branch to enhance output clarity, as shown in SSD in Fig. 2, we introduce a null-text branch aligned with the "no-edit" direction to integrate a "reconstruction" term to explicitly enforce source content preservation, which enhances consistency and produces reliable edits across diverse tasks. Based on above designs, our framework remains streamlined and efficient, achieving both precision and stability without the complexity of additional components.

Our framework integrates seamlessly into existing DDS-based editing pipelines and applications, such as text-driven NeRF editing [22, 24, 36] and 2D image editing [35]. Notably, our approach's "clear" editing direction preserves source content, making a carefully designed identity regularization [22] unnecessary. Moreover, standard DDS methods often lack sufficient editing strength, resulting in minimal or negligible changes in output, particularly in style editing [23]. Our approach, with its streamlined and stable framework, allows for the seamless integration of a prompt enhancement branch to amplify editing capability.

With these improvements, our method achieves faster and

more effective edits during optimization, remains compatible with the Stable Diffusion Model [40] without requiring LoRA [14] or fine-tuning, and integrates effectively with Instructpix2pix [2]. Additionally, by incorporating non-increasing timestep sampling [15], we accelerate convergence, reducing the required iterations to approximately 3,000 for NeRF [33] and 1,500 for Gaussian splatting [21].

In summary, our contributions are as follows:

- We introduce a novel editing framework, Stable Score Distillation, that leverages a single, anchored classifier to achieve targeted and stable edits in 3D scene editing.
- We introduce a prompt enhancement strategy, effectively improve the prompt-alignment, especially style editing in 2D-image editing.
- We demonstrate the effectiveness of our approach across NeRF-editing and image-editing tasks, achieving state-of-the-art results with a streamlined and efficient framework.

2. Related Work

2.1. Diffusion Models

Diffusion models [13, 39, 43, 53] have made significant advancements in generating diverse and high-fidelity images. Starting form a gaussian noise, diffusion models can predict the noise-less sample at each time step, until finally obtaining clear samples. Commonly, the denoising process can utilize U-net model to predict the noise. Some works [13, 44] have observed that is proportional to the predicted score function [17] of the smoothed density. Thus, intuitively, taking steps in the direction of the score function gradually moves the sample towards the data distribution.

To generate images aligned with a target prompt, guidance is typically introduced to explicitly control the weight assigned to the conditioning information. The popular guidance methods include Classifier Guidance [8] and Class-free Guidance (CFG) [12]. While the former rely on a separately learned classifier, the latter directly introduces null-text samples to the model. CFG modifies the score function to steer the process towards regions with a higher ratio of conditional density to the unconditional one. However, it has been observed that CFG trades sample fidelity for diversity [12]. Based on the insights gained from the decomposition of the CFG equation, we propose a novel Stable Score Distillation (SSD) method to *guide* the SDS optimization process.

2.2. Score Distillation Sampling (SDS)

Benefit from the data scale-law, diffusion model [38, 40, 41] achieve high-quality image generation and text-to-image generation. Specifically, Score Distillation Sampling (SDS) [37] leveraging the priors of pre-trained text-to-image models to facilitate text-conditioned generation in 3D content generation. Specifically, SDS is an optimization approach that updates the rendering parameter towards the image distribution of diffusion models by enforcing the noise prediction on noisy rendered images to match sampled noise. While SDS provides an elegant mechanism for leveraging pre-trained text-to-image models, SDS-generated results often suffer from oversaturation and lack of fine realistic details. VSD [47] proposed a particle-based optimization framework that treats the 3D parameter as a random variable of target distribution. Furthermore, by regarding SDS as a reverse diffusion process, decreasing timesteps sampling [15, 58] to imitate the diffusion reverse sampling, which can improve the quality of the generated 3D assets.

In image editing, Delta Denoising Score (DDS) [10] found that Score Distillation Sampling (SDS) introduces noticeable artifacts and over-smoothing in edited images due to inherent bias. To mitigate this bias, DDS employs a subtraction of two SDS scores of the source and target images to obtain a delta score, which is then used to guide the optimization process.

2.3. Text-Driven 3D-Scene Editing

Text-driven 3D scene editing has been a popular research topic [18, 29, 56]. IN2N [9] proposed a Iterative Dataset Update method that can edit 3D scenes from text descriptions. By leveraging advancements in 2D diffusion editing techniques, notably InstructP2P [2] and ControlNet [54], GaussianEditor[6] and GaussCtrl [49] utilize edit multi-view images latent to optimize the 3D scene. We consider utilize score distillation to guide the 3D scene editing, which is more flexible and efficient for the text-driven 3D scene editing.

Building upon the foundational SDS loss introduced by DreamFusion [37], some work has explore SDS loss in the text-driven 3D scene editing. RePaint-NeRF [57] has advanced the application of SDS in 3D editing by integrating a semantic mask to guide and constrain modifications within the background elements. CSD [52] utilize two classifiers to achieve editing. In a similar vein, ED-NeRF [36] has introduced an enhanced loss function specifically designed for 3D editing tasks. PDS [24] proposed a posterior distillation sampling to match stochastic latent [16]. Piva [25] fine-tuned the model while introducing a regularization term to preserve identity. Unfortunately, these methods are still limited to the long-time diffusion reverse sampling process, which is not suboptimal for the text-driven 3D scene editing. DreamCatalyst [22] extends the PDS optimization processing to ID-preserving and edit-ability based on decreasing timesteps sampling.

Different from the above methods, ours firstly improve the DDS optimization process by introducing a single classifier, and further introduce a null-text branch to achieve a more stable and precise editing process.

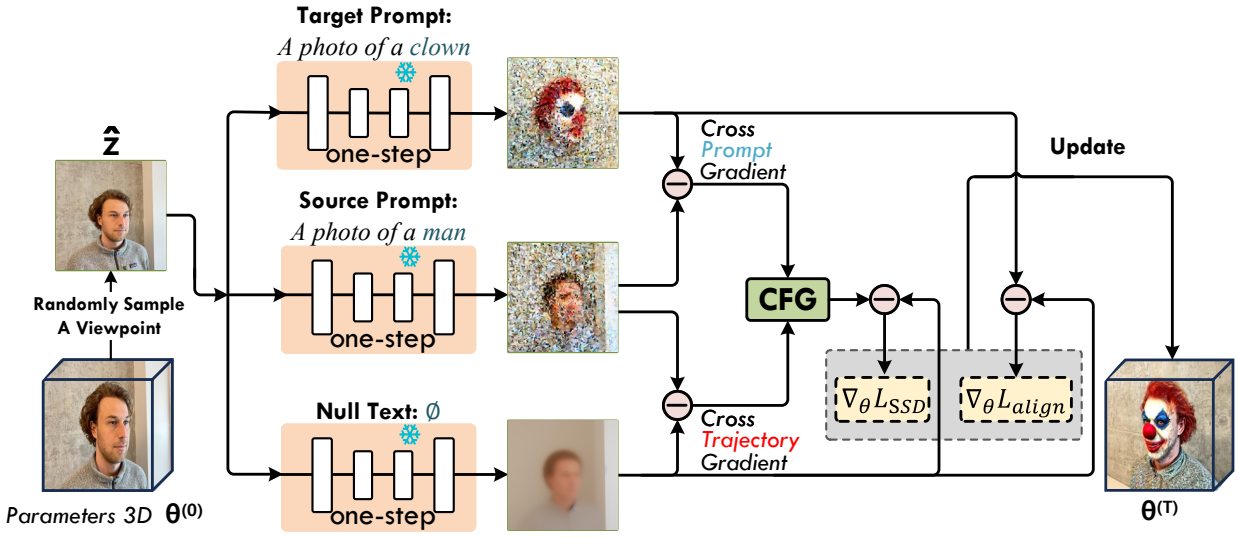

Figure 4. The overview of SSD. Given the parameter 3D-model or image, SSD provides effective editing gradient to guide the optimization process. We utilize CFG equation between the predicted target noise $\epsilon_\phi(z_t, y, t)$ and source noise $\epsilon_\phi(z_t, \hat{y}, t)$, which generate the gradual editing direction. Furthermore, we introduce a null-text branch $\epsilon_\phi(\hat{z}_t, \varnothing, t)$ to regularize the optimization process and achieve stable optimization. We further analyzing and decompose ours design term into three parts: cross-prompt, cross-trajectory, and prompt-enhance.

3. Preliminary

In this section, we first discuss existing optimization-based approaches to handle parametric images. Then, we will introduce our novel parametric image editing method in Section 4.

3.1. Score Distillation Sampling

Score Distillation Sampling (SDS) [37] is proposed to generate parametric images by leveraging the 2D prior of pretrained text-to-image diffusion models. Specifically, given a pretrained diffusion model ϵ_ϕ, SDS optimizes a set of parameters θ of a differentiable parametric image generator g, using the gradient of the loss L_{SDS} with respect to θ:

$$\nabla_\theta L_{\text{SDS}} = w(t)\left(\epsilon_\phi(z_t(x); y, t) - \epsilon\right)\frac{\partial x}{\partial \theta}, \tag{1}$$

where $x = g(\theta)$ is an image rendered by θ, $z_t(x)$ is obtained by adding a Gaussian noise ϵ to x corresponding to the t-th timestep of the diffusion process, and y is a condition to the diffusion model. As Noise-Free Score Distillation (NFSD) [20] has shown, the score $\epsilon_\phi(z_t(x); y, t)$ provides the direction in which this noised version of x should be moved towards a denser region in the distribution of real images.

3.2. Delta Distillation Sampling

Although SDS get excellent generation ability, for editing task, an undesired component from the pretrained model, δ_{bias}, interferes with the process and causes the image to become smooth and blurry in some parts [10]. Based on the observations that a matched source prompt $\hat{y}$ and source latent $\hat{z}_t$ can estimate the noisy direction δ_{bias}, thus, the DDS method aims to remove the δ_{bias} by introducing source branch, as shown in Eq. 2:

$$\nabla_\theta L_{\text{DDS}} = \left(\epsilon^c_\phi(z_t, y, t) - \epsilon^c_\phi(\hat{z}_t, \hat{y}, t)\right)\frac{\partial z}{\partial \theta}, \tag{2}$$

where $\epsilon^c_\phi(z_t, y, t)$ and $\epsilon^c_\phi(\hat{z}_t, \hat{y}, t)$ are pretrained model predictions ϵ, with the superscript c indicating the CFG results. Thus, DDS pushes the optimized image into the direction of the target prompt without the interference of the noise component, namely, $\nabla_\theta L_{\text{DDS}} \approx \delta_{\text{text}}$. Obviously, $\nabla_{\delta_{\text{text}}}$ is contingent on classifier part from $\epsilon^c_\phi(z_t, y, t)$ as discussed in CSD [52] and NFSD [20]. Note that in the following manuscript, we decompose the CFG results without the superscript c and and omit the timestep t for simplicity.

Further exploring prompt editing direction, CSD [52] method proposed a dual-classifier to refine the editing score and achieve more precise editing, as shown in Eq. 3:

$$\begin{aligned}\nabla_\theta L_{\text{CSD}} =& \big(w_a\left(\epsilon_\phi(z_t, y) - \epsilon_\phi(z_t, \varnothing)\right) \\ & - w_b\left(\epsilon_\phi(z_t, \hat{y}) - \epsilon_\phi(z_t, \varnothing)\right)\big)\frac{\partial z}{\partial \theta},\end{aligned} \tag{3}$$

while the $\epsilon_\phi(z_t, y, t)$ and $\epsilon_\phi(z_t, \hat{y}, t)$ are **current** latent z_t predictions for the target prompt y and source prompt $\hat{y}$, respectively. w_a and w_b are weights of classifiers. Simply put, CSD aims to refine the prompt editing direction by determining the difference between two classifiers, which can be regarded as a cross-prompt term.

4. Method

In 3D scene editing process, which requires consideration of both the target prompt and the original source content,

we consider two key aspects: (1) smooth editing direction towards the target prompt (2) and editing results respect the original structure. Based on these, in this section, we introduce our novel editing framework Stable Score Distillation.

4.1. Stable Score Distillation

Firstly, we introduce the design of a cross-prompt editing direction. As discussed about CSD method in Sec. 3.2, the key role in cross-prompt editing is to provide a smooth transition from the source prompt to the target. As the CFG guidance [12] steers the process towards regions with a higher ratio of conditional density to the unconditional one, accordingly, we can modify the SDS score function, as shown below:

$$Grad = \epsilon_\phi(z_t, \hat{y}) + s\left(\epsilon_\phi(z_t, y) - \epsilon_\phi(z_t, \hat{y})\right), \quad (4)$$

where $\epsilon_\phi(z_t, y)$ and $\epsilon_\phi(z_t, \hat{y})$ are pretrained model predictions. The scale factor s is equal to control weight.

Although the cross-prompt term provides a smooth texture transition in the edited region, we observed that the optimization process leads to abrupt structural changes, often resulting in artifacts and unappealing outcomes, similar to CSD in Fig. 1b. To address this, we introduce an additional regularization term to constrain the structural transition. Interestingly, as shown in Fig. 1c, DDS achieves better results than CSD by incorporating a source branch. However, DDS still lacks a mechanism to ensure the original structure remains intact, leading to modification on unedited regions. To address this, we introduce a null-text branch $\epsilon_\phi(\hat{z}_t, \varnothing)$ to regularize the optimization process, as shown in Eq. 5:

$$L_{\text{ssd}} = \epsilon_\phi(z_t, \hat{y}) + s(\epsilon_\phi(z_t, y) - \epsilon_\phi(z_t, \hat{y})) - \epsilon_\phi(\hat{z}_t, \varnothing). \quad (5)$$

Eq. 5 is ours Stable Score Distillation, and we can further decompose above equation into two parts, and the latter is regarded as a cross-trajectory term.

$$L_{\text{ssd}} = \underbrace{w_p\left(\epsilon_\phi(z_t, y) - \epsilon_\phi(z_t, \hat{y})\right)}_{\text{cross-prompt}} + \underbrace{w_t\left(\epsilon_\phi(z_t, \hat{y}) - \epsilon_\phi(\hat{z}_t, \varnothing)\right)}_{\text{cross-trajectory}}, \quad (6)$$

where the w_t and w_p control the strength of the cross-trajectory and cross-prompt, respectively.

The cross-trajectory term can be interpreted as the distance between the transitions of two latents, ensuring that the original structure remains smooth and does not change abruptly (more details are provided in the supplementary material). In Fig. 5, we can see that the cross-trajectory term can provide a strong structure constraint ability, guiding the optimization process to preserve the source image structure. Specifically, when set $w_t = 0$, the optimization process behaves similarly to the CSD[52] method, which fails to retain the original image structure.

Figure 5. The effect of increasing the strength of the prompt enhancement term (w_e) and cross-trajectory term (w_t), with the cross-prompt term fixed at 7.5. Both terms contribute to prompt-aligned results, while setting ($w_t = 0$) leads to saturation and discard source content.

4.2. Improving Prompt Alignment

Although Eq. 5 can achieve gradual editing results, we found Eq. 5 have similar limitation with DDS [7], which have insufficient editing strength. The editing results neither get successful editing nor retain the source image structure, often leads to little or no change in the final. Benefit from the cross-prompt editing design as Eq. 5, we can add a target prompt enhancement branch to guide the optimization process. The target prompt alignment branch will provide the direction of the target prompt, as shown in Eq. 7:

$$L_{\text{align}} = w_e\left(\epsilon_\phi(z_t, y) - \epsilon_\phi(z_t, \varnothing)\right), \quad (7)$$

where w_e is the prompt enhancement scale. As shown in Fig. 5, the synchronous scaling of both the cross-trajectory and prompt-enhancement terms results in effective visual editing outcomes.

4.3. Source Latent Regularization

Empirically, we found that directly using latent-space loss rather than pixel-level loss can lead to optimization difficulties in local regions of 3DGS. For example, the bright spots appearing in Fig. 5. To suppress the steep gradients in these areas, we incorporate ID regularization to guide the stable optimization process. Differ with PDS [24] use source latent $\hat{x}_0$, we can use the noisy latent $\hat{x}_t$ to avoid partial exploding gradient, as shown in Eq. 8:

$$L_{\text{ID}} = w(t) \cdot (x_t - \hat{x}_t), \quad (8)$$

where the $w(t)$ is the iteration-dependent strength, designed as a decreasing function of t. Notably, the $w(t)$ is not necessary to well-designed in our design.Our final loss function

as shown in Eq. 9:

$$L_{\text{final}} = L_{\text{ssd}} + L_{\text{align}} + L_{\text{ID}}. \tag{9}$$

Based on the above design, we achieve a more prompt-aligned editing method, which integrates seamlessly into the Stable Diffusion Model [40] without requiring LoRA [14] or fine-tuning. Moreover, we will further introduce our method's connection with InstructPix2Pix [2].

4.4. Connecting with IP2P

The final design of our method is shown in Eq 9. We found that ours edit gard provide new angle to understand about InstructP2P [2] one-step reverse sampling.

$$\begin{aligned}\epsilon_\theta(z_t, c_I, c_T) = \epsilon_\theta(z_t, \varnothing, \varnothing) \\ + s_I\big(\epsilon_\theta(z_t, c_I, \varnothing) - \epsilon_\theta(z_t, \varnothing, \varnothing)\big) \\ + s_T\big(\epsilon_\theta(z_t, c_I, c_T) - \epsilon_\theta(z_t, c_I, \varnothing)\big),\end{aligned} \tag{10}$$

where c_I and c_T are input-image and instruction prompt separately,s_I and s_T are the source image control and instruction prompt control strength. The Eq. 10 is the InstructP2P one-step reverse sampling, which can provide the direction of the target prompt. We can see that the InstructP2P is the simple version of our method, the middle term of Eq. 10 is cross-trajectory regularization, and the last term of Eq. 10 is ours cross-prompt term. Simply put, as analyzing the Eq. 5, subtracting the constant correction term $\epsilon_\theta(\hat{z}_t; \varnothing; \varnothing)$ is edit grad. Ours method reveal that apply DDS loss in the InstructP2P model can only editing branch, and don't have to provide the source branch.

5. Experiments

In this section, we conduct editing experiments across two types of parameterized images. Section 5.1 evaluates the effectiveness of our method on 3D Scenes Editing, and Section 5.2 evaluates the effectiveness of our method on 2D Image Editing. We also conduct ablation studies to analyze the effectiveness of ours components in Section 5.3.

5.1. 3D Scenes Editing

Dataset. To evaluate the effectiveness of our method, we conduct experiments on the scenes from IN2N [9] and other real-world datasets, including LLFF [32] and Mip-Nerf360 [1].

Baselines. We compare our method with several state-of-the-art inversion methods. We use 3DGS [21] as the 3D representation, and compare our method with InstructNerf2Nerf [9], DDS [10], GS-Edit [6], and DGE [5]. For fairness, we implement the DDS version based on the official GS-Edit code. PDS [24] is designed for addition of objects to unspecified regions, we will provide the comparison results in supplementary material.

Table 1. Quantitative evaluations under 3D editing scenes.

Method	CLIP Sim. ↑	Sim Dir. ↑	User Study ↑
IN2N	0.1676	0.0707	14.54%
DDS	0.1780	0.0401	5.45%
GS-Editor	0.1758	0.0429	14.54%
DGE	0.1758	0.0563	23.63%
Ours	**0.1846**	**0.0773**	**41.81%**

Evaluation Metrics. We follow common practice [5, 6, 9] to evaluate the effectiveness of our method. CLIP Similarity is to evaluate the alignment between the render images and the target prompts, i.e., the cosine similarity between the text and image embeddings encoded by CLIP. Specifically, follow DGE [5], randomly sample 20 camera poses to evaluate. CLIP Directional Similarity is to measure the editing effect, i.e., the cosine similarity between the image and text editing directions (target embeddings minus source embeddings). We evaluate all methods on 6 different scenes and 10 different prompts.

Results. We begin by evaluating our method, starting with a qualitative assessment. In Fig. 6, we present a comparison of results with competing methods. Our approach generates more visually appealing images that are better aligned with the editing instructions. In contrast, methods based on the Iterative Dataset Update (IDU) strategy, such as IN2N [9] and GS-Editor [6], fail to produce the desired editing outcomes, resulting in blurrier or lower-fidelity reconstructions and noticeable artifacts. For example, in the scene of "*Spider-Man with a mask*", IN2N generates a mask with reduced fidelity, while GS-Editor produces a low-detail mask. In the multi-view consistency setup, DGE [5] performs well on common attributes but is constrained to "rainbow" editing and tends to generate artifacts outside the segmentation mask. Our method works seamlessly with masks, producing results with rich details.

In Tab. 1, We present a quantitative comparison. Our method outperforms the baselines in terms of CLIP Similarity and CLIP Directional Similarity. Notably, Dire Sim is not sensitive with the editing quality, much focus on the instruction attributes. We conducted a user study with a survey of 55 participants to evaluate the editing quality. The results show that our method received the most popular votes.

5.2. 2D Image Editing

Dataset. To evaluate the effectiveness of our method, we conduct experiments on the PIE-Bench dataset proposed by PNPInv [19], which consists of 700 images with 9 editing types. Each image is annotated with the source and target prompts.

Baselines. We compare our method with several classical editing methods based on DDIM [43] inversion, including P2P [11], PNP [46] and MasaCtrl [3]. For optimization-

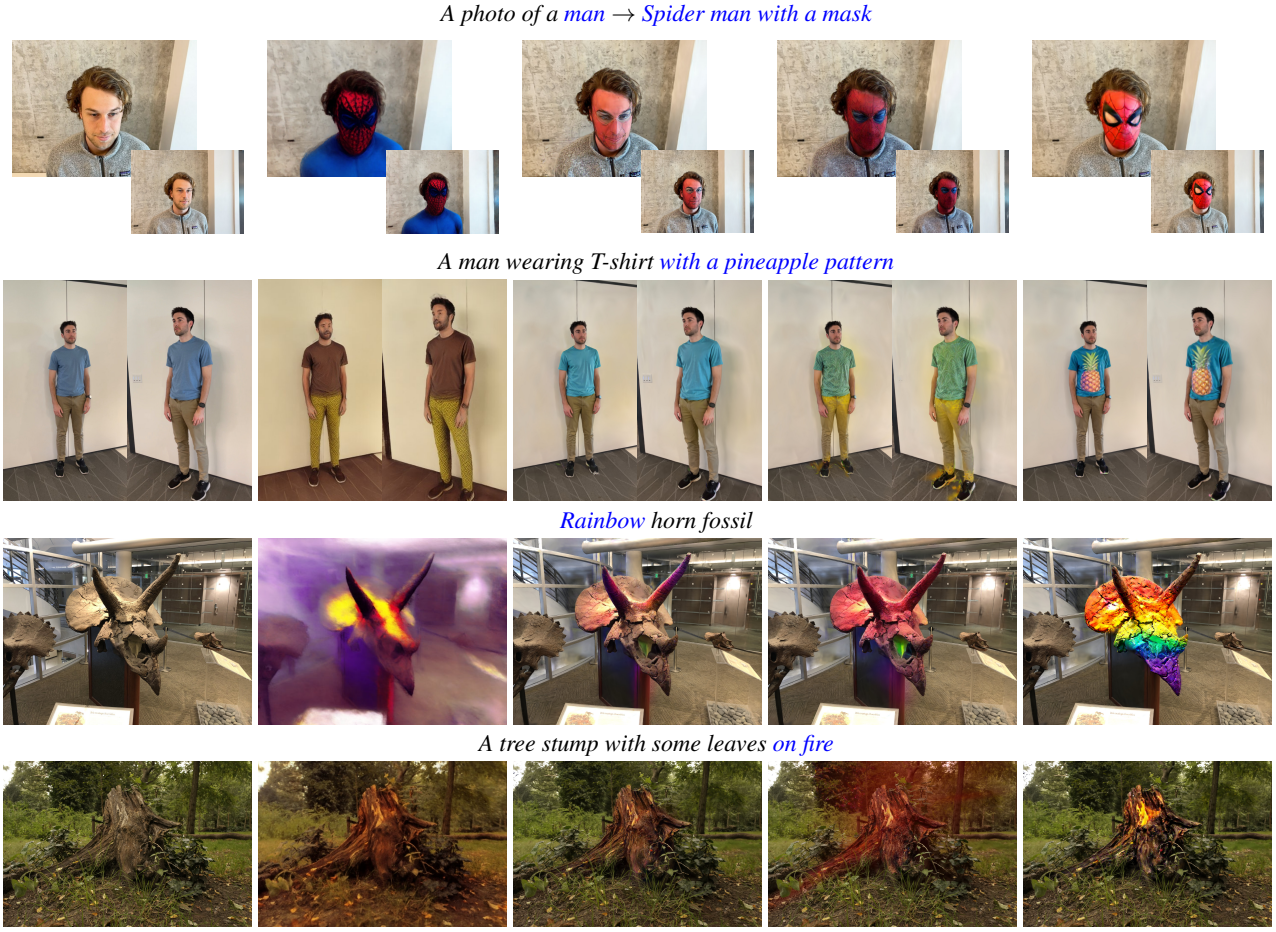

Figure 6. Qualitative comparisons with related works. SDS demonstrates outstanding performance in effectively preserve source structure in the modified region.

based editing method, we compare with NT [34] and StyleD [27]. Besides, we report the comparison with DT [19]. Further, we compare with DDS [10] and its extended method CDS [35].

Evaluation Metrics. We follow DT [19] which uses several metrics to evaluate our method. We use the Structure Distance assessed by DINO score [4] to evaluate the structure distance between original and edited images. We also introduce several metrics to evaluating the background preservation, which includes LPIPS [55] and MSE. Besides, we introduce CLIP Similarity [48] to evaluate the text-image consistency between edited images and corresponding target editing text prompts.

Results. We present a qualitative comparison of our method with competitors in Fig. 7. Our method generates images that are more aligned with the target prompts and preserve the source structure. In "*blue butterfly*", ours successfully changes the color of the butterfly to blue, while DDS [10] and CDS [35] generate similar color from source. Especially, ours method successfully changes the style of the image and generates appealing results, which is challenging for DDS-based methods. Compared to optimization-based methods, NT [34] preserves the general source structure during the inversion process, but tends to discard some content as evident in the distortion of the girl's fingers in Fig. 7. Additionally, due to limitations in the editing methods, the editing results are unsuccessful.

Table 2. Quantitative evaluation in PIE-Bench dataset.

Method	Distance$_{\times 10^3}$	↓LPIPS$_{\times 10^3}$	↓MSE$_{\times 10^4}$	↓CLIP ↑
DDIM + P2P	69.43	208.80	219.88	25.01
DDIM + PNP	28.22	113.46	83.64	25.41
DDIM + MasaCtrl	28.38	106.62	86.97	23.96
NT + P2P	13.44	60.67	35.86	24.75
StyleD + P2P	11.65	66.10	38.63	24.78
DT + P2P	11.65	54.55	32.86	25.02
DDS	14.74	50.58	45.09	25.86
DDS + CDS	7.15	33.14	25.29	24.96
Ours	28.13	82.43	86.64	**26.94**
Ours + CDS	**6.90**	**32.15**	**24.21**	25.12

In Tab. 2, we present a quantitative comparison. Our method strikes a balance between structure distance and editability. Notably, we observe that the distance is much

Figure 7. Comparison of different editing methods on various objects and styles.

lower when no editing occurs, which is particularly visible in DDS-based methods applied to style editing. Our methods achieving better editing score but with a slightly higher structure distance. In terms of precise structure preservation, when combined with CDS [35], it achieves good preservation of the un-edited areas. Our full model achieves the best performance in the CLIP Similarity metric, demonstrating the effectiveness of our prompt enhancement branch. While CDS excels in preserving unedited regions, it suffers from the inferior editability of DDS-based methods. Optimization-based methods [27, 34] refine the inversion process, achieving excellent performance in structure preservation. However, they struggle with editing methods (like P2P), resulting in limited editability.

5.3. Ablation Studies

In this section, we conduct an ablation experiment to analyze different choices in our SSD. Due to space limitations, we first present a qualitative evaluation in the main text. Please refer to the Supp. for quantitative evaluation.

The effectiveness of cross-trajectory. In Sec. 4.1, we have analyzed the necessity of cross-trajectory. This term make the optimization process more stable and provide the source content regularization in ours design, which is also the key difference between ours and Classifier Score Distillation(CSD) [52]. In Fig. 1 and Fig. 5, we present the comparison of the results with and without cross-trajectory. The results show that the cross-trajectory term can provide the direction of generating high-quality images. Please refer to the supplement for more details.

The effectiveness of prompt-enhancement. The enhancement of the target prompt branch is another key component in our method, which is designed to improve the editability aligned with the target prompt in 2D-image task. In Fig. 7, we observe a clear distinction from DDS in style editing. The results show that the prompt-enhancement term effectively overcomes the challenging from style editing.

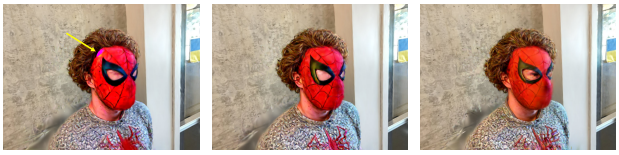

Figure 8. Effect of source latent regularization. In most experiments, the source ID term helps prevent partial gradient explosion. In the left image, the yellow arrow highlights an irregular color. As the weight of the ID term increases, the color becomes more regular, however, the spider on the character's chest is affected.

The effectiveness of ID regularization. ID regularization is designed to ensure stable optimization in 3DGS. In Fig. 8, we compare results with and without ID regularization. The area marked by the yellow arrow highlights its effect in 3D scene editing. However, excessive ID regularization may constrain editing quality by limiting certain attributes, presenting a trade-off in our design.

6. Conclusions, Limitations, and Future Work

In this work, we propose a novel method for text-guided image editing, capable of handling both 3D scenes and 2D images. Our approach is built on a score distillation framework that leverages the powerful priors of diffusion models. For editing tasks, we design an effective optimization strategy that produces high-quality results aligned with target prompts while ensuring stable and consistent optimization.

Our method achieves state-of-the-art performance in both 3D scene and 2D image editing, delivering realistic edits with excellent preservation of the original content. It demonstrates strong adaptability to various editing tasks and target prompts, making it a robust solution for complex scenarios. However, while effective, the optimization process is relatively time-intensive compared to recent one-step methods [50] or few-step approaches [7]. Future work could explore integrating advanced techniques such as LCM [30] or SD-Turbo [42], which show potential for accelerating the optimization process [45].

Acknowledgment: This project is supported by the National Natural Science Foundation of China (62125201, U24B20174, 62102381), the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No.: AISG3-GV-2023-011), the Singapore Ministry of Education AcRF Tier 1 Grant (Grant No.: MSS25C004), the Lee Kong Chian Fellowships.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 6

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 3, 6

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023. 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Dge: Direct gaussian 3d editing by consistent multi-view editing. In *ECCV*, pages 74–92, 2024. 6

[6] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *CVPR*, pages 21476–21485, 2024. 3, 6

[7] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPHA*, 2024. 5, 8

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 3

[9] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *ICCV*, 2023. 3, 6

[10] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *ICCV*, pages 2328–2337, 2023. 1, 2, 3, 4, 6, 7

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 2, 3, 5

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1, 3

[14] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 6

[15] Yukun Huang, Jianan Wang, Yukai Shi, Boshi Tang, Xianbiao Qi, and Lei Zhang. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In *ICLR*, 2023. 3

[16] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *CVPR*, pages 12469–12478, 2024. 3

[17] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *JMLR*, 6(4), 2005. 3

[18] Yutao Jiang, Yang Zhou, Yuan Liang, Wenxi Liu, Jianbo Jiao, Yuhui Quan, and Shengfeng He. Diffuse3d: Wide-angle 3d photography via bilateral diffusion. In *ICCV*, pages 8998–9008, 2023. 3

[19] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *ICLR*, 2024. 2, 6, 7

[20] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. In *ICLR*, 2024. 1, 2, 4

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023. 3, 6

[22] Jiwook Kim, Seonho Lee, Jaeyo Shin, Jiho Choi, and Hyunjung Shim. Dreamcatalyst: Fast and high-quality 3d editing via controlling editability and identity preservation. In *ICLR*, 2025. 2, 3

[23] Hubert Kompanowski and Binh-Son Hua. Dream-in-style: Text-to-3d generation using stylized score distillation. In *3DV*, 2025. 2

[24] Juil Koo, Chanho Park, and Minhyuk Sung. Posterior distillation sampling. In *CVPR*, pages 13352–13361, 2024. 2, 3, 5, 6

[25] Duong H Le, Tuan Pham, Aniruddha Kembhavi, Stephan Mandt, Wei-Chiu Ma, and Jiasen Lu. Preserving identity with variational score for general-purpose 3d editing. *arXiv preprint arXiv:2406.08953*, 2024. 3

[26] Kehan Li, Yanbo Fan, Yang Wu, Zhongqian Sun, Wei Yang, Xiangyang Ji, Li Yuan, and Jie Chen. Learning pseudo 3d guidance for view-consistent texturing with 2d diffusion. In *ECCV*, pages 18–34, 2025. 1

[27] Senmao Li, Joost Van De Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023. 7, 8

[28] Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. In *CVPR*, pages 6743–6752, 2024. 1

[29] Yuqin Lu, Bailin Deng, Zhixuan Zhong, Tianle Zhang, Yuhui Quan, Hongmin Cai, and Shengfeng He. 3d snapshot: Invertible embedding of 3d neural representations in a single image. 46(12):11524–11531, 2024. 3

[30] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 8

[31] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *CVPR*, pages 12663–12673, 2023. 1

[32] 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 TOG*, 38(4):1–14, 2019. 6

[33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 3

[34] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 7, 8

[35] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *CVPR*, pages 9192–9201, 2024. 2, 7, 8

[36] JangHo Park, Gihyun Kwon, and Jong Chul Ye. ED-NeRF: Efficient text-guided editing of 3d scene with latent space nerf. In *ICLR*, 2024. 2, 3

[37] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 3, 4

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

[39] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. In *NeurIPS*, pages 111131–111171, 2024. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3, 6

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 1, 3

[42] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, pages 87–103, 2025. 8

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1, 3, 6

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[45] Feng Tian, Yixuan Li, Yichao Yan, Shanyan Guan, Yanhao Ge, and Xiaokang Yang. Postedit: Posterior sampling for efficient zero-shot image editing. In *ICLR*, 2025. 8

[46] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 6

[47] 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. In *NeurIPS*, 2024. 1, 3

[48] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 7

[49] Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. In *ECCV*, pages 55–71, 2024. 3

[50] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. In *CVPR*, 2024. 8

[51] Yangyang Xu, Shengfeng He, Kwan-Yee K Wong, and Ping Luo. Rigid: Recurrent gan inversion and editing of real face videos and beyond. *IJCV*, 133(6):3437–3455, 2025. 1

[52] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. In *ICLR*, 2024. 1, 2, 3, 4, 5, 8

[53] Yuyang Yu, Bangzhen Liu, Chenxi Zheng, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. Beyond textual constraints: Learning novel diffusion conditions with fewer examples. In *CVPR*, pages 7109–7118, 2024. 3

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[56] Chenxi Zheng, Yihong Lin, Bangzhen Liu, Xuemiao Xu, Yongwei Nie, and Shengfeng He. Recdreamer: Consistent text-to-3d generation via uniform score distillation. In *ICLR*, 2025. 1, 3

[57] Xingchen Zhou, Ying He, F Richard Yu, Jianqiang Li, and You Li. RePaint-NeRF: Nerf editting via semantic masks and diffusion models. In *IJCAI*, 2023. 3

[58] Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. HIFA: High-fidelity text-to-3d generation with advanced diffusion guidance. In *ICLR*, 2024. 1, 3