Target-Balanced Score Distillation

Zhou Xu[1*], Qi Wang[2*], Yuxiao Yang[1], Luyuan Zhang[1], Zhang Liang[2†], Yang Li[1†]

[1]Tsinghua University
[2]Xidian University
xu-z25@mails.tsinghua.edu.cn

Abstract

Score Distillation Sampling (SDS) enables 3D asset generation by distilling priors from pretrained 2D text-to-image diffusion models, but vanilla SDS suffers from over-saturation and over-smoothing. To mitigate this issue, recent variants have incorporated negative prompts. However, these methods face a critical trade-off: limited texture optimization, or significant texture gains with shape distortion. In this work, we first conduct a systematic analysis and reveal that this trade-off is fundamentally governed by the utilization of the negative prompts, where **Target Negative Prompts (TNP)** that embed target information in the negative prompts dramatically enhancing texture realism and fidelity but inducing shape distortions. Informed by this key insight, we introduce the **Target-Balanced Score Distillation (TBSD)**. It formulates generation as a multi-objective optimization problem and introduces an adaptive strategy that effectively resolves the aforementioned trade-off. Extensive experiments demonstrate that TBSD significantly outperforms existing state-of-the-art methods, yielding 3D assets with high-fidelity textures and geometrically accurate shape.

Code — https://github.com/XiaocatMomo/TBSD

1 Introduction

In recent years, text-to-image diffusion models have made remarkable progress, significantly advancing the field of image synthesis (Rombach et al. 2022). These models (Schuhmann et al. 2022), typically trained on large-scale datasets and powered by massive parameter capacities, are capable of generating highly realistic and detail-rich images from simple textual descriptions. However, achieving comparable generation quality remains technically challenging in scenarios with relatively scarce training data, such as 3D generation. The high demand for 3D objects in downstream tasks like computer graphics and robotics has led to the development of methods such as Score Distillation Sampling (SDS) (Poole et al. 2022), which enables knowledge transfer from pre-trained 2D diffusion models to 3D content. The core principle of SDS is to optimize 3D representations (Kerbl et al. 2023; Mildenhall et al. 2021; Müller et al. 2022) such that their rendered images approach high-probability density regions under text conditions, with supervision provided by pre-trained 2D diffusion models. Due to this formulation, these SDS-based methods (Tang et al. 2024; Chen et al. 2023; Yi et al. 2024; Zhu, Zhuang, and Koyejo 2024; Wang et al. 2023; Lukoianov et al. 2024; Liang et al. 2024; Huang et al. 2024a) do not require 3D data for training and can generate 3D results from various text prompts.

Despite its success, existing studies (Liang et al. 2024) have noted that SDS exhibits an averaging effect during generation, leading to color over-saturation and overly smooth textures. Inspired by the success of negative prompts in 2D diffusion models, negative prompts have achieved success in guiding models by specifying "content not to generate" (Ban et al. 2025; Armandpour et al. 2023). Thus, some studies have attempted to introduce negative prompts into score distillation to alleviate the above issues. Existing methods typically leverage negative prompts either as approximate domain correction terms during early training (Katzir et al. 2023; Yu et al. 2023; McAllister et al. 2024), as an auxiliary acceleration mechanism (Yu et al. 2023), or to guide texture optimization via source distribution estimation (Katzir et al. 2023; McAllister et al. 2024). However, these methods suffer from a critical trade-off: either texture optimization remains limited, or significant improvements in texture fidelity come at the cost of shape distortion.

In this paper, we first conduct a comprehensive review of existing SDS variants that incorporate negative prompts and reveal that this trade-off is fundamentally governed by the utilization of the negative prompts. *Interestingly, Our analysis finds that **Target Negative Prompts (TNP)**, which embed only the target information in the negative prompts, can dramatically enhancing the realism and fidelity of generated textures.* However, this overly focused guidance often leads to loss of target information and subsequent global shape distortions, which also reported in prior work (McAllister et al. 2024).

To address this limitation, we further propose a novel score distillation framework termed **Target-Balanced Score Distillation (TBSD)**. Our method formulates SDS as a multi-objective optimization problem: one objective (shape guidance) is optimized using the classifier-free guidance term of standard SDS branch, while the other objective (tex-

*These authors contributed equally.
†Corresponding author.

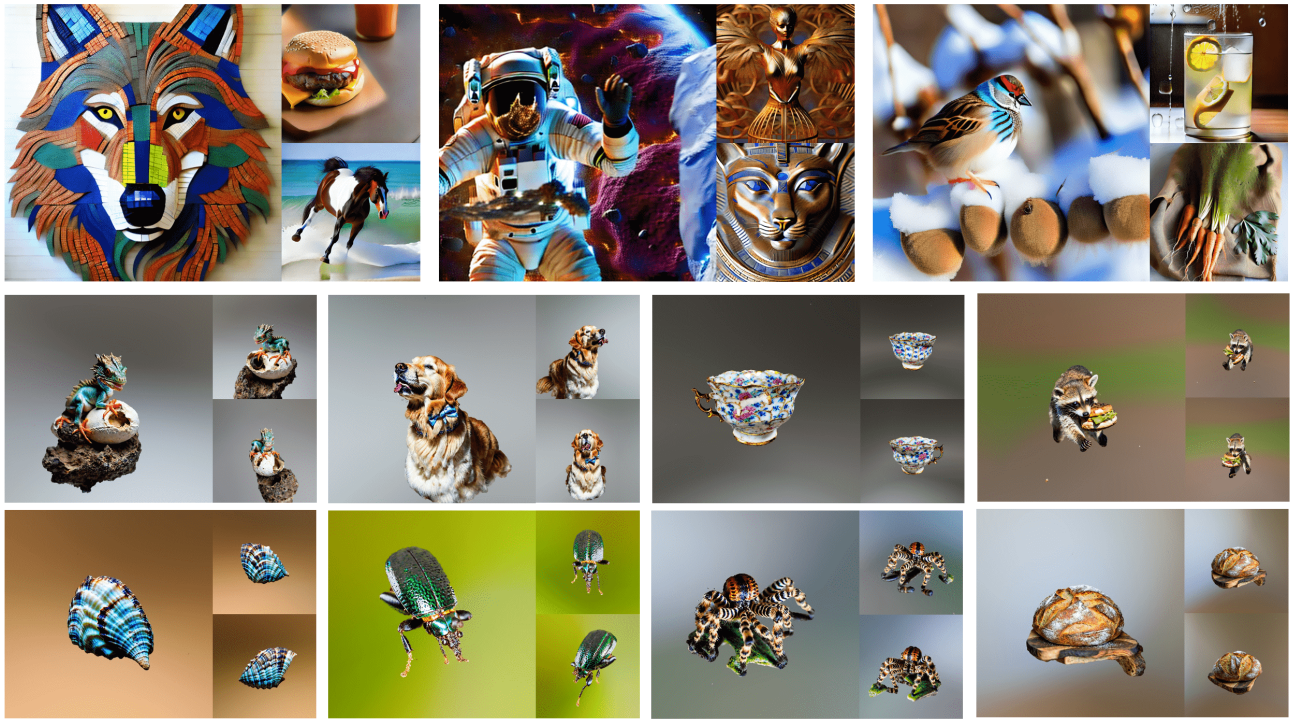

Figure 1: Results obtained with our Target-Balanced Score Distillation (TBSD). Top: a gallery of images optimized with TBSD. Bottom: two rows NeRFs generated by TBSD (other examples are included in the supplementary material).

ture enhancement) is guided by the TNP branch. It introduces an adaptive strategy ensuring that TBSD initially focuses more on shape optimization, enabling the model to generate an accurate shape in the early stages. As training progresses, the optimization gradually shifts toward the texture objective, with TBSD maintaining a good balance between shape and texture to allow texture quality to improve while preserving essential target information. This dynamic balancing strategy effectively resolves the aforementioned trade-off. As shown in Figure 1, our TBSD method can produce 3D assets that preserve geometric correctness while exhibiting extremely realistic and detailed textures.

Our key contributions are as follows:

- We conduct a detailed analysis of existing SDS variants that incorporate negative prompts. We identify and explain the mechanism by which Target Negative Prompts (TNP) enhance texture fidelity while potentially shape distortion.
- To mitigate the shape distortion introduced by TNP, we propose Target-Balanced Score Distillation (TBSD), a dynamic optimization framework that simultaneously balances shape preservation and texture realism.
- Extensive experiments across both 2D and 3D generation demonstrate that TBSD outperforms state-of-the-art methods, achieving the generation with high-fidelity textures and geometrically accurate shape.

2 Related Works

Text-to-3D Generation. DreamFields (Jain et al. 2022) guides NeRF (Mildenhall et al. 2021) optimization using the pre-trained CLIP (Radford et al. 2021a) model. DreamFusion (Poole et al. 2022) proposes Score Distillation Sampling (SDS), which enables generation by leveraging 2D diffusion models. Subsequent studies (Li et al. 2023; Chen et al. 2024, 2023; Lin et al. 2023; Yi et al. 2024; Huang et al. 2024c; Qiu et al. 2024; Liu et al. 2024; Huang et al. 2024b) have improved SDS-based generation from multiple aspects. Given the high reliance on SDS, addressing its issues is crucial. SDS suffers from problems such as over-smoothing and over-saturation. Additionally, it requires a large conditional guidance scale, which further exacerbates over-saturation. Relevant improvements include: HiFA (Zhu, Zhuang, and Koyejo 2024) enhances performance through an iterative process and the introduction of additional loss terms; ProlificDreamer (Wang et al. 2023) proposes Variational Score Distillation (VSD) to alleviate issues like over-saturation; Consistent3D (Wu et al. 2024) presents Consistency Distillation Sampling to reduce over-smoothing and other problems; LucidDreamer (Liang et al. 2024) introduces Interval Score Matching (ISM) to optimize results and recent studies (Katzir et al. 2023; Yu et al. 2023; McAllister et al. 2024; Huang et al. 2024a; Yang et al. 2023; Alldieck, Kolotouros, and Sminchisescu 2024; Yan, Chen, and Wang 2025; Zhuo et al. 2024) have also proposed other improvement methods

for SDS. However, these methods have limitations, such as imposing significant computational burdens or yielding limited performance improvements.

SDS with Negative Prompts. Among SDS-related methods, NFSD (Katzir et al. 2023) uses negative prompts to assist in extracting δ_{D}. For time steps t>200, it approximates δ_{D} by calculating the difference between noise under the null condition and noise under the negative text condition, and constructs a noise-free loss to improve generation quality without the need for a large guidance scale. CSD (Yu et al. 2023) leverages negative prompts to achieve dual-objective optimization, driving the model to approach the target prompt and move away from negative states, thereby accelerating training and improving quality. Bridge (McAllister et al. 2024) uses negative prompts to represent the starting point of the Schrodinger bridge, alleviating the problem of inaccurate optimization starting points and enhancing the clarity of generated results. However, these methods fail to fully utilize the potential of negative text, with limited effectiveness or other issues such as shape distortion. This paper focuses on fully exploiting the advantages of negative prompts.

3 Analysis of Negative Prompts-based SDS Variants

3.1 Revisiting SDS and its variants

Score Distillation Sampling (SDS). SDS leverages a pre-trained text-to-image diffusion model ϕ to guide the 3D representation parameterized by θ. Specifically, for a given camera pose π, $x = g(\theta;\pi)$ represents the image rendered by the differentiable rendering function g. The SDS loss ensures that images obtained via g from any viewpoint are aligned with the prompt y, and its form is as follows:

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbf{E}_{t,\epsilon,\pi}[w(t)(\epsilon_\phi(\mathbf{x}_t; y, t) - \epsilon)\frac{\partial x}{\partial \theta}] \quad (1)$$

$w(t)$ serves as a weighting function, with $\mathbf{x}_t$ standing for a noisy variant of $\mathbf{x}$, which we denote as $\delta_{\mathrm{SDS}} := \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - \epsilon$. In practical implementation, Classifier-free guidance (CFG) is utilized in SDS. Specifically, δ_{SDS} with CFG is expressed as:

$$\begin{aligned}\delta_{\mathrm{SDS}} =& \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - \epsilon + \\ & s \cdot \underbrace{\left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right)\right)}_{\delta^{\mathrm{cls}}}\end{aligned} \quad (2)$$

Noise Free Score Distillation (NFSD).

$$\begin{aligned}\delta_{\mathrm{NFSD}} = & s \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right)\right) + \\ & \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - (t < 0.2) \cdot \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{neg}}, t\right)\right)\end{aligned} \quad (3)$$

NFSD leverages the negative prompt terms to approximate domain correction, enabling noise-free score distillation.

Classifier Score Distillation (CSD).

$$\begin{aligned}\delta_{\mathrm{CSD}} =& w_1 \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right)\right) + \\ & w_2 \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{neg}}, t\right)\right)\end{aligned} \quad (4)$$

CSD gradually reduces w_2 mitigates the negative effects of the latter to improve texture quality, fidelity, and alignment with the target prompt.

Bridge.

$$\delta_{\mathrm{Bridge}} = w \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tnp}}, t\right)\right) \quad (5)$$

Bridge improves source distribution estimation by using negative prompts to describe image corruptions.

The trade-off of these variants. In practice, both NFSD and CSD outperform SDS, but the improvement is limited. Bridge achieves a significantly larger improvement over SDS and can generate rich, clear textures and realistic colors, though it introduces shape distortions. This reflects a clear trade-off: texture optimization either remains limited, or significant improvements in texture fidelity come at the cost of shape distortion. Although these variants analyze SDS from different perspectives and use negative prompts for optimization, they can all be transformed into the same structure. NFSD and CSD share an identical structure, to better understand the differences between Bridge, NFSD, and CSD, we reformulate Bridge into a structure consistent with the latter two, as in Equation 6.

$$\begin{aligned}\delta_{\mathrm{Bridge}} =& w \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right)\right) + \\ & w \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tnp}}, t\right)\right)\end{aligned} \quad (6)$$

By comparing Equations 3, 4, and 6, our comparative experimental analysis reveals that this trade-off is fundamentally governed by the utilization of negative prompt. Details of the comparative analysis are provided in the supplementary material. Specifically, Bridge adopts **Target negative prompts (TNP)**, which contain target information. The form of y_{tnp} is:

$$y_{\mathrm{tnp}} = y_{\mathrm{tgt}} + y_{\mathrm{neg}} \quad (7)$$

For example, if y_{tgt} is "An ice cream sundae" and y_{neg} is ", oversaturated, smooth, pixelated...", then y_{tnp} is "An ice cream sundae, oversaturated, smooth, pixelated...". To avoid ambiguity, negative prompts lacking target information will be uniformly referred to as "General Negative Prompts (GNP)" hereafter.

3.2 The Impact of Target Negative Prompts

SDS optimizes 3D parameters indirectly by refining rendered images of 3D models, which is essentially an optimization of 2D images. In order to analyze why TNP can promote the generation of clear colors while causing shape distortions, we visualize the latter term of this structure (i.e., δ_{post}) in Equations 3, 4, and 6, along with classifier-free guidance terms (i.e., δ^{cls}) and Bridge gradients (i.e., δ_{Bridge}). The form of δ_{post} is:

$$\begin{aligned}\delta_{\mathrm{post}}^{\mathrm{gnp}} &= \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{gnp}}, t\right) \\ \delta_{\mathrm{post}}^{\mathrm{tnp}} &= \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\mathrm{tnp}}, t\right)\end{aligned} \quad (8)$$

As shown in Figure 2, $\delta_{\mathrm{post}}^{\mathrm{gnp}}$ acts globally on the whole image, $\delta_{\mathrm{post}}^{\mathrm{tnp}}$ acts more focused on the target information in the image and can more accurately suppress negative states in the target region, which is referred to as the **focused suppression effect**. Furthermore, a comparison between δ_{Bridge} and δ^{cls} reveals that the target information in δ_{Bridge} is relatively vague, which reflects the **target information loss** caused by TNP. This loss impairs the concentration of gradients on target regions, thereby leading to shape distortion.

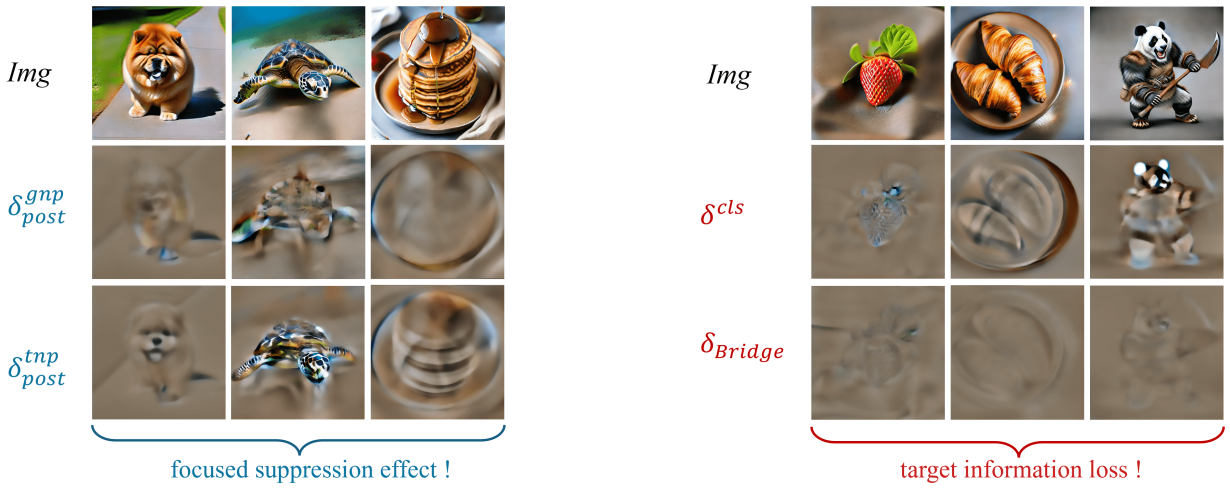

Figure 2: Visualization of $\delta^{\text{gnp}}_{\text{post}}$, $\delta^{\text{tnp}}_{\text{post}}$, δ^{cls} and δ_{Bridge}. Top-row images are generated by TBSD. Visualization is done by decoding each δ with the VAE decoder of Stable Diffusion.

3.3 Analysis of these Variants

From the perspective of negative prompts, NFSD and CSD employ GNP, whose suppression of negative states fails to focus on the target itself, limiting texture improvement. In contrast, Bridge uses TNP, which accurately focuses on and suppresses negative states in target regions to effectively optimize textures but causes target information loss, thereby leading to shape distortion. To achieve generated results with clear textures and stable shapes, we need to leverage TNP for texture optimization while enhancing protection of target information.

4 Method

From the aforementioned analysis, TNP's core focusing suppression effect enhances the realism and detail clarity of texture generation by accurately suppressing negative states in the target region. However, it also leads to shape distortion of the generated object due to the loss of target information.

To address this issue, an intuitive solution is to inject target information into the optimization gradient to supplement geometric constraints. The classifier-free guidance term in SDS maximizes the difference in target information through the gap between the target prompt and empty prompt, thereby effectively providing target information. Based on this, we introduce this guidance term into the optimization framework of TNP and control the injection intensity of target information by adjusting its coefficient a (see Equation 9).

$$\begin{aligned}\delta_x(\mathbf{x}_t; y, t) =& \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tnp}}, t\right) \\ &+ a \cdot \delta^{\text{cls}}\end{aligned} \tag{9}$$

We inject target information of varying intensities into the TNP optimization process. As shown in Figure 3 (a), when the value of a is small, the generated results exhibit clear textures but distorted shapes. when a is too large, the shape becomes accurate but the texture appears overly saturated in color. This indicates that a fixed coefficient a fails to achieve a proper balance between texture and shape.

To address this problem, we propose **Target-Balanced Score Distillation (TBSD)**, inspired by Multiple-Gradient Descent Algorithm (MGDA) (Désidéri 2012). TBSD formulates the generation task as a multi-objective optimization problem, with shape (δ_s) and texture (δ_t) as the two objectives. Since the classifier-free guidance term provides sufficient target information and TNP, with its focus-suppression effect, enables realistic and vivid texture optimization, the optimization objective is defined as in Equation 10.

$$\begin{aligned}\delta_s &= \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) \\ \delta_t &= \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tnp}}, t\right)\end{aligned} \tag{10}$$

The cosine similarity between the texture and shape objectives is always positive, indicating that they share a similar optimization direction. In this case, MGDA tends to favor the objective with a smaller gradient norm if the other is significantly larger. Based on this property and the progressive nature of 3D generation, we observe that a good initial shape helps guide later optimization and improves final quality. Therefore, we introduce a dynamic weighting factor for the texture objective, as shown in Equation 11.

$$\begin{aligned}\delta_{td} &= factor(t) \cdot \delta_t \\ factor(t) &= max(\alpha \cdot (1 - t/\beta), \gamma)\end{aligned} \tag{11}$$

This factor starts with a large value α, decreases over time with rate $\frac{\alpha}{\beta}$, and is bounded below by γ. t is iteration step.

The factor ensures that TBSD initially focuses more on shape optimization, enabling the model to generate an accurate shape in the early stages. As training progresses, the optimization gradually shifts toward the texture objective. During this process, TBSD maintains a good balance between shape and texture, allowing texture quality to improve

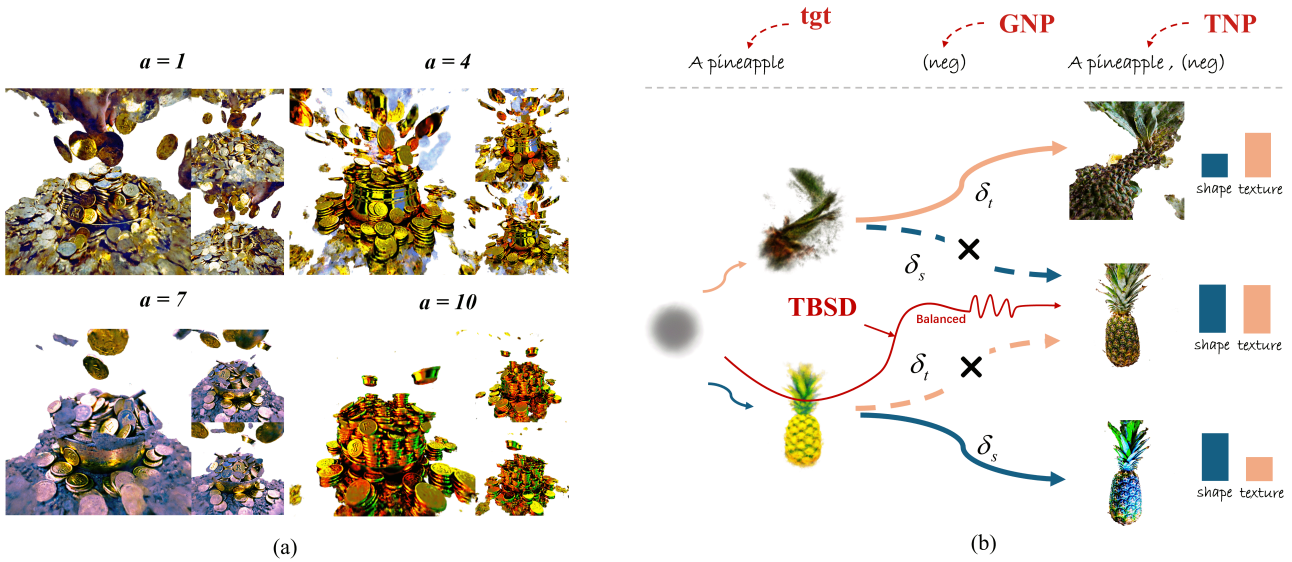

Figure 3: (a) Result images with varying levels of shape information controlled by coefficient a. (b) An overview of the proposed TBSD. Solid lines indicate reachable optimization paths, while dashed lines indicate unreachable ones. The prompts used for Figs. (a), (b) are "A cauldron full of gold coins", "A pineapple", respectively.

while preserving essential target information. Eventually, it reaches an optimal optimization point for both shape and texture. Based on this, the formula of TBSD is shown in Equation 12.

$$\delta_x^{\text{TBSD}} = \mu \cdot \delta_s + (1-\mu) \cdot \delta_{td} \tag{12}$$

Following (Sener and Koltun 2018), the multi-objectives optimization problem can be defined as:

$$\min_{\mu, 1-\mu \in [0,1]} \|\mu \cdot \delta_s + (1-\mu) \cdot \delta_{td})\|_2^2 \tag{13}$$

This is a univariate quadratic function of μ, and considering the value range of μ, its solution is:

$$\mu = \min\left[\max\left[0, \frac{(\delta_{td} - \delta_s)^T \delta_{td}}{\|\delta_{td} - \delta_s\|_2^2}\right], 1\right] \tag{14}$$

As a result, TBSD effectively resolves the aforementioned trade-off and can achieve the generation of 3D objects with accurate shapes and high-fidelity textures. Moreover, TBSD can be directly applied to image generation. TBSD is shown in Figure 3 (b).

5 Experiments

We rigorously evaluate the proposed method on both 2D and 3D generation tasks employing Score Distillation Sampling (SDS), conducting comprehensive benchmarks against SDS and domain-specific state-of-the-art baselines. Extended quantitative analyses, qualitative comparisons, and ablation studies are detailed in the supplementary material.

5.1 Implementation Details

We implement text-based 3D generation using the threestudio (Liu et al. 2023) framework. Unless specified otherwise, all 3D models are trained with the AdamW (Loshchilov and Hutter 2017) optimizer for 20,000 iterations, with a learning rate of 0.01. Our initial rendering resolution is gradually increased from 64×64 to 256×256. Consistent with NFSD (Katzir et al. 2023) and Bridge (McAllister et al. 2024), implicit volumes use an object-centered initialization strategy (Qian et al. 2023; Wang et al. 2023). All experiments employ the pre-trained text-to-image diffusion model Stable Diffusion 2.1-base (Rombach et al. 2022). Additional implementation details are provided in the supplementary material.

5.2 Text-to-3D Generation

Qualitative comparisons. Our comparison with recent methods is presented in Figure 4. Following the same experimental protocol as NFSD (Katzir et al. 2023), ProlificDreamer (Wang et al. 2023), and SDI (Lukoianov et al. 2024), we compare the 3D generation quality of this study with previous works. Selected baseline methods include DreamFusion (Poole et al. 2022), NFSD (Katzir et al. 2023), ProlificDreamer (Wang et al. 2023), ISM (Liang et al. 2024), HiFA (Zhu, Zhuang, and Koyejo 2024), and SDI (Lukoianov et al. 2024), with all comparison results from the original authors. It can be observed that our method achieves comparable or better results without additional computation from model training or multi-step optimization. More comparisons of text-to-3D experimental results are available in supplementary material.

Quantitative Results. We perform quantitative evaluation of generation quality following (Poole et al. 2022; Yu et al.

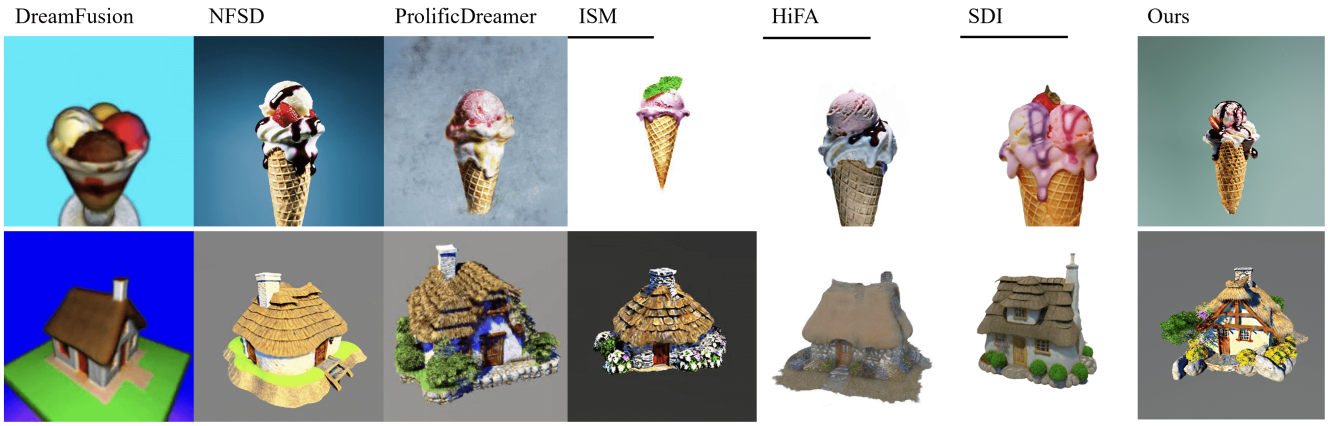

Figure 4: 3D generation comparison with other methods, using their reported results. The prompts employed are "An ice cream sundae" and "A 3D model of an adorable cottage with a thatched roof".

Method	CLIP Score(↑)	User Preference (%) (↑)
SDS	29.81	2.61
VSD	33.31	11.52
HIFA	32.80	7.48
SDI	33.47	13.44
CSD	32.05	7.73
NFSD	31.89	6.56
Bridge	32.36	8.29
TBSD	**33.59**	**42.37**

Table 1: The quantitative comparison of 3D generation between our method and others.

Figure 5: Ablation study of proposed improvements. First and second row results use prompts "An ice cream sundae" and "Bagel filled with cream cheese and lox", respectively.

2023; Lukoianov et al. 2024). The Clip scores in Table 1 are computed using torchmetrics (Detlefsen et al. 2022) and the ViT-B/32 model (Radford et al. 2021b). We test 50 views under 43 prompts (Lukoianov et al. 2024). For fairness, multi-stage methods only run the first stage. All results are compared based on threestudio reproductions. It can be seen that TBSD outperforms SDS in quality and the current state-of-the-art method SDI (Lukoianov et al. 2024), and notably, this improvement is achieved without multi-step optimization. To further assess the perceptual quality of generated results, we conduct a user study comparing our approach with baselines. As shown in Table 1, our method outperforms the baselines in the user study. Details of the user study are presented in the supplementary material.

5.3 Ablation Studies and Analysis

Ablation study of proposed improvements. Figure 5 shows ablation study of our proposed improvements. It can be observed that using TNP significantly improves generation quality, resulting in richer textures and more realistic, vivid colors. Moreover, TBSD effectively resolves the trade-off between shape and texture, ultimately producing 3D outputs with accurate shape and high-fidelity, realistic textures.

Negative Prompts Ablation. We explore the impact of negative prompts in TNP on generation quality. We generate six groups of different negative prompts descriptions via GPT-4 for comparative evaluation. In the experiment, all other hyperparameters are kept unchanged, with only the negative text following the target text replaced. Specific results are shown in Figure 6. No significant differences are observed across the ablation experiments with various negative prompts in TNP, indicating that the design of the TNP prompt format is more critical than the specific wording of negative prompts used. Details of negative prompt variants are provided in supplementary material.

Hyperparameter Ablation. We investigate the impact of two tunable parameters in Equation 11 : α and β. The parameter α controls the strength of TBSD's focus bias toward shape optimization, while β regulates the delay in shifting this focus toward texture. As shown in Figure 7, large values for both α and β result in a strong and persistent shape bias, with a slow transition toward texture optimization. This leads to generated 3D assets with oversaturated colors and

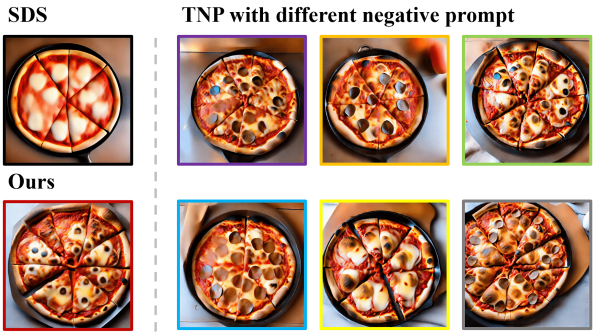

Figure 6: Ablation study on negative prompts in TNP. These images are generated by TBSD using TNP with varying negative prompts. The prompt is "pizza sitting on top of a pan on a table".

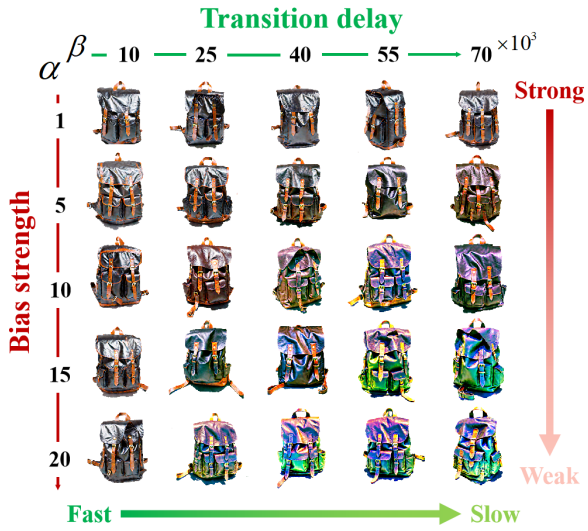

Figure 7: Ablation result of the α and β. The prompt is "Photograph of a black leather backpack".

insufficient texture details. Only with a proper combination of α and β can TBSD gradually and effectively shift its optimization focus from shape to texture, leading to realistic 3D outputs with sharp and detailed textures.

5.4 Text-to-img Generation

To verify our analysis of existing SDS variants and the proposed method, we also conduct text-to-image generation experiments by optimizing images in the Stable Diffusion latent space consistent with previous studies. Compared with text-to-3D tasks, where factors such as initialization strategies, 3D representation methods, and 2D prior models can significantly affect the final results, image generation is less influenced by such confounding variables. To illustrate the advantages of TBSD in 2D experiments, we selected some MS-COCO prompts (Lin et al. 2014) for comparative display. Comparative methods include SDS, VSD, as well as

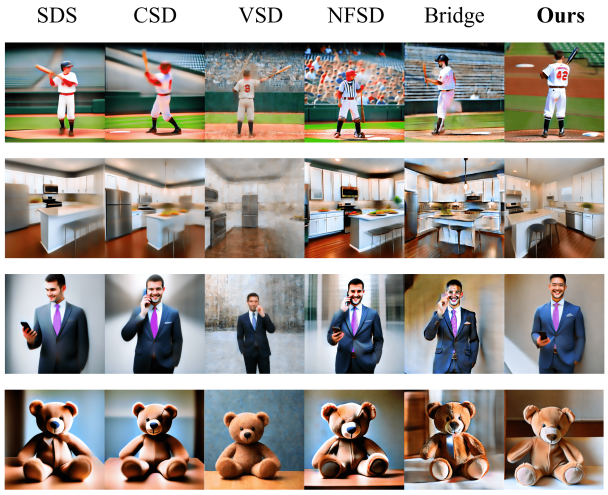

Figure 8: Text-to-image generation results using COCO Captions. We compare various score distillation methods for image generation with COCO captions, where images are optimized from random initializations.

NFSD, CSD, and Bridge that take negative prompts as the core. For each prompt, we randomly initialize the noise and then optimize using score distillation gradients. Figure 8 shows generation examples of different score distillation methods, SDS and CSD exhibit both over-saturation and excessive smoothness. NFSD shows an improved texture, but still tends to be saturated. VSD and Bridge generate samples that are closest to real-world effects, but VSD suffers from global blurriness. Although Bridge exhibits relatively rich details, it has slight saturation issues and shape distortion. In contrast, our method achieves realistic colors, rich textures, and stable shapes, thereby outperforming these approaches. More details of 2D experiments can be seen in the supplementary material.

6 Conclusion

In this paper, we first systematically analyze the trade-off between texture fidelity and shape accuracy in SDS methods utilizing negative prompts and reveal that Target Negative Prompts (TNP) enhance texture realism by suppressing negative states in the target region, but this strong focus can lead to the loss of target information, resulting in shape distortions. To address this, we introduce Target-Balanced Score Distillation (TBSD), a novel multi-objective framework that adaptively balances shape and texture optimization by progressively shifting focus from global shape to detailed texture refinement. Extensive experiments demonstrate that TBSD effectively resolves the trade-off, producing 3D assets with both accurate shape and rich textures, outperforming existing approaches. We believe that our findings can offer a novel insight for the SDS community.

Acknowledgments

This work is supported by the Shenzhen Science and Technology Project under Grant KJZD20240903103210014.

References

Alldieck, T.; Kolotouros, N.; and Sminchisescu, C. 2024. Score distillation sampling with learned manifold corrective. In *European Conference on Computer Vision*, 1–18. Springer.

Armandpour, M.; Sadeghian, A.; Zheng, H.; Sadeghian, A.; and Zhou, M. 2023. Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond. arXiv:2304.04968.

Ban, Y.; Wang, R.; Zhou, T.; Cheng, M.; Gong, B.; and Hsieh, C.-J. 2025. Understanding the Impact of Negative Prompts: When and How Do They Take Effect? In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 190–206. Cham: Springer Nature Switzerland. ISBN 978-3-031-73024-5.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22246–22256.

Chen, Z.; Wang, F.; Wang, Y.; and Liu, H. 2024. Text-to-3D using Gaussian Splatting. arXiv:2309.16585.

Désidéri, J.-A. 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6): 313–318.

Detlefsen, N.; Borovec, J.; Schock, J.; Harsh, A.; Koker, T.; Liello, L.; Stancl, D.; Quan, C.; Grechkin, M.; and Falcon, W. 2022. TorchMetrics-Measuring Reproducibility in PyTorch. *URL https://github. com/Lightning-AI/Torchmetrics*.

Huang, S.; Sun, S.; Wang, Z.; Qin, X.; Xiong, Y.; Zhang, Y.; Wan, P.; Zhang, D.; and Jia, J. 2024a. PlacidDreamer: Advancing Harmony in Text-to-3D Generation. arXiv:2407.13976.

Huang, X.; Shao, R.; Zhang, Q.; Zhang, H.; Feng, Y.; Liu, Y.; and Wang, Q. 2024b. HumanNorm: Learning Normal Diffusion Model for High-quality and Realistic 3D Human Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4568–4577.

Huang, Y.; Wang, J.; Shi, Y.; Tang, B.; Qi, X.; and Zhang, L. 2024c. DreamTime: An Improved Optimization Strategy for Diffusion-Guided 3D Generation. arXiv:2306.12422.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-Shot Text-Guided Object Generation With Dream Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 867–876.

Katzir, O.; Patashnik, O.; Cohen-Or, D.; and Lischinski, D. 2023. Noise-Free Score Distillation. arXiv:2310.17590.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. arXiv:2308.04079.

Li, W.; Chen, R.; Chen, X.; and Tan, P. 2023. SweetDreamer: Aligning Geometric Priors in 2D Diffusion for Consistent Text-to-3D. arXiv:2310.02596.

Liang, Y.; Yang, X.; Lin, J.; Li, H.; Xu, X.; and Chen, Y. 2024. LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6526.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 300–309.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, X.; Zhan, X.; Tang, J.; Shan, Y.; Zeng, G.; Lin, D.; Liu, X.; and Liu, Z. 2024. HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6646–6657.

Liu, Y.-T.; Guo, Y.-C.; Voleti, V.; Shao, R.; Chen, C.-H.; Luo, G.; Zou, Z.; Wang, C.; Laforte, C.; Cao, Y.-P.; et al. 2023. Threestudio: A modular framework for diffusion-guided 3d generation. *cg. cs. tsinghua. edu. cn*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lukoianov, A.; de Ocáriz Borde, H. S.; Greenewald, K.; Guizilini, V. C.; Bagautdinov, T.; Sitzmann, V.; and Solomon, J. 2024. Score Distillation via Reparametrized DDIM. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 26011–26044. Curran Associates, Inc.

McAllister, D.; Ge, S.; Huang, J.-B.; Jacobs, D. W.; Efros, A. A.; Holynski, A.; and Kanazawa, A. 2024. Rethinking Score Distillation as a Bridge Between Image Distributions. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 33779–33804. Curran Associates, Inc.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4).

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. arXiv:2209.14988.

Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Lee, H.-Y.; Skorokhodov, I.; Wonka, P.; Tulyakov, S.; et al. 2023. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*.

Qiu, L.; Chen, G.; Gu, X.; Zuo, Q.; Xu, M.; Wu, Y.; Yuan, W.; Dong, Z.; Bo, L.; and Han, X. 2024. RichDreamer: A Generalizable Normal-Depth Diffusion Model for Detail Richness in Text-to-3D. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9914–9925.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021a. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 25278–25294. Curran Associates, Inc.

Sener, O.; and Koltun, V. 2018. Multi-Task Learning as Multi-Objective Optimization. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2024. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. arXiv:2309.16653.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. arXiv:2305.16213.

Wu, Z.; Zhou, P.; Yi, X.; Yuan, X.; and Zhang, H. 2024. 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 (CVPR)*, 9892–9902.

Yan, R.; Chen, Y.; and Wang, X. 2025. Consistent flow distillation for text-to-3d generation. *arXiv preprint arXiv:2501.05445*.

Yang, X.; Chen, Y.; Chen, C.; Zhang, C.; Xu, Y.; Yang, X.; Liu, F.; and Lin, G. 2023. Learn to Optimize Denoising Scores for 3D Generation: A Unified and Improved Diffusion Prior on NeRF and 3D Gaussian Splatting. arXiv:2312.04820.

Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024. GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6796–6807.

Yu, X.; Guo, Y.-C.; Li, Y.; Liang, D.; Zhang, S.-H.; and Qi, X. 2023. Text-to-3D with Classifier Score Distillation. arXiv:2310.19415.

Zhu, J.; Zhuang, P.; and Koyejo, S. 2024. HiFA: High-fidelity Text-to-3D Generation with Advanced Diffusion Guidance. arXiv:2305.18766.

Zhuo, W.; Ma, F.; Fan, H.; and Yang, Y. 2024. Vividdreamer: invariant score distillation for hyper-realistic text-to-3d generation. In *European Conference on Computer Vision*, 122–139. Springer.