RewardSDS: Aligning Score Distillation via Reward-Weighted Sampling

Supplementary Material

A. Additional Quantitative Results

Complementary qualitative results corresponding to the experiments reported in this section are presented in Sec. B.

A.1. Quantitative Results for GPTEval3D

Several recent works address common issues in SDS, such as over-smoothing and color saturation. To assess the effectiveness of RewardSDS in comparison to these approaches, we evaluated it using the GPTEval3D benchmark [22], which examines text-to-3D methods using GPT-4o [5] as an LLM-based judge across a diverse set of 110 prompts. This benchmark focuses on fine-grained textual and geometric alignment, using evaluation prompts and multi-view renderings and normals.

Table 2. Extended comparison of RewardSDS and other text-to-3D methods on the GPTEval3D benchmark [22], using additional evaluation metrics.

Method	CLIP↑	ImageReward↑	Aesthetic Score↑
RewardSDS	**25.7**	**-0.37**	**5.96**
ProlificDreamer	25.2	-0.51	5.18
MVDream	24.2	-0.57	5.88
LatentNerf	25.8	-0.42	5.15
Fantasia3D	22.2	-1.39	4.55
DreamFusion	22.4	-1.51	4.24

Our method outperforms prior approaches on most of these key aspects, as shown in Tab. 1, which is based on the current benchmark leaderboard [1]. In addition to the main metrics, we report extended results using complementary metrics such as CLIP similarity, ImageReward score, and Aesthetic score in Tab. 2. These metrics, also used in the main paper, provide a broader perspective on image-text alignment and perceptual quality.

Table 3. Comparison of text-guided 3D generation using standard SDS and VSD (with stable-diffusion-2.1-base), with and without reward-based sampling, over 30 randomly sampled prompts from DreamFusion gallery.

Method	CLIPScore↑	Aesthetic↑	LLM Grader↑
SDS	21.49	5.34	3.85
RewardSDS	**22.87**	**5.53**	**4.29**
VSD	21.38	5.14	3.67
RewardVSD	**22.03**	**5.41**	**4.11**

[1] https://huggingface.co/spaces/GPTEval3D/Leaderboard_dev

A.2. Quantitative Results for 3D-GS Generation

To complement our main results in the Experiments section, we provide additional quantitative comparisons for 3D generation with the 3DGS backbone [19]. We evaluated our reward-based method with both SDS and VSD, with stable-diffusion-2.1-base as the 2D prior, and ImageReward is the reward model. Tab. 3 provides a comparison over 30 randomly sampled prompts from DreamFusion gallery, with each score averaged over 10 randomly rendered views. These results further validate the effectiveness of RewardSDS when applied to a 3DGS-based representation.

Table 4. Comparison of SDS-Bridge and ConsistentFlowDistillation (CFD) with and without RewardSDS on zero-shot text-to-image generation.

Method	CLIP↑	Aesthetic↑	LLM-G↑
SDS-Bridge	27.01	5.34	6.44
RewardSDS-Bridge	**27.94**	**5.58**	**7.31**
CFD	28.71	5.25	6.29
RewardCFD	**29.43**	**5.46**	**7.41**

A.3. Quantitative Results for Plug-and-Play Integration

As discussed in the Method section, one of the main advantages of RewardSDS is its plug-and-play compatibility with existing SDS-based optimization frameworks. To demonstrate this, we apply it on top of two recent methods: SDS-Bridge [9] and Consistent Noise Distillation (CFD) [23]. While our primary focus is text-to-3D generation, these experiments are conducted in a text-to-2D setting to reduce computational cost. Quantitative results in Tab. 4 show consistent improvements across key metrics with RewardSDS is used as a drop-in enhancement.

A.4. Evaluation on Counting and Spatial Relationship Prompts

While aggregate metrics such as CLIP, Aesthetic, or LLM-G provide useful quantitative signals, they often fail to capture fine-grained semantic accuracy—particularly for prompts that involve object counting or explicit spatial relations. Moreover, broad user studies can be difficult to interpret when evaluating such specific compositional behaviors. To better assess these aspects, we conducted a focused evaluation using the counting and positional categories of the DrawBench benchmark. We compared 2D

Table 1. Comparison of RewardSDS to baseline text-to-3D methods and on the GPTEval3D benchmark [22]. RewardSDS achieves the highest scores across most evaluation aspects.

Method	Text-Asset Alignment↑	3D Plausibility↑	Text-Geometry Alignment↑	Texture Details↑	Geometry Details↑
RewardSDS	1284.33	**1247.93**	**1340.44**	**1383.77**	**1371.21**
DreamCraft3D [18]	**1336.67**	1224.95	1318.51	1373.89	1288.89
RichDreamer [13]	1294.85	1225.28	1259.99	1355.95	1251.28
MVDream [17]	1270.55	1147.47	1250.57	1324.89	1255.46
ProlificDreamer [21]	1261.8	1058.73	1151.99	1246.37	1180.56
LatentNerf [11]	1222.33	1144.84	1156.7	1180.47	1160.77
OpenLRM [4]	1202.2	1078.78	1188.83	1211.98	1173.86
Fantasia3D [1]	1067.9	891.87	1005.99	1109.29	1027.48
DreamFusion [12]	1000.0	1000.0	1000.0	1000.0	1000.0

generations from SDS and RewardSDS (trained with ImageReward) across all 36 prompts in these categories, which include examples such as "Three cars on the street" and "A banana on the left of an apple."

Table 5. Comparison of SDS and RewardSDS on counting and positional prompts from the DrawBench benchmark [16]. RewardSDS shows improved adherence to quantitative and spatial relationships, while also achieving higher general alignment metrics.

Method	CLIP↑	Aesthetic↑	LLM-G↑	**Counting (%)↑**	**Positional (%)↑**
SDS	26.37	5.46	6.86	29.4	31.6
RewardSDS	**27.92**	**5.70**	**7.08**	**76.4**	**57.8**

For each prompt, we examined whether (i) the number of depicted objects matched the specified count, and (ii) the positional or geometrical relationships were correctly satisfied. The quantitative results are summarized in Table 5, showing that RewardSDS markedly improves adherence to both counting and spatial constraints while also improving general image–prompt alignment metrics. This experiment highlights a concrete and practically meaningful advantage of RewardSDS in interpretable scenarios such as compositional or relational accuracy.

A.5. Effect of Gradient Interpolation on Image Sharpness

To ensure that interpolating gradients across multiple predicted noises does not introduce blurriness, we conducted an additional analysis evaluating both quantitative and perceptual image sharpness. We compared standard SDS with RewardSDS using 6 and 10 noise samples. For each configuration, 25 images were generated from identical prompts and evaluated using two complementary measures: (i) an automated sharpness metric adapted from the LLM-G evaluation framework, scaled to 1–5; and (ii) a user study with 20 participants, who rated image sharpness and image–prompt alignment on a 0–5 Likert scale (higher is better).

Results in Tab. 6 indicate that RewardSDS maintains, and in some cases slightly improves, both perceptual and quantitative sharpness compared to SDS, even when interpolating across multiple noise samples. This suggests that the reward-guided gradient aggregation in RewardSDS preserves high-frequency details rather than averaging them out, mitigating potential over-smoothing effects.

Table 7. Comparison of SDS, SDI [7], and RewardSDS on NeRF-based 3D generation across 22 prompts using stable-diffusion-2.1-base as the 2D prior.

Method	CLIP↑	Aesthetic↑	LLM-G↑	ImageReward↑
SDS	23.19	4.53	3.58	-0.83
SDI	24.06	4.61	**4.29**	-1.19
RewardSDS	**24.52**	**4.71**	4.21	**-0.59**

A.6. Quantitative Comparison to Score Distillation via Inversion (SDI)

As noted in the introduction, recent methods [6, 7] have proposed DDIM inversion as an alternative to random noise sampling in SDS. In this approach, at each training step, a view of the 3D scene is rendered, DDIM inversion is applied up to noise level t, and the image is then denoised to $t-\tau$ before computing gradients with respect to the 3D representation. While this technique improves standard SDS by enforcing consistency with predicted noise, it does not allow controllability based on reward-model alignment.

To compare our method with this line of work, we optimize NeRF-based scenes across 22 prompts (F) using the stable-diffusion-2.1-base as the 2D prior. Tab. 7 shows that our approach outperforms both SDS and the DDIM-based method (SDI), while additionally enabling alignment to arbitrary reward models.

Table 6. Comparison of SDS and RewardSDS using multiple noise samples to evaluate potential effects of gradient interpolation on image sharpness and alignment. Scores are averaged across 25 generated images from identical prompts.

Method	LLM-G (Sharpness)↑	User Study (Sharpness)↑	User Study (Alignment)↑
SDS	2.16	3.06	2.53
RewardSDS (6 noises)	2.17	4.04	3.43
RewardSDS (10 noises)	**2.42**	**4.21**	**3.70**

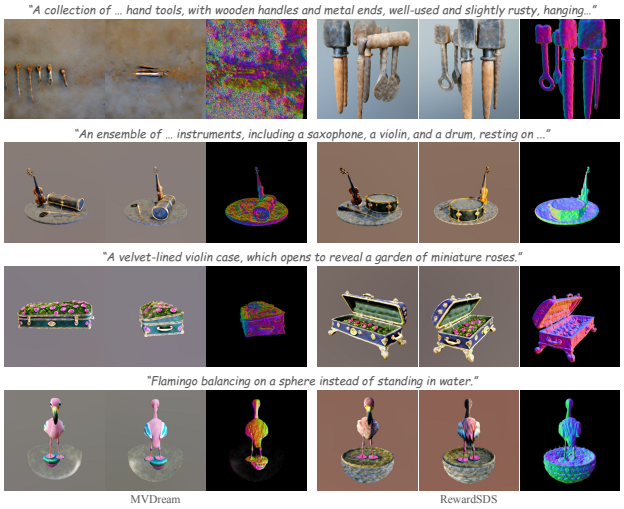

Figure 1. Qualitative comparison of RewardSDS and MVDream on challenging prompts sampled from the GPTEval3D benchmark. [22]

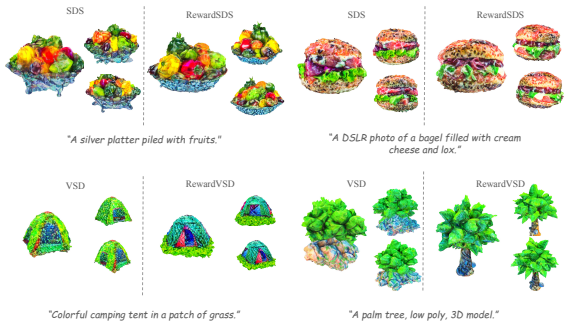

Figure 2. Qualitative results illustrating text-to-3D results for SDS compared to RewardSDS and for VSD compared to our RewardVSD (with stable-diffusion-2.1-base as our 2D prior).

B. Additional Qualitative Results

B.1. Qualitative Results for GPTEval3D Benchmark

Complementing the quantitative analysis in Sec. A.1, Fig. 1 presents qualitative comparisons between RewardSDS and MVDream on challenging prompts from the GPTEval3D benchmark. RewardSDS consistently produces more accurate and detailed results across various cases.

B.2. Qualitative Results for 3D Generation with 3DGS

Fig. 2 presents qualitative examples corresponding to the quantitative results A.2. Additionally, Fig. 3 shows results using MVDream as the 2D prior for 3DGS optimization. These comparisons highlight the improvements in text-to-3D alignment and scene fidelity achieved by RewardSDS. The selected examples also illustrate the ability of our method to produce diverse geometry and fine-grained details, showcasing the robustness of RewardSDS in optimizing the 3DGS backbone.

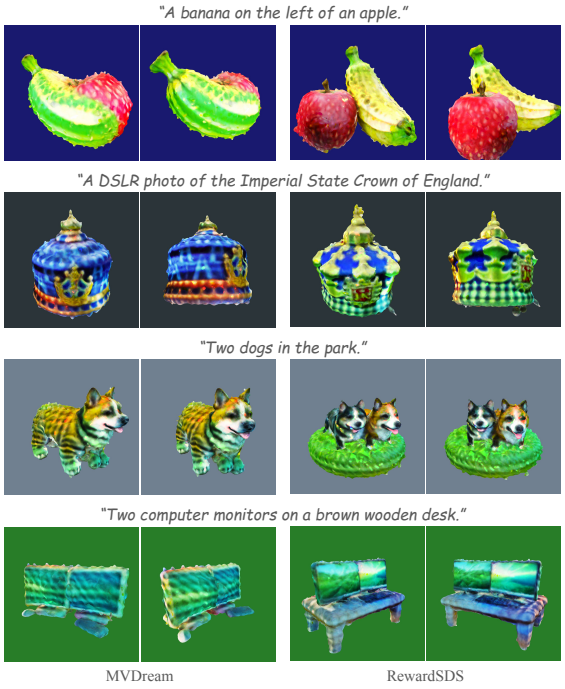

Figure 3. Qualitative comparison of text-to-3D generation based on 3DGS, in comparison to MVDream. Our method demonstrates improved alignment and visual quality, capturing more detailed geometric structures and fine-grained textures.

Figure 4. Qualitative comparison of SDS-Bridge and Consistent-FlowDistillation (CFD) with and without RewardSDS.

B.3. Qualitative Examples of Plug-and-Play Integration

To illustrate RewardSDS improvements qualitatively, Figure 4 presents examples of the plug-and-play integration of reward-based noise sampling. These results further demonstrate the gains in visual fidelity and text alignment when applying RewardSDS to existing SDS-based frameworks.

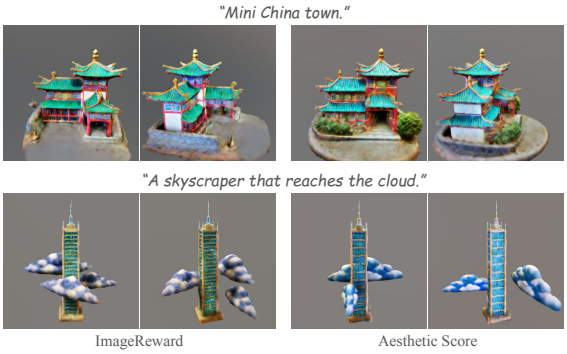

Figure 5. Qualitative comparison of text-to-3D generation using different reward models. We consider a NeRF backbone optimized with RewardSDS, either using the ImageReward reward model or Aesthetic Score reward model. As can be seen, using aesthetic reward results in adding bushes (top row) and a different (more aesthetic) colors (both rows).

Figure 6. Qualitative comparison of generated outputs using different reward models for RewardVSD and the VSD baseline. Each row corresponds to a different reward model, with the input prompts shown at the bottom, taken from Drawbench.

B.4. Qualitative Comparison of Different Reward Models

As a supplement to the Experiments section of the main paper, we present additional qualitative comparisons of SDS-based generations using different reward models. First, in the 3D settings, we show results using a NeRF backbone optimized with RewardSDS, either with the ImageReward reward model or the Aesthetic Score reward model, as illustrated in Fig. 5. Fig. 6 presents 2D outputs generated using different reward models within the RewardVSD framework, alongside the VSD baseline. Together, these comparisons offer additional insight into the impact of reward model selection on the alignment and visual quality of the final generated scenes and images.

Figure 7. Qualitative comparison of SDS and RewardSDS on counting and spatial relationship prompts, showing improved alignment to object counts and spatial layouts.

B.5. Qualitative Evaluation of Counting and Spatial Relationship prompts

Fig. 7 presents qualitative examples corresponding to the quantitative evaluation in Sec. A.4. We observe that RewardSDS produces compositions that more faithfully capture both object counts and spatial arrangements described in the prompts. These examples visually corroborate our quantitative findings, highlighting RewardSDS's advantage in fine-grained compositional alignment.

Figure 8. Qualitative results illustrating the effect of the number of considered noises (N). The top row presents the baseline method, while the input prompts are displayed below the images.

B.6. Qualitative Comparison of the Effect of Number of Considered Noises

The number of noise samples (N) considered at each optimization step plays a crucial role in guiding the generation process. As discussed in the ablation studies, increasing N leads to better alignment with the desired reward model and enhances the overall image quality. Fig. 8 provides qualitative examples illustrating how different values of N affect the final generated outputs. As shown, larger values of N result in more refined and coherent images, whereas smaller values introduce more variability and potential misalignment.

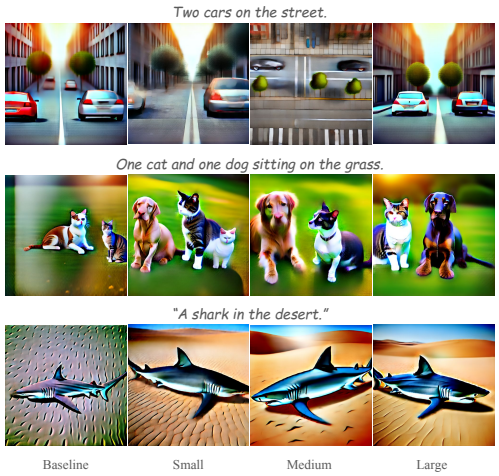

Figure 9. Qualitative comparison across different scale settings (Baseline, Small, Medium, Large). Increasing the scale produces progressively sharper, more aligned, and more coherent generations. Prompts are shown above each row.

B.7. Qualitative Comparison Across Scale Settings

Fig. 9 provides qualitative examples complementing the scaling analysis presented in Tab.5 of the main paper. As the scale increases from Baseline to Large, RewardSDS yields noticeably improved alignment, consistency, and visual fidelity. Even the Small scale setting shows clear gains over the baseline, supporting our claim that modest increases in N, K, and S already produce substantial improvements.

C. Theoretical Analysis of RewardSDS

C.1. Theoretical Connection of RewardSDS to DPO

In addition to the empirical results presented throughout the paper, we provide a formal derivation connecting RewardSDS to previous alignment-related works. The core idea is to optimize the parameters of a 3D scene, θ, (e.g, a Neural Radiance Field) such that rendered images, when processed through a fixed pre-trained 2D diffusion model ϵ_ϕ with different noise samples, align with preferences derived from a reward model R.

C.1.1. Direct Preference Optimization Framework

Direct Preference Optimization (DPO) [14] is a preference-based fine-tuning method originally proposed for language models. Given a pair of annotated samples (x^w, x^l), where x^w is preferred over x^l for a given context c, DPO optimizes a contrastive loss that encourages the model to assign higher log-probability to the preferred output:

$$\mathcal{L}_{\text{DPO}}(\phi) = -\mathbb{E}_{c,x^w,x^l}\Big[\log\sigma\Big(\beta\Big(\log\frac{p_\phi(x^w|c)}{p_{\text{ref}}(x^w|c)} - \log\frac{p_\phi(x^l|c)}{p_{\text{ref}}(x^l|c)}\Big)\Big)\Big] \tag{1}$$

where p_ϕ and p_{ref} are the output distributions of the fine-tuned and reference models, respectively, and β is hyperparameter.

For diffusion models, the DPO objective can be adapted [20] by interpreting x^w and x^l as two denoised predictions (e.g., from different noise samples or guidance trajectories) for the same context. We define the per-sample residual difference:

$$\Delta_t^s = \|\epsilon^s - \epsilon_\phi(x_t^s \mid t, c)\|_2^2 - \|\epsilon^s - \epsilon_{\text{ref}}(x_t^s \mid t, c)\|_2^2 \tag{2}$$

where $x_t^s = \alpha_t x_0^s + \sigma_t \epsilon^s$, $\epsilon^s \sim \mathcal{N}(0, I)$, for $s \in \{w, l\}$. The diffusion-DPO objective becomes:

$$\mathcal{L}_{\text{D-DPO}}(\phi) = -\mathbb{E}_{c,x^w,x^l,t}\Big[\log\sigma\big(-\beta T\,\omega(\lambda_t)(\Delta_t^w - \Delta_t^l)\big)\Big]. \tag{3}$$

where $\lambda_t = \alpha_t^2/\sigma_t^2$ is the signal-to-noise ratio, and $\omega(\lambda_t)$ is a weighting function.

C.1.2. Connecting RewardSDS to DPO in 3D Generation

As the original diffusion loss is closely related to the SDS loss[12], our method can be seen as an extension of DPO to SDS, and is related to the above formulation. Let us consider a specific instance of the general RewardSDS formulation (see Eq.8 in the main paper), corresponding to the "step toward best, away from worst" strategy (scheme (v) in Tab.4 of the main paper). These can be interpreted as annotated examples, ranked by a reward model. Unlike the standard DPO setting, where there is a reference model, the 3D optimization setup involves a single scene being optimized. Nevertheless, RewardSDS can be formulated as a reference-free DPO [10] objective, using the residual:

$$\widetilde{\Delta}_t^s = \|\epsilon_\phi(x_t^s, y, t) - \epsilon^s\|_2^2, \quad s \in \{w, l\},$$

the loss can be written as:

$$L_{\text{R-SDS}}(x_0 = g(\theta)) = \mathbb{E}_t\Big[w(t)\big(\widetilde{\Delta}_t^w - \widetilde{\Delta}_t^l\big)\Big]. \tag{4}$$

where x_0 is an rendered image of the 3D scene, ϵ_ϕ is the 2D prior, y is a context (e.g., text prompt), and ϵ^* denotes either the best or worst noise, corresponding to the highest, or lowest-ranked denoised image (based on scores obtained from R), respectively, obtained from $x_t^* = \alpha_t x_0^* + \sigma_t \epsilon^*$.

This formulation mirrors the structure of the DPO objective in Eq.3, where the model is guided to move closer to the higher-ranked (preferred) noise ϵ^w and away from the lower-ranked (less preferred) noise ϵ^l, aligning gradients with preference signals. Next, we consider the gradient of the loss. Define the gradient residual:

$$r_t^s = \epsilon_\phi(x_t^s, y, t) - \epsilon^s, \qquad s \in \{w, l\}. \tag{5}$$

the gradient becomes:

$$\nabla_\theta L_{\text{R-SDS}}(x_0 = g(\theta)) = \mathbb{E}_t\Big[w(t)\,(r_t^w - r_t^l)\,\frac{\partial x_0}{\partial \theta}\Big] \tag{6}$$

This update rule can be interpreted as a weighted combination of SDS gradients: a positive step in the direction of the high-reward noise and a negative step from the low-reward one. When adapting the DPO framework to optimize a 3D scene θ under a fixed diffusion prior ϵ_ϕ, and using preference signals derived from a reward model R, we arrive at this update, offering a theoretical grounding for RewardSDS strategies that emphasize high-reward noise samples while suppressing those with lower reward.

C.2. Reward Variance Across Gaussian Noise Samples

To further understand the effect of Gaussian noise sampling on optimization outcomes, we also provide additional empirical evidence supporting why certain Gaussian noise samples lead to better outputs. To this end, we analyzed reward variance across noise samples over the course of optimization. Fig. 10 illustrates the evolution of reward distributions for two representative reward models: **ImageReward** and **Aesthetic Score**.

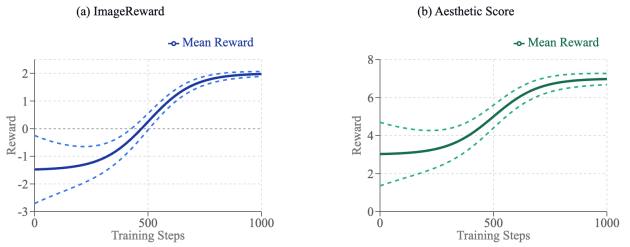

Figure 10. Evolution of reward variance across Gaussian noise samples for (a) ImageReward and (b) Aesthetic Score. ImageReward exhibits reward stabilization, while Aesthetic Score remains sensitive to subtle differences even at convergence.

For ImageReward, which is an unbounded reward model (with most scores in practice ranging from –3 to +3), we observed a consistent and characteristic pattern across ten prompts. The distribution of rewards follows an S-shaped trajectory over the training steps:

- In the early iterations, rewards are mostly negative, widely spread, and highly variable, with values ranging roughly from –3 to 0.
- As optimization progresses, the median and mean rewards steadily increase, passing through a transitional middle stage, where the variance remains high.
- In the final stage, rewards stabilize around positive values (close to +2), and the variance collapses to a very narrow-band (dropping from about 0.4–0.6 initially to below 0.05).

This dynamic shows that as training advances, outputs become both better aligned (higher rewards) and more consistent (lower variability).

Interestingly, for Aesthetic Score the behavior differs: while the variance decreases over time, it remains comparatively higher in the final stages, typically around 0.15 - 0.2. This is expected because prompt-alignment rewards (such as CLIP or ImageReward) focus on coarse structure early in training and converge as alignment improves, whereas rewards emphasizing fine details (such as Aesthetic Score) remain sensitive to subtle variations introduced during the final refinement steps.

This analysis highlights that the reward model itself strongly influences the variance dynamics throughout optimization.

D. Implementation Details

In all of the experiments, including both our method and the baselines, we used a single L40s GPU. For all zero-shot text-to-image, and image editing experiments, optimiza-

Figure 11. In (a), fine details commonly missed by the 2D model are also absent in the resulting 3D scene. In (b), when the 2D prior (MVDream in this case) tends to generate object-centric multi-views, the resulting 3D scene struggles to complete global context such as the background.

Figure 12. In this figure, we visualize the ImageReward (↑) scores. While the reward model generally provides useful guidance, it can occasionally overlook important aspects. (Left) The model fails to accurately assess object count. (Right) The model misses stylistic attributes such as color.

tion is performed using the ADAM optimizer with a learning rate of 0.01. As the text-to-image model, we employ Stable Diffusion 2.1 Base [15], keeping the classifier-free guidance (CFG) scale as it was in the original SDS-based method (100 for SDS, 7.5, etc.). Each 2D image is generated in 1,000 optimization steps using $N = 10$ (the number of noise samples drawn at each iteration), $K = 1000$ (the number of reward-based optimization steps used during generation), and $S = 15$ (the number of inference steps applied during noise selection to obtain refined reward estimates). The weighting strategy assigns a weight of 0.9 to the two highest scoring samples, a weight of -0.1 to the two lowest scoring samples, and 0 to the rest. With those settings, image optimization takes 2227 seconds. For image editing experiments, generation is performed in 200 steps, taking 211 seconds per image, and we use $N = 5$, $K = 200$, and $S = 1$ as the hyperparameters, with the same weighting strategy as above. For text-guided 3D generation, MVDream [17] is used as our text-to-image prior (beside the experiment described in Sec. B.2), and we employ the public implementations of DreamGaussian [19] for 3DGS backbone optimization (leaving their second stage of texture refinement as is) and MVDream [3] (except for the GPTEval3D experiemnts, scenes were optimized without shading) for NeRF-based optimization. We use the same settings as in the public implementations, and in NeRF training, we apply our method only in the first half of the optimization as we found that this is sufficient for convergence with our method. Regarding the weighting strategy, we use the same one as in the 2D. Optimization of 3DGS takes 60 minutes and NeRF takes 7 hours.

E. Limitations, Failure Cases and Future Work

This section is included to provide a more complete evaluation of our method by discussing its limitations, failure cases, and potential directions for future work.

A primary limitation of RewardSDS is the increased runtime, stemming from the denoising process performed at each SDS step to evaluate individual noise samples. This follows recent trends in generation methods that deliberately increase inference time to improve output quality [8], a trade-off that is especially reasonable in 3D, where optimization is inherently slow and not designed for real-time generation, like in text or image synthesis. To address this limitation, as discussed in the Ablations section, users can reduce runtime by selecting hyperparameters that better align with their computational constraints.

Additionally, although our method leverages a reward model to explore different noise samples and improve the distillation process, its performance remains constrained by the expressiveness of both the reward model and the underlying 2D prior. Despite their strong capabilities, the prior can occasionally produce suboptimal generations (see Fig. 11), and the reward model may assign inaccurate scores (see Fig. 12).

Future work can explore learning the alignment scores weights dynamically during optimization, as our current, fixed weighting was chosen through limited-scale experiments and may not generalize across prompts or optimization stages. Furthermore, evaluating noise samples using a combination of reward models, rather than a single one, may further improve alignment [2]. A thorough analysis of how different reward models impact final results could offer valuable insights to users.

E.1. Broader Societal Impacts

Our work on text-to-3D generation has potential positive applications in design, and content creation by lowering barriers to high-quality 3D asset generation. However, like other generative models, it could be misused to produce misleading or harmful 3D content. While our method is

foundational and not deployed directly, we encourage responsible use and support future safeguards to mitigate potential misuse.

F. Hand-Crafted Prompts for NeRF-Based MVDream Evaluation

As described in the Experiments section, to evaluate NeRF-based 3D generation, with and without our method, we utilized a set of 22 hand-crafted prompts. These prompts were carefully designed to be diverse and often complex, covering a wide range of objects, scenes, and subjects, thereby ensuring a comprehensive assessment of text-to-3D generation quality. The full list of prompts is provided in Tab. 8.

Table 8. List of hand-crafted prompts used to evaluate NeRF-based MVDream with and without RewardSDS.

#	Prompt
1	A basketball player dunking a basketball
2	A basketball player in a red jersey, high resolution, 4K
3	A bulldog wearing a black pirate hat
4	A cartoon cat eating a cheesecake, realistic
5	a DSLR photo of a ghost eating a hamburger
6	A guitar player on stage, high quality, realistic, HD, 8K
7	A man with a beard, wearing a suit, holding a pink briefcase, high quality, realistic, HD
8	A penguin with a brown bag in the snow
9	A man with a red scarf, highly detailed, 4K
10	A Shih Tzu with a bowtie, high quality, realistic, HD, 8K
11	A skyscraper that reaches the clouds, high quality, realistic
12	A tiger in the jungle, high quality, realistic, HD
13	A white sofa next to a brown wooden table
14	A young girl flying a kite, high quality
15	An astronaut riding a horse
16	Argentinian football player, celebrating a goal, HD
17	Corgi riding a rocket
18	King Kong climbing the Empire State Building
19	Mini China town, highly detailed, 8K, HD
20	Red drum set, high quality, realistic, HD, 8K
21	Two dogs in the park
22	World cup trophy, high quality, realistic, HD

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 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)*, pages 22246–22256, 2023. 2

[2] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Neural Information Processing Systems (NeurIPS)*, 2024. 7

[3] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 7

[4] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[5] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[6] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. 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*, pages 6517–6526, 2024. 2

[7] Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin M Solomon. Score distillation via reparametrized ddim. *Advances in Neural Information Processing Systems*, 37:26011–26044, 2024. 2

[8] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, T. Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *ArXiv*, abs/2501.09732, 2025. 7

[9] David McAllister, Songwei Ge, Jia-Bin Huang, David W. Jacobs, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions, 2024. 1

[10] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235, 2024. 6

[11] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12663–12673, 2023. 2

[12] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 2, 6

[13] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. 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*, pages 9914–9925, 2024. 2

[14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 5

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 7

[16] 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. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[17] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2024. 2, 7

[18] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 2

[19] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1, 7

[20] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 5

[21] 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, 2023. 2

[22] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *CVPR*, 2024. 1, 2, 3

[23] Runjie Yan, Yinbo Chen, and Xiaolong Wang. Consistent flow distillation for text-to-3d generation, 2025. 1