UNLEASHING 2D REWARDS FOR HUMAN PREFERENCE ALIGNED TEXT-TO-3D GENERATION VIA PREFERENCE SCORE DISTILLATION

Anonymous authors
Paper under double-blind review

Figure 1: Comparisons with state-of-the-art methods RichDreamer (Qiu et al., 2024) and Trellis (Xiang et al., 2025). Even when compared to methods that leverage stronger 3D priors, our method achieves significantly higher text alignment and enhanced visual quality, highlighting the critical role of preference alignment in text-to-3D generation.

ABSTRACT

Human preference alignment presents a critical yet underexplored challenge for diffusion models in text-to-3D generation. Existing solutions typically require task-specific fine-tuning, posing significant hurdles in data-scarce 3D domains. To address this, we propose Preference Score Distillation (PSD), an optimization-based framework that leverages pretrained 2D reward models for human-aligned text-to-3D synthesis without 3D training data. Our key insight stems from the incompatibility of pixel-level gradients: due to the absence of noisy samples during reward model training, direct application of 2D reward gradients disturbs the denoising process. Noticing that similar issue occurs in the naive classifier guidance in conditioned diffusion models, we fundamentally rethink preference alignment as a classifier-free guidance (CFG)-style mechanism through our implicit reward model. Furthermore, recognizing that frozen pretrained diffusion models constrain performance, we introduce an adaptive strategy to co-optimize preference scores and negative text embeddings. By incorporating CFG during optimization, online refinement of negative text embeddings dynamically enhances alignment. To our knowledge, we are the first to bridge human preference alignment with CFG theory under score distillation framework. Experiments demonstrate the superiority of PSD in aesthetic metrics, seamless integration with diverse pipelines, and strong extensibility.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021b;a; Tu et al., 2024a;b; 2025b;a;c), trained on web-scale datasets, demonstrate exceptional capability in generating high-fidelity images (Dhariwal & Nichol, 2021; Rombach et al., 2022). Motivated by this success, researchers have sought

to transfer pretrained 2D generative priors to data-scarce modalities. Score Distillation Sampling (SDS) (Poole et al., 2023) pioneered this cross-modal knowledge transfer, leveraging pretrained text-to-image diffusion models to optimize 3D differentiable representations through gradient-based maximum likelihood estimation. By circumventing the need for 3D training data, SDS has established text-to-3D generation as a prominent research direction and enabled applications beyond 3D synthesis, including one-step diffusion distillation (Yin et al., 2024b;a; 2025), character animation (Li et al., 2025b), and metric depth prediction (Pham et al., 2025). Despite extensive efforts to refine SDS (Wang et al., 2023; McAllister et al., 2024; Li et al., 2025c; Jiang et al., 2025; Yang et al., 2025a), recent studies (Ye et al., 2025; Liu et al., 2025b; Zhou et al., 2025) reveal that SDS-synthesized 3D assets often exhibit misalignment with human preferences — a limitation shared by other diffusion models.

To address this misalignment, Reinforcement Learning from Human Feedback (RLHF) has been incorporated into text-to-3D pipelines. However, existing RLHF-based methods (Ye et al., 2025; Liu et al., 2025a; Zou et al., 2025) typically require training 3D-specific reward models, which fundamentally undermines the core advantage of 3D-data-free synthesis and may induce visual artifacts. Critically, since reward models are exclusively trained on clean images, directly applying their gradients to update 3D representations under high noise levels induces gradient misalignment. Given the established connection between score distillation (Lukoianov et al., 2024; Yan et al., 2025; Li et al., 2025c; Wu et al., 2024) and the *Probability Flow ODE* (PF-ODE) (Song et al., 2021b), we hypothesize this originates from pixel-level conflicts between reward gradients and diffusion dynamics. While DreamDPO (Zhou et al., 2025) attempts to avoid pixel-wise gradients via DPO-inspired objectives, its disconnection from denoising dynamics limits extensibility to iterative refinement processes.

Motivated by DPO's implicit reward modeling and the efficacy of Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) in conditional diffusion, we propose a fundamental rethinking: ***Could the implicit reward in score distillation function as a PF-ODE-compatible guidance signal?*** In this work, we introduce Preference Score Distillation (PSD), a novel framework that harnesses gradients from an implicit reward model to align score distillation with human preferences. To bridge preference learning with guidance mechanisms, we formalize human preference as a binary variable $\mathcal{S}_{\text{pref}}$, to obtain a preference score guidance term $\nabla_{\boldsymbol{x}_t} \log p\left(\mathcal{S}_{\text{pref}} \mid \boldsymbol{x}_t, y\right)$ that decomposes into interpretable gradient components. Crucially, we identify that suppressing pixel-wise artifacts requires theoretically connecting this preference gradient to the score estimates. This insight motivates a reformulation of RLHF under the score distillation paradigm, where we rewrite the KL-divergence with a dynamic reference distribution tied to the current rendering. Eventually, through our derivation and constructing contrastive sample pair on-the-fly, we formulate a CFG-like guidance that is able to increase the likelihood towards preferred completions. Empirical results proves that it is compatible with existing diffusion dynamics and improve various aesthetics scores directly.

Moreover, noting that the pretrained diffusion model is frozen, we design an algorithm that adaptively updates the preference score and negative text embeddings (Ban et al., 2025; Wang et al., 2024; Li et al., 2025a). In each denoising step: the preference score is first computed; subsequently, the negative embedding (projected into the continuous text embedding space) is optimized as trainable parameters via reward score backpropagation, and integrated with CFG. Our experiments demonstrate that our approach has strong compatibility and can generate highly photorealistic, preference-aligned 3D assets.

To show the superiority of our method, we compare with state-of-the-art methods that utilizes stronger 3D priors in Fig. 1. While comparison method (Qiu et al., 2024; Xiang et al., 2025) applies more 3d priors (Normal-Depth diffusion model, physically-based rendering materials) or large-scale training with 3d data, we only distillate diffusion models for image synthesis (MVDream Shi et al. (2024) and Stable Diffusion v2.1 (Rombach et al., 2022)) but still yield better text alignment and aesthetics, which domesticates the novelty of this work.

Our contribution can be summarized as :

- We propose a preference alignment method for score distillation, Preference Score Distillation (PSD). To the best of our knowledge, we are the first to demonstrate that **preference alignment can be directly formulated as a CFG-type guidance** and can produce gra-

dients towards increasing the likelihood of preferred samples via constructing contrastive sample pairs during the optimization process.

- We propose a strategy that alternatively updates the preference score and negative embeddings. In this strategy, optimizing continuous negative embeddings can achieve the effect of updating pretrained diffusion parameters.
- Extensive experiments prove the ability of PSD to improve aesthetics scores. The results shows PSD outperforms other comparing methods in scores of 4 human preference assessment models, 1 visual question answering (VQA) model and delivers highly impressive qualitative comparison.

2 Preliminaries and Notations

Score Based Diffusion Models. The process of diffusing a data sample into random noise can be described as *Probability Flow Ordinary Differential Equation* (PF-ODE) (Song et al., 2021b). For an arbitrary data point $\boldsymbol{x}_0 \sim p_{data}$, if we gradually add noise

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

the PF-ODE that has the same marginal distribution can be written as

$$\begin{aligned} \mathrm{d}(\frac{\boldsymbol{x}_t}{\alpha_t}) &= \mathrm{d}(\frac{\sigma_t}{\alpha_t})(-\sigma_t \nabla_{x_t} \log p(\boldsymbol{x}_t)) \\ &= \mathrm{d}(\frac{\sigma_t}{\alpha_t}) \cdot \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t) \end{aligned} \tag{2}$$

where $\boldsymbol{\epsilon}_\phi(\cdot) \approx -\sigma_t \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ is our trained diffusion models. Our notations of diffusion models are consistent with (Karras et al., 2022; Yan et al., 2024; 2025).

Classifier-free Guidance. In order to generate text-aligned contents, a technique termed *Classifier-free Guidance* (CFG) (Ho & Salimans, 2022) pushes the samples towards higher likelihood through the gradient of an implicit classifier

$$\tilde{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t, y, t) = \underbrace{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t)}_{unconditional} + \underbrace{\gamma(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t))}_{implicit\, classifier\, \delta_{cls}} \tag{3}$$

where y is the conditioning text embedding and γ is a scaling factor, and we denote gradient $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t)$ produced by implicit classifier as δ_{cls}. Additionally, negative prompting has become a common technique to improve generation quality. It replaces $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t)$ with $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, n, t)$ conditioned by negative embedding n. We regard the embedding as a set of the model parameters and thus simplify it as $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t)$.

RLHF on Score Distillation. Typically, *Reinforcement Learning from Human Feedback* (RLHF) fine-tunes the diffusion models by maximizing expected rewards while regularizing the KL-divergence from a reference distribution (Christiano et al., 2017). We define a similar objective for score distillation:

$$\max_{\phi} \mathbb{E}_{\boldsymbol{x}_t \sim p_\phi(\boldsymbol{x}_t|y)}[r(y, \boldsymbol{x}_t)] - \beta \mathbb{D}_{\mathrm{KL}}[p_\phi(\boldsymbol{x}_t|y)||q_\theta(\boldsymbol{x}_t|\boldsymbol{x}_0 = g_\theta(\boldsymbol{c}))], \tag{4}$$

where θ is the learnable parameters of differentiable representation g, $\boldsymbol{c}$ is the rendering camera view, $q_\theta(\cdot)$ is the marginal distribution and $\mathbb{D}_{\mathrm{KL}}[\cdot]$ is the KL-divergence. The difference between our definition and standard RLHF (Christiano et al., 2017) is we modify the reference model in KL term into marginal distribution of current rendering $g_\theta(\boldsymbol{c})$ and we seek to optimize for an arbitrary timestep t. Justifications are presented in Appendix F.2, where existing works (Ye et al., 2025; Zhou et al., 2025) can be related to our definition.

3 Approach

In Section 3.1, we first establish the connection between preference and guidance by deriving the preference score guidance. In Section 3.2, we present the proposed preference score distillation

(PSD) method. In Section 3.3, we introduce a novel adaptive strategy for updating the preference score and negative embedding to improve the quality of generation. Due to the limited space, we present full derivation details in Appendix F and overall pseudo code of PSD in Algorithm 1.

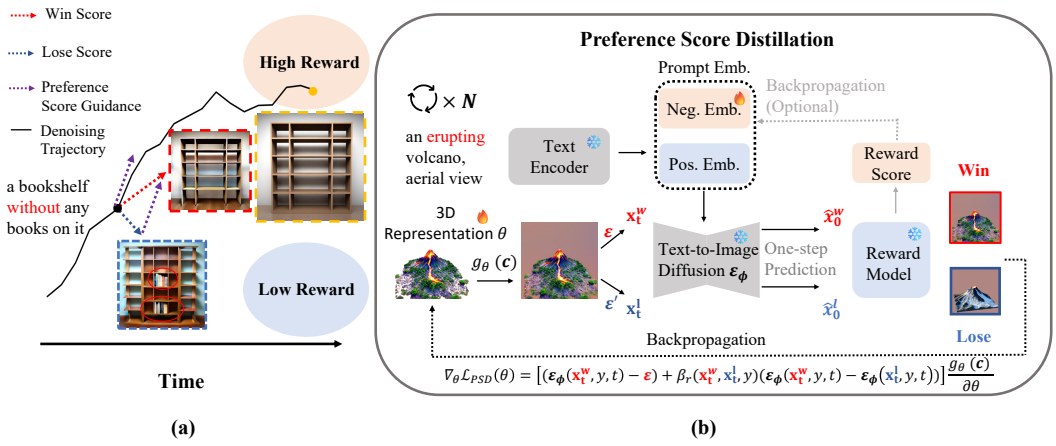


Figure 2: Overall illustration of Preference Score Distillation. a) Win (red) and lose (blue) samples are constructed on-the-fly to calculate win and lose scores, then preference score guidance (purple) pushes the denoising trajectory towards high-reward regions and finally improve alignment with reward. b) In each step, two noise is added to the rendering images $g_\theta(\boldsymbol{c})$ and reward model determines win/lose based on one-step prediction of pretrained diffusion models $\boldsymbol{\epsilon}_\phi$. 3D representation θ and negative embedding n are updated by our objective $\mathcal{L}_{\text{PSD}}(\theta)$ and reward score respectively.

3.1 Linking Preference to Guidance

The foundation of our framework is built upon a recent insight (Li et al., 2025c; Yan et al., 2025; Wu et al., 2024) into the connection between differential representation optimization and the denoising process of diffusion models. In Eq. 5, $\mathrm{d}(\frac{\sigma_t}{\alpha_t})$ can be viewed as the learning rate lr of an optimizer and $\boldsymbol{\epsilon}_\phi(\cdot)$ can be viewed as the gradient $\nabla\mathcal{L}$ of $\mathrm{d}(\frac{\boldsymbol{x}_t}{\alpha_t})$. In order to guiding $\boldsymbol{x}_t$ with preference, we formally introduce a binary variable $\mathcal{S}_{\text{pref}}$ as human-preferred properties for constrained conditions.

$$\mathrm{d}(\frac{\boldsymbol{x}_t}{\alpha_t}) = \underbrace{\mathrm{d}(\frac{\sigma_t}{\alpha_t})}_{-lr} \cdot \underbrace{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, \mathcal{S}_{\text{pref}}, t)}_{\nabla\mathcal{L}}. \tag{5}$$

and apply Bayes' rule

$$\nabla_{\boldsymbol{x}_t} \log p_\phi\left(\boldsymbol{x}_t \mid y, \mathcal{S}_{\text{pref}}\right) = \nabla_{\boldsymbol{x}_t} \log p_\phi\left(\boldsymbol{x}_t \mid y\right) + \nabla_{\boldsymbol{x}_t} \log p_\phi\left(\mathcal{S}_{\text{pref}} \mid \boldsymbol{x}_t, y\right). \tag{6}$$

A naive solution is to train a "classifier" (reward model) to estimate the probability of $\boldsymbol{x}_t$ aligning with human preference $p_\phi\left(\mathcal{S}_{\text{pref}} \mid \boldsymbol{x}_t, \boldsymbol{c}\right)$, but the drawback majorly involves: 1) introducing additional training to produce appropriate gradients and 2) reward models needs to perform at arbitrary timestep, but it can be only trained with clean images so we have to approximate clean data during early steps when noise is large (Dhariwal & Nichol, 2021). Thus, we seek to formulate CFG-type guidance.

To achieve this, our high-level idea is to construct gradient via win-lose pair similar to DPO (Rafailov et al., 2023; Zhu et al., 2025a) such that the denoising process is pushed to increase the likelihood of wining sample. We first introduce Bradley-Terry (BT) model (Bradley & Terry, 1952) for human preference

$$p\left(\mathcal{S}_{\text{pref}} \mid \boldsymbol{x}_t, y\right) = p\left(\boldsymbol{x}_t^{\text{w}} \succ \boldsymbol{x}_t^{\text{l}} \mid \boldsymbol{x}_t, y\right) = \sigma\left(r\left(y, \boldsymbol{x}_t^{\text{w}}\right) - r\left(y, \boldsymbol{x}_t^{\text{l}}\right)\right), \tag{7}$$

where $\sigma(\cdot)$ represents the sigmoid function, $r(\cdot)$ is reward model, $\boldsymbol{x}_t^{\mathrm{w}} \succ \boldsymbol{x}_t^{\mathrm{l}}$ represents $\boldsymbol{x}_t^{\mathrm{w}}$ and $\boldsymbol{x}_t^{\mathrm{l}}$ are win-lose pair examples respectively. Different from off-line preference optimization, we construct $\boldsymbol{x}_t^{\mathrm{w}}$ and $\boldsymbol{x}_t^{\mathrm{l}}$ in Eq. 7 online in order to link preference with inference-time guidance. Consequently, replace $p_\phi\left(\mathcal{S}_{\text{pref}} \mid \boldsymbol{x}_t, y\right)$ in Eq. 6 with Eq. 7 then Eq. 5 becomes

$$\mathrm{d}(\frac{\boldsymbol{x}_t}{\alpha_t}) = \mathrm{d}(\frac{\sigma_t}{\alpha_t}) \cdot \left(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \sigma_t \left(1 - \sigma\left(r\left(y, \boldsymbol{x}_t^{\mathrm{w}}\right) - r\left(y, \boldsymbol{x}_t^{\mathrm{l}}\right)\right) \nabla_{\boldsymbol{x}_t}\left(r\left(y, \boldsymbol{x}_t^{\mathrm{w}}\right) - r\left(y, \boldsymbol{x}_t^{\mathrm{l}}\right)\right)\right)\right) \tag{8}$$

The primary challenge of solving Eq. 8 is term $\nabla_{\boldsymbol{x}_t}\left(r\left(y, \boldsymbol{x}_t^{\mathrm{w}}\right) - r\left(y, \boldsymbol{x}_t^{\mathrm{l}}\right)\right)$ being not tractable. Fortunately, we can rewrite the reward using the unique global optimal solution p_ϕ^* of Eq. 4 varied from (Rafailov et al., 2023), which is

$$r(y, \boldsymbol{x}_t) = \beta \log \frac{p_\phi^*(\boldsymbol{x}_t|y)}{q_\theta(\boldsymbol{x}_t|\boldsymbol{x}_c = g_\theta(\boldsymbol{c}))} + \beta \log Z(\boldsymbol{x}_c) \tag{9}$$

where $Z(\boldsymbol{x}_0)$ is a trivial partition function since it eliminates directly. Then, we employ reparameterization trick by pugging Eq. 9 into Eq. 8, such that we have our proposed preference guided ODE:

$$\mathrm{d}(\frac{\boldsymbol{x}_t}{\alpha_t}) = \mathrm{d}(\frac{\sigma_t}{\alpha_t}) \cdot \left(\underbrace{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t)}_{(A)} + \beta\sigma\left(r\left(y, \boldsymbol{x}_t^l\right) - r\left(y, \boldsymbol{x}_t^w\right)\right) \underbrace{\left(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^w, y, t) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^l, y, t)\right)}_{(B)} \right). \tag{10}$$

Observe Eq. 10, if we force a frozen pretrained diffusion model (substituting p_ϕ^* with p_ϕ), term (A) will be analogous to unconditional score in Eq. 3, and newly introduced term (B) will be analogous to the implicit classifier δ_{cls}. As for term $\sigma\left(r\left(y, \boldsymbol{x}_t^l\right) - r\left(y, \boldsymbol{x}_t^w\right)\right)$, it weights the guidance term by how incorrectly the implicit reward model ranking the win-lose pair. Generally, we name the term (B) as *preference score guidance* and denote as δ_{pref}.

In practice, we calculate the reward on noisy steps $r\left(y, \boldsymbol{x}_t\right)$ with the approximation $r\left(y, \hat{\boldsymbol{x}}_0\right)$ based on Tweedie's formula (Efron, 2011) $\hat{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - \sigma_t \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t)}{\alpha_t}$, and choose the sample with higher score to be $\boldsymbol{x}_t^w$. Furthermore, to reduce the number of forward passes in each denoising step, we replace term (A) with $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^w, y, t)$ and apply CFG to calculate the win/lose score in δ_{pref} as well.

3.2 Preference Score Distillation

A key feature while formulating Eq. 10 is the introduction of on-the-fly win-lose pair. For score distillation, (Zhou et al., 2025) has shown the use of different noise to construct win-lose pair. In our Eq. 10, if we define

$$\boldsymbol{x}_t^w = \alpha_t \boldsymbol{x}_c + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{x}_t^l = \alpha_t \boldsymbol{x}_c + \sigma_t \boldsymbol{\epsilon}', \tag{11}$$

where $\boldsymbol{x}_c$ is a non-noisy sample (whether it is one-step predicted samples from (Lukoianov et al., 2024) or clean space samples from (Yu et al., 2024; Yan et al., 2024)), $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are two independent noise, then our preference guidance δ_{pref} will become a velocity field that pushes the samples towards high reward regions in all conditions. Eventually, with the *change-of-variable* (Lukoianov et al., 2024; Yan et al., 2024; 2025) technique already discussed in previous works, guiding 3D generation using Eq. 10 can be formulated as objective

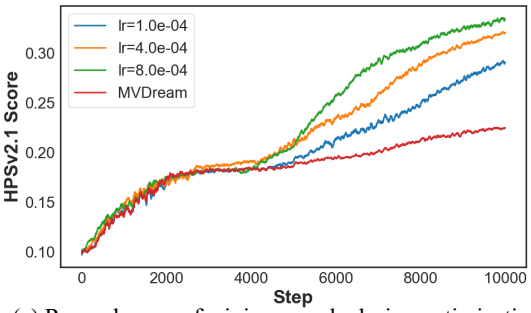
$$\nabla_\theta \mathcal{L}_{\text{PSD}}(\theta) = \mathbb{E}_{t,\boldsymbol{c}}\left[\left(\delta_{gen} + \gamma\delta_{cls} + \beta_r \delta_{pref}\right) \frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}\right], \tag{12}$$

where unconditional prior $\delta_{gen} = \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^w, t) - \boldsymbol{\epsilon}$ and we set $\beta_r = \gamma \frac{||\delta_{cls}||_2}{||\delta_{pref}||_2} \cdot \sigma\left(r\left(y, \boldsymbol{x}_t^w\right) - r\left(y, \boldsymbol{x}_t^l\right)\right)$ to balance the gradient produced by CFG and our preference score guidance. Notice that δ_{gen} is a general formulation for score distillation (Poole et al., 2023; Yan et al., 2025; McAllister et al., 2024; Yu et al., 2024; Zhu et al., 2025b) methods although their derivations may differ. We present the illustration of overall mechanism in Fig. 2.

3.3 Adaptive Update of Preference Score and Negative Embeddings

In Section 3.1, we assumed a frozen diffusion model p_ϕ. However, it can impose limitations on the effectiveness of our proposed preference guidance, especially when 3D generation via score distillation usually requires in thousands of denoising steps. The reason behind is p_ϕ is never updated so that the optimal solution p_ϕ^* can never be approached. On the other hand, since we only perform prompt-specific optimization, it is not rational to update the pretrained model itself even only part of the parameters (e.g. LoRA (Hu et al., 2022) implementations in VSD (Wang et al., 2023)).

Therefore, inspired by ReNeg (Li et al., 2025a), which regards the negative embedding as part of the model parameters, we develop an algorithm which adaptively update the preference score and negative embeddings. Specifically, we initialize the negative embedding with hand-crafted negative descriptors and negative embedding n is updated by maximizing



(a) Reward score of wining sample during optimization

MVDream $lr = 1e^{-4}$ $lr = 4e^{-4}$ $lr = 8e^{-4}$

(b) Results with different learning rate of negative embedding

Figure 3: Effect of negative embedding optimization strategy on single prompt. Employ negative embedding optimization can significantly improve aesthetic score, but overlarge learning rate will harm visual quality.

$$\mathcal{L}_{\text{Emb}}(n) = \mathbb{E}_{\boldsymbol{c}}\left[r(y, \hat{\boldsymbol{x}}_0)\right], \tag{13}$$

To enable training with negative embedding, $\hat{\boldsymbol{x}}_0$ is the same one-step prediction used in Eq. 10, so that negative embedding can be involved while incorporating with CFG. We find it enough to share the same negative embedding for all viewing direction $\boldsymbol{c}$.

For better viewing the effect of our proposed negative embedding optimization technique, Fig. 3 illustrates the curves of target reward score during optimization with different learning rates and their respective results. Comparing the end-point score and the visual quality, larger learning rate will bring benefits of higher score but result in "reward hacking" (Kim et al., 2025b; Gao et al., 2023). Thus, we make several practical trade-off (presented in Appendix H.1). Eventually, we achieve aesthetic score improvements that align with human perception.

4 Experiments

4.1 Experimental Setups

To assess the performance, our experiments include various settings under the same codebase *threestudio* (thr, 2024). For direct comparisons with existing score distillation preference alignment methods, we use 200 test prompts from Eval3d (Duggal et al., 2025) to perform a one-stage distilling of MVDream (Shi et al., 2024). To justify our proposed PSD on high-resolution generation, we further evaluate on several more complex pipelines. In the 2-stage NeRF (Mildenhall et al., 2021) synthesis and 3-stage DMTet (Shen et al., 2021) synthesis, we first distill MVDream and then Stable Diffusion v2.1 (Rombach et al., 2022). Since these pipelines require more optimization time, we evaluate on a more difficult filtered 40-prompt subset presented in Appendix H.6. For target reward, we apply HPSv2.1 (Wu et al., 2023) if without specification. As for evaluation metrics, we assess with human preference reward ImageReward (Xu et al., 2023) (I.R.), PickScore (Kirstain et al., 2023) (Pick.), Aesthetic scores (Schuhmann) (Aes.), Multi-dimensional Preference Score (Zhang et al., 2024) (MPS) and VQA model Qwen2.5-VL-7B (Bai et al., 2025) for text-3D alignment (T.A.) us-

Figure 4: Qualitative comparison of single-stage distillation of MVDream (Shi et al., 2024) (256 × 256). Our PSD significantly improve text alignment (red) and visual quality against comparing method DreamRewad (Ye et al., 2025) and DreamDPO (Zhou et al., 2025).

ing the question-answer pair generated in Eval3d. More implementation details are presented in Appendix H.

4.2 Results

Quantitative comparisons. Shown in Tab. 1 and 2, our PSD outperforms all other methods, which indicates improvements on generation quality as well as alignment with human preference. Specifically, comparing with DreamDPO (Zhou et al., 2025), a method that uses pretrained 2D reward, we achieve a more significant improvements, highlighting our ability to unleash 2D reward. As for cooperating with method that requires additional training represented by DreamReward (Ye et al., 2025), our results still showcases our advantage on text-3D alignment. Besides, Reward3D finetuned in DreamReward enforces 4 input view and is not memory feasible for high-resolution generation. Additionally, we compare with RichDreamer (Qiu et al., 2024), a method that introduce extra multi-view normal, depth, albedo diffusion priors. The results proves that misalignment of human preference commonly exists even when stronger diffusion priors are employed.

Rewards	Methods	I.R. ↑	Pick. ↑	Aes. ↑	MPS ↑	T.A. ↑
HPSV2.1	MVDream	-0.22	20.55	5.79	9.30	53.14
	DreamDPO	-0.28	20.48	5.80	9.00	75.68
	Ours	**0.12**	**20.99**	**5.92**	**10.26**	**75.70**
Reward3D	DreamReward	1.78	21.40	6.15	10.19	74.37
	+Ours	**1.80**	**21.49**	**6.25**	**10.40**	**90.63**

Table 1: Quantitative Comparison of single-stage distillation of MVDream (Shi et al., 2024) across 200 prompts in Eval3d (Duggal et al., 2025) with different rewards. Higher values are better (↑). The best performance is in bold.

Qualitative comparisons. We provide qualitative comparison in Fig. 4 and 5. Baselines and previous methods deviate from given prompt text, while employing our PSD can lead to macroscopic improvements on text alignment (marked in red) and visual details. Also, noticing DreamReward will introduce artifacts, we provide supplementary evaluations of geometry quality in Appendix 13.

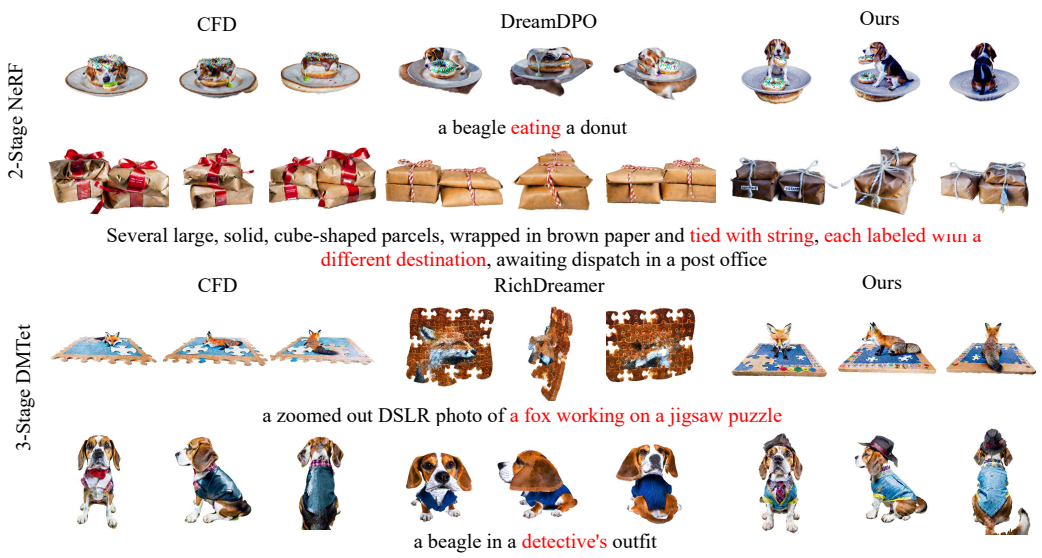


Figure 5: Qualitative comparison of 2-stage NeRF (Mildenhall et al., 2021) (512×512) and 3-Stage DMTet (Shen et al., 2021) (1024×1024) generation. PSD improves alignment with the prompts in red.

Pipeline	Methods	I.R. ↑	Pick. ↑	Aes. ↑	MPS ↑	T.A. ↑
2-Stage NeRF	CFD	-0.09	20.16	5.63	10.03	71.84
	DreamReward			OOM		
	DreamDPO	-0.06	20.18	5.40	10.04	76.18
	Ours	**0.01**	**20.27**	**5.95**	**10.22**	**81.34**
3-Stage DMTet	CFD	-0.44	19.76	5.30	9.45	71.84
	RichDreamer*	**0.02**	19.68	**5.92**	8.24	76.45
	Ours	-0.40	**19.92**	5.34	**9.64**	**81.81**

Table 2: Quantitative Comparison of 2-stage NeRF (Mildenhall et al., 2021) and 3-Stage DMTet (Shen et al., 2021) generation across 40-prompt subset. Higher values are better (↑). The best performance is in bold. Comparison method DreamReward (Ye et al., 2025) suffer from out-of-memory (OOM). * indicates method uses stronger diffusion priors.

Noticing the misalignment of quantitative and qualitative comparisons, we believe this is also due to reward hacking. Limited by the space, more discussion is presented in Appendix G.

User study. To validate the efficacy of our proposed method to real human users, we present a user study involving 24 participants. They are required to choose the better one for each comparison across four dimensions: Appearance Quality, Structure Quality, Text Alignment, and Overall Performance. The survey consists of 30 pairs of videos created from MVDream, DreamDPO, DreamReward, and our PSD. Shown in Fig. 6, our PSD received higher preference score, which is consistent with the results previously given by reward models. More details of this user study is included in Appendix H.3.

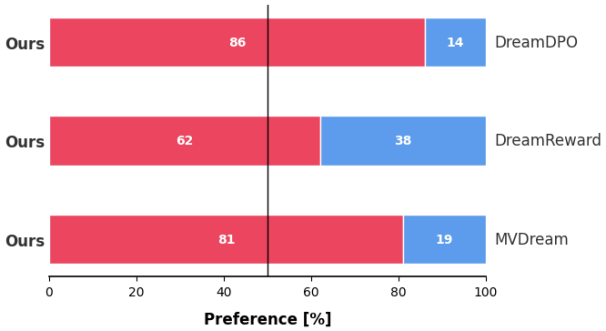


Figure 6: User preference study of comparing PSD with MVDream, DreamDPO, and DreamReward.

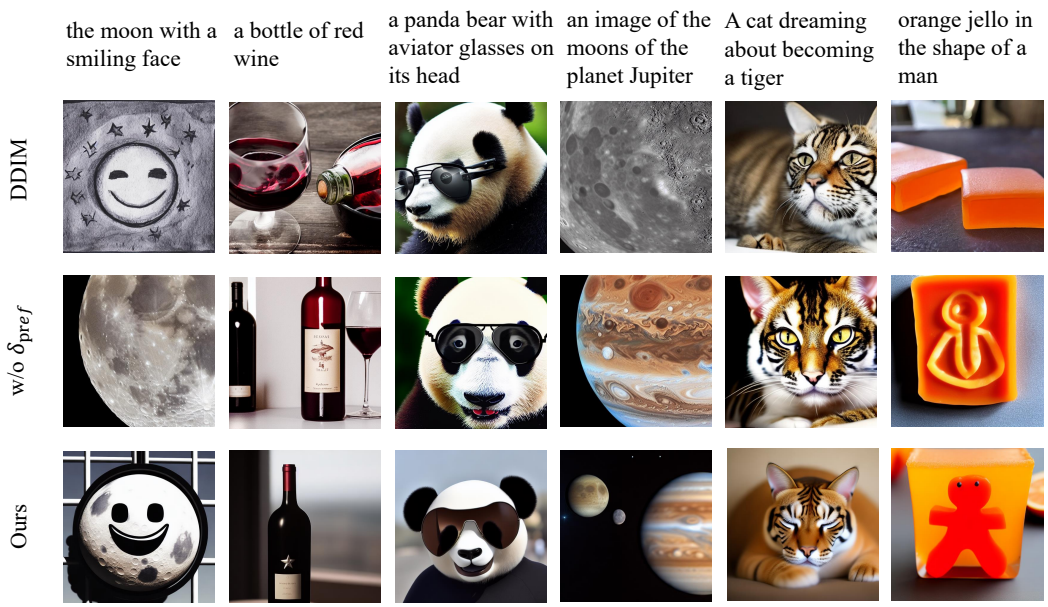


Figure 7: Preference score guidance on image generation.

Methods	Target	Unseen Rewards		
	HPSv2.1 ↑	I.R. ↑	Pick. ↑	MPS ↑
DDIM	0.25	**0.22**	**21.30**	9.77
w/o δ_{pref}	0.25	0.04	21.15	9.51
Ours (Eq. 12)	**0.26**	**0.22**	21.23	**10.30**

Table 3: Ablations of preference score on 2D image generation. Since our final results and experiments without δ_{pref} are performed using optimization, standard DDIM (Song et al., 2021a) are marked in gray.

4.3 Ablation Study

Ablations on negative embedding optimization. In Fig. 3, we analysis the behavior of our negative embedding optimization strategy on single prompt. Quantitative ablation is presented in Tab. 4. We evaluate on the 40-prompt subset with MVDream, PSD with preference score guidance only (w/o n^*) and PSD with negative embedding learning rate $4e^{-4}$ ($lr = 4e^{-4}$). The results verify that our practical trade-off is necessary such that appropriate negative optimization can benefit improving aesthetic scores.

Ablations on different reward models. Theoretically, PSD is compatible with any pretrained 2D reward model. To present the difference, we incorporate ImageReward (Xu et al., 2023). Shown in FIg. 8, different reward models may benefit performance (capturing concept like "streaming" and "growing on a log"), but for the sake of fair comparison, we don't introduce any new models and inherit from the comparison methods instead.

Ablations on the preference score. To justify the compatibility of our preference guidance with PF-ODE, we

Figure 8: Ablation of choosing different reward models.

Experiments	Target	Unseen Rewards			Generation Speed
	HPSv2.1 ↑	I.R. ↑	Pick. ↑	MPS ↑	
MVDream	0.20	-0.63	19.22	8.39	**2.92 it/s**
DreamDPO	0.19	-0.70	19.15	8.58	1.45 it/s
w/o n	0.20	-0.60	19.28	8.51	1.45 it/s
$lr = 4e^{-4}$	**0.24**	-0.49	19.85	9.98	1.19 it/s
$1/5it^{-1}$	0.21	-0.38	19.45	9.35	2.45 it/s
$1/2it^{-1}$	0.22	-0.25	19.72	9.70	1.50 it/s
Ours	0.23	**-0.25**	**19.99**	**10.12**	1.19 it/s

Table 4: Ablations on negative embedding optimization strategy. w/n represents experiments without negative embedding optimization. $lr = 4e^{-4}$ represents experiments with higher learning rate of learnable embedding. $1/5it^{-1}$ and $1/2it^{-1}$ represent performing preference guidance and negative embedding optimization with a interval of 2 and 5 steps repectively.

directly use Eq. 12 to guide the diffusion process of parameterized images. Following common configurations of DDIM (Song et al., 2021a), we update the latents 50 steps but with a Adam (Kingma & Ba, 2014) optimizer. This is based on the fact that image can also be parametrized (Kim et al., 2025a; Luo et al., 2025). Specially, to share the same dynamics with DDIM, we replace one of $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}'$ in Eq. 11 with the noise $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_s, y, s)$ predicted at more noisy timestep s as discussed in (Lukoianov et al., 2024). Evaluations with Stable Diffusion v1.5 on Parti-Prompt dataset (Yu et al., 2022) is presented in Tab. 3 and direct visual comparison is shown in Fig. 7 . Comparing with results of DDIM and optimizing Eq. 12 without δ_{pref}, introducing the preference guidance will significantly improve reward scores and even outperform the original DDIM. It is a strong evidence to prove that our preference guidance can lead the generation process towards high-reward regions. Besides, comparing the scores of target and unseen rewards, using preference guidance alone will reduce the risk of reward hacking. Pseudo code of this experiment is listed in Appendix H.5.

Time Consumption. Since the preference guidance is formulated into a CFG-style term, it can be calculated within one forward pass of diffusion model. As a consequence, if the reward model also supports batch inference, the only additional computational consumption for each step is to update the negative embedding. Considering that the size of negative embedding is comparatively small, times consumption should be acceptable. Results are in Tab. 4, where we provide two additional settings. A larger interval between adjacent reward signal leads to significant accelerations while our method still outperforms baselines, showing the efficacy of our proposal.

5 Conclusion

In this paper, we propose Preference Score Distillation, which basically implies that preference optimization in score distillation can be regarded as a type of guidance. To link preference with guidance, we start by deriving preference score guidance under our modified definition of RLHF. By constructing win-lose pair on-the-fly, we achieve effective guidance to the optimization process of score distillation. We also develop an adaptive update strategy to unleash the potential of preference score guidance, noticing the parameters of pretrained diffusion models are not updated. During the entire procedure, we successfully avoid additional training of reward models using 3d data, and demonstrate significant improvements on generating highly photorealistic, human preference aligned 3d objects.

Limitations. Inheriting computational demands from prior score distillation methods, PSD optimization requires one to several hours per generation, limiting real-time applicability. While this work primarily investigates differentiable rewards, future work should explore non-differentiable objectives (e.g., vision-language model ensembles). Although PSD elevates aesthetic metrics, its performance remains bounded by the reward model's capabilities, risking distribution shift or reward hacking. Moreover, as PSD's output depends directly on the reward model, implementing content filtering modules is essential to prevent malicious content generation.

REFERENCES

threestudio-project/threestudio, 06 2024. URL https://github.com/threestudio-project/threestudio.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *Computer Vision – ECCV 2024*, pp. 190–206, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73024-5.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *International Conference on Learning Representations*, 2025.

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)*, pp. 22246–22256, October 2023.

Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21401–21412, June 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Shivam Duggal, Yushi Hu, Oscar Michel, Aniruddha Kembhavi, William T. Freeman, Noah A. Smith, Ranjay Krishna, Antonio Torralba, Ali Farhadi, and Wei-Chiu Ma. Eval3d: Interpretable and fine-grained evaluation for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13326–13336, June 2025.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885. Curran Associates, Inc., 2023.

Stephanie Fu, Netanel Y. Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Chun Gu, Zeyu Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Tetrahedron splatting for 3d generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 80165–80190. Curran Associates, Inc., 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

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.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chenhan Jiang, Yihan Zeng, Tianyang Hu, Songcun Xu, Wei Zhang, Hang Xu, and Dit-Yan Yeung. Jointdreamer: Ensuring geometry consistency and text congruence in text-to-3d generation via joint score distillation. In *Computer Vision – ECCV 2024*, pp. 439–456, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73347-5.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022.

Jeongsol Kim, Geon Yeong Park, and Jong Chul Ye. Dreamsampler: Unifying diffusion sampling and score distillation for image manipulation. In *Computer Vision – ECCV 2024*, pp. 398–414, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-73007-8.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023.

Min-Seop Kwak, Donghoon Ahn, Inès Hyeonsu Kim, Jin-Hwa Kim, and Seungryong Kim. Geometry-aware score distillation via 3d consistent noising and gradient consistency modeling. *arXiv preprint arXiv:2406.16695*, 2024.

Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, and Emad Barsoum. Reneg: Learning negative embedding with reward guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23636–23645, June 2025a.

Xuan Li, Qianli Ma, Tsung-Yi Lin, Yongxin Chen, Chenfanfu Jiang, Ming-Yu Liu, and Donglai Xiang. Articulated kinematics distillation from video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17571–17581, June 2025b.

Zongrui Li, Minghui Hu, Qian Zheng, and Xudong Jiang. Connecting consistency distillation to score distillation for text-to-3d generation. In *Computer Vision – ECCV 2024*, pp. 274–291, Cham, 2025c. Springer Nature Switzerland. ISBN 978-3-031-72775-7.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, June 2023.

Fangfu Liu, Junliang Ye, Yikai Wang, Hanyang Wang, Zhengyi Wang, Jun Zhu, and Yueqi Duan. Dreamreward-x: Boosting high-quality 3d generation with human preference alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2025a. doi: 10.1109/TPAMI.2025.3609680.

Gaofeng Liu, Zhiyuan Ma, and Tao Fang. Dreamalign: Dynamic text-to-3d optimization with human preference alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5424–5432, 2025b.

Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. Inpo: Inversion preference optimization with reparametrized ddim for efficient diffusion model alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28629–28639, June 2025a.

Yunhong Lu, Qichao Wang, Hengyuan Cao, Xiaoyin Xu, and Min Zhang. Smoothed preference optimization via renoise inversion for aligning diffusion models with varied human preferences. *arXiv preprint arXiv:2506.02698*, 2025b.

Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Campagnolo Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin Solomon. Score distillation via reparametrized ddim. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26011–26044. Curran Associates, Inc., 2024.

Yihong Luo, Tianyang Hu, Weijian Luo, Kenji Kawaguchi, and Jing Tang. Reward-instruct: A reward-centric approach to fast photo-realistic image generation. *arXiv preprint arXiv:2503.13070*, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33779–33804. Curran Associates, Inc., 2024.

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. *Commun. ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250.

Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17060–17069, June 2025.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023.

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 (CVPR)*, pp. 9914–9925, June 2024.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023.

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687625.

Christoph Schuhmann. Laion-aesthetics — laion. URL `https://laion.ai/blog/laion-aesthetics/`.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6087–6101. Curran Associates, Inc., 2021.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International Conference on Learning Representations*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *Computer Vision – ECCV 2024*, pp. 1–18, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73235-5.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024.

Uy Dieu Tran, Minh Luu, Phong Ha Nguyen, Khoi Nguyen, and Binh-Son Hua. Diverse text-to-3d synthesis with augmented text embedding. In *Computer Vision – ECCV 2024*, pp. 217–235, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73226-3.

Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7882–7891, 2024a.

Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024b.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025a.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21096–21106, 2025b.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. *arXiv preprint arXiv:2507.15064*, 2025c.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7280–7290, October 2023.

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 (CVPR)*, pp. 8228–8238, June 2024.

Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. In *International Conference on Learning Representations*, 2025a.

Fu-Yun Wang, Keqiang Sun, Yao Teng, Xihui Liu, Jiaming Song, and Hongsheng Li. Self-npo: Negative preference optimization of diffusion models by simply learning from itself without explicit preference annotations. *arXiv preprint arXiv:2505.11777*, 2025b.

Ruochen Wang, Ting Liu, Cho-Jui Hsieh, and Boqing Gong. On discrete prompt optimization for diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50992–51011. PMLR, 21–27 Jul 2024.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 8406–8441. Curran Associates, Inc., 2023.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023.

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

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, June 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023.

Runjie Yan, Kailu Wu, and Kaisheng Ma. Flow score distillation for diverse text-to-3d generation. *arXiv preprint arXiv:2405.10988*, 2024.

Runjie Yan, Yinbo Chen, and Xiaolong Wang. Consistent flow distillation for text-to-3d generation. In *International Conference on Learning Representations*, 2025.

Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Zuxuan Wu, Yu-Gang Jiang, and Tao Mei. Dreammesh: Jointly manipulating and texturing triangle meshes for text-to-3d generation. In *Computer Vision – ECCV 2024*, pp. 162–178, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-73202-7.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8941–8951, June 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, June 2024b.

Xiaofeng Yang, Yiwen Chen, Cheng Chen, Chi Zhang, Yi Xu, Xulei Yang, Fayao Liu, and Guosheng Lin. Learn to optimize denoising scores: A unified and improved diffusion prior for 3d generation. In *Computer Vision – ECCV 2024*, pp. 136–152, Cham, 2025b. Springer Nature Switzerland. ISBN 978-3-031-72784-9.

JunLiang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference. In *Computer Vision – ECCV 2024*, pp. 259–276, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72897-6.

Po-Hung Yeh, Kuang-Huei Lee, and Jun-cheng Chen. Training-free diffusion model alignment with sampling demons. In *International Conference on Learning Representations*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487. Curran Associates, Inc., 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, June 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, June 2025.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and XIAOJUAN QI. Text-to-3d with classifier score distillation. In *International Conference on Learning Representations*, 2024.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027, June 2024.

Linqi Zhou, Andy Shih, Chenlin Meng, and Stefano Ermon. Dreampropeller: Supercharge text-to-3d generation with parallel sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4610–4619, June 2024.

Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. *arXiv preprint arXiv:2502.04370*, 2025.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. In *International Conference on Learning Representations*, 2025a.

Jiahao Zhu, Zixuan Chen, Guangcong Wang, Xiaohua Xie, and Yi Zhou. Segmentdreamer: Towards high-fidelity text-to-3d synthesis with segmented consistency trajectory distillation. *arXiv preprint arXiv:2507.05256*, 2025b.

Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *International Conference on Learning Representations*, 2024.

Xiandong Zou, Ruihao Xia, Hongsong Wang, and Pan Zhou. Dreamcs: Geometry-aware text-to-3d generation with unpaired 3d reward supervision. *arXiv preprint arXiv:2506.09814*, 2025.

A Disclosure of LLM Usage

In this work, large language model is limited as a general-purpose writing assistive tool. The model only supported us in improving the clarity and readability of the manuscript through suggestions on phrasing and style. It did not contribute to data collection, data analysis, modeling, proofs, experiments, or interpretation of results. All technical content, algorithms, and experimental designs were conceived and validated solely by the authors, and all outputs from the LLM were critically reviewed, revised, or discarded before inclusion. No confidential, proprietary, or non-public data were provided to the LLM, and the authors remain fully responsible for the integrity and accuracy of the work.

B Reproducibility Statement

To facilitate reproducibility, we refer readers to the following resources: the full algorithm description and training procedure in Sections 3,4 of the main text; complete hyperparameter settings, optimization details, and evaluation protocols in Appendix H; anonymous, scripts and configuration files provided in supplementary materials (.zip file); explicit assumptions and complete proofs for theoretical claims in Appendix F. Any deviations from default settings are noted in the code and cross-referenced in the main text. Together, these materials are intended to enable independent verification and replication of our results.

C Ethics Statement

This research was conducted in accordance with the ICLR Code of Ethics. The work does not involve human subjects, personally identifiable information, or sensitive data. All datasets used are publicly available and properly cited, and no proprietary or restricted-access data were employed. Care was taken to ensure that the proposed methods do not intentionally propagate or amplify harmful content, bias, or discrimination. The code and datasets will be released in an anonymized form to support transparency, reproducibility, and accountability. The authors affirm compliance with all relevant ethical standards and accept responsibility for the integrity and potential impact of the research.

D Table of Notations

We use consistent notations across the main paper and supplementary materials, which are listed in Tab. 5

E Related Works

E.1 Diffusion Models

Diffusion model (Ho et al., 2020; Song et al., 2021b;a) is a class of generative models that learn to reverse a diffusion process which gradually adding noise to a data distribution. With the introduction of latent space (Rombach et al., 2022), DM has proven its scaling ability to generate high-dimensional, perceptual data such as image Esser et al. (2024); Sauer et al. (2024) and videos Blattmann et al. (2023). Among the dense theory behind, interrelating the diffusion process into a *Probability Flow Ordinary Differential Equation* (PF-ODE) or *Stochastic Differential Equation* (SDE) (Song et al., 2021b) is an essential step towards a unified framework in pursuit of mathematically guaranteed high-quality generation. In representation, *Stable Diffusion* (Rombach et al.,

Notation	Description
$\boldsymbol{x}_t$	State variable at timestep t
α_t, σ_t	Time-dependent coefficients
ϵ, ϵ'	Noise sampled from Gaussian distribution
$\boldsymbol{\epsilon}_\phi(\cdot)$	Diffusion models parameterized by ϕ
y	Conditioning text prompt embedding
n	Negative prompt embedding
γ	Scaling factor in CFG
β	Scaling factor in RLHF
r	Reward
$g_\theta(\cdot)$	Differential 3D representation parameterized by θ
$q_\theta(\cdot)$	Marginal distribution determined by parameter θ
$\boldsymbol{c}$	Rendering camera view
$\mathbb{D}_{\text{KL}}[\cdot]$	KL-divergence
$\mathcal{S}_{\text{pref}}$	Binary variable represents human preference
$\boldsymbol{x}_c$	Non-noisy sample
$\hat{\boldsymbol{x}}_0$	One-step prediction sample based on Tweedie's formula $\hat{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - \sigma_t \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t)}{\alpha_t}$
$\boldsymbol{x}_t^w, \boldsymbol{x}_t^l$	Win-lose sample pair at timestep t
β_r	Adaptive scaling factor $\beta_r = \gamma \frac{\|\|\delta_{cls}\|\|_2}{\|\|\delta_{pref}\|\|_2} \cdot \sigma\left(r\left(y, \boldsymbol{x}_t^w\right) - r\left(y, \boldsymbol{x}_t^l\right)\right)$

Table 5: List of notations and their descriptions.

2022) has developed a stack of variations that significantly promote the quality and efficiency of photorealistic generation.

E.2 Preference Alignment of Diffusion Models

Although web-scale pretraining enables promising performance to diffusion models, they may deviate users' preference. To overcome the issue, a line of works focuses on using *Reinforcement Learning from Human Feedback* (RLHF) for fine-tuning. For example, Wu et al. (2023) trained a human preference classifier to employ supervised fine-tuning with preference-based reward model. Xu et al. (2023) also trained reward model but they directly maximize reward through backpropagation of differentiable scores. DPOK (Fan et al., 2023) and DDPO (Black et al., 2024) apply policy gradient to the sampling process of diffusion models modeled by Markov decision process. However, the major drawback reward over-optimization (Kim et al., 2025b; Gao et al., 2023) which may harm generation quality and diversity. Then, impacted by the success of Direct Preference Optimization (DPO) (Rafailov et al., 2023) in large language models (LLMs), Diffusion-DPO (Wallace et al., 2024; Yang et al., 2024a) and D3PO adopt denoising steps of diffusion models to perform DPO. Furthermore, several recent works adjust the DPO objective with the essential characteristics of diffusion models. DSPO (Zhu et al., 2025a) modifies the time-dependent score matching objectives to distill the score function of human-preferred image distributions into pretrained score functions. InPO Lu et al. (2025a) and SmPO-Diffusion (Lu et al., 2025b) applies DDIM inversion technique in response to the challenge of the implicit rewards allocation in the long-chain denoising process. Diffusion-NPO (Wang et al., 2025a) and Self-NPO (Wang et al., 2025b) address the efficacy of classifier-free guidance (CFG) and train a model attuned to negative preferences to bias away from the negative-conditional inputs.

Another line of work focuses on test-time alignment. DOODL (Wallace et al., 2023) directly optimizes the diffusion latents at each timestep through the backpropagation of the reward model. Yeh et al. (2025) seeks to synthesize theocratically optimal noises based on multiple evaluations. DNO Tang et al. (2024) turns to optimize the initial noise and develops a zeroth-order optimization algorithm for non-differential rewards. Unfortunately, all these improvements come with significant overload. Due to the special property of score distillation, our work provides a new perspective for preference alignment.

E.3 TEXT-TO-3D GENERATION

The area this work belongs is distilling 2D into 3D. Based on the success of text-to-image diffusion models, Score Distillation Sampling (SDS) (Poole et al., 2023) was first proposed to distill a pretrained diffusion model $\boldsymbol{\epsilon}_\phi$ to generate 3D assets. Instead of guiding the optimization of differentiable 3D shape representation with PF-ODE (SDE) in the main paper, SDS aims to find modes of the score functions across all noise levels. Following the notations in the main paper, it can be expressed as updating the 3D representation with

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta) &= \nabla_\theta \mathbb{E}_{t,\boldsymbol{c},\boldsymbol{\epsilon}} \left[\sigma_t/\alpha_t w(t) \text{KL}\left(q(\mathbf{z}_t|g_\theta(\boldsymbol{c}); y, t) \| p_\phi(\mathbf{z}_t; y, t)\right)\right] \\ &= \mathbb{E}_{t,\boldsymbol{c},\boldsymbol{\epsilon}} \left[w(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \boldsymbol{\epsilon}) \frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta} \right].\end{aligned} \tag{14}$$

Many followup works devote to improve the behavior of SDS from many aspects, including improving view-consistency (multi-face Janus problem) via introducing stronger diffusion priors (Jiang et al., 2025; Qiu et al., 2024), boosting geometry quality through coarse-to-fine training (Chen et al., 2023; Lin et al., 2023; Yang et al., 2025a), and accelerating generation process by applying more advanced 3D representation (Gu et al., 2024; Chen et al., 2024) or parallelization (Zhou et al., 2024). Moreover, despite advancing technically, several recent works build connections between score distillation and PF-ODE through deterministic noising (Wu et al., 2024; Yan et al., 2025; Lukoianov et al., 2024; Yan et al., 2024) or consistency training (Li et al., 2025c; Zhu et al., 2025b). Comparing with mode-seeking objective in Eq. 14, these methods yield significant improvement on fidelity and diversity. In this perspective, our work aims to seek minimum conflicts to these advancements while aligning with preference, but for other relevant existing works, they basically only consider the derivation from Eq. 14. DreamReward (Ye et al., 2025; Liu et al., 2025a) fine-tunes a reward model from ImageReward (Xu et al., 2023) to approximate the shift towards an ideal noise prediction network aligned with human preference. DreamAlign (Liu et al., 2025b) trains a reference noise prediction network using proposed D-3DPO algorithm and derive a preference contrastive loss. DreamDPO (Zhou et al., 2025) completely eliminate the use of reference model, but we find it unstable and will do harm to fidelity. Concurrent work DreamCS (Zou et al., 2025) address the geometry alignment issue through training a new reward model, but it is still under the framework of DreamReward.

Despite generating 3D assets through distilling 2D diffusion priors, other prevailing paradigms such as leveraging large reconstruction models (Hong et al., 2023; Tang et al., 2025) or native 3D generation (Xiang et al., 2025) also shows attractive performance especially on speed and geometry. However, due to the lack of high quality data and expensive computational demands, preference alignment have barely been explored in these fields.

F DERIVATION DETAILS OF PREFERENCE SCORE GUIDANCE

F.1 FROM EQ.8 TO EQ. 12

To get the expansion in Eq. 12, the core is to handle term $\nabla_{x_t}\left(r\left(y, \boldsymbol{x}_t\right) - r\left(y, \boldsymbol{x}_t'\right)\right)$. Our solution is to use the implicit reward rewritten by the global optimal solution p_ϕ^* in Eq. 9, which is

$$\nabla_{x_t}\left(r\left(y, \boldsymbol{x}_t^w\right) - r\left(y, \boldsymbol{x}_t^l\right)\right) = \beta \nabla_\theta (\log \frac{p_\phi^*(\boldsymbol{x}_t^w|y)}{q_\theta(\boldsymbol{x}_t^w|\boldsymbol{x}_c = g_\theta(\boldsymbol{c}))} - \log \frac{p_\phi^*(\boldsymbol{x}_t^l|y)}{q_\theta(\boldsymbol{x}_t^l|\boldsymbol{x}_c = g_\theta(\boldsymbol{c}))}). \tag{15}$$

As already discussed in (Poole et al., 2023), $\nabla_\theta p_\phi(\boldsymbol{x}_t|y)$ is the definition of score function and $\nabla_\theta q_\theta(\boldsymbol{x}_t|\boldsymbol{x}_0 = g_\theta(\boldsymbol{c}))$ is the gradient of the entropy of the forward process with respect to the mean parameter, which is

$$\begin{aligned}&\nabla_\theta \log q_\theta(\boldsymbol{x}_t|\boldsymbol{x}_0) = 0 \\ &\nabla_\theta \log p_\phi(\boldsymbol{x}_t|y) = -\frac{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t)}{\sigma_t} \frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta},\end{aligned} \tag{16}$$

If we approximate p_ϕ^* with p_ϕ for each step, then put everything together in Eq. 8, we have

$$\begin{aligned}
\mathrm{d}(\frac{\boldsymbol{x}_t}{\alpha_t}) &= \mathrm{d}(\frac{\sigma_t}{\alpha_t}) \cdot (\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \sigma_t(1-\sigma(r(y,\boldsymbol{x}_t^{\mathrm{w}}) - r(y,\boldsymbol{x}_t^{\mathrm{l}}))\nabla_{\boldsymbol{x}_t}(r(y,\boldsymbol{x}_t^{\mathrm{w}}) - r(y,\boldsymbol{x}_t^{\mathrm{l}}) \\
&= \mathrm{d}(\frac{\sigma_t}{\alpha_t})(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \beta\sigma_t\sigma(r(y,\boldsymbol{x}_t^{\mathrm{l}}) - r(y,\boldsymbol{x}_t^{\mathrm{w}}))(-\frac{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{w}}, y, t)}{\sigma_t} - (-\frac{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{l}}, y, t)}{\sigma_t}))) \\
&= \underbrace{\mathrm{d}(\frac{\sigma_t}{\alpha_t})}_{-lr} \cdot \underbrace{(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) + \beta\sigma\left(r(y,\boldsymbol{x}_t^{\mathrm{l}}) - r\left(y,\boldsymbol{x}_t^{\mathrm{w}}\right)\right)(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{w}}, y, t) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{l}}, y, t))\Big)}_{\nabla\mathcal{L}}.
\end{aligned} \tag{17}$$

Mechanistically speaking, our preference score guidance perfectly align with DPO (Rafailov et al., 2023) where gradients are in the direction of increasing the likelihood of preferred samples, but more importantly, we successfully build connection between preference and guidance so that our PSD is performed without additional training. Also, thanks to the flexibility of score distillation, we are able to get access to different noise at each denoising step, which enables to implement the crucial win-lose score prediction in our preference score guidance.

Change-of-variable technique. Considering the chain rule in Eq. 16, it is intuitive to replace $\boldsymbol{x}_t$ with $g_\theta(\boldsymbol{c})$ and directly use $\nabla\mathcal{L}$ to update 3D representation. However, as discussed in many previous works (Lukoianov et al., 2024; Yan et al., 2025; Wu et al., 2024), since the rendering images are non-noisy, it will suffer form out-of-distribution issue. Therefore, they develop several noising techniques including fixed noise, DDIM inversion and integral noise. Discussing these noising techniques is out of the scope of this paper, but it is necessary to point out the change-of-variable techniques in our PSD is applied to Eq. 5 and results in affecting the unconditional term so that our objective Eq. 12 can be formulated in a general form without the need of further discussion. Meanwhile, based on the above discussion, a great advantage of our PSD is it can combine with these noising techniques seamlessly which has been shown previously in the ablation of preference score for 2D image generation. For 3D generation, we test with integral noise (Kwak et al., 2024; Chang et al., 2025) modified by CFD (Yan et al., 2025) in the 2-stage NeRF and 3-stage DMTet generation pipeline where wining noise (one of $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}'$) is replaced with

$$G(\mathbf{p}) = \frac{1}{\sqrt{|\Omega_{\mathbf{p}}|}} \sum_{A_i \in \Omega_{\mathbf{p}}} W(A_i). \tag{18}$$

For the query pixel $\mathbf{p}$, $\Omega_{\mathbf{p}} = \mathcal{T}^{-1}(ctw_c(\mathbf{p}))$ is covered by after $\mathbf{p}$ is warped to a pre-set constant reference noise. For more details of this algorithm, interested readers may refer to the original paper of CFD (Yan et al., 2025), what's important is our experiments shown in Tab. 2 proves our compatibility with these nosing techniques.

F.2 Justification of Definition in Eq. 4

Our derivation heavily rely on the definition in Eq. 4. Therefore, discussing its validity is of first priority. We justify it by showing two representative works can also be related to this definition.

DreamReward (Ye et al., 2025):

$$\nabla_\theta\mathcal{L}_{\text{DreamReward}}(\theta) = \mathbb{E}_{t,\boldsymbol{c},\boldsymbol{\epsilon}}\left[\omega(t)\left(\epsilon_\phi(\boldsymbol{x}_t, y, t) - \lambda_r\frac{\partial r_{Reward3D}(y, \hat{\boldsymbol{x}}_0, \boldsymbol{c})}{\partial g_\theta(\boldsymbol{c}))} - \epsilon\right)\frac{\partial g_\theta(\boldsymbol{c})}{\partial\theta}\right]. \tag{19}$$

DreamReward fine-tunes a view-dependent reward model Reward3D which approximates the difference between optimal diffusion model $\boldsymbol{\epsilon}_\phi^*$ and current diffusion model $\boldsymbol{\epsilon}_\phi$ with

$$\boldsymbol{\epsilon}_\phi^*(\boldsymbol{x}_t, y, t)) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t)) = \frac{\partial r_{Reward3D}(\boldsymbol{x}_t, y, t)}{\partial g_\theta(\boldsymbol{c})} = -\frac{\partial r(\boldsymbol{x}_t, y, t)}{\partial g_\theta(\boldsymbol{c})} \tag{20}$$

Note that the purpose of our definition in Eq. 4 is to formulate the preference guidance through implicit reward based on the connection between score distillation and diffusion process built by recent works (Kwak et al., 2024; Yan et al., 2025; Lukoianov et al., 2024). However, DreamReward follows

the framework of SDS, so in order to directly formulate a objective to optimize 3D representation, we make a revision to Eq. 4:

$$\max_{\theta} \mathbb{E}_{t,\boldsymbol{c},\boldsymbol{\epsilon}}[r(y,\boldsymbol{x}_t)] - \beta \mathbb{D}_{\mathrm{KL}}[p_\phi(\boldsymbol{x}_t|y)||q_\theta(\boldsymbol{x}_t|\boldsymbol{x}_0 = g_\theta(\boldsymbol{c}))], \tag{21}$$

Difference is the expectation we take is with respect to the added noise and parameters we tune is 3D representation θ. This modification enables the objective to seek modes of the score functions, which fits in the framework of original SDS. After that, calculating its gradients directly (apply Sticking-the-Landing (Roeder et al., 2017) type gradient suggested in (Poole et al., 2023) and set $\lambda_r = 1/\beta$) leads to

$$\begin{aligned}
&\nabla_\theta \mathbb{E}_t[r(y,\boldsymbol{x}_t)] - \beta \nabla_\theta \mathbb{D}_{\mathrm{KL}}[p_\phi(\boldsymbol{x}_t|y)||q_\theta(\boldsymbol{x}_t|\boldsymbol{x}_0 = g_\theta(\boldsymbol{c}))] \\
&= \mathbb{E}_t[\frac{\partial r(y,\boldsymbol{x}_t)}{\partial g_\theta(\boldsymbol{c})}\frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}] - \beta \mathbb{E}_{t,\boldsymbol{c}}\left[w(t)(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t))\frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}\right] \\
&= \mathbb{E}_{t,\boldsymbol{c}}\left[w(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) + \lambda_r \frac{\partial r(y,\boldsymbol{x}_t)}{\partial g_\theta(\boldsymbol{c})} - \boldsymbol{\epsilon})\frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}\right] \\
&= \mathbb{E}_{t,\boldsymbol{c}}\left[w(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \lambda_r \frac{\partial r_{Reward3D}(\boldsymbol{x}_t, y, t)}{\partial g_\theta(\boldsymbol{c})} - \boldsymbol{\epsilon})\frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}\right]
\end{aligned} \tag{22}$$

DreamDPO (Zhou et al., 2025):

$$\nabla_\theta \mathcal{L}_{\mathrm{DreamDPO}}(\theta) = \mathbb{E}_{t,\boldsymbol{c}}\left[w(t)\left((\epsilon_\phi(\boldsymbol{x}_t^{\mathrm{w}}, y, t) - \epsilon^{\mathrm{w}}) - (\epsilon_\phi(\boldsymbol{x}_t^{\mathrm{l}}, y, t) - \epsilon^{\mathrm{l}})\right)\frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}\right] \tag{23}$$

We ignore the hyperparameter τ in (Zhou et al., 2025) since it is used to resolve numerical instability. Similar to the idea of Diffusion-DPO (Wallace et al., 2024), $r\left(y, \boldsymbol{x}_t\right)$ can also be parameterized by a neural network Φ and estimated via maximum likelihood training for binary classification

$$\mathcal{L}_{\mathrm{BT}}(\Phi) = -\mathbb{E}_{y,\boldsymbol{x}_t^{\mathrm{win}},\boldsymbol{x}_t^{\mathrm{lose}}}\left[\log \sigma\left(r_\Phi(y, \boldsymbol{x}_t^{\mathrm{win}}) - r_\Phi(y, \boldsymbol{x}_t^{\mathrm{lose}})\right)\right], \tag{24}$$

and plug the implicit reward rewritten by the global optimal solution p_ϕ^* in Eq. 9, so that we get

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\mathrm{BT}}(\theta) &= -\mathbb{E}_{y,\boldsymbol{x}_t^{\mathrm{w}},\boldsymbol{x}_t^{\mathrm{l}}}[\sigma\left(r\left(y, \boldsymbol{x}_t^{\mathrm{l}}\right) - r\left(y, \boldsymbol{x}_t^{\mathrm{w}}\right)\right) \\
&\quad \cdot \beta \nabla_\theta (\log \frac{p_\phi^*(\boldsymbol{x}_t^w|y)}{q_\theta(\boldsymbol{x}_t^w|\boldsymbol{x}_c = g_\theta(\boldsymbol{c}))} - \log \frac{p_\phi^*(\boldsymbol{x}_t^l|y)}{q_\theta(\boldsymbol{x}_t^l|\boldsymbol{x}_c = g_\theta(\boldsymbol{c}))})] \\
&= -\mathbb{E}_{y,\boldsymbol{x}_t^{\mathrm{w}},\boldsymbol{x}_t^{\mathrm{l}}}[\sigma(r(y, \boldsymbol{x}_t^{\mathrm{l}}) - r(y, \boldsymbol{x}_t^{\mathrm{w}}))(-\frac{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{w}}, y, t)}{\sigma_t} - (-\frac{\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{l}}, y, t)}{\sigma_t}))] \\
&= \mathbb{E}_{y,\boldsymbol{x}_t^{\mathrm{w}},\boldsymbol{x}_t^{\mathrm{l}}}[\sigma(r(y, \boldsymbol{x}_t^{\mathrm{l}}) - r(y, \boldsymbol{x}_t^{\mathrm{w}}))(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{w}}, y, t) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{\mathrm{l}}, y, t))]
\end{aligned} \tag{25}$$

Then, as suggested by (Poole et al., 2023), if we use Sticking-the-Landing type gradient to control variate by keeping the noise added in Eq. 11, we have an objective similar to Eq. 23. The only difference is the weight $\sigma(r(y, \boldsymbol{x}_t^{\mathrm{l}}) - r(y, \boldsymbol{x}_t^{\mathrm{w}}))$.

G Comparisons with Existing Methods

Comparison with SDS. Although our derivation is from the perspective of PF-ODE in Eq. 2, as already discussed in many previous paper Lukoianov et al. (2024); Yan et al. (2025; 2024); Kim et al. (2025a) SDS can be treated as a special case. Therefore, Eq. 12 can cover all these variants. The main difference is we derive an additional preference score guidance while SDS only consists δ_{gen} and δ_{cls}. Also, in Eq. 15, we don't additionally apply a Sticking-the-Landing type gradient since it is only a guidance.

Comparison with DreamReward. In Eq. 22, DreamReward can be interpenetrated as the gradient of our defined RLHF objective , which implies it's also introducing additional guidance to the generation process. In the formulation of ODE, it can be not strictly written as (omit time-dependent coefficients as well)

$$\mathrm{d}(\frac{\boldsymbol{x}_t}{\alpha_t}) = \mathrm{d}(\frac{\sigma_t}{\alpha_t}) \cdot (\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, y, t) - \nabla_{\boldsymbol{x}_t} r_{Reward3D}(y, \boldsymbol{x}_t, \boldsymbol{c}))\,. \tag{26}$$

Obviously, comparing Eq. 26 with classifier guidance in (Dhariwal & Nichol, 2021), the reward model provides extra guidance that requires pixel-wise gradient directly operating on $\boldsymbol{x}_t$. This enforces retraining of the reward model and is defective when a) 3d data is rare, b) intermediate noisy steps. In contrast, our PSD overcomes this issue completely. At each timestep, two terms in preference score guidance are both the output of the pretrained diffusion models, and the negative embedding optimization strategy also avoids directly operating on $\boldsymbol{x}_t$. Consequently, our results presents much less artifacts. In addition, this perspective directly reveal the source of reward hacking in DreamReward.

Comparison with DreamDPO. Apart from the mainstream derivation we present in the main paper, the novelty of our approach can be supported by another intuition given by the connection between CSD (Yu et al., 2024) and DreamDPO under our framework. While comparing with DreamReward, we can easily write Eq.26, but for DreamDPO, its connection with denoising trajectory is ambiguous. The breakthrough is from the motivation of CSD. CSD notices SDS is heavily relied on high CFG value, so they investigate and discover that using term δ_{cls} $(p(y|\boldsymbol{x}_t))$ alone is enough to generate 3D assets. While under our derivation, DreamDPO is actually $p\left(\mathcal{S}_{\text{pref}} \mid \boldsymbol{x}_t, y\right)$, so this explains the mechanism of DreamDPO in another important perspective.

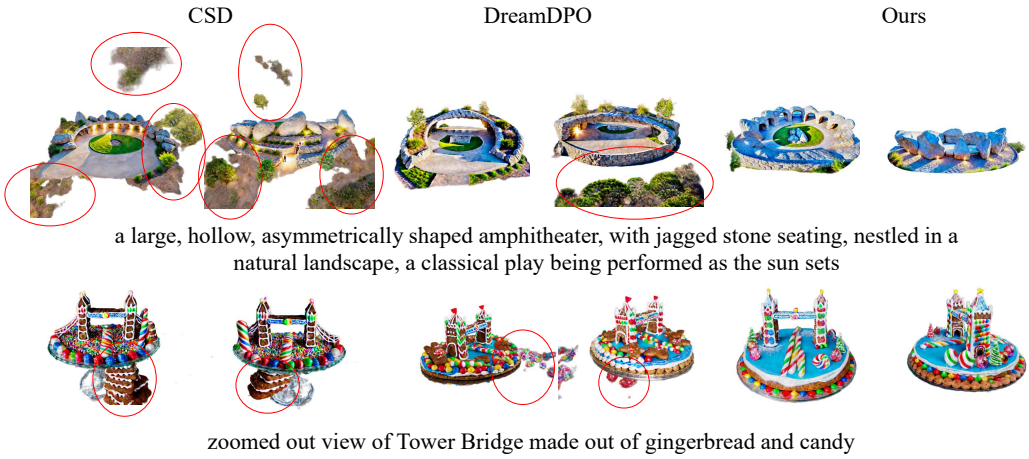


Figure 9: Comaprsion with CSD (Yu et al., 2024) and DreamDPO (Zhou et al., 2025). Artifacts are marked with red circles.

To verify this claim, we perform a simple comparison. A major drawbacks of CSD is it will also result in artifacts (Yu et al., 2024; McAllister et al., 2024). Shown in Fig. 9, CSD and DreamDPO produces similar pattern of artifacts, while our PSD behaves normally. Therefore, based on the above analysis, our work is fundamentally different from DreamDPO even not considering the negative embedding optimization strategy.

Comparison with other works applying prompt embedding optimization. There has been several previous also applying prompt embedding update but for completely different reasons. DiverseDream (Tran et al., 2025) employ HiPer token inversion to augment diversity and update the last several token of prompt embedding y to achieve similar effect of VSD Wang et al. (2023) in a memory-efficient way. LODS (Yang et al., 2025b) also optimize null (negative) embedding, but in order to reduce CFG value via "normalizing" SDS. On the other side, we incorporate CFG to update negative embedding in order to improve network towards higher rewards, which is novel to existing methods.

H Implementation Details

In this appendix, we describe the missing details of implementation in the main paper.

H.1 CONFIGURATION

For fair comparison, we do not propose any new regularizer and maintain same configuration for all methods in each experiment.

Single-stage distillation of MVDream. The configuration of this experiment follows baseline method MVDream (Shi et al., 2024), where only orientation loss proposed in (Poole et al., 2023) is used. The optimization takes 10000 steps with the weight of orientation loss linearly increasing from 10 to 1000 in the first 5000 steps. Training resolution is set to be 64×64 in first 5000 steps and 256×256 in the latter. For our negative embedding optimization strategy, learning rate is set as $1e^{-4}$ constantly. Due to memory limitation, Lambertian shading is not used in this setting. Besides, for experiments using Reward3D ("+Ours"), we combine negative embedding optimization with DreamReward instead of using preference score for fair comparison.

2-stage NeRF generation. The configuration of this experiment is adapted provided in baseline method CFD (Yan et al., 2025), which is performed based on previous experiment. The optimization takes 20000 steps in resolution 512×512 with the weight of "z-varience" loss proposed in (Zhu et al., 2024) being 10 and normal smooth loss being 1000. In the latter stage, negative embedding is optimized with a linearly decreasing learning rate from $1e^{-4}$ to 0 in the first 1000 step.

3-stage DMTet generation. This experiment is also performed after single-stage distillation of MVDream. Geometry optimization takes 15000 steps and texture takes another 20000 steps, with both resolution being 1024×1024. Since reward models are trained in RGB space, we only PSD in the texture space with learning rate being $1e^{-5}$ in the first 1000 step.

H.2 METRICS

Human preference reward models. For ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), Aesthetic scores (Schuhmann), and Multi-dimensional Preference Score (Zhang et al., 2024) we evaluate the average of the scores across 60 equally spaced views and their corresponding text prompts.

VQA models. We apply Qwen2.5-VL-7B (Bai et al., 2025) to calculate the text-3d alignment score. The question-answer pair we utilize is generated in Eval3d (Duggal et al., 2025). We evaluate across 12 renderings. For more details, please refer to (Duggal et al., 2025).

In order to compare geometry quality, we use the following metrics newly proposed from Eval3d:

Geometric Consistency. It measures the consistency between the surface normals analytically derived from the 3D representation and the normals predicted by a dense estimation model from 2D images. Our analytic normals are calculated using PyTorch auto-differentiation and estimated normals are calculated by converting from depth estimation from Depth Anything (Yang et al., 2024b), same as the original paper. The metric is calculated as follows:

$$\text{Geometric consistency} = \frac{1}{N_p} \sum_{p} \mathbb{1}[\arccos(\mathbf{n}_p^{\text{anal}} \cdot \mathbf{n}_p^{\text{pred}}) < \delta^{\text{norm}}] \tag{27}$$

where N_p is number of valid pixel p, $\mathbb{1}(\cdot)$ is indicator function, $\mathbf{n}_p^{\text{anal}}, \mathbf{n}_p^{\text{pred}}$ are analytical and estimated normals respectively. $\delta^{\text{norm}} = 23^\circ$ is a threshold.

Semantic Consistency. This metric measures the change of the underlying content and semantics. It projects 3D point x to 2D DINO feature $\{\mathcal{F}_i^{\text{DINO}}$ of each rendered image via projection π to retrieve its corresponding features. Then calculate

$$\text{Semantic consistency} = \frac{1}{N_{\text{vert}}} \sum_{x^{\text{vert}}} \mathbb{1}[\text{mean}(\text{Var}(\{\mathcal{F}_i^{\text{DINO}}(\pi_{v_i}(x^{\text{vert}}))\})) < \delta^{\text{DINO}}] \tag{2}$$

where x^{vert} is vertices of the 3D mesh. Following the original paper, δ^{DINO} is set to be the 70th percentile of average DINO variance.

Structural Consistency. Comparing with semantic consistency, structural consistency measures whether the overall structure of the generated asset is coherent and plausible. It uses diffusion-based novel view synthesis model Stable-Zero123 (Shi et al., 2023) to predict the image at unobserved viewpoint. Then it applies perceptual metric DreamSim (Fu et al., 2023) for similarity measurement. This metric can be formulated as

$$\text{Structural consistency} = \max_i \frac{1}{N} \sum_{j=1}^{N} \left(1 - f^{\text{DreamSim}}(\mathcal{I}_{i\to j}^{\text{pred}}, \mathcal{I}_j)\right) \tag{28}$$

where $\mathcal{I}_{i\to j}$ is image predicted from viewpoint i.

For interested readers, please refer to the original paper of Eval3d (Duggal et al., 2025) for more details. We choose these metrics because they are especially suitable to localize visual artifacts.

H.3 User Study

To validate that our results are truly preferred by human users, 23 participants are involved to make judgments over anonymous 30 rendered videos generated by our PSD against MVDream, DreamDPO and DreamReward. The instructions are:

- Appearance Quality: Evaluate the clarity and visual appeal of the asset as it appears from any particular viewpoint (ignoring, e.g., inconsistencies in appearance across different viewpoints). Your assessment should focus on the appearance of the foreground object and ignore the background of the video.
- 3D Structure Quality: Assess the detail and realism of the shape of the asset across the multiple viewpoints shown in the video.
- Text Alignment: Determine how accurately each video reflects the content of the text prompt. Consider whether the key elements of the prompt are represented.
- Overall Preference: State your overall preference between the two videos. This is your subjective appraisal of which video, in your view, stands out as better based on appearance quality, 3D structure quality, and text alignment, (i.e., overall quality).

A screen shot of our survey web is shown in Fig. 10.

H.4 Algorithm

In this section, we present the algorithm of PSD.

Figure 10: Screen shot of our survey web.

Algorithm 1 Preference Score Distillation

Input: A prompt y. Negative descriptors y_{neg}. Pretrained text-to-image diffusion model $\boldsymbol{\epsilon}_\phi$. Reward model r. Learning rate lr_1 and lr_2 for 3D representation θ and negative embeddings n, respectively. Annealing time-schedule $t(\tau)$. Classifier-free guidance weight γ.
initialize a 3D representation θ, a negative embeddings n with negative descriptors y_{neg}.
while not converged **do**
  Randomly sample a camera pose c, noise pair $\boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
  Render the 3D representation θ at pose $\boldsymbol{c}$ to get a 2D image $\boldsymbol{x}_c = g_\theta(\boldsymbol{c})$.
  Compute diffusion timestep $t(\tau)$.
  Add noise $\boldsymbol{\epsilon}^1$ and $\boldsymbol{\epsilon}^2$ to $\boldsymbol{x}_c$ and get $\boldsymbol{x}_t^1$ and $\boldsymbol{x}_t^2$
  Compute one-step predicted samples $(\hat{\boldsymbol{x}}_0^1, \hat{\boldsymbol{x}}_0^2)$ and ranking with reward model $(r\left(y, \hat{\boldsymbol{x}}_0^1\right), r\left(y, \hat{\boldsymbol{x}}_0^2\right))$ to get $(\boldsymbol{x}_t^w, \boldsymbol{x}_t^l)$ and their respective added noise $(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}')$
  $\delta_{gen} \leftarrow \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^w, t) - \boldsymbol{\epsilon}$, $\delta_{cls} \leftarrow \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^w, y, t) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^w, t)$, $\delta_{pref} \leftarrow \tilde{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t^w, y, t) - \tilde{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t^l, y, t)$
  $\beta_r \leftarrow \gamma \frac{||\delta_{cls}||_2}{||\delta_{pref}||_2} \cdot \sigma\left(r\left(y, \hat{\boldsymbol{x}}_0^l\right) - r\left(y, \hat{\boldsymbol{x}}_0^w\right)\right)$
  $\theta \leftarrow \theta - lr_1 \cdot \mathbb{E}_{\boldsymbol{c}}\left[\left(\delta_{gen} + \gamma\delta_{cls} + \beta_r\delta_{pref}\right) \frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}\right]$
  $\phi \leftarrow \phi - lr_2 \cdot \nabla_n \mathbb{E}_{\boldsymbol{c}}\left[r(y, \hat{x}_0)\right]$
end while
return

H.5 Details of Toy Experiments on Image Generation

In the main paper, we set up a toy experiments to directly illustrate the effect of our proposed preference score guidance. To simulate the configuration in DDIM, we follow (Lukoianov et al., 2024) to perform a 50-step update of the image parameterized by θ, as presented in Algorithm 2.

Algorithm 2 2D Image Generation using Score Distillation

Input: A prompt y. Pretrained text-to-image diffusion model $\boldsymbol{\epsilon}_\phi$. Learning rate lr. Number of optimization steps N. Time shift interval $[T_1, T_2]$
initialize random initial step $\boldsymbol{x}_T$ and parameterize with $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
while not converged **do**
  Compute timestep $t_i, i = 1, ..., N$ to update.
  for $i = 1$ to n **do**
    Randomly sample noise pair $\boldsymbol{\epsilon}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, time shift $\tau \sim \mathcal{U}(T_1, T_2)$.
    Add noise $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_{t_i}, y, t_i)$ and $\boldsymbol{\epsilon}^2$ to $\boldsymbol{x}_{t_i}$ and get $\boldsymbol{x}^1_{t_i+\tau}$ and $\boldsymbol{x}^2_{t_i+\tau}$
    Compute one-step predicted samples $(\hat{\boldsymbol{x}}^1_0, \hat{\boldsymbol{x}}^2_0)$ and ranking with reward model $(r\left(y, \hat{\boldsymbol{x}}^1_0\right), r\left(y, \hat{\boldsymbol{x}}^2_0\right))$ to get $(\boldsymbol{x}^w_{t_i+\tau}, \boldsymbol{x}^l_{t_i+\tau})$ their respective added noise $(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}')$
    $\delta_{gen} \leftarrow \boldsymbol{\epsilon}_\phi(\boldsymbol{x}^w_{t_i+\tau}, t_i + \tau) - \boldsymbol{\epsilon}$
    $\delta_{cls} \leftarrow \boldsymbol{\epsilon}_\phi(\boldsymbol{x}^w_{t_i+\tau}, y, t_i + \tau) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}^w_{t_i+\tau}, t_i + \tau)$
    $\delta_{pref} \leftarrow \tilde{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}^w_{t_i+\tau}, y, t_i + \tau) - \tilde{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}^l_{t_i+\tau}, y, t_i + \tau)$
    $\beta_r \leftarrow \gamma \frac{||\delta_{cls}||_2}{||\delta_{pref}||_2} \cdot \sigma\left(r\left(y, \hat{\boldsymbol{x}}^l_0\right) - r\left(y, \hat{\boldsymbol{x}}^w_0\right)\right)$
    $\theta \leftarrow \theta - lr \cdot \mathbb{E}_{\boldsymbol{c}}\left[\left(\delta_{gen} + \gamma\delta_{cls} + \beta_r\delta_{pref}\right) \frac{\partial g_\theta(\boldsymbol{c})}{\partial \theta}\right]$
  end for
end while
return

H.6 Test Prompts

For evaluations of single stage distillation of MVDream, we use 200-prompt croup set in Eval3D. For more complex 2-stage NeRF and 3-stage DMTet pipelines, we filter a harder subset consisting 40 prompts with lowest PickScore of MVDream, which is listed in Tab. 8.

I Limitations

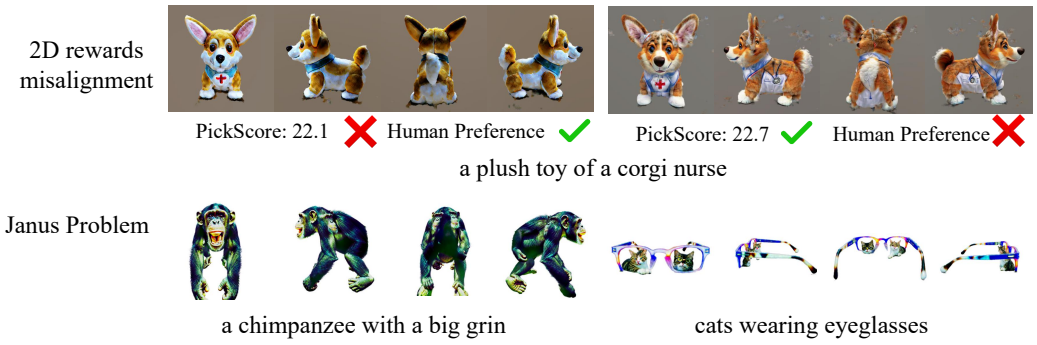


Figure 11: Visual examples of failure cases.

In this appendix, we present failure cases of our proposed PSD. While we successfully overcome the floating artifacts caused by pixel-level conflicts between reward gradients and diffusion dynamics, our method still fails to outperform at certain cases. Besides, Janus problem still exists when distilling Stable Diffusion 2.1. Developing better rewards may become a new solution to this problem.

J Supplementary Results and Comparisons

J.1 VISUALIZATION OF LEARNED NEGATIVE EMBEDDINGS ON DIFFUSION PRIORS

In this section, we conduct an experiment to visualize the difference of sampling diffusion priors with or without our learned negative embedding to illustrate the impact of learned negative embeddings on diffusion priors. Results are list in Tab. 6 and visualized in Fig. 12. Note that although the learned negative embeddings we use here is are from the single stage distillation of MVDream, they are able to transfer to SD 2.1 because they use the same text encoder. This observation is consistent with ReNeg.

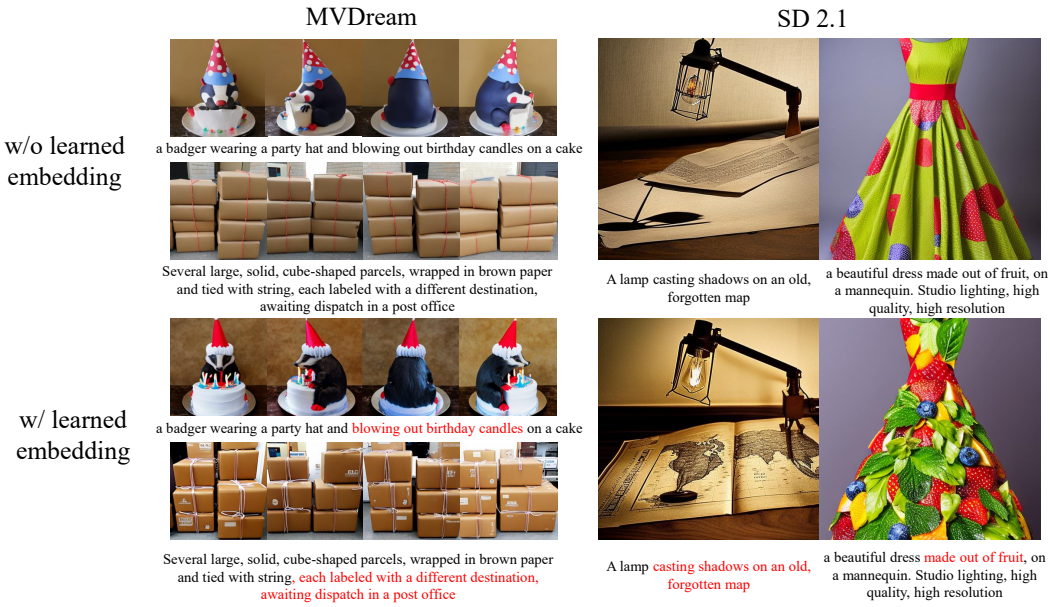


Figure 12: Visualization of learned negative embeddings on diffusion priors.

Experiments	MVDream		Stable Diffusion 2.1	
	Pick. ↑	I.R. ↑	Pick. ↑	I.R. ↑
w/o n	20.23	-0.31	21.07	0.30
w/ n	**20.31**	**-0.11**	**21.12**	**0.41**

Table 6: Impact of learned negative embeddings on diffusion priors. Metrics are evaluated across our 40-prompt subset on the average over 5 random seeds.

J.2 GEOMETRY COMPARISON

One of our major advantage is to avoid the artifacts introduced by directly guiding the 3D representation with gradients produced by reward models. In this section, we present the geometry comparison between DreamReward (Ye et al., 2025) and our proposed PSD using HPSv2.1 as reward model.

Through Tab. 7 and visual examples in Fig. 13 and 14, it's easy to conclude that our results have better geometry, which supports our motivation and claims.

Algorithm	Geometric Consistency ↑	Semantic Consistency ↑	Structural Consistency ↑
DreamReward	70.39	70.74	82.83
Ours	**80.97**	**75.83**	**84.72**

Table 7: Geometry assessment. Higher values are better (↑). Better results are in bold.

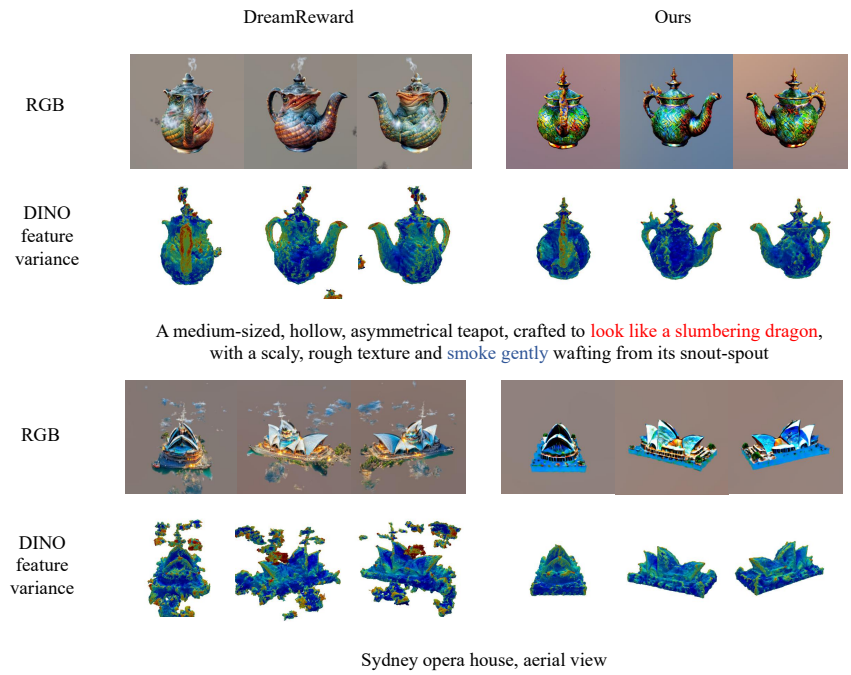


Figure 13: Visual examples of semantic consistency. DINO feature variance is clipped with threshold 0.20 (red).

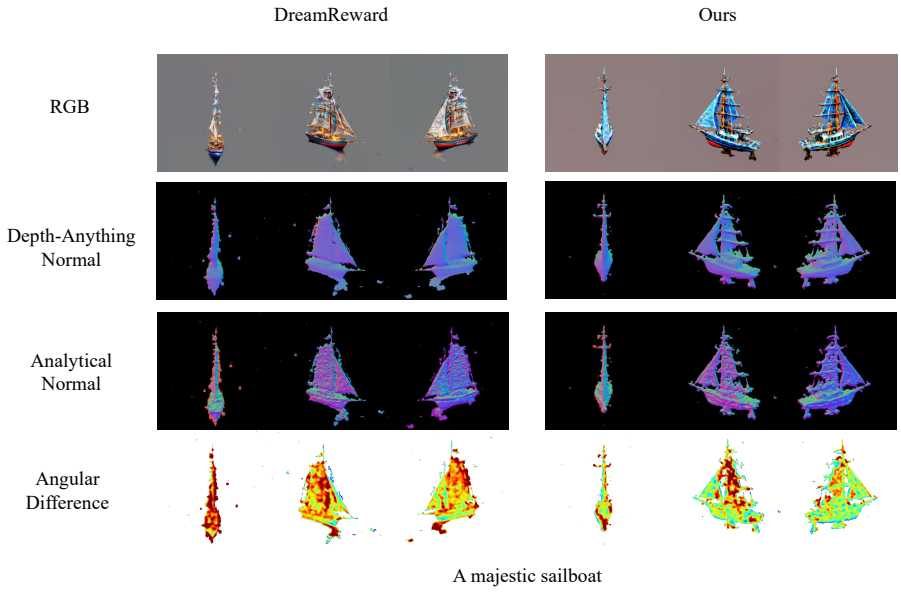


Figure 14: Visual example of geometric consistency. Angular difference is clipped with threshold 23° (red).

J.3 More Qualitative Comparisons

Column 1	Column 2
1. A DSLR photo of a plate of fried chicken and waffles with maple syrup on them	21. A tiger playing the violin
2. A beautiful dress made out of garbage bags, on a mannequin. Studio lighting, high quality, high resolution	22. A wide angle DSLR photo of a colorful rooster
3. A zoomed out DSLR photo of an astronaut chopping vegetables in a sunlit kitchen	23. A lone, ancient tree stands tall in the middle of a quiet field
4. A wide angle zoomed out DSLR photo of A red dragon dressed in a tuxedo and playing chess. The chess pieces are fashioned after robots	24. A squirrel dressed like Henry VIII king of England
5. A zoomed out DSLR photo of a pita bread full of hummus and falafel and vegetables	25. A zoomed out DSLR photo of A punk rock squirrel in a studded leather jacket shouting into a microphone while standing on a stump and holding a beer
6. Several large, solid, cube-shaped parcels, wrapped in brown paper and tied with string, each labeled with a different destination, awaiting dispatch in a post office	26. A dragon-cat hybrid
7. A large, multi-layered, symmetrical wedding cake, with smooth fondant, delicate piping, and lifelike sugar flowers in full bloom, displayed on a silver stand	27. A large, hollow, asymmetrically shaped amphitheater, with jagged stone seating, nestled in a natural landscape, a classical play being performed as the sun sets
8. Jellyfish with bioluminescent tentacles shaped like lightning bolts	28. A compact, cylindrical, vintage pepper mill, with a polished, ornate brass body, slightly worn from use, placed beside a porcelain plate on a checkered tablecloth
9. A Panther De Ville car	29. A zoomed out DSLR photo of a badger wearing a party hat and blowing out birthday candles on a cake
10. A beagle in a detective's outfit	30. A mug filled with steaming coffee
11. A wide angle DSLR photo of a squirrel in samurai armor wielding a katana	31. A zoomed out DSLR photo of a pair of floating chopsticks picking up noodles out of a bowl of ramen
12. A zoomed out DSLR photo of a kangaroo sitting on a bench playing the accordion	32. A wide angle zoomed out DSLR photo of a skiing penguin wearing a puffy jacket
13. A pair of hiking boots caked with mud at the doorstep of a cabin	33. A zoomed out DSLR photo of a fox working on a jigsaw puzzle
14. A zoomed out DSLR photo of cats wearing eyeglasses	34. A zoomed out DSLR photo of a beagle eating a donut
15. A wide angle zoomed out DSLR photo of zoomed out view of Tower Bridge made out of gingerbread and candy	35. A red panda
16. A beautiful dress made out of fruit, on a mannequin. Studio lighting, high quality, high resolution	36. A zoomed out DSLR photo of a kingfisher bird
17. A zoomed out DSLR photo of a bear playing electric bass	37. Clownfish peeking out from sea anemone tendrils
18. A lamp casting shadows on an old, forgotten map	38. A zoomed out DSLR photo of a rainforest bird mating ritual dance
19. An erupting volcano, aerial view	39. A chimpanzee with a big grin
20. A tiger karate master	40. A group of vibrant, chattering parrots perched together

Table 8: 40 prompts for evaluation in 2-stage NeRF and 3-stage DMTet.

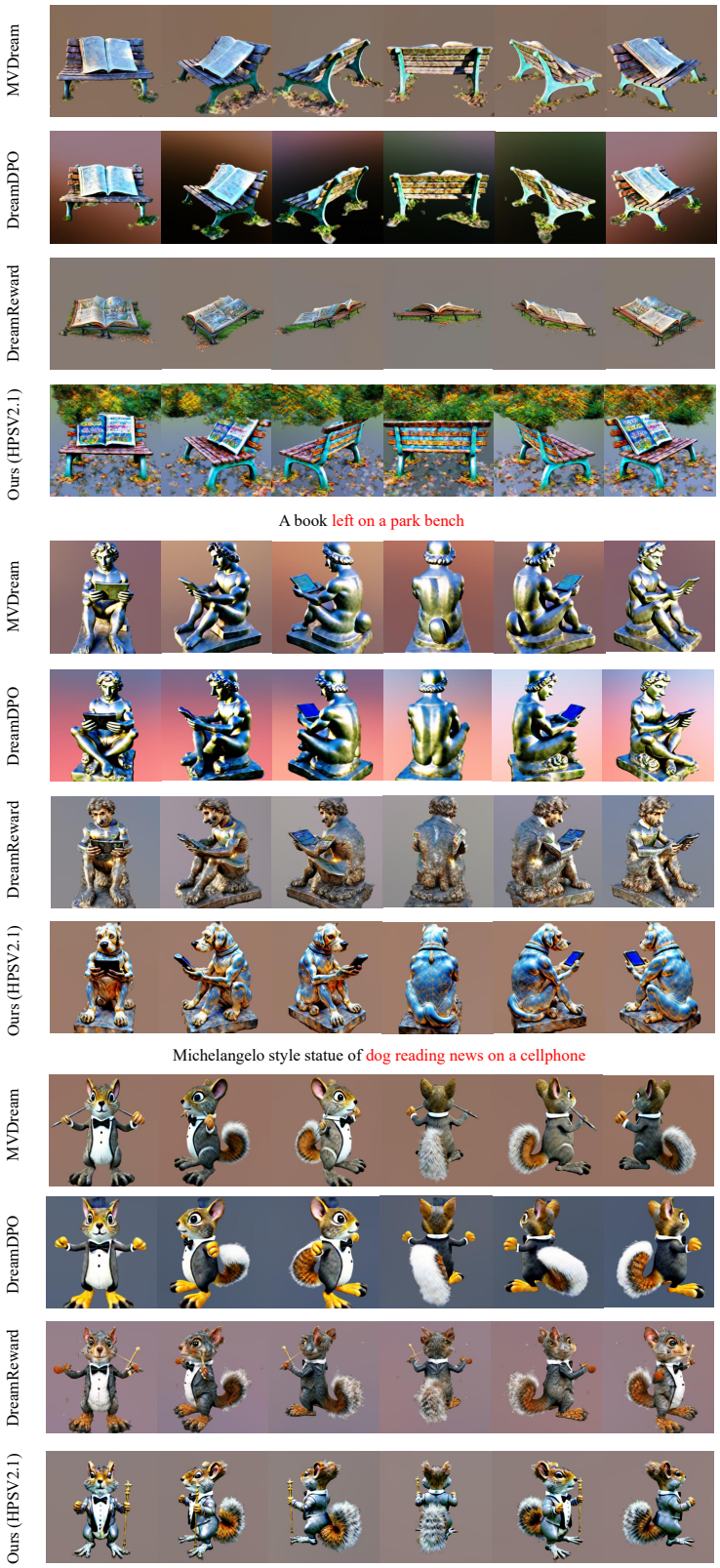
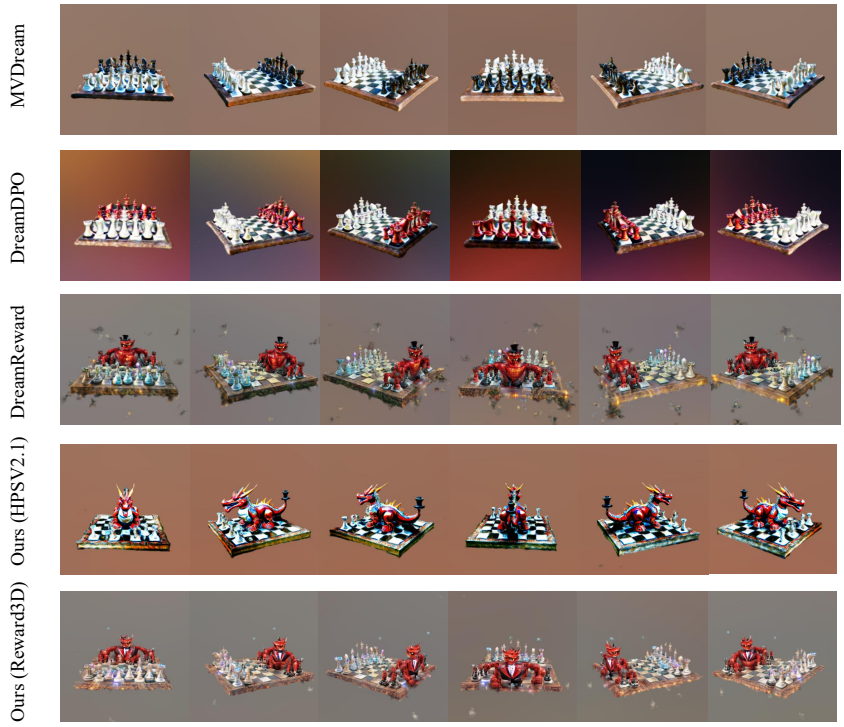
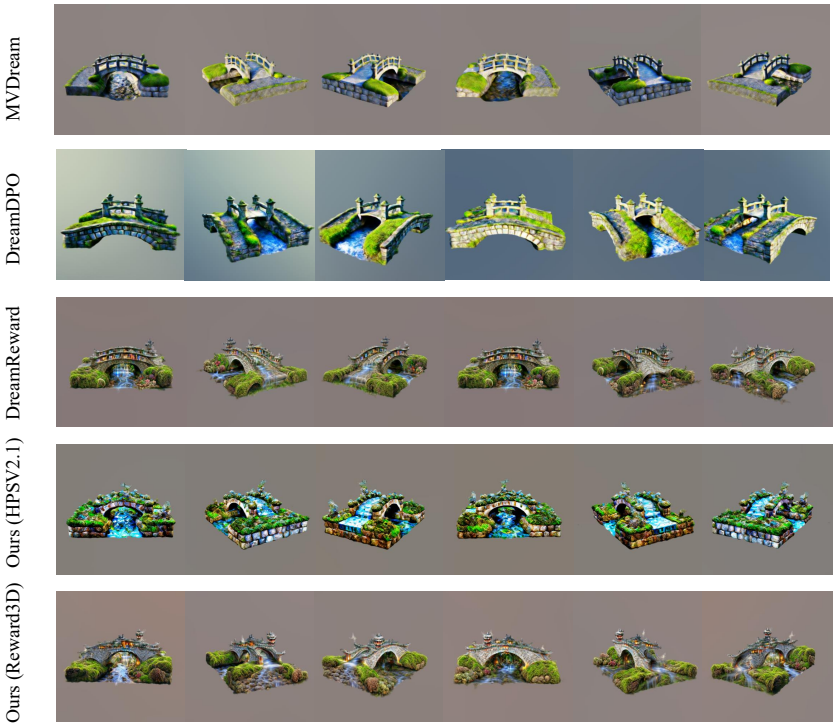


Figure 15: More results of single-stage distillation of MVDream



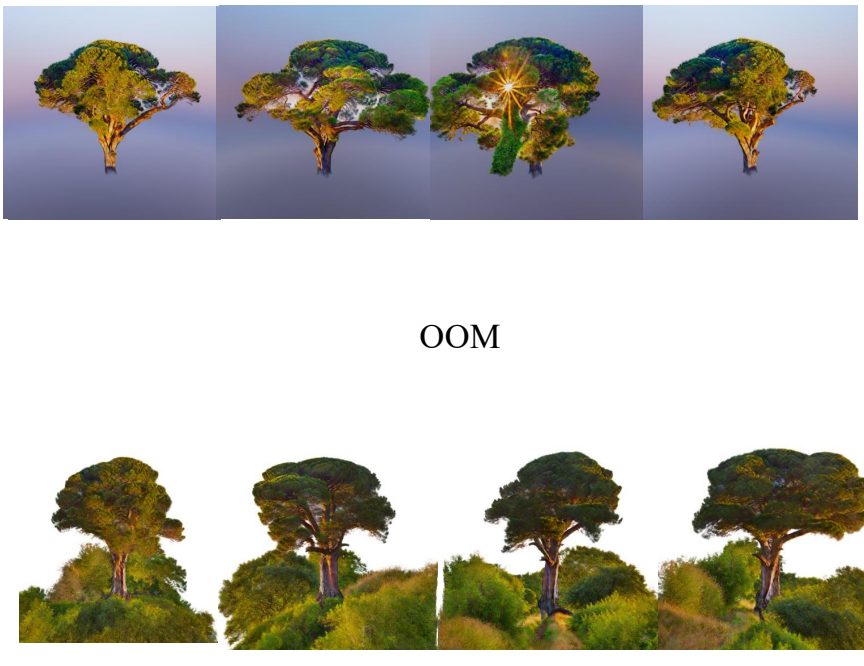
a wide angle zoomed out DSLR photo of A red dragon dressed in a tuxedo and playing chess. The chess pieces are fashioned after robots



A stone bridge arching over a babbling brook, encrusted with moss and echoing with stories

Figure 16: More results of single-stage distillation of MVDream

a wide angle zoomed out DSLR photo of A red dragon dressed in a tuxedo and playing chess. The chess pieces are fashioned after robots



A lone, ancient tree stands tall in the middle of a quiet field

Figure 17: More results of 2-stage NeRF generation.

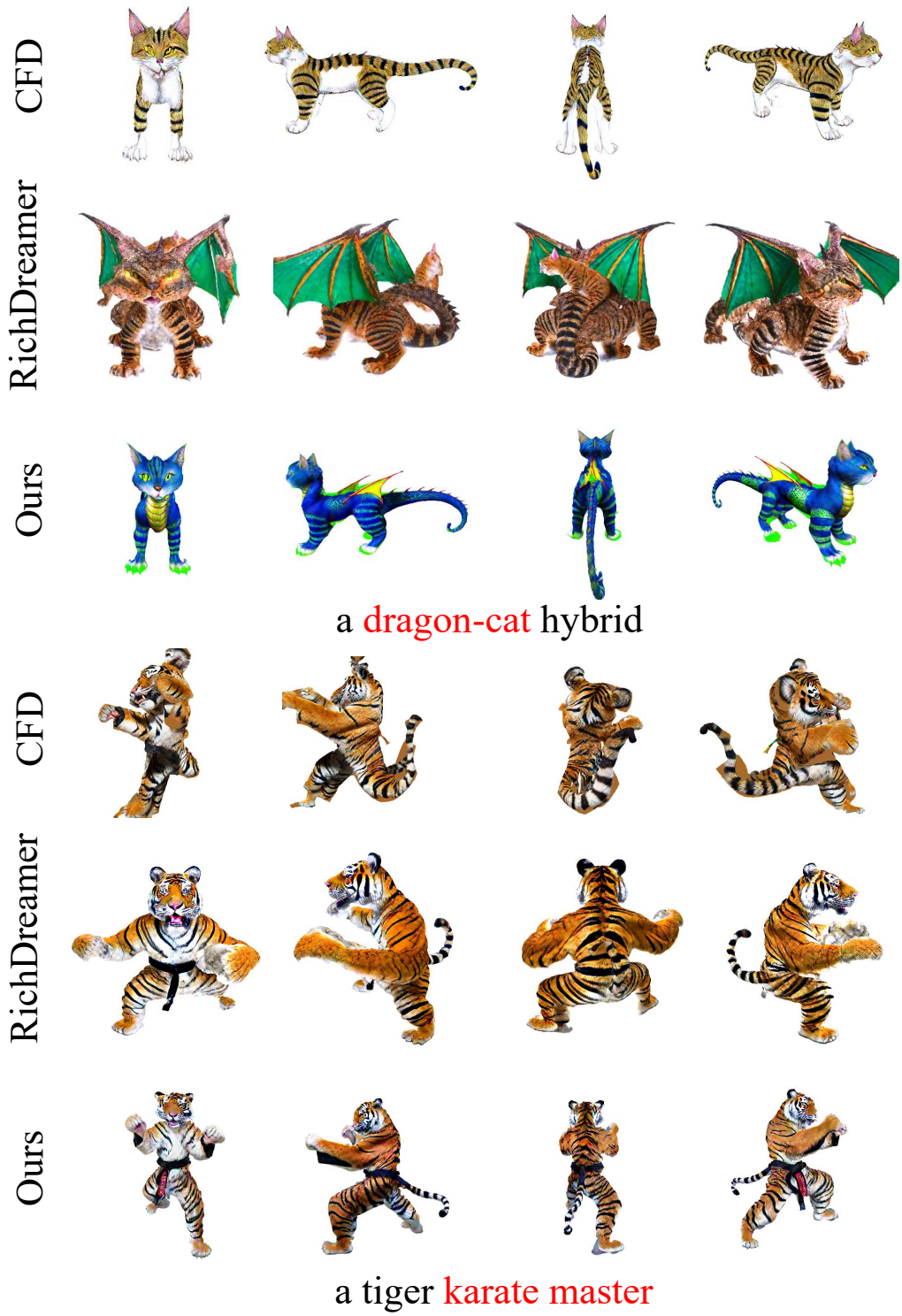


Figure 18: More results of 3-stage DMTet generation.

Figure 19: Extended comparison of 2-stage NeRF generation.