Retrieval-Augmented Score Distillation for Text-to-3D Generation

Junyoung Seo [*1] **Susung Hong** [*1] **Wooseok Jang** [*1] **Inès Hyeonsu Kim** [1]
Min-Seop Kwak [1] **Doyup Lee** [2] **Seungryong Kim** [1]

Abstract

Text-to-3D generation has achieved significant success by incorporating powerful 2D diffusion models, but insufficient 3D prior knowledge also leads to the inconsistency of 3D geometry. Recently, since large-scale multi-view datasets have been released, fine-tuning the diffusion model on the multi-view datasets becomes a mainstream to solve the 3D inconsistency problem. However, it has confronted with fundamental difficulties regarding the limited quality and diversity of 3D data, compared with 2D data. To sidestep these trade-offs, we explore a retrieval-augmented approach tailored for score distillation, dubbed ReDream. We postulate that both expressiveness of 2D diffusion models and geometric consistency of 3D assets can be fully leveraged by employing the semantically relevant assets directly within the optimization process. To this end, we introduce novel framework for retrieval-based quality enhancement in text-to-3D generation. We leverage the retrieved asset to incorporate its geometric prior in the variational objective and adapt the diffusion model's 2D prior toward view consistency, achieving drastic improvements in both geometry and fidelity of generated scenes. We conduct extensive experiments to demonstrate that ReDream exhibits superior quality with increased geometric consistency. Project page is available at https://ku-cvlab.github.io/ReDream/.

1. Introduction

Text-to-3D generation has emerged as an important application that enables non-experts to easily create 3D contents. The conventional approaches for text-to-3D train a generative model directly on 3D data from scratch (Wu et al., 2016; Chen et al., 2019; Zhou et al., 2021). However, their performance is limited due to the insufficient quality and diversity of 3D datasets compared with 2D datasets.

The seminal works for text-to-3D (Poole et al., 2022; Wang et al., 2023a) have introduced Score Distillation Sampling (SDS) to leverage the 2D diffusion models trained on large-scale images (Schuhmann et al., 2022). Given a text prompt, SDS-based frameworks (Chen et al., 2023; Seo et al., 2023; Lin et al., 2023; Wang et al., 2023b) directly optimize a Neural Radiance Field (NeRF) (Mildenhall et al., 2021) by distilling the scores of text-to-image (T2I) diffusion models through the rendered views of the optimizing NeRF. Exploiting the capability of T2I models to synthesize high-quality images (Rombach et al., 2022b; Saharia et al., 2022), SDS-based frameworks have generated high-fidelity 3D models even without 3D datasets. However, the generated scenes often suffer from artifacts and geometric inconsistencies due to the lack of knowledge on 3D geometry (Armandpour et al., 2023; Hong et al., 2023).

Recent approaches (Liu et al., 2023; Shi et al., 2023b) focus on fine-tuning 2D diffusion models on a large 3D dataset for novel view synthesis. Existing approaches (Liu et al., 2023; Shi et al., 2023b) modify and fine-tune a T2I model on Objaverse (Deitke et al., 2023b;a) to incorporate 3D awareness into its parameters for synthesizing novel multi-views. However, compared with 2D images, the insufficiency of high-quality 3D data has consequence of severely limiting and confining the style and fidelity of the generated novel views. For example, MVDream, trained on Objaverse, undergoes a cartoonish style shift (Shi et al., 2023a) , hindering the model from generating photorealistc 3D textures, and Zero123 shows drastically weakened performance when photorealistic images are given as input.

To address these issues, we propose a novel *retrieval-augmented* framework, ReDream, for text-to-3D generation to leverage 3D data information without full fine-tuning of 2D diffusion models. Our key motivation is that 3D assets, which are semantically aligned with a given text, become a minimal yet effective guidance of 3D geometries for SDS-based approaches. Then, ReDream can largely maintain the quality of the pre-trained 2D diffusion model, but also provide an effective geometric prior.

[*]Equal contribution [1]Korea Univeristy, Seoul, Korea [2]Runway, New York, USA. Correspondence to: Seungryong Kim <seungryong_kim@korea.ac.kr>, Doyup Lee <doyup@runwayml.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

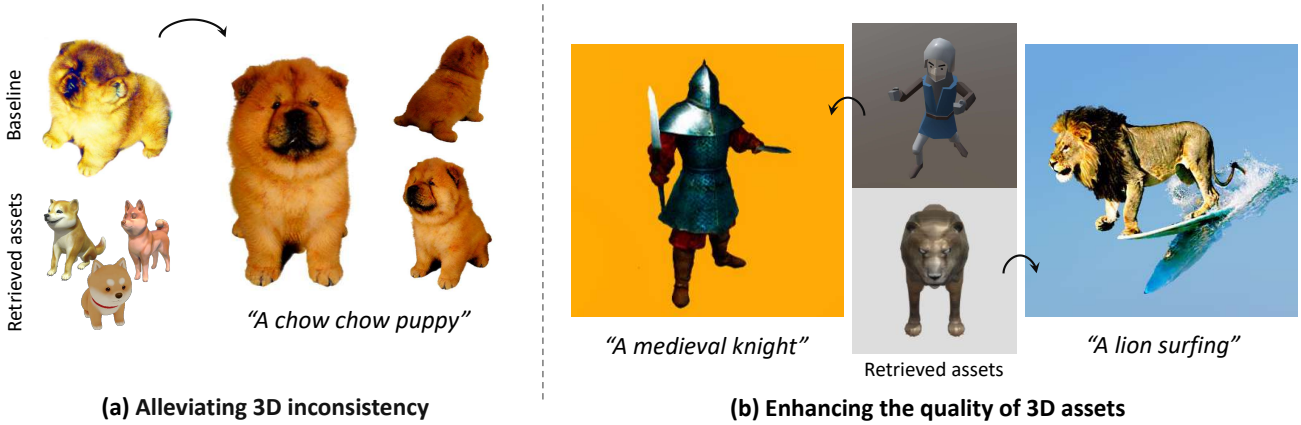

Figure 1. Our framework enables to high-quality generation of 3D contentsby leveraging retrieved assets from external databases, achieving significant enhancement of robust geometric consistency, as demonstrated in (a), and also enhancement of detail and fidelity, as shown in (b), without being bounded by the textural quality of the 3D assets.

Specifically, by interpreting each 3D scene represented by NeRF as sampled particles from a variational distribution, we show that retrieved assets can form a powerful initial variational distribution that incorporates geometric robustness and semantic relevance, grounding the generation process in these desirable qualities that text-to-3D generated scenes oftentimes lack. We also demonstrate that the retrieved assets can be leveraged for lightweight adaptation of 2D prior models, gearing the model towards more view-consistent 3D generation. These elegant and simple approaches effectively facilitates generation of high-quality 3D assets with added controllability and negligible training cost.

Our main contributions are summarized as follows:

- We present an intuitive yet feasible framework, **ReDream**, that effectively integrates the retrieval module with SDS-based frameworks for text-to-3D generation.
- Our framework can exploit both the geometric information of 3D assets and the capability of T2I models to synthesize high-fidelity images without the need of full training of the model parameters.
- We introduce a lightweight approach that significantly reduces viewpoint bias in 2D prior models, which has been plaguing text-to-3D generation.
- We conduct extensive experiments to demonstrate that our proposed methods consistently improve the generation quality and analyze how the retrieval-augmentation affects the 3D generation process.

2. Related work

Generative novel view synthesis. Generative models have been employed to learn a multi-view geometry to synthesize novel views of a 3D scene (Wiles et al., 2020; Rombach et al., 2021). When given a single reference view, (Chan et al., 2023) estimate its 3D volume to condition a model for generating novel views. This process involves incorporating a cross-view attention in a diffusion model to align the correspondences between novel and reference views (Zhou & Tulsiani, 2023; Watson et al., 2023). Zero123 (Liu et al., 2023) adapts the Stable Diffusion model (Rombach et al., 2022a) to fine-tune its entire parameters on Objaverse datasets (Deitke et al., 2023a;b) for generating novel views of 3D objects in the open domain. However, these previous approaches face limitations in fidelity due to the scarcity of high-quality 3D data, which often requires the laborious and specialized work of experts. Additionally, MVDream (Shi et al., 2023b) concurrently proposes a multi-view diffusion model by fine-tuning the Stable Diffusion model.

Text-to-3D generation with score distillation. DreamFusion (Poole et al., 2022) introduced a novel method known as Score Distillation Sampling (SDS) for generating 3D content without relying on 3D data. This method involves optimizing a 3D representation, such as Neural Radiance Fields (NeRF) (Mildenhall et al., 2021), by distilling the prior knowledge of diffusion models to synthesize high-fidelity images. Concurrently, related studies (Wang et al., 2023a) have derived similar loss functions using SDS. Following this, subsequent research (Metzer et al., 2023; Tsalicoglou et al., 2023) has consistently improved text-to-3D generation based on the SDS framework. Other developments in this area include Magic3D (Lin et al., 2023), which utilizes DMTet (Shen et al., 2021) within a coarse-to-fine pipeline to enhance the quality of 3D representation. Fantasia3D (Chen et al., 2023) introduces a two-stage framework to separate geometry and texture in 3D content creation. ProlificDreamer (Wang et al., 2023b) employs a particle-optimization framework for Variational Score Distillation (VSD), significantly improving the fidelity of generated textures. However, a common challenge faced by these

methods, which do not use 3D training data, is the issue of 3D inconsistency. This often results in the unrealistic geometry of the generated contents, highlighting a key area for further improvement in the field of 3D content generation.

Retrieval-augmented generative models. Retrieval-augmented approaches utilize an external database to adapt a generative model for diverse tasks without fine-tuning whole parameters on large-scale data. For example, RETRO (Borgeaud et al., 2022) adapts a large language model for exploiting the external databases and achieves high performances without increasing its parameters. For the task of image synthesis, retrieval-augmented methods have been applied to GANs (Tseng et al., 2020; Casanova et al., 2021) and diffusion models (Blattmann et al., 2022; Sheynin et al., 2022; Chen et al., 2022b), while adapting the models for synthesizing unseen styles such as artistic images (Rombach et al., 2022c). Since retrieval-augmentation is effective when the data scale is insufficient to train the model parameters, (Zhang et al., 2023) and (He et al., 2023) integrate a motion-retrieval module with diffusion models to synthesize motion sequences and videos, respectively.

3. Background: Score distillation sampling

Score distillation sampling (SDS) (Poole et al., 2022) has been proposed as a method to leverage text-to-image diffusion models (Saharia et al., 2022; Rombach et al., 2022b) originally trained on text-paired image datasets for generation of 3D objects. Specifically, 3D scene θ, a differentiable representation such as NeRF (Mildenhall et al., 2021), is optimized so that its renderings at various camera poses follow probability density $p_\phi(x|c)$ which is the 2D distribution conditioned on input text tokens c. The score of this distribution is approximated by the diffusion model ϵ_ϕ, and the practical update rule is derived as follows:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = -\mathbb{E}_{t,\epsilon,\psi}\Big[w(t)\big(\epsilon_\phi(x_t|c,t)-\epsilon\big)\frac{\partial g(\theta,\psi)}{\partial\theta}\Big], \quad (1)$$

where $w(t)$ and x_t are a weighting function and a perturbed image of x with a noise level t, and ϵ is a corresponding Gaussian noise. $g(\cdot)$ and ψ are the differentiable renderer and the camera pose, respectively.

Variational score distillation (VSD) (Wang et al., 2023b) further generalizes this sampling technique by interpreting it as the variational problem of fiding the distribution γ which is represented by the particles θ. Specifically, the variational distribution $q^\gamma\big(x_t|c, x=g(\theta,\psi)\big)$ represents an implicit distribution of rendered images. The VSD framework establishes this implicit relationship by the denoising score matching process leveraging low-rank adaptation (LoRA) (Ryu, 2023), resulting in following approximation: $\nabla_{x_t} q^\gamma\big(x_t|c, x=g(\theta,\psi)\big) \approx -\epsilon_{\phi,\zeta}(x_t|c,t,\psi)/\sigma_t$, where ζ represents a set of parameters for LoRA of the diffusion model. As a consequence, the resulting updating direction corresponds to:

$$\nabla_\theta \mathcal{L}_{\text{VSD}} = -\mathbb{E}_{t,\epsilon,\psi}\Big[w(t)\big(\epsilon_\phi(x_t|c,t) - \epsilon_{\phi,\zeta}(x_t|c,t,\psi)\big)\frac{\partial g(\theta,\psi)}{\partial\theta}\Big]. \quad (2)$$

For the detailed explanation on the background, please refer to Appendix C.

4. Retrieval-augmented score distillation

4.1. Motivation

While previous SDS-based methods have allowed for the flexible, high-quality generation of 3D objects even with complicated prompts, they still tend to produce implausible 3D geometry. Recent studies have mitigated this issue by training multi/novel-view generative models (Shi et al., 2023b; Liu et al., 2023) on existing 3D datasets. Although these methods present viable solutions, the quality and size of existing 3D dataset is inferior in comparison to 2D data, hampering and confining the fidelity and diversity of the models directly trained on these data. This effect can be universally noted in methods that have taken the training-based approach, such as MVDream and Zero123, in which the textures of generated scenes and novel views largely retain clay-like cartoonish styles similar to that of low-quality 3D assets.

To address such issues, we explore a novel retrieval-augmented approach tailored for SDS-based frameworks, which enables the generation of high-quality 3D objects. The fundamental insight is that retrieved 3D assets, which are semantically similar to the specified text, can serve as concise references for abstract 3D appearances and geometries.

4.2. Formulation

We begin by adopting a particle-based variational inference (ParVI) framework (Chen et al., 2018; Liua & Zhub, 2022; Liu & Wang, 2016; Dong et al., 2022), following the convention of (Wang et al., 2023b). Within this framework, a variational distribution γ is composed of particles $\{\theta^{(i)}\}_{i=1}^K$. Each particle is optimized using the gradient of VSD distilled from 2D diffusion models, as described in Eq. 2: $v_{\text{2D}}^{(i)} := \nabla_{\theta^{(i)}}\mathcal{L}_{\text{VSD}}$. Here, $v_{\text{2D}}^{(i)}$ denotes the per-particle velocity derived from the 2D prior of the diffusion model.

Our primary goal is to enable particles to absorb meaningful information from retrieved assets $\{\theta_{\text{ret}}^{(n)}\}_{n=1}^N$, which are conditioned on a text prompt c from the 3D database $\mathcal{D}$, using the retrieval module $\xi_N(c,\mathcal{D})$. To achieve this, we propose a novel method to impose the velocity of each particle with the retrieved assets, as detailed in Sec. 4.3. This approach

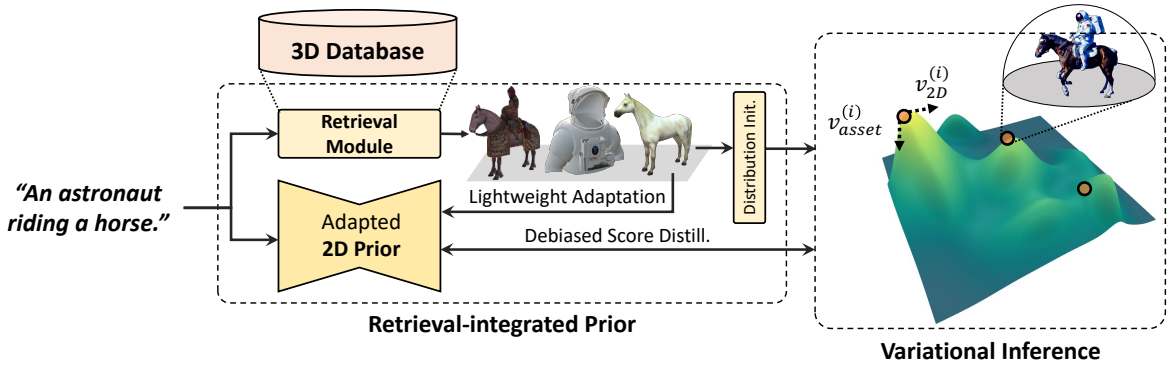

Figure 2. **Overview.** Given a prompt c, we retrieve the nearest neighboring assets from the 3D database. With these assets, we perform initialization of an variational distribution for incorporation of robust 3D geometric prior, as well as conducting lightweight adaptation of 2D prior model for equalize probability density across viewpoints.

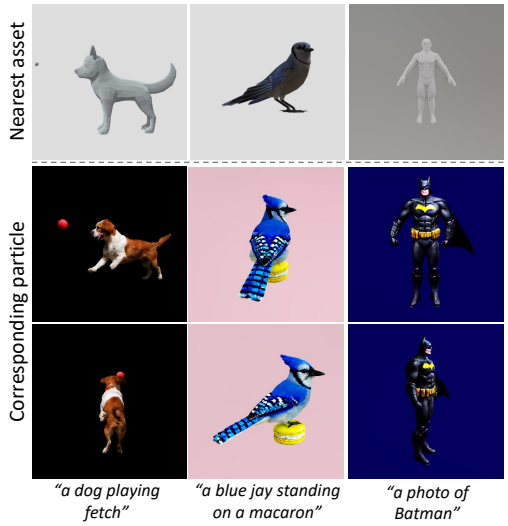

Figure 3. **Generated results and corresponding nearest asset**. The first row shows the first nearest neighbor from the retrieved assets, with the renderings of corresponding particles from the given texts displayed below.

facilitates the subsequent optimization of the distribution γ using the gradient derived from a lightweight-adapted 2D diffusion model, described in Sec. 4.4. Our overall framework is illustrated in Fig. 2.

4.3. Initialized distribution as a geometric prior

Recall that the variational distribution γ is optimized by updating the particles $\theta \sim \gamma(\theta|c)$ as in Eq. 2, and the parametrization of the score of q^γ is additionally tuned with the particles. In this perspective, initializing the particles can be interpreted as providing a *guide* for the variational distribution q^γ.

We find that retrieved neighbors can effectively act as a *guide* for the variational distribution γ, since the ideally selected nearest assets exhibit robust geometry as well as sharing semantic similarity with the optimizing particles. This approach effectively enables our model to achieve geometric robustness while overcoming the weaknesses of methods involving direct training on 3D data such as low-quality data and computation cost described above. To this end, we derive and leverage an auxiliary objective from our retrieval-augmented objective that makes $\gamma(\theta|c)$ and the empirical distribution of retrieved assets similar. The full derivation is shown in Appendix C. Practically, we impose an additional velocity on each particle to coarsely initialize them during the warm-up phase as follows:

$$v_{\text{asset}}^{(i)} := \nabla_{\theta^{(n)}} \frac{\mathbb{1}(s \le \tau)}{\sigma^2} \mathbb{E}_\psi \left[\left\| g(\theta^{(i)}, \psi) - g(\theta_{\text{ret}}^{(a_i)}, \psi) \right\|_2^2 \right], \tag{3}$$

where s, τ, and σ denote the index of iterations, the threshold for the warm-up phase, and the scaling factor, respectively. a_n is a mapping function relating the i-th particle to its corresponding retrieved asset. Note that the particle initialization is reflected in the distribution γ within the framework of VSD, which has the following additional objective:

$$\min_\zeta \sum_{i=1}^{N} \mathbb{E}_{t,\epsilon,\psi} \left\| \epsilon_{\zeta,\phi}\left(x_t, t, c, \psi\right) - \epsilon \right\|_2^2, \tag{4}$$

where $x = g(\theta^{(i)}, \psi)$. Recalling that this denoising score matching objective leads to the following relationship between q^γ and $\epsilon_{\phi,\zeta}$: $\nabla_{x_t} q^\gamma\big(x_t|c, x = g(\theta, \psi)\big) \approx -\epsilon_{\zeta,\phi}(x_t|c, t, \psi)/\sigma_t$, the process in Eq. 4 can be viewed as aligning the distribution γ with the empirical distribution $p_\xi(\theta|c)$.

The effectiveness of our initialization approach is clearly observable in Fig. 3, where we see that the robust geometry of the nearest assets is efficiently leveraged to ensure the robustness and consistency of corresponding particle's 3D structure. We observe that the particle's geometry and texture is not strictly confined to the initialization, allowing for freedom to make sufficient adjustments that enable the

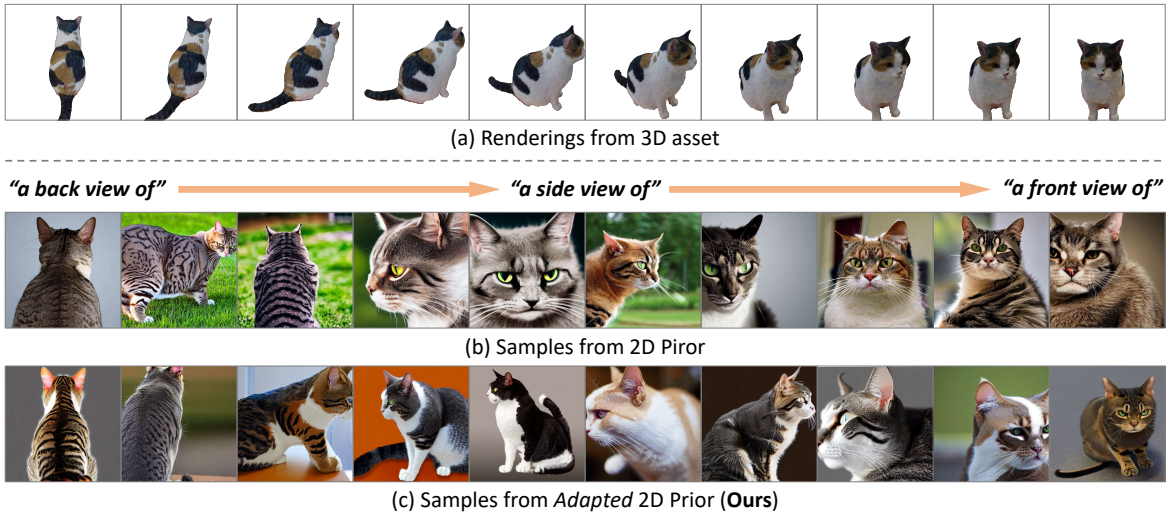

Figure 4. **Lightweight adaptation of 2D diffusion models.** We compare the effectiveness of the adaptation with given rendering from a 3D asset in (a). We linearly interpolate a text embedding from "*a back view of an angry cat*" to "*a front view of an angry cat*" through "*side view*". (b) 2D samples from the prior model. (c) 2D samples from the adapted prior model with learned view prefixes. Compared with (b). The samples from adapted 2D prior in (c) reflect a variety of viewpoints, **not biased towards a single viewpoint**.

particle to faithfully follow the text prompt.

4.4. Lightweight adaptation of 2D prior

During the score distillation process, viewpoint-related bias from diffusion models hinders consistent generation, but at the other extreme, fully fine-tuning diffusion models on 3D assets causes the model to lose its expressiveness. Here, we address the dilemma with *lightweight* adaptation, which mostly maintains the original manifold of pre-trained diffusion models while reducing view-related biases.

A major issue that significantly hampers score distillation based methods is the fact 2D prior models are biased toward certain viewpoints, as shown in Fig. 4(b), leading to text-guided predictions that are misaligned with the initial scenes. The issue of view bias of 2D prior models has been known as one of the cause of Janus problem and has been addressed in other works (Armandpour et al., 2023; Hong et al., 2023). For instance, Perp-Neg (Armandpour et al., 2023) and Debiased-SDS (Hong et al., 2023) addressed this by mainly adopting negative prompt, or removing contradictory words with view prefix, respectively.

Contrary to these works, our method fortunately begins from an advantageous position, as we have access to dense renderings of 3D assets that are semantically close collected with the retrieval module. It allows us to address this issue simply yet effectively. To this end, we introduce a lightweight strategy that adapts 2D prior models by utilizing retrieved 3D assets in test time. This helps balance the probability densities across all viewpoints without a significant drop in the quality of the original 2D prior models, in despite of its simplicity.

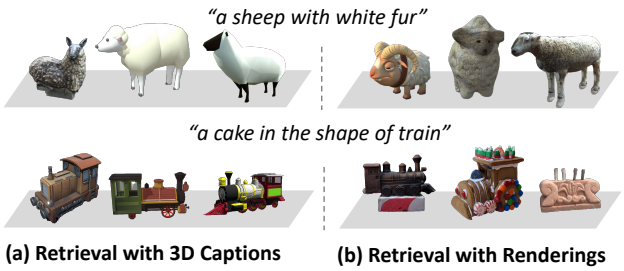

Figure 5. **3D Dataset retrieval.** (a) and (b) show retrieved top-K nearest neighbors on CLIP-text embedding space and CLIP-image embedding space, respectively.

Specifically, we denote $c_{\text{ret}}^{(n)}$ as a ground-truth text caption tokens corresponding to the n-th retrieved asset, and $\{e_\psi\}$ as tokens of view prefixes such as "*front view*". To obtain a adapted 2D prior $\epsilon_{\omega,\phi}$, we densely render the retrieved assets under a uniform camera distribution, and optimize a low-rank adapter (Ryu, 2023; Hu et al., 2021) with the rendered images:

$$\min_{\omega} \sum_{n=1}^{N} \mathbb{E}_{t,\epsilon,\psi} \left\| \epsilon_{\omega,\phi} \left(x_t, t, \text{cat}(e_\psi, c_{\text{ret}}^{(n)}) \right) - \epsilon \right\|_2^2, \quad (5)$$

where $x = g(\theta_{\text{ret}}^{(n)}, \psi)$, and ω is a set of parameters of learnable layers inserted to the diffusion U-net. $\text{cat}(\cdot)$ refers to concatenation function. At the same time, we can additionally optimize the tokens of view prefixes $\{e_\psi\}$ as well as ω using Eq. 5. We empirically find it eliminate the model's viewpoint bias more effectively in the few-shot setting.

After the adaptation, the 2D prior p_ϕ used in v_{2D} is replaced with the adapted prior $p_{\omega,\phi}$ along with the learned view prefixes $\{e_\psi\}$. Our strategy demonstrates encouraging effectiveness as it shows the chronic issue of viewpoint bias in 2D prior models can be efficiently addressed thanks to

Figure 6. **Improved 3D consistency from baseline (Wang et al., 2023b).** We validate the effectiveness of our approach by comparing the baseline. Given challenging prompts that are easy to cause geometric breakdowns, our results show **enhanced performance in terms of 3D**. See Supplementary Materials for videos of these results.

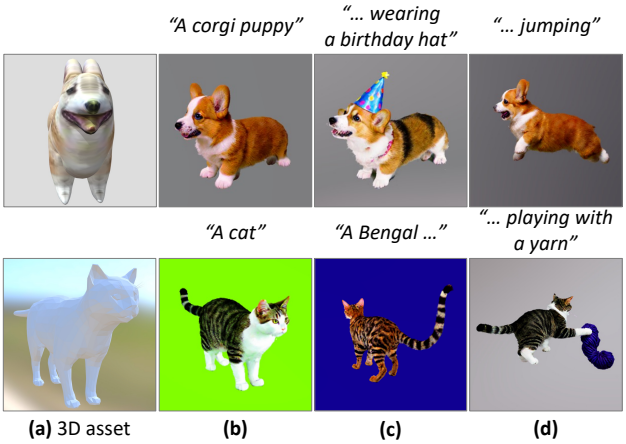

Figure 7. **Variations of text prompts with fixed 3D asset.** Given the retrieved 3D asset shown at leftmost column, each column represents separate optimization results given different text conditions. Note that the scene under optimization is not strictly constrained to the asset, but retains strong capability to generate a 3D scene relevant to the given text prompt and the assets.

the nearest neighbors without any complex technique. As shown in Fig. 4, we can see samples from the adapted 2D prior is capable of generating viewpoints that more closely reflect each given view conditions without severely sacrificing its generation capability.

4.5. Retrieval of 3D assets

We utilize 3D assets from Objaverse 1.0 (Deitke et al., 2023b) dataset and corresponding captions with the help of Cap3D (Luo et al., 2023). We use ScaNN (Guo et al., 2020) to retrieve N nearest neighbors based on CLIP embeddings (Radford et al., 2021) of the captions and the rendered images. The query embedding can be acquired from the prompt c. Specifically, we utilize both image and text embeddings by performing Top-K operation with image embeddings after retrieving $N'(N' > N)$ objects with text embeddings, followed by alignment of orientations as a pre-processing step, described in detail at Appendix B.

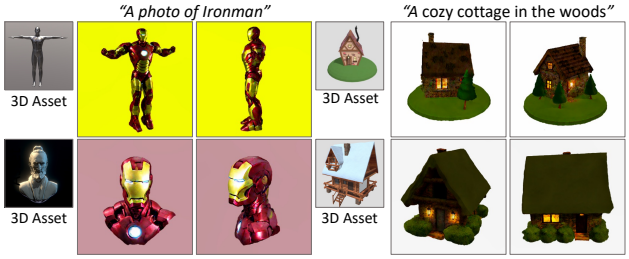

Figure 8. **Variations of 3D assets with fixed text prompt.** We fix the prompts and vary the assets that correspond to each particle. This shows how our method is effected by the retrieved assets.

As we construct a list mapping UIDs of 3D assets to the corresponding CLIP embeddings, end-users can only download the retrieved 3D assets during inference, or download the whole 3D data in advance. The total time spent by the retrieval is under 3 seconds. As shown in Fig. 3 and Fig. 5, in situations where completely matching 3D assets are not retrieved given challenging prompts, we found it still shows sufficient performance to serve as references for generation.

5. Analysis

In this section, we provide extensive analyses on the properties of our approach, including qualitative and quantitative evaluations. The implementation details are described in Appendix A and B.

Does it handle corner cases of the baseline? One of our goals is to alleviate 3D inconsistency, which frequently occurs when given challenging prompts. For example, when testing both the baseline (Wang et al., 2023b) and our method on generating creatures with a face, we observe that our approach can generate more plausible outputs, as

Figure 9. **Comparison with other works.** We compare our framework with novel/multi-view model based frameworks: Zero123 (Liu et al., 2023), MVDream (Shi et al., 2023b), Magic123 (Qian et al., 2023), and image generative model based frameworks: DreamFusion (Poole et al., 2022), ProlificDreamer (Wang et al., 2023b). **When zoomed in, the 3D inconsistency and resulting artifacts are most noticeable**. We provide accompanying **video results** in Supplementary Materials. Also, more qualitative results can be found in Appendix E.1.

illustrated in Fig. 6. Consistent with our claims in Sec. 4, we corroborate that our method alleviates such issues by utilizing a retrieval-integrated prior.

Influence from retrieved assets. In our retrieval-based approach, we address two critical questions: whether there's an over-reliance on retrieved assets leading to overly constrained results, or whether these assets fail to remarkably influence the outcome. To explore the impact of retrieved assets, we set the number of assets for retrieval N to 1, and gradually change the corresponding text prompt to be distanced from the asset. Fig. 7 clearly shows our observation; our approach flexibly operates depending on the similarity between the text prompt and the retrieved asset. In (b) of Fig. 7, where the text prompt aligns best with the asset, we observe minimal geometric changes and textural variations, whereas in (c) and (d), sufficient adjustments are made where necessary. Additionally, Fig. 8 shows the results by changing the assets with the fixed prompts. It also supports our observation, showing the flexibility of our approach.

Qualitative evaluation. We compare our methods with state-of-the-art text-to-3D (Wang et al., 2023b; Poole et al., 2022) and image-to-3D methods (Liu et al., 2023; Shi et al., 2023b; Qian et al., 2023). In the case of image-to-3D methods, we carefully selected appropriate images generated by the text-to-image model (Rombach et al., 2022b). These generated images are delineated in Fig. 12. Comparative results are shown in Fig. 9. In contrast to preceding text-to-image prior based methods, our framework shows enhanced geometric consistency. On the other hand, while methods

Figure 10. **Intermediate renderings in optimization.** We visualize the intermediate renderings of the particle which corresponds to top-1 retrieved asset. Geometric influence of the nearest assets is significant when the 3D representation is coarse, and fine details are generated through the adapted 2D prior. Details are clearest when zoomed in.

employing novel/multi-view models yield plausible geometry, they often suffer from degraded texture, such as overly smoothed surfaces, which detracts from realism, whereas ours generates high-quality textures. Additioanlly, we visualize the optimization process by showing the intermediate renderings of the particle with the corresponding 3D asset in Fig. 10.

Quantitative evaluation. Currently, there is no established metric for evaluating the open-domain text-to-3D field, as text-to-3D is an inherently subjective task and encompasses various aspects that are challenging to quantify. Nevertheless, we align with the practices of quantitative evaluation (Poole et al., 2022; Li et al., 2023a; Yu et al., 2023) in text-to-3D works by utilizing CLIP-based metrics for our quantitative assessments. Specifically, we measure the average CLIP score between text and 3D renderings using variants of the CLIP model, OpenCLIP ViT-L/14 trained on DataComp-1B (Ilharco et al., 2021) and CLIP ViT-L/14 (Radford et al., 2021). The evaluation is done with 50 prompts, each rendered with 120 viewpoints of the corresponding 3D outputs. We note that the CLIP model for retrieval is not used for the evaluation. For view consistency, some works (Li et al., 2023b) manually check their success rate, and (Hong et al., 2023) proposes A-LPIPS, an average LPIPS (Zhang et al., 2018) between adjacent images of generated 3D scenes to measure artifacts caused by view inconsistency. We adopt A-LPIPS as an alternative metric to quantify view consistency and report it alongside the CLIP score in Tab. 1, showing ReDream exhibits superior performance in terms of text-3D alignment and view consistency.

User study. We conduct a user study with 92 participants; the result is shown in Tab. 2. Each participant is asked seven randomly selected questions. Specifically, we inquire about their preference between our method and the baseline, taking into account geometry and textural fidelity. Approximately 75% of the participants express a preference for the results by our method over the baseline. More details are described in Appendix G.

Methods	CLIP-Score ↑		A-LPIPS ↓	
	CLIP L/14	OpenCLIP L/14	VGG	Alex
DreamFusion	0.242	0.185	0.075	0.076
MVDream	0.263	0.217	0.062	0.072
ProlificDreamer	0.218	0.204	0.227	0.135
ReDream	**0.274**	**0.227**	**0.041**	**0.054**

Table 1. **Quantitative evaluation.** We compare our approach with recent text-to-3D works (Poole et al., 2022; Shi et al., 2023b; Wang et al., 2023b). CLIP-score indicates the alignment between text and 3D, while A-LPIPS represents the degree of artifacts due to 3D inconsistency (Hong et al., 2023).

Methods	Preference
Baseline (Wang et al., 2023b)	24.7%
Ours	**75.3%**

Table 2. **User study.** We report the percentage of user preference from 92 participants.

Ablation on each component. We conduct an ablation study on each component of our pipeline, as depicted in Fig. 17 of Appendix E.4. We observe that initializing the variational distribution is crucial for the overall geometry, and lightweight adaptation effectively reduces artifacts such as eyes on the back.

2D experiments on lightweight adaptation. We conduct a 2D experiment to detail the process of lightweight adaptation, which is depicted in Fig. 14. We also report our analysis of how a 3D asset influences the 2D prior model in lightweight adaptation in Fig. 15. Specifically, we progressively change the prompts to describe other objects with different textures while keeping the used asset constant. The results suggest that the adaptation primarily concentrates on general aspects, such as viewpoint, instead of focusing on particular details like texture. The details are described in Appendix E.2 and E.3, respectively.

6. Conclusion

We present a novel retrieval-based framework for text-to-3D generation in which retrieved assets are used as efficient guidance for enhanced fidelity and geometric consistency of generated 3D scenes. We propose simple, elegant methods to leverage the retrieved assets for aforementioned purpose, which use the retrieved asset as initializing point of 3D scene's variational distribution, and also use it for adaptation of 2D diffusion model toward increased faithfulness to given view prompts. Our approach does not compromise the capabilities of 2D diffusion models, not requiring extensive fine-tuning. Through extensive experiments and analysis, both quantitative and qualitative, we demonstrate that our model successfully achieves the goal of quality improvement and geometric robustness in text-to-3D generation.

Impact Statement

This paper presents in the field of AIGC (AI-generated Content) aiming for research advancements. While there may be potential social impacts as a consequence, there is nothing in particular to be highlighted. The framework presented in this paper utilizes data retrieved from an external database; therefore, users employing this framework must verify the copyright of the database they use.

Acknowledgments

This research was supported by the MSIT, Korea (IITP-2024-2020-0-01819, RS-2021-II212068), Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, 25%) and National Research Foundation of Korea (RS-2024-00346597). This work was also supported by a grant-in-aid of HANWHA SYSTEMS.

References

Armandpour, M., Zheng, H., Sadeghian, A., Sadeghian, A., and Zhou, M. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023.

Blattmann, A., Rombach, R., Oktay, K., Müller, J., and Ommer, B. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35: 15309–15324, 2022.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Casanova, A., Careil, M., Verbeek, J., Drozdzal, M., and Romero Soriano, A. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

Chan, E. R., Nagano, K., Chan, M. A., Bergman, A. W., Park, J. J., Levy, A., Aittala, M., De Mello, S., Karras, T., and Wetzstein, G. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023.

Chen, A., Xu, Z., Geiger, A., Yu, J., and Su, H. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022a.

Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. A unified particle-optimization framework for scalable bayesian sampling. *arXiv preprint arXiv:1805.11659*, 2018.

Chen, K., Choy, C. B., Savva, M., Chang, A. X., Funkhouser, T., and Savarese, S. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 100–116. Springer, 2019.

Chen, R., Chen, Y., Jiao, N., and Jia, K. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023.

Chen, W., Hu, H., Saharia, C., and Cohen, W. W. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022b.

Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S. Y., et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023a.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023b.

Dong, H., Wang, X., Lin, Y., and Zhang, T. Particle-based variational inference with preconditioned functional gradient flow. *arXiv preprint arXiv:2211.13954*, 2022.

Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., and Kumar, S. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/1908.10396.

Guo, Y.-C., Liu, Y.-T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.-H., Zou, Z.-X., Wang, C., Cao, Y.-P., and Zhang, S.-H. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023.

He, Y., Xia, M., Chen, H., Cun, X., Gong, Y., Xing, J., Zhang, Y., Wang, X., Weng, C., Shan, Y., et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023.

Hertz, A., Aberman, K., and Cohen-Or, D. Delta denoising score. *arXiv preprint arXiv:2304.07090*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hong, S., Ahn, D., and Kim, S. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., and Bi, S. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023a.

Li, W., Chen, R., Chen, X., and Tan, P. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023b.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023.

Liua, C. and Zhub, J. Geometry in sampling methods: A review on manifold mcmc and particle-based variational inference methods. *Advancements in Bayesian Methods and Implementations*, 47:239, 2022.

Luo, T., Rockwell, C., Lee, H., and Johnson, J. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023.

Metzer, G., Richardson, E., Patashnik, O., Giryes, R., and Cohen-Or, D. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12663–12673, 2023.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Müller, T., Evans, A., Schied, C., and Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15, 2022.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

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

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

Rombach, R., Esser, P., and Ommer, B. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14356–14366, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022a.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022b.

Rombach, R., Blattmann, A., and Ommer, B. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022c.

Ryu, S. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Seo, J., Jang, W., Kwak, M.-S., Ko, J., Kim, H., Kim, J., Kim, J.-H., Lee, J., and Kim, S. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023.

Shen, T., Gao, J., Yin, K., Liu, M.-Y., and Fidler, S. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.

Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., and Taigman, Y. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., and Su, H. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023a.

Shi, Y., Wang, P., Ye, J., Long, M., Li, K., and Yang, X. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023b.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Tancik, M., Mildenhall, B., Wang, T., Schmidt, D., Srinivasan, P. P., Barron, J. T., and Ng, R. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2846–2855, 2021.

Tsalicoglou, C., Manhardt, F., Tonioni, A., Niemeyer, M., and Tombari, F. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*, 2023.

Tseng, H.-Y., Lee, H.-Y., Jiang, L., Yang, M.-H., and Yang, W. Retrievegan: Image synthesis via differentiable patch retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 242–257. Springer, 2020.

Wang, H., Du, X., Li, J., Yeh, R. A., and Shakhnarovich, G. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12619–12629, 2023a.

Wang, Z., Ren, T., Zhu, J., and Zhang, B. Function space particle optimization for bayesian neural networks. *arXiv preprint arXiv:1902.09754*, 2019.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023b.

Watson, D., Chan, W., Brualla, R. M., Ho, J., Tagliasacchi, A., and Norouzi, M. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=HtoA0oT30jC.

Wiles, O., Gkioxari, G., Szeliski, R., and Johnson, J. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7467–7477, 2020.

Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., and Wang, X. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023.

Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., and Kanazawa, A. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2(3): 6, 2021.

Yu, X., Guo, Y.-C., Li, Y., Liang, D., Zhang, S.-H., and Qi, X. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023.

Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., and Liu, Z. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Zhou, L., Du, Y., and Wu, J. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5826–5835, 2021.

Zhou, Z. and Tulsiani, S. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12588–12597, 2023.

A. Experimental Setup

We build our method upon ProlificDreamer (Wang et al., 2023b), and follow (Guo et al., 2023) for details of the implementation. Our experiments were conducted on an NVIDIA RTX A6000 GPU, with a total of 20,000 iterations of optimization for generation. For all our experiments, Instant-NGP (Müller et al., 2022) is used for our NeRF backbone and Stable Diffusion v2 (Rombach et al., 2022b) as the 2D prior. For our method, we retrieve 3 assets and render our retrieved data with 100 uniformly sampled camera poses. We compare our framework with various methods (Poole et al., 2022; Wang et al., 2023b; Liu et al., 2023; Shi et al., 2023b; Qian et al., 2023). For (Wang et al., 2023b; Liu et al., 2023; Shi et al., 2023b; Qian et al., 2023), we utilize author-provided implementations. For (Poole et al., 2022), we used Stable Diffusion as a 2D prior model on Threestudio (Guo et al., 2023), as Imagen (Saharia et al., 2022) used in their implementation is not publicly available.

B. Additional Implementation Details

3D retrieval procedure. 3D datasets such as Objavarse (Deitke et al., 2023b) are too large for users to download; therefore, we construct a list that includes the UIDs of the 3D contents along with the corresponding CLIP embeddings of the renderings and captions (Luo et al., 2023), and we proceed with retrieval employing the ScaNN (Guo et al., 2020) algorithm. In this case, the total time spent by the retrieval is under 3 seconds, which is negligible compared to the time taken for the entire generation process. During inference, we download and load only the essential 3D assets using the UIDs acquired via the retrieval process.

While the Objaverse dataset offers a variety of 3D assets, their orientations are generally not aligned. We observe that this is not necessarily an issue, since score distillation works with misaligned orientations as well. Nevertheless, before employing our nearest neighbors, we find it beneficial to align their frontal views. Specifically, we can categorize 3D objects into (1) those where the front is distinguishable, such as objects with a clear frontal aspect, and (2) those where the front is not distinguishable, such as radially symmetric objects. In the case of the latter, the importance of the view prefix is not high. This is because it is not only difficult to semantically predict the front views but also because the necessity to find their orientations is not significant, allowing it to be disregarded. For the former case, it is relatively important to identify the semantic fronts to assign appropriate view prefixes to their renderings.

For this purpose, we compute the CLIP similarity score between the prompts with view prefixes "*front view*", "*side view*", "*back view*" and the rendered images with different camera poses. Subsequently, we rotate the 3D assets according to the camera poses that exhibit the relatively highest CLIP similarity score. Despite its simplicity, this method effectively aligns our retrieved assets. Note that objects with semantically indistinct front and back differences (such as an ice cream cone) exceptionally demonstrated lower accuracy levels. Nevertheless, due to the nature of such objects, the necessity for orientation alignment is less critical, and we found that this has a minimal impact on the performance of the final results. For the performance of the alignment, refer to Sec. E.5.

Additional regularization. Concerning the degree to which particles may deviate from or overlook their initial state, this divergence becomes apparent when the bias of the 2D prior towards a particular text prompt continues to steer away from v_{asset}. To alleviate this problem, we adopt a variant of the delta denoising score (Hertz et al., 2023), initially employed in image editing, in 3D cases. Specifically, $v_{\text{2D}}(\theta = \theta_0)$ represents the predicted velocity (gradient) of the 2D prior at the point θ_0. Ideally, the combination of a retrieved asset and text should result in minimal gradient or velocity, leading us to identify $v_{\text{2D}}(\theta = \theta_{\text{ret}})$ as a noisy component. To reduce the artifacts, we adjust the original v_{2D} by subtracting from it: $\tilde{v}_{\text{2D}} := v_{\text{2D}} - v_{\text{2D}}(\theta = \theta_{\text{ret}})$. We opt for updates using $\tilde{v}_{\text{2D}}$ in place of v_{2D} for every three iterations. We found that the adjustment strength can be effectively controlled by modulating the frequency of these updates and adjusting the weight.

C. Conceptual Analysis of Our Approaches

Preliminary. Here, we formulate text-to-3D generation with score distillation (Poole et al., 2022), which leverages a diffusion model (Saharia et al., 2022; Rombach et al., 2022b) as a prior to optimize a 3D representation for a given text. We extend the framework of Variational Score Distillation (VSD) (Wang et al., 2023b), which generalizes the original Score Distillation Sampling (SDS) (Poole et al., 2022).

VSD aims to optimize the distribution of 3D representations given a text prompt, while SDS (Poole et al., 2022) aims to optimize an instance of 3D representation for text-to-3D generation. We also define $q^{\gamma}(x|c, \psi)$ as an implicit distribution of the rendered image $x := g(\theta, \psi)$ where $\theta \sim \gamma(\theta|c)$. Then, VSD minimizes the variational objective, $D_{\text{KL}}\big(q^{\gamma}(x|c)||p_{\phi}(x|c)\big)$

to find an optimal γ^*, where $q^\gamma(x|c)$ is marginalized distribution w.r.t. camera viewpoints $p(\psi)$ and $p_\phi(x|c)$ is empirical likelihood of x estimated by a diffusion model ϕ. Since the diffusion model learns noisy distribution $p_\phi(x_t|c,t)$ according to diffusion process (Ho et al., 2020; Song et al., 2020b), the variational objective can be decomposed as follows:

$$\gamma^* := \arg\min_\gamma \mathbb{E}_t\Big[(\sigma_t/\alpha_t)w(t)D_{\mathrm{KL}}(q_t^\gamma(x_t|c)||p_\phi(x_t|c,t))\Big], \tag{6}$$

where $q_t^\gamma(x_t|c)$ is a noisy distribution at noise level t following the diffusion process.

VSD employs the particle-based variational inference (ParVI) (Chen et al., 2018; Liua & Zhub, 2022; Wang et al., 2019; Dong et al., 2022) to minimize Eq. 6. The minimization process proceeds via a Wasserstein gradient flow (Chen et al., 2018). Specifically, N particles $\{\theta^{(i)}\}_{i=1}^N$ are first sampled from initial $\gamma(\theta|c)$, and then updated with the following ODE:

$$v_{\mathrm{2D}} := \frac{d\theta_\eta}{d\eta} = -\mathbb{E}_{t,\epsilon,\psi}\Big[w(t)\big(-\sigma_t\nabla_{x_t}\log p_\phi(x_t|c,t) - (-\sigma_t\nabla_{x_t}\log q_t^{\gamma_\eta}(x_t|\psi,c))\frac{\partial g(\theta_\eta,\psi)}{\partial\theta_\eta}\big)\Big], \tag{7}$$

where η denotes ODE time such that $\eta \geq 0$, and the distribution γ_η converges to an optimal distribution γ^* as $\eta \to \infty$ and θ_η is sampled from γ_η. Note that the first term is a score of noisy real image, approximated by a predicted score of the diffusion model $\epsilon_\phi(x_t,c,t)$. The second term can be regarded as a score of noisy rendered images. They parameterize the second term to a score-predicting U-shaped network. Practically, they train the U-Net network from the pretrained diffusion model with low-rank adaptation (LoRA), $\epsilon_{(\phi,\zeta)}(x_t,t,c,\psi)$, where ζ is a set of parameters of trainable residual layers for LoRA. VSD allows to generate realistic textures of 3D object given a text, but we remark that these method are still vulnerable to generating unrealistic geometry.

Regarding total velocity comprised of v_{asset} and v_{2D}. We first show a total velocity in warm-up phase can be roughly interpreted as minimizing the distance between the variational distribution γ and our retrieval-integrated prior we present in the followings. Let $\xi_N(c,\mathcal{D})$ be a non-parametric sampling strategy to obtain the N nearest neighbors using the retrieval algorithm conditioned on text prompt c in the 3D dataset $\mathcal{D}$. Our goal is to integrate the rich view-dependent information from the retrieved assets with that of 2D prior models, and derive the particle-based optimization process for the variational distribution $\gamma(\theta|c)$. We assume the probability density of 3D content θ by 2D prior is proportional to the expected densities of its multiview images w.r.t. camera viewpoints, following (Wang et al., 2023a):

$$p_\phi(\theta|c) \propto \mathbb{E}_\psi\big[p_\phi^{\mathrm{2D}}(x|c,x=g(\theta,\psi)\big]. \tag{8}$$

Technically, this expectation is set as the geometric expectation (see the last paragraph of this section for details). Subsequently, let us consider a following energy functional for integrating the retrieved assets:

$$\mathcal{E}[\gamma] := D_{\mathrm{KL}}\big(\gamma(\theta|c)||p_{\phi,\xi}(\theta|c)\big), \tag{9}$$

where we present $p_{\phi,\xi}(\theta|c)$ as a retrieval-integrated prior. Based on the intuition that a 3D asset selectively filters a distribution, we simply multiply and normalize the two distributions:

$$p_{\phi,\xi}(\theta|c) := \frac{1}{Z'}p_\phi(\theta|c)p_\xi(\theta|c), \tag{10}$$

where we denote $p_\xi(\theta|c)$ as a 3D likelihood from the retrieved assets, and Z' denotes the normalizing constant. Fig. 11 depicts the intuition behind this; the distribution p_ξ derived from the retrieved nearest neighbor serves as an implicit filter for plausible geometry.

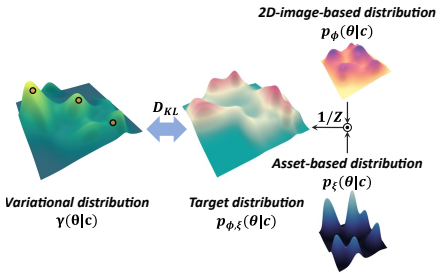

Figure 11. **Conceptual figure of the variational objective.** Geometrically plausible areas by retrieved nearest neighbors have higher density in the target distribution.

Specifically, we derive the distribution $p_\xi(\theta|c)$ from an empirical distribution defined over the top-N nearest neighbors $\{\theta_{\mathrm{ret}}^{(n)}\}_{n=1}^N$ utilizing the sampling strategy $\xi_N(c,\mathcal{D})$, then applying non-parametric kernel K for density estimation. Intuitively, the likelihood $p_\xi(\theta|c)$ depicts how close the particle is to the retrieved assets.

Using the definition of KL divergence, this is further expanded:

$$\mathcal{E}[\gamma] = \mathbb{E}_\psi[D_{\mathrm{KL}}\big(q^\gamma(x|c)||p_\phi^{\mathrm{2D}}(x|c)\big)] + H\big(\gamma(\theta|c);p_\xi(\theta|c)\big) - C \tag{11}$$

$$= \mathbb{E}_\psi[D_{\mathrm{KL}}\big(q^\gamma(x|c)||p_\phi^{\mathrm{2D}}(x|c)\big)] - \mathbb{E}_{\gamma(\theta|c)}\left[\log\sum_n K(\theta-\theta_{\mathrm{ret}}^{(n)})\right] - C', \tag{12}$$

where $x = g(\theta, \psi)$, and H is the joint entropy. C and C' are constants to be unnecessary.

The minimization process then proceeds via a Wasserstein gradient flow (Chen et al., 2018). Given $\mathcal{E}[\gamma_\eta]$ at an optimization step η, the velocity of particles, $v_\eta := \frac{d\theta_\eta}{d\eta} = \nabla_\theta \frac{\delta \mathcal{E}[\gamma_\eta]}{\delta \gamma_\eta}$, is obtained by calculating the functional derivative $\frac{\delta \mathcal{E}[\gamma_\eta]}{\delta \gamma_\eta}$ as follows:

$$v_\eta = \nabla_\theta \frac{\delta \mathcal{E}[\gamma_\eta]}{\delta \gamma_\eta} = v_{\text{2D}} - \nabla_\theta \log \sum_n K(\theta - \theta_{\text{ret}}^{(n)}) \tag{13}$$

$$= v_{\text{2D}} + v_{\text{asset}}. \tag{14}$$

where v_{asset} is the velocity derived from retrieval, and v_{2D} is derived as in Eq. 7. $K(\cdot)$ can be any kernel function. This suggests that the total velocity of particles, derived from the variational objective with the augmented distribution, actually consists of the two components.

As one usual choice would be Gaussian kernel, we start by choosing K as Gaussian kernel with a variance σ^2. However, in practice, strictly computing the derived v_{asset} with Gaussian kernel for all assets remains inefficient, given that it is defined in a high-dimensional space. To address this inefficiency, we turn to our observation that the direction of the velocity of each particle is largely determined by its random initialization as it is drawn towards the nearest mode, which suggest a feasible alternative. Motivated by this observation, instead of computing all terms, we use an efficient surrogate method to compute v_{asset} for each particle as follows:

$$v_{\text{asset}}^{(i)} = \sum_n \frac{\pi_n^{(i)}}{\sigma^2}(\theta^{(i)} - \theta_{\text{ret}}^{(n)}) = \frac{1}{\sigma^2} \sum_n \pi_n^{(i)}(\theta^{(i)} - \theta_{\text{ret}}^{(n)}), \tag{15}$$

where $\theta^{(i)}$ is i-th particle from the variational distribution $\gamma(\theta|c)$ and we assign to them one-hot vectors π whose non-zero indices correspond to a closest random asset when initialized. Intuitively, this property of a particle to follow a specific mode is determined at the time of its creation.

For generality, the particle $\theta^{(i)}$ and 3D asset $\theta_{\text{ret}}^{(n)}$ have not been assumed to have specific representations (*e.g.*, NeRF (Mildenhall et al., 2021), DMTet (Shen et al., 2021), or mesh), and could be different representations. However, some representations can be only partially observed through the differentiable rendering function g. Accordingly, in Eq. 15, the shift term is given in the form of a gradient with respect to the objective (Tancik et al., 2021):

$$(\theta^{(i)} - \theta_{\text{ret}}^{(n)}) \simeq \nabla_{\theta^{(i)}} \mathbb{E}_\psi \Big[\big\| g(\theta^{(i)}, \psi) - g(\theta_{\text{ret}}^{(n)}, \psi) \big\|_2^2 \Big]. \tag{16}$$

Consequently, the velocity of i-th particle towards the retrieved asset in warm-up phase becomes to:

$$v_{\text{asset}}^{(i)} \simeq \nabla_{\theta^{(n)}} \frac{1}{\sigma^2} \sum_n \pi_n^{(i)} \mathbb{E}_\psi \Big[\big\| g(\theta^{(i)}, \psi) - g(\theta_{\text{ret}}^{(a_i)}, \psi) \big\|_2^2 \Big]. \tag{17}$$

Lightweight adaptation as a parametric approach. In the Lightweight adaptation introduced in Sec. 4.4 of the main paper, the adaptor of the 2D prior can be interpreted as a parametric model moderately reflecting $p_\xi(\theta|c)$. Specifically, given the original relationship of the pretrained diffusion models (Song et al., 2020b; Ho et al., 2020),

$$\nabla_{x_t} \big[p_\phi(x_t|c, \psi) \big] \approx -\frac{\epsilon_\phi(x_t, t, c)}{\sigma_t}, \tag{18}$$

and given the (variational) objective of the lightweight adaptation,

$$\sum_{n=1}^{N} \mathbb{E}_{t,\epsilon,\psi} \left\| \epsilon_{\omega,\phi} \left(x_t, t, \text{cat}(e_\psi, c_{\text{ret}}^{(n)}) \right) - \epsilon \right\|_2^2, \tag{19}$$

where $\epsilon_{\omega,\phi}$ represents LoRA, whose initialization is exactly the same function as ϵ_ω. Since the model $\epsilon_{\omega,\phi}$ implicitly matches the empirical distribution p_ξ of retrieved assets, and because we early-stopped the training to maintain quality, p_ξ is moderately reflected. In other words, the score (inclination) of p_ξ is moderately learned by the adapted diffusion model.

Consequently, the resulting velocity from the adapted 2D model can be derived in the same way as in (Wang et al., 2023b):

$$\begin{aligned}\hat{v}_{\text{2D}}^{(n)} &:= \nabla_{\theta^{(n)}} \frac{\delta}{\delta \gamma} \mathbb{E}_{t,\epsilon,\psi}\Big[D_{\text{KL}}\big(q^{\gamma}(x_t|c)||p_{\xi,\omega}(x_t|c,\psi)\big)\Big] \\ &= -\mathbb{E}_{t,\epsilon,\psi}\Big[w(t)\big(\epsilon_{\omega,\phi}(x_t,t,\text{cat}(e_\psi,c)) - \epsilon_{\phi,\zeta}(x_t|c,t,\psi)\big)\frac{\partial g(\theta,\psi)}{\partial \theta}\Big],\end{aligned} \tag{20}$$

With this in mind, as our the previous approach, the velocity attributable to 3D assets can be separated:

$$\hat{v}_{\text{3D}}^{(n)} := \hat{v}_{\text{2D}}^{(n)} - v_{\text{2D}}^{(n)}. \tag{21}$$

Assumption on the density function of 3D content. Several works (Wang et al., 2023a; Hong et al., 2023) have clarified the assumptions on the density function of 3D content, which is an important part in lifting the 2D generative models to do 3D generation. Specifically, SJC (Wang et al., 2023a) proposes to assume it to be proportional to an arithmetic expectation of likelihoods over camera points, *i.e.*, $p_\phi(\theta|c) \propto \mathbb{E}_\psi[p_\phi^{\text{2D}}(x|c, x = g(\theta,\psi))]$, and D-SDS (Hong et al., 2023) finds it more beneficial to define it as a product of likelihoods over a set of camera points. In this paper, we instead use the geometric expectation. Actually, all three premises do not affect the solution of the minimization or maximization problem of the logarithm. Besides, in terms of KL divergence, setting the target distribution to the geometric mean has the following benign property:

$$D_{\text{KL}}(q||\kappa\mathbb{G}_\psi[p_\phi^{\text{2D}}(x|c,x=g(\theta,\psi))]) = \mathbb{E}_\psi[D_{\text{KL}}(q||p_\phi^{\text{2D}}(x|c,x=g(\theta,\psi)))] - \log\kappa, \tag{22}$$

where κ is a constant.

D. Additional Experimental Details

Reference images for image-to-3D works (Liu et al., 2023; Qian et al., 2023) in Fig. 9. In the domain of SDS-based task, some works (Liu et al., 2023; Qian et al., 2023) that address 3D consistency essentially receive images as inputs for image-to-3D, which complicates direct comparisons with text-to-3D works. However, as mentioned in Zero123 (Liu et al., 2023), it is possible to indirectly facilitate text-to-3D by first generating images from text using text-to-image generation models like Stable Diffusion (Rombach et al., 2022a). In this context, we adopt such an approach in Fig. 7 of our main paper, providing a qualitative comparison with Zero123 (Liu et al., 2023) and Magic123 (Qian et al., 2023). For the sake of fairness, we disclose the reference images generated by Stable Diffusion in Fig. 12. These images have undergone processing such as background removal, in accordance with the method described in Zero123 (Liu et al., 2023).

Figure 12. **Reference images for image-to-3D works (Liu et al., 2023; Qian et al., 2023).** These images are generated by Stable Diffusion (Rombach et al., 2022a), followed by processing such as background removal.

In certain cases, we have observed results of image-to-3D methods that do not reach the quality of the qualitative results shown in their paper. We conjecture that this is due to the sensitivity of the input images generated from text-to-image models when they diverge from the domain of the training (or fine-tuning) dataset. In contrast, the results of text-to-3D tasks, including our results, seem not to encounter this issue; they bypass the specific reconstruction objectives of these input images and generate results that align well with the trained domain corresponding to the text.

E. Additional Discussion

E.1. Additional qualitative results.

We present additional qualitative results of our approach in Fig. 13.

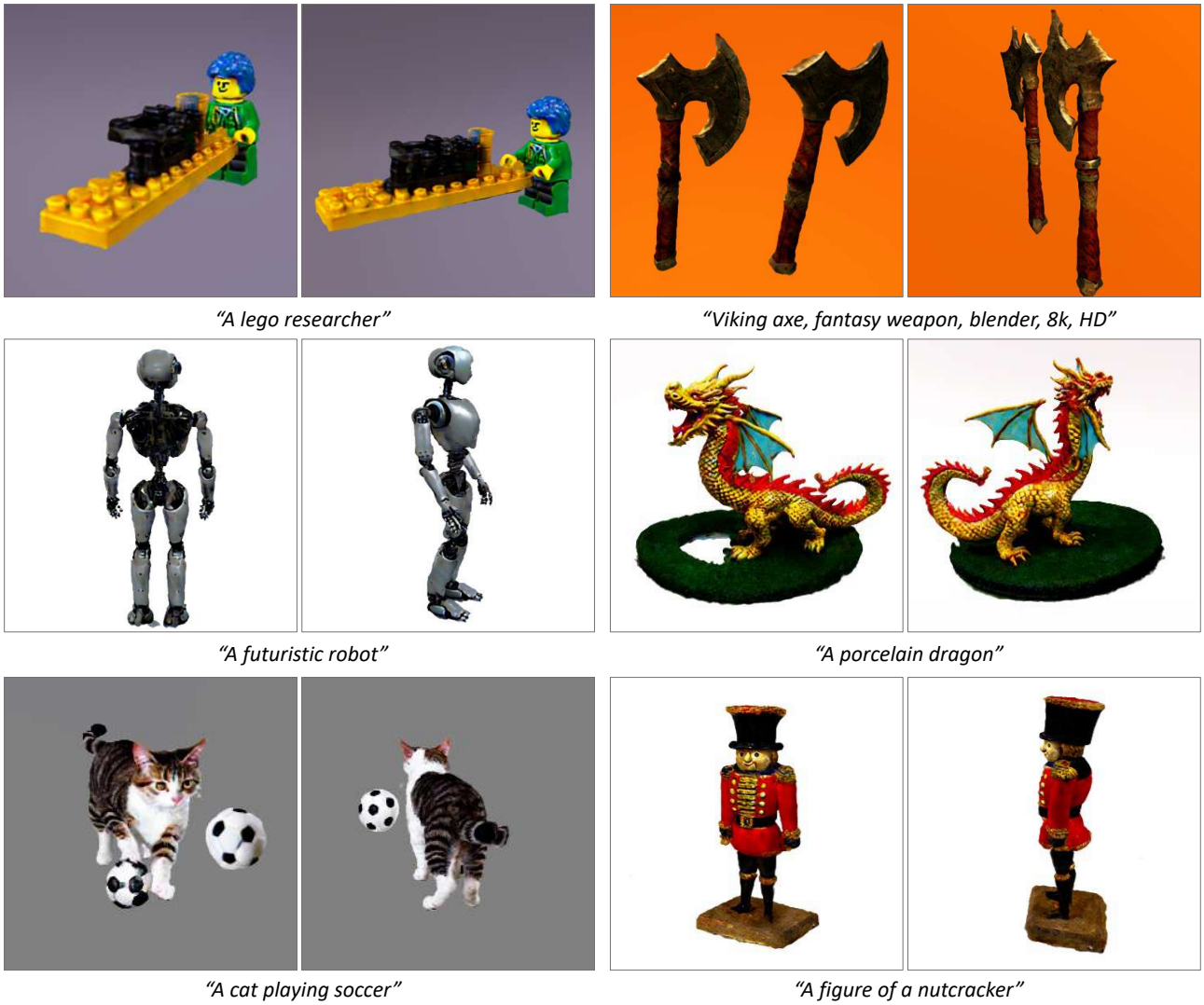

Figure 13. **Additional qualitative results.**

E.2. Ablation on lightweight adaptation

We ablate the components of lightweight adaptation in Fig. 14. In (a) of Fig. 14, the 2D prior model is adapted using only the learnable layers that are embedded within the U-Net. In (b), both the learnable layers and tokens that correspond to the view prefix are adapted. To clearly demonstrate the differences, we present samples by deterministic DDIM (Song et al., 2020a) sampler. We maintain consistency by using the same initial noises for all the samples. We observe that both (a) and (b) effectively mitigate the viewpoint bias inherent in the 2D prior model, indicating that both are capable of guide the model to generate images that are less biased in terms of viewpoint. We also observe that the samples from (b) represent a more diverse range of viewpoints.

E.3. Does lightweight adaptation overfit the model to the retrieved asset?

In this section, we address the concern of potential overfitting to initializing assets during lightweight adaptation. To investigate this, we analyze 2D samples generated using a constant asset with progressively changing prompts. This lets us verify the level of overfitting our model display as it tunes itself to the specific details of the assets. Interestingly, as shown in Fig. 15, our findings indicate that the adaptation focuses more on general aspects, such as viewpoint, rather than specific details like texture. This suggests that lightweight adaptation avoids overfitting to the minor details of of individual assets,

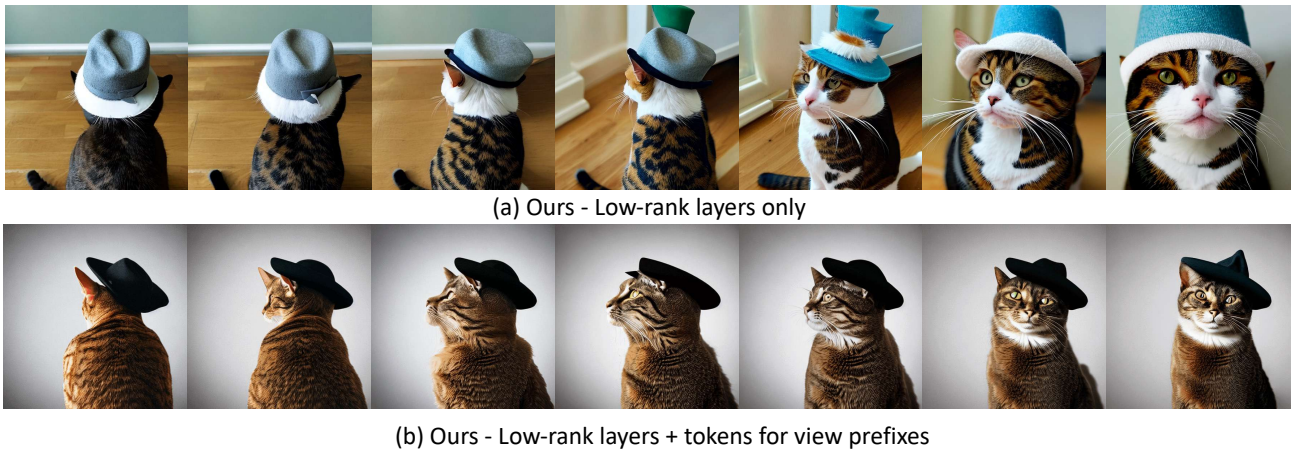

Figure 14. **Ablation on components of lightweight adaptation.** (a): Adaptation using only the learnable layers embedded in the Diffusion U-Net. (b): Adaptation using the learnable layers and tokens corresponding to the view prefix. To clearly demonstrate the difference, we show samples from deterministic DDIM sampler after fixing the initial noise. In both (a) and (b), we observe that the viewpoint bias is effectively removed. However, in the case of (b), it shows samples from a slightly more diverse range of viewpoints.

striking a balance between adapting to the 3D asset and maintaining generalization across various prompts.

E.4. Ablation on each component

We conduct an ablation study on the two primary methodologies proposed in our main paper: initializing the variational distribution, and employing lightweight adaptation. As shown in Fig. 17, initialization of the variational distribution is vital in solidifying coarse geometry, while lightweight adaptation shows its efficacy in preventing Janus problem-like artifacts.

E.5. Orientation alignment

To verify whether retrieved 3D assets are oriented properly and well aligned to our canonical space axis, we measure the success rate of the alignment of assets retrieved for 45 prompts manually. To minimize human error in the measurement process, we follow these principles: 1) If the frontal view is correctly identified, it's a success, and it is considered a failure if its orientation is flipped vertically (upside-down), or flipped sideways, or if the frontal view cannot be identified. 2) We set a reference angle for the frontal view of each asset based on the horizontal axis, and define the failure case as occasions where the frontal view deviates by more than ± 45 degrees from the reference angle. 3) Radially symmetric objects or objects with semantic symmetry that whose frontal view cannot be identified singularly are excluded from the measurement. The results are reported in Tab. 3. Note that failure cases here do not necessarily mean failures in generation, thanks to view prefix optimization in the adaptation.

Success	Vertical / Sideways Inversion	Frontal view not identified
86.7%	6.7%	6.7%

Table 3. **Success rate of orientation alignment.**

E.6. Efficiency

Our method is a retrieval-augmented approach that requires test-time adaptation. Unlike methods such as Zero123 () which involve tuning all of the parameters for 3D awareness with 1,344 GPU hours, our method does not require full fine-tuning in model preparation phase. The aspect of our method not requiring training is similar to DreamFusion () or ProlificDreamer ().

Instead, in inference time, our method includes a few more steps than classic SDS-based methods (); 3D retrieval and lightweight adaptation. The 3D retrieval process takes about 7 seconds. The lightweight adaptation, when measured separately, takes about 7 minutes. However, as it actually proceeds in parallel with the SDS optimization process, the time taken is expected to be less from a total time perspective.

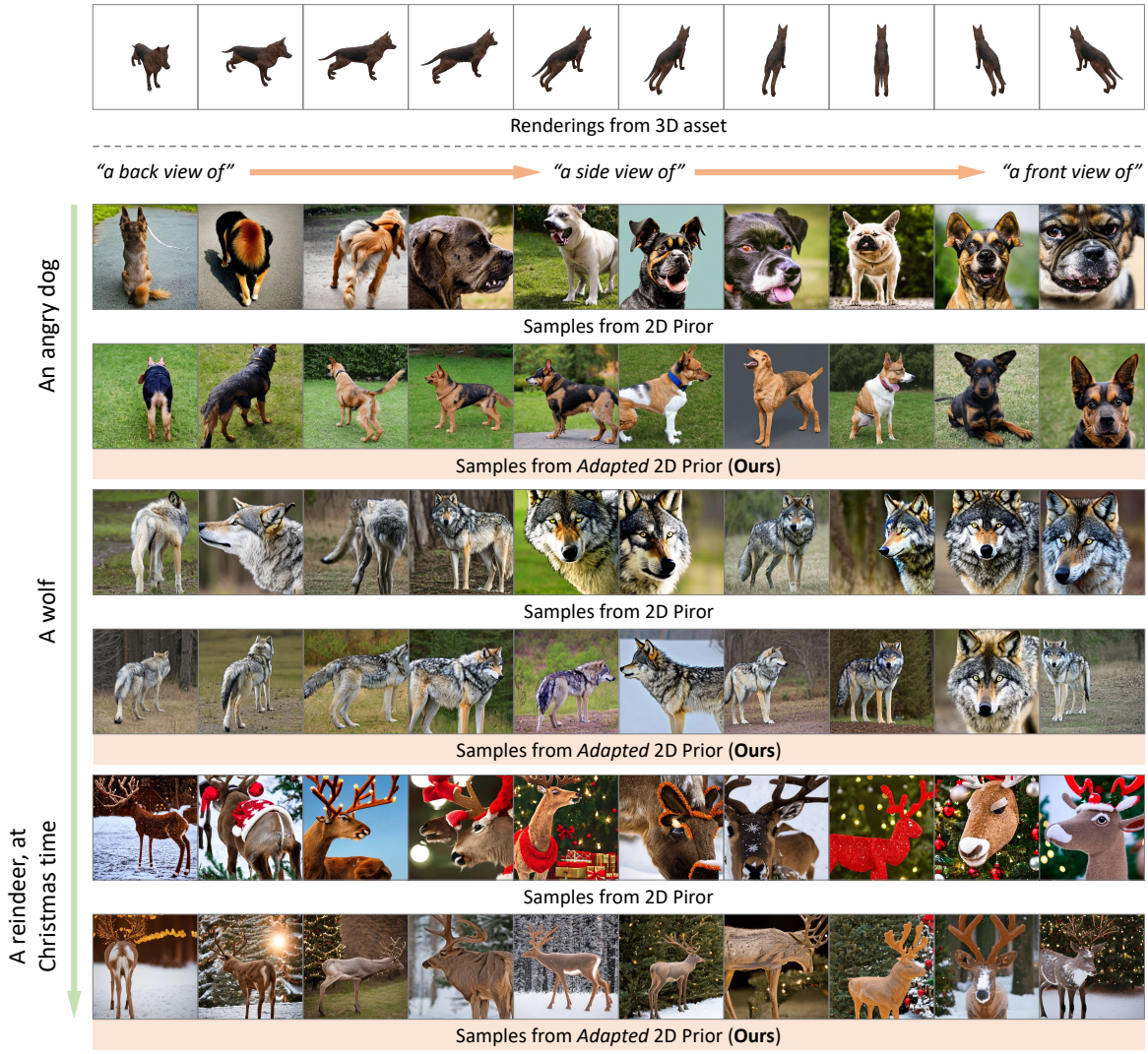

Figure 15. To verify whether the model overfits to the retrieved assets for lightweight adaptation, we report on 2D samples generated by progressively changing prompts, while keeping asset same. It shows that the adaptation focuses more on general aspects like viewpoint, rather than specific details like texture. For more details, see Sec. E.3.

SDS-based methods generate 3D objects through optimization, making it difficult to report exact generation times. This is different from optimization-based methods for 3D reconstruction (Chen et al., 2022a; Müller et al., 2022; Yu et al., 2021) that report time based on reaching a certain PSNR, as it's challenging to know when it is converged. Consequently, we present qualitative results in Fig. 16, the intermediate rendering results at 10,000th iteration. The results show that our method reaches convergence faster than our baseline (Wang et al., 2023b).

Based on this observation, we converge 3D objects over 20,000 iterations, which is 5,000 fewer iterations than the baseline. Therefore, the average time for generation ultimately becomes about 2 hours faster than the baseline.

E.7. Number of particles

In this section, we present an ablation study on the number of particles, as shown in Fig. 18. Similar to the findings in the ablation study for VSD, as described in Appendix E.3 of (Wang et al., 2023b), it has been observed that the impact on quality due to the number of particles is not significant. We have found that there is a tendency for increased diversity with a greater number of particles. This appears to be due to the fact that as the number of particles increases, the ability to absorb a larger number of retrieved assets for distribution initialization becomes more effective, thus enhancing diversity. In the

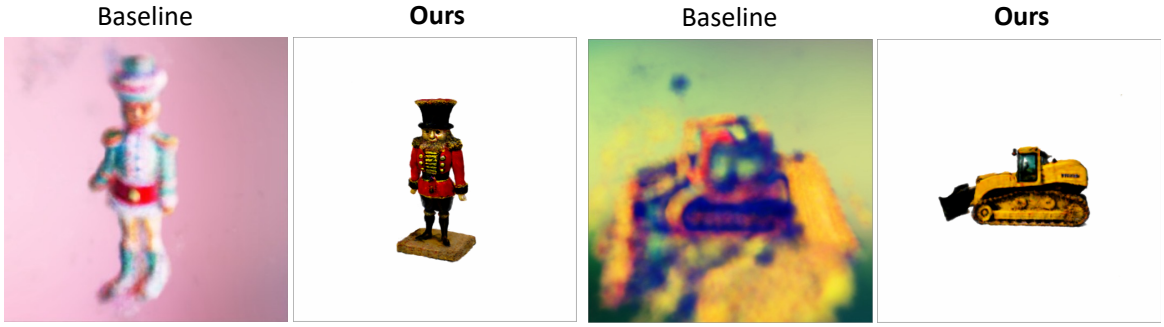

Figure 16. **Intermediate renderings at 10,000th iteration.**

perspective of 3D consistency, even with a smaller number of particles, the number of retrieved assets that can be used in lightweight adaptation is not limited, hence a slight improvement in performance or a similar trend has been observed.

F. Applications

Our retrieval-based text-to-3D generation method can be applied to numerous real-life cases in which 3D models is necessary. In most cases, creating a realistic 3D model requires extensive knowledge of complicated tools and programs, such as CAD, which limits a layman from creatively engaging in 3D scene generation. Our retrieval-based text-to-3D generation method enables the score distillation-based optimization process to be controlled by both text and retrieved asset, giving higher flexibility and diversity to 3D scenes that can be generated. This opens up large possibilities in all areas which requires 3D creation: in AR and gaming. Our retrieval-based methodology can be used to design 3D models of characters or buildings that are more meticulously generated under user control, greatly reducing the redundant time and cost that goes into crafting such 3D assets with hand. Due the realism and fidelity of generated 3D scenes, our model can also be applied to aid 3D design that goes into movies for CGI-based scenes, giving artists more relevant 3D mesh that serves as more efficient template from which they can work and fine-tune upon.

3D data enhancement. Furthermore, our method can also be applied to to specifically enhance the fidelity and details of low-quality 3D assets by simply replacing retrieved asset with the 3D asset to be enhanced. As demonstrated in Sec. E, we show that high-quality 3D model that preserve general geometric structure of the given asset can be created when the assets are not automatically retrieved by hand-picked as input to the given network, despite the low quality of initializing assets. Note also that even when the texts and assets themselves not being completely aligned semantically, (e.g., 3D asset of a plain human figure and text prompt "*A photo of Ironman*"), our model successfully enhances the asset in accordance with the text prompt. This show that our work can be extended to 3D data enhancement in settings where the assets are hand-picked and chosen to be improved by text prompt.

G. Details of User Study

The user study involved a total of 92 participants, each of whom was asked to answer 6 randomly sampled questions. Specifically, each question presented two videos showing our results and baseline's results. It was thoroughly concealed which video is the baseline and which is our result, and the placement of the videos was also randomized. The questions are as follows:

- The text used for this 3D creation is "[TEXT PROMPT]". Considering **texture, shape, geometry**, which result do you find more satisfactory?

This questionnaire was distributed for 3 days in local communities and universities, and stakeholders in this study were strictly excluded, and which result come from which model's results was also strictly blinded. We provide an example of the screen shown to the participants in Fig. 19.

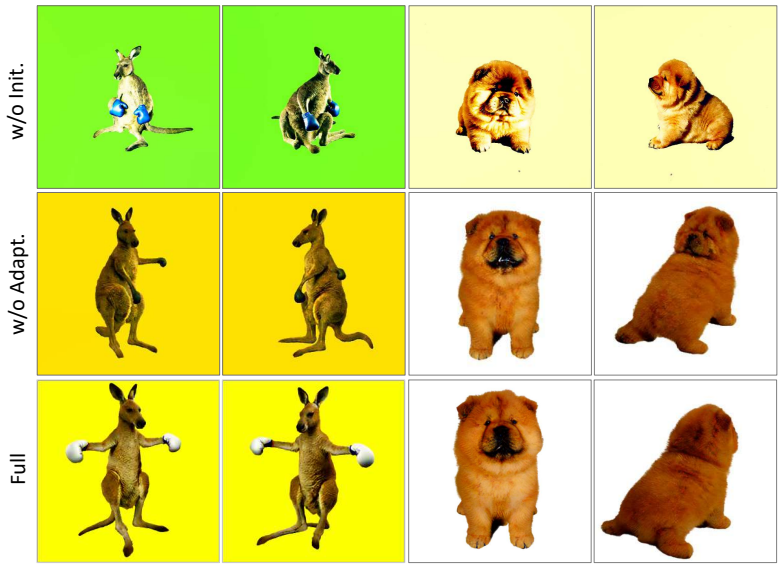

Figure 17. **Ablation on each component.** We drop each component of our framework. Top row and middle row show the results generated by dropping out the initialization of the distribution and lightweight adaptation respectively. The results in the bottom row is produced by our whole framework and they show more consistent geometry compared to upper rows.

H. Limitations

Our generation process approximately takes about 6 hours, which is faster than our implementation baseline, ProlificDreamer (Wang et al., 2023b) which takes about 6 hours. However, it should be noted that, it takes longer time than concurrent works which concentrate on fast inference (Yi et al., 2023), as our goal is to create photorealistic 3D contents like ProlificDreamer, not to make the infernece faster. We believe it would be possible to significantly reduce the time required by applying these techniques in an orthogonal manner as a future work.

Secondly, the receptivity of complex and creative text prompts is bounded by the performance of the 2D prior model, Stable Diffusion (Rombach et al., 2022a). In this paper, while we utilize Stable Diffusion 2.1 for a fair comparison with other work (Wang et al., 2023b), it should be noted that the recent advancements in 2D generative models suggest methods for a better understanding of more complex prompts, which could be pursued in our future work.

Figure 18. **Variation of the number of particles.** We show the generation results with different number of particles.

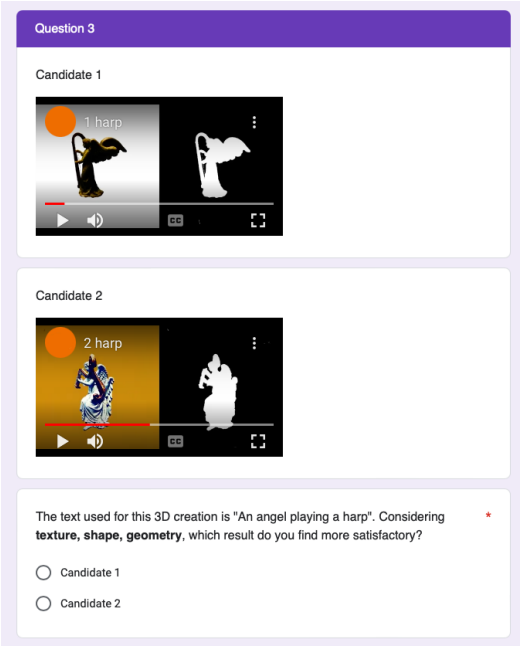

Figure 19. **An example of the screen shown to participants.** For the purposes of anonymization in double-blind review, some contents are obscured in this example.