JointDreamer: Ensuring Geometry Consistency and Text Congruence in Text-to-3D Generation via Joint Score Distillation

Chenhan Jiang[1*†], Yihan Zeng[2*], Tianyang Hu[2], Songcun Xu[2], Wei Zhang[2], Hang Xu[2], and Dit-Yan Yeung[1]

[1] The Hong Kong University of Science and Technology
[2] Huawei Noah's Ark Lab
[*]Equal contribution. [†]Corresponding Author: jchcyan@gmail.com
https://jointdreamer.github.io

Abstract. Score Distillation Sampling (SDS) by well-trained 2D diffusion models has shown great promise in text-to-3D generation. However, this paradigm distills view-agnostic 2D image distributions into the rendering distribution of 3D representation for each view independently, overlooking the coherence across views and yielding 3D inconsistency in generations. In this work, we propose **J**oint **S**core **D**istillation (JSD), a new paradigm that ensures coherent 3D generations. Specifically, we model the joint image distribution, which introduces an energy function to capture the coherence among denoised images from the diffusion model. We then derive the joint score distillation on multiple rendered views of the 3D representation, as opposed to a single view in SDS. In addition, we instantiate three universal view-aware models as energy functions, demonstrating compatibility with JSD. Empirically, JSD significantly mitigates the 3D inconsistency problem in SDS, while maintaining text congruence. Moreover, we introduce the Geometry Fading scheme and Classifier-Free Guidance (CFG) Switching strategy to enhance generative details. Our framework, JointDreamer, establishes a new benchmark in text-to-3D generation, achieving outstanding results with an 88.5% CLIP R-Precision and 27.7% CLIP Score. These metrics demonstrate exceptional text congruence, as well as remarkable geometric consistency and texture fidelity.

Keywords: 3D Vision · 3D Generation · Energy Function

1 Introduction

3D content creation is essential for diverse applications, including gaming, robotics simulation, and virtual reality. However, it is labor-intensive, demanding substantial time for skilled designers to create a single 3D asset. Hence, automating 3D creation with text input has attracted considerable attention. Recently, the score distillation sampling (SDS) algorithm pioneered by DreamFusion [38]

Fig. 1: Text-to-3D generations by JointDreamer from scratch. JointDreamer excels in generating geometrically consistent and high-fidelity 3D assets, adhering to complex textual descriptions that are challenging for previous methods.

shows promise in text-to-3D tasks, which lifts image distribution from a well-trained diffusion model [40] into parameterized 3D representation like NeRF [31]. Compared to 3D generative models [2, 19, 34] that struggle with producing arbitrary objects due to limited text-3D training data, SDS-based methods [6, 24, 38, 45] can generate arbitrary 3D assets with diverse text input.

Although SDS-based methods benefit from the generalizability of diffusion models, they often encounter a common issue known as Janus artifacts [6, 24, 38, 45]. These artifacts manifest as repeated content from different viewpoints of a 3D generation, yielding a lack of realism and coherence in the rendered views. We investigate the Janus artifacts by visualizing image distributions of the diffusion model [40] from multiple viewpoints, as illustrated in Fig. 2(a). The results reveal the view-agnostic nature and the content inconsistency across views of the diffusion model. Consequently, SDS optimizes each rendered view of 3D

representation independently and inherits image distribution without coherent multi-view perspective, leading to the geometric inconsistency of 3D generations.

Existing works [1, 22] address the aforementioned challenges within the SDS framework by employing prompt engineering techniques. However, the effectiveness of such methods remains unsatisfactory, as evident from the results depicted in Fig. 2(b). Alternative methods [16, 43] propose to finetune view-aware diffusion models using rendered images of 3D datasets [7, 8]. Nevertheless, they are prone to overfitting on limited text-3D training data [23], decreasing the text congruence when handling complex text inputs. Based on the above observations, it requires a rethinking of the SDS optimization to enhance the 3D coherence of generations while maintaining generalizability.

In this work, we first present Joint Score Distillation (JSD), which significantly promotes the 3D consistency of generation and inherits generalizability from diffusion models. Specifically, we model the joint image distribution of diffusion model via an energy function measuring coherence across denoised images. It facilitates the extension of the KL-Divergence in SDS from single-view into multi-view. We then derive the joint score distillation function from multi-view KL-Divergence, which ensures inter-view coherence in the optimization process of 3D generation. We show that SDS is a special case of JSD with the energy term omitted, which indicates the absence of coherence constraint across views.

Fig. 2: Illustration of text-conditioned images for different viewpoints, where input texts are augmented with corresponding direction prompts for each view. **(a)** The original generations from 2D diffusion model [40] are view-agnostic and inconsistent across views. **(b)** Text prompt tuning [1] has limited improvement in the directional structure of generated images for each view. **(c)** JSD injects coherence measurement from the proposed binary classifier (refer to Section 4.2), contributing to modified directional structures and semantical consistency across views.

Building upon JSD optimization, we present three view-aware models as energy terms to showcase the compatibility of JSD: the Binary Classification Model, the Image-to-Image Translation Model, and the Multi-view Generation Model. Through empirical analysis, it is observed that different view-aware models introduce distinct coherence measurements, leading to diverse 3D generations, while all contributing to 3D consistency. Furthermore, to facilitate a more comprehensive comparison with existing text-to-3D generation methods, we introduce JointDreamer, an innovative framework capable of producing geometric-consistent and high-fidelity 3D assets adhering to complex text descriptions. Notably, in addition to incorporating a Multi-view Generation model as an energy term in JSD, we introduce two complementary

techniques, namely the Geometry Fading scheme and the Classifier-Free Guidance (CFG) Switching strategy, to enhance generative details.

We systematically assess the quality of our approach, both qualitatively and quantitatively, compared to existing methods. Qualitative results gallery can be found in Fig. 1. Our JointDreamer consistently produces high-fidelity 3D assets and mitigates Janus artifacts in SDS. It maintains text congruence even when confronted with complex text input, achieving 88.5% CLIP R-Precision and 27.7% CLIP Score.

In brief, our contributions are summarized as follows:

- We introduce a novel Joint Score Distillation (JSD) for text-to-3D generation, optimizing multiple views jointly via an energy function to capture inter-view coherence.
- We present three view-aware models as energy functions to show compatibility with JSD, all of them mitigate the Janus problem in SDS.
- We introduce the text-to-3D framework JointDreamer, incorporating complementary Geometry Fading and CFG Switching techniques. Our JointDreamer achieves geometrically consistent and high-fidelity 3D assets even with complex textual inputs.

2 Related Works

Text-to-3D Generation. Existing text-to-3D generation methods can be categorized into two streams: 3D generative models and 2D optimization methods. The former encompasses various deep generative models such as Variational Auto Encoders (VAEs) [12,13], Generative Adversarial Models (GAN) [4,9,10,33,35], diffusion models [5,28,34] and transformer architectures [2,19]. These models are efficient in inference but often struggle with generalizability and training stability, attributed to the limited scope and complexity of available 3D datasets. The latter approach centers around the Score Distillation Sampling (SDS) algorithm proposed by [38], which leverages 2D diffusion model priors [40] for optimizing 3D representations. Subsequent advancements have refined this technique by improving the 3D representations [6, 24], the sampling scheduler [17] and loss design [45]. However, the above approaches overlook the geometric consistency problem, facing inherent multi-face Janus issues in SDS. Prior works try to alleviate the Janus issues with prompt tuning [1] yet they achieve limited effect. Very recent work MVDream [43] addresses the problem by fine-tuning a multi-view diffusion model, but it is susceptible to overfitting on scarce 3D training data, compromising semantic consistency in text-to-3D generations. In this work, we address the fundamental flaw of SDS that optimizes each view independently by introducing a joint optimization function that enforces inter-view consistency, essentially solving the Janus issues in SDS while preserving its generalizability.

Diffusion-based Novel View Synthesis. As an alternative to 3D generation, novel view synthesis models the challenge as a view-conditioned image-to-image translation task. There have been proposals for pose-conditioned image-to-image

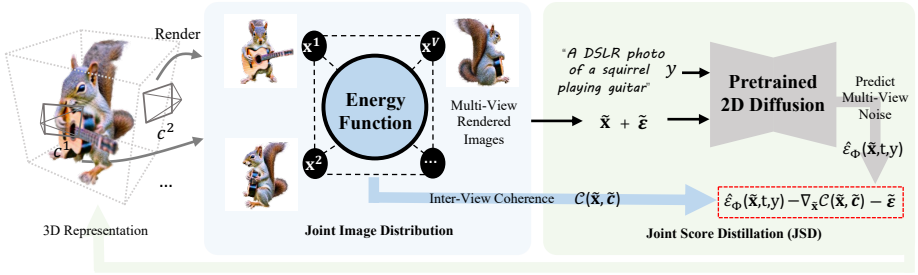


Fig. 3: Overview of JointDreamer Framework. We introduce an energy function to model the joint distribution for multi-view denoised images from 2D diffusion model, facilitating the Joint Score Distillation (JSD) optimization for text-to-3D generation.

diffusion models [46] that generate novel views on synthetic data in 3D. Recently, [26] promoted the generalizability of novel view synthesis by fine-tuning the 2D diffusion model [40] on renderings of 3D dataset, which facilitates images-to-3D tasks with 3D reconstruction or SDS algorithm. To further improve the 3D consistency across generated views and input view, very recent works [27, 29, 42] modify the generation process into multi-view generation and present corresponding architecture designs. Generally, these methods take camera specifications as conditions and enable the viewpoint-aware generations, but they can hardly accurately capture a complete 3D scene consistently and densely. Our method acknowledges the potential of these models to discern relative inter-view relationships, which we harness to provide inter-view coherence for our JSD, thus serving as universal guidance models.

3 Preliminaries

We review the original SDS and address its fundamental limitation: optimizing independently for each single view of 3D representation and distilling with the view-agnostic image distribution of diffusion model. It results in geometric inconsistency issues, dubbed as Multi-Face Janus Problem.

Score Distillation Sampling (SDS). SDS optimization is widely adopted by text-to-3D generation pipelines [6, 24, 30, 38, 45]. Given a 3D representation with learnable parameters θ and a pre-trained 2D diffusion model with noise prediction network $\epsilon_\Phi(\mathrm{x}_t, t, y)$, SDS optimizes θ by minimizing the KL-divergence:

$$\min_\theta D_{KL}(q_t^\theta(\mathbf{x}_t|c,y)||p_t(\mathbf{x}_t|y)). \tag{1}$$

Here, $p_t(\mathrm{x}_t|y)$ is the image distribution sampled from diffusion model, $q_t^\theta(\mathbf{x}_t|c,y)$ is the distribution of rendered image $\mathrm{x}_t = g(\theta, c)$ with respect to camera pose c at timestep t of the forward diffusion process, where g is the renderer. To solve Eq. (1), the score distillation function is derived as:

$$\begin{aligned}\nabla_\theta L_{SDS}(\theta) &\triangleq \mathrm{E}_{t,\mathbf{x}}[w(t)\frac{\sigma_t}{\alpha_t}\nabla_\theta KL(q_t^\theta(\mathbf{x}_t|c,y)||p_t(\mathbf{x}_t|y))] \\ &\triangleq \mathrm{E}_{t,\epsilon_\Phi}[w(t)(\hat{\epsilon}_\Phi(\mathbf{x}_t,t,y)-\epsilon)\frac{\delta g(\theta,c)}{\delta\theta}],\end{aligned} \tag{2}$$

where $\hat{\epsilon}_\Phi := (1+s)\epsilon_\Phi(\mathrm{x}_t, t, y) - s\epsilon_\Phi(\mathrm{x}_t, t, \emptyset)$ is modification of predicted noise with classifier-free guidance (CFG) s, $w(t)$ is time-dependent weighting function.

Multi-Face Janus Problem. To achieve consistency, it is essential that the rendering distributions $q_0^\theta(\mathbf{x}_0|c, y)$ adhere to text condition y and image distribution $p_t(\mathrm{x}_t|y)$ keep consistency across views with different poses. For the image distribution $p_0(\mathbf{x}_0|y)$ of 2D diffusion model, the pose condition can be injected via input text with the corresponding directional prompt [38, 45]. As illustrated in Fig. 2(a), the pre-trained 2D image distribution, trained on individual images, is view-agnostic and lacks identity consistency across views. Even with a text tuning mechanism [1] specifically designed for multi-view image generation, as shown in Fig. 2(b), the above issues are far from resolved. Since SDS minimizes KL-divergence between the image distribution and rendering distribution independently for each rendered view, it can only inevitably inherit the 3D-awareness deficit of the 2D diffusion model, resulting in inconsistent 3D generation, which is commonly referred to as the Multi-face Janus Problem of SDS.

4 Method

In this section, we introduce JointDreamer, a novel text-to-3D generation framework as illustrated in Fig. 3. We first present the derivation of Joint Score Distillation (JSD) in Sec. 4.1, which extends the single-view optimization in SDS into a multi-view KL-Divergence. Then we integrate universal view-aware models into JSD to show the compatibility of JSD in Sec. 4.2, where we instantiate three kinds of view-aware models to capture inter-view coherence. Finally, we elaborate on the overall framework JointDreamer in Sec. 4.3, where we integrate the multi-view generation model into JSD. We also propose a geometry fading scheme and CFG switching strategy to further enhance generative quality.

4.1 Joint Score Distillation (JSD)

To address the multi-face problem arising from SDS, we extend the score distillation from single-view to multi-view settings and promote inter-view coherence across 2D image distribution, and thus derive our JSD optimization function.

Coherence Modeling for Joint Image Distribution. As discussed above, the rendering distributions of the 3D representation should maintain 3D consistency across views $\tilde{\mathbf{x}} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^V\}$ with respect to different poses $\tilde{\mathbf{c}} = \{c^1, c^2, \ldots, c^V\}$. However, for 2D pre-trained diffusion models, different views are generated independently. To ensure consistency, we propose modeling the joint image distribution of multiple views, denoted as $p_0(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}, y)$, within the diffusion model. Following the commonly adopted assumption in energy-based distribution modeling [21, 47, 48], we introduce an energy function that measures the inter-view coherence by $\mathcal{C}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}}) : \mathbb{R}^{Vd} \to \mathbb{R}$ and define:

$$p_0(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}, y) \propto \exp(\mathcal{C}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}})) \prod_{i=1}^{V} p_0(\mathbf{x}^i|c^i, y), \tag{3}$$

where a larger $\mathcal{C}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}})$ indicates greater coherence among the denoised view images. As a result, the joint image distribution is no longer view-independent. In practice, the joint energy function can be implemented via various view-aware models, as long as they can reflect the coherence across multiple views. In Sec. 4.2, we explore different choices in depth. The modeling of coherence across joint image distributions facilitates our JSD on multi-view, integrating inter-view coherence to ensure consistent 3D representation.

KL-Divergence on Multiple Views. We extend the single-view KL-divergence in SDS to a multi-view version, based on the joint image distribution:

$$\begin{aligned} &\min_{\theta} D_{KL}(q_t^{\theta}(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}, y)||p_t(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}, y)) \\ &= \min_{\theta} D_{KL}(q_t^{\theta}(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}, y)|| \exp(\mathcal{C}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}})) \prod_{i=1}^{V} p_t(\mathbf{x}^i|c^i, y), \end{aligned} \tag{4}$$

where the extra energy term $\mathcal{C}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}})$ in Eq. (3) accounts for the inter-view coherence. Without this constraint, e.g., $\mathcal{C}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}}) \equiv 0$, different rendering views are optimized independently with the 2D diffusion model separately. In this sense, the original SDS can be seen as a special case of JSD.

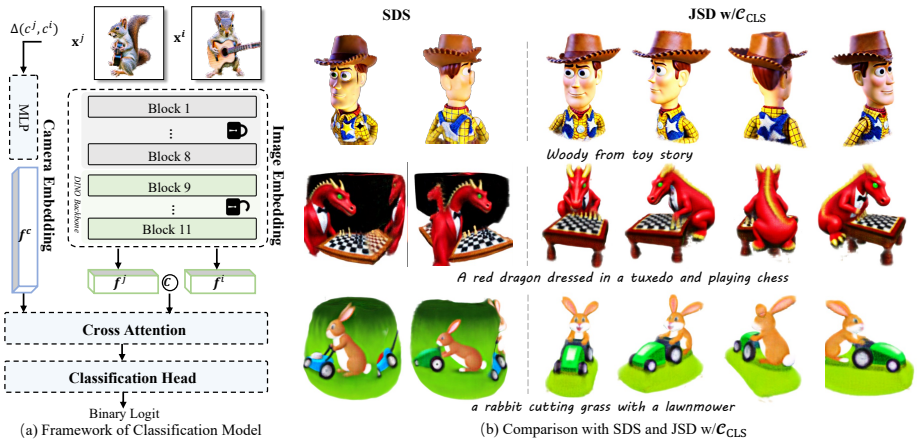
Joint Score Distillation Function. To correspond to the gradient of the new rule of multi-view KL-divergence in Eq. (4), we derive our score distillation function that is jointly conducted on multiple views as follows:

$$\begin{aligned} &\nabla_{\theta} L_{JSD}(\theta) \\ &\triangleq \mathrm{E}_{t,\epsilon_{\Phi}}[w(t)\mathbb{E}(\nabla_{\tilde{\mathbf{x}}} \log q_t(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_0) - \nabla_{\tilde{\mathbf{x}}} \log p_t(\tilde{\mathbf{x}}_t|y))] \\ &= \sum_{i=1}^{V} \mathrm{E}_{t,\epsilon_{\Phi}^i}[w(t)(\hat{\epsilon}_{\Phi}(\mathbf{x}_t^i, y) - \frac{\partial \mathcal{C}(\tilde{\mathbf{x}})}{\partial \mathbf{x}_t^i} - \epsilon^i)\frac{\delta g(\theta, c^i)}{\delta \theta}], \end{aligned} \tag{5}$$

where $\{\epsilon^i\}_{i=1}^{V}$ are noises during score matching for different views. The proof can be found in the Appendix. To intuitively compare SDS with JSD, we sample multi-view images from the forward pass of pre-trained diffusion model [40] as shown in Fig. 2(c). For each view $\mathbf{x}_t$, we randomly select $\mathbf{x}_t'$ from a different view and adopt the binary classification model presented in Sec. 4.2 as the energy function to measure coherence across the two views. The results illustrate that JSD significantly enhances the correspondence with directional prompts and consistency across different views, particularly for the initially biased views such as side and back views. These theoretical and empirical results prove that the multi-face Janus issue in SDS is rooted in the view-agnostic and view-biased 2D image distribution, a challenge effectively mitigated by JSD.

4.2 Universal View-Aware Models as Energy Function

JSD requires an energy function $\mathcal{C}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}})$ to measure coherence across denoised images, as presented in Eq. (3). This energy function plays a crucial role in assessing the consistency between different views in image distribution. To demonstrate the compatibility of JSD, we employ three different types of models trained for various representative multi-view tasks.



(a) Framework of Classification Model (b) Comparison with SDS and JSD w/$\mathcal{C}_{\text{CLS}}$

Fig. 4: Illustration of the binary classification model and qualitative results with JSD. (a) The classification model M_{CLS} produces the binary logit to measure the consistency between two input views x^i and x^j. (b) JSD integrated with the classification model effectively alleviates Janus issues compared to SDS.

Binary classification model M_{CLS}. The binary classification model M_{CLS} is designed to classify the content consistency between two input images based on their relative camera pose. To ensure computational efficiency, we introduce a dedicated classification model, as shown in Fig. 4 (a). The training process takes two days with a single A800 GPU on Objaverse dataset [8]. Further details can be found in the appendix. The classification model M_{CLS} processes pairs of images $\mathbf{x}^i$ and $\mathbf{x}^j$ captured from different viewpoints c^i and c^j. It extracts image features using the DINO-ViT/s16 backbone [3]. These image features are conditioned on the camera feature obtained through MLP layers from the relative camera transformation matrix $\Delta(c^j, c^i)$. Finally, the classification head produces the binary score. To incorporate with JSD, we only consider neighboring views in V as image pairs for coherence measurement, which is denoted as:

$$\mathcal{C}_{\text{CLS}}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}}) = \sum_{i,j \in 1,\dots,V; i \neq j} M_{\text{CLS}}(\mathbf{x}_t^i, \mathbf{x}_t^j, \Delta(c^j, c^i)), \tag{6}$$

where the higher logit indicates the stronger geometric consistency.

Image-to-image translation model M_{I2I}. The image-to-image translation model M_{I2I} is tailored for novel view synthesis [26, 27, 29, 42]. We employ the most recent model, Wonder3D [29] as M_{I2I}, which is a viewpoint-conditioned image translation model and generates consistent content in the target viewpoint. When integrated with JSD, a random reference view $\mathbf{x}^{\mathbf{ref}}$ is selected from the set of 3D rendered images. Then we input the relative camera transformation $\Delta(c^i, c^{\mathbf{ref}})$ and rendered images to M_{I2I}. The measure of consistency is determined by calculating the reconstruction loss between the synthesized new

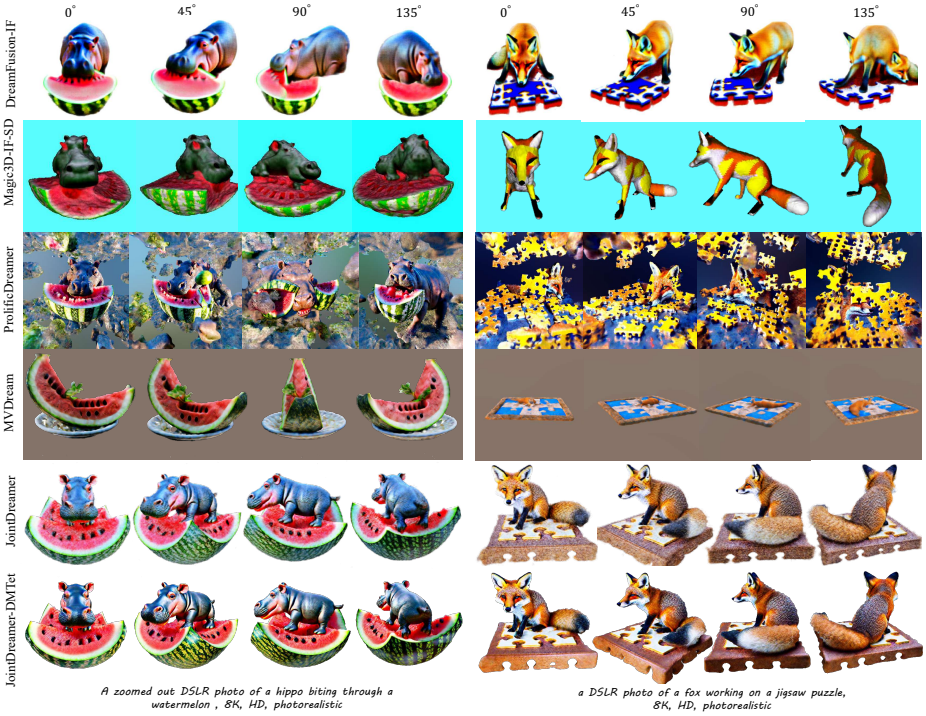


Fig. 5: Comparison of text-to-3D generation. See Appendix for more results.

image and its corresponding rendered image:

$$\mathcal{C}_{\text{I2I}}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}}) = -\sum_{i \in 1,\dots,V} ||M_{\text{I2I}}(\mathbf{x}_t^{\mathbf{ref}}, \Delta(c^i, c^{\mathbf{ref}})) - \mathbf{x}_t^i||_2^2, \tag{7}$$

where a smaller reconstruction loss indicates stronger geometric consistency under the estimation of M_{I2I}.

Multi-view synthesis model $M_{\mathbf{MVS}}$. The multi-view synthesis model is designed to generate multiple images conditioned on text prompts and camera poses, wherein we employ very recent work MVDream [43] as $M_{\mathbf{MVS}}$. We compute the reconstruction loss for multiple generative views and rendered views:

$$\mathcal{C}_{\mathbf{MVS}}(\tilde{\mathbf{x}}, \tilde{\mathbf{c}}) = -||M_{\mathbf{MVS}}(y, \tilde{\mathbf{c}}) - \tilde{\mathbf{x}}||_2^2 \tag{8}$$

The smaller reconstruction loss signifies better geometric consistency within $M_{\mathbf{MVS}}$. These view-aware models measure the coherence across views according to their own 3D-aware insights. Incorporated with JSD, they provide distinct constraints, resulting in different 3D generations that nonetheless contribute to enhanced geometric consistency. We believe that JSD can be adapted to more universal view-aware models, which enables us to progressively redefine the benchmark of 3D generation with the advancement of multi-view tasks.

Fig. 6: Ablations on JSD incorporated with different energy function $\mathcal{C}$. CLS: Binary Classification model; I2I: Image-to-Image Translation model (Wonder3D [29]); MVS: Multi-View Image Synthesis model (MVDream [43]).

4.3 Framework of JointDreamer

Building upon the JSD optimization, we propose the overall framework JointDreamer. The optimization is based on neural radiance field (NeRF) [31], adopting Instant-NGP [32] with volume renderer. We adopt the multi-view synthesis model M_{MVS} as the energy function to integrate with JSD in JointDreamer. During optimization training, we utilize the common techniques including time-annealing and resolution scaling-up following [43, 45]. Besides, we propose two novel techniques to further enhance the generation quality, including a *Geometry Fading* scheme and a *Classifier-Free Guidance Scale (CFG) Switching* strategy.

Geometry Fading. We aim to shift attention between geometric structure and texture details during optimization. Specifically, starting from iteration 5K, we reduce the learning rate of the density network of NeRF from $1e-2$ to $1e-6$ and set orientation loss to 0. Consequently, it benefits geometric convergence in the early phase of optimization, while allowing for decreased attention on geometry and increased attention on texture enhancement in the later stages.

CFG Switching. CFG scheduling strategies have been employed in 2D domain [15, 41] to enhance quality. In this work, we propose to modify the CFG scale s during the training for 3D generation. We are motivated by the observation that a large CFG scale can lead to accelerated geometric convergence but may result in under-optimized geometry and distorted texture. Unlike the annealed CFG approach in CLIP-Sculptor [41], our increasing strategy prioritizes texture while maintaining accurate geometry. Specifically, a smaller $s = 30$ is employed in the early stages to preserve shape integrity, which allows for stronger coherence guidance from JSD. After 5K iterations, we increase the s to 50, enhancing texture fidelity and overall quality.

5 Experiment

In this section, we present the text-to-3D generation results of JointDreamer with qualitative and quantitative evaluations, illustrating state-of-the-art per-

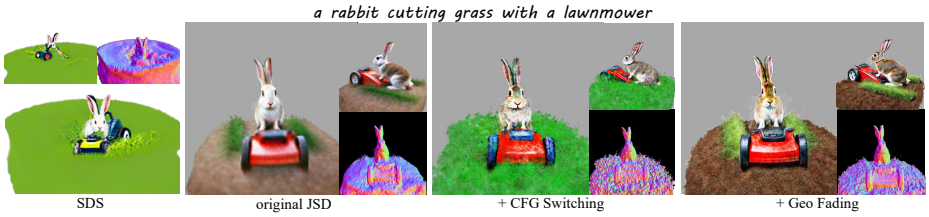


Fig. 7: Ablations on CFG Switching and Geometry Fading techniques in JointDreamer pipeline, demonstrating the effectiveness of quality enhancement.

formance. We also make further ablation analysis on the proposed JSD. More training and evaluation details can be found in the Appendix.

5.1 Text-to-3D Generation

Table 1: Quantitative results on texual consistency and user preference, tested on object-centric subset of MS-COCO [25].

Method	CLIP Score ↑	R-Precision ↑	User Study↑
DreamFusion [38]	20.1	27.7	18.2
ProlificDreamer [45]	25.0	18.7	16.2
MVDream [43]	20.8	33.6	23.5
JointDreamer	**27.7**	**88.5**	**42.1**

Table 2: Ablation study on CFG Switching (CFGS) and Geometry Fading (GF).

SDS	JSD	CFGS	GF	CLIP Score↑	FID↓
✓				20.0	429.2
	✓			27.6	360.7
	✓	✓		28.2	357.6
	✓	✓	✓	28.8	353.9

Qualitative Comparisons. We compare with several representative baselines on threestudio [11] project, which modularizes the text-to-3D framework allowing for fair comparisons by ablating individual components. The generative samples are shown in Fig. 5. *(i) DreamFusion [38]:* Compared to DreamFusion which utilizes traditional SDS optimization, JointDreamer significantly improves the Multi-Face Janus issues in the generations of DreamFusion by introducing an energy term to ensure inter-view coherence. *(ii) Magic3D [24]:* Magic3D introduces a two-stage generation pipeline, transferring NeRF to DMTet in the second stage to enhance generation quality. We also transfer NeRF to DMTet as JointDreamer-DMTet, showcasing consistent superiority in geometry and texture quality by utilizing JSD optimization instead of SDS. *(iii) ProlificDreamer [45]:* ProlificDreamer presents variational score distillation (VSD) as a variant of score distillation function. While VSD enhances photorealism in 3D renderings by introducing a LoRA model, the ill-posed association of pose and images during LoRA training deepens the geometric inconsistency of 3D representation, resulting in severe multi-view artifacts in generations. In contrast, JointDreamer with JSD achieves geometric consistency while maintaining high-fidelity texture quality. *(iv) MVDream [43]:* Compared to direct distillation from a finetuned model as MVDream, JointDreamer employs view-aware models as the coherence constraint in JSD, while still inherit the generalization capabilities of original diffusion models [40]. The results indicate that JointDreamer mitigates the overfitting issues of MVDream, which accurately responds to complex input text and enhances the 3D consistency of generations.

Quantitative Comparisons. Following [18,38], we evaluate CLIP Score [14], CLIP R-Precision [36] and user preference on the object-centric caption subset of MS-COCO [25] with 153 prompts to measure the text congruence and generative quality. For computational efficiency, we generate each 3D asset using 5K iterations of 64×64 rendered images and render 20 images per caption for evaluation. CLIP ViT-B/32 is adopted as the feature extractor for Clip Score and Clip R-Precision. As illustrated by the results in Table 1, JointDreamer can outperform all baselines on CLIP Score and CLIP R-Precision by large margins. Specifically, JointDreamer achieves an improvement of the R-Precision by 60.8% and 54.9% over DreamFusion and MVDream, demonstrating its superior corresponding to textual description. Notably, the severe Janus artifacts in ProlificDreamer compromise the quality of rendering with noisy background and semantic distortion, resulting in the lowest R-Precision. We also conduct a user study about shape preference on these prompts in Table 1.

5.2 Ablation Analyses

Ablations on Energy Functions. Sec. 4.2 discusses the varying inter-view coherency measurements provided by view-aware models trained. When incorporated with JSD, these models have distinct impacts on 3D generations, as shown in Fig. 6. The binary classifier effectively corrects inaccurate geometric structures in SDS. However, it cannot introduce additional imaginative elements as a discriminative model, resulting in oversaturated and monotonous textures. In contrast, as generative models, Wonder3D [29] and MVDream [43] employ reconstruction loss to estimate 3D consistency. Hence, they not only guide geometric structural modifications but also influence texture quality. we further conduct a quantitative comparison using 16 complex multi-Janus prompts, as outlined in Table 3. The experimental setup details can be found in the Appendix. The results indicate that our JSD consistently mitigates Janus artifacts across different view-aware models, with only a slight increase in computational requirements compared to SDS. We find that the energy function $\mathcal{C}_{\text{I2I}}$ derived from the image-to-image translation model exhibits poor performance, likely attributed to a mismatch in camera range and inaccurate translated results.

Table 3: Quantitative comparison of energy functions, showcasing the effectiveness of JSD in mitigating Janus artifacts with comparable computational efficiency to SDS.

Methods	Janus Rate ↓	GPU Memory ↓	Train Time ↓
SDS	100%	16.1 G	50 min.
JSD w/$\mathcal{C}_{\text{CLS}}$	12.5%	22.1 G	80 min.
JSD w/$\mathcal{C}_{\text{I2I}}$	31.2%	16.0 G	119 min.
JSD w/$\mathcal{C}_{\text{MVS}}$	6.2%	19.4 G	54 min.

Analysis on Geometry Fading and CFG Switching Mechanisms. We conduct incremental ablations on our proposed techniques in JointDreamer, including the Geometry Fading scheme and CFG Switching strategy. As shown in Fig. 7, increasing the CFG value enhances texture detail compared to the original JSD, but over-optimizes the geometry, resulting in a bumpier shape. Compared to "+CFG Switching", the Geometry fading effectively protects shape when larger CFG guidance. We also conduct a quantitative evaluation on a 30%

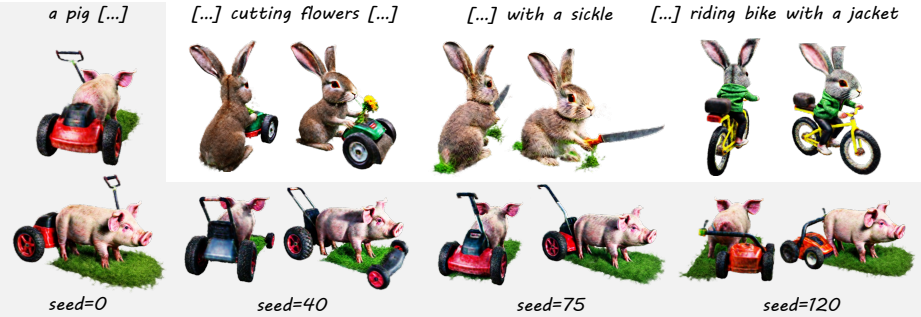


Fig. 9: Discuss the impact of prompts and random seeds, demonstrating the robustness of JointDreamer. *[...]* means the same content as the prompt in Fig. 7.

MS-COCO subset in Table 2. JSD demonstrates superior texture quality compared to SDS, as evidenced by the Clip Score (CS) and FID metrics. The two proposed mechanisms further enhance the texture quality.

Discussions on Training Loss. To make further comparisons with JSD and SDS, we aggregate training losses from multiple prompts on two optimization functions and visualize the training loss curve as illustrated in Fig. 8. We observe that SDS experiences significant fluctuations due to the randomness of single-view optimization. In contrast, JSD can converge gradually and smoothly, demonstrating the introduction of multi-view optimization with inter-view coherence in JSD can reduce the randomness of optimization and contribute to better convergence for 3D representation.

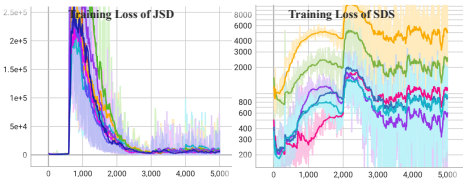


Fig. 8: Training loss comparisons of JSD and SDS. JSD eliminates randomness fluctuations in SDS convergence.

Discussions on Robustness for Prompts and Random Seed. To show the impact of seed and prompt, we modify seeds and key prompt components in Fig. 7, such as subject, object and verb. Results in Fig. 9 demonstrate the robustness of JointDreamer for different seed. Note that the default seed is 0.

The Effectiveness of Classification Model. The classification model surpasses MVDream in training speed by a factor of 48. Fig.4 (b) provides additional high-quality results obtained using JSD in conjunction with the binary classification model M_{CLS}. Furthermore, we conduct an ablation study by replacing $\frac{\partial \mathcal{C}(\tilde{\mathbf{x}})}{\partial \mathbf{x}_t^i}$ in Eq.5 with a randomly generated value between $[0, 1]$. However, it yields shapeless results due to unrelated disturbances in the optimization process. These findings highlight the significance of the proposed classification model in achieving a balance between computational efficiency and generation quality.

More Comparison with MVDream. View-aware models can be incorporated into 3D generation by combining them with JSD or using them directly with

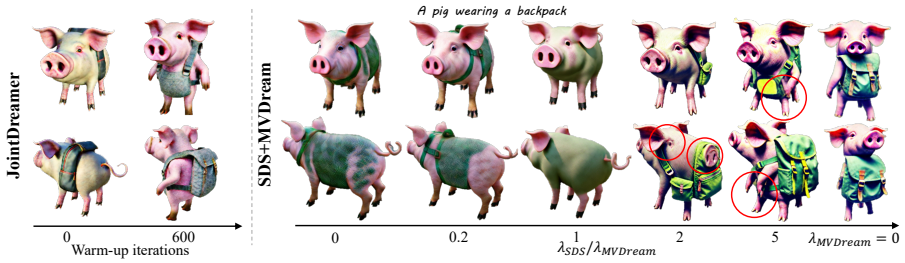


Fig. 10: Comparison of JointDreamer and SDS+MVDream, illustrating the superiority of JointDreamer in mitigating the Janus problem. SDS+MVDream consistently exhibits semantic missing or Janus issues with different weight options.

SDS. In this discussion, we use MVDream [43] as an example to demonstrate the superiority of JSD. MVDream can generate consistent shapes when calculating the SDS loss directly. However, it may miss components in the complete input text due to its fine-tuning on limited 3D data, as illustrated in Fig. 5. To address this limitation, a straightforward approach is to "SDS+MVDream".

However, balancing the impact of SDS and MVDream is challenging. As shown in Fig. 10, setting $\lambda_{\mathbf{SDS}}$ to 0 degrades to MVDream alone, while setting $\lambda_{\mathbf{MVDream}}$ to 0 yields a combination similar to DreamFusion [38]. Achieving a balanced impact between SDS and MVDream through a simple combination is difficult. When $\lambda_{\mathbf{MVDream}}$ is large, textual consistency remains constrained, whereas decreasing $\lambda_{\mathbf{MVDream}}$ leads to the Janus problem. This is due to gradient misalignment across multiple views in the "SDS+MVDream" combination, as SDS lacks multi-view information and cannot derive our objective outlined in Eq. 5. In contrast, JSD based on the joint image distribution provides supervision for text consistency and high-fidelity texture. JSD also promotes inter-view consistency by introducing an energy term with the view-aware model.

6 Conclusion

In this work, we introduce Joint Score Distillation (JSD) as a new paradigm for text-to-3D generation, which conducts multi-view optimization jointly and accounts for inter-view coherence. We demonstrate that JSD can significantly enhance 3D coherence while maintaining generalizability. With other proposed techniques, our overall framework, JointDreamer, is capable of geometric-consistent and high-fidelity 3D generation adhering to complex text input.

Limitations. While the training time of JointDreamer is comparable to existing SDS pipelines, there is room for improvement in terms of acceleration. Future work will explore alternative 3D representations, such as 3D Gaussian [20]. Additionally, JSD utilizes view-aware models to ensure geometry consistency and mitigate the impact of limited data. However, view-aware models still require 3D data for training, and further efficient 3D data collection or reconstruction from multi-view images is also worth investigating.

References

1. Armandpour, M., Zheng, H., Sadeghian, A., Sadeghian, A., Zhou, M.: Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. In: ICLR (2024)
2. Cao, Z., Hong, F., Wu, T., Pan, L., Liu, Z.: Large-vocabulary 3d diffusion model with transformer. In: ICLR (2024)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: CVPR. pp. 16123–16133 (2022)
5. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. arXiv preprint arXiv:2304.06714 (2023)
6. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: ICCV. pp. 22246–22256 (2023)
7. 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. In: NeurIPS (2024)
8. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR. pp. 13142–13153 (2023)
9. Deng, Y., Yang, J., Xiang, J., Tong, X.: Gram: Generative radiance manifolds for 3d-aware image generation. In: CVPR. pp. 10673–10683 (2022)
10. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. NeurIPS **35**, 31841–31854 (2022)
11. 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., Zhang, S.H.: threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio (2023)
12. Henderson, P., Ferrari, V.: Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. IJCV **128**(4), 835–854 (2020)
13. Henderson, P., Tsiminaki, V., Lampert, C.H.: Leveraging 2d data to learn textured 3d mesh generation. In: CVPR. pp. 7498–7507 (2020)
14. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021)
15. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
16. Hu, Z., Zhao, M., Zhao, C., Liang, X., Li, L., Zhao, Z., Fan, C., Zhou, X., Yu, X.: Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion priors. In: CVPR. pp. 4949–4958 (2024)
17. Huang, Y., Wang, J., Shi, Y., Tang, B., Qi, X., Zhang, L.: Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In: ICLR (2023)
18. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: CVPR. pp. 867–876 (2022)

19. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023)
21. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
22. Li, M., Zhou, P., Liu, J.W., Keppo, J., Lin, M., Yan, S., Xu, X.: Instant3d: Instant text-to-3d generation. arXiv preprint arXiv:2311.08403 (2023)
23. Li, W., Chen, R., Chen, X., Tan, P.: Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. In: ICLR (2024)
24. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (2023)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
26. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV. pp. 9298–9309 (2023)
27. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. In: ICLR (2024)
28. Liu, Z., Feng, Y., Black, M.J., Nowrouzezahrai, D., Paull, L., Liu, W.: Meshdiffusion: Score-based generative 3d mesh modeling. In: ICLR (2023)
29. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: CVPR (2024)
30. Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., Zhang, Z.: Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. NeurIPS **36** (2024)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
33. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: Hologan: Unsupervised learning of 3d representations from natural images. In: ICCV. pp. 7588–7597 (2019)
34. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
35. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: CVPR. pp. 11453–11464 (2021)
36. Park, D.H., Azadi, S., Liu, X., Darrell, T., Rohrbach, A.: Benchmark for compositional text-to-image synthesis. In: NeurIPS Datasets and Benchmarks Track (2021)
37. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR (2023)
39. 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)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
41. Sanghi, A., Fu, R., Liu, V., Willis, K.D., Shayani, H., Khasahmadi, A.H., Sridhar, S., Ritchie, D.: Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In: CVPR. pp. 18339–18348 (2023)
42. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
43. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. In: ICLR (2024)
44. Shonenkov, A., Konstantinov, M., Bakshandaeva, D., Schuhmann, C., Ivanova, K., Klokova, N.: Deepfloyd. https://huggingface.co/DeepFloyd (2023)
45. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In: NeurIPS (2024)
46. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. In: ICLR (2023)
47. Weese, J., Kaus, M., Lorenz, C., Lobregt, S., Truyen, R., Pekar, V.: Shape constrained deformable models for 3d medical image segmentation. In: Information Processing in Medical Imaging: 17th International Conference, IPMI 2001 Davis, CA, USA, June 18–22, 2001 Proceedings 17. pp. 380–387. Springer (2001)
48. Zhao, W., Yan, L., Zhang, Y.: Geometric-constrained multi-view image matching method based on semi-global optimization. Geo-spatial information science **21**(2), 115–126 (2018)

This supplementary material consists of five parts, including technical details of the experimental setup (Sec. A), the derivation of Joint Score Distillation (JSD) (Sec. B), additional ablation analysis (Sec. C), additional experimental results (Sec. D) and the Janus prompt list (Sec. E).

A Experimental Setup

A.1 Details of JointDreamer Pipeline.

In our main text, we adopt MVDream $\mathcal{M}_{\text{MVS}}$ as the energy function for the overall JointDreamer pipeline. Since MVDream fine-tunes on SD-V2.1, we retain SD-V2.1 as a diffusion model. The whole training procedure includes 6k iterations, taking around 1.5 h with batch size 4 on 1 Nvidia Tesla A800 GPU. Specifically, we warm up NeRF for the initial 600 training iterations with SDS and adopt JSD for the remaining iterations. We adopt the common time-annealing and resolution-increasing tricks from the open-source implementation, together with the two proposed mechanisms including the Geometry Fading scheme and Classifier-Free Guidance (CFG) Scale switching strategy. We set $t = 0.98$ with resolution 64 for the first 3k iterations and then anneal into $t \sim U(0.02, 0.50)$ with resolution 256 for the extra 2k iterations. Starting from iteration 5k, we scale up the resolution to 512 and conduct the two proposed mechanisms, where the learning rate of the density network is reduced from $1e-2$ to $1e-6$ and the CFG scale is switched from 30 to 50. The Geometry Fading scheme and Classifier-Free Guidance (CFG) Scale switching strategy allow greater influence from coherence guidance in JSD on geometry optimization in the early training stages and enhance the fidelity of textures in later stages.

A.2 Details of Binary Classification Model.

In this part, we will elaborate on the model architecture and training procedure of the binary classification model that is discussed in Sec.4.2 in the main paper.

Model Architecture. We build the model based on the DINO framework. Specifically, we employ ViT-s16 as the backbone for extracting image features. The backbone is initially pre-trained following the DINO method, and during training, the first 9 blocks of the backbone are frozen. Besides, we use a 4-layer MLP with 256 hidden layer channels to extract the relative camera embedding of the transformation matrix between input images, which captures the camera-specific information. Next, we calculate the cross-attention between camera embedding and the concatenated image features of input image pairs. This cross-attention mechanism generates a residual feature input, combined with the concatenated image features as the final feature. Finally, the combined features are fed into the classification head consisting of a 3-layer MLP, which produces the classification logit prediction for input image pairs.

Training Procedure. For training data, we use rendered images from

Objaverse [8] following Zero-1-to-3 [26]. For the binary classification training objective, we adopt the pairs of images from the same object equipped with the correct camera pose as the positive samples and assign the image pairs from different objects or incorrect relative camera poses as negative samples. Before training, we prepare the index list of positive and negative pairs for efficient training. During training, we randomly sample 1 million positive pairs and 1 million negative pairs from the index list as training sets. The design of the training set ensures that the classification model can identify the 3D consistency between rendered images conditioned on relative camera pose. We adopt adamW optimizer with $5e-4$ learning rate and 0.04 weight decay. We also adopt random color jitter, gaussian blur, and polarization following DINO as data augmentation. We use an image size of 224×224 and a total batch size of 640 and train the model for 10 epochs. The training takes about 1 day on 2 Nvidia Tesla A800 GPUs. To validate the classification accuracy, We random sample 5000 pairs as the validation set. The training loss and validation accuracy curve can be found in Fig. A1.

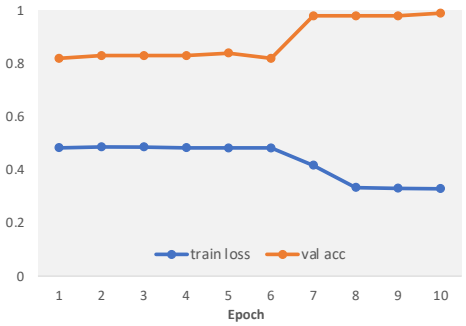


Fig. A1: Training loss and validation accuracy curves of the proposed Binary Classification Model.

A.3 Details of Text-to-3D Generation Comparison

Baseline Setup. We implement the experiments in an open-source three-studio project and reproduce DreamFuion-IF, Magic3D-IF-SD, and Prolific-Dreamer as baselines following the comparisons in the main paper of MVDream. Our MVDream baseline is reproduced by its officially released code. We adopt DeepFloyd-IF [44] as the 2D diffusion model for baseline DreamFuion-IF and the first stage of Magic3D-IF-SD following MVDream. To make a fair comparison with our JointDreamer, we equip the same batch size, resolution, and time annealing strategy with JointDreamer for DreamFuion-IF.

Evaluation Details. We conducted a user study from 100 users on the 153 generated models from the object-centric MS-COCO subset. Each user is given 4 rendered videos with their corresponding text input from generations of different methods. We ask the users to select a preferred 3D model from four options, and then calculate the mean proportion of each method selected over all 153 prompts as the score. The higher score indicates the greater user preference. For the Clip Score and Clip R-Precision, we adopt the CLIP ViT-B/32 as the feature extractor.

A.4 Details of Computational Resource Comparison

We analyze the geometry consistency and computation efficiency of various view-aware models in main paper Table 3, using 16 complex multi-Janus prompts in

Fig. A2: More quality results of JSD with Classification Model.

Sec. E from the DreamFusion [38] library. We maintain consistent experimental parameters, including a batch size of 4, training 5k iterations and a resolution of 64, as well as the same optimizer and time annealing hyperparameters. The only variation is in the camera parameters, which align with each view-aware model's settings. For the baseline SDS model, we adopt the DreamFusion camera parameters. We present some examples showcasing these results incorporating $\mathcal{C}_{\mathbf{CLS}}$ in Figure A2. And the results incorporating $\mathcal{C}_{\mathbf{MVS}}$ can be found in Section D.

B Theory of Joint Score Distillation

Given a well-trained text-to-image diffusion model, like Stable Diffusion, the objective is to distill its knowledge into a 3D representation network parameterized

by θ, such as NeRF and ensures coherent 3D generations. To achieve this, we aim to model the joint rendering distribution across multiple views of θ.

For ease of notation, we define $\tilde{\boldsymbol{x}}$ as the joint random variable comprising $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^V$, which are rendered images sampled from the 3D representation θ. It is important to note that these views are not independent. In a 3D model, the views are inherently connected as they originate from the same underlying 3D object. This means that the rendered images, $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^V$, exhibit dependencies and correlation.

Denote the joint rendering distribution of $\tilde{\boldsymbol{x}}$ as $\tilde{q}^\theta$. We can still define the marginal distributions as

$$q^\theta(\boldsymbol{x}^i) = \int \tilde{q}^\theta(\tilde{\boldsymbol{x}}) d\tilde{\boldsymbol{x}}^{-i},$$

where $\tilde{\boldsymbol{x}}^{-i} = \boldsymbol{x}^1, \ldots, \boldsymbol{x}^{i-1}, \boldsymbol{x}^{i+1}, \ldots, \boldsymbol{x}^V$. This marginal distribution is the same as if only a single view is considered, i.e., $V = 1$.

We can further define the log density ratio as

$$R(\tilde{\boldsymbol{x}}) = \log \frac{\tilde{q}^\theta(\tilde{\boldsymbol{x}})}{\prod_{i=1}^V q^\theta(\boldsymbol{x}^i)}$$

to capture the inter-relationship among different views. Equivalently, we can write

$$\tilde{q}^\theta(\tilde{\boldsymbol{x}}) = \exp(R(\tilde{\boldsymbol{x}})) \prod_{i=1}^V q^\theta(\boldsymbol{x}^i).$$

To get the evaluations of $\tilde{\boldsymbol{x}}$ from the 2D diffusion model, we have

$$\tilde{p}(\tilde{\boldsymbol{x}}) \propto \exp(\mathcal{C}(\tilde{\boldsymbol{x}})) \prod_{i=1}^V p(\boldsymbol{x}^i)$$

since the diffusion model only takes a single image as input and different views are weighted by the introduced joint energy function $\mathcal{C}$.

Now we consider learning $\tilde{q}^\theta(\tilde{\boldsymbol{x}})$ such that the following Integral Kullback–Leibler (IKL) divergence is minimized along the forward diffusion process $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \epsilon$ where ϵ follows standard Gaussian distribution.

$$\begin{aligned}
\min_\theta D_{\mathrm{IKL}}(\tilde{q}^\theta(\tilde{\boldsymbol{x}}) || \tilde{p}(\tilde{\boldsymbol{x}})) &= \min_\theta \int_0^T w(t) \frac{\sigma_t}{\alpha_t} D_{\mathrm{KL}}(\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}) || \tilde{p}_t(\tilde{\boldsymbol{x}})) dt \\
&= \min_\theta \int_0^T w(t) \frac{\sigma_t}{\alpha_t} \mathbb{E}_{\tilde{\boldsymbol{x}}_t \sim \tilde{q}_t^\theta} \left(\log \frac{\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t)}{\tilde{p}_t(\tilde{\boldsymbol{x}}_t)} \right) dt.
\end{aligned}$$

Taking gradient with respect to θ gives

$$\begin{aligned}
&\frac{\partial}{\partial\theta} D_{\mathrm{IKL}}(\tilde{q}^\theta(\boldsymbol{x})||\tilde{p}(\boldsymbol{x})) \\
&= \int_0^T w(t)\frac{\sigma_t}{\alpha_t}\frac{\partial}{\partial\theta}\mathbb{E}_{\tilde{\boldsymbol{x}}_t\sim\tilde{q}_t^\theta}\left(\log\frac{\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t)}{\tilde{p}_t(\tilde{\boldsymbol{x}}_t)}\right)dt \\
&= \int_0^T w(t)\frac{\sigma_t}{\alpha_t}\mathbb{E}_{\tilde{\boldsymbol{x}}_t\sim\tilde{q}_t^\theta}\left[\frac{\partial}{\partial\tilde{\boldsymbol{x}}_t}\left(\log\frac{\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t)}{\tilde{p}_t(\tilde{\boldsymbol{x}}_t)}\right)\frac{\partial\tilde{\boldsymbol{x}}_t}{\partial\theta} + \frac{\partial}{\partial\theta}\log\tilde{q}_t^\theta(\boldsymbol{x})|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}_t}\right]dt \\
&:= A + B.
\end{aligned}$$

The term B vanishes since

$$\begin{aligned}
B &= \int_0^T w(t)\frac{\sigma_t}{\alpha_t}\mathbb{E}_{\tilde{\boldsymbol{x}}_t\sim\tilde{q}_t^\theta}\frac{\partial}{\partial\theta}\log\tilde{q}_t^\theta(\boldsymbol{x})|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}_t}dt \\
&= \int_0^T w(t)\frac{\sigma_t}{\alpha_t}\mathbb{E}_{\tilde{\boldsymbol{x}}_t\sim\tilde{q}_t^\theta}\frac{\frac{\partial}{\partial\theta}\tilde{q}_t^\theta(\boldsymbol{x})|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}_t}}{\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t)}dt \\
&= \int_0^T w(t)\frac{\sigma_t}{\alpha_t}\int\frac{\partial}{\partial\theta}\tilde{q}_t^\theta(\boldsymbol{x})|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}_t}dt \\
&= \int_0^T w(t)\frac{\sigma_t}{\alpha_t}\frac{\partial}{\partial\theta}\int\tilde{q}_t^\theta(\boldsymbol{x})dt \\
&= 0
\end{aligned}$$

The term A is the score distillation loss

$$A = \int_0^T w(t)\frac{\sigma_t}{\alpha_t}\mathbb{E}_{\tilde{\boldsymbol{x}}_0\sim\tilde{q}_0^\theta,\tilde{\epsilon}}\left(\nabla\log\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t) - \nabla\log\tilde{p}_t(\tilde{\boldsymbol{x}}_t)\right)\frac{\partial\tilde{\boldsymbol{x}}_t}{\partial\theta}dt,$$

where $\tilde{\epsilon} = (\epsilon^1, \ldots, \epsilon^V)$ are the noises along the forward diffusion process. Putting things together we have

$$\frac{\partial}{\partial\theta} D_{\mathrm{IKL}}(\tilde{q}^\theta(\boldsymbol{x})||\tilde{p}(\boldsymbol{x})) = \mathbb{E}_{\tilde{\boldsymbol{x}}_0\sim\tilde{q}_0^\theta,\tilde{\epsilon},t}\left[w(t)\frac{\sigma_t}{\alpha_t}\left(\nabla\log\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t) - \nabla\log\tilde{p}_t(\tilde{\boldsymbol{x}}_t)\right)\frac{\partial\tilde{\boldsymbol{x}}_t}{\partial\theta}\right]$$

Notice that the NeRF rendering is a deterministic process given the view information. Therefore, the conditional distribution and marginal distribution coincide, i.e.,

$$\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t) \sim N(\alpha_t\tilde{\boldsymbol{x}}_0, \sigma_t^2), \quad \nabla\log\tilde{q}_t^\theta(\tilde{\boldsymbol{x}}_t) = -\tilde{\epsilon}/\sigma_t.$$

On the other hand, direct score matching tells us that

$$\nabla\log p_t(\boldsymbol{x}_t^i) = \frac{\partial\mathcal{C}(\tilde{\boldsymbol{x}})}{\partial\boldsymbol{x}_t^i} - \hat{\epsilon}_\Phi(\boldsymbol{x}_t^i, t)/\sigma_t.$$

Finally, combining $\frac{\partial\boldsymbol{x}_t^i}{\partial\theta} = \alpha_t\frac{\partial\boldsymbol{x}_0^i}{\partial\theta}$, we have

$$\frac{\partial}{\partial\theta} D_{\mathrm{IKL}}(\tilde{q}^\theta(\boldsymbol{x})||\tilde{p}(\boldsymbol{x})) = \mathbb{E}_{\tilde{\boldsymbol{x}}_0\sim\tilde{q}_0^\theta,\tilde{\epsilon},t}\left[w(t)\sum_{i=1}^V\left(\hat{\epsilon}_\Phi(\boldsymbol{x}_t^i, t) - \frac{\partial\mathcal{C}(\tilde{\boldsymbol{x}})}{\partial\boldsymbol{x}_t^i} - \epsilon^i\right)\frac{\partial\boldsymbol{x}_0^i}{\partial\theta}\right]. \tag{9}$$

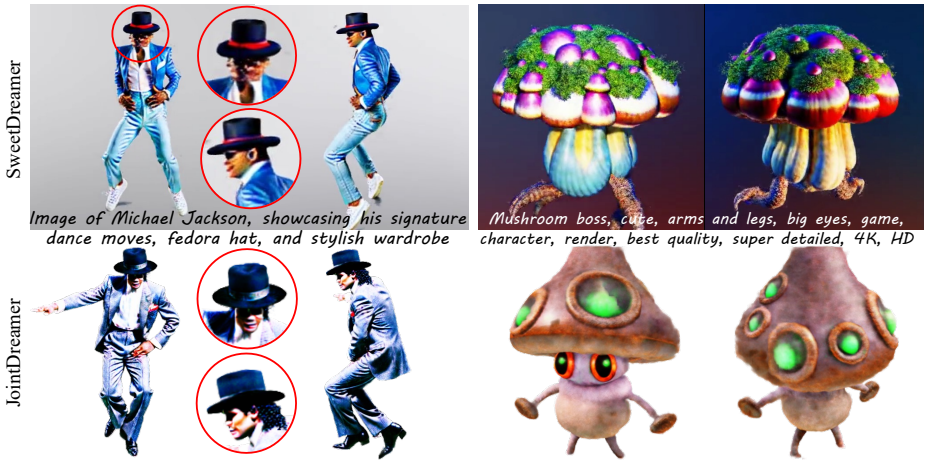


Fig. A3: Comparison with SweetDreamer. SweetDreamer suffers from multi-faces (left) and missing components such as "legs" and "eyes" (right).

Now we have finished extending SDS to multiple views. As it turns out, the joint energy term $R(\tilde{\boldsymbol{x}})$ does not show up in the gradient formula.

C Additional Ablation Study

C.1 Comparison with SweetDreamer

We also conduct a comparison with SweetDreamer [23]. SweetDreamer aligns geometric priors (AGP) in a finetuned diffusion model and combines AGP with SDS to address the Janus issue. In contrast, JSD improves the optimization objective of SDS with various energy functions, and AGP can be one of them. For 3D generation, Fig. 10 shows that a simple combination, like SweetDreamer's, uses more memory and complicates balancing components. Compared to SweetDreamer's demos from its website, our JointDreamer achieves better shape and text congruence without multi-faces and missing components ("arms", "big eyes") in Fig. A5.

C.2 Discussions on Image-to-3D Methods

Since the view-aware models can engage in 3D generation through SDS besides JSD, we make comparisons to showcase the superiority of JSD. Section 5.2 details the comparative use of MVDream, and herein, we extend this comparison to different applications of the image-to-image translation model, Zero-1-to-3 XL, which excels in image-to-3D tasks. Unlike text-to-3D approaches that generate 3D models from textual descriptions, the image-to-3D method uses a reference image to fix the reference view and generate the remaining views. As shown in Fig. A4, we input a reference image, exemplified by the front-view rendered image

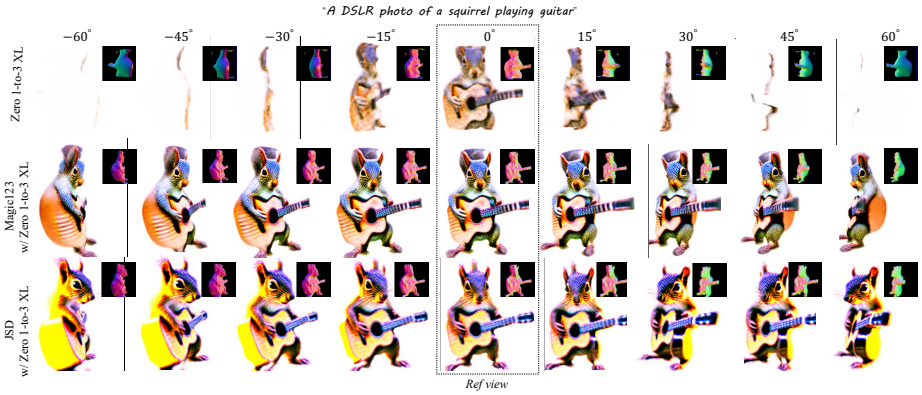


Fig. A4: Comparison with Image-to-3D methods. Compared with two alternative methods, all employing the Zero-1-to-3 XL model, our proposed JSD exhibits superior generative quality in novel view synthesis as evidenced by its geometric consistency.

of the case of "A DSLR photo of a squirrel playing guitar" in Fig. A6 and compare with two alternative utilizations of Zero-1-to-3 XL. *(i)Zero-1-to-3 XL [26]*, which directly utilizes Zero-1-to-3 XL to calculate SDS loss for novel rendered views according to reference view. The overfitting generalizability of Zero-1-to-3 XL reduces the generative quality, especially for the views distant from the reference view. *(ii)Magic123 [39]*, which merges the SDS loss of SD-V2.1 and Zero-1-to-3 XL as objective function. By combining the generalizability from the original diffusion model, it can eliminate the distortion in novel views, but the effect is not satisfactory. By contrast, our JSD achieves better generation quality in novel views, where the overall geometric structure is more reasonable. Notably, when applying JSD in image-to-3D generation, we calculate the inter-view coherence between the reference view and random novel views to fix the reference view, differing from the two random novel views used in text-to-3D generation. The comparisons further illustrate that JSD provides the optimal solution to combine generalizability from 2D models and geometric understanding from 3D-aware models.

C.3 Discussion on Failure Cases

Despite JointDreamer's impressive performance in handling detailed descriptions and multi-object combinations in long texts (as depicted in Fig. 1 of the main paper), it faces difficulties in comprehending complex relationships among objects. Specifically, it struggles to grasp relative spatial arrangements and hierarchical dependencies, as evidenced in Fig. A5. Exploring the use of larger diffusion models, such as SDXL [37], may offer a potential solution to overcome these limitations.

Fig. A5: Failure Cases on MS-COCO Subset.

D Additional Results of JointDreamer

We present more comparisons of text-to-3D generation as shown in Fig. A6, A7 and A8. The results indicate that JointDreamer outperforms current text-to-3D generation methods regarding generation fidelity, geometric consistency, and text congruence. This further validates the effectiveness and generalization of the proposed JSD. We also provide more images and normal maps from additional generated results in Fig. A9, demonstrating the generalizability of JointDreamer with arbitrary textual descriptions.

E Janus Prompts.

Our list of 16 Janus prompts is shown below:

"a blue jay standing on a large basket of rainbow macarons",
"a confused beagle sitting at a desk working on homework",
"Albert Einstein with grey suit is riding a moto",
"a panda rowing a boat in a pond",
"a wide angle zoomed out DSLR photo of a skiing penguin wearing a puffy jacket",
"a zoomed out DSLR photo of a baby monkey riding on a pig",
"a zoomed out DSLR photo of a fox working on a jigsaw puzzle",
"a DSLR photo of a pigeon reading a book",
"a DSLR photo of a cat lying on its side
batting at a ball of yarn"
"A crocodile playing a drum set"
"a rabbit cutting grass with a lawnmower",
"A red dragon dressed in a tuxedo and playing chess",
"a zoomed out DSLR photo of a bear playing electric bass",
"A bald eagle carved out of wood, more detail",
"A pig wearing back pack".
"a lemur drinking boba".

Fig. A6: More comparison of text-to-3D generation.

Fig. A7: More comparison of text-to-3D generation.

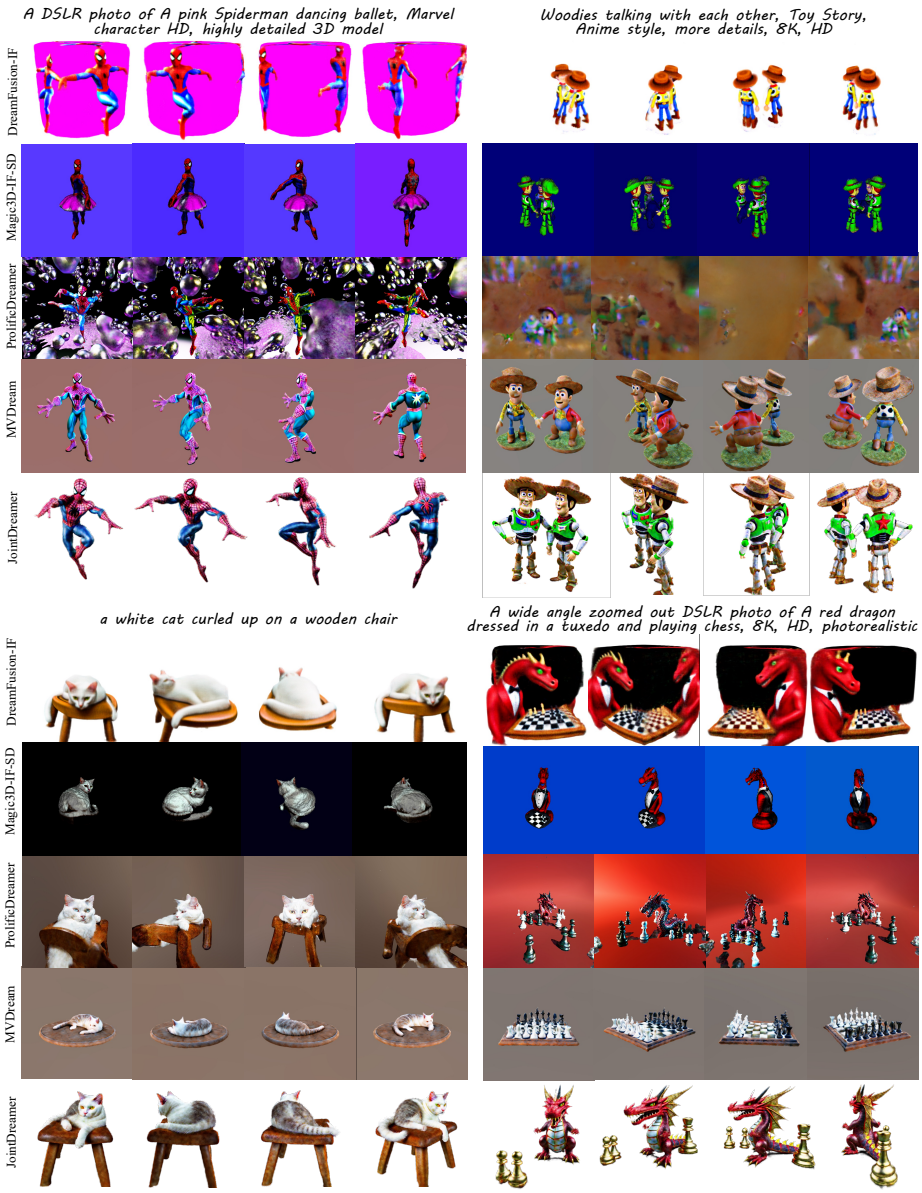


Fig. A8: More comparison of text-to-3D generation.

Fig. A9: More results of JointDreamer.