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

Anonymous CVPR submission

Paper ID 9474

This supplementary material consists of five parts, including technical details of the experimental setup (Sec. 1), the derivation of Joint Score Distillation (JSD) (Sec. 2), additional ablation analysis (Sec. 3), additional experimental results (Sec. 4) and the Janus prompt list (Sec. 5).

1. Experimental Setup

1.1. 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 text.

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 following Zero-1-to-3. 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. During training, we randomly sample 1 million positive pairs and 1 million negative pairs 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. 1.

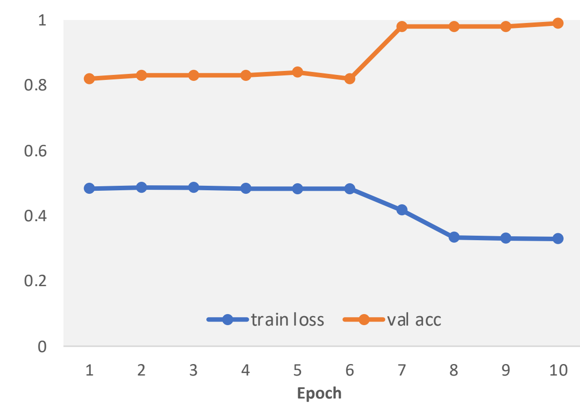

Figure 1. **Training loss and validation accuracy curves** of the proposed Binary Classification Model.

1.2. Details of JointDreamer Pipeline.

In our main text, we adopt MVDream $\mathcal{C}_{(\text{III})}$ as the energy function for the overall JointDreamer pipeline. The whole training procedure includes 7k iterations, taking around 1.5 h with batch size 4 on 1 Nvidia Tesla A800 GPU. Specifically, we warm up NeRF for the initial 500 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)

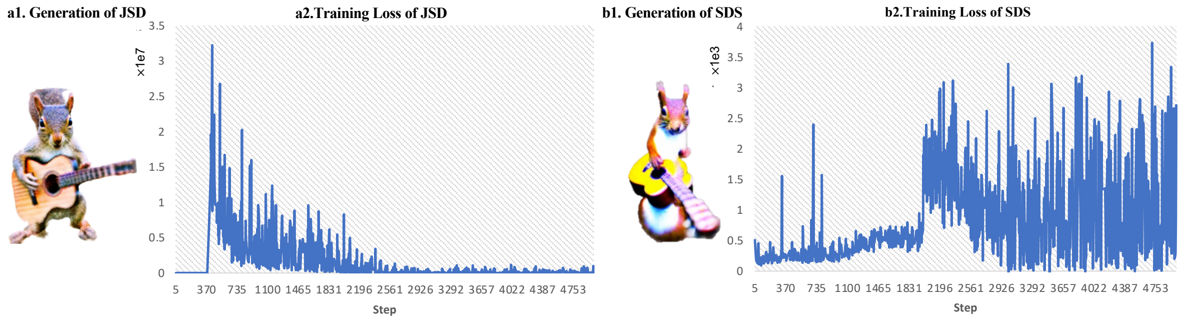

Figure 2. **Comparisons of score distillation training loss**. JSD eliminates the randomness fluctuation in the convergence of SDS and achieves better convergence due to multi-view optimization with inter-view coherence, contributing to enhanced 3D generation quality.

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 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.

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

Baseline Setup. We implement the experiments in an open-source threestudio project and reproduce DreamFuion-IF, Magic3D-IF-SD, and ProlificDreamer 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 [10] 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 10 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.

2. Theory of Joint Score Distillation

We want to match the joint distributions between the well-trained 2D diffusion model and the rendering distribution of 3D representation (NeRF). Recall the notations for multiple views (V views) that we denote $\tilde{\mathbf{x}} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_V})$ and $\tilde{\mathbf{c}} = (c_1, c_2, \ldots, c_V)$. The score information learned from the 2D diffusion model is denoted as $\nabla_{\tilde{\mathbf{x}}} \log p_t(\tilde{\mathbf{x}}_t|y)$, which can be directly factored as

$$\begin{aligned} &\nabla_{\tilde{\mathbf{x}}} \log p_t(\tilde{\mathbf{x}}_t|y) \\ &= \mathrm{diag}(\nabla_{\mathbf{x_1}} \log p_t(\mathbf{x}_1|y), \ldots, \nabla_{\mathbf{x_K}} \log p_t(\mathbf{x}_V|y)). \end{aligned}$$

Though the 2D diffusion model is biased across views, we don't want to modify it. Instead, the consistency requirement is applied to the rendering distribution of 3D representation, without which, we are basically doing SDS for different views separately and independently. We consider an inter-view coherency measure (generalized to accommodate the diffusion process)

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

where $q_t(\tilde{\mathbf{x}}$ denotes the joint distribution along the forward diffusion path and the joint energy term $\mathcal{C}_t$ is also written as diffusion time-dependent. In practice, the universal view-aware models do not have to adapt to noisy samples and align with the diffusion process. As is shown in [1], pre-trained models on noiseless data can also provide effective guidance along the diffusion generation process. Ma et al. [6] further demonstrated that with proper designs, off-the-shelf discriminative models can even be better at guiding diffusion generation than specifically fine-tuned ones. With a slight abuse of notation, we use $\mathcal{C}_t$ and $\mathcal{C}$ interchangeably.

We extend the single-view KL-divergence in SDS to a multi-view version, based on the joint rendering distribution:

$$\begin{aligned} &\min_{\theta} D_{KL}(q_t^{\theta}(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}, y)||p_t(\tilde{\mathbf{x}}|y)). \\ &= \min_{\theta} \mathbb{E}_{q_t^{\theta}(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}, y)} \left(\mathcal{C}(\tilde{\mathbf{x}}|\tilde{\mathbf{c}}) + \sum_{i=1}^{V} \log \frac{q_t^{\theta}(\mathbf{x}^i|c^i, y)}{p_t(\mathbf{x}^i|y)} \right) \end{aligned}$$

Directly extending the derivations in Poole et al. [7], we have our score distillation function that is jointly conducted

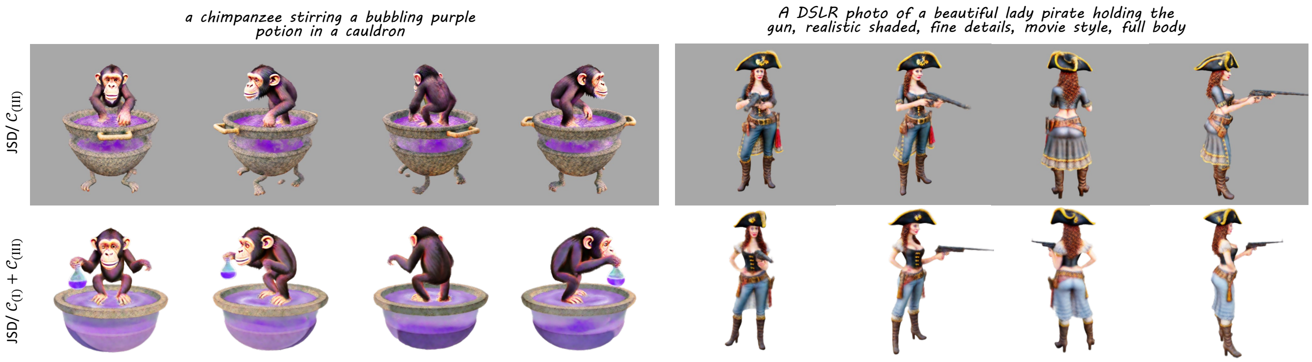

Figure 3. **Comparisons on energy function combination**. The combination of two energy functions further improves the geometry structure, demonstrating that JSD can effectively use the view-aware knowledge from diverse multi-view models.

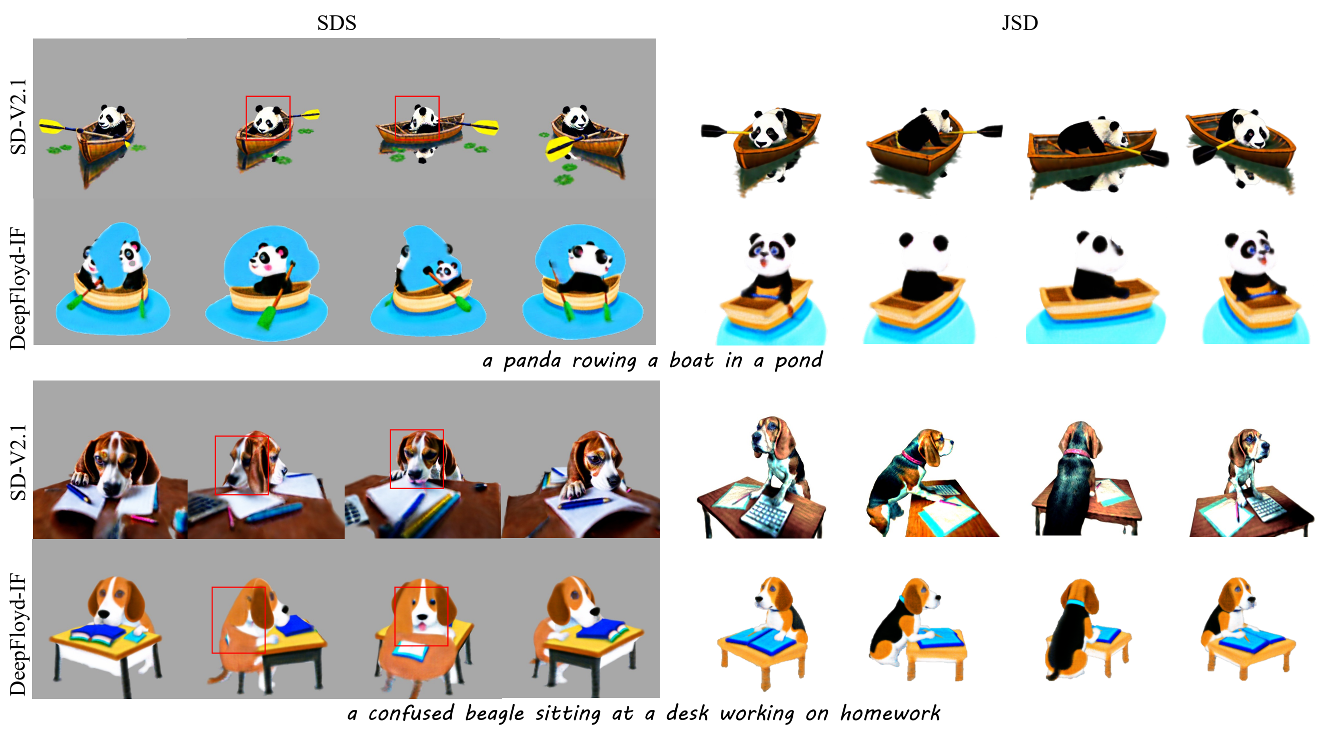

Figure 4. **Comparisons on 2D diffusion models,** including Stable-Diffusion-V2.1 (SD-V2.1) and DeepFloyd-IF. Different diffusion models have distinct impacts on the texture and geometry of generations, but both suffer the Janus issues. JSD incorporated with the binary classification model can consistently enhance the geometric consistency for both diffusion models.

on multiple views as follows:

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

where $\{\epsilon^i\}_{i=1}^V$ are noises during score matching for different views.

3. Additional Ablation Study

3.1. Discussions on Training Loss

To make further comparisons with JSD and SDS, we conduct training on two optimization functions with the text prompt "A DSLR photo of a squirrel playing guitar" and visualize the training loss curve as illustrated in Fig. 2. We observe that the training loss of SDS demonstrates serious fluctuation, which results from the randomness introduced by single-view optimization. By contrast, JSD can converge

CVPR #9474

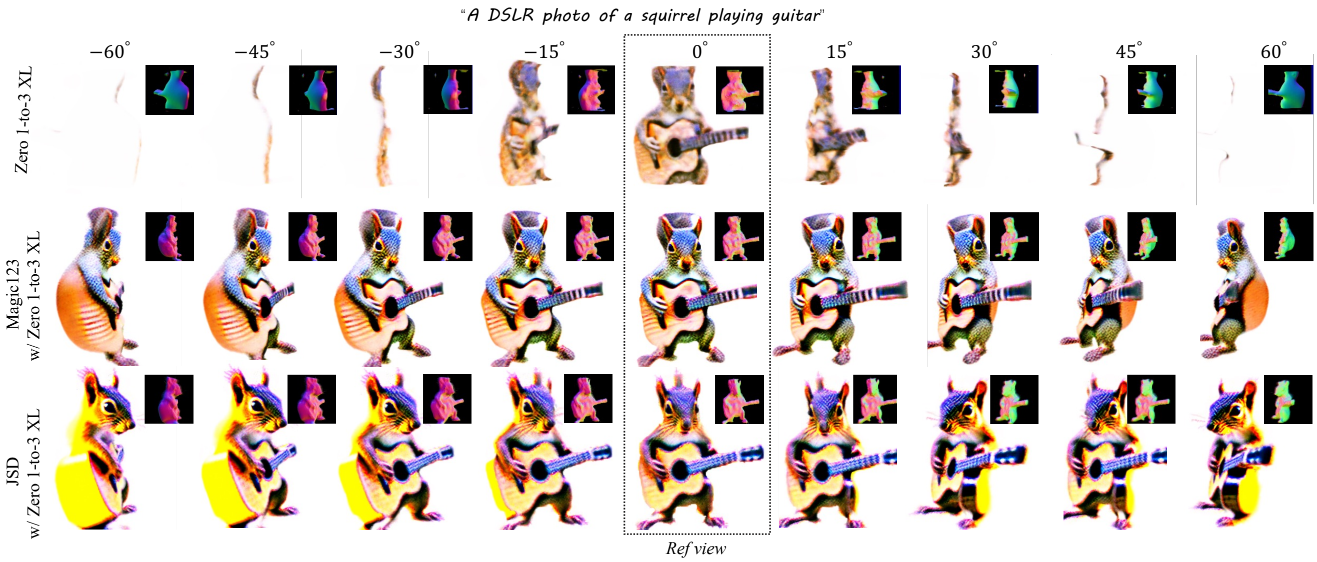

Figure 5. **Comparison with Image-to-3D methods.** Compared with two alternative methods, all employing the Zero-1-to-3 XL model, our proposed JSD exhibits superior generative quality in novel view synthesis as evidenced by its geometric consistency.

gradually and smoothly, which indicates that 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.

3.2. Discussions on Energy Function Combination

As discussed in our main paper, our proposed JSD can incorporate universal view-aware models as energy functions. Since the universal models are trained with different multi-view tasks, their inter-view coherence measurements are distinct, resulting in different 3D generations when incorporated with JSD. We have presented three representative view-aware models (Sec. 4.2 in the main paper) and demonstrated their different impacts on generations (Sec. 5.2 in the main paper). For computational efficiency, we adopt only a multi-view generation model as an energy function as JSD w/$\mathcal{C}_{(\mathrm{III})}$ for the final result of JointDreamer in our main text. To combine the complementary view-aware knowledge from different models, we incorporate JSD with the combination of the binary classification model and multi-view generation model as JSD w/$\mathcal{C}_{(\mathrm{I})} + \mathcal{C}_{(\mathrm{III})}$. As demonstrated in Fig. 3, the combination of two energy functions further improve the geometry structure, where the weird feet of the cauldron are eliminated. Since the classification model is a discrimination model, the texture quality remains similar. The comparison demonstrates that JSD can effectively take advantage of the view-aware knowledge from diverse multi-view models. Thus it can consistently enhance the benchmark of text-to-3D generation with the advancement of multi-view tasks and the combination of different multi-view models.

3.3. Discussions on Diffusion Models

Earlier works [4, 7] typically apply Stable Diffusion V1.5 (SD-V1.5) or Stable Diffusion V2.1 (SD-V2.1) as the 2D diffusion model in the SDS pipeline. However, more recent works [3, 9] have popularized the utilization of Deepfloyd-IF [10] . To align with recent works, we adopt Deepfloyd-IF for the baselines and JSD w/$\mathcal{C}_{(\mathrm{I})}$ and JSD w/$\mathcal{C}_{(\mathrm{II})}$. While MVDream fine-tunes on SD-V2.1, we retain SD-V2.1 as diffusion model in JSD w/$\mathcal{C}_{(\mathrm{III})}$. Notably, we observe that Deepfloyd-IF and SD-V2.1 have different impacts on 3D generations, as shown in the results of Fig. 4. SD-V2.1 leads to a high-fidelity and more detailed texture than Deepfloyd-IF, while Deepfloyd-IF contributes to better geometric structure in 3D generations as discussed in recent work [3] and open-source community [2]. Nevertheless, both SD-V2.1 and Deepfloyd-IF suffer from Janus issues in the SDS pipeline, as highlighted in the red box in Fig. 4. By substituting JSD for SDS and maintaining identical settings, including the resolution and time annealing strategy, we significantly enhance the 3D consistency of generations. We implement JSD w/$\mathcal{C}_{(\mathrm{I})}$ in Fig. 4, where the binary classification model can reduce the impact on texture quality to enable a more equitable comparison between Deepfloyd-IF and SD-V2.1. The results further demonstrate the compatibility of JSD to incorporate with various diffusion models to boost 3D consistency.

3.4. Discussions on Image-to-3D Methods

Since the view-aware models can engage in 3D generation through SDS besides JSD, we make comparisons to show-

CVPR #9474

case 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. 5, we input a reference image, exemplified by the front-view rendered image of the case of "A DSLR photo of a squirrel playing guitar" in Fig. 6 and compare with two alternative utilizations of Zero-1-to-3 XL. *(i)Zero-1-to-3 XL [5]*, 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 [8]*, 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.

4. Additional Results of JointDreamer

We present more comparisons of text-to-3D generation as shown in Fig. 6, 7 and 8. 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. 9, demonstrating the generalizability of JointDreamer with arbitrary textual descriptions.

5. Janus Prompts.

Our list of 20 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 plush dragon toy",

"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 squirrel playing guitar",

"a DSLR photo of a cat lying on its side batting at a ball of yarn"

"A crocodile playing a drum set"

"A pig wearing a back pack"

"A ceramic lion",

"a rabbit cutting grass with a lawnmower",

"Corgi riding a rocket",

"A bulldog wearing a black pirate hat",

"a zoomed out DSLR photo of a bear playing electric bass",

"A bald eagle carved out of wood, more detail",

"a lemur drinking boba".

References

[1] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 2

[2] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 4

[3] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 4

[4] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022. 4

[5] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 5

[6] Jiajun Ma, Tianyang Hu, Wenjia Wang, and Jiacheng Sun. Elucidating the design space of classifier-guided diffusion generation. *arXiv preprint arXiv:2310.11311*, 2023. 2

[7] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 4

[8] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123:

One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 5

[9] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 4

[10] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd. https://huggingface.co/DeepFloyd, 2023. 2, 4

Figure 6. **More comparison of text-to-3D generation.**

CVPR
#9474

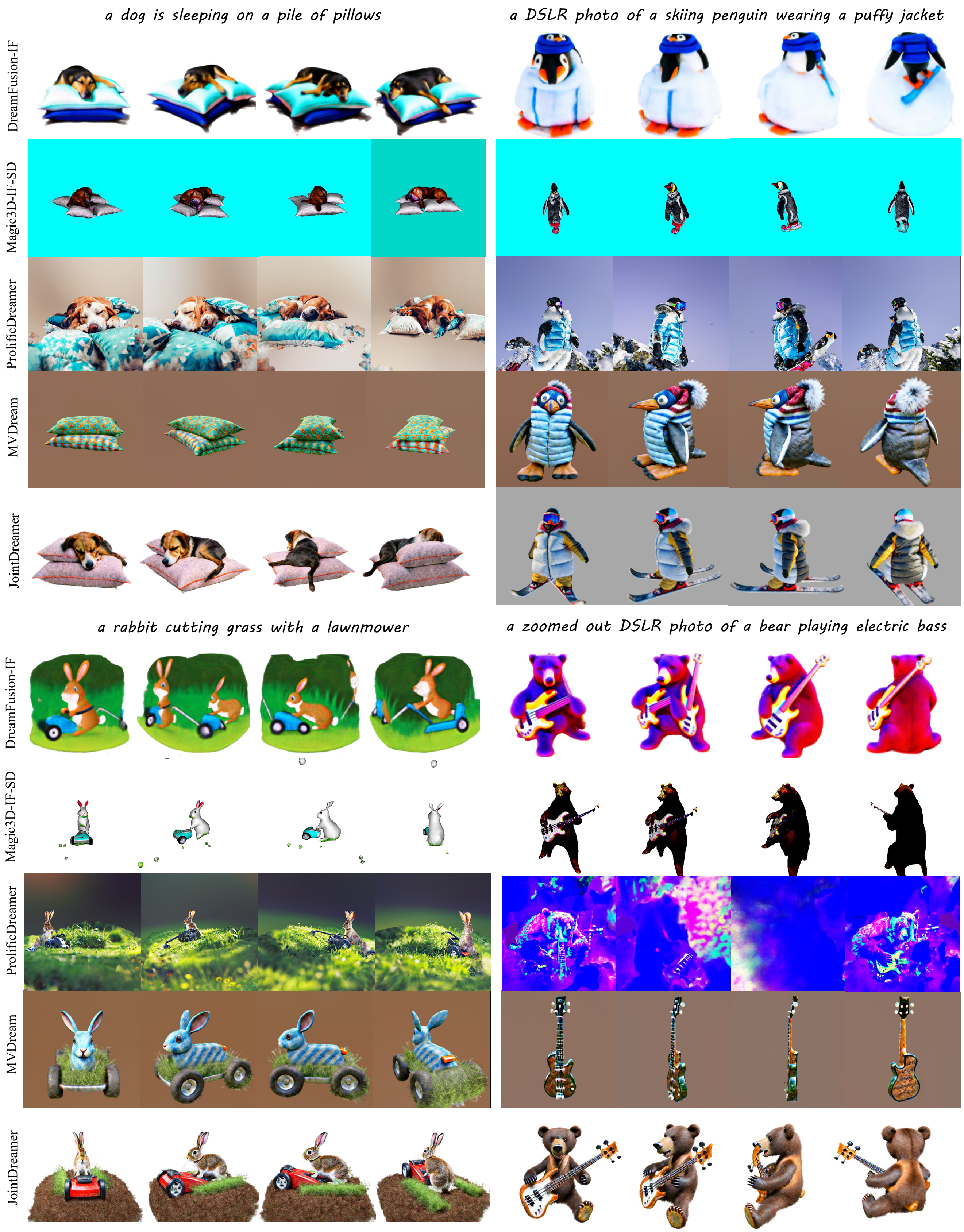

Figure 7. **More comparison of text-to-3D generation.**

CVPR
#9474

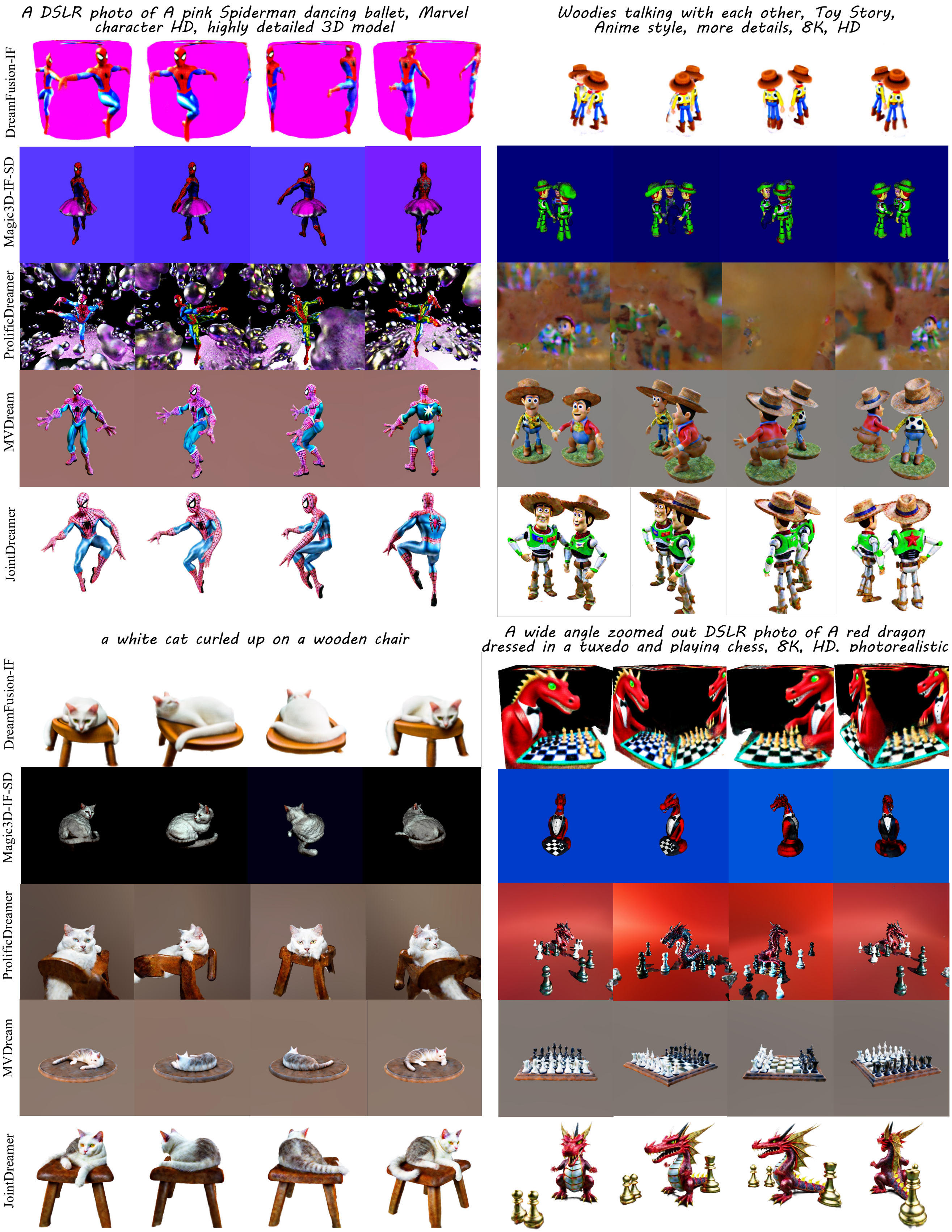

Figure 8. **More comparison of text-to-3D generation.**

CVPR
#9474

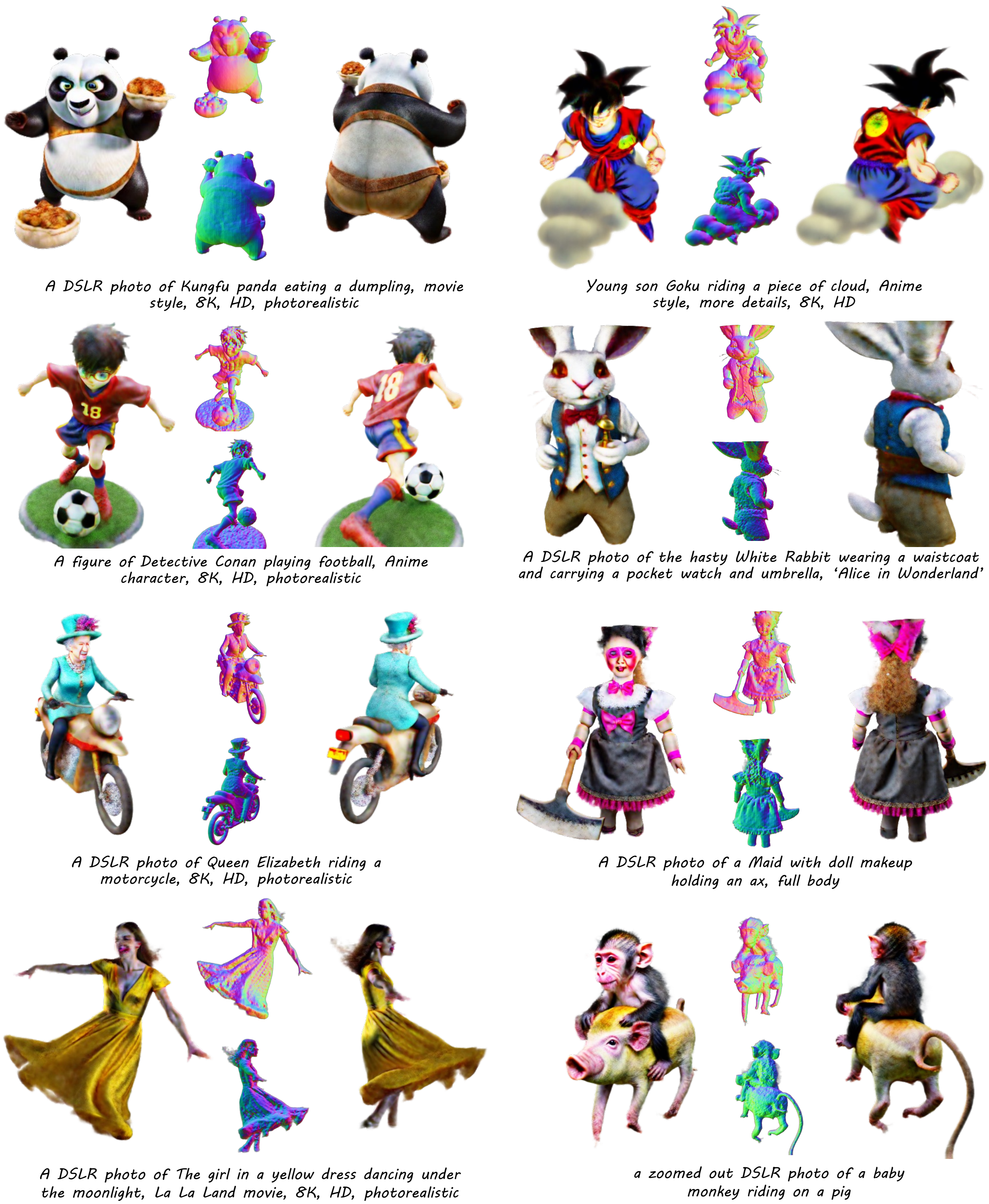

Figure 9. **More results of JointDreamer.**