Supplementary Material for Diffusion Time-step Curriculum for One Image to 3D Generation

Xuanyu Yi[1], **Zike Wu**[1], **Qingshan Xu**[1], **Pan Zhou**[3*],
Joo-Hwee Lim[4], **Hanwang Zhang**[2,1]

[1]Nanyang Technological University, [2]Skywork AI

[3]Singapore Management University, [4]Institute for Infocomm Research

xuanyu001@e.ntu.edu.sg, zike001@e.ntu.edu.sg, qingshanxu@hust.edu.cn,
panzhou@smu.edu.sg, joohwee@i2r.a-star.edu.sg, hanwangzhang@ntu.edu.sg

The *Appendix* is organized as follows:

- **Section A:** gives a theoretical justification of the proposed *Diffusion Time-step Curriculum*.
- **Section B:** further provides the experimental justification and discussion on the collaboration of the teacher and student with *Diffusion Time-step Curriculum*.
- **Section C:** elaborates more details about our DTC123 pipeline. Specifically, we detailed the implementation of instruct-LLM design and geometry smoothness regularization.
- **Section D:** showcases more immerse experiment results and ablation studies with Level50 benchmark.

A. Theoretical Justification

This section elaborates on our *diffusion time-step curriculum* motivation, where larger time steps capture coarse-grained concepts and smaller time steps learn fine-grained details. We first show that a diffusion time-step curriculum is necessary which further induces an annealed sampling strategy. Upon that, we explain why teacher and student models should collaborate with each other to achieve such a time-step curriculum.

Specifically, SDS employs the de-noised $\hat{\mathbf{x}}_0$ generated by the teacher diffusion model $\boldsymbol{\epsilon}_{\boldsymbol{\phi}}(\mathbf{x}_t; \boldsymbol{y}, t)$ to guide the student-rendered $\mathbf{x}_\pi$. Thus, the crux of this teacher-student optimization process is determined by the quality of the teacher guidance $\hat{\mathbf{x}}_0$, which motivates us to explore an appropriate strategy to ensure the valid guidance $\hat{\mathbf{x}}_0$ during any training iterations. We first formalize the definition of our target from the perspective of student data corruption.

Definition 1. (Data Corruption Reduction) *Given the camera pose π and the condition y, we consider a student-rendered image $\mathbf{x}_\pi = g(\theta, \pi)$ at an arbitrary training iteration k. Suppose that there exists a real data point $\mathbf{x}^*$ drawn from the data distribution $p_{data}(\mathbf{x})$ and an unknown data corruption $\delta_k = D_{\mathrm{KL}}[\delta(\mathbf{x}-\mathbf{x}_\pi)||p_{data}(\mathbf{x})]$, such that we can express $\mathbf{x}_\pi$ as $\mathbf{x}_\pi = \mathbf{x}^* + \delta_k$. Our **objective** is to iteratively reduce the corruption δ_k inherent in $\mathbf{x}_\pi$ as $k \to \infty$, guided by the conditional score function $\nabla \log p_t(\mathbf{x})$, such that $\mathbf{x}_\pi$ increasingly resembles a sample from $p_{data}(\mathbf{x})$.*

To estimate the real sample $\mathbf{x}^*$ in $\mathbf{x}_\pi$ for good guidance, one can resort to the score function $\nabla \log p_t(\mathbf{x})$, which, however, is practically inaccessible. To solve this issue, a pre-trained diffusion model $\boldsymbol{\epsilon}_\phi(\mathbf{x}, t)$ is often used to estimate the score function as shown in previous works [4, 5, 20]. But to better denoise and thus produce quality $\hat{x_0}$, the teacher model $\boldsymbol{\epsilon}_\phi(\mathbf{x}, t)$ needs a certain time step t to inject noise $\sigma_t\boldsymbol{\epsilon}$ into $\mathbf{x}_\pi$. In this way, the noisy sample $\mathbf{x}_t = \mathbf{x}_\pi + \sigma_t\boldsymbol{\epsilon}$ can approximately lie in the forward diffusion distribution $p_t(\mathbf{x}_t) = \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x}, \sigma_t^2\mathbf{I}) p_{\text{data}}(\mathbf{x})\mathrm{d}\mathbf{x}$ in the diffusion model $\boldsymbol{\epsilon}_\phi(\mathbf{x}, t)$, *i.e.*, the marginal distribution at time-step t in the forward diffusion process which satisfies $p_0(\mathbf{x}_0) = p_{\text{data}}(\mathbf{x})$. For this point, we provide a formal analysis, and derive a *proper* time-step sampling for better diffusion de-noising.

Theorem 1. (Diffusion Time-step Lower bound) *Assume $p_t(\mathbf{x})$ is the noisy data distribution and $q_t(\mathbf{x}_t|\mathbf{x}_\pi) = \mathcal{N}(\mathbf{x}_t; \alpha_t\mathbf{x}_\pi, \sigma_t^2\mathbf{I})$, for any $\mathbf{x}_t \sim q_t(\mathbf{x}_t|\mathbf{x}_\pi)$, we have $\|\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t) - \nabla \log p_t(\mathbf{x}_t)\|_2^2 = \mathcal{O}(\varepsilon)$, if these two conditions hold :*
a) the pretrained teacher diffusion model $\boldsymbol{\epsilon}_\phi(\mathbf{x}, t)$ satisfies $\|\boldsymbol{\epsilon}_\phi(\mathbf{x}, t) - \nabla \log p_t(\mathbf{x})\|_2^2 < \varepsilon$;
b) $t \sim \mathcal{U}[\tilde{T}_{\delta_k,\varepsilon}, T]$ where $\tilde{T}_{\delta_k,\varepsilon} = \mathcal{O}(\frac{\|\delta_k\|}{\epsilon})$.

Proof. We assume a) always hold since $\boldsymbol{\epsilon}_\phi(\cdot)$ is a well trained diffusion model. Assume that the diffusion model $\boldsymbol{\epsilon}_\phi(\cdot)$ satisfies the Lipschitz condition with a constant $L > 0$. Specifically, we have:

$$\|\boldsymbol{\epsilon}_\phi(\mathbf{x}, t) - \boldsymbol{\epsilon}_\phi(\mathbf{x}', t)\| \leq L\|\mathbf{x} - \mathbf{x}'\|. \tag{1}$$

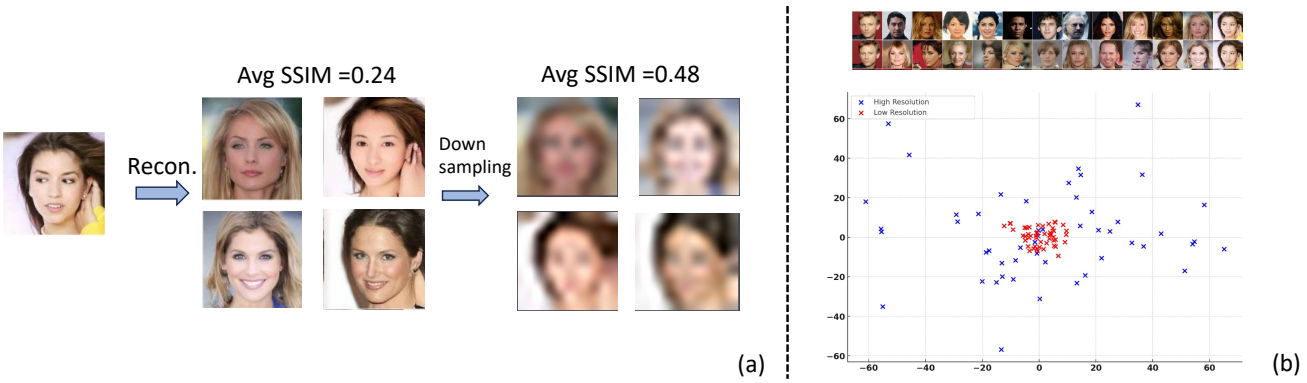

Figure 1. **(a)** The low-resolution generated set embraces similarity compared to its high-resolution counterparts. **(b)** The T-SNE visualization of high/low resolution generated sets on CelebA dataset.

From condition a) we have for any t:

$$\|\boldsymbol{\epsilon}_\phi(\mathbf{x}, t) - \nabla \log p_t(\mathbf{x})\|_2^2 < \varepsilon. \tag{2}$$

Recall that $\mathbf{x}_\pi = \mathbf{x}^* + \delta_k$ where $\mathbf{x}^* \sim p_{\text{data}}(\mathbf{x})$, we have:

$$\begin{aligned} &\|\boldsymbol{\epsilon}_\phi(\alpha_t \mathbf{x}_\pi + \sigma_t \epsilon, t) - \nabla \log p_t(\alpha_t \mathbf{x}^* + \sigma_t \epsilon)\| \\ \leq &\|\boldsymbol{\epsilon}_\phi(\alpha_t \mathbf{x}_\pi + \sigma_t \epsilon, t) - \boldsymbol{\epsilon}_\phi(\alpha_t \mathbf{x}^* + \sigma_t \epsilon, t)\| \\ + &\|\boldsymbol{\epsilon}_\phi(\alpha_t \mathbf{x}^* + \sigma_t \epsilon, t) - \nabla \log p_t(\alpha_t \mathbf{x}^* + \sigma_t \epsilon)\| \\ \leq &L \alpha_t \|\delta_k\| + \epsilon \end{aligned} \tag{3}$$

To achieve the desired accuracy, we have to let t sufficient large such that $L\alpha_t\|\delta_k\| = \mathcal{O}(\epsilon)$, which means $\alpha_t = \mathcal{O}(\frac{\epsilon}{L\|\delta_k\|})$. Recall that in conventional diffusion models [5, 7, 17], we have $\alpha_t \propto \frac{1}{t}$, thus we derive

$$t \geq \mathcal{O}(\frac{L\|\delta_k\|}{\epsilon}) = \mathcal{O}(\frac{\|\delta_k\|}{\epsilon}). \tag{4}$$

□

Theorem 1 shows that the teacher diffusion model can accurately estimate the desired score function $\nabla \log q_t(\mathbf{x}_t|\mathbf{x}_\pi)$ under the condition a) and b). For condition a), it often holds, since the pretrained teacher diffusion model $\boldsymbol{\epsilon}_\phi(\mathbf{x}, t)$ can (approximately) converge to the forward diffusion distribution $p_t(\mathbf{x})$. Thus, if one can sample a proper time step t such that $\mathbf{x}_t$ satisfies condition b), then the teacher diffusion model $\boldsymbol{\epsilon}_\phi(\mathbf{x}, t)$ can well denoise $\mathbf{x}_t$ and provides quality guidance $\hat{\mathbf{x}}_0$ to supervise the student 3D model. Since in the early training iterations, the student-rendered $\mathbf{x}_\pi$ contains high corruption δ_k and $\tilde{T}_{\delta_k,\varepsilon}$ positively depends on the corruption level as shown in Theorem 1, a large time step $t \geq \tilde{T}_{\delta_k,\varepsilon}$ is needed to inject more noise $\sigma_t\epsilon$ into $\mathbf{x}_\pi$ so that condition b) holds and thus guarantees the quality of the denoising.

However, there is a dilemma that the marked divergence of the teacher-generated $\hat{\mathbf{x}}_0$ [5] at large time-steps could negatively compromise the student coherent modeling that possesses consistent geometric and photometric properties, resulting in geometry distortion and mode collapse [19, 21]. From the perspective of information theory [1], given a set of teacher-generated $\hat{\mathbf{x}}_0$ in certain training iterations, coarse-grained information (*e.g.*, blur contour) tends to have less variance than its fine-grained counterpart (*e.g.*, texture nuances), which is also empirically justified in Part B. This motivates us to first focus on fundamental, low-variance student-teacher knowledge transfer with large time-steps. As the coarse-grain converges along with the training iteration k, the corruptions δ_k in the student-rendered $\mathbf{x}_\pi$ diminish. From Theorem 1, smaller noise can counteract corruption δ_k and help $\mathbf{x}_t$ conform to distribution $p_t(\mathbf{x}_t)$. Thus, the teacher diffusion model often only refines $\mathbf{x}_t$ to improve the fine-grains without destroying the course-grains. Accordingly, we can gradually derive a more accurate estimation of the score function $\nabla \log p_t(\mathbf{x})$, thus ensuring the refinement of intricate details (*e.g.*, texture nuances) at smaller time-steps.

B. Experimental Justification and Discussion

This section elaborates on the experimental justification for the collaboration of the teacher and student model with the *diffusion time-step curriculum*. We first intuitively indicate that the low-resolution representation has lower variance and is more robust compared to its high-resolution counterpart. Then we quantitatively analyze the view-conditioned and text-conditioned teacher diffusion model by comparing their reconstructed results with the multi-view ground-truth rendering images among different levels of perturbed noise.

Student progressive representation. We first generated a

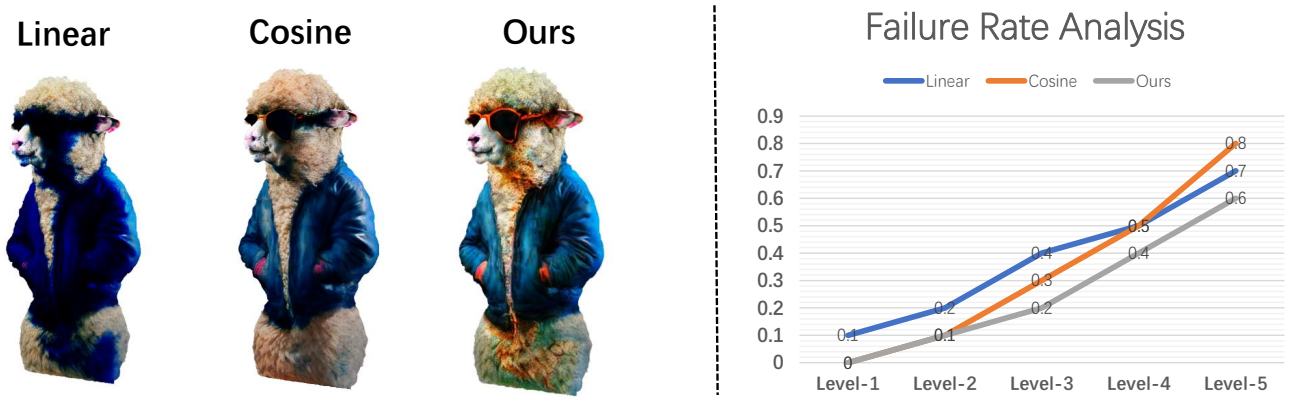

Figure 2. Ablations on different time-step sampling strategy.

set of diverse teacher guidance $\hat{\mathbf{x}}_0$, following [5] to use the CelebA [12] images by perturbing them with large noise (1000 diffusion time-steps) and reconstructing them during the reverse process. We then leveraged image down-sampling to simulate low-resolution modeling. As depicted in Figure 1(a), the down-sampled generated set is more similar to each other than its high-resolution counterpart: the high-resolution images scored a structural similarity index (SSIM) of *0.24*, contrasting sharply with the low-resolution set's *0.48*. This significant difference highlights the reduced variance and increased similarity among down-sampled images, which is more suitable for student coherent modeling in the initial iteration. Figure 1(b) further shows the T-SNE visualization of the high/low resolution generated set, which vividly demonstrates the low-variance and clustering effect of low-resolution representation. This toy experiment indicates that though the teacher diffusion model provides diverse guidance with a large time step [5], we can ensure coherent modeling of 3D models by coarse-to-fine representation (resolution constraint) with the annealed time-step.

On the other hand, from the perspective of parameter sensitivity of multi-resolution modeling [14], rays intersecting the scene at coordinates x and the hash function defined as $h(x) = \bigoplus_{i=1}^{d} x_i \pi_i \mod T$, even minor deviations in x lead to notable variations in encoded values. Denoting δx as the noise-induced deviation, the change in hash values, Δh, escalates with higher resolutions, i.e., $\frac{\partial \Delta h}{\partial \delta x} \propto$ resolution, indicating that higher resolutions inherently magnify the sensitivity to noise and variance in the ground truth, which also inspires us to first capture the low-resolution representation for more stable 3D model optimization.

Teacher coarse-to-fine prior. Here, we conducted a quantitative experiment to answer the question about the suitable teacher for diffusion time-step curriculum. Given a 3D object from the *high-quality real-scanned* dataset OmniObject3D [22], we used Cycles Engine in Blender to randomly (with fixed elevation and different azimuth) render 16 multi-view images and transparent backgrounds with pure grey color. We then perturb them with different levels of noise and then compare [1] the quality of the reconstructed image with the ground-truth renderings by investigating the contour exploration consistency via MaskIoU and the perceptual generation quality computed by CLIP-similarity.

As illustrated in Table 1, (1) At large time-steps, Zero-1-to-3 generated outputs tend to have better MaskIoU than Stable Diffusion outputs, which suggests that Zero-1-to-3 serves as a coarse-grained teacher by providing a more accurate contour or boundary at large t; (2) At smaller time-steps, both Zero-1-to-3 and Stable Diffusion have relatively high MaskIoU since the small scale perturbed noise doesn't corrupt the overall geometry structure. Considering the quality of perceptual generation (CLIP similarity), we notice that Stable Diffusion surpasses Zero-1-to-3 to some extent, indicating that Stable Diffusion is more suitable for a fine-grained teacher, since it produces more realistic texture details at smaller t.

C. Background

C.1. Student 3D Model

We aim to learn an underlying 3D representation θ (*e.g.*, NeRF, mesh), which uses a differentiable renderer $g(\cdot)$ to generate the relative image from any desired camera pose π by $\mathbf{x}_\pi = g(\theta, \pi)$. For computation memory concerns, we leverage NeRF [13] for low-resolution scene modeling, and then adopt DMTet [16] for high-resolution mesh fine-tuning.

(NeRF) [13] is a differentiable volumetric representation. It characterizes the scene as a volumetric field by density and color with a neural network θ. Given a camera pose π,

[1] Note that the condition of Zero-1-to-3 is the default front view image and the camera parameters, while the condition of Stable Diffusion is the caption of the 3D object.

Quantitative analysis among different time-steps								
Time-step	**200**		**400**		**600**		**800**	
Metrics	**MaskIoU**	**CLIP-S**	**MaskIoU**	**CLIP-S**	**MaskIoU**	**CLIP-S**	**MaskIoU**	**CLIP-S**
Zero-1-to-3	0.92	0.84	0.87	0.82	0.84	0.79	0.82	0.80
Stable Diffusion	0.89	0.90	0.81	0.84	0.72	0.82	0.63	0.74

Table 1. Quantitative compassion of Zero-1-to-3 and Stable Diffusion among different time-steps, where CLIP-S denotes the CLIP similarity between de-noising output and the ground truth renderings.

the rendered image can be computed by alpha compositing the color density field. Considering rendering efficiency, multi-resolution hash grids [14] are usually utilized to parameterize the scene. This representation helps to achieve high-quality rendering results with a faster training speed.

Hybrid SDF-Mesh Field (DMTET) is a differentiable surface representation. It parameterizes the Signed Distance Function (SDF) by a deformable tetrahedral grid (V_T, T), where T represents the tetrahedral grid and V_T corresponds to its vertices. By assigning every vertex $v_i \in V_T$ with a SDF value $s_i \in R$ and a deformation vector $\Delta v_i \in R^3$, this representation allows recovering a explicit mesh through differentiable marching tetrahedra.

C.2. Teacher Diffusion Model

Conditional Diffusion Model (CDM) [6,8] basically generates the desired samples given a certain condition y. In the context of SDS-based 3D generation, we mainly utilize the following two types of teacher diffusion prior with different conditions:

Text-conditioned Prior. Large-scale pre-trained text-to-image diffusion models, *e.g.*, Stable Diffusion [15], are often leveraged with the text description condition of the reference image. In practice, one often employs textual inversion [3, 23] or large vision-language models *e.g.*, BLIP2 [10] to generate the text description of the reference image.

View-conditioned Prior. Zero-1-to-3 [11] is fine-tuned from the Stable Diffusion image variations [9] on the 3D synthetic dataset Objaverse [2], and integrates viewpoint control to conduct novel view synthesis given the camera pose and reference image as condition.

D. Implementation Details

This section elaborates on the details of implementation details of instruct-LLM design and geometry smoothness regularization.

Instruct-LLM Design. In the second stage, we noted that employing finer, more precise linguistic prompts to significantly narrow the image distribution in Stable Diffusion, complementing the mode-seeking SDS algorithm. Utilizing large language models (LLMs), we converted BLIP2-derived prompts into comprehensive, view-specific descriptions. This approach intensifies detail in each orthogonal view, avoiding superfluous structural descriptions and resolving perspective conflicts. The detailed instruction is as follows:

> "I am about to begin a series of image-3D generation tasks and need your help to create prompt descriptions. I'll provide the frontal description first. You will then give a one-sentence description for each the left, right, and rear sides, ensuring: (1) Alignment with the frontal description, (2) Conciseness with rich textural detail, (3) 3D consistency across descriptions, using DALLE-3 for validation."

Figure 3 illustrates the prompts created using this method, demonstrating our technique's effectiveness in generating consistent, detailed multi-view descriptions.

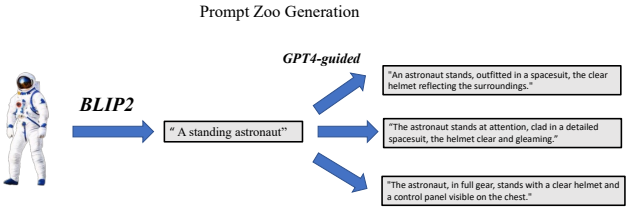

Figure 3. Examples of viewpoint-augmentation with LLM. Given an abstract prompt, our instruct-LLM outputs complement the initial prompts with insufficient information.

Geometry smoothness regularization. We observe that the 3D model occasionally generated high-frequency artifacts on the crisped surface and edge contour. Following [18], we leverage the normal vector regularization $\mathcal{L}_{\text{reg}}$:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\mathbf{a}} \left[\|\mathbf{n}(\mathbf{a}) - \mathbf{n}\left(\mathbf{a} + \beta \cdot \mathcal{N}(0, I)\right) \|_1 \right] \quad (5)$$

where $\mathbf{n}$ denotes the normal vector at a point $\mathbf{a}$ in the 3D space, β is a small perturbation scale and $\mathcal{N}(0, I)$ is standard Gaussian noise.

In the second stage of DMTet, we implement Laplacian smoothing regularization where the Laplacian matrix

$\mathbf{L} \in \mathbb{R}^{V \times V}$ is computed by identifying adjacent vertex pairs from each face in $\mathbf{F}$. Entries in $\mathbf{L}$ are set to -1 for adjacent vertices and to the degree of the vertex on the diagonal. The smoothing loss $\mathcal{L}_{\text{reg}}$ is defined as the mean norm of the product of the Laplacian matrix $\mathbf{L}$ and vertex positions $\mathbf{V}$:

$$\mathcal{L}_{\text{reg}} = \text{mean}\left(\|\mathbf{L} \cdot \mathbf{V}\|_2\right), \tag{6}$$

which ensures the uniformity and smoothness of the mesh by minimizing deviations in the vertex positions, leading to a more regular and smooth 3D structure.

E. More Experimental Results

This section presents more qualitative results of DTC123 and ablation studies on the time-step sampling strategy.

E.1. Ablation on Time-step Sampling Strategy

To better analyze the DTC123 premium sampling strategy, we still adopted the failure rate metric in the manuscript on Level50. As illustrated in Figure 2, our proposed sampling strategy consistently exhibits a lower failure rate compared to other methods at different difficulty levels. We justify that the robustness should be attribute to the introduced local randomness with the annealed interval, since (1) even within the same training iteration, the corruption level of 3D models varies across camera poses, and (2) in contrast to [20], it is nearly impossible to pinpoint the exact corruption level without the ground-truth of unseen view, we need some randomization for self-calibration of the teacher-student symbiotic cycle, like SDE sampling in conventional DMs [7], which can partially correct the cumulative error introduced by the optimization process and alleviate 'floaters' effects.

E.2. More Qualitative Results

In Figure 4 and Figure 5, we present additional qualitative outputs with high fidenity and multi-view consistency. Please check out the video demos for more results.

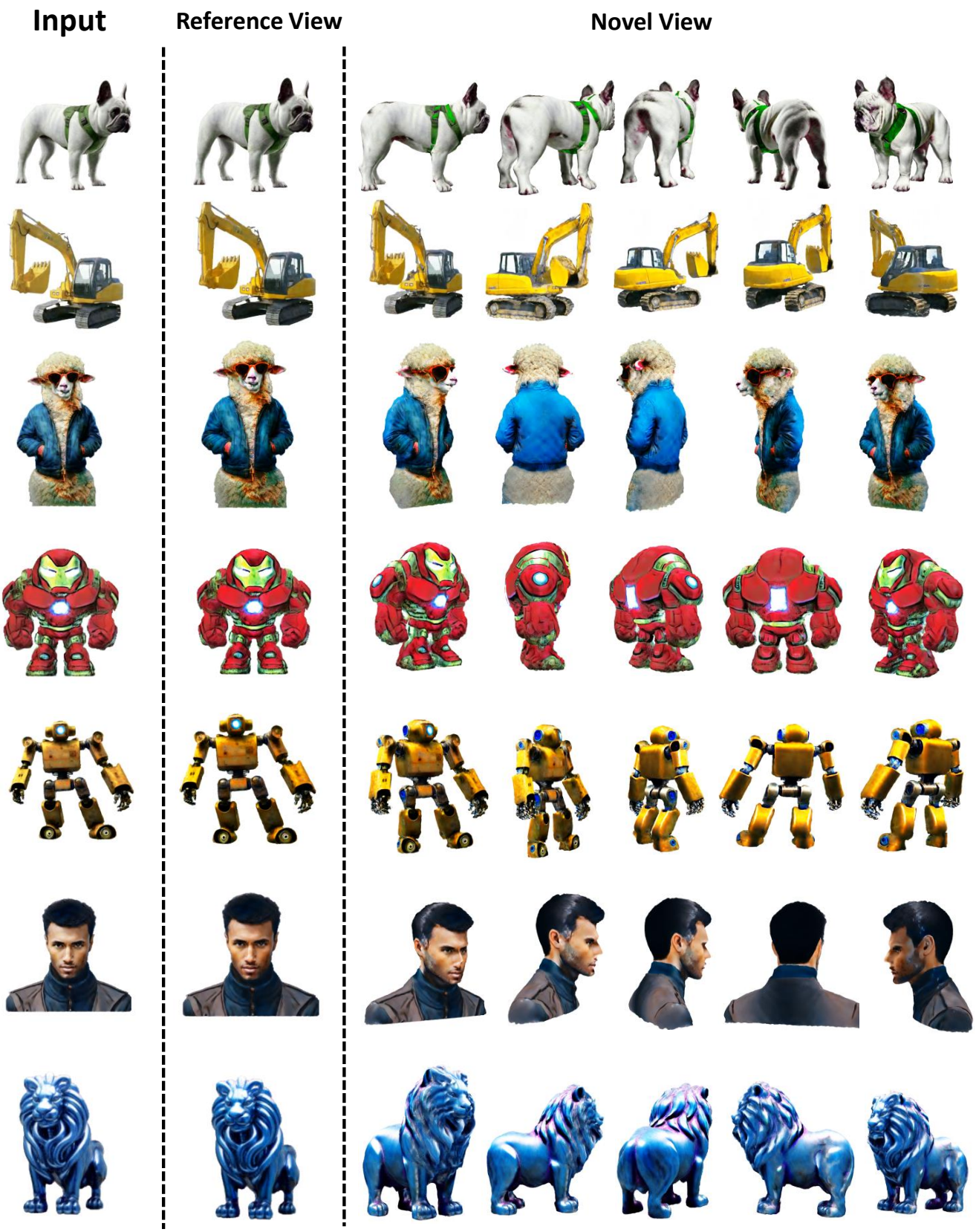

Figure 4. More DTC123-generated results. Our approach yields results with improved fidelity and more robust geometry.

Figure 5. More DTC123-generated results. Our approach yields results with enhanced fidelity and more robust geometry.

References

[1] Robert B Ash. *Information theory*. Courier Corporation, 2012. 2

[2] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 4

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 4

[4] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[7] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 2, 5

[8] Bahjat Kawar, Roy Ganz, and Michael Elad. Enhancing diffusion-based image synthesis with robust classifier guidance. *arXiv preprint arXiv:2208.08664*, 2022. 4

[9] Lambda Labs. Stable diffusion image variations, 2023. Accessed: 2023-11-17. 4

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4

[11] 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. 4

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 3

[13] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[14] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3, 4

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

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

[17] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[18] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion. 4

[19] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[20] Luozhou Wang, Shuai Yang, Shu Liu, and Ying-cong Chen. Not all steps are created equal: Selective diffusion distillation for image manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7472–7481, 2023. 1, 5

[21] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[22] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 3

[23] Jianan Yang, Haobo Wang, Ruixuan Xiao, Sai Wu, Gang Chen, and Junbo Zhao. Controllable textual inversion for personalized text-to-image generation. *arXiv preprint arXiv:2304.05265*, 2023. 4