Sherpa3D: Boosting High-Fidelity Text-to-3D Generation via Coarse 3D Prior

Supplementary Material

1. More Discussion of Preliminaries

In this section, we provide more preliminaries and details of our implementation for Score Distillation Sampling (SDS).

1.1. Diffusion Models

The diffusion model, which is a type of likelihood-based generative model used to learn data distributions, has been studied extensively in recent years [6, 21–24]. Given an underlying data distribution $q_0(\boldsymbol{x})$, a diffusion model composes two processes: (a) a forward process $\{q_t\}_{t\in[0,1]}$ to gradually add noise to the data point $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$; (b) a reverse process $\{p_t\}_{t\in[0,1]}$ to denoise data (*e.g.*, generation). Specifically, the forward process is defined by $q_t\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) := \mathcal{N}\left(\alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\right)$ and $q_t\left(\boldsymbol{x}_t\right) := \int q_t\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) q_0\left(\boldsymbol{x}_0\right) \mathrm{d}\boldsymbol{x}_0$, where $\alpha_t, \sigma_t > 0$ are hyperparameters. On the other hand, the reverse process is described with the transition kernel $p_t(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) := \mathcal{N}(\mu_\phi(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I})$ from $p_1(\boldsymbol{x}_1) := \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The training objective is to optimize μ_ϕ by maximizing a variational lower bound of a log-likelihood. In practice, μ_ϕ is reparameterized as a denoising network $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t)$ [6] to predict the noise added to the clean data $\boldsymbol{x}_0$, which is trained by minimizing the MSE criterion [6, 9]:

$$\mathcal{L}_{\text{Diff}}\left(\phi\right) := \mathbb{E}_{\boldsymbol{x}_0, t, \boldsymbol{\epsilon}}\left[\omega(t)\left\|\boldsymbol{\epsilon}_\phi\left(\alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}\right) - \boldsymbol{\epsilon}\right\|_2^2\right], \quad (1)$$

where $\omega(t)$ is the time-dependent weights. Besides, the noise prediction network $\boldsymbol{\epsilon}_\phi$ can be applied for approximating the score function [23] of the perturbed data distribution $q(\boldsymbol{x}_t)$, which is defined as the gradient of the log-density:

$$\nabla_{\boldsymbol{x}_t} \log q_t\left(\boldsymbol{x}_t\right) \approx -\boldsymbol{\epsilon}_\phi\left(\boldsymbol{x}_t, t\right) / \sigma_t. \quad (2)$$

This means that the diffusion model can estimate a direction that guides $\boldsymbol{x}_t$ towards a high-density region of $q(\boldsymbol{x}_t)$, which is the key idea Score Distillation Sampling (SDS) [15, 25] for optimizing the 3D scene via well 2D pre-trained models.

1.2. SDS with Classifier-Free Guidance

As one of the most successful applications of diffusion models, text-to-image generation [16–18] generate samples $\boldsymbol{x}$ based on the text prompt y, which is also fed into the $\boldsymbol{\epsilon}_\phi$ as input, denoted as $\boldsymbol{\epsilon}_\phi\left(\boldsymbol{x}_t; t, y\right)$. An important technique to improve the performance of these models is Classifier-Free Guidance (CFG) [5]. CFG modifies the original model by adding a guidance term, *i.e.*, $\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t; y, t) := (1+s)\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; y, t) - s\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; t, \varnothing)$, where $s > 0$ is the guidance weight that controls the balance between fidelity and diversity, while $\varnothing$ denotes the "empty" text prompt for the unconditional case. Recall the SDS gradient form to update θ:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \boldsymbol{x}) = \mathbb{E}_{t,\epsilon}\left[\omega(t)\left(\boldsymbol{\epsilon}_\phi\left(\boldsymbol{x}_t; y, t\right) - \epsilon\right)\frac{\partial \boldsymbol{x}}{\partial \theta}\right], \quad (3)$$

and denote $\delta_{\boldsymbol{x}}(\boldsymbol{x}_t; y, t) := \epsilon_\phi(\boldsymbol{x}_t; y, t) - \boldsymbol{\epsilon}$. In principle, $\boldsymbol{\epsilon}(\boldsymbol{x}_t; y, t)$ should represent the pure text-conditioned score function in Eq. (3). But in practice, CFG is employed in it with a guidance weight s to achieve high-quality results, where we rewrite

$$\delta_{\boldsymbol{x}}(\boldsymbol{x}_t; y, t) = [\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; y, t) - \boldsymbol{\epsilon}] + s[\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; y, t) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; t, \varnothing)]. \quad (4)$$

As DreamFusion [15] uses $s = 100$ for high fidelity, our implementation adopts $s = 50$ with the enhancement of structural and semantic guidance to preserve some diversity. The two types of guidance can also be seen as another form of prompt guidance that is more generalizable and robust. Therefore, there is a gap between the original formulation in Eq. (3) and the practical coding implementation in Eq. (4).

2. More Discussion of Our Method

Discussion of Our Framework. Existing 3D generation methods are based on single diffusion models, such as a 3D diffusion model for 3D native methods [8, 14], a 2D diffusion model for 2D-lifting methods [3, 12, 15, 26], or a multi-view diffusion model for multi-view-based methods [13, 19, 20]. However, these methods suffers from limited diversity, Janus problems, or limited generation styles. Instead of using single diffusion models, we find that a hybrid text-to-3D pipeline with carefully-designed guidance has potential to simultaneously address these concerns and achieve both SOTA 3D fidelity and 2D texture richness only using basic diffusion models (e.g., SD-2-1-base [17]) instead of larger ones (e.g., SDXL [1]), also reducing the generation process from several hours to 20 minutes. While a simple combination yields undesired results, our focus is to design a unified framework by setting 3D diffusion as auxiliary and bridging them with our structural and semantic guidance and the balanced annealing strategy.

Discussion of the CLIP Input. In our framework, we utilize normal images instead of traditional RGB images as inputs for CLIP due to their capability to richly convey geometric details, including both local nuances and silhouette features of shapes. This choice significantly enhances geometry learning. Furthermore, the inclusion of normal images in the CLIP-filtered dataset LAION-5B suggests that these images are not deemed out-of-distribution (OOD) for

CLIP. Fig. 1 shows that CLIP extracts discriminative semantic features from normal images.

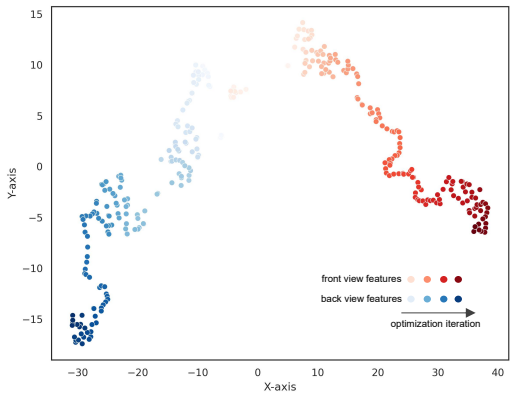


Figure 1. t-SNE visualization of CLIP features. Prompt: "A head of Terracotta Army."

3. Additional Implementation Details

Training details. Our geometry model $\mathcal{F}_\theta$ and appearance model $\mathcal{T}_\eta$ is approximated by three-layer MLPs and we apply adam [10] optimizer to update them with an initial learning rates of 1×10^{-3} to decaying to 5×10^{-4}. In particular, our method is optimized for 2500 iterations about 15 minutes to learn $\mathcal{F}_\theta$ and 2500 iterations about 10 minutes to learn $\mathcal{T}_\eta$. For geometry modeling, we utilize the Open3D library [27] to calculate the signed distance function (SDF) value for each point in Equations 2 and 3 in the main paper. In our experiments, the DMTet-based coarse 3D prior building stage is critical as it not only provides coarse 3D knowledge with consistency but also boosts the speed of the convergence of generation. For appearance modeling, since our focus in this paper is to fully exploit easily obtained coarse 3D knowledge that serves as guidance for 2D lifting optimization (as discussed in Section 3.3 of our paper), we do not design a specific appearance model for our framework. Note that our geometry model is plug and play and we can leverage different models [2, 3, 11], we leverage the same PBR materials approach in Fantasia3D [3] to achieve photorealistic surface renderings and better aligns with our geometry modeling.

Hyperparameter settings. We select the camera positions (r, κ, φ) in the spherical coordinate system, where r denote radius, κ is the elevation and φ is the azimuth angle respectively. Specifically, we sample random camera poses at a fixed $r = 2.5$ with the $\kappa \in [-30°, 30°]$. In a batch of $b \times l$ images, we partition φ into l intervals in $[-180°, 180°]$ and uniformly sample b azimuth angles in each interval. For structural guidance, we set $\sigma = 1$ in Eq. (4) in the main paper as the standard deviation of the Gaussian filter. We tune λ_{struc} and λ_{sem} in $\{0.01, 0.1, 1, 5, 10, 20, 30, 100\}$. We find that often $\lambda_{\text{struc}} = 10$ and $\lambda_{\text{sem}} = 30$ works well with $\beta = 0.5$ in the step annealing technique, which may balance the magnitude of SDS losses and better guide the 2D lifting to refine the 3D contents with multi-view coherence. We assigned the value of m to the epoch at around 1000 iterations. For the guidance weight $\omega(t)$, we follow the DreamTime [7] to achieve higher fidelity results. Our codes for implementation will be available upon acceptance.

4. Additional Experiments and Analysis

4.1. Additional User Study

To further demonstrate the effectiveness and impressive visualization results of our Sherpa3D, we conducted a more intuitive user study (Figure 2) on 20 text prompts of five baselines (ShapE [8], DreamFusion [15], Magic3D [12], ProlificDreamer [26], Fantasia3D [3]) and ours. The study engaged 50 volunteers to assess the generated results in 20 rounds. In each round, they were asked to select the 3D model they preferred the most, based on quality, creativity, alignment with text prompts, and consistency. We also compare our method with recent finetuning-based techniques, such as Zero123 [13] and MVDream [20], which utilize more 3D data [4] to retrain a costly 3D aware diffusion model from Stable Diffusion [17]. We use the same text prompts and settings as mentioned above. As shown, we

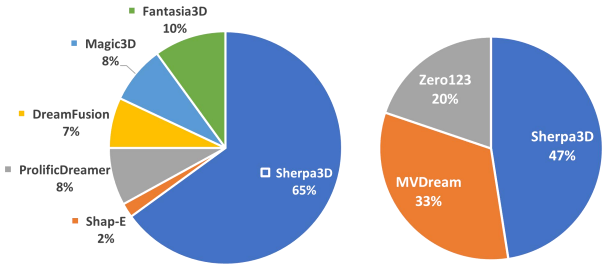


Figure 2. **User study** of the rate from volunteers' preference for each method in the inset pie chart.

observe that Sherpa3D is preferable (65%) by the raters on average. In other words, our model is preferred over the best of all baselines in most cases. What's more, our Sherpa3D also outperforms than fine-tuning based method in terms of overall performance as they easily suffer from styles (lightning, texture) overfitting [13, 20]. We believe this is strong proof of the robustness and quality of our proposed method.

4.2. More Qualitative and Quantitative Results

Sherpa3D. In Figure 7, 8, 9, we present more text-to-3D results obtained with Sherpa3D, which can generate high-fidelity, diverse, and 3D-consistent results within 25 minutes. Besides the impressive 3D consistency and high fidelity, we can also change the style of generated 3D content

(Figure 3) by only modifying a small part of the prompt, while preserving the basic structure of 3D content, which is more convenient for users to flexibly edit generated objects.

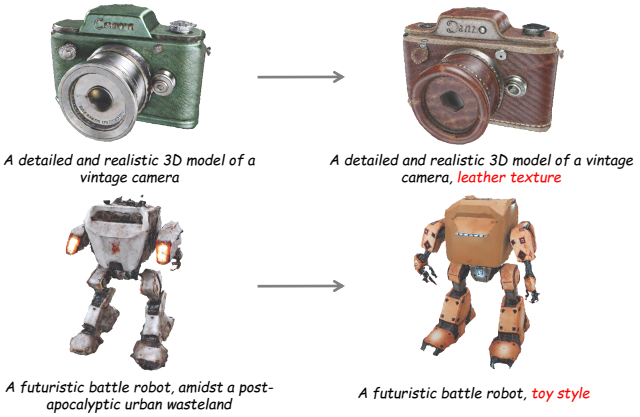


Figure 3. Sherpa3D can be used for flexible editing through a small part of the prompt modification.

Generalizability and diversity. The preservation of generalizability is particularly attractive, which leads to more diverse and realistic results. We achieve this by maximizing retention of the original 2D diffusion priors in text-to-3D pipeline, while incorporating the geometric information in a soft manner compared with re-training or fine-tuning with 3D data. Fig. 4 shows that our method obtains good results for complex prompts.

Figure 4. Generalizable results of our method.

Evaluation of adjacency-based metric. In the main paper, we employ the quantitative metrics presented in most SDS-based methods [12, 15, 26] for fair comparisons. In this section, we expand our analysis through the inclusion of additional results from adjacency-based metrics to provide a more comprehensive comparison for multi-view consistency. Tab. 1 shows better 3D coherence of Sherpa3D.

More comparison results. We provide more comparisons with baselines in Figure 10, 11. To further demonstrate the robustness and generalization of our method, we compare

Table 1. Evaluation of adjacency-based metric. Prompt:'A head of Terracotta Army."

Method	A-LPIPS $_{VGG}$ ↓	A-LPIPS $_{Alex}$ ↓
Shap-E [8]	0.2113	0.1725
DreamFusion [15]	0.2030	0.1453
Magic3D [12]	0.1940	0.1372
ProlificDreamer [26]	0.1843	0.1322
Fantasia3D [3]	0.2350	0.1933
Debiased-SDS [**?**]	0.1920	0.1340
Ours	**0.1081**	**0.0748**

our Sherpa3D with Zero123 [13] and MVDream [20] in Figure 5. Although the concurrent work MVDream and Zero123 can also resolve the multi-view inconsistency issues via fine-tuning a costly viewpoints-aware model, we observe that it is prone to overfit the limited 3D data [4]. Specifically, MVDream generates strange color styles while Zero123 fails in such open-vocabulary prompts.

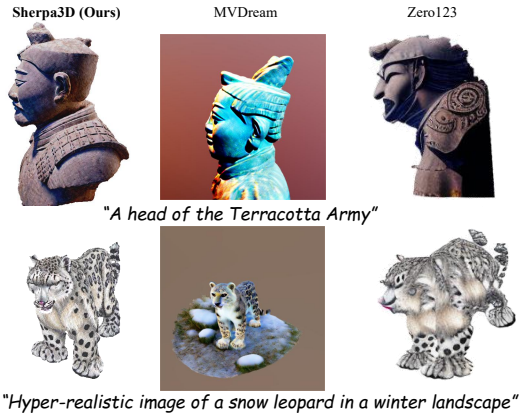


Figure 5. Comparison with MVDream [20] and Zero123 [13].

More ablation study results. In this section, additional ablation study results of our methods are presented. Figure 6 offers a qualitative illustration with the prompt, "A DSLR photo of an adorable Corgi dog with a wagging tail." Moreover, quantitative analyses are conducted and displayed in Table 2, demonstrating the efficacy of our proposed designs.

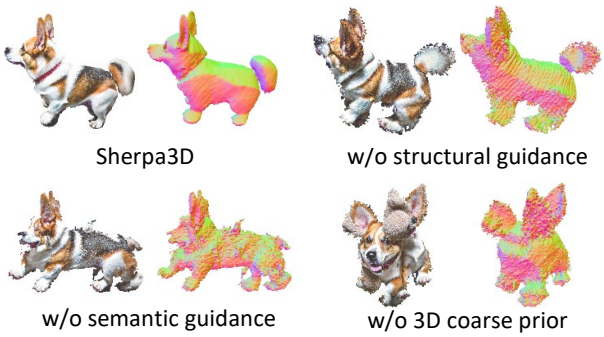


Figure 6. More qualitative results of ablation study. Prompt: "A DSLR photo of an adorable Corgi dog with a wagging tail."

"A statue of a angel"

"A futuristic battle robot, heavily armed, amidst a post-apocalyptic urban wasteland"

"A cybernetic biomechanical arm, with a blend of organic and mechanical elements"

"A luxurious sky-blue leather handbag with a sleek and elegant design, highlighted by its vibrant blue color"

"Iron Man in his state-of-the-art suit, confidently standing, looking ahead, ready for action"

"Commercial airliner in flight, sleek and modern design"

Figure 7. More generated results using our Sherpa3D within 25 minutes. Our work can generate high-fidelity and diversified 3D results from various text prompts, free from the multi-view inconsistency problem.

"Detailed portrait of a noble knight, full armor, intricate helmet design"

"Hyper-realistic image of a snow leopard, capturing its camouflage and majestic stance"

"A detailed and realistic 3D model of a vintage camera"

"Spaceship,futuristic design,sleek metal,glowing thrusters, flying in space"

"A DLSR Photo of the Leaning Tower of Pisa"

"An ultra-detailed illustration of a mythical Phoenix, rising from ashes, vibrant feathers in a fiery palette"

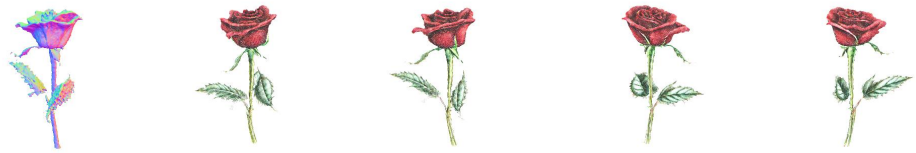
Figure 8. More generated results using our Sherpa3D within 25 minutes. Our work can generate high-fidelity and diversified 3D results from various text prompts, free from the multi-view inconsistency problem.

"A carved wooden Bodhisattva from China's Song dynasty"

"A futuristic-style motorcycle with sleek design, neon lights, and a sci-fi aesthetic in an urban setting"

"Vintage wooden race car, polished mahogany finish, classic design with spoked wheels"

"A head of the Terracotta Army"



"A blooming red rose, with velvety petals, delicate green leaves, and a captivating fragrance that fills the air"

"A DSLR photo of an adorable Corgi dog with a wagging tail"

Figure 9. More generated results using our Sherpa3D within 25 minutes. Our work can generate high-fidelity and diversified 3D results from various text prompts, free from the multi-view inconsistency problem.

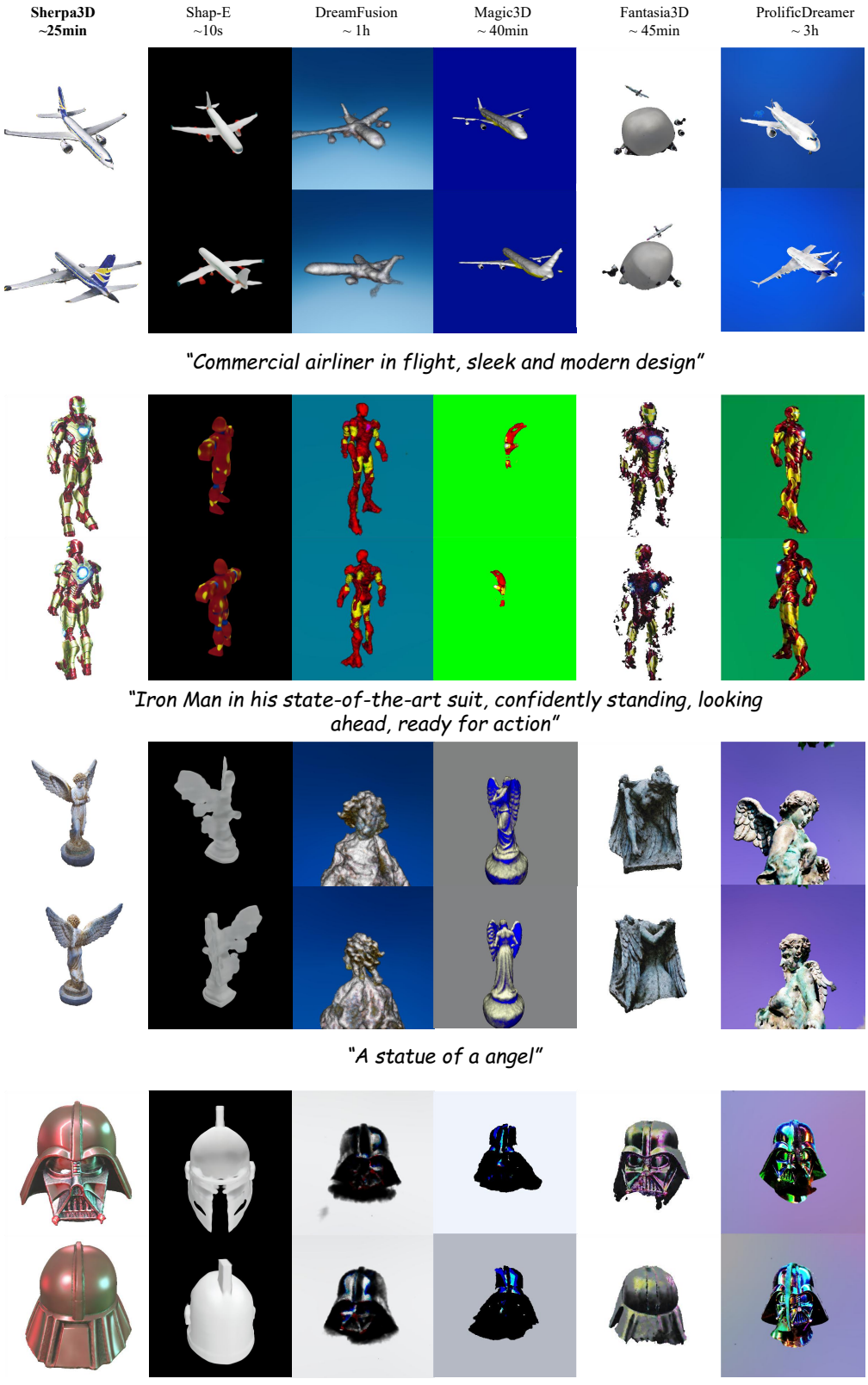


Figure 10. Qualitative comparisons with baseline methods across different views. All methods use *stabilityai/stable-diffsuion-2-1-base* for fair comparison. We observe that baselines suffer from severe multi-face issues while Sherpa3D achieves better quality and 3D coherence.

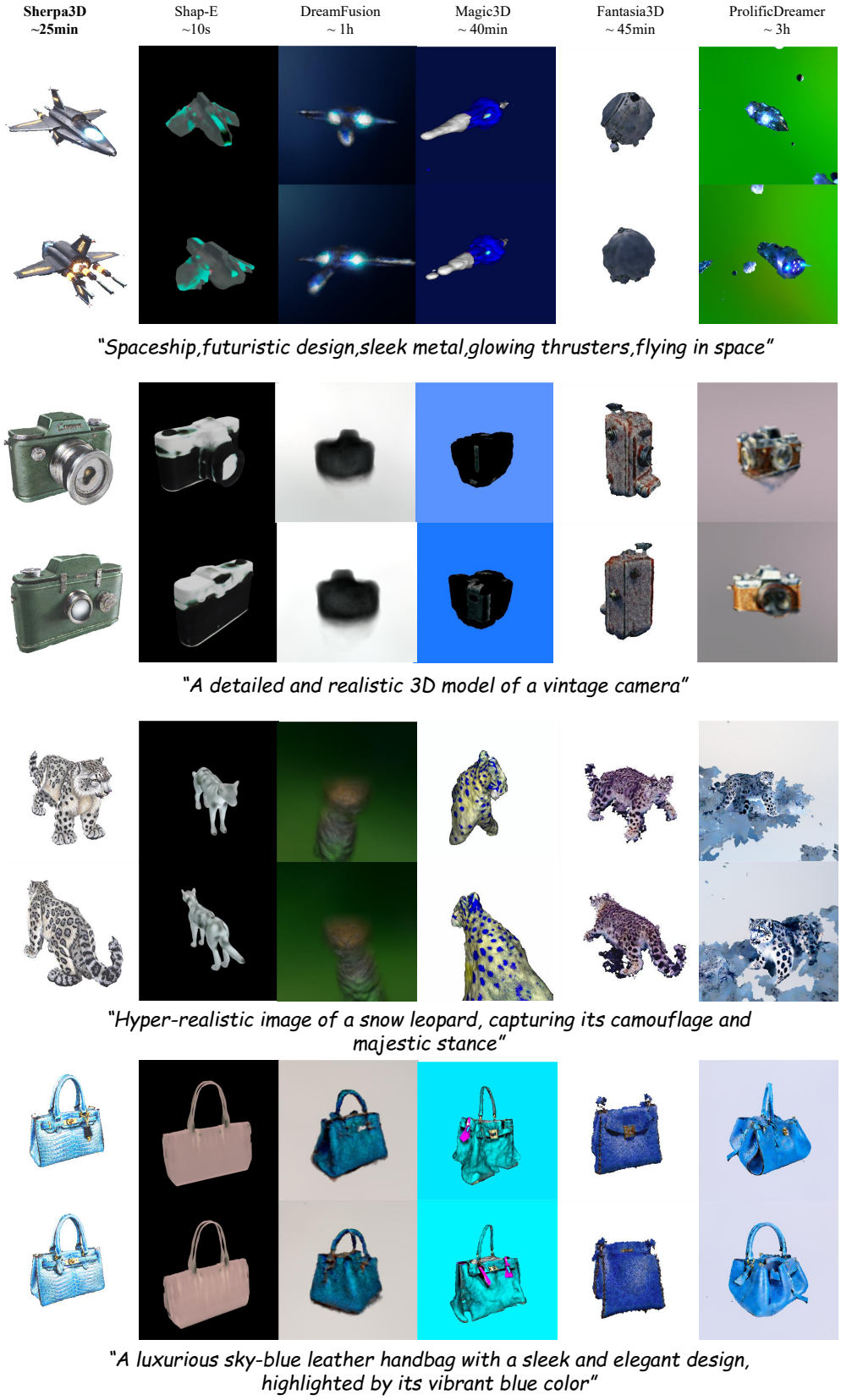


Figure 11. Qualitative comparisons with baseline methods across different views. All methods use *stabilityai/stable-diffsuion-2-1-base* for fair comparison. We observe that baselines suffer from severe multi-face issues while Sherpa3D achieves better quality and 3D coherence.

Table 2. Quantitative results of ablation study. Prompt: "A head of Terracotta Army."

Method	A-LPIPS $_{\text{VGG}}$ ↓	A-LPIPS $_{\text{Alex}}$ ↓	Multi-view consistency ↑	Overall quality ↑
w/o Structural Guidance	0.1543	0.1215	6.22	6.55
w/o Semantic Guidance	0.1662	0.1302	5.45	5.85
w/o Step Annealing	0.1210	0.1175	6.75	7.35
Full	**0.1081**	**0.0748**	**8.95**	**8.74**

References

[1] stable-diffusion-xl-base-1.0. https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0. Accessed: 2023-08-29. 1

[2] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 2

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 1, 2, 3

[4] 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. 2, 3

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

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

[7] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 2

[8] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 2, 3

[9] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 1

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[11] Jiabao Lei, Yabin Zhang, Kui Jia, et al. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *Advances in Neural Information Processing Systems*, 35:30923–30936, 2022. 2

[12] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. 1, 2, 3

[13] 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. 1, 2, 3

[14] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts, 2022. 1

[15] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 3

[16] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2

[18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 1

[19] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1

[20] 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. 1, 2, 3

[21] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[23] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[25] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022. 1

[26] 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. 1, 2, 3

[27] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 2