Fast-T3D: Fast Text-to-3D Object Generation via Scarce 3D Prior and Multi-Gradient Optimization –Supplementary Material–

Anonymous submission

This supplementary material consists of three parts, including additional implementation details, experimental results, and additional relative works. We also provide video demos in https://fastt3d.github.io.

Additional Implementation Details

In this section, we will provide more details about the experimental implementation of Fast-T3D as follows, which are omitted in the main paper due to the page constraint.

Text-Position Co-Encoder.

Specifically, we involve the utilization of hash-grid positional encoding (Müller et al. 2022) as the position encoder f_{P} to obtain a high-dimensional position embedding $\mathbf{e_P} \in \mathrm{R}^{N_{\mathrm{P}} \times d_{\mathrm{P}}}$ for any 3D position coordinate $\mathbf{pos} \in P$, where $d_{\mathrm{p}} = 32$. Besides, another MLP is applied to map the original text embeddings from CLIP into $\mathbf{e_T}$, where $d_{\mathrm{T}} = 32$. Then we implement a cross-attention layer following (Rombach et al. 2022), with 32 head dimensions and 0.2 dropout. It is noteworthy that no extra position embedding is added to the text embeddings. The feed-forward network utilizes GeLU activation after input layer.

NeRF Network.

Following (Poole et al. 2022), the implicit function network F_Θ of Neural Radiance Field (NeRF) in our paper is 5-layer MLPs with ReLU and sigmoid activation for color and NeRF density respectively, in which the hidden layer has 32 dimensions and a ReLU activation. As described in the main paper, we apply the concatenated embedding as the input of F_Θ, which has a feature dimension of 96.

Prompts Interpolation.

To facilitate out-of-domain tests, we conduct training with prompts interpolation. Specifically, in the latter stages of multi-gradient optimization, we randomly select a prompt with a 70% probability and interpolate α between the text embeddings of the selected prompt and the current prompt:

$$\alpha \mathbf{e_T}_1 + (1-\alpha)\mathbf{e_T}_2, \alpha \sim \mathcal{U}(0,1) \tag{1}$$

For text-to-image samples in the stable diffusion model, we employ text embeddings interpolation $\hat{\alpha}$ following,

$$\hat{\alpha} y_1 + (1-\hat{\alpha}) y_2, \hat{\alpha} \sim Bern(1/2) \tag{2}$$

More training Details.

We use AdamW optimizer for multiple gradients training with a learning rate 0.001 for NeRF and 0.01 for position embedding f_P. The L2 norm regulation for rendered opacity from NeRF during training. In the geometry refine phrase, we introduce a surface normal smoothness loss and laplacian smoothness loss to ensure the smoothness of the extracted mesh. We use Adam optimizer with a learning rate 0.005. In the texture refinement phrase, we use VSD loss following (Wang et al. 2023) with 2000 iterations. Note that texture refinement only needs about 10 minutes per object.

Additional Experimental Results

In this section, we evaluate the user study for additional quantitative results, present the qualitative results for prior binding and provide more visualizations of generations.

User Study

We conduct an additional quantitative evaluation of user study utilizing the 4-alternative force-choice paradigm to evaluate 20 text-3D generative tasks produced by four methods, as depicted in Table 1. The study protocol adheres to the guidelines outlined in (Rombach et al. 2022). Human preference scores are utilized to ascertain the relative performance of each method. Specifically, our generation tasks cover several representative prompts, including 5 main paper results and 15 results in Figure 2, 3, 4, 5 and 6. We also calculate the probability of Janus problem occurrence to evaluate rational geometry, where results with multiple faces, limbs, excessive hollowness, and meaningless mesh as Janus problem.

Comparing the qualitative results, we observe that Shap-E has a low Janus rate benefiting from a 3D model dataset pretraining and denser mesh representation. However, it lacks realistic texture and the open-world text generation capability is weak, especially in fine-grained prompts. ProlificDream has a realistic texture but its issue of having a multiface Janus problem is particularly severe. The findings of this study demonstrate that our Fast-T3D exhibits superior performance compared to the other methods, particularly in the figures and animal categories. Fast-T3D produces rational shapes and realistic textures, which are highly preferred according to the human preference scores obtained.

Figure 1: **Results after prior binding training.** In this work, we introduce three representative 3D prior shapes.

Method	Preference	Janus Rate
Shap-E (Jun and Nichol 2023)	5%	20%
Magic3D (Lin et al. 2022a)	9%	55%
ProlificDreamer (Wang et al. 2023)	27%	65%
Fast-T3D-refine	59%	5%

Table 1: **Quantitative comparisons** on 20 generation tasks. **Preference:** 10 users with different preferences, **Janus Rate:** The probability of Janus occurrence includes multiple faces, limbs, excessive hollowness, and meaningless mesh.

Prior Binding Results.

In this work, to obtain the geometric structure prior, we introduce three representative 3D prior shapes for the prior binding training. The outcomes obtained after the prior training stage are demonstrated in Figure 1. Based on our text-position co-encoder, a fast-fitting representation can be obtained, which is devoid of any overlap. Benefiting from the shape prior, our 3D generations can effectively alleviate the multi-face problems in the typically difficult generation tasks of human figures and animal models. Besides, the shape prior can help constrain the layout and orientation of 3D generation, contributing to better convergence.

More Generation Results.

We present more comparisons of generation results Magic3D (Lin et al. 2022a), ProlificDreamer (Wang et al. 2023) and Shap-E (Jun and Nichol 2023) in Figure 2, 3, 4, 5 and 6. It is worth noting the inference time of Fast-T3D is about a few seconds, which is comparable with Shape-E. The inference time of Fast-T3D-refine is about 30 minutes, which significantly reduces the time cost compared to Magic3D and ProlificDreamer. Besides, our generation quality obviously excel in the geometrically consistency.

Additional Related Works

For conciseness and precision, we only referenced the most relevant text-to-3D generation works on open-world text in the main paper. To provide a more comprehensive background, we will elaborate on relative works in this section.

3D shape generation. The field of 3D generative modeling has witnessed extensive research, commonly exploring explicit 3D representations including voxel grids (Smith and Meger 2017; Xie et al. 2018; Gadelha, Maji, and Wang 2017), point clouds (Achlioptas et al. 2018; Luo and Hu 2021; Zeng et al. 2022; Zhou, Du, and Wu 2021; Yang et al. 2019), meshes (Gao et al. 2022; Liu et al. 2023), etc. However, a common limitation has been their generation scalability in format of 3D assets. Recent advances in implicit representation (Mildenhall et al. 2021) lead to a series of generative models on neural field (Chan et al. 2022, 2021; Kosiorek et al. 2021; Schwarz et al. 2020). In light of the neural volume rendering, those works have explored adopting 2D images as training sources to achieve 3D-aware image synthesis. Notable attempts like EG3D (Chan et al. 2022) have been made to achieve 3D face or object generation based on 2D training data. Despite the reasonable 3D generation results, the above methods are typically restricted to the single-object category due to the limitation of training data and 3D representation scalability.

Text-to-3D generation. More recent works have leaned to text-to-3D generation tasks for more controllable 3D generations (Chen et al. 2019; Sanghi et al. 2023; Fu et al. 2022; Wei et al. 2023; Gupta et al. 2023; Cheng et al. 2023; Nichol et al. 2022; Jun and Nichol 2023). Inherited from the traditional 3D shape generation pipeline, such methods predominantly follow encoder-decoder architecture and condition the text embedding into the latent space. Specifically, TAPS3D (Wei et al. 2023) concatenates the text embedding with sampled noise to condition the generator of GAN. While some recent approaches embed the text embedding together with the latent code of 3D representation. Differing in embedding strategies, 3DGen (Gupta et al. 2023) adopt diffusion process, SDFusion (Cheng et al. 2023) apply concatenation, while Shap-E (Jun and Nichol 2023) and Point-E (Nichol et al. 2022) apply transformer backbone. In the main paper, we summarize the above works as 3D data-guided learning-based methods. Such methods rely on the text-3D data pair for training, which restricts the scalability of textual input to the finite scope of training data (Chen et al. 2019; Sanghi et al. 2023; Fu et al. 2022; Wei et al. 2023; Cheng et al. 2023). Though recent works have expanded the scale of training data (Gupta et al. 2023; Nichol et al. 2022; Jun and Nichol 2023), the generation scalability is still limited. Therefore, a series of methods conduct optimization based on the gradients (Sanghi et al. 2022; Mohammad Khalid et al. 2022; Poole et al. 2022; Lin et al. 2022b; Xu et al. 2023; Metzer et al. 2023; Wang et al. 2023) to eliminate the requirement of 3D data and inherit the generation capabilities from 2D models, which are summarized as 2D gradient-guided optimization-based methods in main paper.

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, 40–49. PMLR.

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. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Chan, E. R.; Monteiro, M.; Kellnhofer, P.; Wu, J.; and Wetzstein, G. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5799–5809.

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

Cheng, Y.-C.; Lee, H.-Y.; Tulyakov, S.; Schwing, A. G.; and Gui, L.-Y. 2023. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4456–4465.

Fu, R.; Zhan, X.; Chen, Y.; Ritchie, D.; and Sridhar, S. 2022. Shapecrafter: A recursive text-conditioned 3d shape generation model. *Advances in Neural Information Processing Systems*, 35: 8882–8895.

Gadelha, M.; Maji, S.; and Wang, R. 2017. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, 402–411. IEEE.

Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; and Fidler, S. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35: 31841–31854.

Gupta, A.; Xiong, W.; Nie, Y.; Jones, I.; and Oğuz, B. 2023. 3DGen: Triplane Latent Diffusion for Textured Mesh Generation. *arXiv preprint arXiv:2303.05371*.

Jun, H.; and Nichol, A. 2023. Shap-E: Generating Conditional 3D Implicit Functions. *arXiv preprint arXiv:2305.02463*.

Kosiorek, A. R.; Strathmann, H.; Zoran, D.; Moreno, P.; Schneider, R.; Mokrá, S.; and Rezende, D. J. 2021. Nerf-vae: A geometry aware 3d scene generative model. In *International Conference on Machine Learning*, 5742–5752. PMLR.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2022a. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv preprint arXiv:2211.10440*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2022b. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv preprint arXiv:2211.10440*.

Liu, Z.; Feng, Y.; Black, M. J.; Nowrouzezahrai, D.; Paull, L.; and Liu, W. 2023. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2837–2845.

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

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

Mohammad Khalid, N.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*, 1–8.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv preprint arXiv:2212.08751*.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sanghi, A.; Chu, H.; Lambourne, J. G.; Wang, Y.; Cheng, C.-Y.; Fumero, M.; and Malekshan, K. R. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18603–18613.

Sanghi, A.; Fu, R.; Liu, V.; Willis, K. D.; Shayani, H.; Khasahmadi, A. H.; Sridhar, S.; and Ritchie, D. 2023. CLIP-Sculptor: Zero-Shot Generation of High-Fidelity and Diverse Shapes From Natural Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18339–18348.

Schwarz, K.; Liao, Y.; Niemeyer, M.; and Geiger, A. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33: 20154–20166.

Smith, E. J.; and Meger, D. 2017. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, 87–96. PMLR.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213*.

Wei, J.; Wang, H.; Feng, J.; Lin, G.; and Yap, K.-H. 2023. TAPS3D: Text-Guided 3D Textured Shape Generation from Pseudo Supervision. *arXiv preprint arXiv:2303.13273*.

Xie, J.; Zheng, Z.; Gao, R.; Wang, W.; Zhu, S.-C.; and Wu, Y. N. 2018. Learning descriptor networks for 3d shape synthesis and analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8629–8638.

Xu, J.; Wang, X.; Cheng, W.; Cao, Y.-P.; Shan, Y.; Qie, X.; and Gao, S. 2023. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20908–20918.

Yang, G.; Huang, X.; Hao, Z.; Liu, M.-Y.; Belongie, S.; and Hariharan, B. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4541–4550.

Zeng, X.; Vahdat, A.; Williams, F.; Gojcic, Z.; Litany, O.; Fidler, S.; and Kreis, K. 2022. LION: Latent point diffusion models for 3D shape generation. *arXiv preprint arXiv:2210.06978*.

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

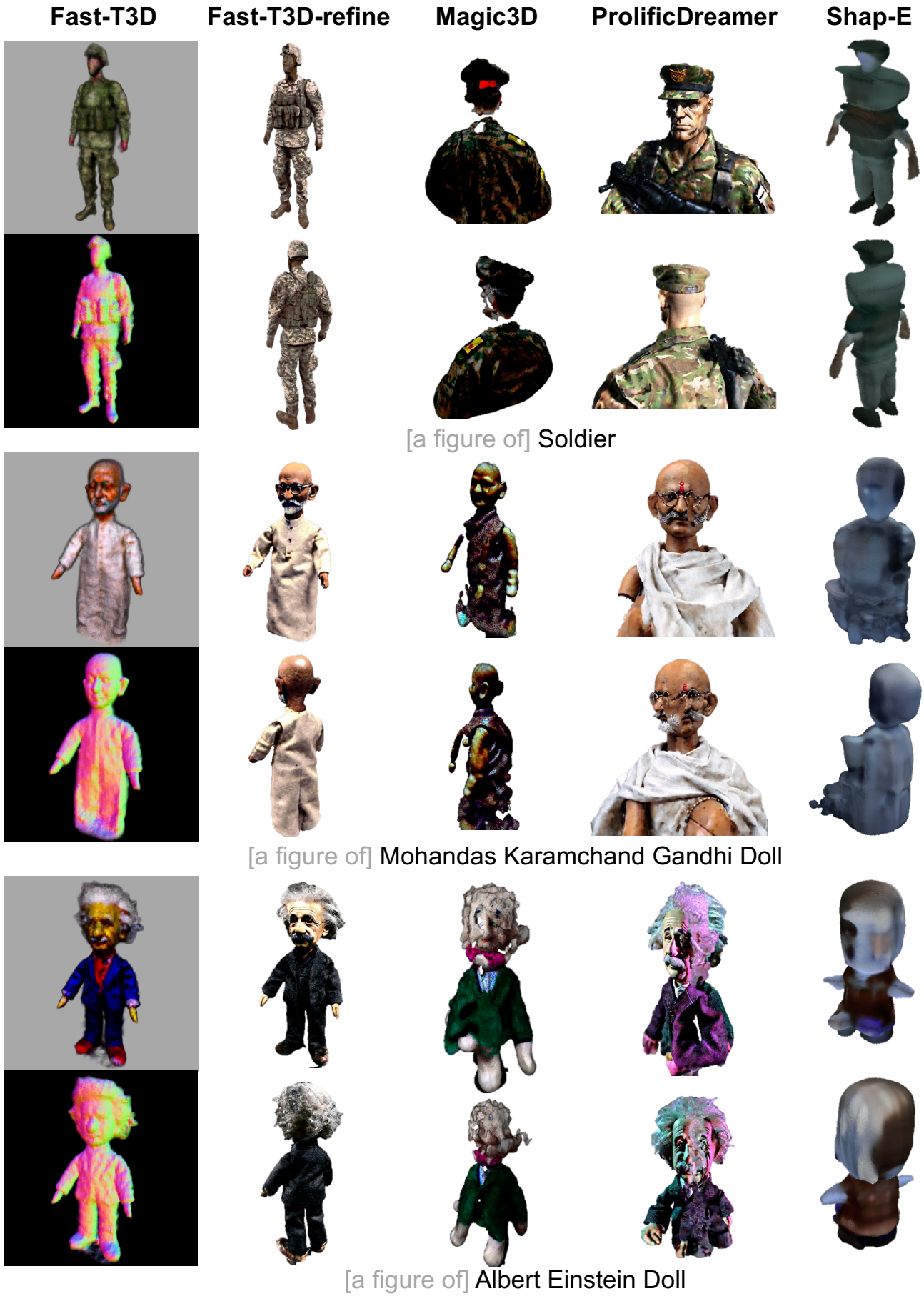

Figure 2: **Qualitative results comparisons** with the open-source or open-model methods.

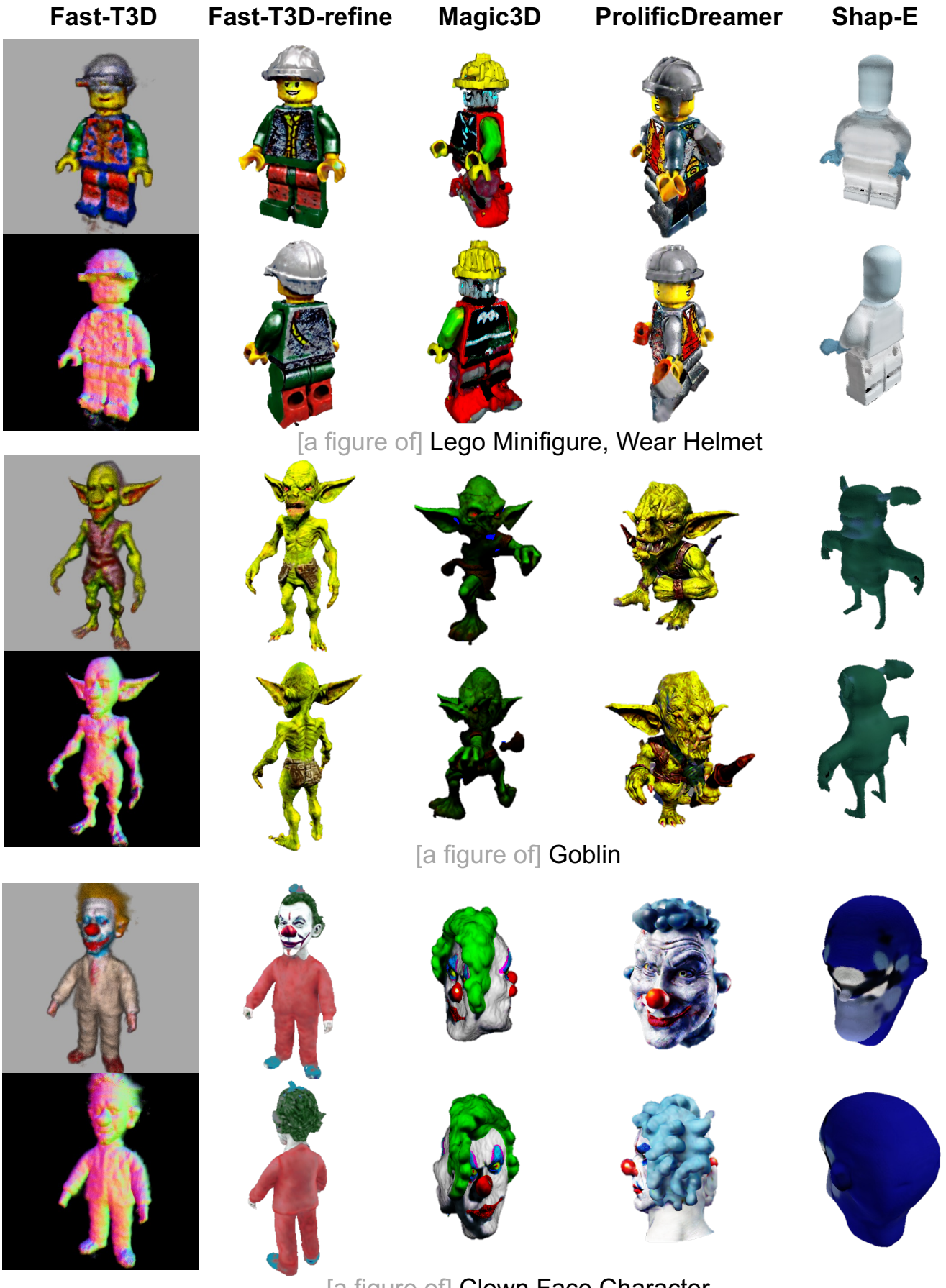

Figure 3: **Qualitative results comparisons** with the open-source or open-model methods.

Figure 4: **Qualitative results comparisons** with the open-source or open-model methods.

Figure 5: **Qualitative results comparisons** with the open-source or open-model methods.

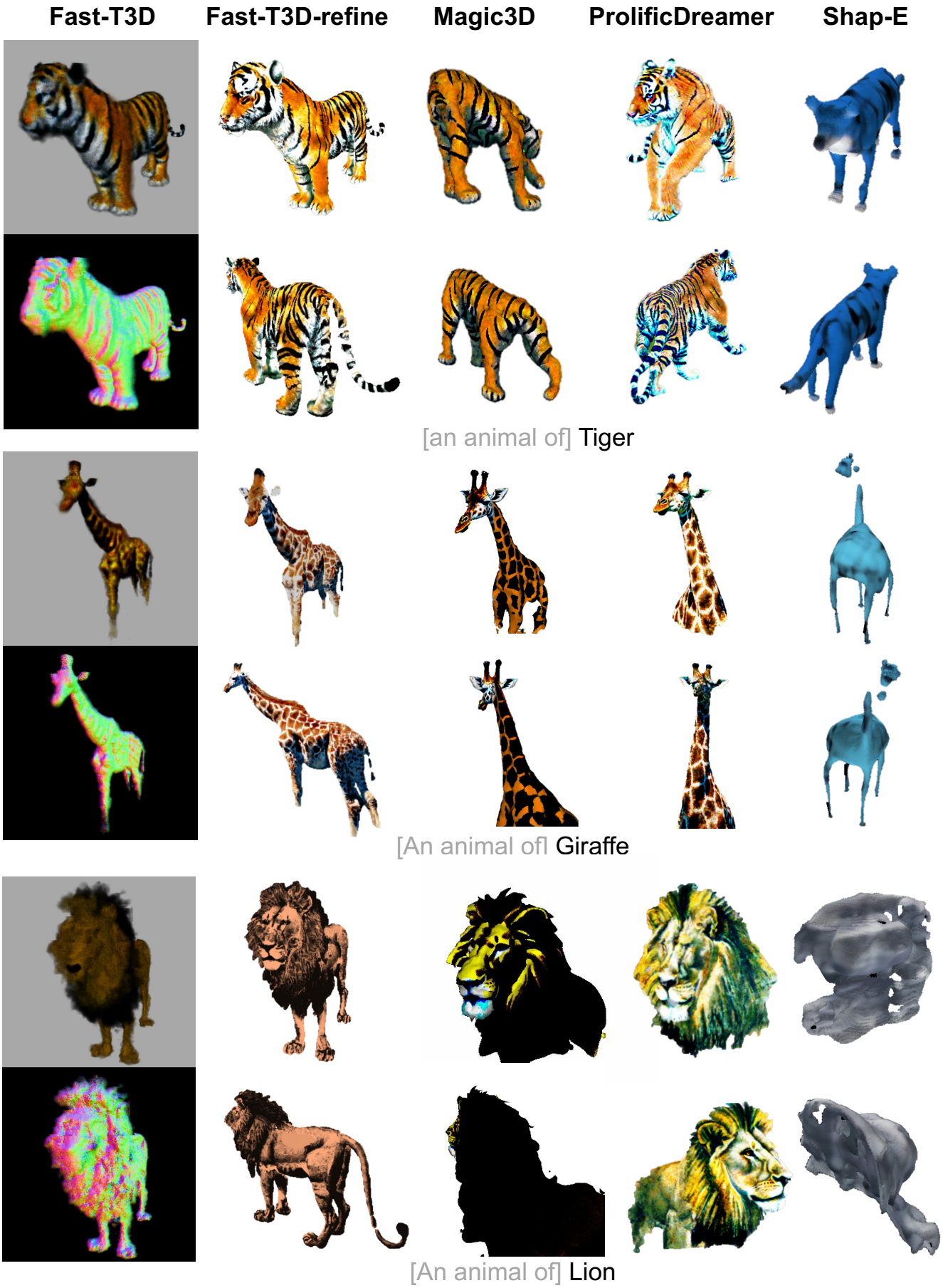

Figure 6: **Qualitative results comparisons** with the open-source or open-model methods.