Make-Your-3D: Fast and Consistent Subject-Driven 3D Content Generation

Fangfu Liu, Hanyang Wang, Weiliang Chen,
Haowen Sun, and Yueqi Duan*

Tsinghua University
{liuff23,hanyang-21,chen-wl20,shw22,duanyueqi}@mails.tsinghua.edu.cn

A Additional Implementation Details

In this section, we provide additional details about the implementation of our subject-driven 3D generation framework **Make-Your-3D**.

A.1 Hyperparameter Settings

In our subject-driven optimization, We retain the optimizer settings and ϵ-prediction strategy from the pretrained process, with a 0.9 adam β_1, a 0.999 adam β_2, a $1e-2$ adam weight decay, and a $1e-8$ adam ϵ. During the optimization, we use a reduced image size of 256×256.

A.2 Training Details

The time consumption for the three stages in our framework is as follows: identity-aware optimization takes $\sim$ 1 minute, subject-driven optimization takes $\sim$ 3 minutes, and mesh conversion takes approximately $\sim$ 1 minute. The mesh conversion module is designed to be adaptable to a variety of mesh extraction models [3,8,9]. For the sake of improved efficiency, we have chosen to utilize the LGM [8], which is capable of generating 3D Gaussians of objects within a mere 7 seconds. After that, we train an efficient NeRF (*i.e.*, Instant-NGP [4]) by using the rendered images from 3D Gaussians, and then convert the NeRF to polygonal meshes [9]. Specifically, we train two hash grids to reconstruct the geometry and appearance from Gaussian renderings. Please refer to LGM [8] for more details of generated Gaussians. With adequately optimized implementation, it takes extra 1 minute to perform this Gaussians to NeRF to mesh conversion. Our codes for implementation will be available upon acceptance.

A.3 Designing Prompts for One-Shot Personalization

Our goal is to let the diffusion model's be deeply aware of a new subject's identity. As mentioned in DreamBooth [7], a common way to personalize a diffusion

* Corresponding author.

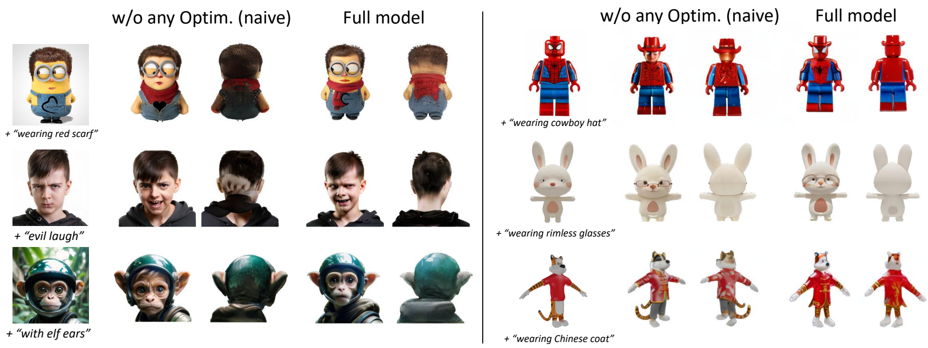

Fig. 1: Comparison results with the naïve combination.

model is to "implant" a new (*unique identifier, subject*) pair into the model's "dictionary" and label input multi-view images of the subject "a [identifier] [class noun]", where [identifier] is a unique identifier (*e.g.*, "xxy5syt00") linked to the subject and [class noun] is a coarse class descriptor of the subject (*e.g.*, cat, dog, watch, etc.). However, this simple method loses all the position information in multi-view images, which is strongly important for identity awareness. To fully leverage the concealed information, we propose a novel *(unique identifier, subject, direction)* pair inspired by view-dependent prompting in DreamFusion [5] and automatically label each image "a [identifier] [class noun], [direction]", where "direction" is one of the six directions (*e.g.*, front, back, left, etc.) for the corresponding multi-view image. By employing this design, we ensure that positional cues are incorporated into our identity-aware optimization process. We utilize the resulting six matched pairs for the optimization, further enhancing the model's ability to capture and comprehend the subject's identity.

B Additional Comparison Results

To further demonstrate the effectiveness and impressive visualization results of our Make-Your-3D, we conducted more experiments including additional comparison results against DreamBooth3D [6] and visual results (*e.g.*, multi-view images, textured meshes, normals, etc.).

B.1 More Qualitative Comparisons

Fig. 3 demonstrates additional qualitative comparisons with DreamBooth3D [6]. We observed that DreamBooth3D tends to generate coarse results for limited subjects, which lack sufficient identity consistency and also suffer from overfitting issues. For example, the red backpack in Fig. 3 exhibits three small signs at the right bottom, which are not preserved by DreamBooth during its generation process. On the other hand, our proposed method successfully maintains this distinct feature while generating high-resolution outputs.

B.2 More Quantitative Results

As there is no "Ground-Truth" 3D content corresponding to our prompts, we choose another deterministic adjacency-based metric as in Instruct-NeRF2NeRF [2] for multi-view consistency evaluation as shown in Tab. 1.

Table 1: Deterministic metrics for multi-view consistency.

Method	DreamBooth3D	MV DreamBooth	Ours
A-LPIPS $_{\text{VGG}}\downarrow$	0.2045	0.1856	**0.1172**
A-LPIPS $_{\text{Alex}}\downarrow$	0.1715	0.1340	**0.0859**

B.3 Comparisons with the Naïve Method

We compare with the naïve combination method (without any optimization) in Fig. 1 and Tab. 2, which shows the results from our full model are much better than the naïve combination in terms of identity preservation and visual quality.

Table 2: More comparisons with the naïve method and others.

Method	w/o any Opt. (naïve)	w/o Identity-aware Opt.	w/o Subject prior Opt.	Full model
ViT-B/16 $\uparrow$	0.802	0.807	0.812	**0.817**
ViT-B/32 $\uparrow$	0.730	0.741	0.745	**0.764**
ViT-L-14 $\uparrow$	0.790	0.798	0.805	**0.826**
A-LPIPS $_{\text{VGG}}\downarrow$	0.1612	0.1563	0.1550	**0.1172**
A-LPIPS $_{\text{Alex}}\downarrow$	0.1435	0.1120	0.1095	**0.0859**

B.4 Comparion with 3D Editing Method

We also compare our customized method with these 3D editing methods. We first generated multiple views with Wonder3D, trained a NeRF as its 3D representation, and performed text-based 3D editing with Instruct-NeRF2NeRF [2]. Fig. 2 shows the hat is with lower quality. It is worth noting that most text-based 3D editing methods require 2D diffusion models as per-image supervision, which would hurt original 3D quality with lower efficiency. In contrast, our method operates editing directly in 2D space followed by 3D reconstruction, making it more efficient and with lower iterative quality degradation.

C Additional Visual Results

Fig. 4, 13 shows multiple views of the assets rendered for the subjects with different text prompts. Fig. 9,10 provides additional results with associated normals

Fig. 2: Comparisons with 3D editing method. It is interesting to compare with single-view 3D reconstruction + text-based 3D editing. For fair comparison, we only change IP-Adapter to Instruct-NeRF2NeRF for text-based 3D editing while keeping Wonder3D model.

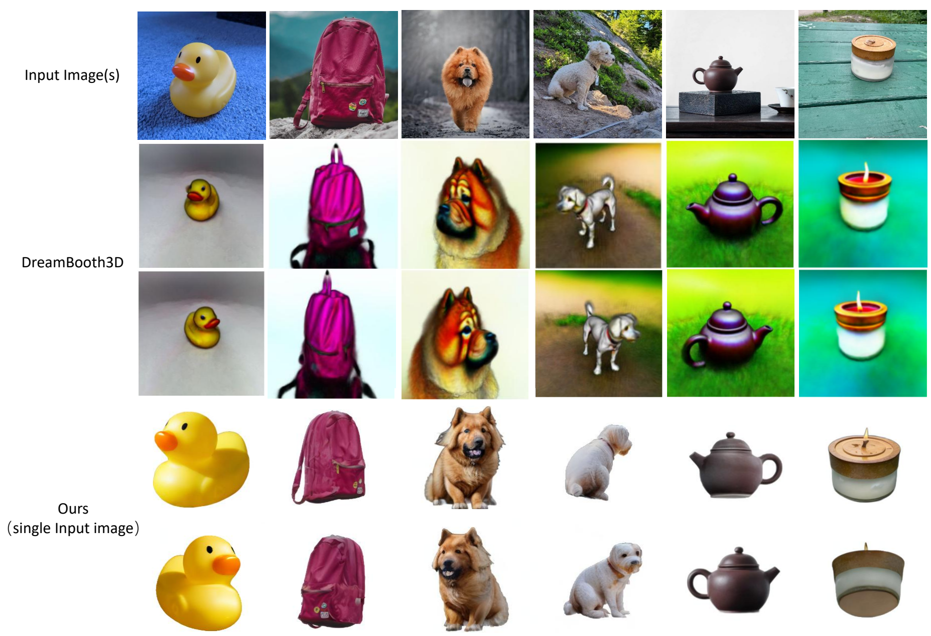

Fig. 3: More qualitative comparisons with DreamBooth3D on subjects of DreamBooth dataset. Notice ours perform better on the object's identity consistency with less input images.

and meshes to demonstrate the 3D consistency of our results on a variety of customized subjects.

C.1 More Human Personalization Results

Fig. 11, 12 presents additional personalized examples of human faces generated by our co-evolution framework. Our approach enables modification of attributes such as hairstyle, clothing, and more, as well as the capability to alter expressions, makeup, and styling. These remarkable results provide further evidence of the effectiveness of the co-evolution framework in our Make-Your-3D application and demonstrate its broad potential for various areas, such as customizing 3D characters, virtual reality, online clothing try-on, and beyond.

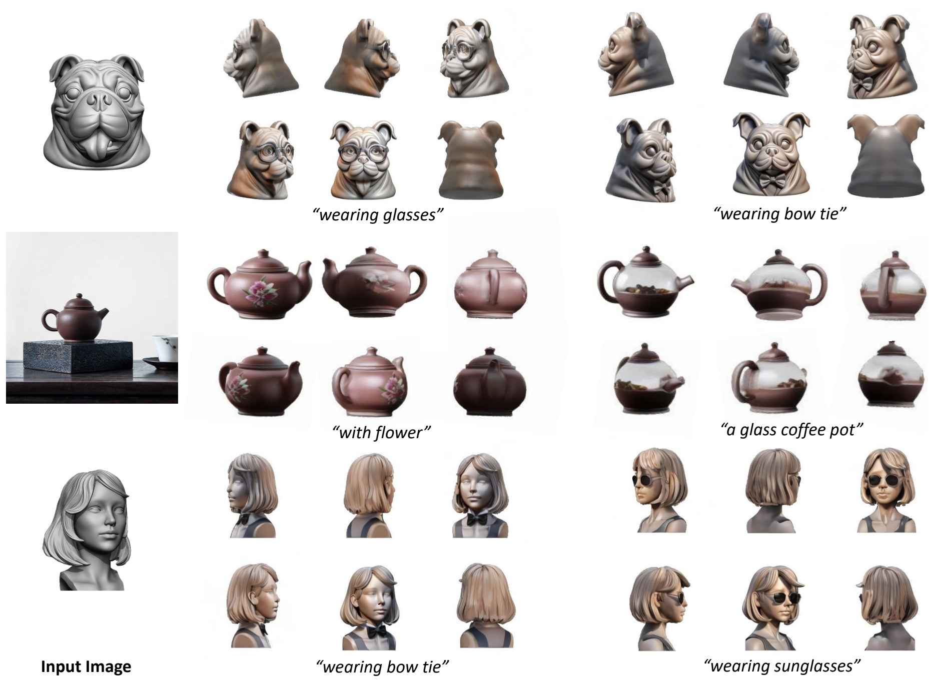

Fig. 4: More personalization results for a subject with different text inputs.

C.2 More Ablation Results

We present a comprehensive comparison in Fig. 5. We observe that finetuning Wonder3D maintains similar performance in meshes and normals. We also evaluate the number of reference multi-views in two-stage optimization as ablation (Fig. 6) shows. More reference views lead to better identity preservation and quality.

D GPT4-V for 3D Evaluation

We choose a recent automatic and versatile evaluation metric GPTEval3D [10] based on GPT-4Vision (GPT-4V) [1] for additional pairwise comparison. For the two 3D assets, we render them from four or nine viewpoints. These two images will be concatenated together before passing into GPT-4V along with the text

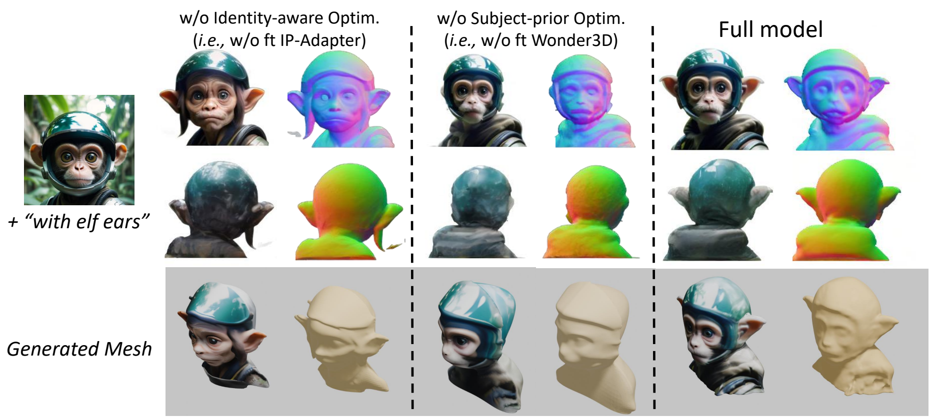

Fig. 5: More comparisons on meshes and normal maps. The fine-tuned RGB results are superior to results without finetuning while maintaining similar visual performance in normal maps

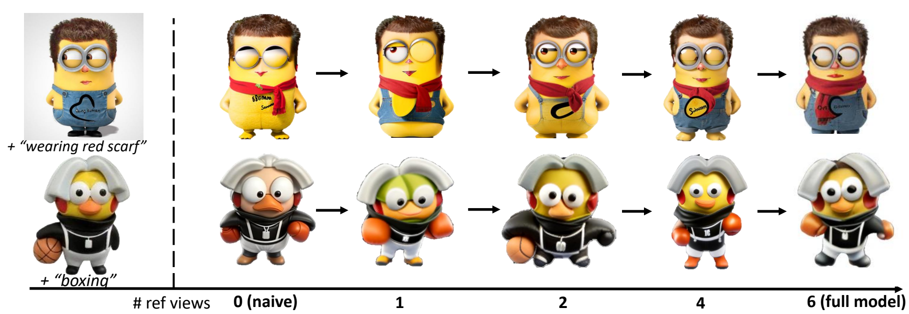

Fig. 6: Ablations on the number of mutli-views in optimization.

instructions. GPT-4V will return a decision of which of the two 3D assets is better according to the instruction. As shown in Fig. 7, We evalute our results in three main criteria: text–asset alignment, 3D plausibility and texture details.

E Ethical Statement

We confirm that all images used in this paper for research and publication have been obtained and used in a manner compliant with ethical standards. The individuals depicted in these images have given consent for their use, or the images are sourced from publicly available datasets and were used in accordance with the terms of use and permissions. Furthermore, the publication and use of these images do not pose any societal or ethical harm. We have taken necessary precautions to ensure that the research presented in this paper respects individual rights, including the right to privacy and the fundamental principles of ethical research conduct.

Fig. 7: Examples of the analysis by GPT-4V [1]. Given two 3D assets, we ask GPT-4V to compare them on various aspects and provide an explanation. We find that GPT-4V's preference closely aligns with that of humans in our user study.

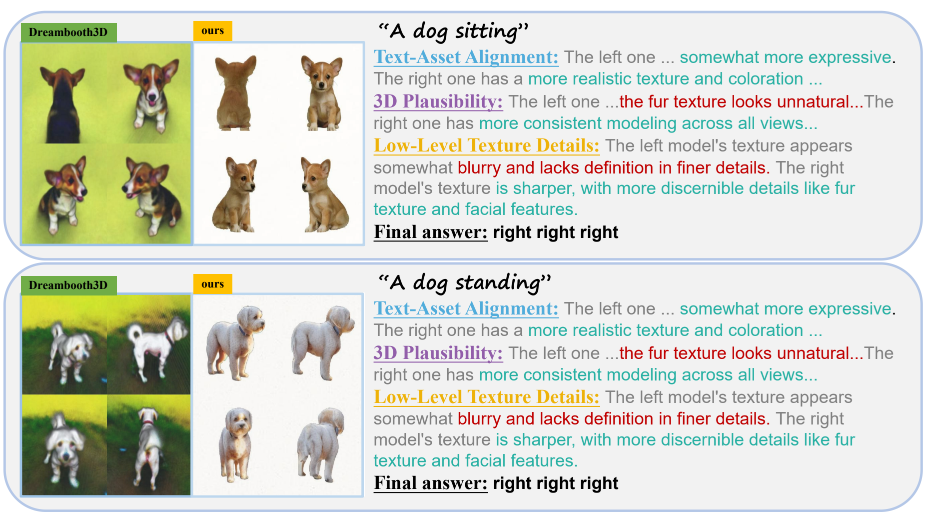

Fig. 8: More examples of the analysis by GPT-4V [1].

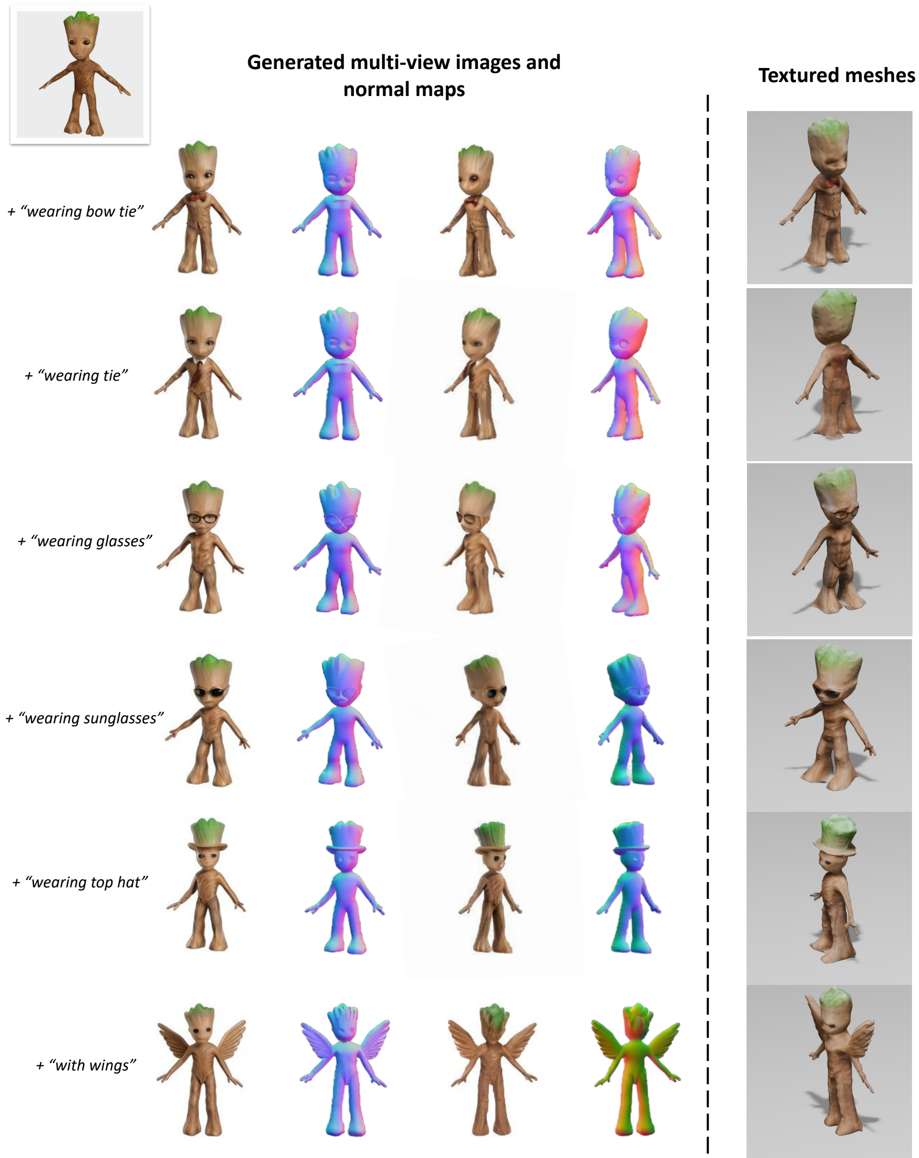

Fig. 9: Textured meshes and normal maps on a subject "Gelute" with various customized text inputs.

Fig. 10: More textured and untextured mesh results.

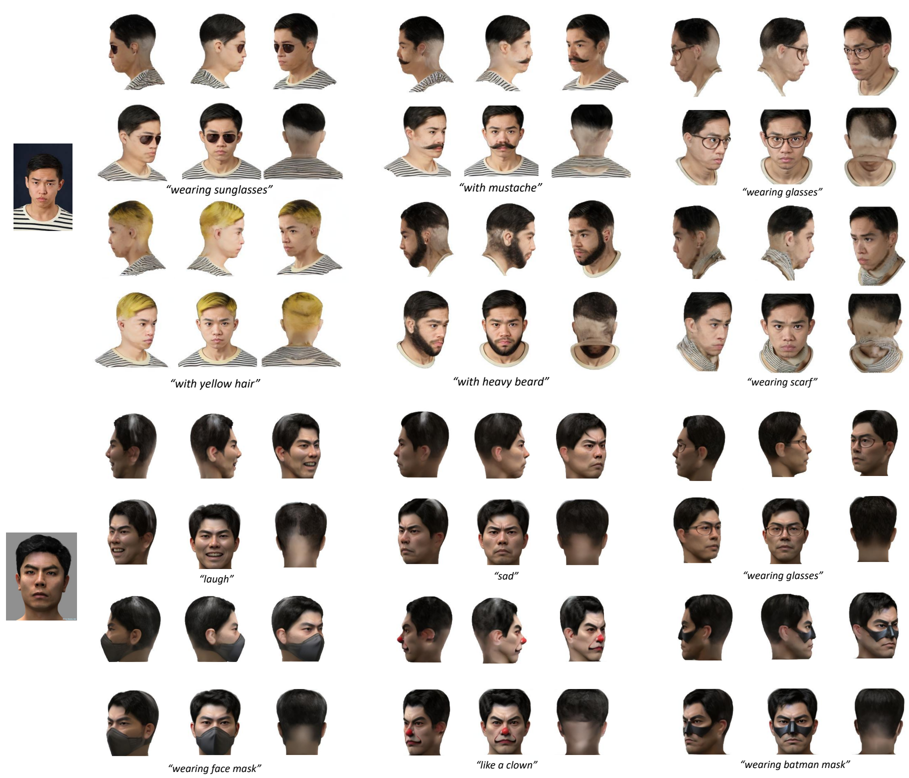

Fig. 11: More personalization results for human faces.

Fig. 12: More personalization results for one person with multiple customized text inputs.

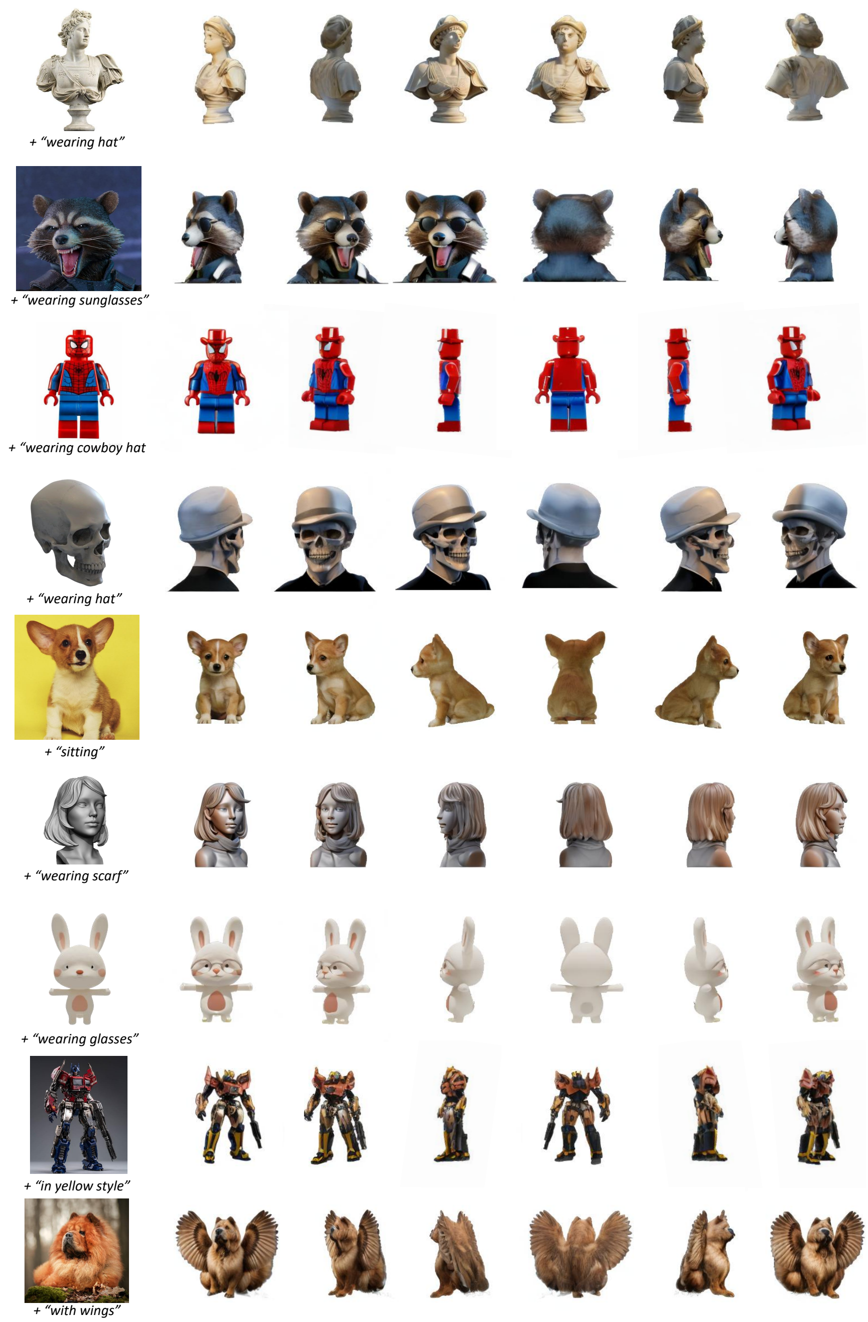

Fig. 13: More visual results of Make-Your-3D on different subjects with customized text inputs.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Haque, A., Tancik, M., Efros, A.A., Holynski, A., Kanazawa, A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19740–19750 (2023)
3. 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. arXiv preprint arXiv:2310.15008 (2023)
4. 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)
5. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion (2022)
6. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., et al.: Dreambooth3d: Subject-driven text-to-3d generation. arXiv preprint arXiv:2303.13508 (2023)
7. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
8. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation (2024)
9. Tang, J., Zhou, H., Chen, X., Hu, T., Ding, E., Wang, J., Zeng, G.: Delicate textured mesh recovery from nerf via adaptive surface refinement. arXiv preprint arXiv:2303.02091 (2023)
10. Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., Wetzstein, G.: Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. arXiv preprint arXiv:2401.04092 (2024)