Photo3D: Advancing Photorealistic 3D Generation through Structure-Aligned Detail Enhancement

Supplementary Material

1. More Details for Photo3D-MV

As mentioned in the main paper, we developed a structure-aligned multi-view synthesis pipeline to construct the detail-enhanced dataset Photo3D-MV. In this pipeline, we first process the text prompts, then generate the corresponding images, construct 3D models from these images, and finally refine the rendered results to achieve greater realism.

Specifically, for the text prompt processing stage, we employ LLaMA-3-8B [2] to transform the input text from DiffusionDB [12] into object-centered descriptions, while appending unified realistic constraints to ensure photorealistic image generation. For the final realistic multi-view generation stage, we use GPT-4o-Image to generate detail-enhanced multi-views on the rendered images with an editing prompt. The textual prompts we used are as follows:

Text Prompt Processing (Input for LLaMA-3-8B)

"Optimize this prompt into a single, high-quality, photorealistic physical object description, focusing on realistic materials, detailed textures, and authentic visual qualities: {*Raw_Text*}."

Realistic Constraints (Input for Flux.1-Dev [7])

"{*Text_Prompt*}, real camera shot, real photograph, pure white background with no shadows, complete object, high-quality photography, macro lens detail, professional studio lighting."

Realistic Multi-View Generation (Input for GPT-4o-Image [6])

"Edit Image, photorealistic micro-refinement only, make it a real object; strictly preserve exact composition and framing (NO recomposition); lock camera parameters (position, rotation, FOV, focal length); lock scale and subject position; preserve exact geometry, silhouette and perspective; fix tiny artifacts; refine textures and micro-details; keep colors and lighting exactly the same."

Besides, we also evaluated several recent advanced image generators for realistic multi-view generation, including GPT-4o-Image, Gemini-2.5-Flash [3], and Flux.1-Kontext [9], which are capable of both text and image conditioned generation. Notably, GPT-4o-Image achieves superior performance in fine-grained realistic detail enhancement and is therefore adopted in our framework. Since the Trellis-generated 3D models will have some color distortions, we further align the generated views with the rendered views by performing per-channel histogram matching in the CIE L*a*b* [4] color space, so that the luminance and color distributions can match the original rendered views while preserving the 3D structure.

Compared to original Trellis outputs, after GPT-4o-Image enhancement, multi-view consistency of Photo3D-MV defined by Probe3D [5] slightly changes from 0.854 to 0.842 (1%↓), while realism measured by MANIQA [15] substantially improves from 0.438 to 0.655 (**50%**↑), indicating that the enhancement boosts realism without severe consistency degradation.

2. Analyses on Training Strategies

2.1. Test-Time Optimization Baseline

We show the test-time optimization results in Fig. 1 to explain our proposal of training the 3D generators instead of using a 2D generator for multi-view test-time enhancement. The results highlight that improving the 3D model itself leads to more consistent and geometry-aware realism. Compared to Photo3D, test-time optimization baseline significantly degrades both quality (MUSIQ: 76.6→ 71.5) and time efficiency (10 s→ >4 min). This is mainly caused by the mismatches with the initial 3D geometry (see Fig. 1 **Mix Trellis and GT**) in the enhanced views, which inevitably lead to texture distortions during test-time optimization, whereas Photo3D aligns realistic details with the 3D-native distribution, effectively preserving geometric consistency.

2.2. Training-Time Optimization Strategies

We analyze different training strategies for diffusion-based geometry–texture coupled 3D generation. In the original training paradigm of the geometry–texture coupled 3D-native generator Trellis, each 3D asset is first rendered into about 150 multi-view images, which are then projected onto the asset's voxelized representation and subsequently encoded into a structured 3D latent. Following the original Trellis paradigm, we attempt to construct GT 3D latents by projecting the 4 realism-enhanced multi-view images onto the sparse structures generated in the preceding 3D sparse structure generation stage of Trellis, and jointly train the 3D VAE models on the resulting structured 3D latents under the supervision of realism-enhanced multi-view images. However, since the reconstructed structured 3D latent appearance mainly depends on the information provided during projection, four orthogonal views cannot fully cover the latent volume, leading to blurred regions in the unseen areas (Fig. 2(a)), or ensure detail consistency across views (Fig. 2(b)), resulting in distorted 3D structures. Con-

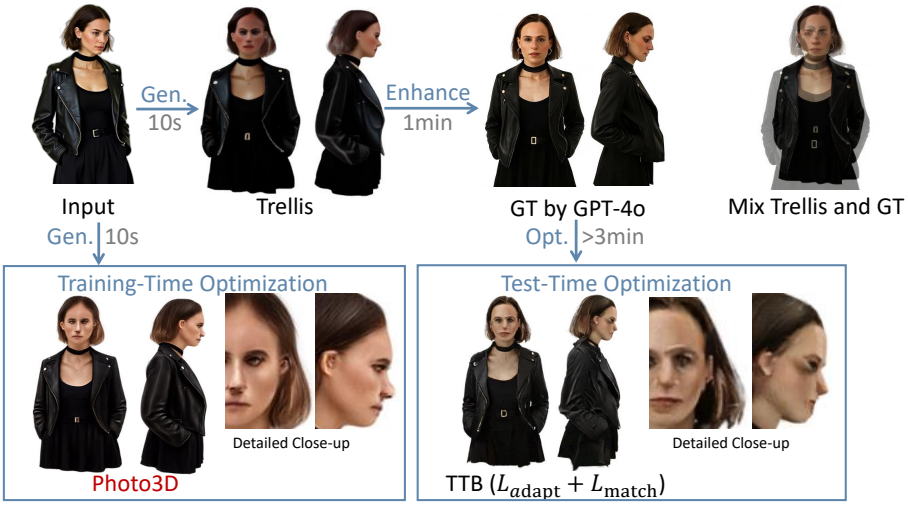

Figure 1. Comparison between the test-time optimization baseline and Photo3D (training-time optimization).

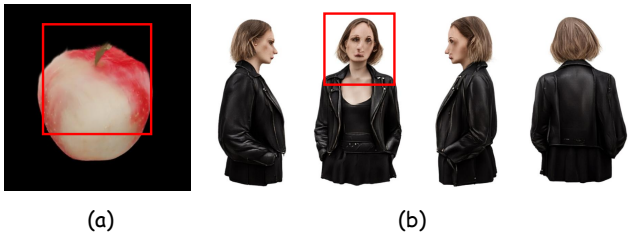

Figure 2. (a) Incomplete view coverage causes blurred regions. (b) Inconsistent details across views lead to distorted 3D structures.

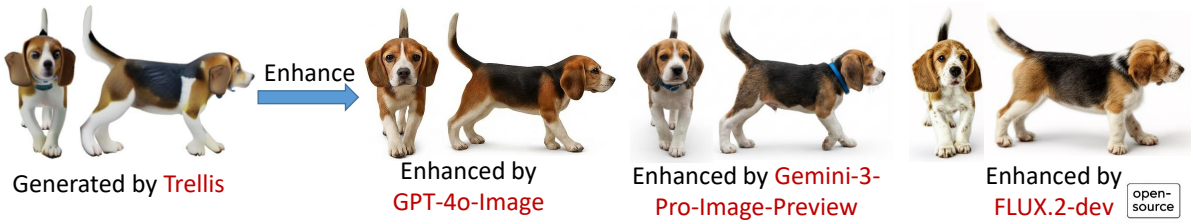

Figure 3. Enhanced multi-views of different 2D generators.

sequently, such structured GT 3D latents provide insufficient supervision for realistic 3D generation. Our proposed realism-enhancement scheme offers a more robust solution that overcomes this limitation.

3. Choices of 2D generators

Our framework is not sensitive to specific 2D generators and can work with any high-quality image model. More recent models, such as Gemini-3-Pro-Image-Preview [1] and the open-source model FLUX.2-dev [8], also work well with our proposed realistic multi-view synthesis pipeline, as shown in Fig. 3.

4. More Results

We present more 3D generation results produced by Photo3D, together with those of its baseline counterparts, in Fig.5, Fig. 6 and Fig.7. Note that for the 3D-native texturing model TexGaussian [14], we first input the images into Step1X-3D to obtain untextured 3D meshes. Subsequently, we use BLIP-2 [10] to generate captions for each image, which are then adopted as text conditions for both TexGaussian and Photo3D (TexGaussian).

5. Geometric Correction

For Photo3D (Trellis), our method can correct geometry flaws (see examples in Fig. 4). This is achieved via the learned realistic appearance priors under coupled geometry–texture optimization. For geometry–texture decoupled methods (TexGaussian and Step1X-3D), the geometry is fixed in the texturing stage and thus requires a good base geometry generated in the previous stage.

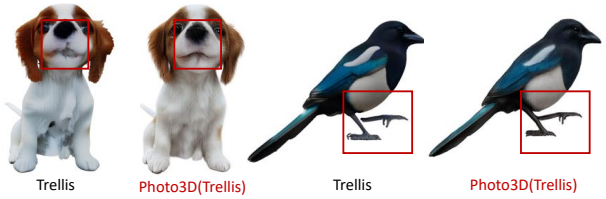

Figure 4. Geometric correction ability of Photo3D (Trellis).

Figure 5. Comparison of 3D generation results between TexGaussian [14] and Photo3D trained on the TexGaussian's 3D-native texturing model. The results demonstrate the realistic appearance of geometry-texture decoupled 3D-native generation achieved by Photo3D.

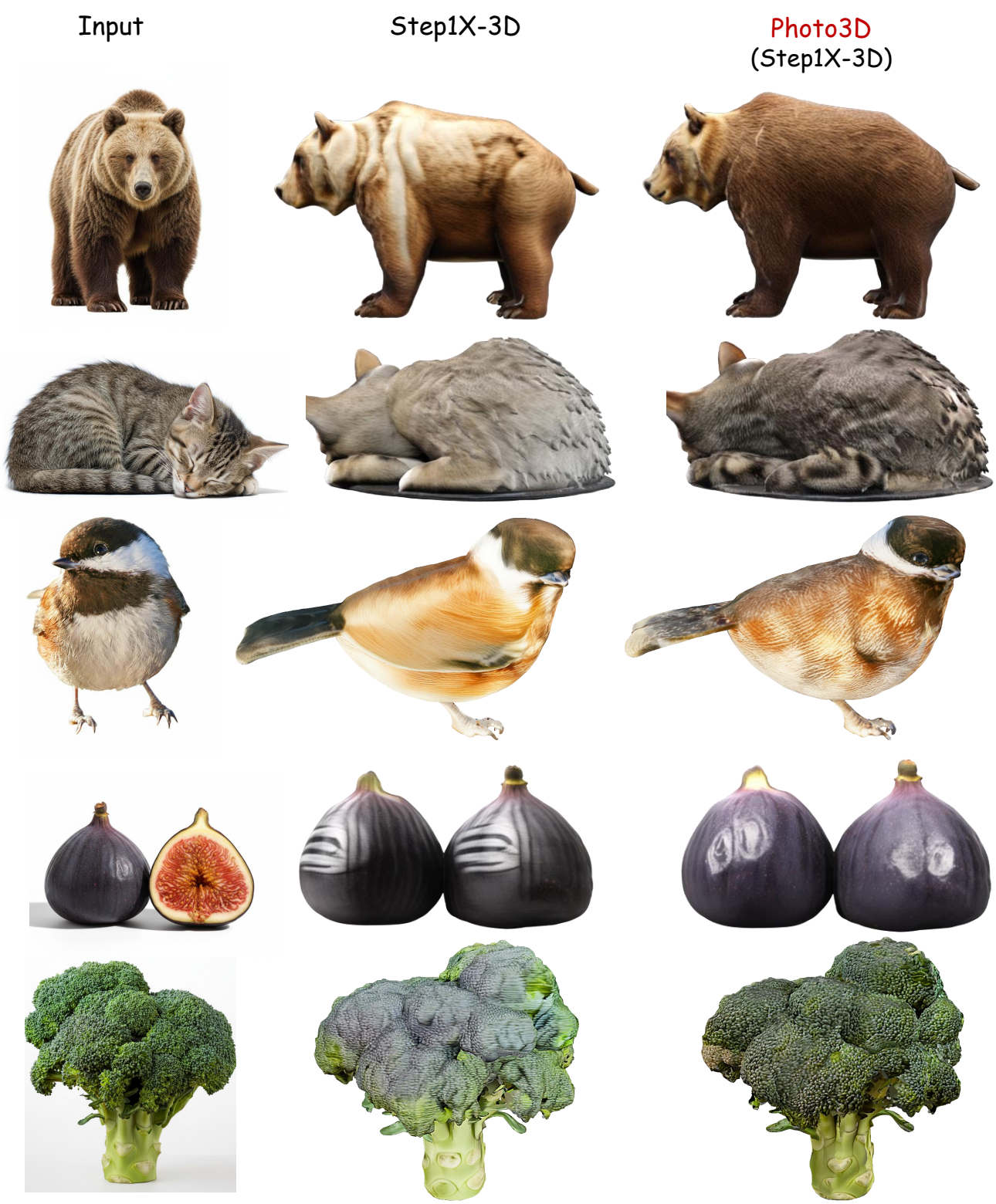

Figure 6. Comparison of 3D generation results between Step1X-3D [11] and Photo3D trained on the Step1X-3D's multi-view texturing model. The results demonstrate the realistic appearance of geometry-texture decoupled 3D-native generation achieved by Photo3D.

Figure 7. Comparison of 3D generation results between Trellis [13] and Photo3D trained on the Trellis model. The results demonstrate the realistic appearance of geometry-texture coupled 3D-native generation achieved by Photo3D.

References

[1] Nano banana pro. https://deepmind.google/models/gemini/pro/, 2025. 2

[2] AI@Meta. Llama 3 model card. 2024. 1

[3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[4] Ernst DmenuACl and R HS. Commission internationale de l'eclairage c. 1

[5] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D Awareness of Visual Foundation Models. In *CVPR*, 2024. 1

[6] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[7] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[8] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025. 2

[9] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1

[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. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[11] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 5

[12] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 1

[13] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 6

[14] Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 551–561, 2025. 2, 4

[15] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022. 1