VP3D: Unleashing 2D Visual Prompt for Text-to-3D Generation — Supplementary Material

Yang Chen[†], Yingwei Pan[†], Haibo Yang[§], Ting Yao[†], and Tao Mei[†]
[†]HiDream.ai Inc.
[§]School of Computer Science, Fudan University, China
{c1enyang, pandy}@hidream.ai, yanghaibo.fdu@gmail.com, {tiyao, tmei}@hidream.ai

Figure 1. Diversity comparison between SDS models and VP3D.

This supplementary material contains: 1) diversity comparisons between SDS models and our VP3D [1]; 2) visualization of visual prompts used in Figure 3 of the main paper; 3) comparisons against image-to-3D methods.

1. Diversity Comparisons

As discussed in the DreamFusion [5] paper, SDS is not a perfect loss fusion that has mode-seeking property (tends to seek the most common visual appearance of text prompt at high noise levels). Thus existing SDS-based works [2, 5] inevitably result in oversmoothing 3D generations with lower diversity across random seeds (see Figure 1 (a-b)). Instead, our VP3D aligns SDS optimization with additional visual appearance knowledge in 2D visual prompt. This way encourages 3D generations to match the commonly diverse and semantically relevant 2D visual prompt, thereby nicely retaining the diversity capability of 2D diffusion model in 3D generation. Figure 1 (c) shows more diverse results of VP3D across random seeds.

2. Visualization of Visual Prompts

Figure 2 illustrates the visual prompts that we used in Figure 3 of the main paper.

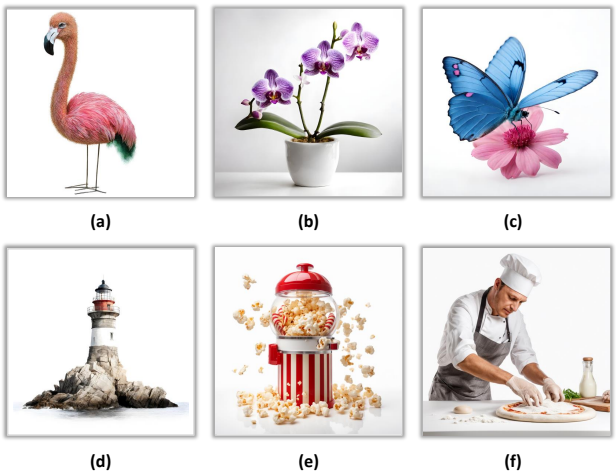

Figure 2. Visualization of visual prompts used in Figure 3 of main paper. The corresponding prompts are (a) "A fuzzy pink flamingo lawn ornament", (b) "A blooming potted orchid with purple flowers", (c) "A blue butterfly on a pink flower", (d) "A lighthouse on a rocky shore", (e) "Hot popcorn jump out from the red striped popcorn maker", (f) "A chef is making pizza dough in the kitchen".

3. Comparisons with Image-to-3D Methods

Recall that our VP3D generally decomposes the typical text-to-3D process into two cascaded stages: first text-to-image generation, and then (text plus image)-to-3D generation. Notably, during the (text plus image)-to-3D generation stage, our target is different from conventional image-to-3D task that precisely converts the input image (i.e., visual prompt in our context) into 3D content. Instead, we exploit the visual prompt as additional guidance in conjunction with the input text prompt to jointly supervise score distillation sampling (SDS) optimization of the underlying 3D model.

It is worthy noting that a degraded cascaded solution of text-to-3D generation is to simply integrate existing text-to-image and image-to-3D approaches. To further evaluate

the effectiveness of our VP3D for text-to-3D generation, we additionally include two state-of-the-art image-to-3D methods (i.e., One-2-3-45 [3] and [6]) for comparison. Specifically, One-2-3-45 first uses a view-conditioned 2D diffusion model (Zero-1-to-3 [4]) to generate multi-view images for the reference image and then reconstructs a 3D model from these multi-view images via an SDF-based generalizable neural surface reconstruction model. Magic123 also leverages Zero-1-to-3 as 3D prior and simultaneously utilizes 2D diffusion model for optimization. For fair comparison, here we feed the same visual prompt of VP3D into these two image-to-3D approaches for generating 3D assets.

Figure 3 and Figure 4 showcase the comparison results. It is easy to see that both One-2-3-45 and Magic123 generate distorted geometry and unrealistic textures when feeding intricate text/visual prompts (e.g., "A chef is making pizza dough in the kitchen"). In contrast, our VP3D manages to generate much better 3D scenes in terms of both geometry and texture, which again validates the effectiveness of our proposed VP3D.

References

[1] Yang Chen, Yingwei Pan, Haibo Yang, Ting Yao, and Tao Mei. Vp3d: Unleashing 2d visual prompt for text-to-3d generation. In *CVPR*, 2024. 1

[2] 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 *CVPR*, 2023. 1

[3] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 2, 3

[4] 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 *ICCV*, 2023. 2

[5] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1

[6] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *arXiv preprint arXiv:2306.17843*, 2023. 2, 3

Figure 3. Comparisons with image-to-3D methods: One-2-3-45 [3] and Magic123 [6]. (Example 1/2)

Figure 4. Comparisons with image-to-3D methods: One-2-3-45 [3] and Magic123 [6]. (Example 2/2)