

Supplementary Material for WonderTurbo: Generating Interactive 3D World in 0.72 Seconds

Chaojun Ni^{1,2*} Xiaofeng Wang^{2*} Zheng Zhu^{2*†} Weijie Wang^{2,3*} Haoyun Li^{2,4}
Guosheng Zhao^{2,4} Jie Li² Wenkang Qin² Guan Huang² Wenjun Mei^{1†}
¹Peking University ²GigaAI ³Zhejiang University
⁴Institute of Automation, Chinese Academy of Sciences

In the supplementary material, we provide further implementation details. We present additional qualitative results, where we use panoramic camera path maps to automate the generation process.

1. Implementation Details

Time cost evaluation. To effectively evaluate the generated scenes, we compare the time required to generate scenes of the same size. Specifically, for offline methods, we follow the setups of LucidDreamer [1] and Text2Room [3], first generating multiple images of new scenes and then converting them into 3D scenes according to their respective methods. For DreamScene360 [7] and Text2Room [3], we use the Diffusion360 [2] model to generate panorama images of the input images, which are then lifted to 3D. We compute the total time for this process and compute the time cost of generating the test scenes based on the size of the generated and test scenes. For online methods, we directly calculate the time required to generate a new scene.

Metrics. To evaluate the quality of 3D interactive scene generation, we utilize several metrics: CLIP scores (CS) [4], CLIP consistency (CC) [4], CLIP-IQA+ (CIQA) [5], Q-Align [6], and CLIP aesthetic scores (CA) [4]. These metrics not only assess the quality of appearance modeling but also evaluate the quality of geometry modeling, as inaccurate geometry can lead to severe distortions when rendering novel views, which can significantly impact metrics such as CLIP-IQA+ (CIQA), Q-Align, and CLIP aesthetic scores (CA). The CLIP score (CS) measures the relevance between the scene prompt and the rendered image by computing the cosine similarity between their respective CLIP embeddings. CLIP consistency (CC) is assessed by measuring the cosine similarity between the CLIP embeddings of each novel view and the central view, ensuring semantic consistency across views. CLIP-IQA+ (CIQA) is an enhanced image quality metric that combines perceptual quality models with deep learning techniques to evaluate attributes. Finally, the CLIP aesthetic score (CA)

captures the aesthetic quality of the image, considering elements like composition, contrast, and color harmony.

2. Qualitative results.

As shown in Fig. 1 and Fig. 2, we provide additional scenes to demonstrate the superiority of *WonderTurbo*.

References

- [1] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 1
- [2] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023. 1
- [3] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 1
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1
- [5] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 1
- [6] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching llms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 1
- [7] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suyu You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision*, pages 324–342. Springer, 2024. 1



Figure 1. Qualitative examples.



Figure 2. Qualitative examples.