
Video Perception Models for 3D Scene Synthesis

Supplementary Material

Anonymous Author(s)

Affiliation

Address

email

Appendix

A Additional Qualitative Results

We provide additional qualitative results in the supplementary video, including experiments on scene synthesis from multi-modal inputs, text-based scene synthesis. Please refer to the video for details.

B Implementation Details

Rescaling the Scene To refine the overall scale of the reconstructed scene, we estimate depth maps for each view using UniDepth [7], and compare them against the corresponding reconstructed point clouds. For each point, we compute the ratio between the estimated depth and the depth derived from the reconstructed geometry. The global scale factor is then determined as the median of these per-point ratios across all views, providing a robust estimate that mitigates the influence of outliers. This median-based scaling approach ensures consistency across views and improves the alignment of the reconstructed scene with real-world metric dimensions.

Orientation Estimation of the Object Without loss of generality, we assume that object bounding boxes are aligned with the ground plane. To estimate their orientation, we first determine the ground plane equation. This process begins by extracting the ground point cloud, using a method analogous to that employed for object extraction. Specifically, we prompt Grounded-SAM [8] with the label “ground” for outdoor scenes or “floor” for indoor scenes to generate ground masks. These masks are then used to extract the corresponding ground points from the reconstructed scene, and a least squares fitting is applied to estimate the ground plane. With the ground plane established, each object’s point cloud P_i is transformed into a new coordinate system that retains the origin of the original camera coordinate system C , but aligns its horizontal plane with the estimated ground plane. The transformed point cloud is then projected onto the ground plane, and Principal Component Analysis (PCA) is applied to identify the principal axes of the point distribution. The direction of greatest variance is taken as an approximation of the object’s orientation θ_i . A tight bounding box is subsequently aligned with this estimated direction.

C Prompts Details

Prompts for Scene Synthesis. We utilize GPT-4o [5] to generate text prompts for four types of indoor scenes: living room, bedroom, kitchen, and bathroom. Each prompt specifies the room type along with the objects intended to furnish the space. For example, “A bedroom with a large bed, two nightstands, a floor lamp, a wardrobe, and a big window.”

I’m working on an interior design project and would like to generate video scenes of a {room type} using a text-to-video model. Please help me create detailed prompts to feed into the model.

Guidelines:

1. Based on the typical function and layout of a {room type}, list the furniture, appliances, decorations, and other items commonly found in the space. 2. Prompts should describe the room’s contents clearly and in detail.

Example: “A bedroom with a large bed, two nightstands, a floor lamp, a wardrobe, and a big window.”

32

33 **Prompts for FPVSCORE.** To facilitate consistent and goal-driven evaluations in FPVSCORE, we
34 design structured prompts that include: (1) task-specific instructions for multi-scene comparison,
35 (2) clearly defined evaluation criteria, and (3) standardized formatting requirements. These prompts
36 guide the model to assess each scene in terms of semantic fidelity, spatial layout accuracy, and overall
37 coherence, while also requiring concise justifications to support its ratings and enhance transparency.

Task: Compare the room layout rationality of three methods, all generated from the same text description. From top to bottom, the video sequences display a 360-degree view of each method’s generated scene. Decide which method performs best according to the criteria below.

Text Description: {text_description}

Instructions:

1. Semantic Correctness

Does the generated layout accurately reflect the text description?

Check whether all described objects are present and correctly represented.

2. Layout Correctness

Is the room design physically plausible and functional?

Evaluate if the layout supports practical use, space efficiency, and proper object functionality.

Consider object positions, orientations, and user convenience.

3. Overall Preference

Does the room layout look realistic and natural?

Consider the visual coherence and harmony of the scene.

Evaluation process:

Carefully examine the multi-view images of all three 3D scenes. Focus on one criterion at a time and make independent judgments for each.

Output format:

Provide a clear, concise analysis for each criterion. Avoid vague terms like “realistic,” or “spacious.” Instead, specify exact issues or strengths. For example:

- For Semantic Correctness, indicate which objects are missing or inaccurately depicted.

- For Layout Correctness, specify which objects are misplaced or poorly oriented, and explain how this impacts usability or functionality.

After the analyses, assign ranks (1–3) to each method per criterion (1 = best, 3 = worst).

Summarize your final ranking in the format: <rank for criterion 1> <rank for criterion 2> <rank for criterion 3>

for each method.

Example:

Analysis:

1. Semantic Correctness: The first one ...; The second one ...; The third one ...

2. Layout Correctness: The first one ...; The second one ...; The third one ...

3. Overall Preference: The first one ...; The second one ...; The third one ...

Final answer:

The first one: x x x

The second one: x x x

The third one: x x x

(where x denotes ranks 1–3)

(Please strictly follow the format above. Do not include extra symbols like **, quotation marks, or bullet points.)

38

39 **Prompts for Top-Down View Scores.** Following the approach of Architect [10], we design targeted
 40 prompts to guide GPT-4o in evaluating room layouts based solely on top-down views. To ensure
 41 a fair comparison, the prompts also emphasize spatial structure, semantic fidelity, and functional
 42 usability, consistent with our own.

Task: Compare the room layout rationality of three methods, all generated from the same text description. The top-down views of the scenes produced by the three methods are presented from left to right. Identify which method performs best based on the criteria below.
 Text Description: {text_description}

Instructions:

1. Semantic Correctness

Does the generated layout accurately reflect the text description?

Check whether all described objects are present and correctly represented.

2. Layout Correctness

Is the room design physically plausible and functional?

Evaluate if the layout supports practical use, space efficiency, and proper object functionality.

Consider object positions, orientations, and user convenience.

3. Overall Preference

Does the room layout look realistic and natural?

Consider the visual coherence and harmony of the scene.

Provide only your final ranking of the three methods in the format below:

Final answer:

x x x

(where x denotes ranks from 1 to 3)

43

44 **D User Study Details**

45 We conducted a thorough user study to evaluate the quality of the generated scenes, involving 30
 46 participants, including both undergraduate and graduate students. All participants took part voluntarily
 47 and received no compensation. At the start of the study, participants were given five minutes to
 48 read through the instructions, as illustrated in Fig. 1. An example evaluation page presented to the
 49 participants is shown in Fig. 2.

50 **E Limitations**

51 Currently, VIPSCENE focuses on generating semantically coherent scene layouts rather than modeling
 52 the fine-grained details of individual objects. Furniture and other elements are directly retrieved
 53 from the Objaverse [2] dataset. Although these objects are richly annotated, some textures still lack
 54 photorealistic quality. In future work, we plan to improve object quality in two complementary
 55 directions. First, we will adopt advanced 3D object generation techniques such as text-to-3D and
 56 image-to-3D methods [11, 12, 14, 16] to produce higher-quality assets. Second, we will incorporate
 57 state-of-the-art physically-based rendering (PBR) techniques [1, 3, 4, 15] to enable realistic material
 58 representations and lighting interactions, making the objects appear more photorealistic. These
 59 improvements aim to further enhance both the diversity and realism of the generated scenes.

60 **F Broader Impact**

61 Our work facilitates more accessible and coherent 3D scene generation from multi-modal inputs, with
 62 potential applications in virtual reality (VR), augmented reality (AR), education, robotics simulation,
 63 and digital content creation. By lowering the barrier to structured 3D environment design, it enables
 64 broader participation from creators with limited technical or design expertise, allowing them to
 65 generate complex scenes using intuitive prompts. Additionally, our human-aligned evaluation metric
 66 offers a valuable benchmark for assessing semantic fidelity and spatial consistency, promoting more
 67 rigorous and meaningful comparisons in the field. Although our approach does not explicitly amplify
 68 dataset biases, it inherits common limitations of machine learning systems and may still reflect
 69 underlying biases in the data. To support responsible deployment, future work should incorporate
 70 fairness-aware data practices, model auditing, and content provenance tracking.

Text-to-Room Layout Algorithm Evaluation

In this study, we invite you to evaluate three different text-to-room layout algorithms. You will see a series of room layout images and videos generated by different algorithms and compare them based on the following three metrics:

Prompt Adherence (PA)

"To what extent does the generated scene align with the input prompt?"
Check if the generated content represents the room description.

Layout Correctness (LC)

"Do the object placements make sense both physically and functionally?"
Evaluate if the layout meets practical requirements, optimizes space usage, and ensures objects can be used as intended.

Overall Preference (OP)

"Does this solution resemble a real scene?"
Rate your overall satisfaction with the layout design. Does the room layout appear realistic and natural?

Evaluation Process

Click the "Start Evaluation" button to enter the main evaluation page.
Please read the background description and the text to be depicted at the top of each page. Then, based on your assessment of the three images and corresponding videos below, rate the different metrics. The higher the score, the more it meets the requirements.
Use the "Previous" and "Next" buttons to continue evaluating the next set of samples.
On the summary page, click the "Download Results" button to download a JSON file and send it to us.

Key Tips

We strongly recommend using a computer as the terminal, as you can zoom the page for the best experience.
This webpage supports resuming the evaluation from where you left off. If you accidentally close the page, don't worry, just reopen it in the same browser to continue.
The order of scenes generated by different methods has been randomized.
Thank you very much for your support of this experiment!

[Start Evaluation](#)


Figure 1: **User Study Instructions.** This page was shown to participants at the beginning of the study to explain the task, interface, and evaluation criteria.

71 G Licenses


Table 1: Licenses of assets used.

Asset	License
Cosmos [6]	Apache 2 License
MASt3R [9]	CC BY-NC-SA 4.0
Fast3R [13]	FAIR Noncommercial Research License
Grounded-SAM [8]	Apache License 2.0
Unidepth [7]	CC-BY-NC 4.0
GPT-4o [5]	OpenAI Terms of Use
Gemini [9]	Google Terms of Service

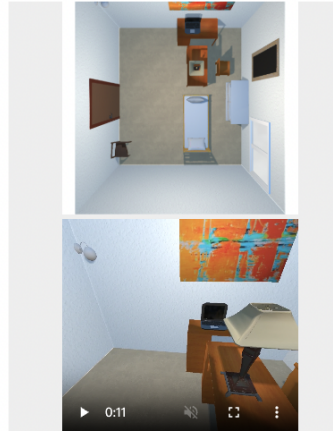
A bedroom with a bed, wardrobe, chair, dressing table, and armchair.



Method 1



Method 2



Method 3

Prompt Adherence: To what extent does the generated scene align with the input prompt?

Method 1

Method 2

Method 3

Layout Correctness: Is the room design both physically plausible and functional?

Method 1

Method 2

Method 3

Overall Preference: Does the room layout appear realistic and natural?

Method 1

Method 2

Method 3

Figure 2: **Example Page.** Participants were shown a 360-degree video captured from the center of each scene, along with a top-down rendered image. This setup allowed them to evaluate both the global structure and fine details. Each scene was rated on a 3-point scale (1 = lowest, 3 = highest) across three criteria.

References

- [1] Zilong Chen, Yikai Wang, Wenqiang Sun, Feng Wang, Yiwen Chen, and Huaping Liu. Meshgen: Generating pbr textured mesh with render-enhanced auto-encoder and generative data augmentation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [2] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023.
- [3] Zebin He, Mingxin Yang, Shuhui Yang, Yixuan Tang, Tao Wang, Kaihao Zhang, Guanying Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, et al. Materialmvp: Illumination-invariant material generation via multi-view pbr diffusion. *arXiv preprint arXiv:2503.10289*, 2025.
- [4] Xin Huang, Tengfei Wang, Ziwei Liu, and Qing Wang. Material anything: Generating materials for any 3d object via diffusion. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [5] 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.
- [6] Nvidia. Cosmos, 2024. <https://www.nvidia.com/en-us/ai/cosmos/>.
- [7] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024.
- [8] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.
- [9] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [10] Yian Wang, Xiaowen Qiu, Jiageng Liu, Zhehuan Chen, Jiting Cai, Yufei Wang, Tsun-Hsuan Johnson Wang, Zhou Xian, and Chuang Gan. Architect: Generating vivid and interactive 3d scenes with hierarchical 2d inpainting. *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- [11] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025.
- [12] 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. *arXiv preprint arXiv:2412.01506*, 2024.
- [13] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. 2025.
- [14] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024.
- [15] Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al. Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models. *ACM Transactions on Graphics (TOG)*, 2024.
- [16] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.