# SPATIALGEN: Layout-guided 3D Indoor Scene Generation

## Supplementary Material

In the appendix, we provide more details of SPATIALGEN dataset in Section 1, additional experimental results in Section 2, and ablation studies in Section 3.

## 1. SPATIALGEN Dataset

### 1.1. Dataset Construction

**Data Curation.** Our dataset is sourced from an online platform in the interior design industry, providing a large-scale collection of professional designs intended for real-world applications. We employ a rigorous multi-stage filtering pipeline to ensure both the quality and diversity of the dataset.

The curation process begins by selecting scenes based on four key criteria: (i) professional designer ratings, (ii) the number of renderings generated by the design, (iii) a total floor area exceeding $20\text{m}^2$, and (iv) the presence of more than 35 unique objects.

Then, we extract individual rooms from each selected scene and apply additional filters to retain only those rooms that (i) have a floor area greater than $8\text{m}^2$ and (ii) contain more than 3 unique objects.

For rendering, we use an industry-leading rendering engine to generate images. We simulate physically plausible camera trajectories that navigate smoothly within each room while avoiding obstacles. After rendering, we implement strict quality control measures by discarding low-quality images—specifically those with camera-object intersections, overexposure, or inadequate lighting, as illustrated in Figure 2.

The final dataset consists of 12,328 distinct scenes, 57,431 individual rooms covering a variety of room types, and 4.7M photo-realistic panoramic renderings. The total floor area across all scenes is approximately $914,687\text{m}^2$.

**Camera configuration.** We capture panoramic renderings at intervals of $0.5\text{m}$ to ensure comprehensive scene coverage, as shown in the top-left of Figure 1. Each panoramic rendering is generated at a resolution of $1024 \times 2048$ and includes color, albedo, depth, normal, semantic, and instance maps. The entire rendering process requires approximately 54K GPU hours.

Following an obstacle-avoiding camera trajectory within each room, we obtain dense sequences of panoramic images. Thanks to the $360^\circ$ field-of-view (FoV) of panoramas, we can simulate an unlimited number of perspective images with varying camera configurations. For each panoramic viewpoint, we generate perspective views with different fields-of-view and rotation angles using equirectangular-to-perspective projection [4], as illustrated in Figure 3.

Furthermore, we introduce four distinct camera trajectories with varying amounts of view overlap and distances between input and target views: (i) *Forward*: a linear path with minimal directional variation, simulating steady camera movement; (ii) *Inward Orbit*: both input and output views are oriented toward the center of the room, ensuring significant view overlap; (iii) *Outward Orbit*: the input and output views share the same location but have different orientations, resulting in less than $45^\circ$ overlap between adjacent views; and (iv) *Random Walk*: input and output views are sampled along a continuous random-walk path, with minimal view overlap.

### 1.2. Dataset Statistics

**Room type statistics.** The resulting dataset contains 12,592 living and dining rooms, 2,179 living rooms, 2,524 study rooms, 8,540 kitchens, 8,460 bathrooms, 1,464 balconies, 9,049 master bedrooms, 8,603 secondary bedrooms, 2,793 children's rooms, and 4,418 other room types, as illustrated in Figure 4 representing a diverse and substantial collection of indoor environments.

**Object category statistics.** The raw online designs initially contained approximately 65,000 object categories. We filtered out niche object classes specific to interior design and mapped the remaining objects to 62 common categories from ADE20K [12]. We then curated the object bounding boxes according to the following criteria: (i) objects outside the room layout were discarded; (ii) objects with any edge shorter than $0.1\text{m}$ or longer than $1.8\text{m}$ were excluded. This process yielded a total of 1,046,637 object bounding boxes. Figure 5 shows the distribution of object categories throughout our dataset, excluding the spotlight and other categories (containing 250K and 240K instances, respectively) to improve visualization of the remaining categories.

### 1.3. Dataset Visualization

As shown in Figure 1, our dataset provides high-quality panoramic renderings accompanied by precise 2D annotations and comprehensive 3D structural layouts, including architecture elements (*e.g.*, walls, windows, and doors), which distinguishes it from existing datasets like Hypersim [7], offering extensive evaluation opportunities for scene generation and spatial understanding tasks.

## 2. Additional Experiments and Results

In this section, we show more results of text-to-3D scene generation in Section 2.1, and conduct comprehensive experiments of image-to-3D scene generation in Section 2.2.

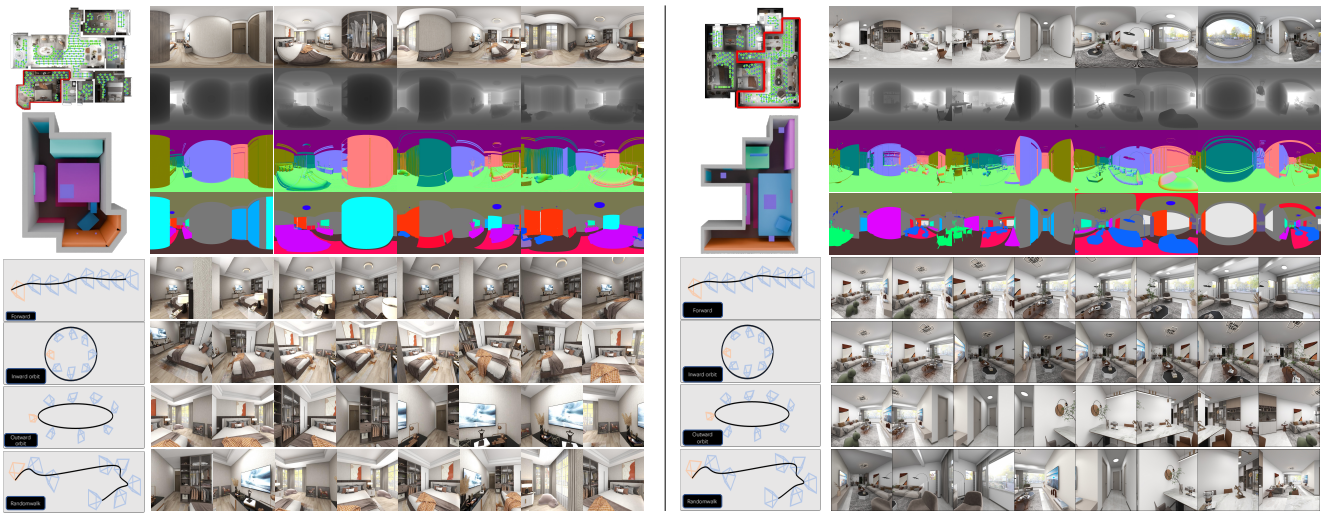

Figure 1. Examples of SPATIALGEN dataset.


Figure 2. Example of low-quality renderings.

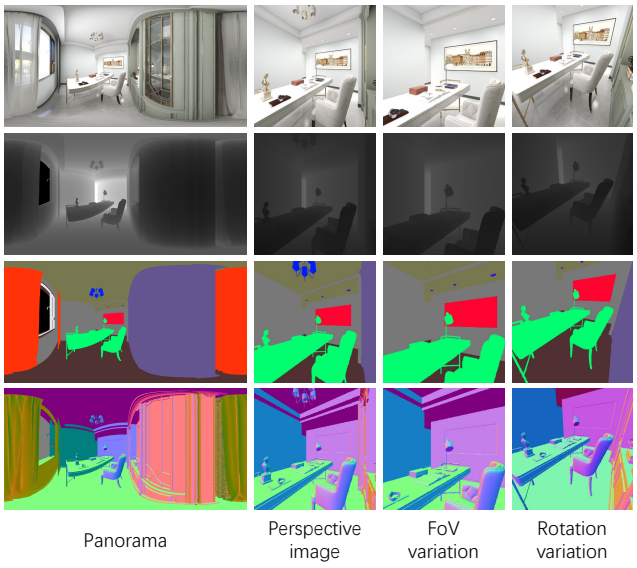


Figure 3. Camera configuration.

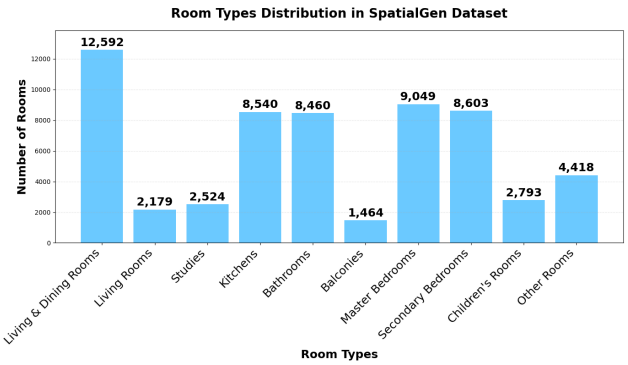


Figure 4. Room type distribution.

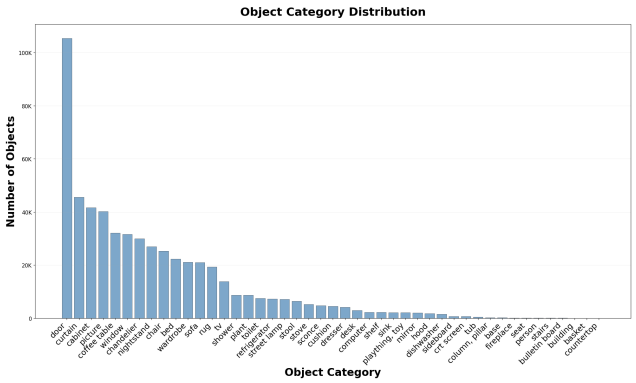


Figure 5. Object category distribution.

Furthermore, we show generation results from common videos (unposed) captured in indoor scenes in Section 2.3.

## 2.1. Text to 3D Scene Generation

As demonstrated in Figure 6, our method outperforms SDS-based baselines on the Hypersim dataset [7], producing photorealistic and layout-faithful scenes with superior detail. This advantage is evident even against our model trained only on Hypersim (SPATIALGEN$^{\dagger}$), which produces blurring and ambiguous results, highlighting the benefits of

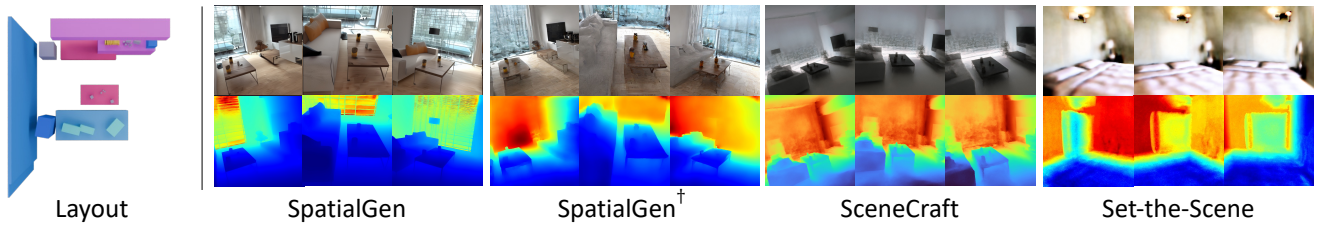


Figure 6. Qualitative comparison of text-to-3D scene on Hypersim [7] dataset. In each case, we show the generated color images and depth map.

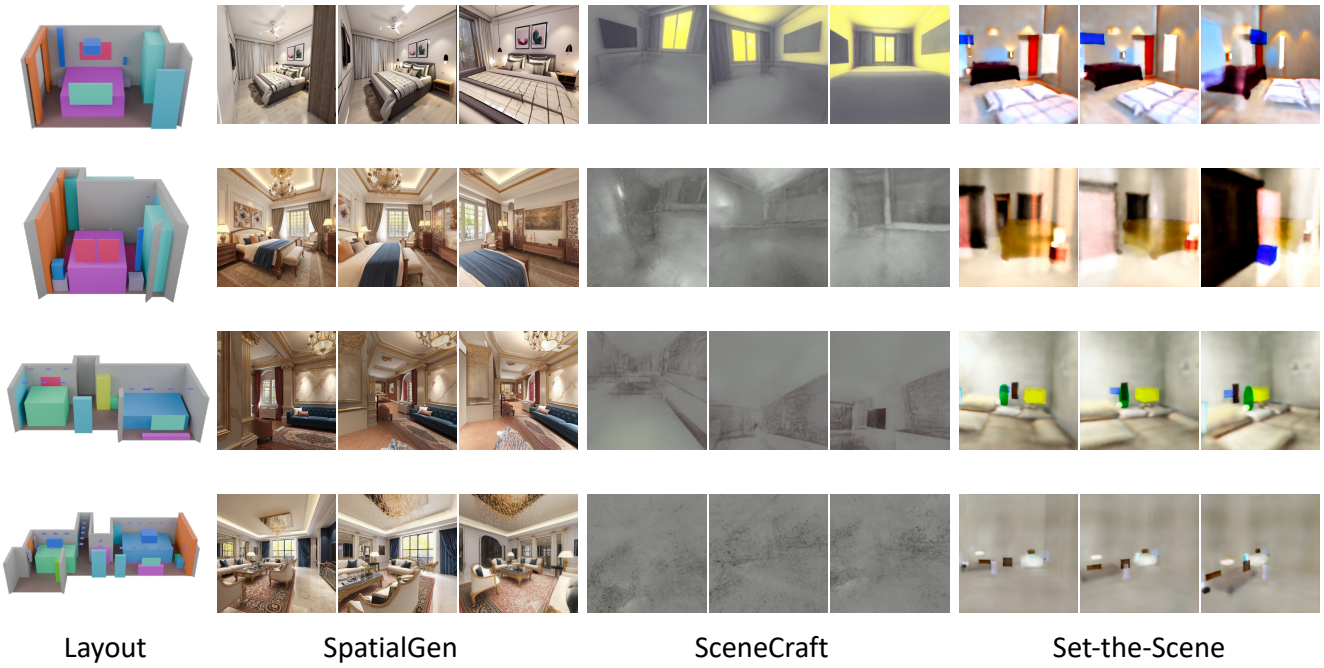


Figure 7. Qualitative comparison of text-to-3D scene on SPATIALGEN dataset. In each case, we show the generated color images.

our dataset.

In Figure 7, we compare against SceneCraft [9] and Set-The-Scene [1] under diverse 3D layouts on SPATIALGEN dataset. As the increase of layout complexity, while competing methods fail to generate meaningful radiance fields or capture scene details, our approach consistently delivers realistic and coherent results for complex scenes like living and dining rooms.

We further compare our method against the panorama-as-proxy based method, Ctrl-Room [2], on both the Structured3D [11] and SPATIALGEN dataset. We split the panoramic image into 8 perspective images for a direct comparison, as shown in Figure 8.

For the SPATIALGEN dataset, we render a layout-semantic panorama from a random viewpoint to use as input for Ctrl-Room. We then spatially align its resulting mesh with our generated scene for a fair comparison. The results, presented in Figure 9, demonstrate that Ctrl-Room exhibits severe stretching artifacts and scale misalignment at novel viewpoints. In contrast, our method consistently produces photorealistic and fully 3D-consistent renderings from all views.

## 2.2. Image to 3D Scene Generation

In this section, we conduct additional image-to-3D scene generation experiments with a focus on two key aspects: (i) generation capability – the ability to synthesize missing regions for large viewpoint changes; (ii) semantic consistency – the ability to produce semantically consistent views aligned with the 3D scene layout. Given the lack of accessible literature on the layout-conditioned image-to-3D scene generation task, we compare our multi-view generation model against the version without utilizing layout priors. We employ PSNR, SSIM [8], LPIPS [10], and FID [5]


Figure 8. Qualitative comparison with Ctrl-Room on Structured3D for panorama generation. We split the panorama into eight perspective images for a direct comparison. Our method achieves competitive RGB synthesis compared with Ctrl-Room, resulting in photo-realistic scenes that are well-aligned with the provided layout.

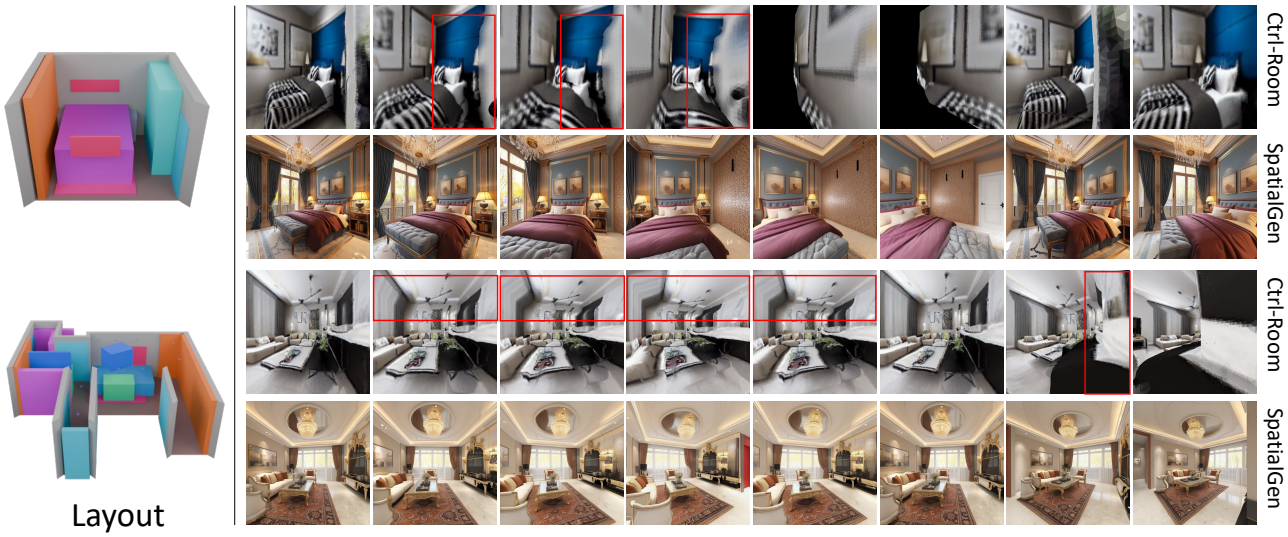


Figure 9. Qualitative comparison with Ctrl-Room on SPATIALGEN dataset. Ctrl-Room exhibits severe stretching artifacts and scale misalignments at novel viewpoints. In contrast, our method consistently produces photorealistic and fully 3D-consistent renderings from all views.

to evaluate the quality of image generation.

**Quantitative Results.** Table 1 reports quantitative results of our method under different camera trajectories. Under all trajectories, the semantic layout improves the results across all metrics. Furthermore, the improved FID shows that our method with layout guidance can capture the underlying data distribution more effectively. These results collectively underscore the critical role of incorporating 3D layout information in novel view synthesis.

**Qualitative Results.** Figure 10 shows example outputs of our method, including RGB images, scene coordinate maps, and semantic maps. Removing the layout input leads to severe artifacts in occluded regions, revealing the limitations of image diffusion models in capturing 3D scene structures. In addition, the semantic map contains unknown content, suggesting degraded semantic prediction without layout input. In contrast, our method with layout guidance generates better novel view images and achieves more reasonable se-

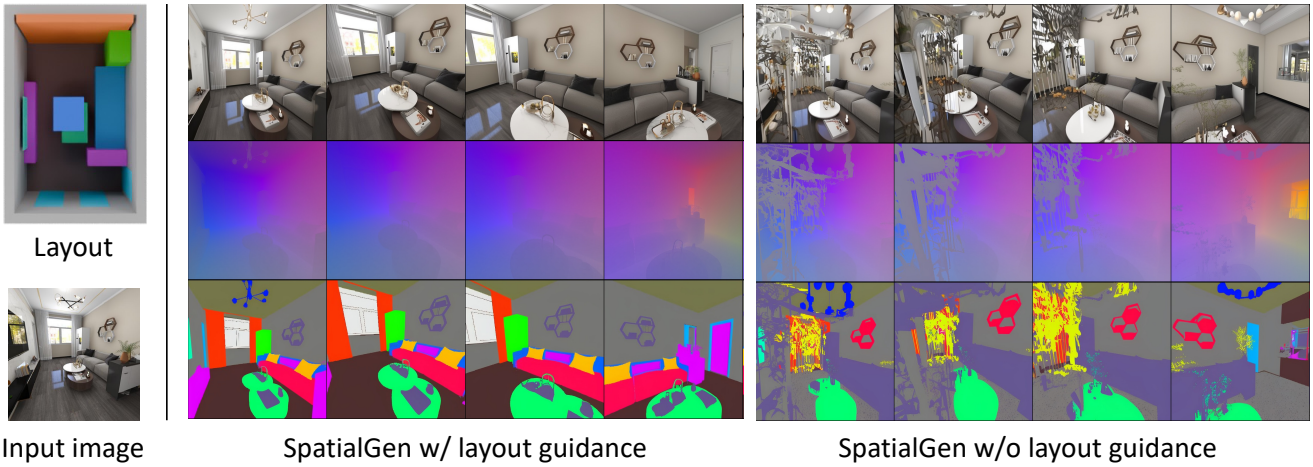


Figure 10. Qualitative comparison of image-to-3D scene generation on our dataset. Given a single input image, our method with layout guidance consistently generates better color images, scene coordinate maps, and semantic maps.

Table 1. Experimental results on image-to-3D scene generation under four distinct camera trajectories: *Forward*, *Inward Orbit*, *Outward Orbit*, and *Random Walk*, with gradually reduced view overlaps.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| *Forward* | | | | |
| SPATIALGEN (w/o layout) | 11.47 | 0.49 | 0.59 | 67.96 |
| SPATIALGEN (w/ layout) | **17.59** | **0.69** | **0.32** | **34.98** |
| *Inward orbit* | | | | |
| SPATIALGEN (w/o layout) | 12.57 | 0.48 | 0.54 | 64.14 |
| SPATIALGEN (w/ layout) | **17.30** | **0.66** | **0.33** | **35.57** |
| *Outward orbit* | | | | |
| SPATIALGEN (w/o layout) | 11.14 | 0.60 | 0.47 | 76.73 |
| SPATIALGEN (w/ layout) | **13.32** | **0.59** | **0.46** | **57.76** |
| *Random walk* | | | | |
| SPATIALGEN (w/o layout) | 11.26 | 0.45 | 0.59 | 98.42 |
| SPATIALGEN (w/ layout) | **14.07** | **0.62** | **0.45** | **52.10** |

mantic and geometric predictions.

In Figure 11, given a reference image (highlighted in orange box) across four diverse camera trajectories , our method successfully generates 3D-consistent novel views and synthesizes semantically plausible content for areas beyond the original input view.

### 2.3. Video-to-3D Scene Generation

Given the fact that a well-defined 3D layout is not easy to obtain, we apply SPATIALGEN to the challenging task of generating novel 3D scenes from videos. By leveraging a state-of-the-art layout estimation model, SpatialLM [6], we get the reconstructed 3D layout from the video. Then, we perform text-to-3D scene generation conditioned on this layout and additional user-provided text prompts. This approach allows us to generate entirely new scenes that preserve the structural layout of the original video while altering its stylistic and semantic content based on the text description. We validate this video-to-new-scenes application on the SpatialLM test set. Figure 16 shows qualitative results.

### 2.4. Failure Cases

Because SPATIALGEN relies on an initial RGB image, if the images happen to be inconsistent with the given 3D layout, then it fails to generate consistent results. In this example of text-to-3D scene generation in Figure 12, we first convert the provided layout and textual description into an initial RGB image. However, the initialization process yields an RGB output (highlighted in red) that does not align with the given layout—specifically, the orientation of the bed is entirely reversed compared to the ground truth. Consequently, during the subsequent generation stage, the resulting outputs exhibit inconsistencies and fail to align properly with both the layout and the initial RGB image.

## 3. Ablation Studies

In this section, we conduct additional ablation studies to verify the layout guidance, design choice of network architecture, and the number of input views.

**Ablation on layout guidance.** We first study the effect of layout guidance to validate our design. We compare our full model (denoted as *W/ Layout*) against a variant that removes layout priors (denoted as *W/O Layout*). The *W/O Layout* variant is implemented similarly to CAT3D [3] but incorporates our multi-view multi-modal alternating attention module to enable multi-modal output. Both models are trained identically for single-image 3D scene generation.

Figure 14 presents a faithful comparison, showing generated RGB outputs, scene coordinate maps, and seman-

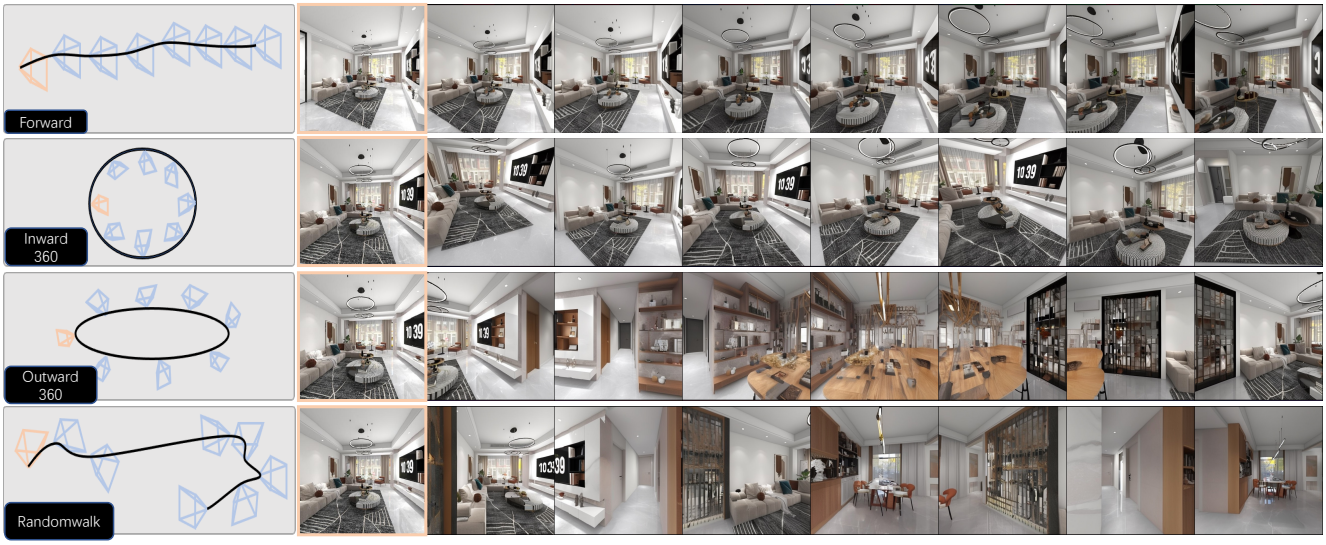


Figure 11. Qualitative results on SPATIALGEN dataset under various camera trajectories. From left to right: input view and target views. *First Row (forward)*: sampled views follow a progressive forward-moving path. *Second row (inward orbit)*: views are directed toward the center of the room, ensuring substantial overlap between them. *Third row (outward orbit)*: views are positioned at the center of the room, looking outward, with an angle of less than $45^\circ$ between two adjacent views. *Bottom (random walk)*: views are selected from a continuous random-walk camera trajectory, producing aggressive viewpoint changes.

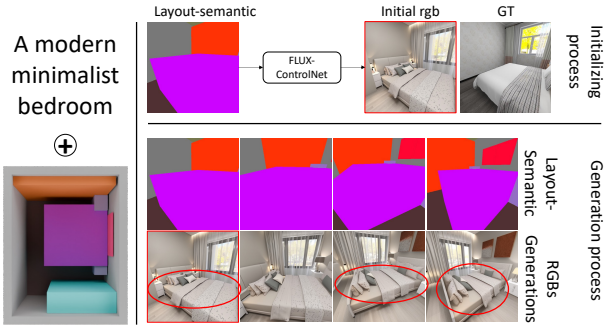


Figure 12. Failure case analysis. The initial RGB image, generated from a layout and text, misorients the bed (red, vs. ground truth). This initial error causes subsequent generations to conflict with the layout.

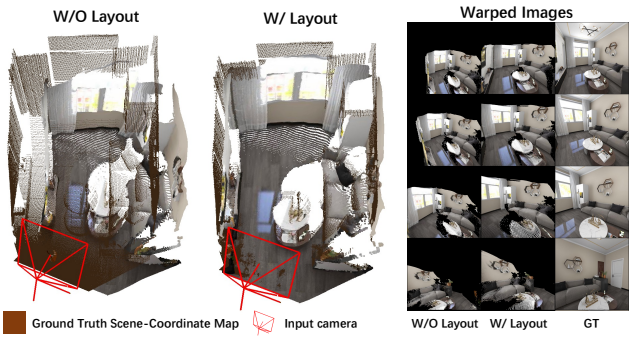


Figure 13. Comparing geometric prediction quality between our *(W/ layout)* and *(W/O layout)*. The first two columns show the predicted scene coordinate maps, where our method (W/ layout) achieves better alignment with ground-truth geometry (brown color point cloud) compared to the counterpart without layout guidance (W/O layout). Correspondingly, the warped images projected by the predicted scene coordinates demonstrate improved spatial consistency and reduced artifacts.

tic maps (top to bottom) from a given input (left-most column). As the red circles highlight, the *W/O Layout* variant produces artifacts in occluded regions, exhibits imperfect image-pose alignment, and generates degraded dense predictions. These failures indicate the inherent limitations of relying solely on image diffusion priors for 3D scene generation. In contrast, our full model *W/ Layout* leverages explicit layout guidance to achieve superior novel-view synthesis and more accurate geometry and semantic predictions.

Furthermore, Figure 13 provides an in-depth analysis of 3D consistency. We visualize the predicted scene coordinates for the input view, demonstrating that the *W/ layout* predictions achieve better alignment with the ground truth. This superior alignment provides more accurate warped images for all target viewpoints, explaining its clear superiority over the *W/O layout* baseline.

**Ablation on Alternating Attention mechanism.** We further evaluate the design choice of Multi-view Multi-modal Alternating Attention (AA). We compare our full model (denoted as *W/ AA*) against a variant that disables the multi-

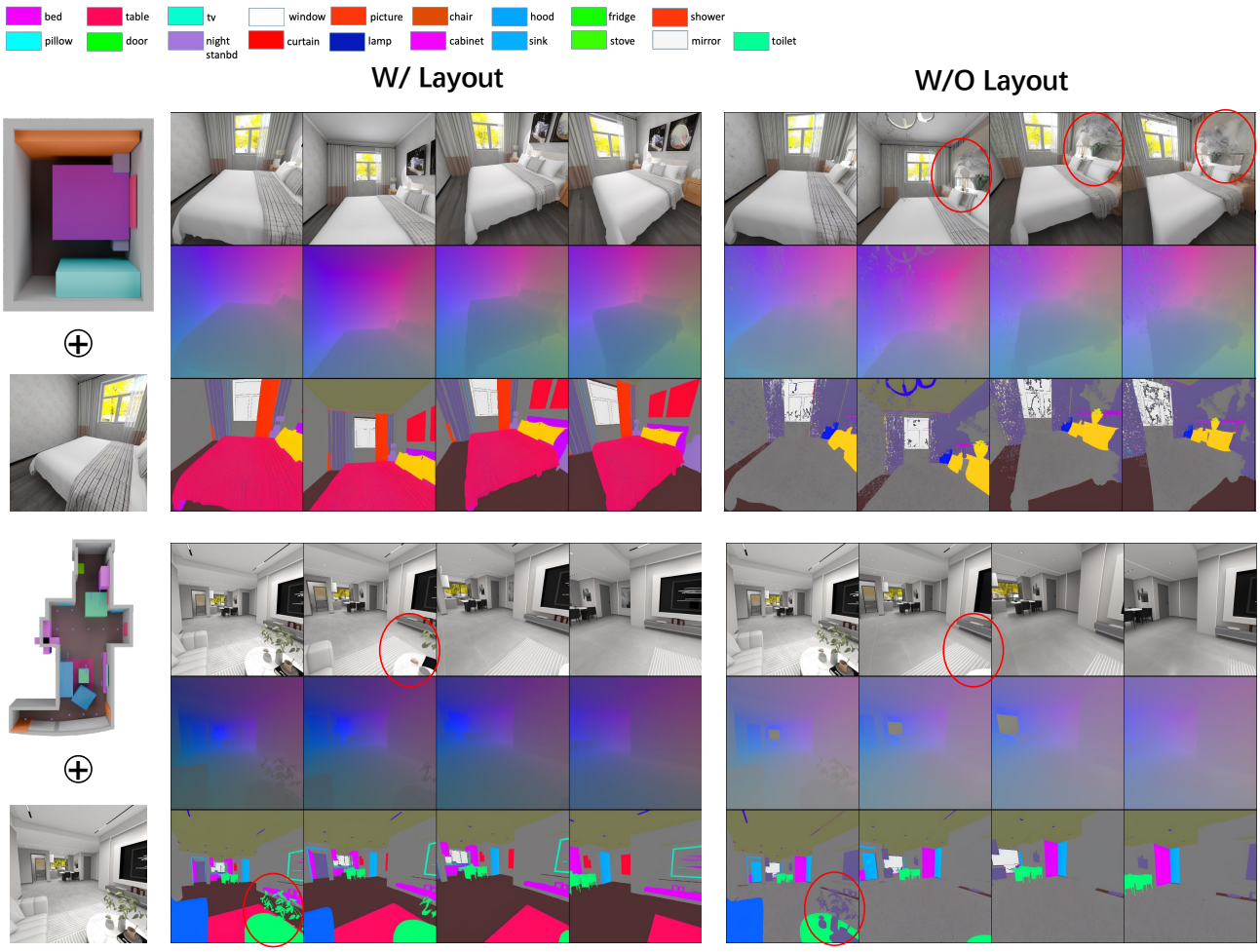


Figure 14. Ablation on the effectiveness of layout as guidance.

modal attention (denoted as *W/O AA*). Both models share the same training protocol as the aforementioned.

Qualitative results in Figure 15 reveal a critical weakness in the ablated model. As indicated by the red circles, the *W/O AA* model fails to produce semantically meaningful segmentations. This deficiency corrupts the geometry in the scene coordinate maps and resulting in less coherent multi-view RGB generation. This demonstrates that without an explicit mechanism to align and refine information across modalities (RGB, geometry, semantics), the model cannot effectively leverage their synergies.

Our complete model **W/ AA** directly addresses this limitation. By facilitating cross-modal interaction, the AA mechanism enables more precise semantic labels and geometrically consistent scene coordinates. This improvement subsequently elevates the fidelity and view-consistency of the final RGB outputs, confirming that the alternating attention is pivotal for high-quality, multi-modal scene generation.

Table 2. Effect of the number of input views in the inward orbit setting.

| #input views | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| 1 | 17.30 | 0.66 | 0.33 | 35.57 |
| 3 | 17.83 | 0.67 | 0.31 | 28.72 |
| 6 | **18.33** | **0.67** | **0.31** | **21.93** |

**Ablation on number of input views.** In Table 2, we evaluated SPATIALGEN using different numbers of input views in the *inward orbit* camera configuration. Increasing the number of input views enhances all metrics, particularly the FID score; this implies that a greater input views enhances semantic consistency.

## References

[1] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-Scene: Global-local training

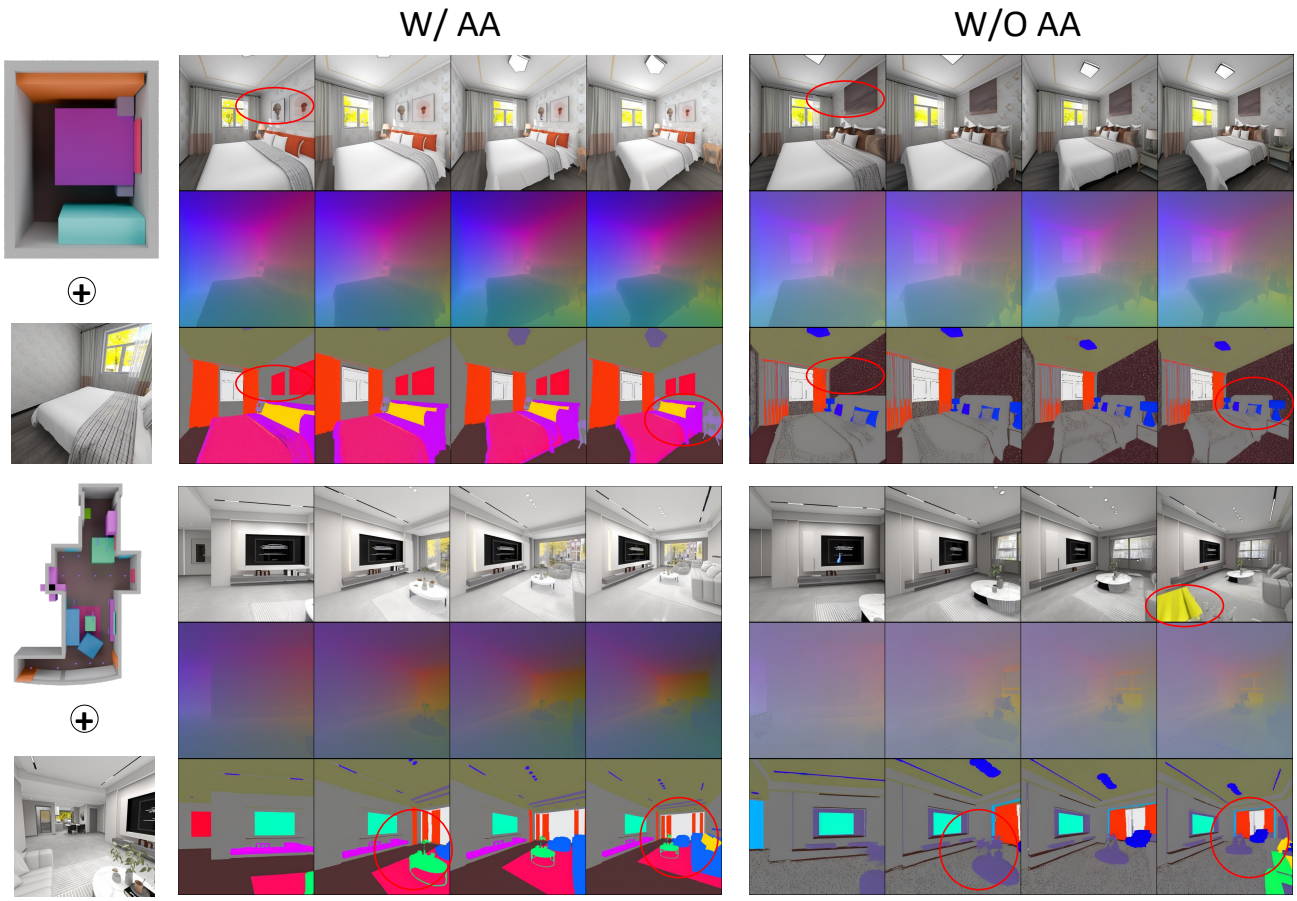


Figure 15. Ablation on the effectiveness of the Multi-view Multi-modal Alternating Attention mechanism.

for generating controllable nerf scenes. In *IEEE Int. Conf. Comput. Vis. Worksh.*, pages 2920–2929, 2023. 3

[2] Chuan Fang, Yuan Dong, Kunming Luo, Xiaotao Hu, Rakesh Shrestha, and Ping Tan. Ctrl-Room: controllable text-to-3d room meshes generation with layout constraints. In *IEEE Int. Conf. 3D Vis.*, 2025. 3

[3] Ruiqi Gao, Aleksander Hoł yński, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create anything in 3d with multi-view diffusion models. In *Adv. Neural Inform. Process. Syst.*, pages 75468–75494, 2024. 5

[4] haruishi43. Equilib, 2020. 1

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 3

[6] Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. SpatialLM: Training large language models for structured indoor modeling. In *Adv. Neural Inform. Process. Syst.*, 2025. 5, 9

[7] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *IEEE Int. Conf. Comput. Vis.*, pages 10912–10922, 2021. 1, 2, 3

[8] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 3

[9] Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. SceneCraft: Layout-guided 3d scene generation. *Adv. Neural Inform. Process. Syst.*, 37:82060–82084, 2024. 3

[10] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 3

[11] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3d modeling. In *Eur. Conf. Comput. Vis.*, pages 519–535, 2020. 3

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. ADE20K: Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. 1

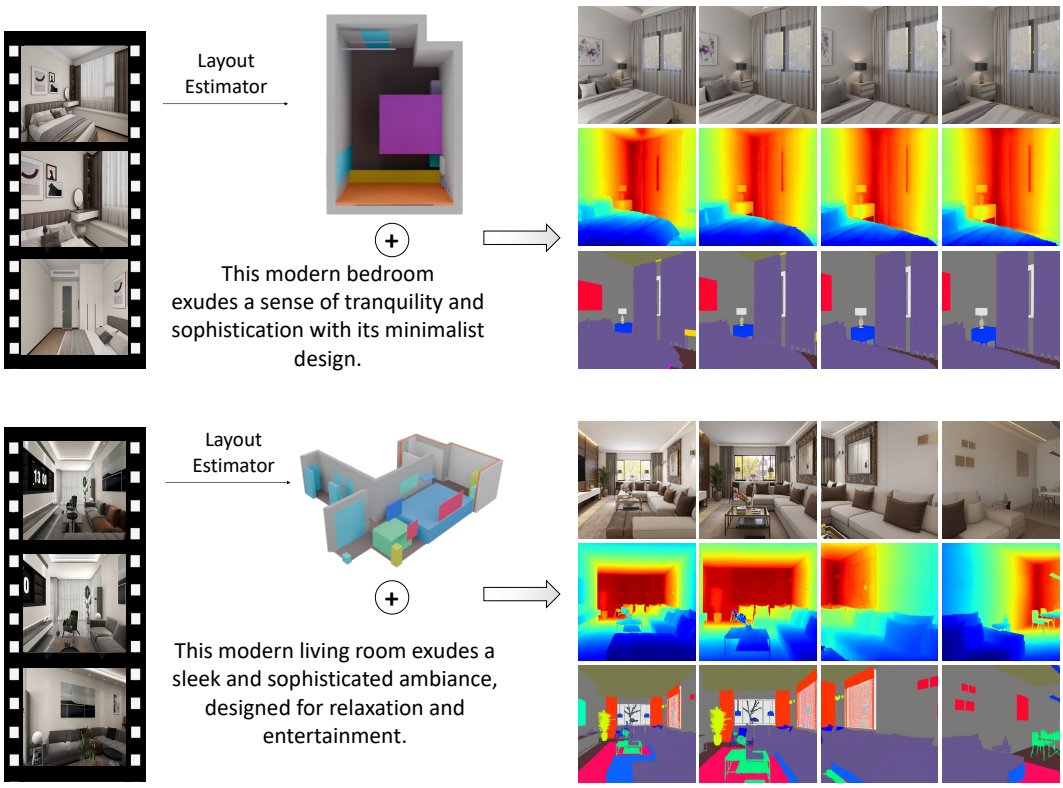


Figure 16. Video-to-New-3D Scene Generation on the SpatialLM Test set [6]. By leveraging the state-of-the-art scene layout estimation method, SpatialLM [6], we get the reconstructed 3D layout from the video. Then, we perform text-to-3D scene generation conditioned on this layout and additional user-provided text prompts. For clearer visualization of 3D consistency and multi-modal prediction capabilities, we put depth maps here instead of displaying the coordinate maps directly.