SPATIALGEN: Layout-guided 3D Indoor Scene Generation

Chuan Fang[1]*, Heng Li[1], Yixun Liang[1], Jia Zheng[2],
Yongsen Mao[2], Yuan Liu[1], Rui Tang[2], Zihan Zhou[2], Ping Tan[1]
[1]Hong Kong University of Science and Technology, [2]Manycore Tech Inc.
https://manycore-research.github.io/SpatialGen

Figure 1. Given a 3D semantic layout, SPATIALGEN can generate a 3D indoor scene conditioned on either a textual description (left) or a reference image (middle). Furthermore, it can transform a real-world scene, where its 3D layout is estimated from a video by a layout estimator [26], into some brand new scenes.

Abstract

Creating high-fidelity 3D models of indoor environments is essential for applications in design, virtual reality, and robotics. However, manual 3D modeling remains time-consuming and labor-intensive. While recent advances in generative AI have enabled automated scene synthesis, existing methods often face challenges in balancing visual quality, diversity, semantic consistency, and user control. A major bottleneck is the lack of a large-scale, high-quality dataset tailored to this task. To address this gap, we introduce a comprehensive synthetic dataset, featuring 12,328 structured annotated scenes with 57,431 rooms, and 4.7M photorealistic 2D renderings. Leveraging this dataset, we present SpatialGen, a novel multi-view multi-modal diffusion model that generates realistic and semantically consistent 3D indoor scenes. Given a 3D layout and a reference image (derived from a text prompt), our model synthesizes appearance (color image), geometry (scene coordinate map), and semantic (semantic segmentation map) from arbitrary viewpoints, while preserving spatial consistency across modalities. SpatialGen consistently generates supe-

*Work done during an internship at Manycore Tech Inc.

rior results to previous methods in our experiments. We are open-sourcing our data and models to empower the community and advance the field of indoor scene understanding and generation.

1. Introduction

Indoor scene generation aims to produce spatially coherent and photorealistic 3D indoor environments. As a fundamental challenge in computer vision, this task underpins diverse applications, including immersive films and games, interior design, and augmented/virtual reality (AR/VR). Moreover, it also provides diverse and physically realistic environments in robotic simulation for training and evaluating robot navigation and interaction capabilities.

A major consideration in developing 3D scene generation methods is the trade-off between *realism* and *scene diversity*. Procedural modeling methods [9, 32, 53] leverage hand-crafted heuristic rules and geometric constraints in the graphics engines, which produce highly realistic and physically plausible indoor environments. However, these scenes lack diversity. Recent 3D generative methods automatically generate scene layouts [27, 45] or other 3D representations like NeRFs [1] and 3D Gaussians [20]. But these methods exhibit limited layout and appearance realism, primarily due to the scarcity of annotated 3D data. In comparison, image-based methods utilize diffusion models to generate panoramas [38, 46] or multi-view images [11, 44] followed by 3D reconstruction. By leveraging powerful 2D priors, these methods show promise in striking a better balance between realism and scene diversity. Image-based methods, however, face additional challenges in multi-view *semantic consistency*. While recent video generation methods [25, 33, 54] have improved temporal coherence, synthesizing semantically consistent content when exploring beyond input views remain highly challenging.

To this end, **3D semantic layout** prior (Figure 1) has been employed in the literature to guide the generation process. But due to the lack of a large-scale dataset with paired 3D layout and images (or videos), existing layout-conditioned methods resort to one of the following two strategies: *score distillation* [4, 6, 50, 63] and *panorama-as-proxy* [8, 38]. The former directly distills powerful 2D pre-trained models for 3D content creation, avoiding the need for large-scale training data. But due to the inherent limitation of the SDS method [29], results produced by these methods suffer from severe visual artifacts (*e.g.*, over-saturation, lack of details). In contrast, the latter makes use of a special type of data, namely panoramas, for which large datasets with diverse scenes and annotations are available (*e.g.*, Structured3D dataset [60]). However, since panorama images are captured at fixed camera locations, models trained on such data have limited ability to extrapolate to novel viewpoints, restricting their application in real-world tasks.

To overcome these limitations, we collect a new indoor scene dataset on a much larger scale. Our dataset features 4.7M panoramic images with precise 2D and 3D layout annotations, spanning 57,431 rooms and 12,328 scenes. With this dataset, we take a new approach to 3D scene synthesis by building a scalable multi-view diffusion (MVD) model conditioned on 3D layout priors, which achieves high semantic consistency while maintaining the realism and scene diversity in the results.

We introduce SPATIALGEN, a novel framework for high-fidelity 3D indoor scene generation from a 3D room layout. *First*, we convert the 3D semantic layout into view-specific representations comprising coarse semantic maps and scene coordinate maps [39]. *Second*, we design a layout-guided attention mechanism that alternatively operates through: (i) cross-view attention for consistent information propagation across different viewpoints; (ii) cross-modal attention for fine-grained feature alignment between appearance, semantic, and geometric representations. This mechanism enables the joint synthesis of photorealistic RGB images, precise object semantic maps, and accurate scene coordinates for both input and novel viewpoints. *Finally*, we employ an iterative multi-view generation strategy to ensure complete scene coverage, followed by 3D Gaussian splatting optimization that reconstructs an explicit radiance field to enable free-viewpoint rendering.

Our main contributions are summarized as follows:

- We introduce a new large-scale dataset featuring over 4.7M panoramic images of 57,431 rooms and precise 2D and 3D layout annotations. This dataset fills a critical gap in 3D scene modeling by providing comprehensive multi-view data with structural annotations.
- We present SPATIALGEN, a new framework for layout-guided indoor scene generation. At the core of this framework is a novel multi-view multi-modal image diffusion method conditioned on a given layout prior, which generates semantically and geometrically consistent images from arbitrary viewpoints.
- Extensive evaluations conducted on text or image to 3D scene generations demonstrate that our method generates substantially more realistic and plausible 3D scenes.

2. Related Work

Procedural & 3D-based Scene Generation. Procedural generation (PCG) methods [31, 32] create 3D scenes with hand-crafted rules or constraints. Recent approaches integrate large language models (LLMs), either to generate scene layouts for subsequent object retrieval or shape synthesis [9, 43, 51], or to act as agents that produce Python scripts controlling procedural frameworks [16, 42].

3D-based methods generate 3D scene representations us-

Table 1. Statistics of the datasets for indoor scene generation. †: object annotations are provided by Ctrl-Room [8].

Dataset (year)	source	#scenes	#images	#objects	image type	annotations layouts	annotations objects
SUN R-GBD (2015)	real	-	10.3K	59K	perspective image	•	•
ScanNet (2017)	real	1,513	2.5M	36K	regular video		•
Matterport3D (2017)	real	90	10.8K	41K	sparse panoramas		•
ScanNet++ v2 (2024)	real	1,006	11.1M	111K	regular video		•
Structured3D (2020)	syn.	3,500	196.5K	150K†	panorama image	•	•†
Hypersim (2021)	syn.	461	77.4K	58K	regular video		•
SPATIALGEN dataset (ours)	syn.	**12,328**	**4.7M**	**1M**	panoramic video	•	•

ing generative models trained on datasets with 3D annotations. ATISS [27] and DiffuScene [45] predict compact layout parameters for scene objects. DiffInDScene [18], PDD [22], and SceneFactor [2] introduce a semantic layout as an intermediate guide to generate the indoor scene with an explicit geometric representation. But the lack of annotated 3D scene datasets results in subpar performance and limited generalization of such methods.

Image-based Scene Generation. In contrast, image-based methods exploit strong 2D priors in pretrained diffusion models to obtain photorealistic and diverse results. MVDiffusion [46] and PanoFusion [56] finetune a latent diffusion model [35] to generate a 360-degree panorama of a scene. Text2Room [15] and LucidDreamer [5] start with an initial RGB image and iteratively build the 3D scene by progressively warping and inpainting. CAT3D [11] and Bolt3D [44] trained a multi-view LDM to generate novel views from input images, followed by a 3D reconstruction. Despite these advances, existing methods struggle to synthesize large viewpoint changes [48, 54] and semantically coherent scenes [11, 37, 44] beyond observed areas.

The line of work that most closely relates to ours employs 3D layout prior to guide the generation process. Set-the-Scene [6], SceneCraft [50], and Layout2Scene [4] generate 3D scenes by distilling the pretrained image diffusion models conditioned on a given semantic layout. These methods achieve better view and semantic consistency, but the realism and controllability remain limited. While Ctrl-Room [8] and ControlRoom3D [38] generate panoramas with high visual fidelity, they struggle to extrapolate the scene beyond a fixed camera location without resorting to a dedicated room completion procedure. We argue that these limitations stem from the scarce scale and diversity of available 3D scene datasets, hindering the learning of robust 3D priors.

Indoor Scene Dataset. Existing indoor datasets are either captured from real-world scenes using RGB [21, 62] or RGB-D [3, 7, 40, 52] sensors or professionally designed with curated 3D CAD furniture models [10, 34, 41, 60]. The real-world dataset provides a physically realistic appearance observation of 3D scenes; however, collecting and annotating these data typically requires significant resources in terms of cost and labor. On the other hand, indoor synthetic datasets address the constraints of real-world data by supplying extensive, varied, and richly annotated scenes. In addition, Structured3D [60] and Hypersim [34] utilize the advanced render engine for photorealistic image rendering with accurate 2D labels. However, the camera view is limited, which restricts the downstream application.

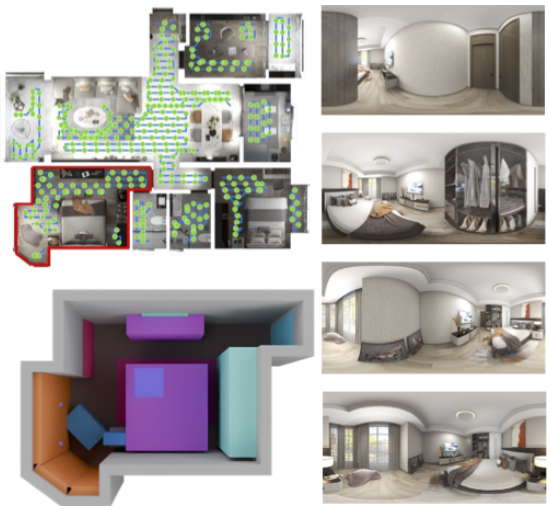

Figure 2. Illustration of our dataset. For each scene, we provide comprehensive panoramic renderings and 3D layout annotation.

3. SPATIALGEN Dataset

We summarize the commonly used indoor scene dataset for layout-conditioned scene synthesis in Table 1. As one can see, the real-world datasets suffer from a limited number of scenes, incomplete 3D annotations, and inconsistent annotation quality. Synthetic datasets are easier to annotate with ground-truth 3D labels, but they still have limitations in scene diversity (for example, Hypersim only has 461 scenes) or camera viewpoints (for example, Structured3D provides a single panorama for each room).

In this paper, we build a new dataset to train generative

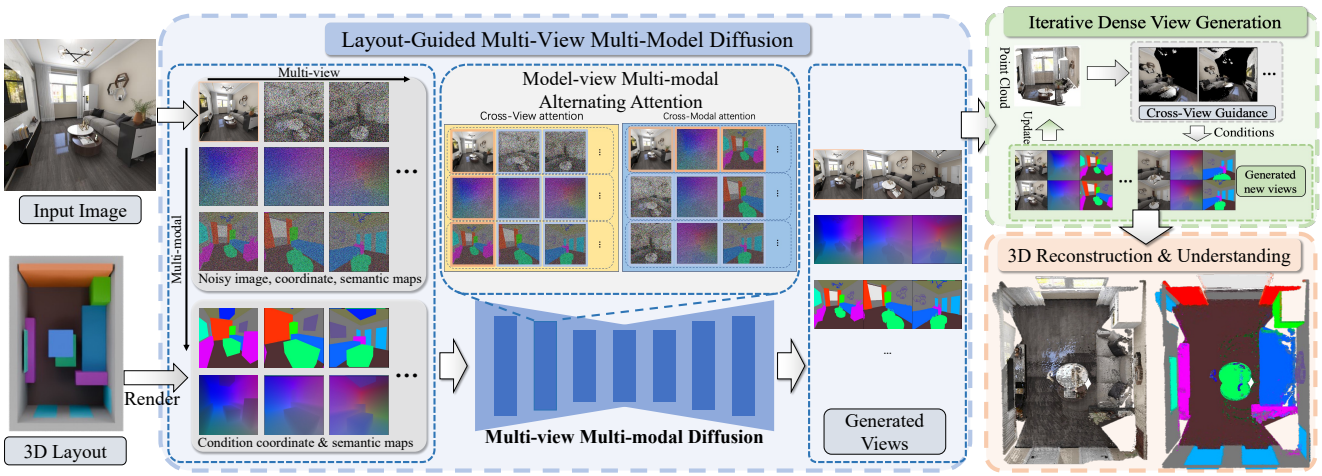

Figure 3. Overall pipeline. SPATIALGEN takes as input a 3D semantic layout and one or more posed images, to create a 3D scene. First, we generate per-view RGB images, scene coordinate maps, and semantic segmentation maps from a Layout-Guided Multi-view Multi-modal diffusion model. Then, we adopt an iterative dense view generation strategy to generate images at more sampled viewpoints. Finally, these images are fed into a 3D reconstruction method to produce the final result.

models for 3D indoor scenes. Our dataset is based on a large repository of house designs sourced from an online platform in the interior design industry. Most of these designs are designed by professional designers and are intended for real-world production.

As shown in Figure 2, we create physically plausible camera trajectories that navigate smoothly through each scene while avoiding obstacles. These trajectories are sampled at 0.5m intervals to ensure comprehensive spatial coverage. For each viewpoint, we generate photorealistic panoramic renderings using an industry-leading rendering engine, capturing color, depth, normal, semantic, and instance segmentation data. We further convert the panoramic image into multiple perspective images using equilib [14]. To ensure both quality and diversity, we apply rigorous filtering criteria during dataset curation, resulting in 12,328 distinct scenes encompassing 57,431 individual rooms with diverse room types. Each scene is annotated with precise 3D layouts and divided into 57,381/50 scenes for training/testing, respectively.

Figure 2 also shows some panoramas and 3D layout annotation from our dataset. Our dataset offers comprehensive structural layout annotations, including architecture elements (*i.e.*, walls, doors, and windows). We further simulate diverse camera motion patterns from panoramic video data to train and evaluate the generation capabilities of existing approaches – a crucial advantage over limited rule-based trajectories from existing dataset. The empirical benefits of this dataset are demonstrated in the experiments.

4. Method

We introduce SPATIALGEN, a novel method that generates Gaussian Splatting [19] scenes with semantics conditioned on a 3D layout with reference images. Specifically, given a semantic layout and one or more source images, SPATIALGEN first utilizes a layout-guided multi-view multi-modal diffusion model to generate dense views of the target scene via Iterative Dense View Generation (detailed in Section 4.3). Then, we recover those dense views to a unified semantic Gaussian Splatting via an off-the-shelf reconstruction method [55].

We start by providing a brief overview of multi-view diffusion models in Section 4.1. Then, we introduce our layout-guided latent diffusion model in Section 4.2, followed by the iterative generation scheme in Section 4.3 and the 3D reconstruction process in Section 4.4.

4.1. Preliminaries

Multi-view Diffusion Model. A multi-view latent diffusion model takes a single or multiple posed source views as input and generates multiple novel images in some target camera views. To incorporate multi-view conditioning, it typically involves two designs: (1) 2D attention layers are improved to a 3D-aware or multi-view aware attention mechanism, such as epipolar constraint [13], to capture multi-view features across different source views. (2) Camera poses are encoded by Plucker coordinate maps [28, 57] and then processed by a Transformer to compute view-conditioned embeddings.

Given M input views $\mathbf{I}_M = \{I_1, \ldots, I_M\}$ with camera poses $\mathbf{C}_M = \{C_1, \ldots, C_M\}$. Multi-view diffusion aims to predict N new view images $\mathbf{I}_N = \{I_{M+1}, \ldots, I_{M+N}\}$ in camera poses $\mathbf{C}_N = \{C_{M+1}, \ldots, C_{M+N}\}$. In other words, the multi-view latent diffusion model aims to learn the following joint distribution,

$$p(\mathbf{I}_N \mid \mathbf{I}_M, \mathbf{C}_{M+N}), \tag{1}$$

where $\mathbf{C}_{M+N} = \{C_1, \ldots, C_{M+N}\}$ includes camera poses for both input and output views.

Layout Condition in MVD. A 3D semantic layout provides an informative description of the scene. Following previous works [8, 38, 50], we represent the layout as a set of semantic bounding boxes of objects $\{\mathbf{b}_k\}_{k=1}^{K}$, where each box $\mathbf{b}_k$ includes the center location $l_k \in \mathbb{R}^3$, the size $s_k \in \mathbb{R}^3$, the orientation $r_k \in \mathbb{R}$ around the vertical axis, and the category label z_k. For each viewpoint with camera parameter $C_n = (K_n, T_n)$, we render a semantic map S_n^{layout} and a depth map D_n^{layout} of the bounding box of the 3D layout. The D_n^{layout} is then converted to the scene coordinate map $P_n^{\text{layout}} = T_n \cdot (K_n^{-1} \cdot D_n^{\text{layout}})$. In this way, we obtain the input layout conditions $\mathbf{S}_{M+N}^{\text{layout}}, \mathbf{P}_{M+N}^{\text{layout}} = \{S_n^{\text{layout}}, P_n^{\text{layout}}\}_{n=1}^{M+N}$ for the latent diffusion model. We use scene coordinate maps instead of depth maps to represent 3D scene geometry because they encode the scene in a globally consistent manner. As discussed in previous work [44, 59], they facilitate learning multi-view geometric consistency in the latent diffusion model. Therefore, extending Eq. (1), the joint distribution to be learned for layout-conditioned multi-view image generation can be formulated as follows,

$$p(\mathbf{I}_N \mid \mathbf{I}_M, \mathbf{S}_{M+N}^{\text{layout}}, \mathbf{P}_{M+N}^{\text{layout}}, \mathbf{C}_{M+N}). \tag{2}$$

Note that the layout conditions S_n and P_n only provide a coarse description of the bounding box without pixel-level details, as shown in Figure 3.

4.2. Layout-guided Multi-view Diffusion Model

Since the input layout maps do not contain pixel-level details, we further predict a pixel-wise semantic map S_n, scene coordinate map P_n for each viewpoint. This joint learning scheme improves 3D consistency in two ways:

- **Explicit 3D supervision.** By explicitly integrating both geometric and semantic maps into the latent diffusion model, SPATIALGEN leverages direct 3D supervision to achieve high-fidelity novel view synthesis results while maintaining cross-view consistency.
- **Cross-view guidance.** With the additional pixel-wise scene coordinate maps, we provide fine-grained guidance for diffusion by computing a warped image at any target view from the input images. Specifically, we adopt the point cloud based render [17] to obtain the warped image $I_n^{\text{warp}}, \forall n \in \{M+1, \ldots, M+N\}$. The warped image is encoded and concatenated with the original noise map I_n to form a conditioning signal for the target view, which we denote as augmented target views $\hat{I}_n = [I_n; I_n^{\text{warp}}]$.

The joint distribution to be learned now becomes,

$$p(\hat{\mathbf{I}}_N, \mathbf{S}_{M+N}, \mathbf{P}_{M+N} \mid \mathbf{I}_M, \mathbf{S}_{M+N}^{\text{layout}}, \mathbf{P}_{M+N}^{\text{layout}}, \mathbf{C}_{M+N}). \tag{3}$$

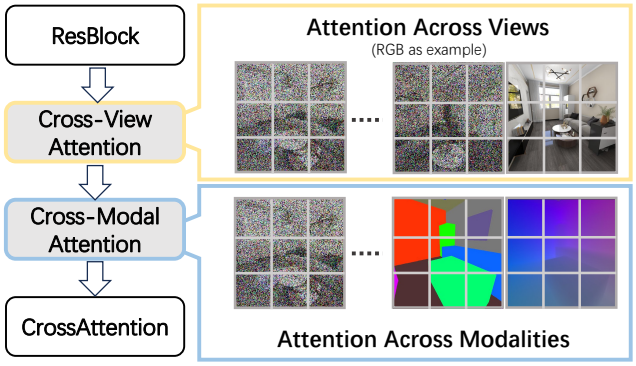

Figure 4. Multi-view and multi-modal alternating attention. It alternates between enforcing multi-view consistency and multi-modal fidelity within a unified attention mechanism.

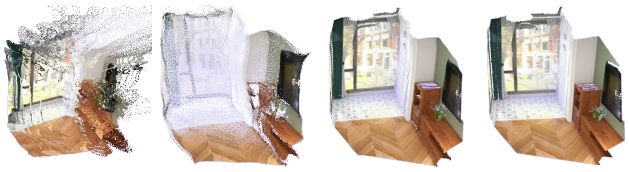

Figure 5. Comparison of reconstruction results for scene coordinate map. The image VAE (a) generates noisy results, and the SCM-VAE without gradient loss (b) produces distorted results. Our SCM-VAE (c) accurately reconstructs the scene geometry.

We use a v-parametrization and a v-prediction loss for the diffusion model [36]. Following CAT3D [11], the model is trained on a total of 8 views, with randomly sampled $\{1, 3, 7\}$ views as source views. We use the ground-truth scene coordinate map for warping in training, and use the predicted scene coordinates during inference.

Multi-view Multi-modal Alternating Attention. With our formulation Eq. (3), our objective is to generate output that is consistent with multiple viewpoints and modalities. We observe that a simple modification to the standard architecture, namely an alternating attention mechanism, is effective in preserving the desired consistency. As illustrated in Figure 4, the new architecture operates through complementary attention pathways: cross-view attention and cross-modal attention. Inspired by previous works [11, 23], our cross-view attention processes reshaped tokens along the view dimension (*e.g.*, $\{\mathbf{t}_1^I, \mathbf{t}_2^I, \ldots, \mathbf{t}_{M+N}^I\}$ for all RGB images), allowing feature aggregation across multiple views in each modality. While cross-modal attention operates within each view, observing modality-specific tokens (*e.g.*, $\{\mathbf{t}_n^I, \mathbf{t}_n^S, \mathbf{t}_n^P\}$ for image, semantics, and geometry) to achieve fine-grained feature alignment. This design achieves a balance between integrating information across different views and different modalities.

Scene Coordinate Map VAE (SCM-VAE). A standard image VAE pretrained on RGB images generalizes well to se-

mantic maps, but fails to accurately reconstruct scene coordinate maps, leading to poor geometric fidelity. See Figure 5(a) for an example. To address this, we introduce SCM-VAE, which encodes a scene coordinate map P into a latent representation z as $\mathbf{z} = \xi(P)$ and reconstructs z into a scene coordinate map with an uncertainty map as $\{\hat{P}, \mathbf{c}\} = \mathcal{D}(\mathbf{z})$, where ξ denotes the encoder and $\mathcal{D}$ is the decoder. The SCM-VAE is trained by fine-tuning the decoder $\mathcal{D}$ with an additional output dimension $\mathbf{c}$ from an image diffusion VAE, while keeping the encoder ξ frozen. $\mathbf{c}$ is activated by $\mathbf{c} = 1 + \exp(\mathbf{c})$ to ensure a strictly positive confidence [47]. The training objective combines standard VAE reconstruction with geometry-specific loss:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{grad}}, \tag{4}$$

$$\mathcal{L}_{\text{rec}} = c \odot \|\hat{P} - P\| - \alpha \log c, \tag{5}$$

$$\mathcal{L}_{\text{grad}} = \sum_{s=1}^{4} \|(\nabla \hat{P}_i^s - \nabla P_i^s)\|, \tag{6}$$

where $\alpha = 0.2$ and $\odot$ denotes element-wise multiplication. Here, we follow previous monocular depth estimation works [12] to use a multiscale gradient loss $\mathcal{L}_{\text{grad}}$ to improve boundary sharpness in the decoded scene coordinate map. As we can see in Figure 5, our SCM-VAE with $\mathcal{L}_{\text{grad}}$ outperforms the one without the term, especially around complex object boundaries and flat areas.

4.3. Iterative Dense View Generation

Our goal is to generate a complete 3D scene aligned to the given layout and text or image prompt. Although our layout-guided MVD model can generate an arbitrary number of views in principle, it is limited by GPU memory constraints. Thus, instead of generating all views at once, we adopt an iterative view synthesis strategy, a similar approach is also used in previous work [25, 54]. The main idea is to incrementally maintain a colored global point cloud of the scene to enforce appearance consistency between iterations. During each iteration, the point cloud $\mathcal{P}$ is projected onto the target views $\mathbf{I}^{warp}$ to provide pixel-aligned guidance for consistent generation. The SCMs of the target view generated by our diffusion model will be inserted into $\mathcal{P}$. In this way, we can effectively reduce error accumulation. Furthermore, by incorporating the uncertainty map $\mathbf{c}$, we filter out 3D points with uncertainty below a predefined threshold, resulting in cleaner warped images.

The iterative process, detailed in Algorithm 1, proceeds as follows: *First*, an initial point cloud is built by obtaining the scene coordinate maps $\mathbf{P}_M$ for the input views $\mathbf{I}_M$. *Second*, at the beginning of each iteration, we render warped images for the target views from the point cloud. Then, we perform inference with not only the input images $\mathbf{I}_m$, but also the warped images to ensure global consistency. We then update the point cloud accordingly. *Finally*, we collect the images generated from all iterations as output.

Algorithm 1: Iterative Dense View Generation

Input: input views $\mathbf{I}_M$, camera poses for all iterations $\{\mathbf{C}_M, \mathbf{C}_{N_1}, \ldots, \mathbf{C}_{N_K}\}$, global point cloud $\mathcal{P}$ initialized as empty, layout-guided multi-view diffusion model $\mathcal{U}(\cdot)$.

Initialization:

Obtain the $\mathbf{S}_M$ and $\mathbf{P}_M$: $\{\mathbf{S}_M, \mathbf{P}_M\} \leftarrow \mathcal{U}(\mathbf{I}_M, \mathbf{C}_M)$;

Update global point cloud: $\mathcal{P} \leftarrow \mathbf{P}_M$;

for $k \leftarrow 1$ **to** K **do**

  Render warped images: $\mathbf{I}_{N_k}^{\text{warp}} \leftarrow \text{Render}(\mathcal{P}, \mathbf{C}_{N_k})$;

  Augment the target views $\hat{\mathbf{I}}_{N_k} = [\mathbf{I}_{N_k}; \mathbf{I}_{N_k}^{\text{warp}}]$;

  Run inference: $\{\hat{\mathbf{I}}_{N_k}, \mathbf{S}_{N_k}, \mathbf{P}_{N_k}\} \leftarrow \mathcal{U}(\mathbf{I}_M, \mathbf{C}_{M+N_k})$;

  Update global point cloud: $\mathcal{P} \leftarrow \mathcal{P} \bigcup \mathbf{P}_{N_k}$;

end

Output: all generated views: $\{\mathbf{I}_{N_1}, \mathbf{I}_{N_2}, \ldots, \mathbf{I}_{N_K}\}$.

4.4. Scene Reconstruction and Understanding

Building upon RaDe-GS [55], we reconstruct a 3D scene representation from the densely generated color, geometric, and semantic images. Following Feature-3DGS [61], We augment the standard 3D Gaussians with a semantic feature in each point. The scene is initialized from the predicted point cloud $\mathcal{P}$. During differentiable rendering optimization, we employ a depth supervision loss that utilizes the predicted scene coordinate maps, enabling rapid convergence in just 7,000 steps. As shown in Figure 6, the pipeline produces high-fidelity RGB renderings and geometrically accurate depth reconstructions.

5. Experiments

5.1. Experiment Setup

Benchmark Datasets. We use both existing datasets (*i.e.*, Hypersim [34] and Structured3D [60]) and our new dataset. As discussed in Section 3, one key difference of these datasets is in the diversity of camera viewpoints. Since our dataset provides abundant panorama images at dense locations in each room, we can design various camera movement patterns to conduct an extensive evaluation.

Specifically, we introduce four distinct camera trajectories with different amounts of view overlap and distance between input and target views: (i) *Forward*: the trajectory follows a linear path with minimal directional variation, simulating steady camera movement. (ii) *Inward Orbit*: both the input and output views are directed toward the center of the room, ensuring substantial view overlap; (iii) *Outward Orbit*: the input and output views are at the same location, but oriented differently, with less than 45° overlap at adjacent views. (iv) *Random Walk*: input and output views are sampled from a continuous random-walk path,

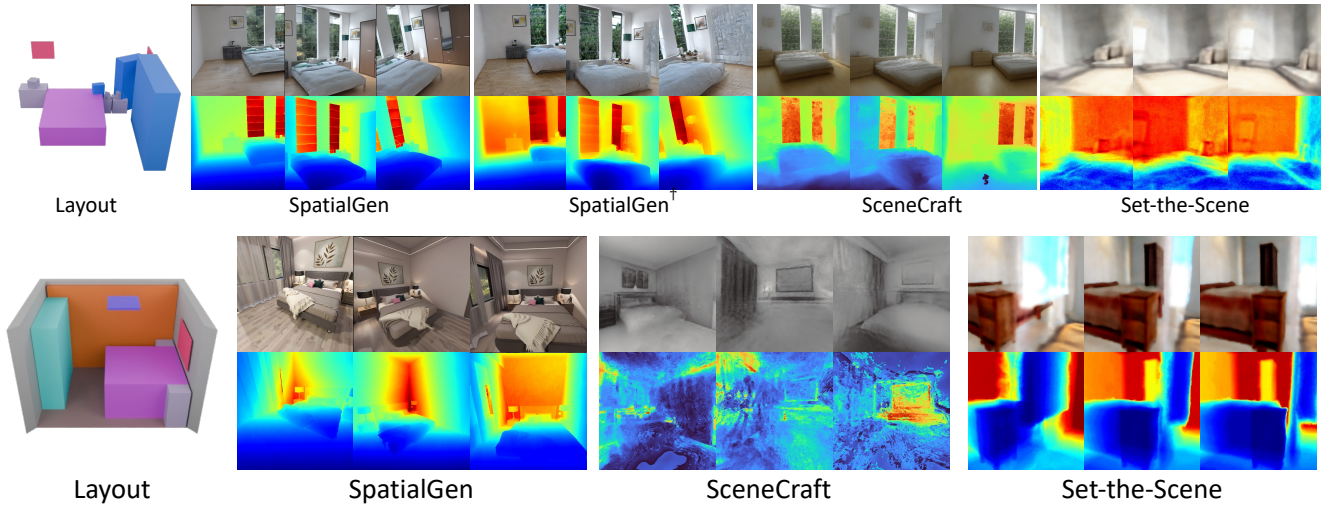

Figure 6. Qualitative comparison to score distillation methods on Hypersim [34] (top row) and our dataset (bottom row), In each case, we show the generated color images and depth maps.

Table 2. Comparison to score distillation methods.

Dataset	Method	CLIP Sim. (%) ↑	Image Reward ↑
Hypersim [34]	Set-the-Scene [6]	25.18	-2.005
	SceneCraft [50]	26.94	-1.096
	SPATIALGEN†	25.93	-1.168
	SPATIALGEN	**27.59**	**-0.285**
Our Dataset	Set-the-Scene [6]	25.23	-2.100
	SceneCraft [50]	18.93	-2.267
	SPATIALGEN	**27.89**	**-0.123**

resulting in minimal view overlap. Please refer to the appendix for the visualization of these settings.

Implementation Details. We implement SPATIALGEN in PyTorch. Both the SCM VAE and the latent diffusion model are fine-tuned from stable diffusion 2.1 (SD-2.1) [35]. We use AdamW optimizer [24]. For SCM VAE, we freeze the encoder and only fine-tune the decoder for 10K steps with a batch size of 64. For the latent diffusion model, we fine-tune it for 35K steps with a batch size of 128. The first 16K steps use resolution 256×256, whereas the remaining steps use resolution 512×512. All models are fine-tuned on 64 NVIDIA RTX 4090 GPUs. The learning rate starts at 10^{-4} and decays by a factor of 0.01 at 90% of the total training process. We render warped images using PyTorch3D [17].

For text-to-3D scene generation, we further train a layout ControlNet [58] to generate the reference image for our latent diffusion model.

5.2. Text-to-3D Scene Generation

As discussed before, existing layout-conditioned text-to-3D methods can be grouped into two categories: *score distillation* methods [4, 6, 50, 63] and *panorama-as-proxy* methods [8, 38]. In the following, we compare to these two groups separately.

5.2.1 Comparison to Score Distillation Methods

For this experiment, we compare to two open-source methods: Set-the-Scene [6] and SceneCraft [50]. The other methods such as Layout2Scene [4] and GALA3D [63] do not release their codes. We conduct experiments on both Hypersim [34] and our new datasets. Evaluation is performed on the test scenes of each target dataset following standard protocols.

For *Hypersim dataset*, we directly use the official checkpoints of Set-the-Scene and SceneCraft. To illustrate the benefit of our proposed large-scale dataset, we include two training configurations for our model: SPATIALGEN† which is trained solely on Hypersim, and SPATIALGEN which is trained on a combination of Hypersim and our datasets.

For *our dataset*, we fine-tune the layout ControlNet of SceneCraft and adapt the layouts to match the input of Set-the-Scene. All methods on our dataset use the *Inward Orbit* camera trajectory for consistency.

Quantitative Results. Table 2 reports the performance of all methods with established 2D rendering metrics: CLIP similarity score [30] to measure text-image alignment and Image Reward [49] to assess human aesthetic preference.

As one can see, our method performs slightly worse than SceneCraft when trained solely on Hypersim. We hypothesize that Hypersim is too small for powerful latent multi-view diffusion models like the one employed by our method. When trained on both Hypersim and our datasets, our method outperforms both SDS methods on all metrics. SPATIALGEN achieves a significantly higher image-reward score than models trained solely on Hypersim, validating the benefit of our large-scale dataset for high-quality 3D scene generation. Furthermore, when tested on our dataset, SPATIALGEN consistently outperforms the baselines, with a significantly higher image-reward score.

Qualitative Results. The advantage of our method can be

Table 3. Comparison to panorama-as-proxy method.

Dataset	Method	CLIP Sim. (%) ↑	Image Reward ↑
Structured3D [60]	Ctrl-Room [8]	**25.03**	**-1.016**
	SPATIALGEN	23.90	-1.405
Our Dataset	Ctrl-Room [8]	22.63	-1.546
	SPATIALGEN	**27.89**	**-0.123**

Figure 7. Qualitative comparison between our method and the panorama-as-proxy baseline [8] on Structured3D [60] (top row) and our dataset (bottom row).

better observed in Figure 6. On both Hypersim and our datasets, SPATIALGEN generates photorealistic scenes with superior details that are well-aligned with the specified layout. In contrast, SceneCraft struggles to balance layout adherence with text-prompt fidelity, and Set-the-Scene takes nearly two hours to synthesize a radiance field that still lacks fine-grained details. While SPATIALGEN† produces blurry and artifact images, further underscores the benefit of our proposed dataset to the final output quality.

5.2.2 Comparison to Panorama-as-Proxy Methods

Next, we compare our method to panorama-as-proxy methods. We choose Ctrl-Room [8] as the baseline because the source code and checkpoints are available. We conduct experiments on both Structured3D and our datasets. For *Structured3D dataset*, we use ground-truth layouts to ensure consistent inputs for both methods. For *our dataset*, we use the *Inward Orbit* camera trajectory for consistency.

Quantitative Results. Table 3 reports the performance of both methods, we take the renderings of the generated mesh from Ctrl-Room for comparison. On Structured3D dataset (which provides only a single panorama per scene), our method still achieves competitive performance. The relatively lower scores are expected as Ctrl-Room is specifically trained to synthesize a single panorama at a fixed camera location. In contrast, our method generates multiple perspective images without explicitly exploiting the fact that all images are taken from a single location.

Nevertheless, the critical advantage of our method is revealed on our dataset: when rendering from novel viewpoints, the performance of Ctrl-Room degrades significantly, whereas our method consistently produces high-quality results from arbitrary views.

Qualitative Results. Figure 7 further highlights the advantage our method over panorama-as-proxy method. On our dataset, Ctrl-Room fails to synthesize coherent novel views, exhibiting severe distortions and artifacts. In contrast, our method is not limited to a single camera position (*i.e.*, panorama generation). It achieves high-quality panorama generation on Structured3D while also enabling photorealistic novel view synthesis on our dataset.

5.3. More Experiments and Ablations

Additionally, we evaluate SPATIALGEN on image-to-3D scene generation to validate its generative capability and 3D consistency. Recognizing that well-defined 3D semantic layouts are often unavailable in practice, we further extend the evaluation to video inputs. In this setting, we leverage the state-of-the-art layout estimation model [26] to parse a 3D layout directly from the video. Additional results, implementation details, and ablations are provided in Appendix.B and Appendix.C.

6. Conclusion & Limitations

We present SPATIALGEN, a novel framework for layout-guided 3D indoor scene synthesis. At the core of our pipeline is a multi-view multi-modal diffusion model, which generates images with high visual quality and geometric consistency. To train this model, we collect a new synthetic indoor scene dataset with 4.7M panoramic renderings of 57,431 rooms and the 3D layout annotations. These advancements open new possibilities for downstream applications such as interior design, embodied AI, and AR/VR.

Limitations. First, the cross-view and cross-modal attention introduces additional computational cost to the diffusion model, which limits SPATIALGEN to generate more images at a time. Second, the camera sampling strategy might affect the generation quality. Moreover, since our method relies on an initial RGB image, it is sensitive to any misalignment between the image and the provided 3D layout. We plan to address these challenges in the future.

Acknowledgments

This work was supported in part by the Key R&D Program of Zhejiang Province (2025C01001) and HKUST Project 24251090T019. We thank Yingqi Shen, Liangbin Hu, and Fuchun Dong (Manycore Tech) for dataset support; Chenfeng Hou for SDS-based experiments; Zhiwei Wang for ControlNet testing; and Kunming Luo for figure suggestions.

References

[1] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Joshua Susskind. GAUDI: A neural architect for immersive 3d scene generation. In *Adv. Neural Inform. Process. Syst.*, pages 25102–25116, 2022. 2

[2] Aleksey Bokhovkin, Quan Meng, Shubham Tulsiani, and Angela Dai. SceneFactor: Factored latent 3D diffusion for controllable 3D scene generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 628–639, 2025. 3

[3] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *IEEE Int. Conf. 3D Vis.*, pages 667–676, 2017. 3

[4] Minglin Chen, Longguang Wang, Sheng Ao, Ye Zhang, Kai Xu, and Yulan Guo. Layout2Scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. *arXiv preprint arXiv:2501.02519*, 2025. 2, 3, 7

[5] 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. 3

[6] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-Scene: Global-local training for generating controllable nerf scenes. In *IEEE Int. Conf. Comput. Vis. Worksh.*, pages 2920–2929, 2023. 2, 3, 7

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017. 3

[8] 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. 2, 3, 5, 7, 8

[9] Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. In *Adv. Neural Inform. Process. Syst.*, pages 18225–18250, 2023. 2

[10] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3d furnished rooms with layouts and semantics. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10933–10942, 2021. 3

[11] 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. 2, 3, 5

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 270–279, 2017. 6

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[14] haruishi43. Equilib, 2020. 4

[15] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3D meshes from 2D text-to-image models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7909–7920, 2023. 3

[16] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. SceneCraft: An llm agent for synthesizing 3d scenes as blender code. In *Int. Conf. Mach. Learn.*, 2024. 2

[17] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3d deep learning with pytorch3d. In *SIGGRAPH Asia 2020 Courses*, 2020. 5, 7

[18] Xiaoliang Ju, Zhaoyang Huang, Yijin Li, Guofeng Zhang, Yu Qiao, and Hongsheng Li. DiffInDScene: Diffusion-based high-quality 3D indoor scene generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4526–4535, 2024. 3

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 4

[20] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3D: Real-world camera trajectory and 3D scene generation from text. In *Adv. Neural Inform. Process. Syst.*, pages 75125–75151, 2024. 2

[21] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22160–22169, 2024. 3

[22] Yuheng Liu, Xinke Li, Xueting Li, Lu Qi, Chongshou Li, and Ming-Hsuan Yang. Pyramid diffusion for fine 3d large scene generation. In *Eur. Conf. Comput. Vis.*, pages 71–87, 2024. 3

[23] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9970–9980, 2024. 5

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2017. 7

[25] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2016–2029, 2025. 2, 6

[26] 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. 1, 8

[27] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. In *Adv. Neural Inform. Process. Syst.*, pages 12013–12026, 2021. 2, 3

[28] Julius Plucker. On a new geometry of space. *Phil. Trans. R. Soc*, 155:725–791, 1865. 4

[29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *Int. Conf. Learn. Represent.*, 2023. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021. 7

[31] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12630–12641, 2023. 2

[32] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21783–21794, 2024. 2

[33] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6121–6132, 2025. 2

[34] 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. 3, 6, 7

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 3, 7

[36] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *Int. Conf. Learn. Represent.*, 2022. 5

[37] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9420–9429, 2024. 3

[38] Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, and Ji Hou. ControlRoom3D: Room generation using semantic proxy rooms. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6201–6210, 2024. 2, 3, 5, 7

[39] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. SceneCoordRegression: Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2930–2937, 2013. 2

[40] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Int. Conf. Comput. Vis.*, pages 567–576, 2015. 3

[41] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica Dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2024. 3

[42] Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3D-GPT: Procedural 3D modeling with large language models. *arXiv preprint arXiv:2310.12945*, 2023. 2

[43] Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. LayoutVLM: Differentiable optimization of 3D layout via vision-language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 29469–29478, 2025. 2

[44] Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3D: Generating 3D scenes in seconds. In *IEEE Int. Conf. Comput. Vis.*, 2025. 2, 3, 5

[45] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. DiffuScene: Scene graph denoising diffusion probabilistic model for generative indoor scene synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20507–20518, 2024. 2, 3

[46] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Adv. Neural Inform. Process. Syst.*, pages 51202–51233, 2023. 2, 3

[47] Sheng Wan, Tung-Yu Wu, Wing H. Wong, and Chen-Yi Lee. Confnet: Predict with confidence. In *Int. Conf. Acoust. Speech Signal Process.*, pages 2921–2925, 2018. 6

[48] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 2024. 3

[49] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *Adv. Neural Inform. Process. Syst.*, 36: 15903–15935, 2023. 7

[50] 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. 2, 3, 5, 7

[51] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3D embodied AI environments. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16227–16237, 2024. 2

[52] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *IEEE Int. Conf. Comput. Vis.*, pages 12–22, 2023. 3

[53] Lap Fai Yu, Sai Kit Yeung, Chi Keung Tang, Demetri Terzopoulos, Tony F Chan, and Stanley J Osher. Make it home: automatic optimization of furniture arrangement. *ACM Trans. Graph.*, 30(4), 2011. 2

[54] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2, 3, 6

[55] Baowen Zhang, Chuan Fang, Rakesh Shrestha, Yixun Liang, Xiaoxiao Long, and Ping Tan. RaDe-GS: Rasterizing depth in gaussian splatting. *arXiv preprint arXiv:2406.01467*, 2024. 4, 6

[56] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6347–6357, 2024. 3

[57] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *Int. Conf. Learn. Represent.*, 2024. 4

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE Int. Conf. Comput. Vis.*, pages 3836–3847, 2023. 7

[59] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3D modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21685–21695, 2025. 5

[60] 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. 2, 3, 6, 8

[61] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D gaussian splatting to enable distilled feature fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21676–21685, 2024. 6

[62] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), 2018. 3

[63] Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. GALA3D: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. In *Int. Conf. Mach. Learn.*, 2024. 2, 7