Seen2Scene: Completing Realistic 3D Scenes with Visibility-Guided Flow

Quan Meng[1], Yujin Chen[1], Lei Li[2], Matthias Nießner[1], and Angela Dai[1]

[1]Technical University of Munich [2]University of Virginia

https://quan-meng.github.io/projects/seen2scene/

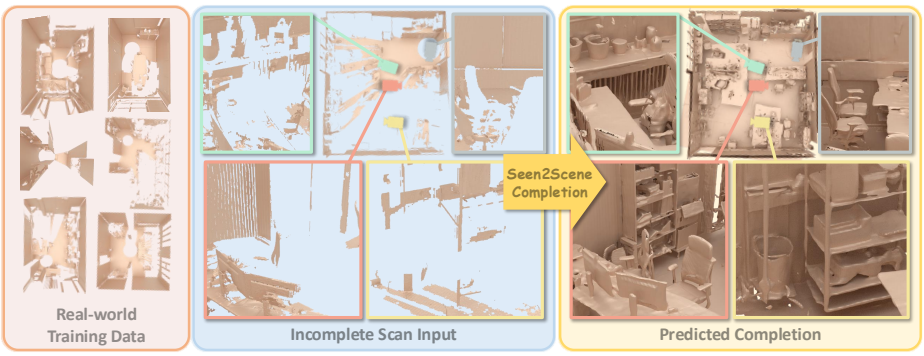

Fig. 1: Seen2Scene is the first visibility-guided flow matching approach for 3D scene completion and generation, trained directly on incomplete, real-world 3D scan data. Our approach predicts high-fidelity, geometrically complete, and structurally coherent scene geometry in unobserved regions, conditioned on observed areas and 3D box-based layouts.

Abstract. We present Seen2Scene, the first flow matching-based approach that trains directly on incomplete, real-world 3D scans for scene completion and generation. Unlike prior methods that rely on complete and hence synthetic 3D data, our approach introduces visibility-guided flow matching, which explicitly masks out unknown regions in real scans, enabling effective learning from real-world, partial observations. We represent 3D scenes using truncated signed distance field (TSDF) volumes encoded in sparse grids and employ a sparse transformer to efficiently model complex scene structures while masking unknown regions. We employ 3D layout boxes as an input conditioning signal, and our approach is flexibly adapted to various other inputs such as text or partial scans. By learning directly from real-world, incomplete 3D scans, Seen2Scene enables realistic 3D scene completion for complex, cluttered real environments. Experiments demonstrate that our model produces coherent, complete, and realistic 3D scenes, outperforming baselines in completion accuracy and generation quality.

Keywords: 3D Scene Completion · Real-world Scene Generation · Multi-Modality

1 Introduction

Modern 3D content creation is rapidly shifting from synthetic, isolated objects toward realistic, immersive, and interactive world models. A significant challenge in this transition is generating complete 3D scenes from partial observations – a capability increasingly central to content creation, simulation, and planning. Addressing this challenge calls for 3D generative models that can imagine missing structure while respecting observed geometry and scaling to complex, multi-object indoor scenes.

Recent progress has been made in 3D scene synthesis leveraging the power of deep generative models, in particular, through diffusion-based [4, 41, 47, 59] and flow matching approaches [58, 61, 62]. However, training these models heavily relies on synthetic 3D datasets [20, 21] that provide complete ground-truth scene geometry for supervision. This fundamentally limits their ability to generalize to real-world settings. Unlike synthetic data, real-world scans [3, 12, 65] are inherently incomplete and noisy due to occlusions and limited sensor coverage. Methods trained exclusively on complete and synthetic 3D scenes often fail to capture the variability, complexity, incompleteness, and off-axis alignments inherent in real scans.

To address these challenges, we propose Seen2Scene, a masked flow matching–based approach capable of training directly on incomplete, real-world scans for 3D scene completion and generation. Our method operates on partial 3D scenes represented as TSDF voxel grids, which naturally encode surfaces while handling unobserved space. We first train a masked sparse variational autoencoder (VAE) to compress observed TSDF voxel grids into compact geometry latents, masking out unknown regions during training to learn clean latent representations of the observed scene geometry. A sparse transformer then models the distribution of partial geometry latents using visibility-guided flow matching [37, 38], synthesizing coherent 3D structure conditioned on input 3D box-based layouts. While designed for partial scan completion, Seen2Scene naturally extends to other conditions, such as text prompts, to generate high-fidelity 3D scenes from scratch.

We evaluate our method on the large-scale ScanNet++ [65], ARKitScenes [3, 30], and 3D-FRONT [20] datasets, where we reconstruct complete indoor scenes from partial TSDF scans and generate 3D scenes from text prompts or layouts. Our experiments demonstrate that Seen2Scene produces complete and realistic 3D scene geometry, outperforming baselines in both geometric completeness and surface fidelity. These results establish visibility-guided flow matching as an effective paradigm for 3D scene completion from real-world partial data.

In summary, we present the following contributions:

- We present the first masked flow-matching method that can be trained directly on complex, incomplete real-world 3D scans and learn their underlying geometry distributions for accurate scene completion.
- We introduce a masked sparse VAE compression that faithfully encodes partial TSDF scans into clean and compact geometry latents, enabling our masked

visibility-guided generative modeling to focus on observed scene geometry while effectively handling unknown regions.
– Our method can be easily adapted to generate complete, high-fidelity 3D scenes from scratch conditioned on text prompts and/or object layouts, demonstrating strong flexibility and generalization beyond scene completion.

2 Related Work

Score-Based 3D Shape Generation. Following the success of diffusion [23,48] and flow matching [37, 38] in the image domain, 3D shape generative methods have adopted this generative paradigm for various 3D shape representations: point clouds [71], learned latent codes [54], and implicit function parameters [9, 18, 29, 69]. More recently, structured sparse latents [35, 47, 61, 62] and scalable 3D diffusion transformers [8, 58] have pushed both shape quality and resolution further. However, these methods focus on object-centric generation, relying on datasets of curated synthetic object assets. Our method builds upon flow matching and scales 3D generative models to unstructured, incomplete, large-scale real-world data such as RGB-D scans.

3D Scan Completion. 3D scan completion aims to reconstruct the full geometry representing a scene from partial observations such as RGB-D inputs or sparse LiDAR scans. Early approaches employed volumetric 3D CNNs [13,14,26], with SG-NN [13] introducing self-supervised sparse generative modeling from partial scans. Subsequent works improved structural priors and robustness via sketch-aware supervision [7], anisotropic convolutions [32], and uncertainty modeling [5]. To mitigate occlusion, camera-based methods exploit temporal context [31], virtual multi-view augmentation [27, 50], and pseudo-future modeling [40]. [22, 34, 60] infer a complete 3D volumetric representation of both semantics and geometry from a single RGB image. However, these works view completion from a regression standpoint. Generative completion works [10,49,67] incorporate strong generation ability for completion task. We treat scene completion as sparse-scan-conditioned generation, which enables more direct fine-tuning of pretrained generative foundation models, achieving significant generalization to unobserved regions while keeping fidelity to observations.

3D Scene Generation. 3D scene generation aims to capture global layout structure and local geometry [45, 46]. Layout-conditioned approaches [6, 19, 52, 53,66] learn to synthesize scenes given spatial arrangements of objects. To achieve scalability in expansive environments, recent methods focus on chunk-based generation and outpainting [4, 25, 28, 41, 47, 59]. While effective in high-resolution synthesis, these models are trained using complete, synthetic 3D scene datasets limited in diversity and asset categories, limiting their ability to bridge the domain gap to real-world 3D scan data. Recently, agentic frameworks [36,43,57,64] have introduced LLM/VLM planners to produce object-based indoor 3D scenes. Although these methods excel in high-level planning, they often rely on asset retrieval or simplified layout assumptions (*e.g.*, axis-aligned), struggling to syn-

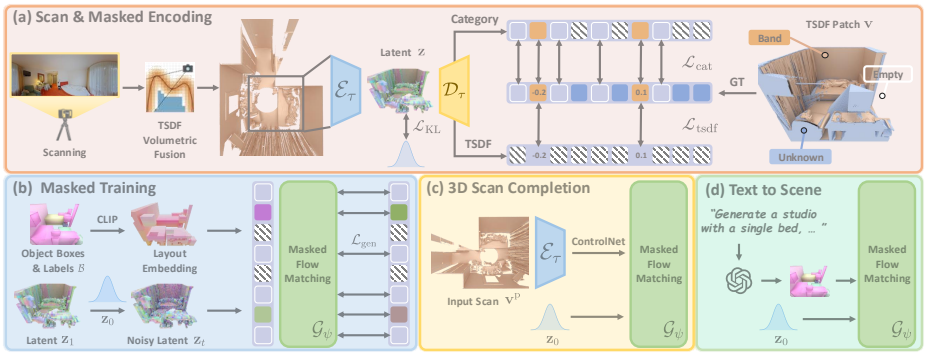

Fig. 2: Overview of Seen2Scene. We introduce visibility-guided flow matching for modeling the distribution of TSDF partial scans. **(a)** Partial scan TSDF patches $\mathbf{v}$ are encoded by a masked sparse VAE $(\mathcal{E}_\tau, \mathcal{D}_\tau)$ into latent representations $\mathbf{z}$, masking out unknown regions unseen by the camera. **(b)** A sparse transformer $\mathcal{G}_\psi$ conditioned on 3D layout boxes $\mathcal{B}$ is trained with masked flow matching on surface and empty region tokens. **(c)** We fine-tune $\mathcal{G}_\psi$ for scan completion by injecting partial scan inputs $\mathbf{v}^{\mathrm{p}}$ via ControlNet [68]. **(d)** $\mathcal{G}_\psi$ can also be flexibly adapted for text or layout-conditioned 3D scene generation from scratch.

thesize complex, high-fidelity geometry reflective of real-world, cluttered environments [3,12,65]. In contrast, we introduce visibility-aware masked flow matching that learns directly from incomplete, noisy real scans.

3 Method

We propose a new generative approach, Seen2Scene, which synthesizes complete, high-fidelity 3D scenes from partial, real-world scans. Key to our method is visibility-guided flow matching, which enables generative modeling on incomplete 3D scan data by focusing on observed scene geometry while remaining agnostic to unknown areas.

An overview of our approach is illustrated in Fig. 2. Input 3D scans are represented as TSDF volumes reconstructed via volumetric fusion [11] (Sec. 3.1). We then sample and encode patches, denoted as $\mathbf{v}$, from these partial TSDF into compact geometry latents $\mathbf{z}$ using a masked sparse VAE (Sec. 3.2). As the TSDF representation inherently encodes which regions are unknown (*i.e.*, unseen by the scanning camera, with signed distance of $-3 \times$ voxel size), these regions are masked out during their encoding. A sparse transformer $\mathcal{G}_\psi$ is then trained with visibility-aware masked flow matching to generate geometry latents from Gaussian noise conditioned on 3D bounding box layouts, with supervision only on observed regions, while generated geometry latents for unknown regions are masked and excluded from the flow matching loss (Sec. 3.3). This flow matching model then serves as the backbone for 3D scene completion and generation. To complete partial scans, we fine-tune the pretrained generative model by injecting input partial scan TSDFs via a ControlNet [68] branch, which guides generation of unknown regions while preserving observed geometry (Sec. 3.4).

3.1 Visibility-Aware Structured Scene Representation

To learn directly on partial scans of real-world environments, at each iteration, we randomly sample a patch $\mathbf{v}$ of resolution 256^3 within the scene. The patch $\mathbf{v}$ is stored as a TSDF grid reconstructed from raw depth sequences of the environment via volumetric fusion [11], as shown in Fig. 2(a).

This volumetric representation naturally handles the incompleteness of real-world scans and encodes visibility information. Voxels with TSDF values greater than the negative truncation bound encode known geometry observed by the camera, while voxels at the sentinel value of $-3 \times$ voxel size indicate unknown regions unseen by the camera. We explicitly leverage this visibility information to mask out unknown regions during both the scene latent encoding and generative modeling stages. This ensures that the model is supervised only by valid geometry observations, preventing it from learning the ambiguity from unobserved areas to enable more generalized completion.

3.2 Latent Encoding with Masked Sparse VAE

To efficiently model high-resolution 3D scene geometry, we first compress the TSDF chunks $\mathbf{v}$ into a compact latent representation $\mathbf{z}$. We design a visibility-aware masked sparse VAE, as illustrated in Fig. 2(a), which explicitly handles the inherent sparsity and incompleteness of real-world scans.

This sparse VAE consists of a sparse encoder $\mathcal{E}_\tau$ and a sparse decoder $\mathcal{D}_\tau$, parameterized by τ, both built with residual sparse convolutional blocks [56]. Given the TSDF patch $\mathbf{v}$, the encoder $\mathcal{E}_\tau$ progressively downsamples the sparse voxel features and maps them to a latent distribution $q(\mathbf{z} \mid \mathbf{v})$, regularized toward a standard normal prior $p(\mathbf{z})$ via a KL divergence loss $\mathcal{L}_{\text{KL}}$. The decoder $\mathcal{D}_\tau$ then reconstructs the input partial geometry from sampled latents through consecutive upsampling. To predict sparse 3D surfaces, the decoder has two prediction heads: a *category head* that classifies each voxel as surface (i.e., voxels with distance smaller than $3 \times$ voxel size, colorized as orange in Fig. 2) or empty (i.e., voxels with signed distance of $3 \times$ voxel size, colorized as white in Fig. 2), and a *TSDF head* that regresses signed-distance values only for surface voxels.

We adopt a masked training strategy [13] to accommodate the incompleteness of real-world scans in unobserved regions (highlighted in blue in Fig. 2). Such incompleteness arises from inevitable occlusion and reflective surfaces — for instance, regions beneath tables and in front of windows are largely unscanned, as illustrated in Fig. 2 (top right). Training losses are computed only over valid (known observed) voxels, encouraging the model to learn a more complete latent representation from partial observations while ignoring unobserved regions. The total loss is

$$\mathcal{L}_{\text{vae}} = \mathcal{L}_{\text{tsdf}} + \mathcal{L}_{\text{cat}} + \lambda_{\text{KL}}\, D_{\text{KL}}\big(q(\mathbf{z} \mid \mathbf{v}) \,\|\, p(\mathbf{z})\big), \tag{1}$$

where $\mathcal{L}_{\text{tsdf}} = \frac{1}{|m_s|} \sum_{j \in m_s} |\hat{v}_j - v_j|$ is the masked ℓ_1 reconstruction loss, $\mathcal{L}_{\text{cat}} = -\frac{1}{|m|} \sum_{j \in m} [y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)]$ is the masked binary cross-entropy for voxel occupancy classification, where j indexes voxels, 1 represents surface,

and 0 represents empty space. The mask m_s denotes the set of surface voxels, m denotes all known voxels (surface m_s and empty space m_e), the loss weight λ_{KL} weights the KL regularization toward the standard normal prior.

3.3 Visibility-Aware Masked Flow Matching

With the observed scene regions encoded into a compact geometry latent space, we train a conditional generative model $\mathcal{G}_\psi$, parameterized by ψ, to generate these geometry latents, conditioned on a 3D semantic scene layout, as shown in Fig. 2(b). By observing many partial examples of objects (*e.g.*, chairs, cabinets) from different angles across diverse real-world 3D scenes, the model learns the underlying distribution of object and scene geometry. The 3D semantic layout, specifying object spatial extents and categories, provides crucial coarse structural cues for inferring plausible geometry in unobserved regions.

We represent the 3D semantic layout condition as a set of axis-aligned object bounding boxes with semantic labels $\mathcal{B} = \{\mathbf{b}_k = (\mathbf{c}_k, \mathbf{s}_k, l_k)\}_{k=1}^{K}$, where each box $\mathbf{b}_k$ is defined by its centroid $\mathbf{c}_k$, size $\mathbf{s}_k$, and semantic label l_k, and k indexes objects in the scene. For each scene chunk, we collect all intersecting object bounding boxes to form $\mathcal{B}$. We encode each instance's semantic label with a CLIP text encoder [44]. The resulting semantic features are painted into a 3D layout map according to each box's spatial extent. This layout map is spatially aligned with the geometry latent grid and embedded into a layout token sequence, serving as the conditioning signal.

We formulate scene geometry generation with flow matching [37, 38], using a sparse transformer as the generative model backbone $\mathcal{G}_\psi$. The model learns a velocity field that transports samples from a Gaussian prior p_0 to the target geometry distribution p_{geo}. Self-attention [15, 16] is applied across all geometry and layout tokens jointly. The training loss is

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\mathbf{z}_0 \sim p_0,\, \mathbf{z}_1 \sim p_{\text{geo}},\, t \sim \mathcal{U}(0,1)} \left[\left\| \mathcal{G}_\psi\left(\mathbf{z}_t, t, \mathcal{B}\right) - \left(\mathbf{z}_1 - \mathbf{z}_0\right) \right\|^2 \right], \tag{2}$$

where $\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\mathbf{z}_1$ is the linear interpolant between noise $\mathbf{z}_0$ and target latent $\mathbf{z}_1$, and the model predicts the velocity $\mathbf{z}_1 - \mathbf{z}_0$. During training, geometry tokens corresponding to unobserved regions (as indicated by the visibility mask from the sparse VAE) are excluded from the flow matching objective, so the model learns to generate geometry only for observed regions while remaining agnostic to missing areas. Following classifier-free guidance [24], we randomly drop $\mathcal{B}$ during training, making the layout condition optional at test time, as validated by our experiments. We provide more comprehensive training details in the supplementary material.

3.4 Scan Completion and Multi-modal Generation

3D Scan Completion. To infer the missing structure of input 3D scans, we leverage the generative priors learned from partially observed, real-world scene

geometry (Sec. 3.3), as shown in Fig. 2(c). We adopt a self-supervised formulation [13] by fine-tuning the pretrained generative model $\mathcal{G}_\psi$ on existing 3D scan training data, but with varying levels of incompleteness. Concretely, we remove a portion of the depth frames from an existing partial scan $\mathbf{v}$ to create a more incomplete variant $\mathbf{v}^{\mathrm{p}}$, serving as input condition. The original partial scan $\mathbf{v}$ is used as the target for supervision. This encourages the model to learn to correlated different levels of incompleteness within the same scene and predict the removed geometry.

We add a ControlNet [68] branch $\mathcal{C}_\phi$, parameterized by ϕ which is initialized from the pretrained weights of $\mathcal{G}_\psi$, to inject the more incomplete scan condition $\mathbf{v}^{\mathrm{p}}$ into the flow matching model $\mathcal{G}_\psi$ during fine-tuning, without modifying the pretrained flow matching weights. This conditioning provides the scene context and geometric cues for the synthesis of unobserved regions, while ensuring observed areas are preserved.

Formally, the more partial scan $\mathbf{v}^{\mathrm{p}}$ is first encoded by the frozen VAE encoder $\mathcal{E}_\tau$ into a latent condition $\mathbf{z}^{\mathrm{p}} = \mathcal{E}_\tau(\mathbf{v}^{\mathrm{p}})$. The ControlNet branch $\mathcal{C}_\phi$ processes $\mathbf{z}^{\mathrm{p}}$ and injects per-layer control signals into the frozen base model [68]. The fine-tuning objective follows the same flow matching loss:

$$\mathcal{L}_{\mathrm{sc}} = \mathbb{E}_{\mathbf{z}_0,\,\mathbf{z}_1,\,t}\left[\left\|\mathcal{G}_\psi(\mathbf{z}_t, t, \mathcal{B}, \mathcal{C}_\phi(\mathbf{z}^{\mathrm{p}})) - (\mathbf{z}_1 - \mathbf{z}_0)\right\|^2\right], \tag{3}$$

where only the ControlNet parameters ϕ are optimized while the pretrained weights of $\mathcal{G}_\psi$ remain frozen. During training, the layout condition $\mathcal{B}$ is derived from the ground-truth semantic annotations available in the datasets. At inference on novel scans, $\mathcal{B}$ is optional as generation in Sec. 3.3, as validated by our experiments.

Text-to-3D Scene Generation. While optimized for real-world scene completion, our generative model backbone $\mathcal{G}_\psi$ is flexible and generalizes to 3D layouts converted from other conditioning modalities, such as natural language prompts, enabling high-fidelity scene generation from scratch, as shown in Fig. 2(d). For instance, a given text prompt like "*Generate a studio with a single bed, ...*" can be first analyzed by a large language model (LLM) [1] and then translated into a structured set of semantic labels and 3D bounding boxes. To improve robustness to diverse object labels, we augment each label with LLM-generated synonyms, abbreviations, plurals, and typos. Alternatively, users can directly specify or adjust 3D layout information to define scene structure.

Generating Coherent, Large-Scale 3D Scenes. While our approach is trained on chunks for efficiency, we can directly generalize to large-scale scan completion or generation of arbitrary-sized scenes by applying MultiDiffusion [2], similar to LT3SD [41]. Specifically, we employ a tiled generation strategy across multiple overlapping 256^3 chunks. At each timestep, we split the 3D scene into chunks with an overlap ratio of 0.2 with respect to the chunk size and take each denoising step on all chunks simultaneously. Then, an average-based blending strategy is applied for the overlapping regions to ensure seamless fusion between adjacent chunks. Finally, the generated chunk-wise latent representations are decoded and fused into a global large-scale and multi-room 3D geometry.

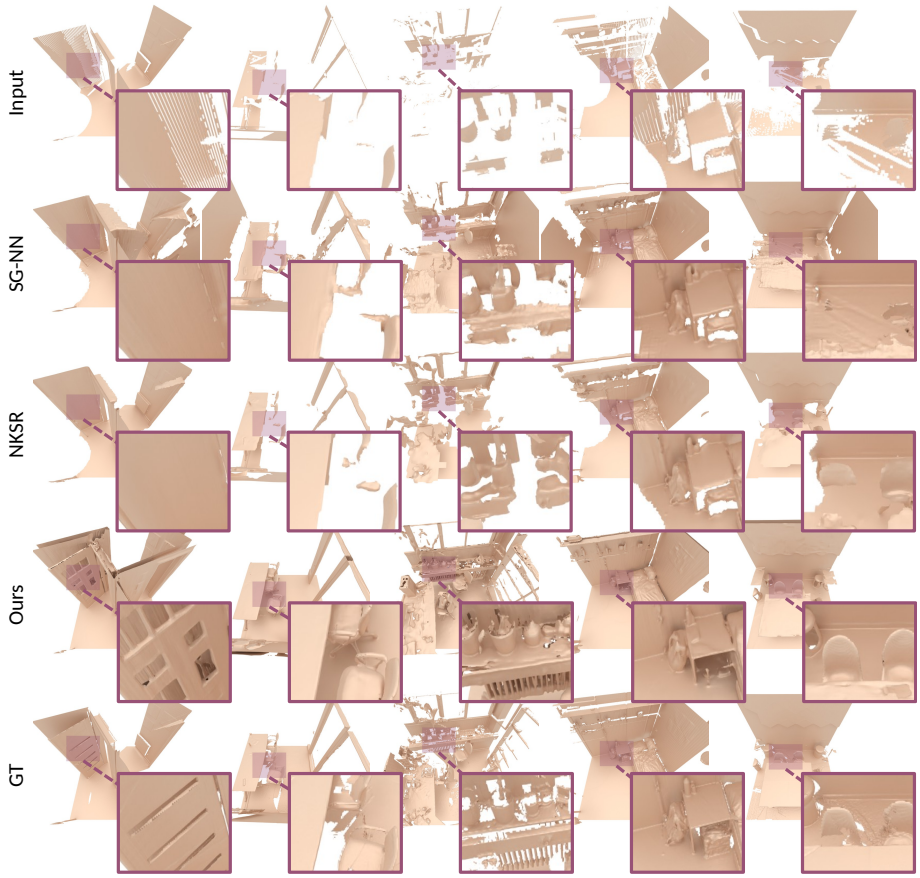

Fig. 3: Qualitative comparison on 3D scan completion. Given real-world partial depth scans from ScanNet++ [65] and ARKitScenes [3], Seen2Scene learns realistic, high-fidelity priors from incomplete real-scan data to produce much more complete, detailed geometry than baselines.

4 Experiments

4.1 Implementation Details

We train on ScanNet++ [65], ARKitScenes [3, 30], and 3D-FRONT [20], which together provide large-scale real-world and synthetic indoor scenes with object annotations. We apply masked training on all three datasets. For 3D-FRONT, we generate synthetic partial scan data by randomly sampling cameras [17] and render depth maps to simulate real-world scanning. TSDF volumes [11] are reconstructed using VDBFusion [55] with a voxel size of 1.1 cm and truncation of 3× voxel size, then divided into chunks of resolution 256^3. Object names are encoded with CLIP ViT-B/32 [44]. The sparse VAE uses fVDB [56] residual convolutional blocks with 8× downsampling, and 8 latent channels. The SparseDiT has 28 transformer blocks and uses RoPE [51] positional encoding. Both stages

Table 1: Quantitative evaluation of scan completion. We compare Seen2Scene with SG-NN [13] and NKSR [26]: CD on 3D-FRONT [20], and L2, TMD, and U3D-FPD on ScanNet++ [65], ARKitScenes [3], and 3D-FRONT [20].

Method	CD$_{\times 10^{-2}}\downarrow$	L2$_{\times 10^{-4}}\downarrow$	TMD$_{\times 10^{-2}}\uparrow$	U3D-FPD$_{\times 10^{-2}}\downarrow$
SG-NN [13]	10.77	1.90	0.00	15.41
NKSR [26]	5.22	2.55	0.00	21.78
Seen2Scene (w/o bbox)	**1.92**	0.53	1.69	9.57
Seen2Scene	2.05	**0.51**	**3.33**	**7.79**

are trained with AdamW [39], a learning rate of 10^{-4}, cosine annealing with 1000 warmup steps, and a batch size of 64 on 4 NVIDIA H100 GPUs with BF16 mixed precision. At inference, we use 50-step Euler sampling with classifier-free guidance [24] at scale 3.0. More data processing details are in the supplementary.

Evaluation Metrics. For scan completion, we compare with the overall $4,000$ samples from the three datasets. We compute Chamfer distance (CD) between uniformly sampled pointclouds from the meshes, and metrics that can be computed only in the observed region of target scans: L2 distance and Total Mutual Difference (TMD) which computes diversity across two completions via pairwise CD. Note that the CD metric is only computed on 3D-FRONT [20] which contains complete mesh ground truth, while other metrics are computed on ScanNet++ [65], ARKitScenes [3, 30], and 3D-FRONT [20]. For scene generation, we report DINOv2-FID [42] computed from $4,000$ patches with 30 views per patch with different elevations, Uni3D Fréchet Point Cloud Distance (U3D-FPD) [70], which measures distributional similarity in a semantically aligned 3D feature space, and a vision-language model (VLM)-based perceptual score that aggregates geometric quality, structural connectivity, semantic coherence, and content diversity using Qwen3-VL-8B-Instruct [63] (prompt details in the supplementary). Note that all the baselines are developed for 3D-FRONT only, so we compute all the metrics only on 3D-FRONT.

4.2 3D Scan Completion

Given partial depth scans, Seen2Scene completes the scene by generating plausible geometry for unobserved regions via the ControlNet branch. We compare SG-NN [13], NKSR [26], ours without object bounding boxes $\mathcal{B}$, and ours with $\mathcal{B}$ in Tab. 1. Our model without bounding boxes achieves comparable performance to its bounding box-conditioned counterpart, and even yields better overall structural correctness (CD). This may be because bounding box information is redundant and potentially disruptive for completing small missing regions (e.g., holes on object surfaces), yet beneficial for recovering large unobserved regions — as illustrated by the two office chairs at the top of the second sample in Fig. 4. Since SG-NN and NKSR are both deterministic, their TMD values are zero. Fig. 3 shows qualitative comparisons: our method produces more complete and structurally coherent completions. Seen2Scene can also complete large-scale

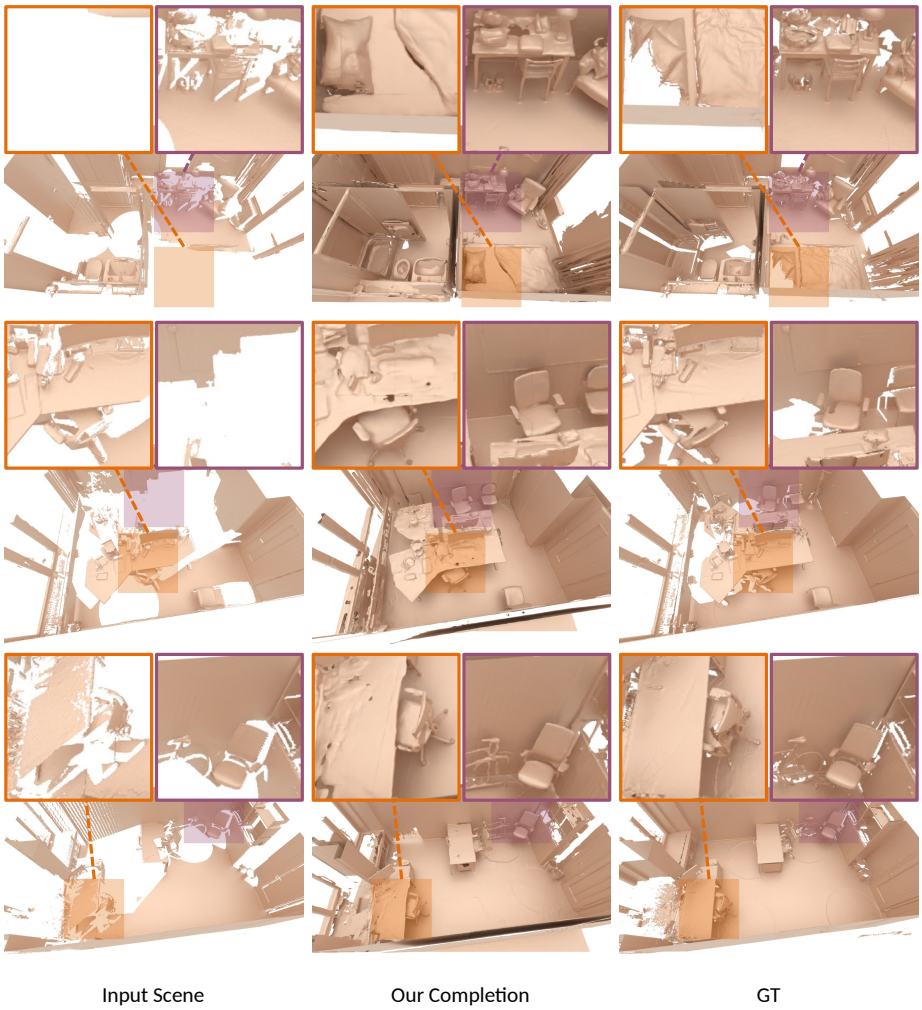

Fig. 4: 3D scene completion. Seen2Scene can complete large-scale scenes by generating geometry for unobserved regions across multiple chunks, guided by partial scan conditions injected into the ControlNet branch. Results are shown on ScanNet++ [65] and ARKitScenes [3].

scenes by generating geometry for unobserved regions across multiple chunks. Fig. 4 shows that our method generates high-fidelity, geometrically complete, and globally consistent reconstructions in multi-room environments. We provide more results in the supplementary.

4.3 3D Scene Generation

We evaluate layout-conditioned 3D scene generation against BlockFusion [59], LT3SD [41], and WorldGrow [33]. Since LT3SD and WorldGrow are unconditional generators while BlockFusion and ours are layout-conditioned, we ensure

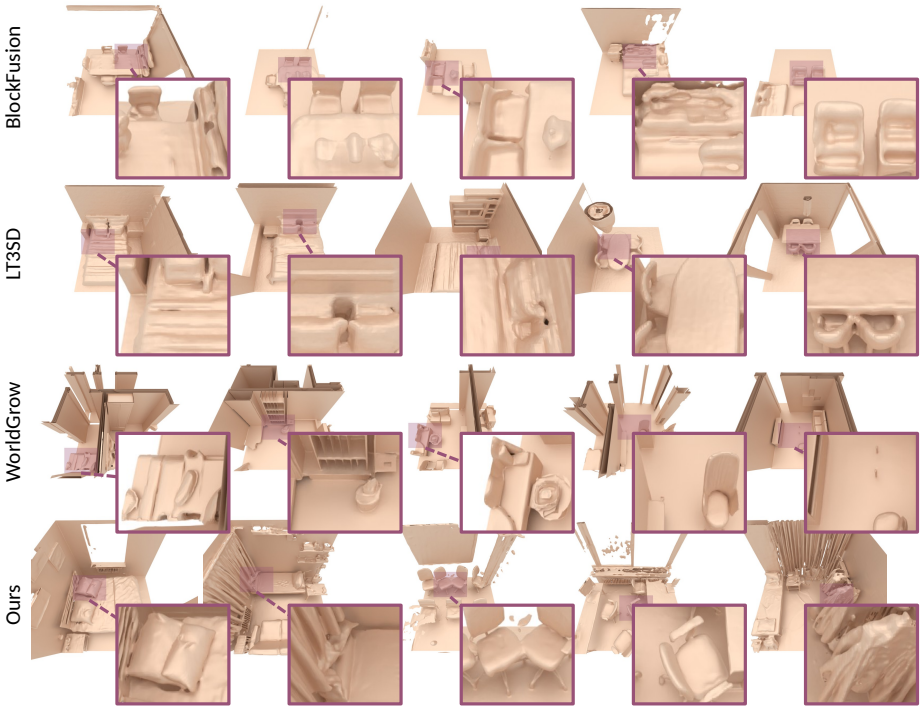

Fig. 5: Qualitative comparison on layout-conditioned 3D scene generation. Our method produces more geometrically detailed and semantically coherent scenes compared to BlockFusion [59], LT3SD [41], and WorldGrow [33]. Layouts are from ScanNet++ [65] and ARKitScenes [3].

a fair comparison by splitting the 3D-FRONT test set into two disjoint subsets: layouts from the first subset serve as conditions for BlockFusion and Seen2Scene, while the second subset is used as the reference distribution for all methods.

Table 2: Quantitative comparison on 3D scene generation. We report DINOv2 Fréchet Inception Distance (FID), Uni3D Fréchet Point Cloud Distance (FPD), and VLM Quality Score assessed by Qwen3-VL on a 0–10 scale 3D-FRONT [20] only.

Method	DINOv2-FID$_{\times 10^2}\downarrow$	U3D-FPD$_{\times 10^2}\downarrow$	VLM Score$\uparrow$
BlockFusion [59]	12.92	28.30	3.51
LT3SD [41]	3.51	28.55	5.20
WorldGrow [33]	4.44	50.83	4.69
Seen2Scene (Synthetic only)	**2.73**	**11.55**	5.61

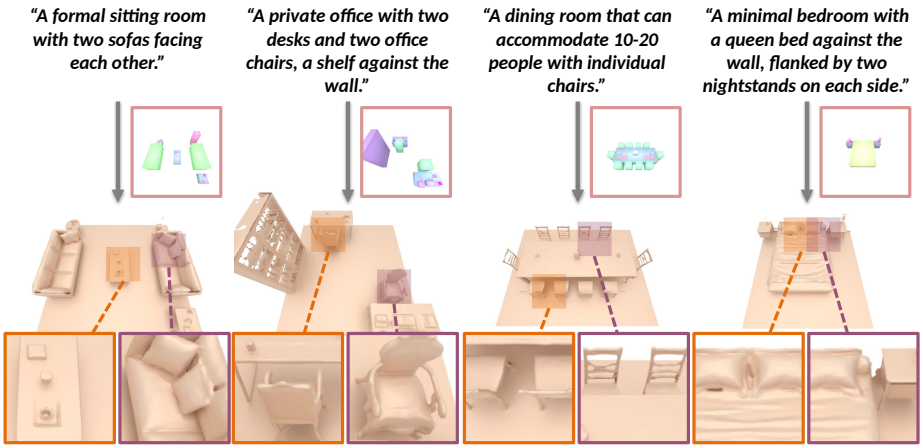

Fig. 6: Text to scene. Given a text description, Seen2Scene generates a 3D object layout first via an LLM and then synthesizes high-fidelity scene geometry from scratch conditioned on the layout.

As shown in Tab. 2, Seen2Scene achieves the best performance across all three metrics. Our method obtains the lowest DINOv2-FID, indicating that our generated scenes are perceptually closest to high-quality ground-truth meshes. The U3D-FPD gap demonstrates that the generated 3D geometry is distributionally well-aligned with ground-truth scenes in a semantically meaningful feature space. The combination of our sparse transformer operating on compact TSDF latents and CLIP-based layout conditioning jointly captures both fine-grained geometric structure and high-level semantic consistency. This improvement can be attributed to our visibility-guided flow matching, which learns geometry distributions directly from real-world partial scans rather than relying solely on synthetic data, producing more realistic surface details and object arrangements. Qualitatively, Fig. 5 shows that our method generates more geometrically detailed, semantically coherent, and realistic scenes, whereas baselines tend to produce over-smoothed surfaces, noticeable artifacts, axis-aligned objects, and clean layout.

Seen2Scene generates 3D scenes from scratch conditioned on text descriptions by first producing an object layout (3D bounding boxes with semantic labels) and then synthesizing geometry conditioned on this layout (Sec. 3.4). Fig. 6 shows that our approach robustly generalizes to 3D layouts inferred by an LLM.

4.4 Ablation Studies

Impact of Masked Training. We ablate the effect of masked training on our final generated scenes in Tab. 3 and Fig. 7. Masked training is crucial to extend generative models to real scans. With masked training, the sparse VAE better reconstructs input partial 3D scans (Tab. 3 - Reconstruction), and the SparseDiT synthesizes scene geometry closer to the reference distribution (Tab. 3

Table 3: Masked training ablation. Reconstruction metrics (L1, L2, CD) assess the autoencoder and generation metrics (U3D-FPD, VLM) assess the flow matching model; CD is computed on 3D-FRONT only, while others are evaluated across datasets.

Configuration	Reconstruction			Generation	
	L1$\times 10^{-4}\downarrow$	L2$\times 10^{-6}\downarrow$	CD$\times 10^{-3}\downarrow$	U3D-FPD$\times 10^{2}\downarrow$	VLM Score$\uparrow$
w/o masked training.	3.24	7.1	10.64	20.28	5.66
Seen2Scene (full)	**2.01**	**2.0**	**9.04**	**14.56**	**5.70**

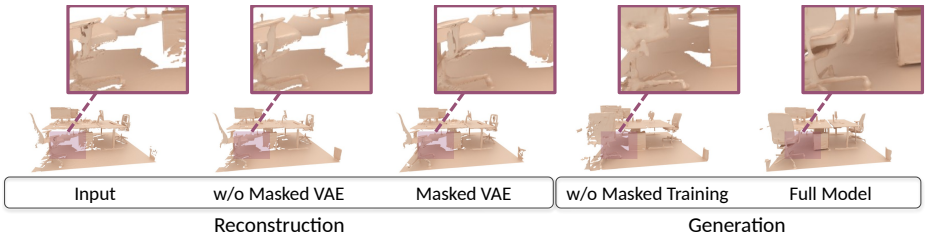

Fig. 7: Qualitative ablation on masked training. Comparison of generated scenes with and without masked flow matching during training. Without two stages of masked training, the model tends to generate holes (under the table) following the training incomplete scans.

- Generation). The reason is that without masked training, the model wrongly treats unobserved space (under the table) as empty space and is trained to generate a wrong distribution.

Effect of Training on Real Data. We ablate the effect of real-world scan data in Tab. 4 and Fig. 8. For the synthetic-only setting, we train the model only on the synthetic 3D-FRONT dataset, which has limited object categories and mostly axis-aligned scenes and objects, leading to a large domain gap with the real world. As shown in Fig. 8, when provided with a complex real layout (*e.g.*, "clothes on an office chair" as shown in the ground-truth), the model trained with only synthetic data fails to generalize. The VLM-based score confirms that the full model trained on all datasets exhibit superior overall quality in terms of semantic realism, geometric plausibility, and structural coherence.

Table 4: Ablation study on generation. We evaluate the contribution of each design choice by removing it from the full model on ScanNet++ [65], ARKitScenes [3], and 3D-FRONT [20].

Configuration	DINOv2-FID$\times 10^{2}\downarrow$	U3D-FPD$\times 10^{2}\downarrow$	VLM Score$\uparrow$
Synthetic only	2.90	21.49	5.61
Discrete Category	3.17	21.46	5.64
Without Label Synonym	2.14	18.77	5.62
Seen2Scene (full)	**1.90**	**14.56**	**5.70**

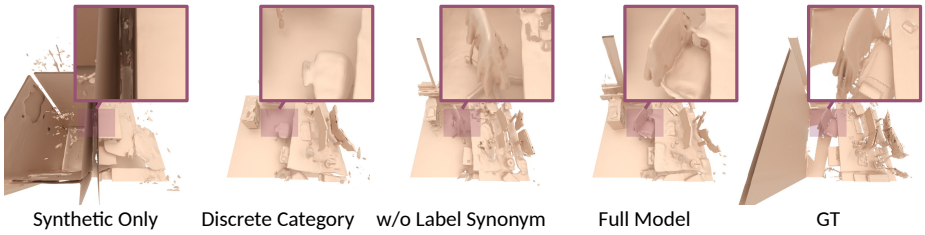

Fig. 8: Qualitative ablation on real data. Training the model only on 3D-FRONT [20], or with limited categories, or without label synonym augmentation, reduces the model's generalization ability to diverse object layouts, like "clothes on an office chair."

Open Vocabulary Semantic Encodings. We ablate the effect of using CLIP embedding [44] instead of one-hot semantic labels which has a fixed number of categories in Tab. 4 and Fig. 8. To encode discrete category labels across datasets, we need to map different object labels to the shared object categories. This category mapping is challenging because the same object labels can be categorized into different categories, i.e., "TV" can also be categorized as "screen", "monitor", or "television". We instead use the CLIP model to embed each object label into a feature vector. This significantly improves the model performance on real datasets. As shown in Fig. 8, the model using discrete category fails to generate the chair with clothes on it.

Label Synonym. We ablate the effect of using label synonym augmentation in Tab. 4 and Fig. 8. To augment the textual object labels, we prompt an LLM [1] to provide 20 variations for each object label in the dataset (prompt details in the supplementary), including: synonyms ("sofa" → "couch"), abbreviations ("mobile device" → "M"), hyphenated ("bag luggage" → "bag-luggage"), plural ("plant" → "plants"), and typos ("ceiling" → "cielin"). Without the augmentation, the model struggles with diverse object labels as shown in Fig. 8, where the model only generates a partial office chair.

5 Conclusion

We presented Seen2Scene, which introduces visibility-aware masked flow matching for 3D scan completion and generation. Our approach combines a masked sparse VAE that compresses TSDF volumes into compact geometry tokens with a sparse transformer that generates these tokens via flow matching conditioned on semantic layouts. By training jointly on synthetic and real-world scan data with a masked training strategy that excludes unseen regions from supervision, our model learns to complete missing geometry from incomplete observations. Experiments on 3D-FRONT, ScanNet++, and ARKitScenes demonstrate that Seen2Scene produces coherent, complete scenes and supports controllable generation including text-conditioned 3D scene synthesis in addition to scan completion.

Acknowledgements

This work was supported by the ERC Starting Grant SpatialSem (101076253). Matthias Nießner was supported by the ERC Starting Grant Scan2CAD (804724).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 7, 14
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: fusing diffusion paths for controlled image generation. In: Proceedings of the 40th International Conference on Machine Learning. pp. 1737–1752 (2023) 7
3. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. arXiv preprint arXiv:2111.08897 (2021) 2, 4, 8, 9, 10, 11, 13
4. Bokhovkin, A., Meng, Q., Tulsiani, S., Dai, A.: Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 628–639 (2025) 2, 3
5. Cao, A.Q., Dai, A., De Charette, R.: Pasco: Urban 3d panoptic scene completion with uncertainty awareness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14554–14564 (2024) 3
6. Chen, M., Wang, L., Ao, S., Zhang, Y., Xu, K., Guo, Y.: Layout2scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. arXiv preprint arXiv:2501.02519 (2025) 3
7. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4193–4202 (2020) 3
8. Chen, Z., Tang, J., Dong, Y., Cao, Z., Hong, F., Lan, Y., Wang, T., Xie, H., Wu, T., Saito, S., et al.: 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26576–26586 (2025) 3
9. Chou, G., Bahat, Y., Heide, F.: Diffusion-sdf: Conditional generative modeling of signed distance functions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2262–2272 (2023) 3
10. Chu, R., Xie, E., Mo, S., Li, Z., Nießner, M., Fu, C.W., Jia, J.: Diffcomplete: Diffusion-based generative 3d shape completion. Advances in neural information processing systems **36**, 75951–75966 (2023) 3
11. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312 (1996) 4, 5, 8
12. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 2, 4
13. Dai, A., Diller, C., Nießner, M.: Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 849–858 (2020) 3, 5, 7, 9

14. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Nießner, M.: Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2018) 3
15. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations (ICLR) (2024) 6
16. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 6
17. Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K.H., Humt, M., Triebel, R.: Blenderproc2: A procedural pipeline for photorealistic rendering. Journal of Open Source Software **8**(82), 4901 (2023). https://doi.org/10.21105/joss.04901, https://doi.org/10.21105/joss.04901 8
18. Erkoç, Z., Ma, F., Shan, Q., Nießner, M., Dai, A.: Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14300–14310 (2023) 3
19. Fang, C., Li, H., Liang, Y., Zheng, J., Mao, Y., Liu, Y., Tang, R., Zhou, Z., Tan, P.: Spatialgen: Layout-guided 3d indoor scene generation. arXiv preprint arXiv:2509.14981 (2025) 3
20. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021) 2, 8, 9, 11, 13, 14
21. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision pp. 1–25 (2021) 2
22. Han, Z., Higashita, R., Liu, J.: Voic: Visible-occluded decoupling for monocular 3d semantic scene completion. arXiv preprint arXiv:2512.18954 (2025) 3
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3
24. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 6, 9
25. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7909–7920 (2023) 3
26. Huang, J., Gojcic, Z., Atzmon, M., Litany, O., Fidler, S., Williams, F.: Neural kernel surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4369–4379 (2023) 3, 9
27. Huang, J., Artemov, A., Chen, Y., Zhi, S., Xu, K., Nießner, M.: Ssr-2d: semantic 3d scene reconstruction from 2d images. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 8486–8501 (2024) 3
28. Ju, X., Huang, Z., Li, Y., Zhang, G., Qiao, Y., Li, H.: Diffindscene: Diffusion-based high-quality 3d indoor scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4526–4535 (2024) 3
29. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023) 3
30. Lazarow, J., Griffiths, D., Kohavi, G., Crespo, F., Dehghan, A.: Cubify anything: Scaling indoor 3d object detection. arXiv preprint arXiv:2412.04458 (2024) 2, 8, 9

31. Li, B., Deng, J., Zhang, W., Liang, Z., Du, D., Jin, X., Zeng, W.: Hierarchical temporal context learning for camera-based semantic scene completion. In: European Conference on Computer Vision. pp. 131–148. Springer (2024) 3
32. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3351–3359 (2020) 3
33. Li, S., Yang, C., Fang, J., Yi, T., Lu, J., Cen, J., Xie, L., Shen, W., et al.: Worldgrow: Generating infinite 3d world. arXiv preprint arXiv:2510.21682 (2025) 10, 11
34. Li, Y., Li, S., Liu, X., Gong, M., Li, K., Chen, N., Wang, Z., Li, Z., Jiang, T., Yu, F., Wang, Y., Zhao, H., Yu, Z., Feng, C.: Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2024) 3
35. Li, Z., Wang, Y., Zheng, H., Luo, Y., Wen, B.: Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. arXiv preprint arXiv:2505.14521 (2025) 3
36. Ling, L., Lin, C.H., Lin, T.Y., Ding, Y., Zeng, Y., Sheng, Y., Ge, Y., Liu, M.Y., Bera, A., Li, Z.: Scenethesis: A language and vision agentic framework for 3d scene generation. arXiv preprint arXiv:2505.02836 (2025) 3
37. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 2, 3, 6
38. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022) 2, 3, 6
39. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 9
40. Lu, H., Su, Y., Zhang, X., Hu, H.: One step closer: Creating the future to boost monocular semantic scene completion. arXiv preprint arXiv:2507.13801 (2025) 3
41. Meng, Q., Li, L., Nießner, M., Dai, A.: Lt3sd: Latent trees for 3d scene diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 650–660 (2025) 2, 3, 7, 10, 11
42. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 9
43. Pfaff, N., Cohn, T., Zakharov, S., Cory, R., Tedrake, R.: Scenesmith: Agentic generation of simulation-ready indoor scenes. arXiv preprint arXiv:2602.09153 (2026) 3
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021) 6, 8, 14
45. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., Newell, A., Law, H., Goyal, A., Yang, K., Deng, J.: Infinite photorealistic worlds using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12630–12641 (2023) 3
46. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., Ma, Z., Deng, J.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21783–21794 (June 2024) 3

47. Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4209–4219 (2024) 2, 3
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3
49. Schaefer, S., Galvis, J.D., Zuo, X., Leutengger, S.: Sc-diff: 3d shape completion with latent diffusion models. arXiv preprint arXiv:2403.12470 (2024) 3
50. Selvakumar, A., Bharadwaj, M.: Fake it to make it: Virtual multiviews to enhance monocular indoor semantic scene completion. arXiv preprint arXiv:2503.05086 (2025) 3
51. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 8
52. Sun, X., Goel, D., Chang, A.X.: Semlayoutdiff: Semantic layout generation with diffusion model for indoor scene synthesis. arXiv preprint arXiv:2508.18597 (2025) 3
53. Tang, J., Nie, Y., Markhasin, L., Dai, A., Thies, J., Nießner, M.: Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20507–20518 (2024) 3
54. Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al.: Lion: Latent point diffusion models for 3d shape generation. Advances in Neural Information Processing Systems **35**, 10021–10039 (2022) 3
55. Vizzo, I., Guadagnino, T., Behley, J., Stachniss, C.: Vdbfusion: Flexible and efficient tsdf integration of range sensor data. Sensors **22**(3) (2022). https://doi.org/10.3390/s22031296, https://www.mdpi.com/1424-8220/22/3/1296 8
56. Williams, F., Gojcic, Z., Schindler, K., Litany, O., Fidler, S., Museth, K.: fVDB: A deep-learning framework for sparse, large-scale, and high-performance spatial intelligence. In: ACM SIGGRAPH Asia (2024) 5, 8
57. Wu, Q., Iliash, D., Ritchie, D., Savva, M., Chang, A.X.: Diorama: Unleashing zero-shot single-view 3d indoor scene modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8896–8907 (2025) 3
58. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Yang, Y., Bao, Y., Qian, J., Zhu, S., Cao, X., Torr, P., et al.: Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. arXiv preprint arXiv:2505.17412 (2025) 2, 3
59. Wu, Z., Li, Y., Yan, H., Shang, T., Sun, W., Wang, S., Cui, R., Liu, W., Sato, H., Li, H., et al.: Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. ACM Transactions on Graphics (ToG) **43**(4), 1–17 (2024) 2, 3, 10, 11
60. Xi, Z., Chen, H.X., Xue, N., Yan, H., Feng, Q.Y., Kara, L.B., Jorge, J., Xu, Q.C.: Flowssc: Universal generative monocular semantic scene completion via one-step latent diffusion. arXiv preprint arXiv:2601.15250 (2026) 3
61. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. arXiv preprint arXiv:2512.14692 (2025) 2, 3
62. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025) 2, 3

63. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 9
64. Yang, Y., Jia, B., Zhang, S., Huang, S.: Sceneweaver: All-in-one 3d scene synthesis with an extensible and self-reflective agent. arXiv preprint arXiv:2509.20414 (2025) 3
65. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the International Conference on Computer Vision (ICCV) (2023) 2, 4, 8, 9, 10, 11, 13
66. Zhai, G., Örnek, E.P., Chen, D.Z., Liao, R., Di, Y., Navab, N., Tombari, F., Busam, B.: Echoscene: Indoor scene generation via information echo over scene graph diffusion. In: European Conference on Computer Vision. pp. 167–184. Springer (2024) 3
67. Zhang, D., Williams, F., Gojcic, Z., Kreis, K., Fidler, S., Kim, Y.M., Kar, A.: Outdoor scene extrapolation with hierarchical generative cellular automata. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20145–20154 (2024) 3
68. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) 4, 7
69. Zhang, X., Li, N., Dai, A.: Dnf: Unconditional 4d generation with dictionary-based neural fields. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26047–26056 (2025) 3
70. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. In: The Twelfth International Conference on Learning Representations (2024) 9
71. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5826–5835 (2021) 3