ROOMTOKENS: MASKED DIFFUSION OVER STRUCTURED TOKENS FOR INDOOR SCENE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating 3D indoor scenes requires a representation that captures object semantics, instance identity, height structure, and orientation while respecting room geometry. Methods that model scenes as sequences of object vectors learn spatial constraints implicitly, leaving collisions and boundary violations to losses or post-processing; semantic-map methods make occupancy explicit, but collapse instance and directional structure. We introduce RoomTokens, a sliced structured-token representation in which each token stores semantic class, instance-color identity, and facing field. This representation allows a scene to be generated as a discrete grid while preserving cues for instance grouping, orientation estimation, and 3D extent prediction. To learn the RoomToken distribution, we propose MDIRT (Masked Diffusion over RoomTokens), a masked discrete diffusion model that denoises semantic/instance-color/facing channels jointly with a shared cross-slice backbone. Architectural tokens are clamped as explicit floor-plan constraints, so the model samples furniture layouts that respect room geometry while retaining instance and object-attribute structure. Experiments on 3D-FRONT show that RoomTokens with MDIRT improve layout fidelity and physical consistency over sequence/vector- and map-based methods, suggesting structured spatial tokens as a practical choice for discrete 3D scene synthesis.

1 INTRODUCTION

Automatic generation of 3D scenes is increasingly important, as it allows for scaling up of training data for 3D scene understanding and embodied AI, and provides content for applications such as augmented reality, gaming, and entertainment. Compositional 3D indoor scene generation (Tam et al., 2026b) formulates the problem as an object arrangement problem, where we must decide on object categories, individual object identities, placement, and orientation while respecting the geometry of the room. To ensure that generated scenes are plausible, it is important to model the spatial structure and object-level semantics together.

Most existing methods represent scenes as sets or sequences of object vectors (Paschalidou et al., 2021; Tang et al., 2024; Hu et al., 2026; Lin and Mu, 2024), where each object vector encodes the object category and bounding box of the object. While this representation is compact, it does not make occupancy constraints explicit and cannot ensure that two generated boxes do not occupy the same space. As a result, generated scenes may contain inter-object collisions, wall penetrations, and inadequate clearance for accessing objects and room regions. Prior systems therefore add auxiliary mechanisms, such as intersection penalties in the training objective (Tang et al., 2024), rejection-style resampling of invalid insertions (Ritchie et al., 2019), or optimization-based layout refinement after an initial prediction (Luo et al., 2020). Another line of work (Feng et al., 2023; Yang et al., 2025; Xia et al., 2026) leverages the broad knowledge encoded in LLMs/VLMs for more extensive semantic priors, but their coordinate outputs still require geometric cleanup and potentially multiple passes with a visual critic to ensure a physically plausible scene.

Approaches that generate scenes based on semantic-maps (Yang et al., 2021; Sun et al., 2026) capture the spatial arrangements explicitly by assigning semantic labels to cells on a dense grid. A recent work, SemLayoutDiff (Sun et al., 2026) demonstrated improved floor-plan-conditioned generation with fewer collisions and out-of-boundary objects. However, a single top-down semantic map fails to capture important information needed for 3D scene generation. It collapses vertical information, merges adjacent objects of the same category under connected-component decoding, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

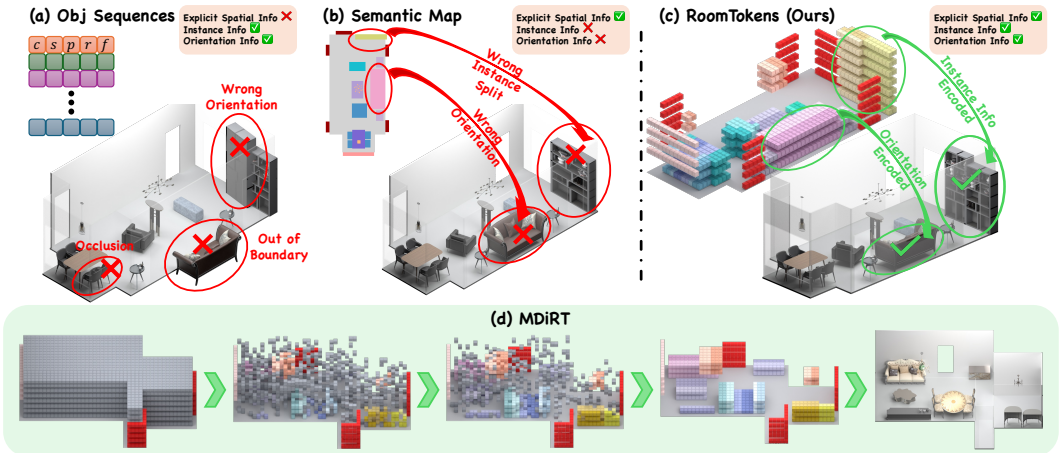

Figure 1: *Top.* Comparison of scene representations for indoor scene generation. Object sequences (a) encode object attributes but lack explicit spatial relationship information. Semantic maps (b) make spatial structure explicit, but lack vertical, instance, and orientation information. We propose RoomTokens (c) to address these missing factors by encoding semantic occupancy, instance identity, and facing cues in a sliced spatial grid. *Bottom.* We introduce MDiRT (d), a masked discrete diffusion model over RoomTokens, which denoises missing token fields while keeping architectural tokens fixed so generated scenes respect room boundaries.

provides no directional cue for estimating object orientation. In this work, we tackle the challenge of how to introduce instance-level and vertical information into these spatial maps.

We introduce **RoomTokens**, a sliced structured-token representation for indoor scenes. We represent a room as a stack of height-aware grids, with each grid cell storing a RoomToken with three fields that capture the semantic class, the instance identity, and facing cues. Encoding the instance identity is challenging as the assignment of a particular ID to an object is arbitrary (i.e., can be permuted or re-assigned), and the number of objects in a scene is not fixed. As our aim is to distinguish spatially adjacent object instances on a planar map (vs. providing an instance ID), we can use four 'colors' (following the four-color theorem (Robertson et al., 1997)) to identify different instances. For orientation, the facing field encodes whether a furniture token occupies the body or front half of an object; non-furniture tokens use none. We find that these two furniture labels are sufficient for decoding orientation. Together, these properties provide sufficient information for a model to generate a discrete spatial layout while preserving cues for instance grouping, orientation estimation, and 3D extent prediction.

To learn the RoomToken distribution, we propose **MDiRT** (Masked Diffusion over RoomTokens), a single-stage masked discrete diffusion model for indoor scene generation. Each token property is embedded as a semantic, instance-color, or facing channel. The channels are processed by a shared cross-slice UNet and decoded jointly. The denoising process maps masked RoomTokens to structured 3D scene layouts while conditioning on and clamping the architectural tokens from the floor plan (Fig. 1 (bottom)), so furniture tokens is generated within the valid interior region defined by the architectural condition. Instead of predicting semantics and then instance-level attributes as in prior work (Sun et al., 2026), this multichannel parameterization learns semantic, instance-color, and facing dependencies through shared features in a single reverse process. Experiments on 3D-FRONT (Fu et al., 2021a) show that MDiRT generates more plausible layouts compared with SemLayoutDiff (Sun et al., 2026). Generated layouts better match reference rooms (FID 49.84 vs. 71.06) and capture more plausible object relations (Pair-Rel 12.13 vs. 20.79). They also contain fewer collisions (COL 12.75 vs. 16.13) and leave doors and windows more unobstructed (OC 82.04 vs. 56.15).

In summary, our contributions are: 1) **A sliced structured RoomToken representation** that stores semantic class, graph-color instance identity, and facing information in each height-aware grid token. 2) **A single-stage masked discrete diffusion model, MDiRT, for RoomTokens** that clamps the architectural condition throughout denoising so furniture is generated within the valid room interior, while jointly modeling semantic/instance-color/facing channels with a shared cross-slice backbone. 3) **An empirical analysis** showing that explicit instance-color and facing fields improve instance recovery, orientation estimation, and provides for physically plausible scene layouts.

2 Related Work

Object-vector and relational scene synthesis. Many indoor scene synthesis methods represent a room as a set or sequence of objects with attributes specifying the semantic class, position, and orientation of objects. Early neural and graph-based approaches learn spatial priors over object arrangements (Wang et al., 2018; Ritchie et al., 2019; Li et al., 2019; Wang et al., 2019; Zhang et al., 2020), while recent autoregressive and diffusion methods predict the object attributes directly (Wang et al., 2021; Paschalidou et al., 2021; Sun et al., 2024b; Tang et al., 2024; Hu et al., 2026; Maillard et al., 2024; Sun et al., 2025b). Other recent work leverages scene graphs, relational triples, or mixed graph diffusion for improved modelsing of object layouts (Lin and Mu, 2024; Zhai et al., 2023; 2024; Bai et al., 2025; Ruiz et al., 2025), and physical or human-aware constraint objectives and optimization (Yi et al., 2023; Yang et al., 2024a; Sun et al., 2024a; Su et al., 2025; Lin et al., 2026). These methods differ in conditioning and priors, but their generated layout is ultimately a set of object bounding boxes. As a result, generated scenes may contain object collisions and boundary overlaps, and additional post-processing is required as the representation does not natively prevent these. MDIRT instead generates a discrete occupancy grid, where each cell has a single semantic assignment and grid-level overlap is therefore explicit. By modeling object occupancy explicitly across height-aware slices, the representation helps reduce inter-object collisions.

LLM and agent-based scene generation. By leveraging LLMs and VLMs, it is possible to condition on language input and access the common sense knowledge of how objects are positioned, stored in these pretrained models. LayoutGPT (Feng et al., 2023) predicts layouts with language models, while Holodeck (Yang et al., 2024b), AnyHome (Wen et al., 2024), SceneTeller (Öcal et al., 2024), and open-universe systems (Aguina-Kang et al., 2024) connect text prompts with object retrieval, constraint solving, or scene assembly. More recent LLM/VLM and agentic methods use program synthesis, visual-language feedback, or multi-agent refinement for broader controllability (Hu et al., 2024; Çelen et al., 2024; Sun et al., 2025a; Yang et al., 2025; Pfaff et al., 2026; Gao et al., 2025). These systems provide flexible interfaces, but coordinate- or program-level outputs still require geometric cleanup for dense indoor layouts. Our focus is complementary: we learn the layout distribution from annotated 3D-FRONT rooms and make the representation itself carry local occupancy, instance, and orientation structure.

Discrete spatial representations for scene synthesis. Discrete spatial representations encode the scene by assigning labels directly to locations on a 2D or 3D lattice, making occupancy constraints local and explicit. Yang et al. (Yang et al., 2021) use voxel labels, Lee et al. (Lee et al., 2023) model scene-scale categorical 3D data, and SemLayoutDiff (Sun et al., 2026) applies multinomial discrete diffusion to a top-down semantic map. This family is closest to ours because it moves generation from unordered boxes toward explicit spatial fields. However, semantic-only spatial grids do not distinguish touching same-class instances, top-down maps collapse vertical structure, and a single class label per cell provides no direct orientation cue. Our representation keeps the grid advantage but upgrades each sliced cell into a structured RoomToken carrying semantic, instance-color, and facing fields. Graph coloring is used to assign instance colors during preprocessing, with the four-color theorem (Appel and Haken, 1977; Robertson et al., 1997) providing the minimum number of colors needed for separating adjacent same-class instances. Similar coloring ideas have been studied for discriminative instance segmentation (Kulikov et al., 2018; Zhang et al., 2025), but remain largely unexplored as compact instance channels for generative modelling.

Discrete diffusion and masked token modeling. Masked discrete diffusion corrupts categorical data by replacing tokens with absorbing mask states and trains a denoiser to reverse that process. Multinomial and masked variants (Hoogeboom et al., 2021; Shi et al., 2024; Lou et al., 2024; Sahoo et al., 2024; Nie et al., 2025) have become effective for discrete sequences and token grids. Layout-specific diffusion models also use discrete or mixed discrete-continuous variables for structured 2D layouts and indoor scene attributes (Chai et al., 2023; Inoue et al., 2023; Levi et al., 2023; Hui et al., 2023; Choi et al., 2026), but these variables are often organized as layout tokens or object attributes rather than as per-cell scene occupancy fields. We extend masked discrete diffusion from a single semantic map or object-attribute tokenization to structured RoomTokens. MDIRT uses a multichannel parameterization for training, where semantic, instance-color, and facing values are denoised together under validity constraints and floor-plan conditioning.

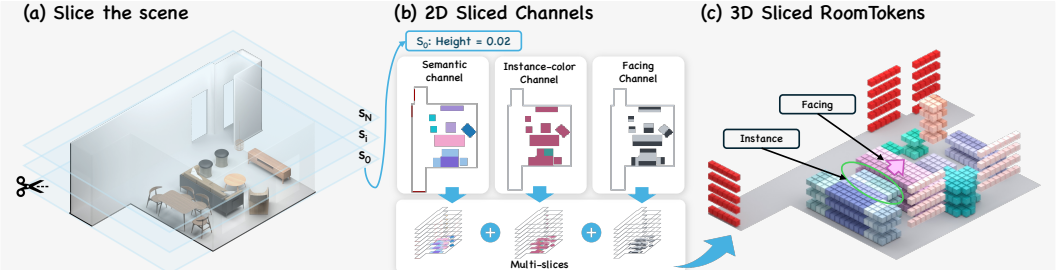

Figure 2: RoomToken data representation. (a) A 3D scene is represented as horizontal slices taken at selected vertical heights. (b) Each slice encodes three aligned 2D channels: semantic class, instance-color identity, and the facing field. Stacking these channels over all heights forms a sliced token volume. (c) The resulting 3D representation of stacked slices (RoomTokens) store semantic, instance, and orientation information, with height derived from the slice index.

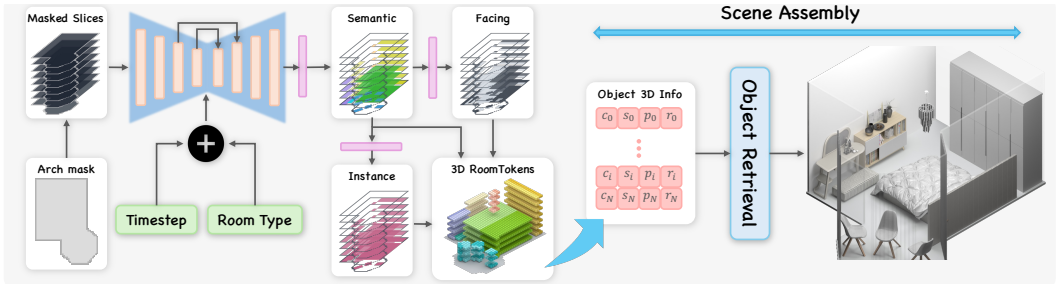

Figure 3: Overview of MDIRT. Given selected height slices and an architectural mask, we clamp the fixed room structure and denoise masked interior RoomTokens with a shared UNet with cross-slice attention. The denoiser predicts semantic, instance-color, and facing slices under semantic-gated validity constraints: architecture/background tokens carry no instance-color or facing cues, while furniture tokens do. The valid RoomToken grid is then grouped into object records, whose category, size, and pose are used to retrieve and place matched 3D assets.

3 METHOD

Given a room type r and an input architectural mask (arch mask), MDIRT generates a complete object-level 3D layout (Fig. 3). With our RoomToken representation, slices at different heights store semantic occupancy with instance and facing cues, allowing objects to be recovered from the grid tokens (Sec. 3.1). We then use a masked diffusion model to clamp the architectural tokens and generate interior tokens (Sec. 3.2). The denoising network couples this process across slices with a shared UNet and cross-slice attention, producing consistent semantic, instance-color, and facing slices (Sec. 3.3). Finally, during scene assembly, the token grid is converted into object records and matching assets are retrieved for the final scene (Sec. 3.4).

3.1 SLICED ROOMTOKENS

To recover object instances and poses from a discrete occupancy map, we encode each room as a stack of sliced RoomTokens. We discretize the floor plane into an $H \times W$ grid and use S selected height slices $\mathcal{Z} = \{\zeta_1, \ldots, \zeta_S\}$ to capture the vertical extent (Fig. 2). We choose these heights by greedy coverage over furniture height intervals in the training set: each object covers a candidate height h if its vertical interval intersects $[h - \tau, h + \tau]$, with τ=0.10m matching the rasterization tolerance. Starting from fixed floor and ceiling anchors, we repeatedly add the candidate height that covers the most previously uncovered objects. We provide dataset-specific grid size, slice heights, and other parameters in Sec. 4.1.

We index a token by $i = (u, v, s)$, where $u \in \{1, \ldots, H\}$, $v \in \{1, \ldots, W\}$, and $s \in \{1, \ldots, S\}$. Each grid location stores a token $\mathbf{x}_i = (y_i, c_i, p_i)$ with semantic class y_i, instance-color label c_i, and facing label p_i. Semantic labels use K_{sem} classes. The instance-color label is either one of C instance colors or the none value $\varnothing_c$; the facing label is body, front, or the none value $\varnothing_p$. The three values must agree with the semantic class. Let Y_{arch} denote architecture/background classes and

Y_{furn} denote furniture classes. A RoomToken $\mathbf{x}_i = (y_i, c_i, p_i)$ is valid only if

$$\text{valid}(y_i, c_i, p_i) \iff \begin{cases} c_i = \varnothing_c \text{ and } p_i = \varnothing_p, & y_i \in Y_{\text{arch}}, \\ c_i \in \{0, \dots, C-1\} \text{ and } p_i \in \{\text{body}, \text{front}\}, & y_i \in Y_{\text{furn}}. \end{cases} \tag{1}$$

This rule ties instance and facing cues to semantic occupancy: architecture/background tokens carry no attributes, while furniture tokens carry both instance color and facing labels.

3.2 Masked Diffusion over RoomTokens

Because the arch mask already specifies the room boundary, the diffusion process should generate only the unknown interior while leaving the observed architecture unchanged. To accomplish this, we use masked diffusion (Shi et al., 2024) In contrast to regular diffusion models that slowly add noise to the input, masked diffusion gradually replaces and hides the interior RoomToken labels with a mask token during the forward process. See Fig. 1 bottom for the reverse process.

Forward process. We formulate the forward masking process from a clean RoomToken cell $\mathbf{x}$ to the partially masked $\mathbf{z}_t$ at time step t as repeatly sampling whether to keep the original token or replace it with a mask token. For our three channel RoomToken $\mathbf{x}_i = (y_i, c_i, p_i)$, we will use three mask tokens $[\text{M}]_y, [\text{M}]_c, [\text{M}]_p$. We draw independent keep indicators m_i for the semantic label, n_i for the instance-color and facing labels from a Bernoulli distribution with keep probability $\alpha(t)$: $m_i, n_i \sim \text{Bernoulli}(\alpha(t))$. Both instance-color and facing labels are revealed only when the semantic label is visible. The indicator $a_i = 1$ marks fixed architectural cells, while $a_i = 0$ marks cells to generate. For $\rho \in \{c, p\}$, let $x_i^c = c_i$ and $x_i^p = p_i$. The forward mask process $q_t(\mathbf{z}_t \mid \mathbf{x}, \mathbf{a})$ is

$$z_{t,i}^y = \begin{cases} y_i, & \text{if } a_i = 1 \text{ or } m_i = 1, \\ [\text{M}]_y, & \text{otherwise,} \end{cases} \qquad z_{t,i}^\rho = \begin{cases} \varnothing_\rho, & \text{if } a_i = 1, \\ x_i^\rho, & \text{if } a_i = 0 \text{ and } m_i = n_i = 1, \\ [\text{M}]_\rho, & \text{otherwise.} \end{cases} \tag{2}$$

Equation (2) masks both attributes whenever an unfixed cell's semantic label is masked, regardless of its ground-truth category. Sampling initializes all unfixed cells with all three fields masked, while the supplied architecture remains fixed.

Reverse process. The reverse process fills masked tokens by stepping from time t to an earlier denoising time $s < t$ under the same constraints. We condition on $\kappa = (r, \mathbf{a})$, where r is the room type and $\mathbf{a}$ is the arch-mask indicator. A learned model $p_\theta(\hat{\mathbf{z}}_s \mid \mathbf{z}_t, t, \kappa)$ proposes semantic, instance-color, and facing labels: $\mathbf{z}_t$ gives the current partially observed grid, while t tells the model the current mask level. We then restore fixed arch-mask tokens and project generated tokens to the valid RoomToken support:

$$\mathbf{z}_{s,i} = \begin{cases} \mathbf{x}_i, & a_i = 1, \\ \Pi_{\text{valid}}(\hat{\mathbf{z}}_{s,i}), & a_i = 0. \end{cases} \tag{3}$$

Here $\hat{\mathbf{z}}_s$ is the model proposal before projection, and Π_{valid} enforces the validity rule in Eq. (1). This projection removes only illegal semantic–attribute combinations; it does not merge instances or change placement. We report invalid-token rates in App. C and ablate Π_{valid} in App. B

3.3 Multichannel Denoising Network

The denoising network maps a masked RoomToken grid to a slice-consistent distribution over semantic, instance-color, and facing slices. Its design shares evidence across token channels, conditions on the arch mask, and exchanges information across height slices before decoding token logits.

Joint RoomToken denoising. The main denoising step predicts the RoomToken values from one shared hidden representation, conditioned on the room type and arch mask through κ. With channel embeddings E_y, E_c, E_p, shared UNet U_θ, and output heads D_y, D_c, D_p, the denoising map is

$$\begin{aligned} \mathbf{e}_{t,i} &= E_y(z_{t,i}^y) + E_c(z_{t,i}^c) + E_p(z_{t,i}^p), \quad \text{for each grid location } i, \\ \mathbf{h}_\theta &= U_\theta(\mathbf{e}_t; t, \kappa), \quad \boldsymbol{\ell}_i^\rho = D_\rho(\mathbf{h}_{\theta,i}), \quad \rho \in \{y, c, p\}. \end{aligned} \tag{4}$$

Here $\mathbf{e}_{t,i}$ is the embedded token, $\mathbf{h}_{\theta,i}$ the hidden feature, $\boldsymbol{\ell}_i^\rho$ the logits for channel ρ, and κ the conditioning input. The clean $\mathbf{x}_i$ and noisy $\mathbf{z}_{t,i}$ variables are discrete categorical states; $\mathbf{e}_{t,i}$ and $\mathbf{h}_{\theta,i}$ are continuous vectors, and the three decoders map the shared vector to categorical channel logits.

Summed embeddings and a shared hidden state parameterize one RoomToken distribution rather than a semantic-then-attribute pipeline.

Cross-slice consistency. Cross-slice attention makes height slices at the same floor location communicate. The UNet treats the $H \times W$ grid as the spatial domain after concatenating S slice embeddings per location. At attention resolutions, a cross-slice block reshapes each (u, v) feature into S tokens, applies attention along the slice dimension, and writes the result back through a residual projection. Thus, predictions for $\{(u, v, s) : s = 1, \ldots, S\}$ exchange evidence about object extent and body/front alignment.

Cross-channel consistency. Cross-channel consistency makes instance-color and facing predictions depend on the semantic prediction. Instance-color and facing logits share the semantic backbone features, and the attribute branch is modulated by a semantic feature branch through learned gates and biases. This makes attributes semantic-aware within the same denoising step, while Sec. 3.2 enforces legal channel combinations during sampling.

Training objective. The training objective supervises both token denoising and the geometric quantities needed for scene assembly. The main terms denoise the masked RoomToken channels, while auxiliary terms calibrate occupancy and the geometric heads used downstream. We optimize

$$\mathcal{L} = \lambda_y \mathcal{L}_y + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_c \mathcal{L}_c + \lambda_p \mathcal{L}_p + \lambda_{\text{occ}} \mathcal{L}_{\text{occ}} + \lambda_{\text{attr}} \mathcal{L}_{\text{attr}}. \tag{5}$$

Here each λ is a scalar weight. $\mathcal{L}_y$ is masked cross-entropy over semantic classes, while $\mathcal{L}_{\text{cls}}$ is an object-count-balanced semantic class loss for the long-tailed furniture distribution. It applies fixed inverse-square-root weights computed from training object counts, keeps architecture/background weights at one, averages semantic cross-entropy over the cells of each ground-truth object, and then averages over objects. This gives rare categories and small instances comparable influence to common or large furniture. $\mathcal{L}_c$ and $\mathcal{L}_p$ are masked cross-entropy losses for instance-color and facing labels on visible furniture cells. $\mathcal{L}_{\text{occ}}$ is binary cross-entropy with logits on foreground occupancy derived from the semantic logits. $\mathcal{L}_{\text{attr}}$ groups instance-level mean-squared regression losses for yaw $(\cos\theta, \sin\theta)$, height, and vertical offset from instance-pooled features. See Sec. D.2 for details.

3.4 Inference and Scene Assembly

At inference, architectural cells from the input arch mask are clamped throughout the reverse process, and MDIRT samples semantic, instance-color, and facing slices for the furnished interior. Because these architectural tokens are never overwritten, the masked discrete diffusion design preserves the architectural condition throughout generation, enabling the model to learn furniture layouts within the valid room interior and yielding zero OOB in our experiments.

We convert the sampled RoomTokens into object records before asset retrieval. After validity projection, furniture cells are grouped by 3D connected components over equal semantic and instance-color labels (y_i, c_i). For each component, the object category is given by the majority semantic label, and the position (x-y center) and planar extents of the object is given by the occupied floor cells. The component geometry also provides slice-span and body/front cues , which are pooled with denoiser features by three attribute heads to predict continuous yaw, height, and vertical offset for the object record. We instantiate the final scene by retrieving appropriate 3D assets from 3D-FUTURE (Fu et al., 2021b). For each object record, we query the 3D database with the object category and select the closest asset based on the 3D extents. The retrieved asset is scaled along each axis to match the estimated extents, rotated to the predicted yaw, and translated to the recovered object position.

4 Experiments

We evaluate architecture-conditioned generation against recent diffusion methods. Additional unconditioned and per-room results appear in App. B, and representation checks in App. C. Ablations and a user study assess design choices and perceived quality.

Table 1: Quantitative comparison on 3D-FRONT. KID is reported as KID$\times 10^3$ and CKL as CKL$\times 10^2$. Best values are bold and second-best values are italic.

	Distribution / Composition					Spatial / Physical Plausibility				
Method	FID↓	KID↓	CKL↓	Pair-Rel↓	#Obj.	OOB_S↓	OOB_O↓	COL↓	NAV↑	OC↑
DIFFUSCENE (Tang et al., 2024)	88.47	30.22	20.42	32.49	8.65	66.47	20.03	40.97	94.34	59.46
MIDIFFUSION (Hu et al., 2026)	93.51	31.06	36.97	40.65	7.35	60.68	34.66	51.62	96.19	*75.95*
SEMLAYOUTDIFF (Sun et al., 2026)	*71.06*	**8.65**	*4.78*	*20.79*	7.89	16.63	6.62	*16.13*	*96.36*	56.15
SEMLAYOUTDIFF (arch. clamped)	73.58	10.54	**4.63**	22.02	7.23	*0.50*	*0.26*	40.77	95.37	65.37
Ours	**49.84**	*10.41*	6.43	**12.13**	8.36	**0.00**	**0.00**	**12.75**	**99.30**	**82.04**

4.1 EXPERIMENTAL SETUP

Dataset. We train and evaluate on 3D-FRONT (Fu et al., 2021a) following splits established in prior work (Paschalidou et al., 2021) with 4,496 train / 274 val / 390 test rooms over three room types (bedroom, living room, and dining room).

Model details. Following SemLayoutDiff (Sun et al., 2026), we train all methods as a single unified model across room types rather than one model per room type. Each scene is rasterized on a 128×128 floor-plane grid, where one pixel/token has side length 0.10 m in the horizontal plane. The grid is repeated over $S{=}8$ fixed height slices at $(0.02, 0.45, 0.80, 1.20, 1.60, 2.00, 2.40, 2.70)$ m, which are selected once by greedy coverage analysis with a 0.10 m vertical tolerance. For MDIRT, RoomTokens use $K_{\text{sem}}{=}39$ semantic classes, $C{=}3$ instance colors plus $\varnothing_c$, and body/front/$\varnothing_p$ facing labels; valid semantic–instance-color–facing triples are flattened into a 209-id palette for visualizations and the single-channel ablation. See Sec. D.1 for model details and App. E for data processing details.

Baselines. We compare DIFFUSCENE (Tang et al., 2024) (continuous diffusion), MIDIFFUSION (Hu et al., 2026) (mixed discrete-continuous diffusion), and SEMLAYOUTDIFF (Sun et al., 2026) (semantic-map diffusion), all trained as unified models on the same 3D-FRONT split. We also retrain a SEMLAYOUTDIFF variant that retains multinomial diffusion while fixing exterior, door, and window pixels during training and sampling, restricting furniture to valid floor regions.

Metrics. Following prior work (Sun et al., 2026), we use FID and KID to assess rendered layout fidelity and CKL to compare object-category distributions against the 3D-FRONT test set. We introduce Pair-Rel to compare object-pair categories and spatial relations. For physical plausibility, we report out-of-boundary rates for scenes (OOB_S) and objects (OOB_O), object collisions (COL), navigability (NAV), and our door/window opening-clearance metric (OC). We also report the mean number of objects per scene (#Obj.). Metric definitions and computation details are in App. A.

4.2 RESULTS AND ANALYSIS

Physical validity with competitive layout quality. Table 1 shows that MDIRT achieves a better FID than SemLayoutDiff (49.84 vs. 71.06) and comparable KID/CKL (10.41/6.43 vs. 8.65/4.78), demonstrating competitive distribution learning. It also performs best across all spatial plausibility metrics: lower Pair-Rel (12.13) indicates more plausible object relations, zero OOB confirms boundary compliance, lower COL (12.75) reflects fewer intersections, higher NAV (99.30) indicates better free-space connectivity, and higher OC (82.04) indicates better clearance around openings. The average number of objects of 8.36 objects per scene for MDIRT, is comparable to that of 7.89 for SemLayoutDiff and 9.39 in the ground truth, confirms that the gains is not due to sparse layouts.

SemLayoutDiff with architecture clamping reduces OOB but raises COL from 16.13 to 40.77 because its reverse process denoises random semantic values rather than masked furniture under a fixed architecture. Clamping also cannot account for geometry added during grouping, retrieval, and rotation, so some objects remain partially outside the room. By jointly modeling arch-conditioned semantic, instance-color, and facing slices, MDIRT achieves zero OOB while retaining low collision.

MDIRT reduce visible layout failures. Figure 4 shows that the quantitative improvements are visible in the generated layouts: baseline methods can place objects across boundaries, collide

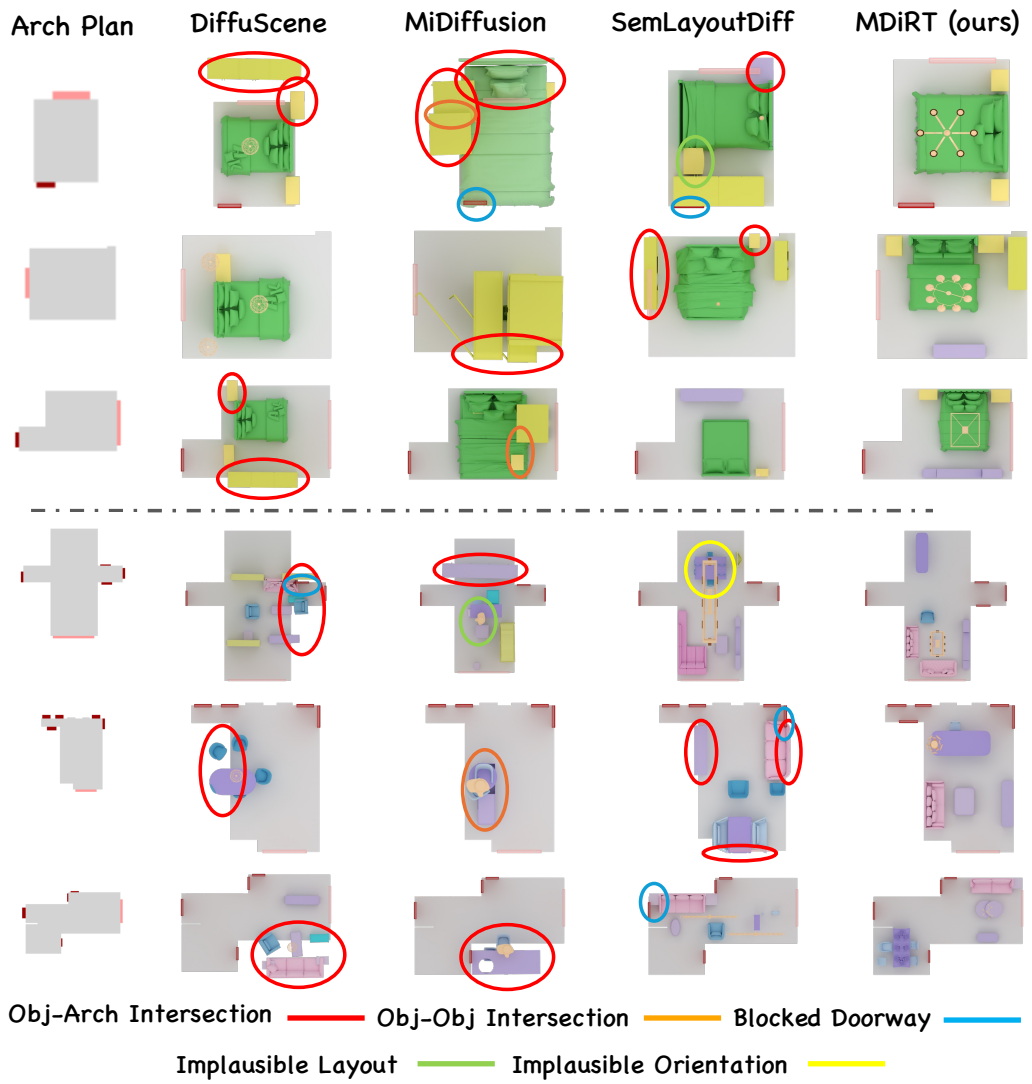

Figure 4: Arch-mask-conditioned qualitative comparison on 3D-FRONT. We use semantic orthographic views to focus the comparison on layout evaluation. Baselines show boundary violations, collisions, or blocked door/window openings, while MDiRT preserves clear openings, room geometry, and structured object cues.

Table 2: Ablations for MDiRT. Best values are in bold and second-best values are in italic.

	Distribution / Composition					Spatial / Physical Plausibility				
Configuration	FID↓	KID↓	CKL↓	Pair-Rel↓	#Obj.	OOB_S↓	OOB_O↓	COL↓	NAV↑	OC↑
Multichannel training (MDiRT)	**49.84**	**10.41**	**6.43**	**12.13**	8.36	**0.00**	**0.00**	**12.75**	**99.30**	**82.04**
w/o class balancing	74.77	34.18	*21.31*	*20.03*	6.31	**0.00**	**0.00**	*37.73*	*96.58*	*76.20*
Single-channel training	95.35	34.91	30.05	48.32	7.12	**0.00**	**0.00**	49.30	95.74	66.40
w/o semantic-priority sampling	88.21	28.74	32.40	34.54	7.84	**0.00**	**0.00**	58.81	73.07	54.90
w/o instance-color field	98.73	36.42	51.96	37.32	9.41	**0.00**	**0.00**	88.77	83.52	48.30
w/o facing field	*84.92*	*26.18*	28.62	36.63	8.96	**0.00**	**0.00**	89.07	75.69	50.70
w/o cross-slice attention	91.05	31.27	37.82	34.78	9.18	**0.00**	**0.00**	89.01	87.16	55.60

nearby furniture, or block architectural openings such as doors and windows, while MDiRT better preserves clear openings, room geometry, and object structure. See App. G for additional results.

Ablation and failure analysis. We conduct ablation experiments on our model to determine the impact of different components (see Tab. 2). Without class balancing ($\mathcal{L}_{\text{cls}}$), mean object count falls

from 8.36 to 6.31 and CKL rises from 6.43 to 21.31, indicating sparser scenes with less accurate category frequencies. Collapsing the three RoomToken fields raises FID from 49.84 to 95.35 and Pair-Rel from 12.13 to 48.32, indicating poorer layout fidelity and object-pair relations. Without semantic-priority sampling , Pair-Rel rises to 34.54 and COL to 58.81, suggesting that updating categories before attributes improves object arrangements and reduces overlap. Removing the instance-color field raises COL to 88.77, indicating less reliable separation of object instances. Removing the facing field raises COL to 89.07 and lowers NAV from 99.30 to 75.69, indicating more overlap and less navigable scenes. Without cross-slice attention, FID rises to 91.05 and COL to 89.01, consistent with less coherent object geometry across heights. OOB remains zero in every variant because architecture clamping is retained.

Figure 5 shows representative failures: implausible furniture arrangements, object–object intersections, and blocked doorways. Valid RoomTokens and in-boundary placement do not guarantee functional arrangements or sufficient clearance in the generated scene.

User study. We ask participants to rank scenes based on architectural fidelity and feasibility of object arrangements. Each participant is shown 20 sets of four generated scenes from different methods. The scenes are randomized, and users are asked to rank the scenes from best to work. Overall, participants ranked MDIRT best on average (Table 3). This preference is significant against every baseline after Holm correction (p=0.0015), and indicates that find MDIRT to have stronger perceived architectural fidelity and object arrangement than baselines.

Table 3: User study with 12 participants ranking generated scenes. MDIRT was ranked 1st almost 80% of the time.

Method	Mean rank↓	First↑	Top-2↑
MDIRT (Ours)	**1.25**	**79.6%**	**96.3%**
SEMLAYOUTDIFF	2.15	18.3%	73.3%
DIFFUSCENE	3.15	2.1%	17.9%
MIDIFFUSION	3.45	0.0%	12.5%

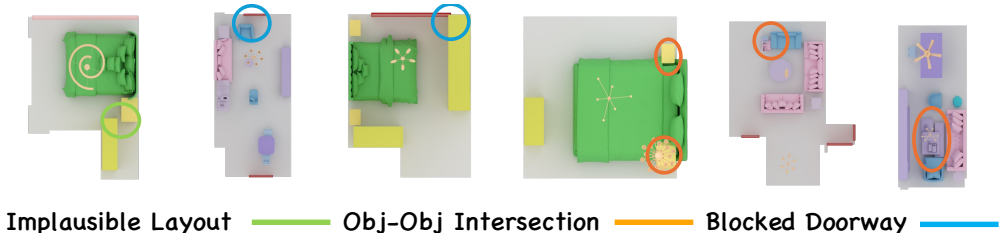

Figure 5: Failure cases: implausible layouts (green), object intersections (orange), and blocked doorways (blue).

5 CONCLUSION

We presented MDIRT, a masked discrete diffusion framework over sliced RoomTokens for indoor scene layouts. RoomTokens add height slices, instance-color cues, and facing cues to semantic occupancy maps. Clamping architectural tokens during denoising preserves the supplied floor plan, keeping room geometry explicit while supporting object grouping, orientation estimation, and scene assembly.

On 3D-FRONT, MDIRT achieves competitive distribution quality and the strongest overall spatial plausibility among evaluated methods. Controlled comparisons show gains beyond architecture clamping and sparse generation; a blinded human ranking study favors our scenes. Representation checks show little information loss; ablations support factorized RoomToken fields, semantic-priority sampling, and cross-slice attention. Structured occupancy tokens could also support outdoor scenes, objects with parts, and other semantic occupancy tasks.

Limitations. The fixed grid and height slices can miss fine details and narrow gaps. The small 3D-FRONT training set is uneven across object categories and room types. RoomTokens and validity projection cannot prevent mesh collisions from grouping, continuous pose estimation, scaling, or asset retrieval. Larger, more diverse data and stronger scene assembly could improve distribution matching and retrieval; adaptive-resolution slices and geometry-aware refinement could capture details and reduce collisions.

AI USE STATEMENT

We used AI tools for the following tasks:

- Help develop conceptual frameworks: Coding agents helped brainstorm alternatives for the scene representation and related method design. The authors selected and refined the final approach.
- Design or provide feedback on research methodology or experiments: Coding agents provided feedback on the evaluation protocol, metrics, and ablation experiments.
- Implement methods: Coding agents assisted with implementing and debugging the model, training, scene reconstruction, and evaluation code.
- Support qualitative and thematic data analysis: Coding agents helped examine experimental results and ablations, identify inconsistencies, and interpret patterns in the reported metrics.

Additional disclosures include the use of AI for brainstorming, creating and editing software code, searching for information and relevant literature, and drafting and editing manuscript text to improve clarity. AI also helped clarify the written and mathematical descriptions of our training objectives and evaluation metrics.

We have reviewed and validated all AI-assisted work and take responsibility for the final content of the work, including text, claims, code, and other artifacts produced with the aid of AI.

REFERENCES

Rio Aguina-Kang, Maxim Gumin, Do Heon Han, Stewart Morris, Seung Jean Yoo, Aditya Ganeshan, R Kenny Jones, Qiuhong Anna Wei, Kailiang Fu, and Daniel Ritchie. Open-universe indoor scene generation using LLM program synthesis and uncurated object databases. *arXiv preprint arXiv:2403.09675*, 2024.

Kenneth Appel and Wolfgang Haken. Every planar map is four colorable. Part I: Discharging. *Illinois Journal of Mathematics*, 21(3):429–490, 1977.

Tongyuan Bai, Wangyuanfan Bai, Dong Chen, Tieru Wu, Manyi Li, and Rui Ma. FreeScene: Mixed graph diffusion for 3D scene synthesis from free prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5893–5903, 2025.

Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, and Xi Wang. I-design: Personalized LLM interior designer. *arXiv preprint arXiv:2404.02838*, 2024. doi: 10.48550/arXiv.2404.02838.

Shang Chai, Liansheng Zhuang, and Fengying Yan. LayoutDM: Transformer-based diffusion model for layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18349–18358, 2023.

Jeongjun Choi, Yeonsoo Park, and H. Jin Kim. SceneNAT: Masked generative modeling for language-guided indoor scene synthesis. *arXiv preprint arXiv:2601.07218*, 2026.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 2023.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3D-Front: 3D furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021a.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3D-FUTURE: 3D furniture shape with texture. *International Journal of Computer Vision*, 129(12):3313–3337, 2021b.

Jialin Gao, Donghao Zhou, Mingjian Liang, Lihao Liu, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. DisCo-Layout: Disentangling and coordinating semantic and physical refinement in a multi-agent framework for 3D indoor layout synthesis. *arXiv preprint arXiv:2510.02178*, 2025.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

Siyi Hu, Diego Martin Arroyo, Stephanie Debats, Fabian Manhardt, Luca Carlone, and Federico Tombari. Mixed diffusion for 3D indoor scene synthesis. In *Winter Conference on Applications of Computer Vision*, 2026.

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 scene as Blender code. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

Mude Hui, Zhizheng Zhang, Xiaoyi Zhang, Wenxuan Xie, Yuwang Wang, and Yan Lu. Unifying layout generation with a decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1942–1951, 2023.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. LayoutDM: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10167–10176, 2023.

Victor Kulikov, Victor Yurchenko, and Victor Lempitsky. Instance segmentation by deep coloring. *arXiv preprint arXiv:1807.10007*, 2018.

Jumin Lee, Woobin Im, Sebin Lee, and Sung-Eui Yoon. Diffusion probabilistic models for scene-scale 3D categorical data. *arXiv preprint arXiv:2301.00527*, 2023.

Elad Levi, Eli Brosh, Mykola Mykhailych, and Meir Perez. DLT: Conditioned layout generation with joint discrete-continuous diffusion layout transformer. In *International Conference on Computer Vision*, 2023.

Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. GRAINS: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-driven 3D indoor scene synthesis with semantic graph prior. In *International Conference on Learning Representations*, 2024.

Guying Lin, Kemeng Huang, Michael Liu, Ruihan Gao, Hanke Chen, Lyuhao Chen, Beijia Lu, Taku Komura, Yuan Liu, Jun-Yan Zhu, and Minchen Li. PAT3D: Physics-augmented text-to-3D scene generation. In *International Conference on Learning Representations*, 2026.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2024.

Andrew Luo, Zhoutong Zhang, Jiajun Wu, and Joshua B Tenenbaum. End-to-end optimization of scene layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3754–3763, 2020.

Léopold Maillard, Nicolas Sereyjol-Garros, Tom Durand, and Maks Ovsjanikov. DeBaRA: Denoising-based 3D room arrangement generation. *Advances in Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2409.18336.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoglu, and Theo Gevers. SceneTeller: Language-to-3D scene generation. In *European Conference on Computer Vision*, 2024.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems*, 34:12013–12026, 2021.

Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. SceneSmith: Agentic generation of simulation-ready indoor scenes. *arXiv preprint arXiv:2602.09153*, 2026.

Daniel Ritchie, Kai Wang, and Yu-an Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6182–6190, 2019.

Neil Robertson, Daniel Sanders, Paul Seymour, and Robin Thomas. The four-colour theorem. *Journal of Combinatorial Theory, Series B*, 70(1):2–44, 1997.

Antonio Ruiz, Tao Wu, Andrew Melnik, Qing Cheng, Xuqin Wang, Lu Liu, Yongliang Wang, Yanfeng Zhang, and Helge Ritter. GeoSceneGraph: Geometric scene graph diffusion model for text-guided 3D indoor scene synthesis. *arXiv preprint arXiv:2511.14884*, 2025.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, 2024.

Chong Su, Yingbin Fu, Zheyuan Hu, Jing Yang, Param Hanji, Shaojun Wang, Xuan Zhao, Cengiz Öztireli, and Fangcheng Zhong. CHOrD: Generation of collision-free, house-scale, and organized digital twins for 3D indoor scenes with controllable floor plans and optimal layouts. *arXiv preprint arXiv:2503.11958*, 2025.

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 Conference on Computer Vision and Pattern Recognition*, 2025a.

Jia-Mu Sun, Jie Yang, Kaichun Mo, Yu-Kun Lai, Leonidas Guibas, and Lin Gao. Haisor: Human-aware indoor scene optimization via deep reinforcement learning. *ACM Transactions on Graphics*, 43(2):1–17, 2024a.

Kaifan Sun, Bingchen Yang, Peter Wonka, Jun Xiao, and Haiyong Jiang. RelTriple: Learning plausible indoor layouts by integrating relationship triples into the diffusion process. *arXiv preprint arXiv:2503.20289*, 2025b. doi: 10.48550/arXiv.2503.20289.

Qi Sun, Hang Zhou, Wengang Zhou, Li Li, and Houqiang Li. Forest2Seq: Revitalizing order prior for sequential indoor scene synthesis. In *European Conference on Computer Vision*, 2024b. doi: 10.48550/arXiv.2407.05388.

Xiaohao Sun, Divyam Goel, and Angel X. Chang. SemLayoutDiff: Semantic layout generation with diffusion model for indoor scene synthesis. In *International Conference on 3D Vision*, 2026.

Hou In Ivan Tam, Hou In Derek Pun, Austin T Wang, Angel X Chang, and Manolis Savva. SceneEval: Evaluating semantic coherence in text-conditioned 3D indoor scene synthesis. In *Winter Conference on Applications of Computer Vision*, 2026a.

Hou In Ivan Tam, Hou In Derek Pun, Austin T Wang, Xiaohao Sun, Qirui Wu, Han-Hung Lee, Angel X Chang, and Manolis Savva. Survey on compositional 3D indoor scene generation. *Computer Graphics Forum*, page e70591, 2026b.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kai Wang, Manolis Savva, Angel X Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. PlanIT: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019.

Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *International Conference on 3D Vision*, pages 106–115. IEEE, 2021.

Zehao Wen, Zichen Liu, Srinath Sridhar, and Rao Fu. AnyHome: Open-vocabulary generation of structured and textured 3D homes. In *European Conference on Computer Vision*, 2024.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, et al. SAGE: Scalable agentic 3D scene generation for embodied AI. *arXiv preprint arXiv:2602.10116*, 2026.

Ming-Jia Yang, Yu-Xiao Guo, Bin Zhou, and Xin Tong. Indoor scene generation from a collection of semantic-segmented depth images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15203–15212, 2021.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically interactable 3D scene synthesis for embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16262–16272, 2024a.

Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. SceneWeaver: All-in-one 3D scene synthesis with an extensible and self-reflective agent. In *Advances in Neural Information Processing Systems*, 2025.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3D embodied AI environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 30, pages 20–25. IEEE/CVF, 2024b.

Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12965–12976, 2023.

Guangyao Zhai, Evin Pinar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. CommonScenes: Generating commonsense 3D indoor scenes with scene graphs. *Advances in Neural Information Processing Systems*, 2023.

Guangyao Zhai, Evin Pınar Örnek, Dave Zhenyu Chen, Ruotong Liao, Yan Di, Nassir Navab, Federico Tombari, and Benjamin Busam. Echoscene: Indoor scene generation via information echo over scene graph diffusion. In *European Conference on Computer Vision*, pages 167–184. Springer, 2024.

Ye Zhang, Yu Zhou, Yifeng Wang, Jun Xiao, Ziyue Wang, Yongbing Zhang, and Jianxu Chen. The four color theorem for cell instance segmentation. In *Forty-second International Conference on Machine Learning*, 2025.

Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics (TOG)*, 39(2):1–21, 2020.