# M3DLAYOUT: A MULTI-SOURCE DATASET OF 3D INDOOR LAYOUTS AND STRUCTURED DESCRIPTIONS FOR 3D GENERATION

**Anonymous authors**
Paper under double-blind review

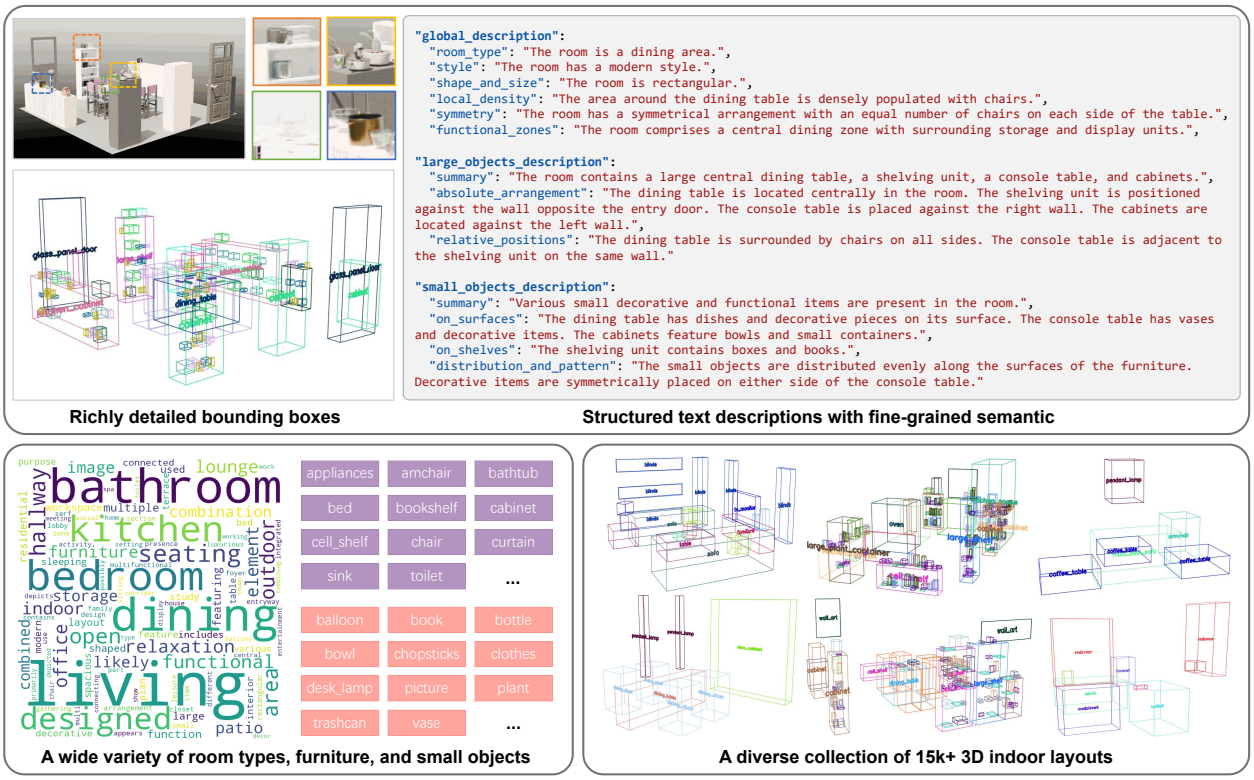


Figure 1: **The M3DLayout dataset — A multi-source benchmark for text-to-3D indoor scene generation.** Top: An example from our dataset showing a detailed 3D indoor layout with richly annotated bounding boxes and its corresponding structured textual description. Bottom-left: Word cloud visualization demonstrating the diversity of room types, furniture, and objects in the dataset. Bottom-right: Overview of the large-scale collection containing 15,080 diverse 3D layout scenes with various styles.

## ABSTRACT

In text-driven 3D scene generation, object layout serves as a crucial intermediate representation that bridges high-level language instructions with detailed geometric output. It not only provides a structural blueprint for ensuring physical plausibility but also supports semantic controllability and interactive editing. However, the learning capabilities of current 3D indoor layout generation models are constrained by the limited scale, diversity, and annotation quality of existing datasets. To address this, we introduce **M3DLayout**, a large-scale, multi-source dataset for 3D indoor layout generation. M3DLayout comprises **15,080 layouts** and over **258k object instances**, integrating three distinct sources: real-world scans, professional CAD designs, and procedurally generated scenes. Each layout is paired with detailed structured text describing global scene summaries, relational placements of large furniture, and fine-grained arrangements of smaller items. This diverse and richly annotated resource enables models to learn complex spatial and semantic patterns across a wide variety of indoor environments. To assess the potential of M3DLayout, we establish a benchmark using a text-conditioned diffusion model. Experimental results demonstrate that our dataset provides a solid foundation for training layout generation models. Its multi-source composition enhances diversity, notably through the Inf3DLayout subset which provides rich small-object information, enabling the generation of more complex and detailed scenes. We hope that M3DLayout can serve as a valuable resource for advancing research in text-driven 3D scene synthesis.

# 1 INTRODUCTION

Recent advances in 3D generative modeling have enabled remarkable progress in synthesizing 3D objects and scenes from various modalities, such as text or images. These developments have great potential for downstream applications in areas such as content creation, robotics, and virtual reality Höllein et al. (2023); Yang et al. (2024a;b); Schult et al. (2024). In particular, text-to-3D generation has attracted increasing attention due to its intuitive and flexible interface for controlling complex scene content. For example, LucidDreamer Chung et al. (2023) and Text2Immersion Ouyang et al. (2023) adopt point-based or Gaussian-splatting representations to generate detailed scene geometry directly from text or other modalities. Other approaches, such as ATISS Paschalidou et al. (2021), SceneFormer Wang et al. (2021), EchoScene Zhai et al. (2025), and MIDI Huang et al. (2024), perform joint layout-object generation by autoregressively predicting room structures and furnishing objects in a unified framework. LayoutGPT Feng et al. (2023), HoloDeck Yang et al. (2024b), and InstructScene Lin & Mu (2024) employ large language models to plan scene layouts from free-form descriptions, showcasing a promising direction in LLM-driven compositional layout generation.

While these methods demonstrate strong capabilities, they also reveal key limitations. Some of them generate scenes as inseparable volumetric representations, which limits modularity and downstream controllability. Layout-aware models such as CommonScenes Zhai et al. (2023) and DiffuScene Tang et al. (2024) tend to produce relatively simple scenes with few object types and limited diversity. Meanwhile, LLM-based planners show promise in parsing natural language but often struggle with spatial consistency and accurate physical arrangements.

As an essential and intermediate representation of 3D scene, 3D layout data is crucial for high-quality generation. Its importance can be summarized in three key points: $(i)$ 3D layout data define the position, orientation, and scale of objects within a scene, forming the foundation of its structure and spatial coherence. This ensures objects are placed logically and functionally, greatly enhancing the generated scene's realism and credibility while preventing chaotic or illogical visual outcomes. $(ii)$ As powerful conditional information, 3D layout data guides and constrains the 3D scene generation process. It significantly reduces the model's degrees of freedom and ambiguity, allowing for more efficient and accurate detail filling. This also enables users to precisely control the scene's layout, meeting personalized and diverse generation demands. $(iii)$ Real-world 3D scenes typically serve specific functions, and 3D layout data directly reflect this functionality. Through logical arrangements, generated scenes can better fulfill practical application needs (e.g., interior design, game levels) and provide interactive spaces that align with human cognitive habits, thereby significantly improving the end-user experience.

We argue that a major bottleneck limiting further progress in controllable and high-quality scene generation is the lack of large-scale, richly annotated 3D layout datasets that provide structured, semantic-level supervision. Existing datasets either focus on scene geometry from real-world scans (e.g., ScanNet Dai et al. (2017), Matterport3D Chang et al. (2017)) or offer object-level annotations based on professional designs (e.g., 3D-Front Fu et al. (2021), Structured3D Zheng et al. (2020)). However, none of them provide comprehensive layout annotations that include both large-scale furniture and small functional or decorative objects, paired with structured descriptions capturing global scene organization and fine-grained spatial relations.

To address this gap, we introduce M3DLayout, a multi-source dataset for 3D indoor layout generation from structured language. M3DLayout contains 15,080 layouts and over 258k object instances, collected from three complementary sources: real-world scans, professional CAD designs, and procedurally generated scenes. Each layout is paired with structured text descriptions that include global scene summaries, relational placements of large furniture, and fine-grained arrangements of smaller items. These annotations are constructed using a combination of rule-based extraction, GPT-assisted generation, and human verification.

To demonstrate the utility of M3DLayout, we train a text-conditioned diffusion model as a baseline for layout generation. Our experiments show that progressively incorporating the three data sources during training leads to consistent improvements in generation quality, supporting the complementary value of each data type. Furthermore, our model outperforms recent layout generation methods in terms of semantic alignment, spatial plausibility, and controllability.

We summarize our main contributions as follows:

1. We propose M3DLayout, a large-scale, richly annotated dataset of 3D indoor layouts with structured text descriptions, collected from multiple complementary sources.
2. We dedicate efforts to creating the Inf3DLayout subset, which fills a gap in high-quality data for common scene types and substantially increases the level of detail and diversity within the dataset.
3. We establish a benchmark for text-to-layout generation using a diffusion-based baseline. Results validate the utility of M3DLayout, showing that the dataset enhances the diversity and detail of generated scenes.

# 2 Related Work

## 2.1 Datasets for Indoor Scenes

Large-scale datasets play a crucial role in learning-based 3D indoor layout generation and scene synthesis. Early efforts to capture real-world environments using 3D scans led to the creation of datasets such as ScanNet Dai et al. (2017), Matterport3D Chang et al. (2017), and SceneNN Hua et al. (2016). These datasets provide high-fidelity mesh reconstructions, reflect real-world object distributions and spatial constraints. However, they often suffer from noisy geometry, incomplete object coverage, and a lack of fine-grained annotations suitable for generative tasks, as well as limited layout variability due to constrained capture environments.

To address these limitations, synthetic datasets such as SUNCG Song et al. (2017), 3D-FRONT Fu et al. (2021), and Structured3D Zheng et al. (2020) were introduced, offering structured 3D layouts with complete object metadata and annotations from professional CAD designs. However, these professional designs typically lack object variety and fine-grained detail.

Recent hybrid datasets attempt to bridge this gap. FurniScene Zhang et al. (2024) enriches layout realism with more diverse furniture arrangements. OpenRooms Li et al. (2021) provides photorealistic rendering with physical material properties and lighting. Despite these advances, a key limitation persists across nearly all prior datasets: the lack of scene-level textual annotations, which limits their use for conditional or multimodal generation tasks.

In this context, we introduce M3DLayout, a multi-source dataset that combines real-world scans, professional designs, and procedurally generated layouts. We further enrich these sources with structured textual annotations, creating a foundational resource aimed at propelling research in controllable, text-driven 3D scene synthesis.

## 2.2 Indoor Scene Synthesis

The evolution of indoor scene synthesis methods underscores the critical need for richer, more descriptive layout data. Procedural approaches generate indoor scenes using predefined rules, templates, or simulation engines. Systems such as ProcTHOR Deitke et al. (2022) and Infinigen Raistrick et al. (2023; 2024) rely on large-scale procedural scene grammars and asset libraries to create diverse, physically realistic environments. These methods offer high controllability and scalability, especially for generating synthetic training data for embodied agents. However, they are ultimately constrained by their handcrafted rules.

In parallel, a large body of work uses learning-based generative models trained on indoor scene datasets to learn object layouts and spatial relationships in a data-driven fashion. Early methods relied on auto-encoding architectures (e.g., SG-VAE Purkait et al. (2020), SceneHGN Gao et al. (2023), CommonScenes Zhai et al. (2023)) and autoregressive models (e.g., ATISS Paschalidou et al. (2021), SceneFormer Wang et al. (2021)). More recently, diffusion-based models such as DiffuScene Tang et al. (2024) and EchoScene Zhai et al. (2025) have shown superior performance in capturing complex spatial dependencies. SemlayoutDiff Sun et al. (2025)optimized the Diffuscene Tang et al. (2024) by using semantic-mediated 2D distribution maps to finally generate 3D layouts, and is capable of generating different room types with the same model. However, precise control of small object generation cannot be achieved through functional zoning alone. These methods improve the plausibility and diversity of generated layouts but often face challenges in optimization, attribute disentanglement, and generalization to out-of-distribution scenes.

Inspired by the success of large language models (LLMs), several works explore text-guided 3D scene generation. Architect Wang et al. (2024) and HoloDeck Yang et al. (2024b) leverage world knowledge to interpret user instructions and generate spatial constraints. Similarly, FlairGPT Littlefair et al. (2025) controls layout generation through more detailed object descriptions using LLMs, but the overly complex conversational process greatly reduces the efficiency of generating large amounts of data. While these methods support more flexible and intuitive interactions, they often struggle with precise 3D spatial reasoning, often producing physically implausible layouts. This core limitation stems from their lack of grounding in large-scale, physically plausible 3D scene data. M3DLayout is designed to address this critical gap, providing the missing link and a robust foundation to train and benchmark the next generation of these powerful synthesis models.

# 3 The M3DLayout Dataset

To advance controllable and generalizable text-to-3D scene generation, we introduce M3DLayout, a large-scale, multi-source dataset of 3D indoor layouts. This dataset integrates 15,080 layouts from three complementary types of sources:

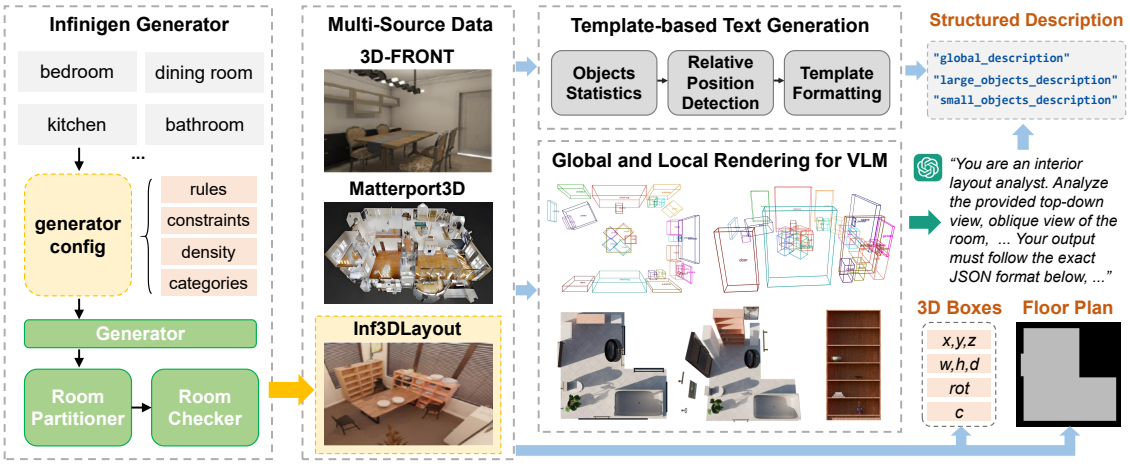


Figure 2: **Pipeline for Constructing the M3DLayout Dataset.** Our framework integrates multi-source data, including the professional designs dataset 3D-FRONT, real-world scans from Matterport3D, and procedurally generated scenes from Infinigen. The construction process involves: meticulously generating, partitioning, and filtering layouts to create the Inf3DLayout subset; performing template-based rules to produce formatted text; and employing global and local rendering for vision-language models (VLM) to produce structured descriptions. This pipeline results in a large-scale, richly-annotated text-3D layout paired dataset.

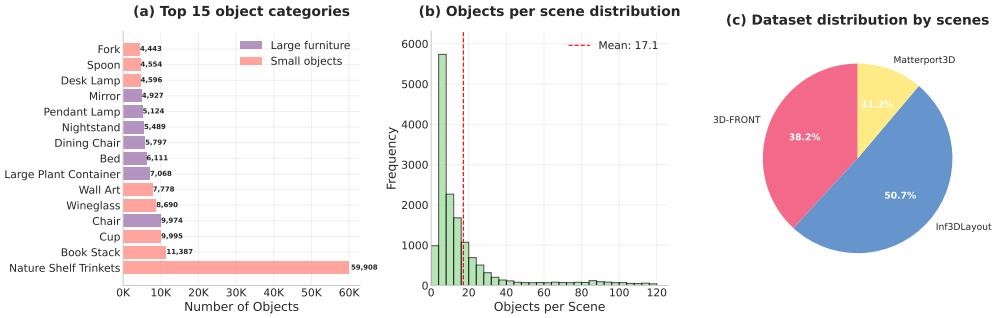


Figure 3: **Dataset statistics of M3DLayout.** (a) Top 15 most frequent object categories. (b) Distribution of the number of objects per scene. (c) Proportion of scenes contributed by each source dataset.

real-world scans, professional interior designs, and procedurally generated scenes. Each layout is annotated with detailed structured textual descriptions to support fine-grained, text-conditioned layout generation.

### 3.1 Data Sources and Curation

M3DLayout is built upon three distinct types of data sources, each contributing unique characteristics to the dataset:

**Real-world Scans.** We incorporate layouts from the Matterport3D dataset Chang et al. (2017), which are derived from real environment scans. These layouts reflect realistic, and often cluttered, spatial arrangements found in actual indoor settings. To ensure data quality, we performed a cleanup of the object category list by merging and removing low-frequency categories. Scenes containing fewer than two object instances were filtered out. This curated subset provides a wide variety of scene types and offers important cues for learning robust and realistic layout patterns.

**Professional Interior Designs.** We integrate high-quality layouts from 3D-FRONT Fu et al. (2021), which contains professionally designed indoor scenes. These layouts are characterized by well-organized spatial semantics and adhere to tidy, minimalist design principles. They typically feature sparser object arrangements, providing strong supervision for generating structurally coherent and aesthetically plausible scenes with clean layouts. Following established practices in prior work Paschalidou et al. (2021); Tang et al. (2024), we applied filters to remove layouts with uncommon object configurations or unnatural room proportions, ensuring a focus on typical, well-structured interior designs.

**Procedurally Generated Scenes.** To systematically enhance object diversity, particularly for small and decorative items, we generate the Inf3DLayout subset using the procedural generator Infinigen Raistrick et al. (2024). A key part

Table 1: Quantitative analysis of three data sources (3D-FRONT, Matterport3D, Inf3DLayout) in M3DLayout.

| Source | Scenes | Total Objects | Avg Objs/Scene | Large Furniture | Small Objects | Small % |
|---|---|---|---|---|---|---|
| 3D-FRONT | 5,754 | 39,494 | 6.9 | 39,407 | 87 | 0.2% |
| Matterport3D | 1,684 | 21,212 | 12.6 | 12,859 | 8,353 | 39.4% |
| Inf3DLayout | 7,642 | 197,707 | 25.9 | 57,125 | 140,582 | 71.1% |
| | **15,080** | **258,413** | **17.1** | **103,391** | **149,022** | **57.7%** |

Table 2: **Comparisons between existing 3D indoor scene datasets**, where "N/A" denotes "not available", "L" and "S" denote Large and small objects in the scene, respectively.

| Dataset | Scenes | Objects | Layout Collection | Layout Complexity | Variation in Object Sizes | Structured Descriptions |
|---|---|---|---|---|---|---|
| SUN3D Xiao et al. (2013) | 254 | N/A | Real scan | Low | N/A | ✗ |
| SceneNNHua et al. (2016) | 100 | N/A | Real scan | Low | N/A | ✗ |
| Matterport3DChang et al. (2017) | 1,684 | N/A | Real scan | Medium | L-S | ✗ |
| ScanNet Dai et al. (2017) | 1,506 | N/A | Real scan | Low | L-S | ✗ |
| Scan2CAD Avetisyan et al. (2019) | 1,506 | N/A | Real scan | Low | N/A | ✗ |
| OpenRooms Li et al. (2021) | 1,068 | 97,607 | Real scan | Low | N/A | ✗ |
| SceneNet Handa et al. (2016) | 57 | 3,699 | Professional | Low | N/A | ✗ |
| Structured3D Zheng et al. (2020) | N/A | N/A | Professional | Low | N/A | ✗ |
| 3D-FRONT Fu et al. (2021) | 5,754 | N/A | Professional | Low | L | ✗ |
| M3DLayout | 15,080 | 258,413 | Mixture | High | L-S | ✓ |

of our curation involved carefully configuring the generator to produce plausible layouts for five common room types: bedrooms, living rooms, dining rooms, kitchens, and bathrooms. The generated houses were then programmatically partitioned into individual rooms. Finally, we applied a filtering step to remove rooms with abnormal layouts or spatial inconsistencies. This curated subset significantly increases the variety and granularity of object arrangements, covering numerous long-tail scenarios that are underrepresented in scan-based or design-based data.

The combination of these sources ensures that M3DLayout encompasses a wide spectrum of indoor environments, balancing realism, design integrity, and compositional diversity.

### 3.2 Structured Description Annotation

To support fine-grained text-conditioned layout generation, we annotate each 3D layout in the M3DLayout dataset with a comprehensive structured description. The annotation schema is designed to capture spatial and semantic information at multiple levels of detail, comprising three key components.

**Global Scene Description.** This part captures the overall properties and organization of the scene. Each layout is labeled with the room type (e.g., dining area), stylistic attributes (e.g., modern style), and geometric features such as room shape and object density. It also includes high-level functional zoning (e.g., central dining zone with surrounding storage units) and global spatial patterns like symmetry (e.g., equal number of chairs on both sides of the table).

**Large Furniture Description.** We describe the presence and arrangement of major furniture pieces such as dining tables, shelves, consoles, and cabinets. The annotation includes both absolute positioning (e.g., "The shelving unit is placed opposite the entry door") and relative spatial relations (e.g., "The console table is adjacent to the shelving unit"). A summary of large furniture composition is also provided for each room.

**Small Object Description.** This component focuses on decorative and functional small items like dishes, bowls, vases, books, and boxes. These are annotated based on their placement (e.g., on furniture surfaces or shelves) and distribution patterns (e.g., evenly distributed or symmetrically placed). Such fine-grained annotations support more detailed spatial reasoning and realistic scene generation.

The structured descriptions are generated through a multi-stage pipeline, as illustrated in Figure 2. As outlined in Section 3.1, we begin by generating the Inf3DLayout subset using a carefully configured procedural generation pipeline based on Infinigen. For layouts sourced from Matterport3D and the generated Inf3DLayout subset, we render top-down and side views, along with close-up images highlighting the placement of small objects. These multi-view renders are subsequently processed by the GPT-4o model to produce the structured textual descriptions. For scenes from 3D-FRONT, we adopt a rule-based approach: after extracting object-level bounding boxes and semantic labels, we detect relative spatial relationships between objects and format this structured information into coherent descrip-

tions using predefined templates. This methodology is suitable for 3D-FRONT due to the typically simpler and more regular spatial arrangements in its professionally designed scenes, allowing for accurate and comprehensive coverage via template-based generation. Finally, all automatically generated descriptions undergo a sampling-based manual review to ensure annotation quality.

### 3.3 Dataset Statistics and Analysis

M3DLayout encompasses a diverse collection of indoor environments, covering 26 scene categories. Among them, five core room types receive the most focus: bedroom, living room, dining room, kitchen, and bathroom. These constitute the most common residential spaces. In addition to these, the dataset includes functional areas such as office, entryway, closet, toilet, and balcony, as well as specialized spaces including gym, library, and home theater.

**Data Composition and Source Characteristics.** As detailed in Table 1, M3DLayout integrates 15,080 layouts with 258,413 object instances, averaging 17.1 objects per scene. The three data sources exhibit complementary characteristics: 3D-FRONT provides professionally designed layouts with clean structural regularity but limited small objects (0.2%); Matterport3D offers realistic scanned environments with moderate object density (12.6 objects/scene) and a balanced object distribution (39.4% small objects); while Inf3DLayout significantly enriches the dataset with high scene complexity (25.9 objects/scene) and abundant small objects (71.1%).

**Object Distribution and Scene Complexity.** Figure 3(a) shows the top 15 most frequent object categories, where small decorative items (e.g., Nature Shelf Trinkets, Book Stack) dominate the distribution, reflecting the dataset's fine-grained annotation richness. The objects-per-scene distribution in Figure 3(b) reveals that M3DLayout covers a wide spectrum of scene complexities, from minimalistic arrangements to densely populated environments. This variation enhances the generalization capability of trained models for real-world scenarios.

**Comparative Advantages.** As shown in Table 2, M3DLayout surpasses existing datasets in scale (15,080 scenes, 258k+ objects), layout complexity, and object size variation. Unlike datasets limited to either large furniture (L) or simple layouts, M3DLayout provides comprehensive coverage of both large and small objects (L-S) with structured textual descriptions, addressing a critical gap in current data resources for detailed scene generation.

The broad coverage of room types, the detailed structured descriptions, and the variation in scene complexity make this dataset a valuable resource for downstream tasks such as layout prediction, text-conditioned scene synthesis, and embodied AI simulation.

## 4 Benchmark

To evaluate the capabilities and limitations of the M3DLayout dataset, we establish a comprehensive benchmark using a diffusion model to assess the dataset's potential in supporting text-conditioned 3D indoor layout generation.

### 4.1 Problem Formulation

We formulate 3D indoor layout generation as a conditional denoising diffusion process. Given a scene-level natural language description $c^{\text{text}}$, our goal is to generate a structured 3D layout $x$ consisting of $N$ objects. Each object $o_i$ in the layout is parameterized as a 3D oriented bounding box $o_i = (c_i, x_i, y_i, z_i, w_i, h_i, d_i, \theta_i)$, where $c_i$ denotes the semantic class label, $(x_i, y_i, z_i)$ is the object center in 3D space, $(w_i, h_i, d_i)$ represent the width, height, and length, and $\theta_i$ is the yaw angle. The full layout is then denoted as $x_0 = \{o_i\}_{i=1}^{N}$.

Following the Denoising Diffusion Probabilistic Model (DDPM) framework, we define a forward diffusion process that gradually adds Gaussian noise to the layout representation $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$, where $\beta_t$ is a predefined noise schedule. The reverse process is modeled by a neural network $\epsilon_\theta$ that predicts the noise $\epsilon$ from the noisy layout $x_t$ and the conditioning information $c^{\text{text}}$:

$$p_\theta(x_{t-1} \mid x_t, c^{\text{text}}) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c^{\text{text}}), \Sigma_\theta(x_t, t)). \tag{1}$$

The model is trained to minimize the noise prediction objective $\mathcal{L}_{\text{DM}} = \mathbb{E}_{x_0, c^{\text{text}}, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, c^{\text{text}})\|_2^2\right]$.

### 4.2 Model Architecture

Our model architecture follows a denoising diffusion framework similar to DiffuScene, employing a U-Net style backbone with 1D convolutional layers and attention mechanisms, with conditional inputs injected via cross-attention.

Table 3: **Quantitative comparison of layout generation methods and ablation studys of our model trained on different datasets.** Lower FID/KID ($\times 0.001$) and higher Clip Score indicate better synthesis quality. FID and KID are computed with respect to the real layouts from 3D-FRONT, Matterport, and Inf3DLayout. We train InstructScene following the public implementation.

| Method | FID ↓ | | | KID ↓ | | | CLIP-Score ↑ |
|---|---|---|---|---|---|---|---|
| | 3D-FRONT | Matterport | Inf3DLayout | 3D-FRONT | Matterport | Inf3DLayout | |
| DiffuSceneTang et al. (2024) | **29.47** | 98.03 | 102.12 | **10.32** | 47.92 | 75.49 | 0.1982 |
| InstructSceneLin & Mu (2024) | 68.58 | 100.54 | 159.27 | 54.70 | 49.23 | 156.62 | 0.1944 |
| Ours (M3DLayout) | 57.64 | **87.89** | **70.85** | 36.80 | **34.62** | **50.94** | **0.2001** |
| Ours (3D-FRONT) | 27.33 | 83.88 | 110.98 | 10.59 | 21.80 | 83.45 | 0.2083 |
| Ours (Matterport) | 81.31 | 69.61 | 114.58 | 46.82 | 18.41 | 94.45 | 0.1916 |
| Ours (Inf3DLayout) | 93.51 | 115.07 | 54.36 | 55.67 | 55.53 | 34.95 | 0.1969 |

For the training objective, we employ the standard noise prediction loss combined with additional regularization terms. The primary diffusion loss minimizes the L2 distance between predicted and actual noise components. We also incorporate an IoU loss term to penalize object intersections and encourage physically plausible arrangements.

### 4.3 Experimental Settings

We conduct experiments to evaluate the effectiveness of our conditional diffusion model on the proposed dataset. The goal is to assess both the plausibility and controllability of the generated 3D layouts under different scene conditions. The detailed dataset train/val splits and implementation details are provided in the Appendix A.1. We describe our evaluation settings as follows:

**Baselines.** We compare our method with two state-of-the-art scene generation approaches: DiffuScene Tang et al. (2024) and InstructScene Lin & Mu (2024), both of which are text-driven methods based on diffusion models. More details are provided in the Appendix A.2.

**Metrics.** Following prior works Tang et al. (2024); Lin & Mu (2024), we adopt Fréchet Inception Distance (FID) Heusel et al. (2017) and Kernel Inception Distance (KID) Bińkowski et al. (2018), and the CLIP-Score Hessel et al. (2021) to measure the controllability and fidelity, respectively. We employ a self-designed object retrieval method to realize instance filling from layout to scene and rendering. More details about rendering and retrieval are provided in the Appendix A.3 and Appendix A.6.

### 4.4 Quantitative and Qualitative Comparison

**Quantitative Comparison.** The quantitative comparison results are shown in the upper part of Table 3, where DiffuScene and InstructScene are trained on 3D-FRONT and Ours on M3DLayout. In Table 3, the horizontal axis represents ground-truth renderings (real images) from three different datasets. As described in Appendix A.1, the same set of 1,500 prompts is used as conditioning input for all comparative methods to generate layouts, which are then rendered as synthetic images and used together with real images to compute FID and KID. Typically, real images are taken from the training set; however, since our method and the baselines are trained on different datasets with varying data distributions, direct comparison would be unfair. To address this, we select different datasets as the source of real images, allowing for both a fair comparison and an evaluation of fidelity and controllability.

As demonstrated, for FID and KID, our method drastically outperforms the state-of-the-art by 10%–32% on the reference Matterport and Inf3DLayout dataset, demonstrating superior generalization compared with DiffuScene and InstructScene. On 3D-FRONT, however, it falls behind DiffuScene on these metrics. This is primarily because the number of objects per scene in 3D-FRONT typically ranges from 5 to 12, whereas our method generates scenes with more than 12 objects in most cases, leading to a mismatch in the distribution of scene complexity. As a result, the visual statistics of our generated layouts deviate more from the ground truth (3D-FRONT), which negatively impacts FID and KID. Conversely, the richer object counts and variety in the scenes generated by our model provide another advantage of our method relative to the multi-source datasets M3DLayout, which is also confirmed by the visualizations in Figure 5. Moreover, our method surpasses the baselines in CLIP-Score, demonstrating enhanced controllability and a stronger alignment of the generated scenes with the given prompts.

**Qualitative Comparison.** The qualitative comparison results for different methods in the bedroom, dining room, and living room are shown in Figure 5. Across all settings, our method trained on the M3DLayout dataset demonstrates improved semantic controllability and visual fidelity compared to Diffuscene and Instructscene. Specifically, DiffuScene can generate scenes that appear visually neat at first glance but struggles to produce small objects and exhibits lim-

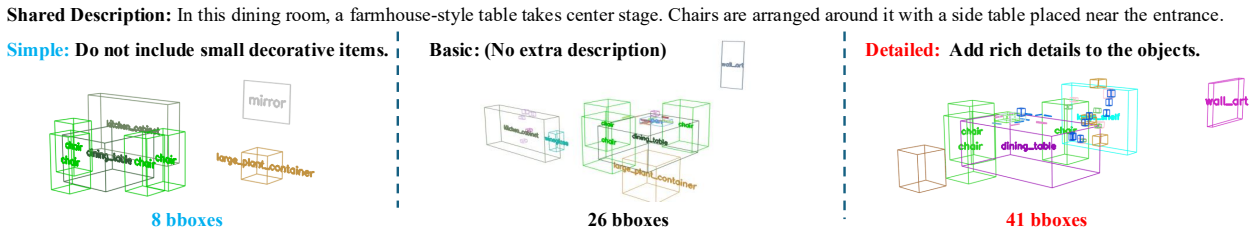


Figure 4: **Density Controllability in Layout Generation with Different Input Prompts.** The first row presents input prompts for our layout generation model, showcasing variations in objects density from low to high, with minor changes in the last sentence. The second row illustrates the corresponding output results generated by our model, which adapt based on the prompt density.

ited prompt controllability, as illustrated by the living room case. InstructScene, on the other hand, fails to accurately model spatial relationships among instances, leading to disorderly object placement, as shown in the dining room case. By contrast, our method generates precise and diverse objects in accordance with the prompt semantics—such as the small items on the shelf and six chairs in the dining room case—while effectively capturing 3D spatial arrangements, exemplified by "The bed is located near the wall corner" in the bedroom case.

### 4.5 Density Controllability

We verify the controllability of object density in our layout generation model using different input prompts. As shown in Figure 4, the first row displays the different types of input prompts, which only differ in the final sentence. These prompts adjust the density of objects within the layout, from a minimal setup to one with richer details. This highlights the model's ability to control scene density based on the granularity of the input description, thus offering flexibility in layout customization for various applications.

### 4.6 Ablation Study

We perform the ablation experiments to validate the effectiveness of a single training dataset and report the results in the lower part of Table 3. As shown in the table, when the training data and ground truth come from the same dataset, our model trained on a single dataset achieves the best FID and KID compared with the other two methods. However, its performance drops significantly when evaluated on data from different datasets. This indicates that, while models trained on a single dataset can effectively fit the distribution of that dataset, they struggle to generalize to varied data. For example, a model trained on the professional CAD designs dataset (3D-FRONT) encounters difficulties in generating scenes that align with real-world scans (Matterport) or procedurally generation (Inf3DLayout) dataset. In contrast, our method trained on the multi-source M3DLayout dataset achieves balanced performance across data types, producing more realistic and controllable layouts.

### 4.7 User Case Study

We conducted a perceptual study to evaluate the quality of our generated layouts against DiffuScene and InstructScene. We recruited 42 participants to rate 15 scenes across three room types (dining room, bedroom, and living room). For each scene, participants were shown a text description, a top-down rendering, and the generated layout for each of the three methods, rating them on a 1-to-5 scale across six metrics. All participants were volunteers without compensation. The overall results are summarized in Figure 6. As shown, our method outperformed both baselines in the vast majority of metrics and room categories. The most significant advantage was observed in the Scene Richness metric. The performance gap was less pronounced in living rooms, likely because key layouts are harder to assess from the top-down view used compared to the iconic beds or dining sets in other rooms. These results confirm that users perceive our generated layouts as more detailed, coherent, and of higher quality.

## 5 Conclusion

We introduced M3DLayout, a large-scale, multi-source dataset for 3D indoor layout generation from structured text descriptions. It integrates real-world scans, professional designs, and procedurally generated scenes. Each layout is paired with structured descriptions that cover global scene summaries, relational placements of large furniture, and fine-grained arrangements of small objects. This multi-source, richly annotated structure enables the learning of diverse spatial and semantic patterns across a wide variety of indoor environments. We established a benchmark using

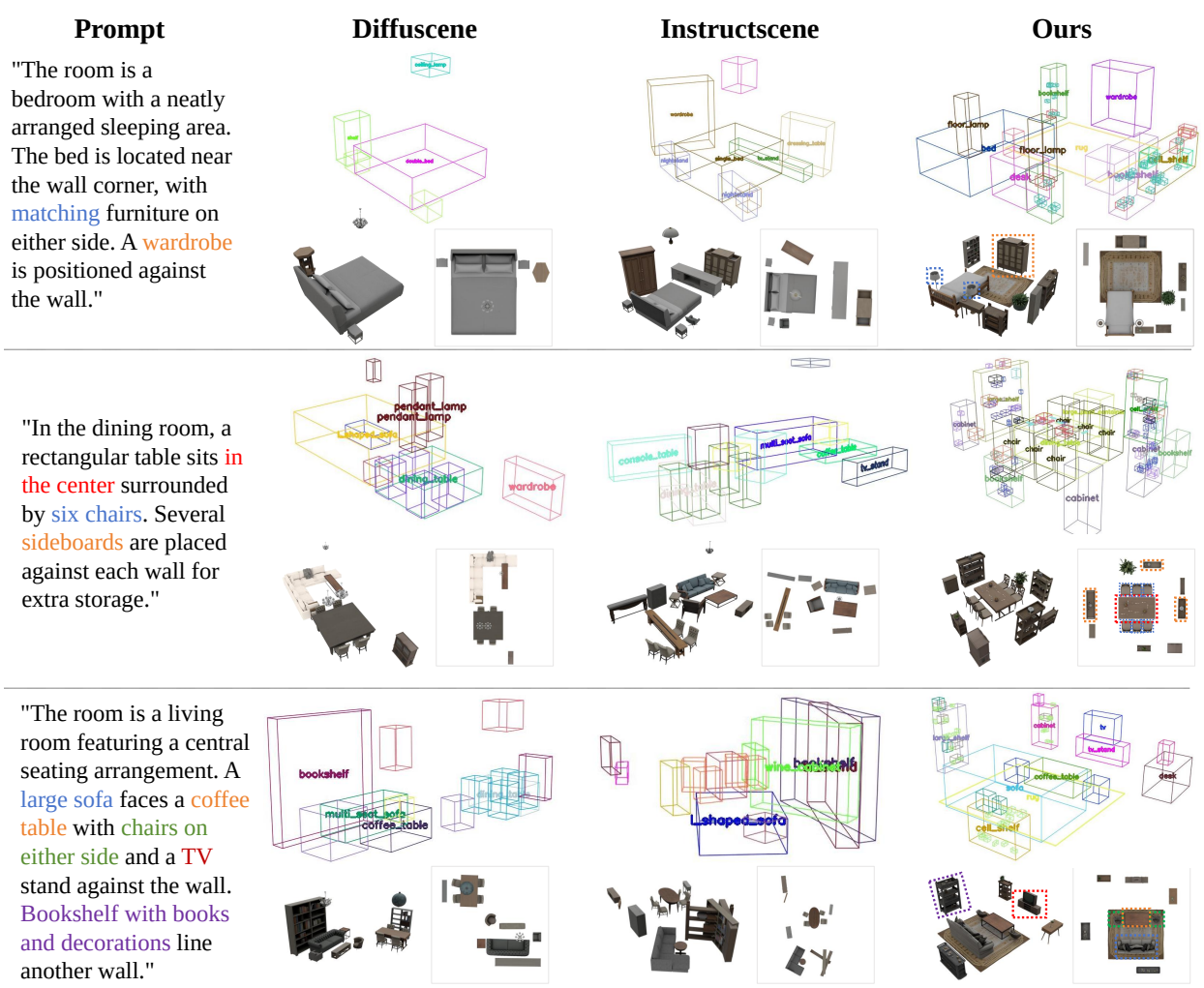


Figure 5: **Qualitative comparison of different methods on diverse room types.** From top to bottom: bedroom, dining room, and living room generation results. Each row shows the input prompt and generated layouts from Diffuscene, Instructscene, and our method. Trained on the M3DLayout dataset, our method produces richer layout details from text descriptions.

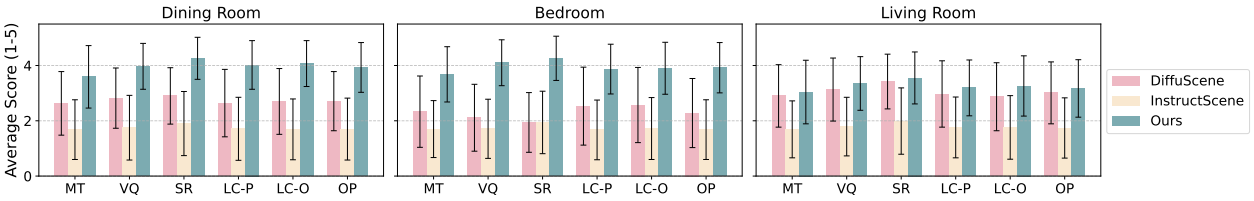


Figure 6: **User case study results.** The charts compare our method against DiffuScene and InstructScene across dining rooms, bedrooms, and living rooms. Bars represent the average user score for six metrics: Match with Text (MT), Visual Quality (VQ), Scene Richness (SR), Layout Coherence (Position) (LC-P), Layout Coherence (Orientation) (LC-O), and Overall Preference (OP).

a text-conditioned diffusion model, and experimental results validate that training on our dataset enhances the diversity and detail of generated layouts, with the Inf3DLayout subset in particular enabling more complex and richly annotated scenes. Despite these advances, our dataset has certain limitations. The structured descriptions, while comprehensive, are partly generated via language models and may contain noise. We hope M3DLayout serves as a valuable resource for advancing research in text-driven 3D scene synthesis.

## REFERENCES


Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 2614–2623, 2019.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *2017 International Conference on 3D Vision (3DV)*, pp. 667–676, October 2017. doi: 10.1109/3DV.2017.00081.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. LucidDreamer: Domain-free Generation of 3D Gaussian Splatting Scenes, November 2023.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.261.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation, June 2022.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, S. Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. *Advances in Neural Information Processing Systems*, 36:18225–18250, December 2023.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10913–10922, October 2021. doi: 10.1109/ICCV48922.2021.01075.

Lin Gao, Jia-Mu Sun, Kaichun Mo, Yu-Kun Lai, Leonidas J. Guibas, and Jie Yang. SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation with Fine-Grained Geometry, February 2023.

Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4077–4085, 2016.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models, September 2023.

Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. SceneNN: A Scene Meshes Dataset with aNNotations. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 92–101, Stanford, CA, USA, October 2016. IEEE. ISBN 978-1-5090-5407-7. doi: 10.1109/3DV.2016.18.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation, December 2024.

Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Zexiang Xu, Hong-Xing Yu, Kalyan Sunkavalli, Miloš Hašan, Ravi Ramamoorthi, and Manmohan Chandraker. OpenRooms: An End-to-End Open Framework for Photorealistic Indoor Scene Datasets, September 2021.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior, February 2024.

Gabrielle Littlefair, Niladri Shekhar Dutt, and Niloy J. Mitra. FlairGPT: Repurposing LLMs for Interior Designs. *Computer Graphics Forum*, 44(2):e70036, May 2025. ISSN 0167-7055, 1467-8659. doi: 10.1111/cgf.70036. URL http://arxiv.org/abs/2501.04648. arXiv:2501.04648 [cs].

Hao Ouyang, Kathryn Heal, Stephen Lombardi, and Tiancheng Sun. Text2Immersion: Generative Immersive Scene with 3D Gaussians, December 2023.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive Transformers for Indoor Scene Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12013–12026. Curran Associates, Inc., 2021.

Pulak Purkait, Christopher Zach, and Ian Reid. SG-VAE: Scene Grammar Variational Autoencoder to Generate New Indoor Scenes. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, volume 12369, pp. 155–171. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58585-3 978-3-030-58586-0. doi: 10.1007/978-3-030-58586-0_10.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite Photorealistic Worlds Using Procedural Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12630–12641, June 2023. doi: 10.1109/CVPR52729.2023.01215.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen Indoors: Photorealistic Indoor Scenes using Procedural Generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, June 2024. doi: 10.1109/CVPR52733.2024.02058.

Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, and Ji Hou. ControlRoom3D: Room Generation Using Semantic Proxy Rooms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6201–6210, June 2024. doi: 10.1109/CVPR52733.2024.00593.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic Scene Completion from a Single Depth Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 190–198, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.28.

Xiaohao Sun, Divyam Goel, and Angle X. Chang. SemLayoutDiff: Semantic Layout Generation with Diffusion Model for Indoor Scene Synthesis, August 2025. URL http://arxiv.org/abs/2508.18597. arXiv:2508.18597 [cs] version: 1.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. DiffuScene: Denoising Diffusion Models for Generative Indoor Scene Synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20507–20518, June 2024. doi: 10.1109/CVPR52733.2024.01938.

Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. SceneFormer: Indoor Scene Generation with Transformers. In *2021 International Conference on 3D Vision (3DV)*, pp. 106–115, December 2021. doi: 10.1109/3DV53792.2021.00021.

Yian Wang, Xiaowen Qiu, Jiageng Liu, Zhehuan Chen, Jiting Cai, Yufei Wang, Tsun-Hsuan Wang, Zhou Xian, and Chuang Gan. Architect: Generating Vivid and Interactive 3D Scenes with Hierarchical 2D Inpainting. *Advances in Neural Information Processing Systems*, 37:67575–67603, December 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21469–21480, 2025.

Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pp. 1625–1632, 2013.

Bangbang Yang, Wenqi Dong, Lin Ma, Wenbo Hu, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. DreamSpace: Dreaming Your Room Space with Text-Driven Panoramic Texture Propagation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 650–660. IEEE Computer Society, March 2024a. ISBN 979-8-3503-7402-5. doi: 10.1109/VR58804.2024.00085.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli Vanderbilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language Guided Generation of 3D Embodied AI Environments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16277–16287, June 2024b. doi: 10.1109/CVPR52733.2024.01536.

Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. CommonScenes: Generating Commonsense 3D Indoor Scenes with Scene Graph Diffusion. *Advances in Neural Information Processing Systems*, 36:30026–30038, December 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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 167–184, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72664-4. doi: 10.1007/978-3-031-72664-4_10.

Genghao Zhang, Yuxi Wang, Chuanchen Luo, Shibiao Xu, Junran Peng, Zhaoxiang Zhang, and Man Zhang. FurniScene: A Large-scale 3D Room Dataset with Intricate Furnishing Scenes, January 2024.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling, July 2020.

# A APPENDIX

## A.1 Dataset Split and Implementation Details

**Dataset Split.** To validate the fidelity and controllability of the generated 3D layouts, we randomly split the M3DLayout dataset into 12062 layouts for training and 3018 layouts for validation. For ablation studies, models were trained on three independent datasets with corresponding splits (training/validation): 4603/1151 for 3DFront, 1347/337 for Matterport, and 6112/1530 for Inf3DLayout, respectively. Regarding the testing, the same set of text prompts ($500 \times 3$), generated by GPT-4o and restricted to the bedroom, dining room, and living room, is used in all experiments to ensure fairness.

**Implementation Details.** The model is trained for 30k epochs using AdamW optimizer with a learning rate of $2 \times 10^{-4}$ and linear noise schedule.

## A.2 Baseline

We compare our method with two state-of-the-art scene generation approaches: (1) DiffuScene Tang et al. (2024), a diffusion model for 3D indoor scene synthesis that denoises unordered object attributes to produce physically plausible layouts. (2) InstructScene Lin & Mu (2024), a graph diffusion model that integrates a semantic graph with a layout decoder to synthesize 3D indoor scenes from natural language instructions. Both methods allow for the conditioning on text prompts. For inference with DiffuScene, we employ the officially released model weights for the bedroom, dining room, and living room, and follow the public implementation to train InstructScene on the same room types.

## A.3 Metrics

Following prior works Tang et al. (2024); Lin & Mu (2024), we adopt Fréchet Inception Distance (FID) Heusel et al. (2017) and Kernel Inception Distance (KID) Bińkowski et al. (2018) to quantify the fidelity of scenes synthesized from layouts by measuring the similarity between generated and ground-truth top-down renderings. Meanwhile, we employ the CLIP-Score Hessel et al. (2021) to evaluate the controllability of generated layouts by computing the cosine similarity between CLIP-encoded features of the generated renderings and the given prompts. To this end, we first employ a text2mesh model Xiang et al. (2025) to generate the required object instances and then retrieve both the ground-truth and synthesized scenes conditioned on the layout. More details about self-designed object retrieval process and visualization are provided in the Appendix A.6.

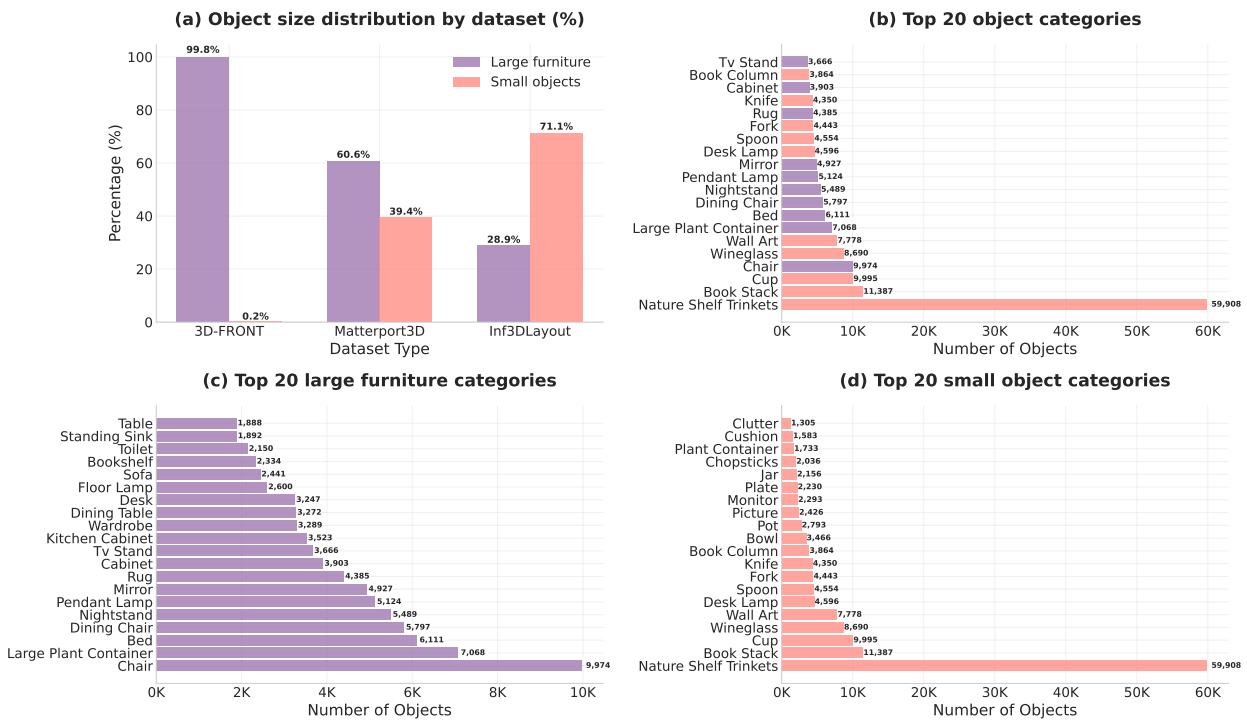


Figure 7: **Object distribution statistics of the M3DLayout dataset.** (a) Size distribution (large/small) of objects by source. (b) Overall ranking of the top 20 object categories. (c-d) Rankings for large and small objects, respectively.

### A.4 Dataset Layout Visualization

We provide diverse types of 3D scene data from our M3DLayout dataset in Figure 8, which includes scenes from CAD designs sourced from the 3D-FRONT dataset at the first row, scenes derived from real-world scans, specifically from the Matterport3D dataset at the second row and procedurally generated scenes from Infinigen at the third row. These images demonstrate the flexibility of the M3DLayout dataset in representing a wide spectrum of interior environments, from synthetic CAD designs to real-world captures and generative models.

### A.5 Generated Layout Visualization

We visualize more generated layouts by our model trained on the M3DLayout dataset, involving bedroom, living room, and dining room in Figure 9. From the table, it is evident that our method achieves remarkable performance in both layout coherence and the richness of objects in the generated scenes. These qualitative visualizations further highlight that our method surpasses prior state-of-the-art approaches in both fidelity and controllability.

### A.6 Object Retrieval

To effectively visualize the generated layouts and meet the evaluation requirements, such as FID (Fréchet Inception Distance), KID (Kernel Inception Distance), and CLIP score, we present a simple, effective and scalable pipeline for layout-to-scene object retrieval.

In Figure 10, we first build our retrieval dataset by constructing huge amounts of *prompts* for 95 object categories (See details in Table 4), which covers all objects for our dataset, as input for *Text-to-3D generation model* (TRELLIS Xiang et al. (2025)). By delicatly designing our prompts, we can obtain 3D assets with different scales, textures and application scenarios, which can be generalizable to handle with intricate object retrieval process.

In the *Object Selection* phase, the retrieved object, such as a bed, undergoes attribute extraction through the *Attribute Solver*. This solver precomputes attributes like width, height, and depth ratios for each object in the retrieval dataset, and extracts scalar and categorical information from bounding box (BBox) of each object in the generated layout. The

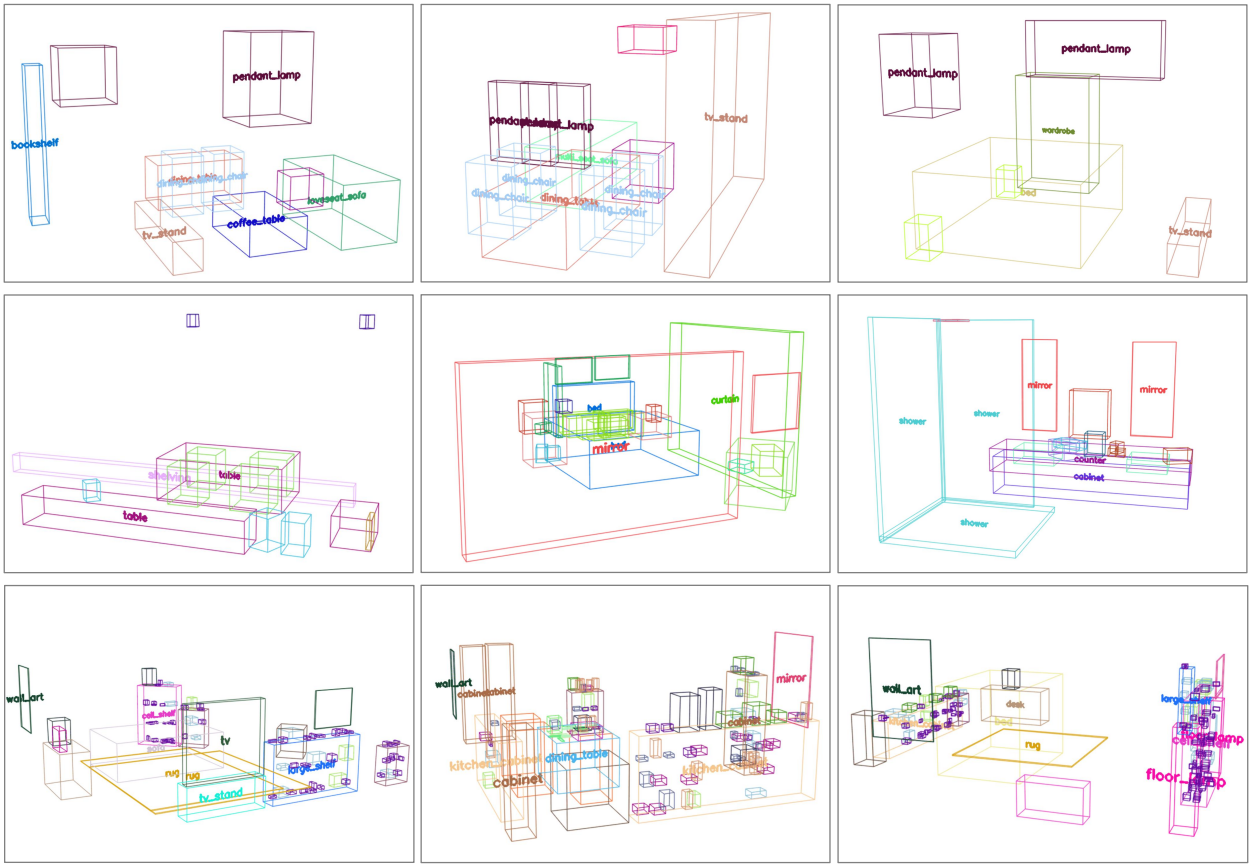


Figure 8: Samples from the M3DLayout dataset. The first row shows scenes from CAD designs (3D-FRONT), the second row from real-world scans (Matterport3D), and the third row from procedurally generated scenes (Infinigen).

BBox, along with additional dataset information, is passed to a *Shape & Category Similarity Solver* to match the most appropriate object.

Finally, after iterating all objects in the scene, all best matching objects are chosen and their properties (such as translations, sizes, and rotation) are determined, culminating in the retrieval of the desired 3D scene. This multi-step process ensures accurate retrieval and selection of 3D objects for following applications.

After retrieving successfully, the visualizations provided in this Figure 11 aim to assess both the quality of the generated object retrievals and the performance of the evaluation metrics. The pure color renderings eliminate the influence of textures, making it easier to assess the layout's alignment and object retrieval accuracy using metrics like FID and KID. Meanwhile, the textured renderings offer a visually richer evaluation and applications for users. This approach allows for a comprehensive understanding of the model's effectiveness from both a metric-based and visual quality perspective.

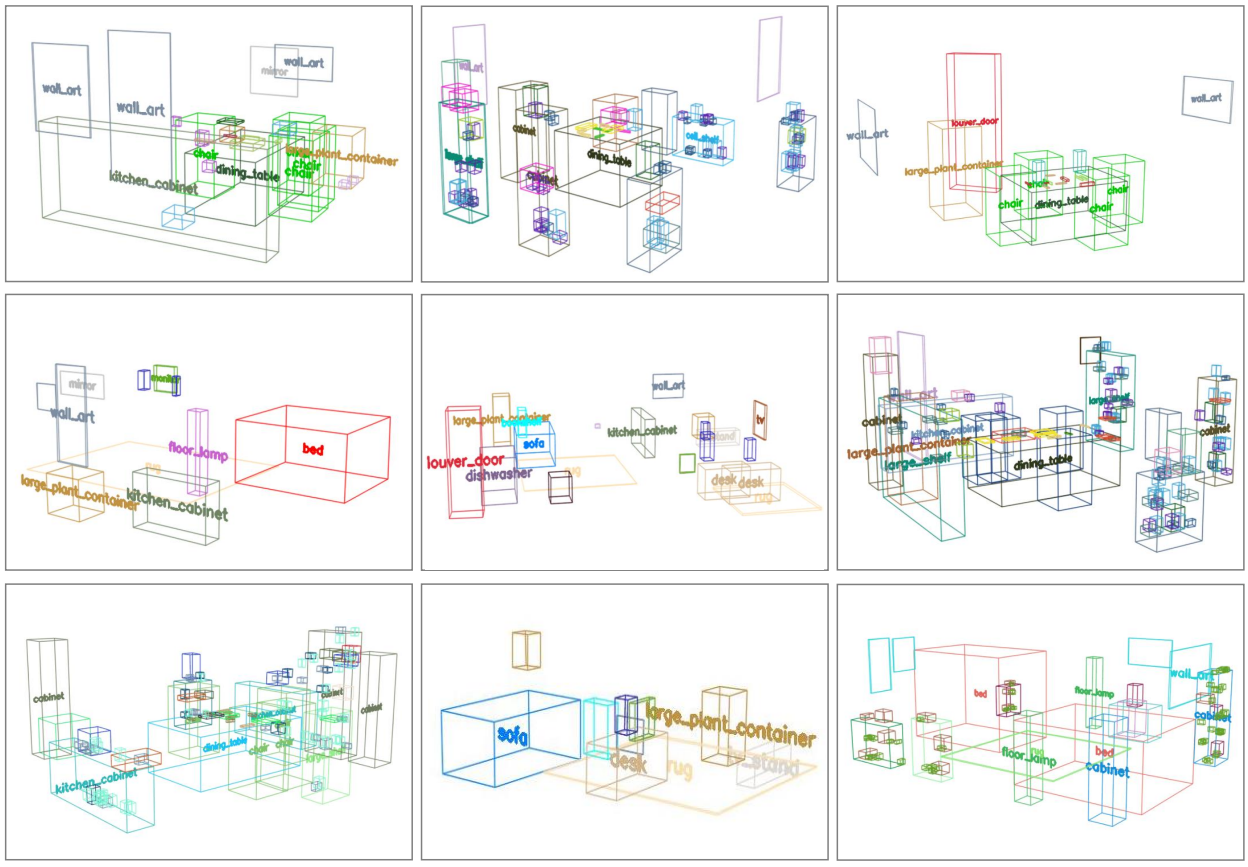


Figure 9: Generated layouts by our model trained on the M3DLayout dataset, visualized for randomly selected bedroom, living room, and dining room.

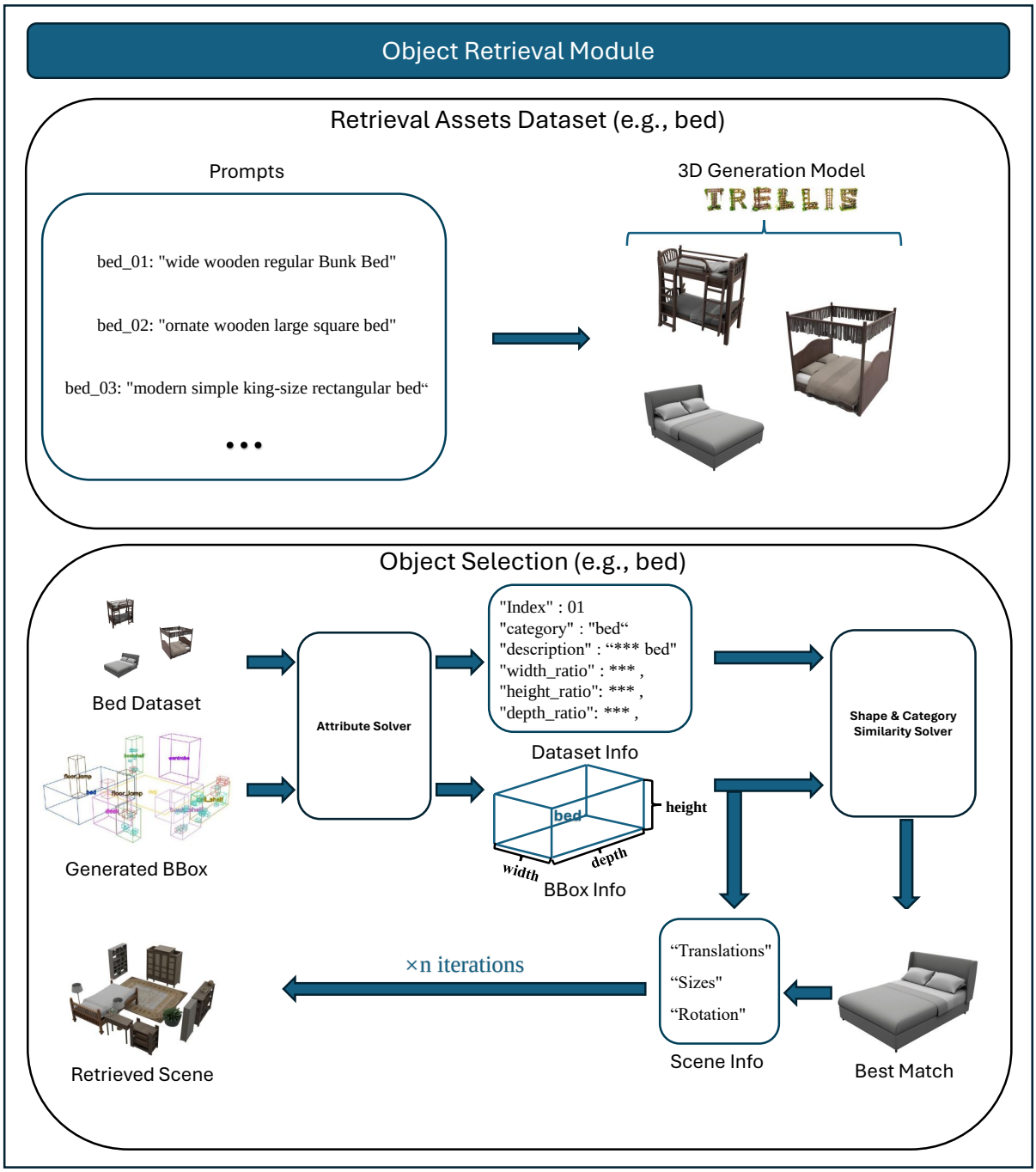


Figure 10: **Retrieval process of layout-to-scene.** This flowchart illustrates the process of retrieving and selecting 3D objects based on generated BBox information.

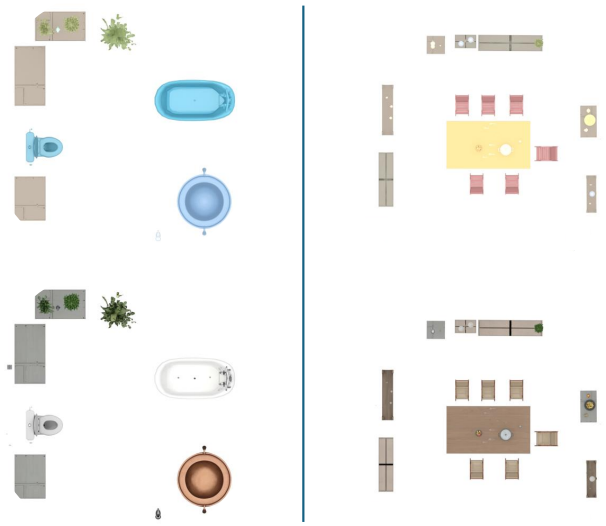

Figure 11: **Retrieval visualization of generated layouts.** The first row displays the retrieved 3D scenes' renderings with pure color schemes. The second row shows the retrieved 3D scenes' renderings with original textures applied.

| Category | Objects (95 in total) |
|---|---|
| **Lighting** | lighting, ceiling_lamp, pendant_lamp, floor_lamp, desk_lamp, fan |
| **Tables** | table, coffee_table, console_table, corner_side_table, round_end_table, dining_table, dressing_table, side_table, nightstand, desk, tv_stand |
| **Seating** | seating, chair, armchair, lounge_chair, chinese_chair, dining_chair, dressing_chair, stool, sofa, loveseat_sofa, l_shaped_sofa, multi_seat_sofa |
| **Beds** | bed, kids_bed |
| **Shelves & Book storage** | shelf, shelving, large_shelf, cell_shelf, bookshelf, book, book_column, book_stack, nature_shelf_trinkets |
| **Cabinets & Wardrobes** | cabinet, kitchen_cabinet, children_cabinet, wardrobe, wine_cabinet |
| **Appliances & Electronics** | appliances, microwave, oven, beverage_fridge, tv, monitor, tv_monitor |
| **Kitchen & Tableware** | pan, pot, plate, bowl, cup, bottle, can, jar, wineglass, chopsticks, knife, fork, spoon, food_bag, food_box, fruit_container |
| **Bathroom fixtures** | bathtub, shower, sink, standing_sink, toilet, toilet_paper, toiletry, faucet, towel |
| **Doors, Windows & Coverings** | glass_panel_door, lite_door, window, blinds, curtain, vent |
| **Hardware & Controls** | hardware, handle, light_switch |
| **Decor** | plant, large_plant_container, plant_container, vase, wall_art, picture, mirror, statue, basket, balloon, cushion, rug, decoration |
| **Containers & Waste** | bag, box, container, clutter, trashcan |
| **Architecture & Elements** | counter, fireplace, pipe, furniture |
| **Clothes** | clothes |
| **Spaces** | kitchen_space |
| **Gym & Misc** | gym_equipment |

Table 4: **Category list of retrieval objects.** Our retrieval dataset includes 95 objects which covers nearly all common indoor objects.