M3DLayout: A Multi-Source Dataset of 3D Indoor Layouts and Structured Descriptions for 3D Generation

Supplementary Material

1. Supplementary Material Overview

This supplementary document provides additional technical details and visualization results to support the main paper. The supplementary is organized into four sections:

Sec. 2 Implementation Details. We provide detailed explanations of the dataset splitting strategy in 2.1, both diffusion and autoregressive based model details in 2.2, training details in 2.3, compared baselines details in 2.4 and metric details in 2.5.

Sec. 3 More Dataset Analysis. We further evaluate the quality of our dataset through extensive analysis and validation. Specifically, we quantitatively compare more metrics with 3D-FRONT in 3.1, conduct a human evaluation via a dedicated website in 3.2 to demonstrate its overall complexity, diversity and high quality,.

Sec. 4 More Visualization Results. We provide additional examples that further demonstrate the quality of our dataset, the fidelity of our generative model, and its coherence with textual inputs.

Sec. 5 Ablation. We conduct some ablation studies to show that models trained on a single dataset overfit to its distribution and generalize poorly, whereas training on the multi-source M3DLayout dataset yields consistently balanced, realistic, and controllable layouts across diverse data types.

Sec. 6 Object Retrieval. We present a scalable and self-developed layout-to-scene object retrieval pipeline that accurately maps generated layouts to 3D assets, facilitating reliable metric evaluation and high-fidelity visualizations.

2. Implementation Details

2.1. Dataset Split

Our experiments are conducted on a subset of the M3DLayout dataset containing 15,080 scenes. This subset is randomly divided into 12,062 layouts for training and 3,018 layouts for validation. For the ablation studies in Table 4, models are additionally trained on three independent datasets with the following splits (training/validation): 4,603/1,151 for 3D-FRONT, 1,347/337 for Matterport3D, and 6,112/1,530 for Inf3DLayout. For all evaluations, and to ensure a fair comparison with prior work, we do *not* use the scene description texts provided in M3DLayout. Instead, we generate 500 scene descriptions each for *bedroom*, *dining room*, and *living room* using GPT-4o, resulting in a total of 1,500 test descriptions. This unified test set is used across all experiments for fairness.

2.2. Model Details

Diffusion-based Model. Our diffusion-based model represents a scene as a sequence of N objects, where the maximum number of objects N is fixed at 120. Padding objects with a PAD category label is applied to maintain a consistent sequence length. The sequence is ordered by the (x, z, y) coordinates of the bounding box centers, following recent best practices in the 3D part-level generation community for layout generation. Each object is parameterized by a concatenated vector $[l_i, s_i, \cos\theta_i, \sin\theta_i, c_i]$, which includes its location ($l_i \in \mathbb{R}^3$), size ($s_i \in \mathbb{R}^3$), orientation ($\theta_i \in \mathbb{R}$, encoded using $\cos\theta_i$ and $\sin\theta_i$), and category label ($c_i \in \mathbb{R}^C$).

The denoiser is a UNet-based architecture built upon 1D convolutions with skip connections. It independently encodes and predicts the distinct attributes of the object. We use a pre-trained BERT encoder to extract text embeddings z from the input description. This language guidance is injected into the denoiser network through cross-attention layers, enabling the network to predict the noise ϵ_ϕ conditioned on the timestep t and the text embedding z.

The training objective consists of the scene loss L_{sce}, which minimizes the difference between the actual and predicted noise, supplemented with an IoU regularization term L_{iou}. For two 3D axis-aligned bounding boxes A and B, we define

$$L_{\text{IoU}}(A, B) = \frac{I}{V_A + V_B - I + \epsilon},$$

$$I = \prod_{k \in \{x,y,z\}} \max\big(0, \min\big(k_2^A, k_2^B\big) - \max\big(k_1^A, k_1^B\big)\big)\,.$$

This term penalizes object intersections and encourages physically plausible arrangements.

The key difference between DiffuScene and Our DIFF-M3DLayout is the use of shape codes (object geometric embedding) during training. We trained DIFF-M3DLayout (our diffusion-based trained on M3DLayout) only with layout data (without shape codes) because a subset of M3DLayout, Matterport, comes from real-world scans and lacks complete object shapes to extract shape codes. Table 1

shows metrics of the two methods trained on 3D-FRONT. The performance is quite similar.

Autoregressive Model. Our autoregressive model shares the same scene representation as our diffusion-based model for consistency. The autoregressive approach does not have a strict upper limit on the sequence length, but we still utilize data containing up to 120 objects per scene for training. A scene is represented as a sequence of objects, bookended by special start-of-sequence (SOS) and end-of-sequence (EOS) tokens.

The input text description is similarly encoded into text embeddings using a pre-trained BERT model. At each generation step, the input sequence to the model consists of the text token embeddings and the embeddings of all previously generated objects. A Transformer encoder serves as the backbone network. The attributes of each previously generated object are first projected by an MLP into an embedding space. The network then attends to this combined sequence to predict the parameters of the next object.

The training objective is the scene loss L_{sce}, which minimizes the negative log-likelihood of the ground-truth object parameters given the input context.

2.3. Training Details.

Both our diffusion and autoregressive based models are trained for 30k epochs using the Adam optimizer. For the diffusion model, we use a learning rate of 2×10^{-4} with a step-wise decay factor of 0.5 every 10k steps and a linear noise schedule. For the autoregressive model, we use a learning rate of 1×10^{-4}.

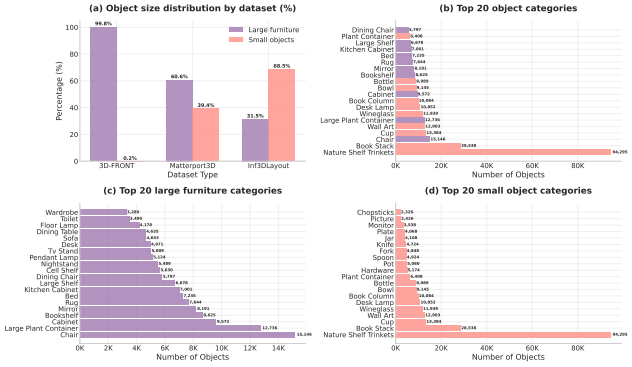


Figure 1. **Object distribution statistics of the M3DLayout dataset.** (a) Size distribution (large/small) of objects by source. (b) Overall ranking of the top 20 object categories. (c-d) Rankings for large and small objects, respectively.

2.4. Compared Baselines

We compare our method with two state-of-the-art scene generation approaches: (1) DiffuScene [5], a diffusion model for 3D indoor scene synthesis that denoises unordered object attributes to produce physically plausible layouts. (2) InstructScene [4], 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.

2.5. Metrics

Following prior works [4, 5], we adopt Fréchet Inception Distance (FID) [3] and Kernel Inception Distance (KID) [1] 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 [2] 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 [6] to generate the required object instances and then retrieve both the ground-truth and synthesized scenes conditioned on the layout.

As for the calculation of FID and KID, the real images are typically 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.

3. More Dataset Analysis

3.1. Layout Complexity And Diversity

We further strengthened our quality validation by quantitatively benchmarking our dataset. We evaluated the data across key dimensions, including scale, layout complexity, and diversity. The quantitative results in Table 3 demonstrate that our M3DLayout significantly surpasses 3D-FRONT across all these metrics, confirming the higher quality and richness of our data. To provide a robust assessment, we evaluate layout complexity using the Average Number of Large Objects (identified by categories, to ensure independence from small-object density) and Average Token Counts (BPE sequence length) as a proxy for description to model the layout. For dataset diversity, we benchmarked 500 scenes using Category Entropy (measuring object class uniformity), Intra-Category Spatial Distance (calculating average weighted nearest-neighbor distance for geometric variation), and CLIP Embedding Distance (averaging cosine distances of text embeddings for semantic diversity, $D_{\text{text}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \big(1 - \langle \hat{\mathbf{e}}_i, \hat{\mathbf{e}}_j \rangle\big)$).

Table 1. DiffuScene and Our DIFF-M3DLayout.

Method	**FID ↓**			**KID ↓**			**CLIP-Score ↑**
	3D-FRONT	Matterport	Inf3DLayout	3D-FRONT	Matterport	Inf3DLayout	
DiffuScene (trained on 3D-FRONT)	29.47	98.03	**102.12**	**10.32**	47.92	**75.49**	0.1982
Ours diffusion-based trained on 3D-FRONT	**27.33**	**83.88**	110.98	10.59	**21.80**	83.45	**0.2083**

Table 2. Comparison of generation efficiency with state-of-the-art methods.

Method	Generation Time per Scene (s)	Generation Time per BBox (s)	Average BBoxes per Scene
Diffuscene	17.4067	1.7674	9.849
InstructScene	1.3438	0.1280	10.502
Ours (DIFF-M3DLayout)	30.18	1.639	**18.4**
Ours (AR-M3DLayout)	**0.2348**	**0.01587**	14.8

3.2. Human Evaluation

Besides User Case Study, we conduct a human evaluation on 1,000 (4.68% of total) randomly sampled scenes from M3DLayout. Evaluators (around 200) are asked to rate three aspects: layout rationality, overall description accuracy, and object-level description accuracy, using a 5-point Likert scale (5: completely correct, 1: completely incorrect). Each sample is independently evaluated by two participants, with cross-validation applied to ensure consistency. The results show that the average scores for the three criteria are 4.423, 4.462, and 4.436, respectively. Moreover, the proportions of samples receiving scores ≥ 4 are 94.05%, 94.08%, and 93.84%, respectively. These results further validate the high quality of the M3DLayout dataset and demonstrate its strong alignment with human judgments

4. More Visualization Results

4.1. Dataset Layout Visualization

We provide diverse types of 3D scene data from our M3DLayout dataset in Figure 2, 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 Inf3DLayout 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.

4.2. Generated Layout Visualization

We visualize more generated layouts by both our diffusion and autoregressive model trained on the M3DLayout dataset, involving bedroom, living room, and dining room in Figure 3 and Figure 4 respectively. 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. We also provide more visualization to indicate strong text coherence of our generated layouts in Figure 5. These qualitative visualizations further highlight that our method surpasses prior state-of-the-art approaches in both fidelity and controllability.

5. Ablation Study

We perform the ablation experiments using the diffusion- and autoregressive-based methods to validate the effectiveness of a single training dataset and report the results in Table 4. As shown in the table, for the diffusion-based method, when the training data and ground truth come from the same dataset, the model trained on a single dataset achieves the best FID and KID compared with the other two models. 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 examplprovided textual guidancee, 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. Furthermore, the autoregressive-based method exhibits the same characteristics. In contrast, these methods trained on the multi-source M3DLayout dataset achieves balanced performance across data types, producing more realistic and controllable layouts (see Table 3 in Section 5.2 of the present work for comprehensive data).

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

Table 3. M3DLayout with 3D-FRONT comparison.

Dataset	Layout Complexity			Dataset Diversity		
	Avg. # Objs	Avg. # Large	Avg. Tokens	Entropy ↑	Spatial Dist. ↑	CLIP Dist. ↑
3D-FRONT	6.9	6.8	19.5	0.635	1.416	0.133
M3DLayout (Ours)	**26.8**	**8.4**	**77.9**	**0.720**	**3.854**	**0.203**

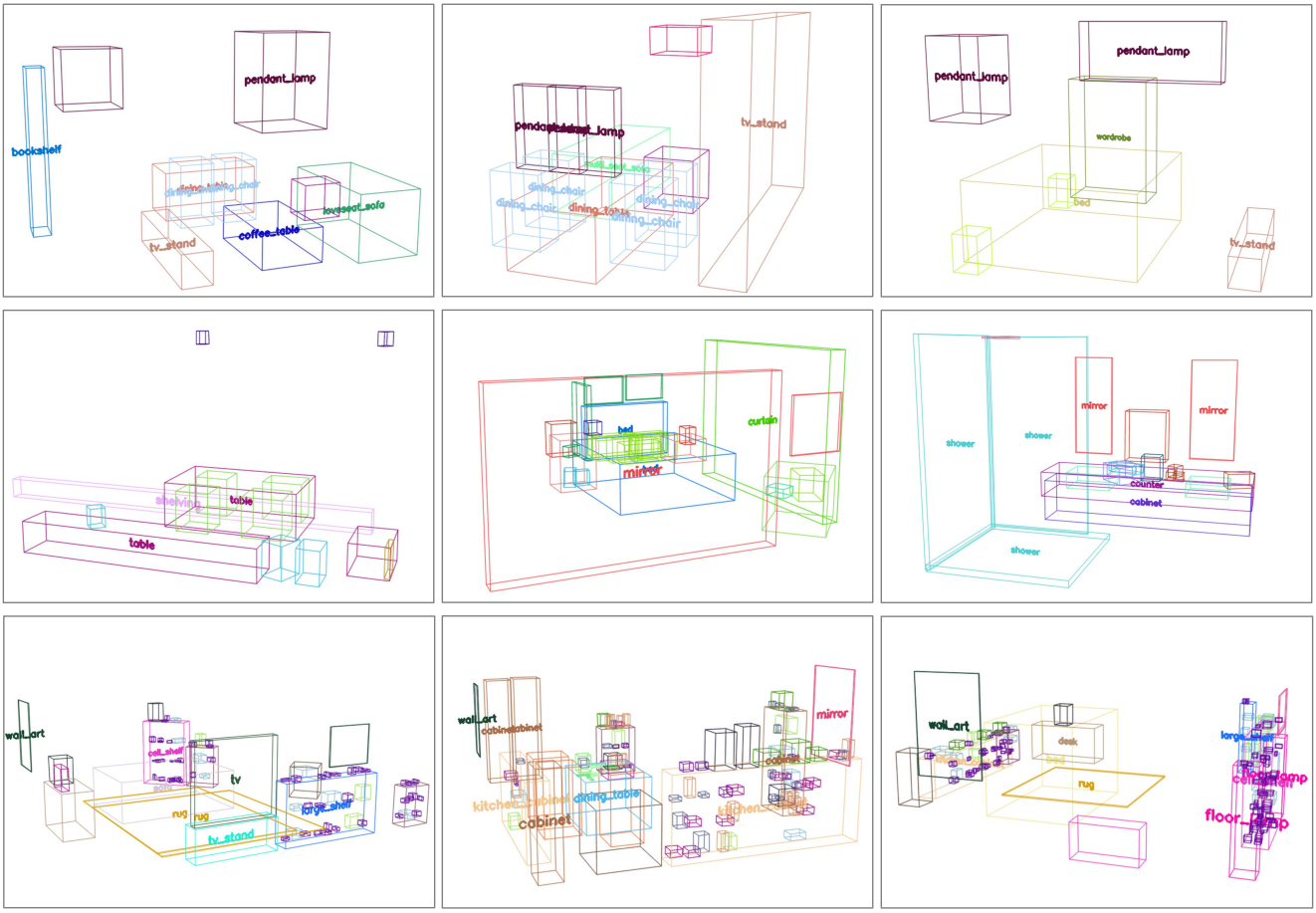


Figure 2. Samples from the M3DLayout dataset. The first row shows scenes from CAD designs (3D-FRONT), the second row from real-world scans (Matterport3D), and the third row from procedurally generated scenes (Infinigen).

score, we present a simple, effective and scalable pipeline for layout-to-scene object retrieval.

In Figure 6, we first build our retrieval dataset by constructing huge amounts of *prompts* for 95 object categories (See details in Table 5), which covers all objects for our dataset, as input for *Text-to-3D generation model* (TRELLIS [6]). 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 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

Table 4. **Ablation studys of diffusion-based methods trained on different datasets.** Lower FID/KID (×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.

Method		FID ↓			KID ↓			CLIP-Score ↑
		3D-FRONT	Matterport	Inf3DLayout	3D-FRONT	Matterport	Inf3DLayout	
Diffusion-based	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
Autoregressive-based	Ours (3D-FRONT)	30.02	85.08	117.39	12.09	24.46	95.11	0.2056
	Ours (Matterport)	73.36	61.08	132.54	45.11	12.58	122.21	0.1959
	Ours (Inf3DLayout)	105.24	122.19	60.73	70.81	65.21	43.21	0.2026

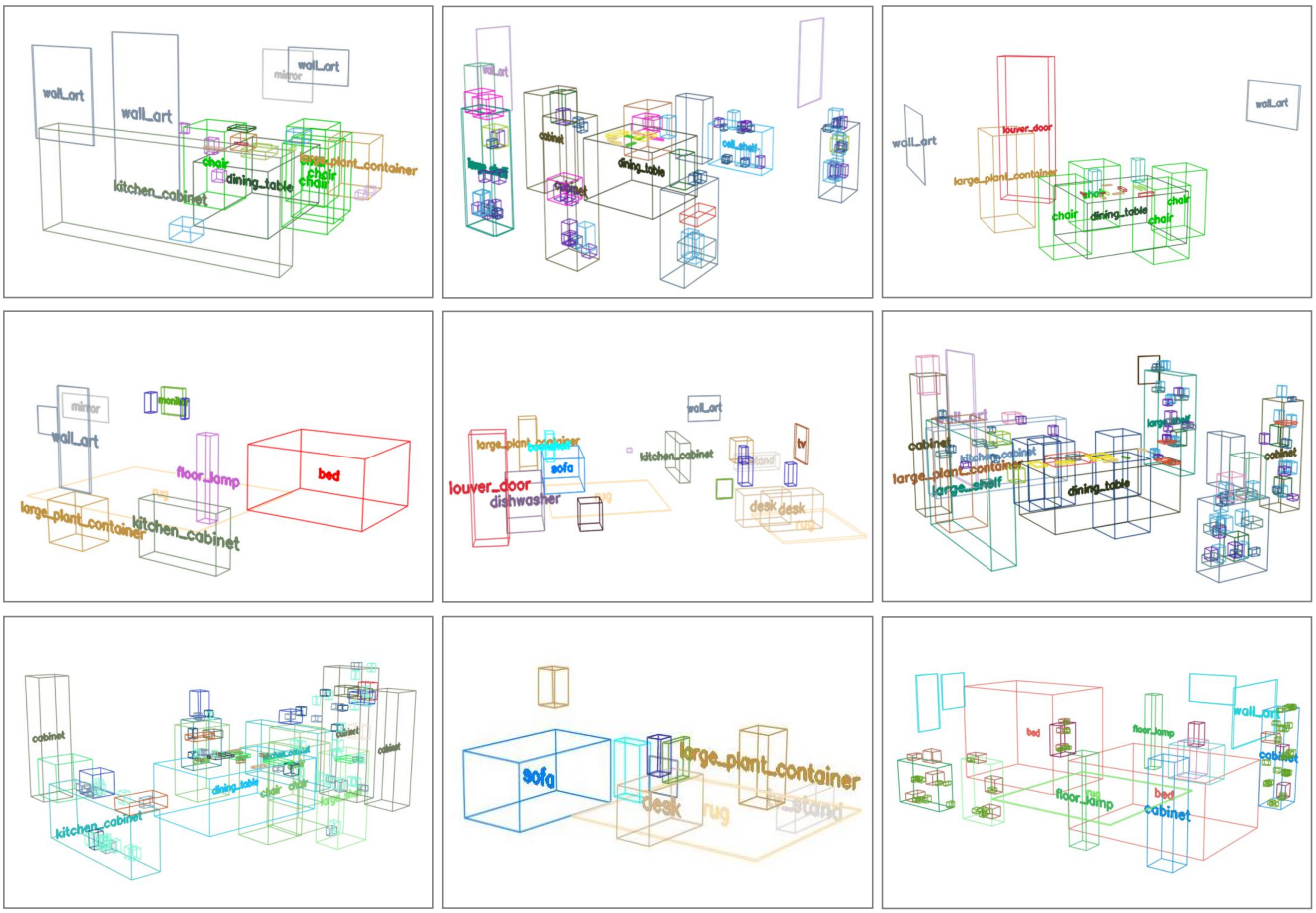


Figure 3. Generated layouts by our model trained on the M3DLayout dataset, visualized for randomly selected bedroom, living room, and dining room.

in this Figure 7 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.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 2

[2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric

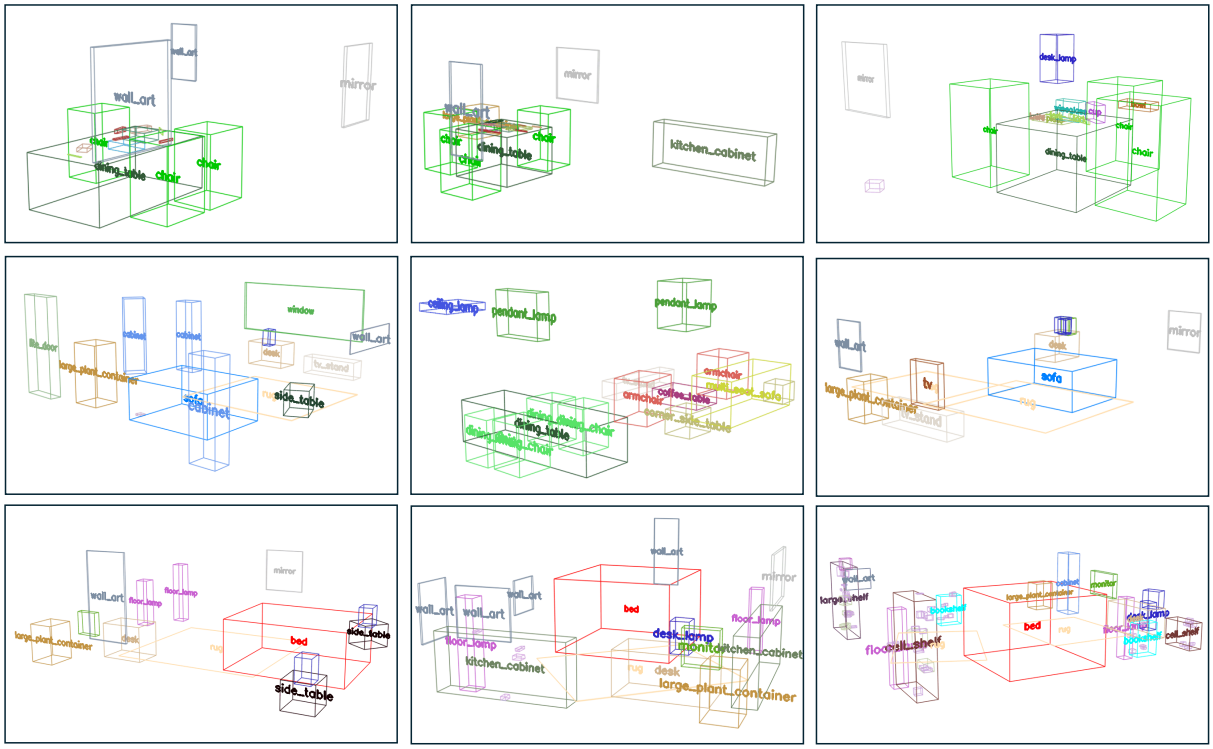


Figure 4. Generated layouts by our autoregressive model trained on the M3DLayout dataset, visualized for randomly selected bedroom, living room, and dining room.

for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2

[3] 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. 2

[4] Chenguo Lin and Yadong Mu. InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior, 2024. 2

[5] 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)*, pages 20507–20518, 2024. 2

[6] 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*, pages 21469–21480, 2025. 2, 4

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

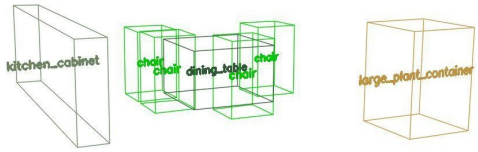
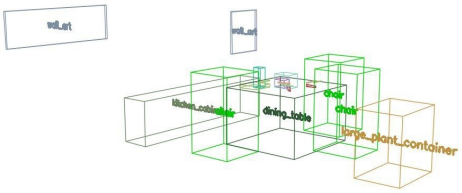
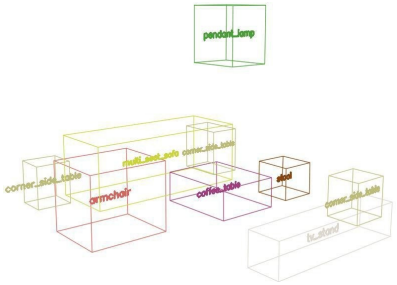
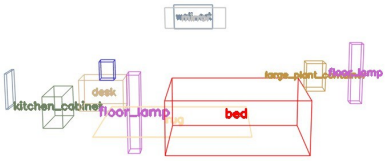
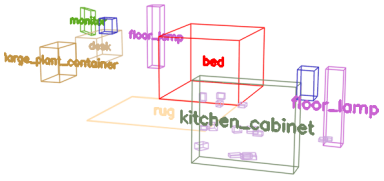
Input	Output
A cozy dining room with a square table in the center. Chairs rest on all sides, and a planter adds greenery to one corner.	
A dining room with a circular table has evenly placed chairs around it. A cabinet lines the wall, enhancing the dining experience.	
This living room has a cozy corner with a sectional sofa and a small round coffee table. A tv stand lines the wall opposite the sofa, and a side table is placed beside it.	
In this bedroom, the bed is against the back wall, with a floor lamp at its foot. A small desk and table lamp create a work area.	
The room is a bedroom and includes a single bed pushed into a corner. Across the room, a cabinet provide a place to store things.	

Figure 5. Our generated layouts exhibit strong adherence to the provided textual guidance.

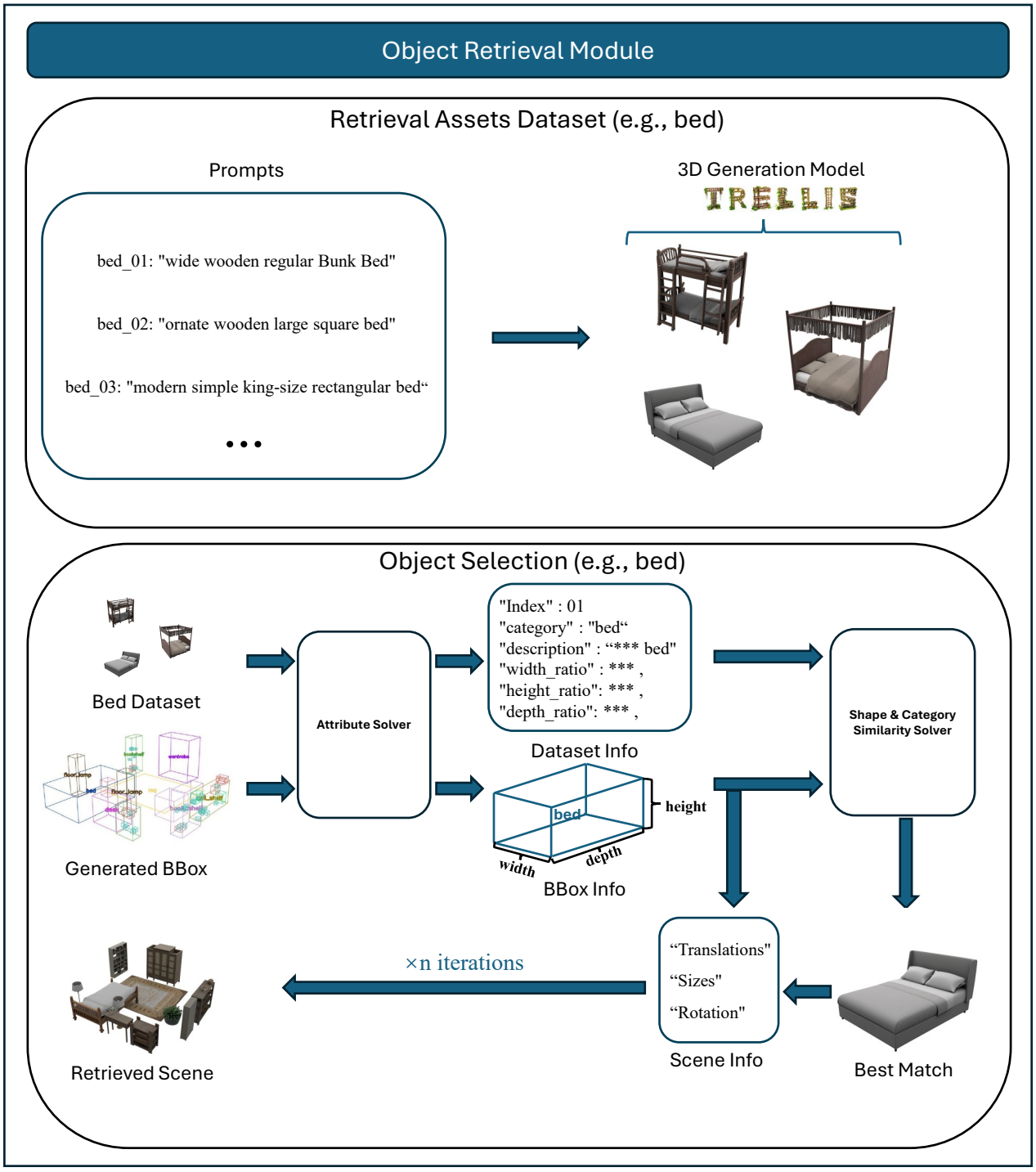


Figure 6. **Retrieval process of layout-to-scene.** This flowchart illustrates the process of retrieving and selecting 3D objects based on generated BBox information.

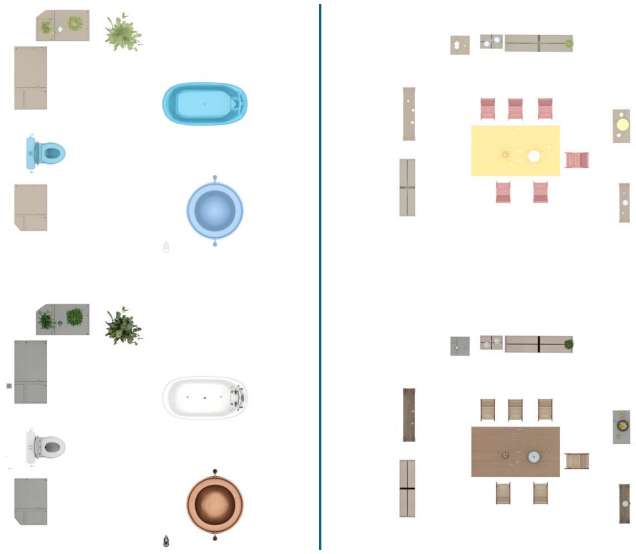


Figure 7. **Retrieval visualization of generated layouts.** The first row displays the retrieved 3D scenes' renderings with pure color schemes. The second row shows the retrieved 3D scenes' renderings with original textures applied.