UNLEASHING CHAIN-OF-THOUGHT REASONING FOR 3D SCENE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, 3D Scene Synthesis (3DSS) has attracted growing interest for its applications in autonomous intelligent systems. However, conventional methods rely heavily on manual effort and expert knowledge, and are often criticized for their limited quantity and diversity. On the other hand, large language models (LLMs) have achieved remarkable performance across a wide range of tasks, making automatic 3DSS from textual conditions feasible. However, due to the lack of spatial reasoning capabilities, they still face significant challenges in generating coherent 3D scenes, often resulting in inappropriate objects, incorrect arrangements, and spatial conflicts. In this paper, for the first time, we explore the potential of chain-of-thought (CoT) reasoning in 3DSS and propose an innovative approach to enhance the spatial reasoning and generation capabilities of LLMs. Specifically, we introduce a cascading Scene-CoT generation pipeline that decomposes complex design tasks into manageable subtasks through hierarchical agents, complemented by an iterative spatial optimization strategy to resolve conflicting constraints commonly encountered by LLMs. Through the interaction between semantic reasoning agents and spatial constraint optimization modules, our approach is capable of generating 3D scenes accompanied by reasoning traces and computational processes. Furthermore, we propose a two-stage progressive training framework that first distills a base model using Scene-CoT to acquire initial scene generation capabilities, and then improves the coherence and plausibility of 3D scene generation through reinforcement learning. Extensive experiments demonstrate the effectiveness of our Scene-CoT dataset and model in enabling high-quality automatic 3D scene synthesis.

1 INTRODUCTION

Synthesizing physically-plausible and human-preferred 3D indoor scenes remains a long-standing challenge in computer vision, playing a pivotal role in applications such as interior design, gaming development, and embodied AI (Deitke et al., 2022; Gao et al., 2023; Du et al., 2024). Traditionally, the creation of such scenes has relied on professional designers with extensive domain knowledge and proficiency in specialized tools, which is both labor-intensive and time-consuming. Consequently, the increasing demand for scalable and efficient solutions has spurred significant advancements in automated 3D Scene Synthesis (3DSS) (Gao et al., 2024; Yang et al., 2024d; Wang et al., 2024; Wu et al., 2024; Yi et al., 2023; Shi et al., 2022; Zhang et al., 2024; Öcal et al., 2024).

Recent approaches proposed to synthesize 3D scenes by incorporating additional input conditions, such as images (Chen et al., 2023; Huang et al., 2018), scene graphs (Gao et al., 2024; Dhamo et al., 2021; Ost et al., 2021), or semantic layouts (Yang et al., 2024c; Chen et al., 2025). While these methods achieve impressive performance, the requirement for these conditions imposes an additional burden on users. In contrast, text-driven 3D indoor scene generation (Hwang et al., 2023) requires only natural language descriptions as input. For example, DiffuScene(Tang et al., 2024) introduces a denoising diffusion model that generates diverse and realistic 3D indoor scenes by learning to denoise unordered sets of object attributes such as location, orientation, and semantics. However, the reliance of such methods on predefined scene and furniture categories during training restricts their generalization to various room types and unseen object distributions.

Motivated by the remarkable capabilities of large language models (LLMs) (Achiam et al., 2023; Bai et al., 2023; Guo et al., 2025a; Hurst et al., 2024; Jaech et al., 2024; Yang et al., 2024a; Bai et al., 2025; Team et al., 2023) across various tasks, recent approaches have explored their application to 3D scene generation. For example, LayoutGPT (Feng et al., 2024) employs a pre-trained LLM to predict the positions of objects within a 3D space directly. However, due to the limited spatial reasoning ability of LLM, it often generates layouts with object collisions or placements that exceed the room boundaries. Furthermore, LayoutGPT (Feng et al., 2024) operates as a black box, failing to provide explanations regarding the scene generation process. More recently, methods such as HOLODECK (Yang et al., 2024f) and I-Design (Çelen et al., 2024) attempt to address these limitations by introducing multi-LLM systems. While this paradigm improves the scale and diversity of generated scenes, it remains inherently brittle: any failure or noisy output from a single LLM can disrupt the system and degrade the quality of the final scene. Moreover, 3DSS inherently requires models to possess reasoning abilities, including understanding users' instructions, selecting aesthetically pleasing and stylistically coherent objects, and placing them in appropriate positions. However, the potential of LLMs in facilitating step-by-step reasoning for 3D scene generation remains unexplored.

In this paper, we propose a novel framework designed to unleash LLMs' spatial reasoning capabilities in 3DSS by explicitly modeling intermediate steps toward the final layout. To this end, we introduce a cascading data generation pipeline that decomposes complex 3D scene synthesis tasks into four structured subtasks: (1) user preference analysis, (2) functional zone partition, (3) object recommendation, and (4) object placement. Firstly, the system interprets and expands upon the user's instruction by inferring how the user is expected to interact with the scene, along with a comprehensive set of requirements from the perspectives of aesthetics, functionality, and manufacturing. Subsequently, the scene is partitioned into several zones (e.g., resting areas, conversation spaces) based on contextual cues. For each zone, we employ LLMs to recommend relevant objects, where the agent not only incorporates explicitly mentioned items but also infers implicit needs that are not directly stated before. Instead of directly predicting absolute coordinates (Feng et al., 2024), our method predicts the relative spatial relationships among all objects and organizes them into a hierarchical scene tree. We then apply a symbol-centric interface to resolve the dependencies between nodes in the tree and iteratively place objects within the scene, ensuring both spatial feasibility and semantic coherence.

Furthermore, we propose a two-stage training paradigm, called Scene-Aware Group Relative Policy Optimization, to enhance the capabilities of the off-the-shelf LLM in 3D scene synthesis. Specifically, we first fine-tune the model using scene reasoning data to empower it with initial layout generation abilities. However, the standard token-level objective is insufficient for 3D scene generation, which demands higher diversity and adaptability. To address this, we introduce an online policy optimization stage that enables the model to self-improve through interaction with a feedback-rich environment. We conduct extensive experiments to evaluate the effectiveness of our methodology. The results demonstrate significant improvements in both spatial validity and user alignment, demonstrating a potential way for future research in interpretable and reasoning-driven 3D scene synthesis.

Our key contributions can be summarized as follows:

- We explore chain-of-thought reasoning for 3D scene synthesis, demonstrating its potential to enhance LLMs' spatial generation capabilities through fine-grained reasoning.
- We propose a cascading data pipeline that decomposes complex indoor design tasks into manageable subtasks, along with a progressive training paradigm that encourages the model to predict intermediate steps toward the final layout.
- We conduct extensive experiments to validate the effectiveness of our method in synthesizing 3D indoor scenes.

2 Related Work

2.1 Chain-of-Thought Reasoning

Chain-of-Thought (CoT) reasoning is a prompting technique designed to enhance LLMs' performance on complex tasks by encouraging them to articulate intermediate logical steps before predict-

ing final answers (Wei et al., 2022). DDCoT (Zheng et al., 2023) proposes a dual-process prompting framework, dividing reasoning into reasoning and recognition modules. X-CoT (Chai et al., 2025) adopts the online CoT strategy to enhance the multilingual reasoning ability of the large language model. Visual CoT (Chen et al., 2024b) introduces visual-language iterative reasoning for multi-modal tasks. Additionally, recent work (Guo et al., 2025b) explores the potential of CoT reasoning in enhancing image generation. However, LLMs often struggle with 3DSS due to inadequate spatial reasoning, such as inappropriate object placements and spatial conflicts. In this paper, we propose integrating CoT reasoning into 3D scene synthesis, demonstrating its effectiveness in addressing the limitations of LLMs in spatial reasoning.

2.2 LLM-Driven 3D Reasoning

Recent advancements in LLMs have catalyzed significant progress in 3D scene understanding and reasoning, fostering the emergence of models capable of scene captioning, question answering, and embodied navigation. For example, 3D-LLM (Hong et al., 2023) proposes injecting 3D point clouds into LLMs, enabling a wide range of 3D-related tasks such as captioning, dense captioning, 3D question answering, task decomposition, visual grounding, and so on. LL3DA (Chen et al., 2024a) introduces a framework that directly processes point clouds for text-based instruction following and visual interaction. LiDAR-LLM (Yang et al., 2025) reformulates outdoor 3D scene cognition as a language modeling task, encompassing 3D captioning, 3D grounding, and 3D question answering. LLM-Grounder (Yang et al., 2024b) further introduces a zero-shot and open-vocabulary pipeline for 3D visual grounding. Our work fully harnesses the inherent strengths of large language models to tackle the challenges of 3D scene synthesis.

2.3 Text-driven Indoor Scene Synthesis

Text-driven indoor scene synthesis aims to generate furniture layouts based on user instructions, which can be categorized into two main directions. One involves using generative models to derive scene information through denoising. DiffuScene (Tang et al., 2024) employs a diffusion model that learns holistic scene configuration priors and generates 3D instance properties through iteratively denoising of unordered object attributes, such as semantics, locations, sizes, and geometry features. InstructScene (Lin & Mu, 2024) combines a semantic graph prior with a layout decoder to enhance controllability and fidelity of scene synthesis. These methods may be limited by the dataset when generating complex scenes or style-specific scenes, leading to a lack of diversity in the results.

Recently, researchers have utilized LLMs to synthesize 3D indoor scenes based on prompt engineering or multi-turn conversational interactions. HOLODECK (Yang et al., 2024f) leverages GPT-4 (Achiam et al., 2023) to generate spatial relational constraints and uses constraint-based optimization to position objects correctly, thereby creating diverse 3D environments from textual descriptions. I-Design (Çelen et al., 2024) applies LLM agents to transform user inputs into feasible scene graph designs, followed by an effective placement algorithm to determine optimal object locations within the scene. LLplace (Yang et al., 2024e) employs an open source LLM fine-tuned to allow efficient and credible room layout generation through multi-agent conversation, eliminating the need for spatial relationship priors or in-context exemplars. LayoutGPT (Feng et al., 2024) uses LLMs to generate layouts directly from text conditions through simple prompts, enhancing visual planning capabilities. These methods are either unable to avoid collisions between objects or are prone to failing the entire scene generation due to a small error in multi-LLM systems. Compared with these methods, our approach reflects the logical reasoning process of scene arrangement and effectively reduces conflicts in the generated results.

3 Scene-CoT

3.1 Problem Formulation

Following prior LLM-based works on 3DSS, our goal is to synthesize semantically coherent and geometrically valid 3D indoor scenes from natural language instructions T. Formally, the generated scene should contain $N \in \mathbb{R}$ objects, each represented as $\mathbf{o}_i = \{\alpha_i, c_i, s_i, r_i, p_i\}$, where $1 \leq i \leq N$, α_i denotes the object category, c_i describes the furniture's material and style, $s_i \in \mathbb{R}^3$

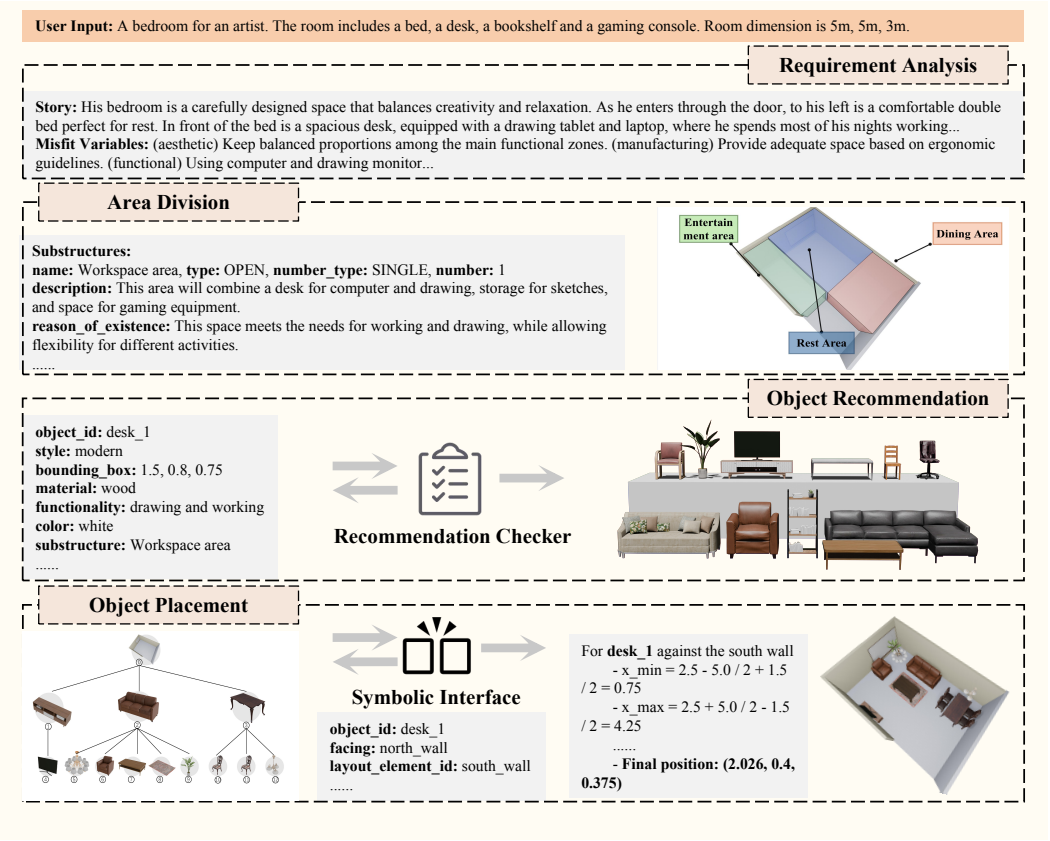

Figure 1: **Overview of data generation pipeline**. We introduce a novel method that decomposes complex 3D scene synthesis tasks into four structured subtasks: (1) user preference analysis, (2) functional zone partition, (3) object recommendation, and (4) object placement.

represents the object size, $r_i \in \mathbb{R}$ is the rotation angle around the z-axis, and $p_i \in \mathbb{R}^3$ indicates the 3D position of the object in the scene. The input description is highly flexible, allowing users to specify scene styles, object categories, and quantities, or provide ambiguous instructions such as "a minimalist bedroom". Furthermore, Scene-CoT requires the model to explicitly output a step-by-step reasoning trace before generating the scene. We formalize this process using dedicated XML-style tags: reasoning steps are marked within `<think>...</think>`, while the final scene representation is marked within `<answer>...</answer>`.

3.2 Dataset Construction

We propose a novel method to generate 3D indoor scenes via multi-agent conversational interactions. During this process, we record the chain of thought of LLMs and specific symbolic operations for object placement as part of the dataset. Our dataset includes not only the final scene graph but also the logical reasoning behind the scene generation, thereby offering significant benefits for training LLMs to achieve more reasonable and accurate indoor scenes. The dataset, not tied to specific categories or scenes, aims to enhance the model's spatial understanding and reasoning abilities in object arrangement.

Our method mainly involves four core agents: user preference analysis, functional zone partition, object recommendation, and object placement. These agents engage in iterative dialogues to determine the objects needed for the scene and their relative positions. Considering object collisions with the room and between objects themselves, we calculate viable placement positions within the room and determine object coordinates according to their hierarchical relationships.

User Preference Analysis. The agent first imagines the text input T as a story by envisioning the interactions between characters and the scene. The story incorporates specific objects that drive the plot and integrates common human knowledge, accurately reflecting scene information provided in the input. The agent then identifies the misfit variables for the scene as forces to improve and optimize the design, which represent the discrepancies between the scene and human habits or needs. Misfit variables are described from three perspectives, aesthetics, functionality and manufacturing. Aesthetic misfit variables refer to the visual and atmospheric needs, functional ones are determined by practical uses of the space, while manufacturing ones ensure that spatial arrangements comply with ergonomic principles. The final output consists of a story related to the instruction T and three categories of misfit variables.

Functional Zone Partition. According to the preferences and scenario descriptions provided, the entire scene is partitioned into distinct rectangular areas. Each area represents a specific physical space suitable for object placement and satisfies a set of misfit variables. The agent analyzes and groups these variables based on functional similarities and spatial relationships. For instance, bedside tables should be grouped with beds. Depending on each group of requirements, the agent will divide one or more zones from the room, each described in terms of name, type, quantity, corresponding description, and rationale for the division.

Object Recommendation. This agent recommends multiple objects vital for enhancing functionality and aesthetic appeal, while each zone may correspond to more than one object. In addition to the explicitly mentioned objects, the agent actively identifies implicit needs and provides some additional items. We have developed a verification agent to ensure that the results encompass all the objects specified in T. The properties of objects include name, style, size, material, color, etc.

Object Placement. The agent can iteratively place objects and determine their relative positions, including ground contact, orientation, spatial relationships with the room and other objects. For each new item, it incrementally reasons about its spatial position in the room based on the existing items and their placements, avoiding any conflicts between the new item and the current scene.

Following object placement, collision detection is performed to identify any overlaps between objects or instances where objects exceed room boundaries. Upon detecting a conflict, the conflicting objects are repositioned. If a suitable arrangement cannot be achieved, the object is removed.

When locating the exact positions, we consider dimensional constraints based on the spatial relationships between objects. For those near walls or in the middle of a room, possible coordinate ranges are calculated on the basis of room dimensions and object sizes. After an initial position is established, collision detection is conducted by comparing the boundaries of the objects to determine if there is any overlap, and the positions are adjusted accordingly. For objects with hierarchical relationships, the position of the child object is adjusted based on the parent object to ensure logical spatial arrangement:

$$(x_c, y_c, z_c) = (x_p, y_p, z_p) + \left(\frac{w_p - w_c}{2}, \frac{d_p - d_c}{2}, \frac{h_p - h_c}{2} \right) + (\alpha, \beta, \gamma) \tag{1}$$

where (x_c, y_c, z_c) represents the final position of the child object, (x_p, y_p, z_p) means the position of the parent object. (w_p, d_p, h_p) , (w_c, d_c, h_c) denote the dimensions of the parent object and the child object. (α, β, γ) means adjustment coefficients used to adjust the position of the child object based on specific constraints.

3.3 Scene-aware Group Relative Policy Optimization

Despite their strong performance across various tasks, LLMs often generate 3D scenes with geometrically invalid placements due to insufficient spatial knowledge. To enhance LLMs' spatial reasoning capabilities, we propose Scene-aware Group Relative Policy Optimization (as shown in fig. 2), a novel reinforcement training framework consisting of two key stages: Spatial Prior Knowledge Injection and Self-improving with Online Policy Optimization. Specifically, we fine-tune the base LLM using 3DSS-oriented reasoning data (introduced in Section 3.2), empowering the model to generate initial scene layouts. Furthermore, we introduce Self-Improvement with Online Policy Optimization, a reinforcement-learning-based mechanism that allows LLMs to enhance their reasoning capabilities for 3DSS through interaction with the designed reward environment. This reinforce-

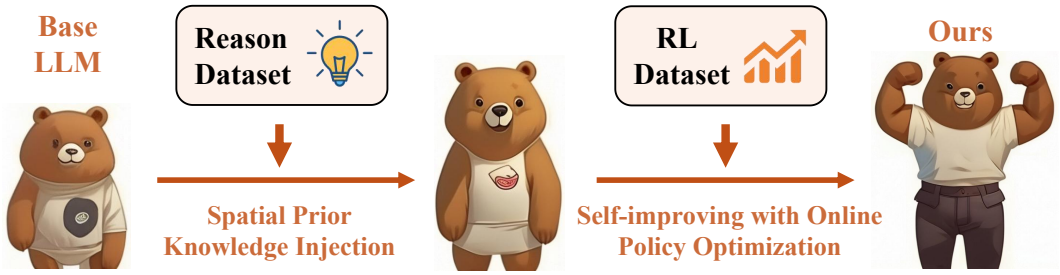

Figure 2: **Overview of Scene-aware Group Relative Policy Optimization.** We present a progressive training framework to unlock the reasoning capabilities of LLMs in 3D scene synthesis.

ment learning mechanism enhances the model's ability to generate spatially valid and aesthetically coherent 3D scenes.

Stage I: Spatial Prior Knowledge Injection. To endow the model with foundational reasoning capabilities before generating initial scene layouts, we perform end-to-end supervised training on a base LLM, utilizing Chain-of-Thought (CoT) data tailored for 3DSS. During this stage, the LLM receives a user instruction T as input and autoregressively generates both the intermediate reasoning steps and the final layout solution. The primary objective is to optimize trainable parameters θ such that the likelihood of producing the target response sequence $S = \{s_i \mid i \in [1 : L]\}$ given the instruction T is maximized:

$$\theta^* = \arg\max_{\theta} P(S \mid T; \theta). \tag{2}$$

In practice, we train the model using a token-level cross-entropy loss $\mathcal{L}_s$, defined as:

$$\mathcal{L}_s = -\sum_{i=1}^{L} \log P(s_i \mid T, S_{<i}, \theta), \tag{3}$$

where $S_{<i}$ denotes all previously generated tokens before the current prediction s_i.

Stage II: Self-improving with Online Policy Optimization. While supervised fine-tuning equips the model with preliminary layout reasoning abilities, the token-level objective often leads to degenerate solutions that lack the diversity and adaptability required for spatial reasoning. Specifically, models trained in this manner tend to produce rigid outputs, which are insufficient for capturing the rich variations in object arrangements and user preferences. Therefore, we argue that LLMs should not only infer a layout through initial reasoning but also iteratively refine their outputs by interacting with the environment.

Inspired by DeepSeek-R1 (Guo et al., 2025a), we treat 3D scene synthesis as an **online policy optimization** task, in which the model generates multiple candidate responses and evaluates their quality through interaction with an environment. To implement this, we curate a policy optimization dataset containing 1,248 scene generation instructions covering diverse room types. Unlike structured reasoning tasks such as mathematical problem-solving, where correctness can be objectively measured, 3D scene synthesis is inherently open-ended. For example, a single instruction may admit multiple valid and semantically coherent object arrangements. As a result, relying on fixed labels or accuracy-oriented reward is insufficient. To this end, we propose a new reward environment for 3DSS that provides *spatial* and *semantic* feedback signals through LLM-environment interaction.

Given a textual instruction T, our policy model π_θ first generate M candidate responses $\{s_1, s_2, \ldots, s_M\}$ through multiple rollouts, where each s_m is decoded by the LLM to extract the corresponding layout. These layouts are then evaluated within the environment $\mathcal{E}$ from two key aspects: spatial feasibility and semantic consistency. Specifically, we extract object dimensions and positions to construct a scene graph that incorporates layout boundaries (e.g., walls, ceilings, and floors) for collision detection and spatial constraint validation. To this end, we adopt two metrics: *Object Overlap Rate (OOR)* and *Out-of-Boundary Score (OOB)*, which quantify the probability of object-object and object-boundary collisions, respectively. The spatial reward is defined as:

$$R_{\text{spa}} = (1 - OOR) + (1 - OOB), \tag{4}$$

where higher rewards are assigned to layouts with lower object collision and boundary violation rates. However, relying solely on spatial validity is insufficient as it fails to detect semantic or commonsense errors. For example, a floating cup that should be placed on a table rather than on the floor. To address this, we further introduce the **semantic consistency reward** R_{sem}, which encourages generated layouts to align with human preferences and aesthetic principles. Specifically, we retrieve the most relevant object for each layout node from the database. Then, the scene is rendered into 2D images using an automated renderer from four complementary camera views to reduce occlusion and viewpoint bias. Following the evaluation criteria proposed by I-Design (Çelen et al., 2024), we employ GPT-4 (Achiam et al., 2023) as the 3D content evaluator and assess each scene across five key dimensions: (1) functionality and plausibility, (2) alignment with user preferences T, (3) layout and furniture arrangement, (4) color scheme and material appropriateness, and (5) overall aesthetics and atmospheric consistency.

Finally, we normalize rewards using their mean and standard deviation to compute the advantage value A_m for each candidate response s_m:

$$A_m = \frac{r_m - \text{mean}\{r_1, r_2, \ldots, r_M\}}{\text{std}\{r_1, r_2, \ldots, r_M\}}, \quad r_m = R_{\text{spa}}(s_m) + R_{\text{sem}}(s_m). \tag{5}$$

Here, A_m represents the relative advantage of candidate s_m compared to other responses in the same group. The reward r_m combines the spatial constraint reward R_{spa} and semantic alignment reward R_{sem}. During training, we follow GRPO and optimize the policy to generate responses with higher intra-group advantages.

4 Experiments

4.1 Data, Metrics, and Implementation Details

Data. While Objaverse (Deitke et al., 2023) provides a vast collection of 3D assets, its objects often suffer from misaligned orientations or incorrect scaling, making precise placement within the scene difficult. To address this, we construct an asset library containing 2,032 indoor objects, each with rich semantic descriptions and meticulously standardized dimensions and orientations. Benefiting from our data generation pipeline, we further curate two dedicated datasets: the *Reason Dataset* and the *RL Dataset*. The Reason Dataset includes 443 annotated samples, each consisting of a natural language instruction, a complete reasoning path, and the corresponding final scene layout. This dataset is used to warm up the base model and activate its structured generation capabilities. The RL Dataset contains 1,248 entries, where only user preferences are provided. This dataset is designed to improve scene diversity and spatial plausibility through online policy optimization. To quantitatively assess the model's performance on 3DSS, we curate an evaluation using the same data generation pipeline. Each scene in the test set is manually inspected and refined to ensure spatial accuracy and semantic coherence. The final test set contains eight room types, including bedrooms, living rooms, dining rooms, studies, and more, with approximately six samples per category.

Metrics. We evaluate the proposed method in terms of visual quality, spatial validity, and semantic coherence with respect to the provided language instructions. Specifically, we adopt Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) to assess the visual realism of generated scenes. To measure geometric plausibility, we employ two additional metrics: Out-of-Boundary Rate (OOB) and Object Overlap Rate (OOR). For each scene, OOB is calculated as the ratio of objects whose bounding boxes fall outside the defined room boundaries to the total number of objects. OOR is computed by checking each object pair for intersecting bounding boxes. If any overlap is detected, the volume of the overlapping region is divided by the smaller volume of the two objects. The final OOR is obtained by averaging these ratios across all overlapping pairs in the scene. Moreover, we employ GPT-4o (Hurst et al., 2024) as an evaluator to assess 3D scenes based on rendered multi-perspective images. The GPT-4o (Hurst et al., 2024) score is reported as the average across multiple dimensions, including realism, functionality, etc.

Implementation Details. For scene reasoning data construction, we utilize GPT-4o-2024-08-06 (Hurst et al., 2024) as the base LLM, with temperature 0.8 and top-p 0.9. To facilitate retrieval, we encode object category, color, and material information into textual embeddings. For objects in Objaverse (Deitke et al., 2023), we extract embeddings using OpenShape (Liu et al., 2023), while

Table 1: **Performance comparison on 3D scene synthesis.** We conduct a comprehensive evaluation on 3D scene synthesis. Our approach exhibits notable effectiveness and outperforms all baselines across multiple metrics, achieving superior spatial validity (OOB, OOR), visual quality (FID, KID), and alignment with human intent (GPT-4o (Hurst et al., 2024) Criteria).

Methods	$Params.$	FID↓	KID↓	OOB↓	OOR↓	GPT-4o Criteria↑
Non-reasoning Model						
LayoutGPT (Feng et al., 2024)	-	106.65	0.0232	0.1926	0.2745	5.20
Deepseek-V3 (Liu et al., 2024)	660B	100.28	0.0297	0.1384	0.2711	5.05
OpenAI GPT-4.1 (Achiam et al., 2023)	-	96.56	0.0257	0.0891	0.1691	5.10
Reasoning Model						
OpenAI o1 (Jaech et al., 2024)	-	97.94	0.0337	0.1078	0.1811	5.05
Deepseek-R1 (Guo et al., 2025a)	671B	92.51	0.0213	0.1317	0.2222	5.22
OpenAI o3 (Jaech et al., 2024)	-	90.24	0.0171	0.1092	0.2166	5.35
Ours	7B	**85.31** (±**2.23**)	**0.0035** (±**0.0012**)	**0.0320** (±**0.0069**)	**0.1759** (±**0.0415**)	**5.57** (±**0.0314**)

for both the local asset library and user-provided descriptions, we use CLIP-ViT-bigG-14-laion2B-39B-b160k (Radford et al., 2021). The most semantically relevant asset is selected by computing the cosine similarity between these embedding vectors. In addition, we perform full-parameter supervised fine-tuning of Qwen-Instruct-7B (Bai et al., 2023) on eight NVIDIA A100 (80GB) GPUs based on the rllm codebase(Luo et al., 2025). The model is trained for 3 epochs using a sequence length of 24,576 tokens, a batch size of 8, and a learning rate of 1×10^{-5}. In the second stage, we perform policy optimization with a mini-batch size of 16. To ensure training stability and prevent over-optimization, we incorporate KL regularization with a coefficient of $\beta = 0.001$, along with dynamic clipping ratios ranging from 0.2 to 0.28. The model is trained for one epoch, where each prompt is constrained to a maximum of 4,096 tokens and the corresponding response to 16,384 tokens. For each prompt, four rollout samples are generated using a temperature of 0.7.

4.2 Evaluation on 3D Scene Synthesis

As reported in table 1, we conduct a comprehensive comparison between our method and a suite of strong baselines, including both non-reasoning models (e.g., LayoutGPT (Feng et al., 2024), Deepseek-V3 (Liu et al., 2024), and GPT-4.1 (Achiam et al., 2023)) and state-of-the-art reasoning models (e.g., Deepseek-R1 (Guo et al., 2025a), OpenAI o1/o3 (Jaech et al., 2024)). All reported metrics are the means of five independent runs, and the corresponding variances of our model are also shown.

Specifically, compared to LayoutGPT(Feng et al., 2024), our approach reduces Out-of-Boundary Rate (OOB) by 83.4% (from 0.1926 to 0.0320) and Object Overlap Rate (OOR) by 35.9% (from 0.2745 to 0.1759), while achieving 20.0% decrease in FID (from 106.65 to 85.31) and 84.9% decrease in KID (from 0.0232 to 0.0035). Against the strong OpenAI o3 (Jaech et al., 2024), our model further improves spatial validity by 70.7% in OOB (from 0.1092 to 0.0320) and 18.8% in OOR (from 0.2166 to 0.1759), with the GPT-4o (Hurst et al., 2024) score rising from 5.35 to 5.57 (+4.1%). Notably, these gains are achieved using a 7B-parameter model, outperforming Deepseek-V3 (Liu et al., 2024) (660B parameters) and black-box LLMs like GPT-4.1 (Achiam et al., 2023).

4.3 Ablation Studies

As shown in table 2, we conduct a comprehensive ablation study on the proposed scene-aware group relative policy optimization. When neither spatial prior knowledge injection nor online policy optimization is applied, the model performs poorly, with an FID of 110.75, a KID of 0.0290, an OOB of 0.0617, an OOR of 0.3586, and a GPT-4o (Hurst et al., 2024) Criteria score of 4.74. Additionally, without scene-prior knowledge injection, applying online policy optimization alone results in performance degradation, as the model lacks adequate scene understanding and struggles to improve effectively. In contrast, our scene-aware group relative policy optimization achieves significant improvements, yielding an FID of 85.31, a KID of 0.0035, an OOB of 0.0320, an OOR of 0.1759, and a GPT-4o (Hurst et al., 2024) score of 5.57.

Table 2: **Effectiveness of scene-aware group relative policy optimization.**

Spatial Prior Knowledge Injection	Self-improving with Online Policy Optimization	FID↓	KID↓	OOB↓	OOR↓	GPT-4 Criteria↑
×	×	110.75	0.0290	0.0617	0.3586	4.74
×	✓	119.45	0.0285	0.0527	0.3345	4.58
✓	✓	**85.31**	**0.0035**	**0.0320**	**0.1759**	**5.57**

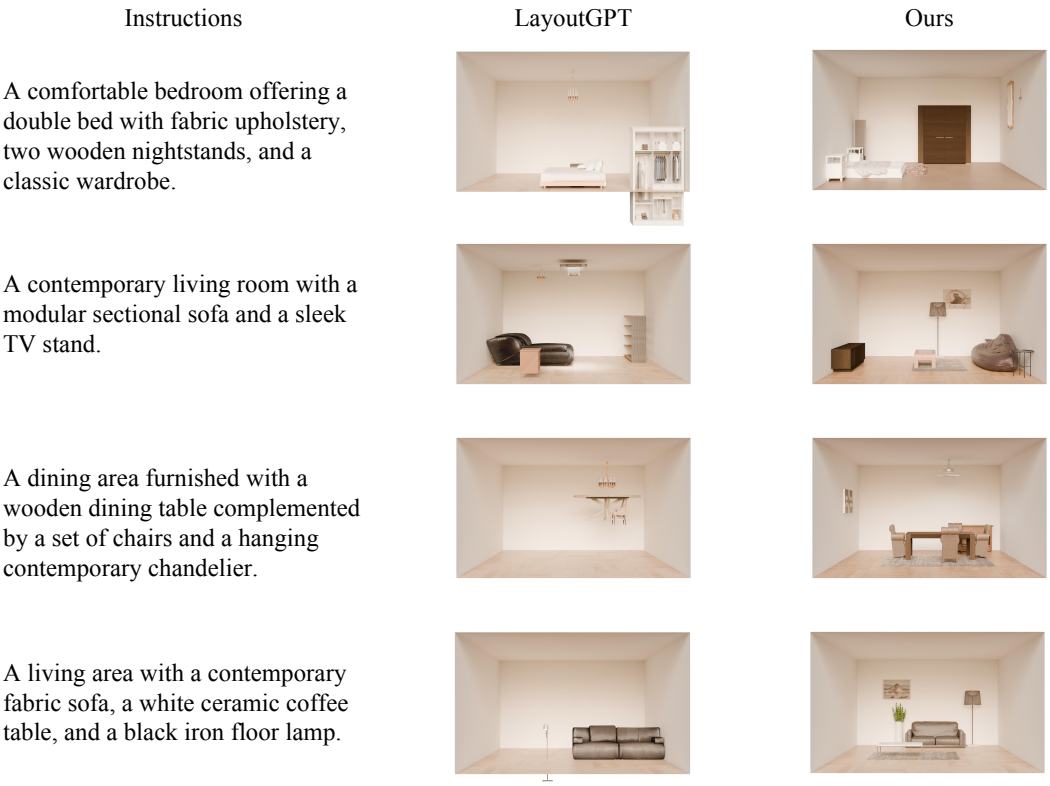

Figure 3: **Qualitative Results.** We provide qualitative comparisons between LayoutGPT (Feng et al., 2024) and our method on 3DSS. The results demonstrate that our approach generates scenes with better semantic alignment to user instructions and improved spatial structure.

4.4 Qualitative Results

We present qualitative results in fig. 3 to demonstrate the effectiveness of our model in generating physically plausible 3D scenes aligned with user instructions. Notably, our approach accurately interprets object relationships and faithfully incorporates specified attributes.

5 Conclusion

In this work, we have introduced Scene-CoT, a novel chain-of-thought framework tailored for 3D Scene Synthesis (3DSS) that bridges the gap between large language models' generative power and the spatial reasoning demands of coherent scene generation. By automatically generating structured reasoning data through a cascading pipeline of hierarchical agents, Scene-CoT decomposes complex layout tasks into manageable subtasks and supplies explicit reasoning traces. This structured guidance enables large language models to plan object placement, respect spatial constraints, and resolve conflicts that typically plague end-to-end text-driven 3D synthesis. Our approach has both strong initial scene construction capabilities and the flexibility to improve coherence and plausibility in 3D Scene Synthesis adaptively.

6 Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation is involved. All datasets used are sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information is used, and no experiments are conducted that may raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

7 Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code and datasets have been made publicly available in an anonymous repository to facilitate replication and verification. We place the code in the repository at `https://github.com/DeepScene-Generation/scene-cot`, and the Scene-CoT dataset is available at `https://huggingface.co/datasets/DeepSceneGen/scene-cot`. The experimental setup, including training steps, model configurations, and hardware details, is described in section 4.1. We believe these measures will enable other researchers to reproduce our work and further advance the field.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 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.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23550–23558, 2025.

Minglin Chen, Longguang Wang, Sheng Ao, Ye Zhang, Kai Xu, and Yulan Guo. Layout2scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. *arXiv preprint arXiv:2501.02519*, 2025.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26428–26438, 2024a.

Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Scenedreamer: Unbounded 3d scene generation from 2d image collections. *IEEE transactions on pattern analysis and machine intelligence*, 45 (12):15562–15576, 2023.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1254–1262, 2024b.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5982–5994. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/27c546ab1e4f1d7d638e6a8dfbad9a07-Paper-Conference.pdf.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16352–16361, 2021.

Weihua Du, Qiushi Lyu, Jiaming Shan, Zhenting Qi, Hongxin Zhang, Sunli Chen, Andi Peng, Tianmin Shu, Kwonjoon Lee, Behzad Dariush, and Chuang Gan. Constrained human-ai cooperation: An inclusive embodied social intelligence challenge. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 44526–44553. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4eb8e997fc91086225b7484cf8eac341-Paper-Datasets_and_Benchmarks_Track.pdf.

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*, 36, 2024.

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, 2021.

Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21295–21304, 2024.

Qiaozi Gao, Govind Thattai, Suhaila Shakiah, Xiaofeng Gao, Shreyas Pansare, Vasu Sharma, Gaurav Sukhatme, Hangjie Shi, Bofei Yang, Desheng Zhang, Lucy Hu, Karthika Arumugam, Shui Hu, Matthew Wen, Dinakar Guthy, Shunan Chung, Rohan Khanna, Osman Ipek, Leslie Ball, Kate Bland, Heather Rocker, Michael Johnston, Reza Ghanadan, Dilek Hakkani-Tur, and Prem Natarajan. Alexa arena: A user-centric interactive platform for embodied ai. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19170–19194. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3d0758f0b95e19abc68c1c8070d36510-Paper-Datasets_and_Benchmarks.pdf.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025b.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 187–203, 2018.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Inwoo Hwang, Hyeonwoo Kim, and Young Min Kim. Text2scene: Text-driven indoor scene stylization with part-aware details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1890–1899, 2023.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Chenguo Lin and Yadong Mu. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. In *International Conference on Learning Representations (ICLR)*, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. `https://pretty-radio-b75.notion.site/...`, 2025. Notion Blog.

Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoğlu, and Theo Gevers. Sceneteller: Language-to-3d scene generation. In *European Conference on Computer Vision*, pp. 362–378. Springer, 2024.

Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2856–2865, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Zifan Shi, Yujun Shen, Jiapeng Zhu, Dit-Yan Yeung, and Qifeng Chen. 3d-aware indoor scene synthesis with depth priors. In *European Conference on Computer Vision*, pp. 406–422. Springer, 2022.

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*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 107326–107349. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/c2166d01fe4bcd694aba89f608737678-Paper-Conference.pdf`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A. Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey R. Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 76289–76318. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8bc74514d554a90c996576f6c373f5f3-Paper-Conference.pdf.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7694–7701. IEEE, 2024b.

Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, et al. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9247–9255, 2025.

Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. *Advances in Neural Information Processing Systems*, 37:82060–82084, 2024c.

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*, pp. 16262–16272, 2024d.

Yixuan Yang, Junru Lu, Zixiang Zhao, Zhen Luo, James J. Q. Yu, Victor Sanchez, and Feng Zheng. Llplace: The 3d indoor scene layout generation and editing via large language model, 2024e. URL https://arxiv.org/abs/2406.03866.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16227–16237, 2024f.

Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12965–12976, June 2023.

Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. Towards text-guided 3d scene composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6829–6838, June 2024.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

A Appendix

A.1 Dataset Analysis

We conduct a statistical analysis of the dataset. As shown in fig. 4, prominent words include "functionality", "user", "design", "color", "material", and "object". These words highlight key considerations and elements in the scene generation process, such as ensuring functionality, meeting user requirements, and focusing on design aspects like color and material.

Figure 4: **Word cloud generated based on the content of the dataset.**

We also draw a bar chart to illustrate the frequency of different furniture items in our dataset. As reported in fig. 5, the most frequent items are ”rug” and ”chair”, with ”rug” appearing 218 times and ”chair” appearing 170 times. Other items like ”floor_lamp”, ”mirror”, ”side_table”, and ”plant” have relatively lower frequencies, ranging from 53 to 77.

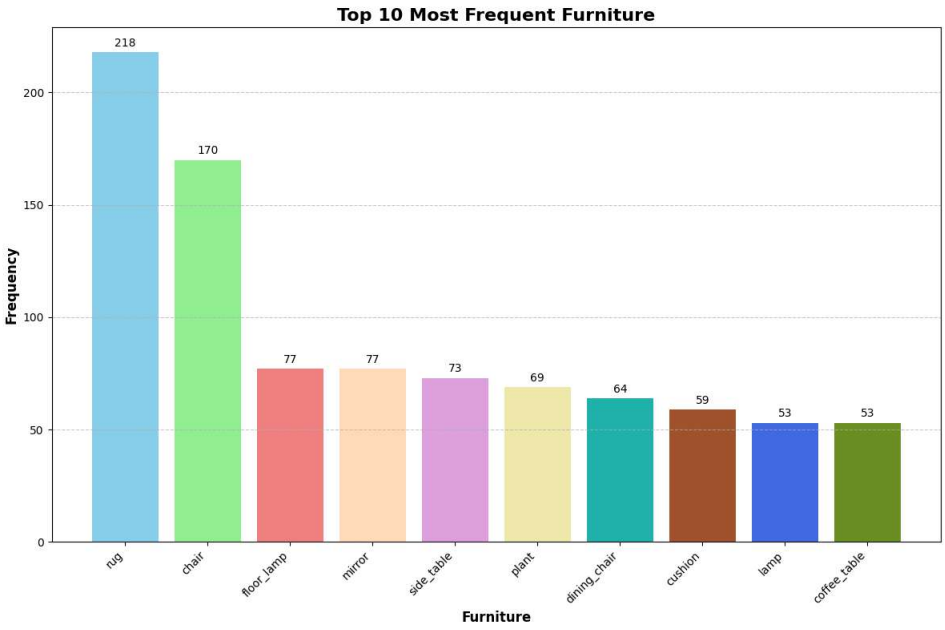

Figure 5: **Number of object occurrences in Scene-CoT.**

A.2 Quantitative Comparison on 3D-FUTURE Dataset

To demonstrate the consistency of our dataset and approach across benchmarks, we compare our method with existing approaches on 3D-FUTURE (Fu et al., 2021) dataset. As shown in table 3, our model outperforms LayoutGPT (Feng et al., 2024) across all metrics on 3D-FUTURE (Fu et al., 2021), proving that our Scene-CoT dataset and methods can generalize to standardized benchmarks.

Table 3: **Quantitative comparison of different methods on 3D-FUTURE (Fu et al., 2021) dataset.** ↓ indicates lower is better, ↑ indicates higher is better.

Method	FID↓	KID↓	OOB↓	OOR↓	GPT-4o Criteria↑
LayoutGPT (Feng et al., 2024)	225.20	0.1675	0.2017	0.2783	5.12
Ours	**202.42**	**0.1283**	**0.0283**	**0.2091**	**5.47**

A.3 The Impact of Reasoning Steps on Layout Quality

To validate the contribution of the reasoning steps to the final layout quality, we additionally train the model with data that excludes reasoning steps (containing only structured outputs). Without reasoning steps, the model generates scenes with no more than three objects. Consequently, the objects never collide and the OOR is zero. As reported in table 4, it is evident that incorporating reasoning steps allows the model to significantly improve layout quality by enhancing alignment with visual aesthetic criteria.

Table 4: **Comparison of layout quality with and without reasoning steps.** ↓ indicates lower is better, ↑ indicates higher is better.

Method	FID↓	KID↓	OOB↓	OOR↓	GPT-4o Criteria↑
Ours (without reasoning steps)	161.71	0.1136	0.0800	0.0000	4.33
Ours (with reasoning steps)	**85.31**	**0.0035**	**0.0320**	0.1759	**5.57**

A.4 User Study

We conduct a user preference evaluation by anonymously packaging results generated by LayoutGPT (Feng et al., 2024), HOLODECK (Yang et al., 2024f), and our model, asking users to choose the one they perceive as most effective. Out of 142 responses received, 137 users (96.5%) find the scenes generated by our model more realistic and aesthetically pleasing, 5 users (3.5%) prefer HOLODECK, and no users select LayoutGPT. The results indicate that scenes generated by our model align with real human preferences and design standards in practice.

A.5 Complete Reasoning Process of the Model

A.5.1 Chain-of-Thought Demonstration

We present a comprehensive demonstration of our trained model's workflow, from user input to the generation of the final scene graph. This includes detailed insights into the model's reasoning processes at each step and the iterative placement of objects, as shown in fig. 6 and fig. 7. Our input text is "A functional home office space equipped with a wooden desk, a leather upholstered swivel chair, and a black metal filing cabinet".

User Preference Analysis: The user envisions a functional home office that incorporates specific furniture pieces such as a wooden desk, a leather upholstered swivel chair, and a black metal filing cabinet. The room measures 5.0 meters by 5.0 meters with a height of 3.0 meters, providing ample space for these items. The primary focus is on creating a workspace that is both efficient and aesthetically pleasing, with additional elements like a rug, wall art, and a plant suggested to enhance the room's warmth and comfort. The user also expressed interest in adding a desk lamp and a monitor to support work activities.

Functional Zone Partition: The room is divided into several functional areas based on the user's requirements. The central area is designated for the desk and chair setup, forming the core of the home office. Adjacent to this is the filing cabinet area, which provides storage while maintaining accessibility. The east wall is reserved for wall art and a plant, adding aesthetic value and warmth to the space. Additional elements like a rug and a desk lamp are strategically placed to enhance the overall functionality and aesthetic of the room.

Object Recommendation: For the central workspace, a classic wooden desk and a leather upholstered swivel chair are recommended to create a cohesive and functional setup. A modern black metal filing cabinet is suggested for storing office supplies. To enhance the ambiance, contemporary wall art and a plant are proposed for the east wall. A minimalist grey rug is recommended to add texture and warmth under the desk area. A modern desk lamp and a monitor are also included to improve lighting and support work activities.

Object Placement: The wooden desk, a central element of the home office, is placed against the north wall, facing the south wall. This placement maximizes space usage and creates a functional workspace, aligning with the user's preference for a classic style. The desk's dimensions are 1.8 meters by 0.9 meters by 0.75 meters, fitting comfortably against the wall and leaving ample space for other furniture.

The leather upholstered swivel chair is positioned directly in front of the desk, facing the north wall. This placement ensures ergonomic seating and complements the desk's orientation, maintaining balance and proportion in the room. The chair's dimensions are 0.7 meters by 0.7 meters by 1.2 meters, allowing it to fit seamlessly in front of the desk without spatial conflicts.

The black metal filing cabinet is placed against the west wall, facing the east wall. This location ensures accessibility from the desk without obstructing the chair's movement, maintaining an open and functional workspace. The filing cabinet's dimensions are 0.8 meters by 0.6 meters by 1.8 meters, fitting well along the wall and complementing the desk and chair arrangement.

The wall art is centrally placed on the east wall, facing the west wall. This placement avoids spatial conflicts with existing furniture and adds visual interest without cluttering the room. The wall art's dimensions are 1.2 meters by 0.05 meters by 0.8 meters, making it suitable for wall display.

The plant is placed on the floor against the east wall, facing the west wall. This placement enhances the room's aesthetic without obstructing functional areas. The plant's dimensions are 0.5 meters by 0.5 meters by 1.5 meters, ensuring it fits comfortably against the wall.

The rug is centrally placed under the desk, oriented parallel to the room's dimensions. This placement enhances the aesthetic appeal and provides a cohesive look to the workspace. The rug's dimensions are 2.0 meters by 1.5 meters by 0.01 meters, fitting comfortably under the desk area without overlapping other objects.

A conflict was identified with the desk lamp initially intended to be placed on the desk. The width of the desk was too small to accommodate the lamp without causing spatial issues. To resolve this, the desk lamp was removed, as it was deemed less critical compared to the other elements specified by the user, ensuring the room remains functional and aesthetically pleasing.

Figure 6: **Chain-of-Thought for user preference analysis, functional zone partition, object recommendation and part of object placement.**

For rug_1
- parent object: swivel_chair_1
  - calculation_steps:
    1. reason: Calculate size constraint for 'under' relation
       - calculation:
         - rug_1 size: 2.0x1.5x0.01
         - Cluster size (under): 0.0 (non-directional)
       - conclusion: No directional constraint applied
    2. reason: Calculate possible positions based on 'middle of the room' constraint
       - calculation:
         - x_min = x_max = 2.5
         - y_min = y_max = 2.5
         - z_min = z_max = 0.005
       - conclusion: Possible position: (2.5, 2.5, 2.5, 2.5, 0.005, 0.005)
    3. reason: Adjust for 'under desk_1' constraint
       - calculation:
         - x_min = max(2.5, 2.0426 - 1.8/2 - 2.0/2) = 1.0
         - y_min = max(2.5, 4.55 - 0.9/2 - 1.5/2) = 3.25
       - conclusion: Final position: x: 1.0, y: 3.25, z: 0.005
    4. reason: Collision check with swivel_chair_1
       - calculation:
         - Overlap detection: $1.0 \leq 1.9482 \leq 4.0 \rightarrow$ No collision
       - conclusion: No collision detected
    5. reason: Final position calculation
       - calculation:
         - Selected position within overlap: x=1.8337, y=3.9774, z=0.005
       - conclusion: Final position: x: 1.8337, y: 3.9774, z: 0.005

For filing_cabinet_1
- calculation_steps:
  1. reason: Calculate rotation difference with other objects
     - calculation:
       - No other objects in proximity
     - conclusion: No rotation difference applicable
  2. reason: Calculate size constraint for 'west_wall' relation
     - calculation:
       - filing_cabinet_1 size: 0.8x0.6x1.8
       - Cluster size (west_wall): 0.0 (non-directional)
     - conclusion: No directional constraint applied
  3. reason: Calculate possible positions based on 'west_wall' constraint
     - calculation:
       - filing_cabinet_1 size: length=0.8, width=0.6, height=1.8
       - Room size: 5.0x5.0x3.0
       - x_min = 0 + 0.6/2 = 0.3
       - x_max = 0 + 0.6/2 = 0.3
       - y_min = 2.5 - 5.0/2 + 0.8/2 = 0.4
       - y_max = 2.5 + 5.0/2 - 0.8/2 = 4.6
       - z_min = z_max = 1.8/2 = 0.9
     - conclusion: Possible position: (0.3, 0.3, 0.4, 4.6, 0.9, 0.9)
  4. reason: Adjust boundaries for valid placement
     - calculation:
       - Adjusted cluster constraint: x(0.3-0.3), y(0.4-4.6)
       - Final coordinates: x=0.3, y=1.5713, z=0.9
     - conclusion: Final position: x: 0.3, y: 1.5713, z: 0.9
  5. reason: Collision check with other objects
     - calculation:
       - No other objects in proximity
     - conclusion: No collision detected
  6. reason: Final position calculation
     - calculation:
       - Selected position within overlap: x=0.3, y=1.5713, z=0.9
     - conclusion: Final position: x: 0.3, y: 1.5713, z: 0.9

Figure 7: **Part of Chain-of-Thought for determining the specific positions of objects in object placement.**

A.5.2 VISUALIZATION

By integrating with the local asset library, we render the generated scene graph to obtain images from four complementary camera views, which are shown in fig. 8.

Figure 8: **Visualization of the generated scene graph of our model.**

A.6 MORE VISUAL RESULTS

We present additional visualizations to demonstrate that our method can generate diverse and complex scenes that align with human aesthetics.

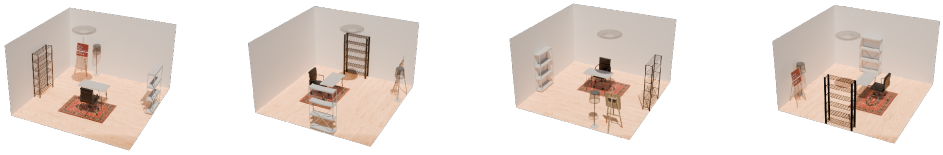

Figure 9: **A bright artist's studio with an easel, a canvas storage rack, and a drafting table.**

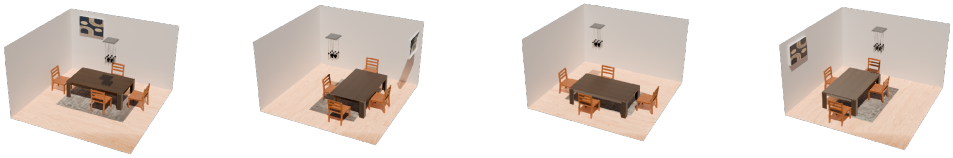

Figure 10: **A mid-century modern dining room with a table, a set of chairs, and a pendant light.**

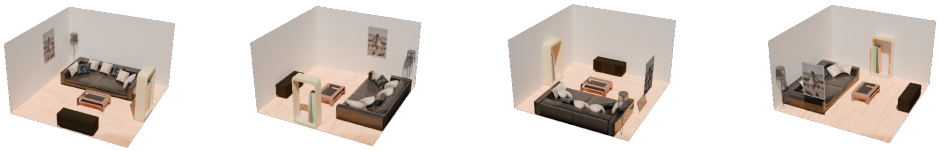

Figure 11: **A minimalist living room with a sectional sofa, a coffee table, and a floor lamp for ambient lighting.**

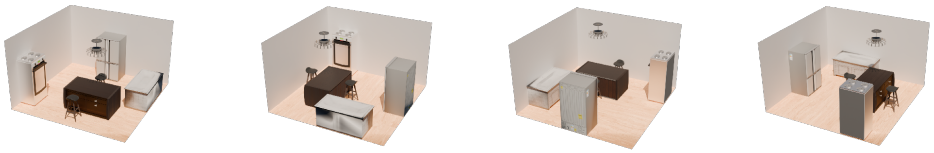

Figure 12: **A modern kitchen featuring a stainless steel refrigerator, an oven, and a ceramic sink counter.**

Figure 13: **A playful children's playroom with a toy storage unit, a child-sized table, and a set of colorful chairs.**

Figure 14: **A luxurious bathroom with a ceramic white bathtub, a marble sink, and a metal towel rack.**