Text-to-Scene with Large Reasoning Models

Frédéric Berdoz, Luca A. Lanzendörfer, Nick Tuninga, Roger Wattenhofer

ETH Zurich
{fberdoz, lanzendoerfer, ntuninga, wattenhofer}@ethz.ch

Abstract

Prompt-driven scene synthesis allows users to generate complete 3D environments from textual descriptions. Current text-to-scene methods often struggle with complex geometries and object transformations, and tend to show weak adherence to complex instructions. We address these limitations by introducing Reason-3D, a text-to-scene model powered by large reasoning models (LRMs). Reason-3D integrates object retrieval using captions covering physical, functional, and contextual attributes. Reason-3D then places the selected objects based on implicit and explicit layout constraints, and refines their positions with collision-aware spatial reasoning. Evaluated on instructions ranging from simple to complex indoor configurations, Reason-3D significantly outperforms previous methods in human-rated visual fidelity, adherence to constraints, and asset retrieval quality. Beyond its contribution to the field of text-to-scene generation, our work showcases the advanced spatial reasoning abilities of modern LRMs. Additionally, we release the codebase to further the research in object retrieval and placement with LRMs.

Code — https://github.com/ETH-DISCO/reason-3d
Extended version — https://arxiv.org/abs/2509.26091

1 Introduction

The demand for customizable 3D scenes is rapidly growing across a wide range of fields, including interior design (Çelen et al. 2024; Fu et al. 2021; Dai et al. 2017), video game development (Raistrick et al. 2024; Yang et al. 2024c; Unity Technologies 2025), autonomous driving (Gao et al. 2024; Mao et al. 2025), robotics (Lee, Park, and Kim 2025; Bu et al. 2024; Po and Wetzstein 2024; Cen et al. 2024), embodied AI (Yang et al. 2024a; Lee, Park, and Kim 2025; Deitke et al. 2022; Yang et al. 2024c), and more (Wen et al. 2025). Generating plausible and physically-feasible arrangements of objects from high- to low-level descriptions is therefore at the core of text-to-scene models. Traditionally, scene synthesis has been addressed by specialized models trained on annotated layout datasets or by engineered pipelines that encode spatial priors (Paschalidou et al. 2021; Tang et al. 2024). But these methods tend to constrain the ability of the model in some way, be it context confinement, visual fidelity, physical plausibility, or object transformation freedom. Recent advances in training strategies, such as Group Relative Policy Optimization (GRPO), have led to the emergence of Large Reasoning Models (LRMs) (Shao et al. 2024). Unlike standard LLMs, LRMs are trained to perform multi-step reasoning by leveraging test-time compute through long reasoning traces, enabling them to reason about geometry, context, physical affordances, and object-function relationships (Xu et al. 2025). This shift enables context-free scene synthesis, where systems can generate and arrange 3D environments directly from natural language, without relying on predefined templates or training distributions. In this work, we present Reason-3D, a modular and highly flexible scene synthesis pipeline that leverages LRMs to construct indoor and outdoor 3D scenes from open-ended textual descriptions. Unlike prior approaches, Reason-3D requires no domain-specific training or architectural constraints. Given a scene prompt, our system extracts relevant objects using a combination of embedding-based retrieval and LLM-based semantic voting across three dimensions: physical, functional, and contextual relevance. We use a dual-stage placement process, where each object is placed in an optimal order autoregressively with the help of an LRM. Finally, Reason-3D refines outputs by making the LRM aware of potential collisions and resolving them. This architecture enables Reason-3D to operate entirely through natural language instructions. It supports complex spatial compositions, unusual object configurations, and even outdoor or hybrid environments without requiring any manual scripting or handcrafted scene rules. We evaluate Reason-3D on established indoor scene synthesis tasks where we significantly outperform prior methods in terms of plausibility and structural coherence (see Fig. 1). Furthermore, we highlight the generalization capabilities of Reason-3D by demonstrating applications in outdoor scenes and composition scenarios. Through an ablation study, we find that certain LRMs consistently outperform others across a diverse set of spatial reasoning tasks. We summarize our contributions as follows:

- We introduce Reason-3D, a novel text-to-scene pipeline that leverages Large Reasoning Models (LRMs) to retrieve and place objects in 3D space, based only on natural language descriptions of the scene, without requiring fine-tuning or pretraining.

Holodeck LayoutVLM **Reason-3D (ours)**

Figure 1: Showcase comparison for object retrieval and placement between Reason-3D and baseline approaches for the instruction *"A cozy living room of size 5 by 5 units. There is a plant on a small table in front of the L-shaped sofa."*

- Reason-3D significantly outperforms previous baselines by using a dual-phase placement strategy combining autoregressive layout with collision-aware refinement, enabling physically coherent scene synthesis.
- Reason-3D generalizes to any scene setting, as demonstrated by our outdoor scene synthesis and open-world compositions. By being able to run out-of-the-box on any object library without requiring scripting or hand-crafted rules.

2 Related Work

Recent progress in indoor scene synthesis has primarily followed two distinct approaches. One line of work capitalizes on the strong generative capabilities of image-based models, often employing Neural Radiance Fields (NeRFs) or 3D Gaussian splats as output representations (Schult et al. 2024; Zhou et al. 2024; Epstein et al. 2024; Po and Wetzstein 2024). While these methods produce visually realistic results, the generated scenes lack object-level separability, making them unsuitable for downstream tasks requiring precise object manipulation or interaction. A second research direction focuses on structured scene generation using intermediate representations, such as scene graphs or layout templates, coupled with curated asset libraries to synthesize 3D environments with discrete objects (Feng et al. 2023; Yang et al. 2024c).

Neural-3D generation. Within this structured generation paradigm, learning-based generative models such as diffusion models have emerged as powerful tools for modeling spatial priors (Hu et al. 2024). DiffuScene (Tang et al. 2024) applies a denoising diffusion process to generate unordered sets of object attributes, synthesizing layouts from the 3D-FRONT dataset, which comprises 19,000 annotated indoor scenes (Fu et al. 2021). Other methods use model training with scene graphs (Yang et al. 2025; Zhai et al. 2024; Lin and Yadong 2024; Zhai et al. 2023) and without scene graphs (Paschalidou et al. 2021; Yang et al. 2024a; Ritchie, Wang, and Lin 2019). Despite their effectiveness, these models require task-specific training and often struggle to generalize beyond the distributions they were trained on. In contrast, our method performs scene synthesis in a zero-shot setting, relying purely on language-based reasoning without additional fine-tuning or supervision.

Language-based room synthesis. The advent of Large Language Models (LLMs) has enabled open-vocabulary 3D scene synthesis, supporting the flexible generation of scenes without dependence on predefined labels or categories. These systems treat the LLM as a "design assistant" that outputs a high-level layout or scene graph, which is then materialized into 3D. Holodeck (Yang et al. 2024c) and others (Çelen et al. 2024; Aguina-Kang et al. 2024) orchestrate multi-agent LLM systems to generate scene graphs or DSLs (Domain Specific Language), which are later optimized through layout engines. These prompt-based LLM planners are able to interpret open-ended descriptions but often require post-processing or optimization to convert text into numeric layouts, which is an inherent limitation. LayoutVLM (Sun et al. 2025) represents the latest advancement in object placement, employing VLMs to generate two mutually reinforcing representations from visually marked images, and a self-consistent decoding process to improve spatial planning.

Direct LLM-to-layout methods. A straightforward approach to scene synthesis with LLMs involves directly querying the model for object positions and rotations. LayoutGPT (Feng et al. 2023) pioneered this strategy by retrieving relevant example layouts from a database to serve as in-context demonstrations. The GPT-based model then generates a new layout by predicting object bounding boxes conditioned on these exemplars. However, the reliance on a fixed library of example scenes limits generalization, as the diversity of layouts that can be produced is constrained by the database contents. Other approaches fine-tune pretrained LLMs for scene synthesis or editing (Bucher and Armeni 2025; Yang et al. 2024b), but similarly inherit dataset biases that hinder their generalizability. Concurrent work such as DirectLayout (Ran et al. 2025) prompts an LLM with chain-

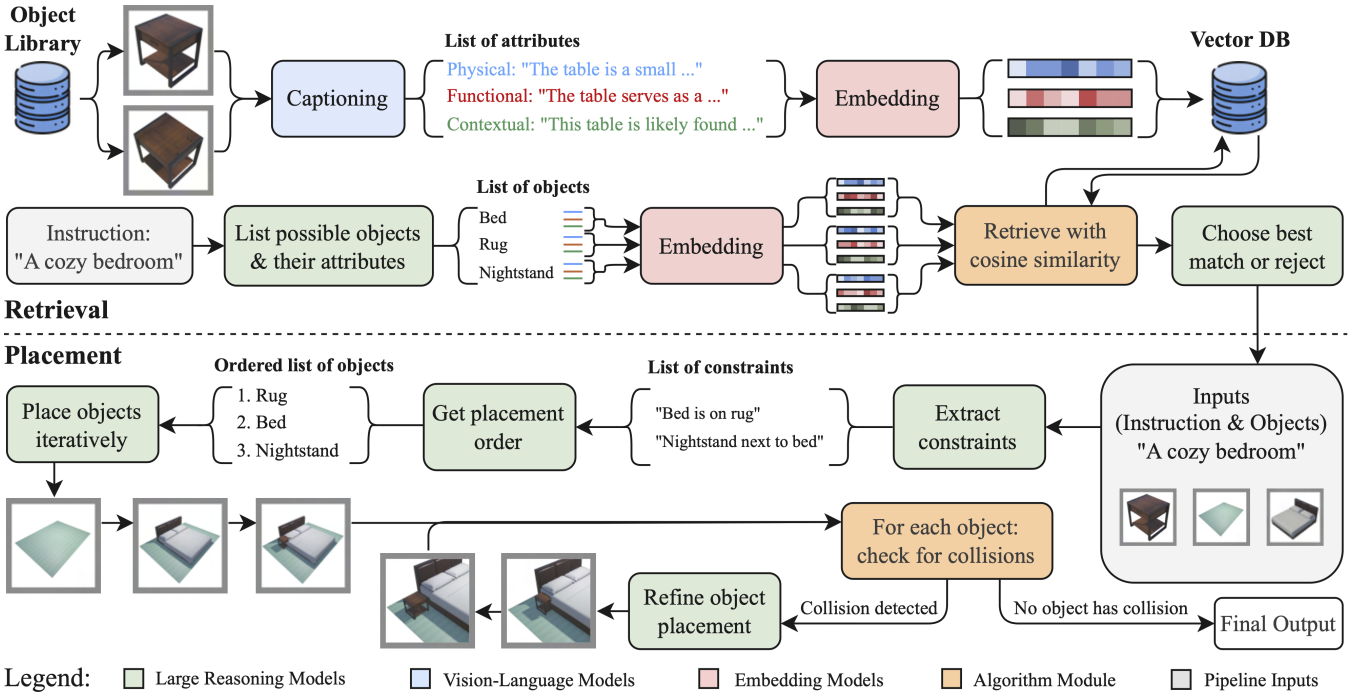

Figure 2: **Overview of our proposed architecture**. *Retrieval*: We start by processing objects from an asset library, generating images of these objects for captions and orientation. A captioning model creates object descriptions, which are then turned into embedding vectors and stored in a vector database. Given an instruction, we extract a list of objects that would be feasible according to the instruction. and subsequently query the database for such objects. *Placement*: Given the instruction and retrieved objects, we extract a set of constraints to determine an ordered sequence for object placement. Once all objects are placed, the scene is refined by calculating and adjusting for object collisions.

of-thought reasoning to generate a 2D bird's-eye view layout, which is then lifted into 3D and refined. While promising, this method is fine-tuned on the 3D-FRONT dataset, again limiting its applicability outside the training distribution. Naively prompting standard LLMs that lack spatial multi-step reasoning capabilities to generate object coordinates often results in physically implausible outputs (e.g., overlapping objects, unrealistic placements, or violations of spatial constraints) because general-purpose non-reasoning LLMs do not inherently model geometry, scale, or collision dynamics. As a result, most recent LLM-based scene synthesis pipelines incorporate auxiliary mechanisms, such as example-based in-context retrieval, spatial reasoning decomposition, or fine-tuned scene grammar models, to compensate for limitations in raw layout prediction and ensure plausible, physically consistent scenes. These hybrid frameworks confirm that LLMs are well-suited for capturing high-level design semantics and following open-ended user instructions, but also underscore the limitations of purely text-based models when tasked with precise spatial generation, in the absence of reasoning steps. In contrast, our method operates entirely with off-the-shelf LLMs exhibiting multi-step spatial reasoning, also referred to as LRMs, requiring no task-specific fine-tuning. This enables generalization to more diverse and open-ended settings, including outdoor environments.

3 Methodology

Our proposed method, Reason-3D, consists of multiple stages (see Fig. 2). Broadly speaking, the pipeline is divided into two phases: object retrieval and object placement.

Dataset and preprocessing. The objects used in our pipeline are sourced from the Objaverse dataset (Deitke et al. 2023). To ensure reliable operation, these 3D objects must conform to a standardized structure: specifically, each object is required to be upright and oriented consistently along a canonical forward direction. While the system includes runtime mechanisms to detect and correct certain rotational misalignments, overall robustness is significantly improved when operating on a geometrically consistent dataset. Since objects imported from Objaverse do not always adhere to these conventions, we apply a preprocessing step involving a vision-language model (VLM) to analyze the object's appearance and infer its intended orientation. The VLM receives four rendered views of the object and selects the image that most likely represents the front-facing orientation (see the extended version for the prompting strategy). The identified view is then used to reorient the object accordingly. The remaining discrepancies must be addressed dynamically during layout synthesis by the LRM, provided the object's dimensions and context allow for such adjustments.

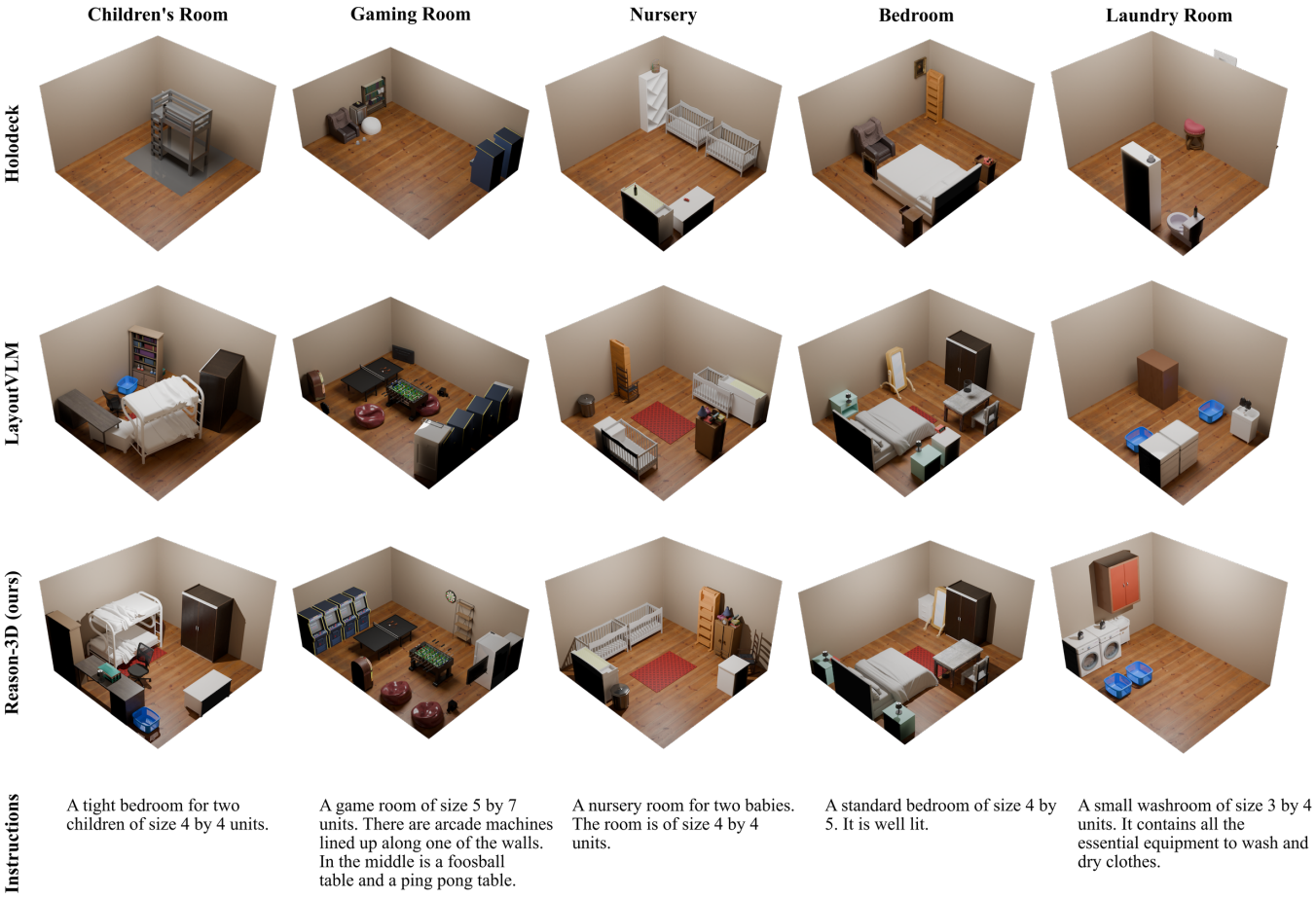

Figure 3: Qualitative comparison for object retrieval and placement across various scenes. We find that overall, Reason-3D can better follow instructions and place objects reasonably. Compared to Holodeck and Reason-3D, LayoutVLM was not designed to retrieve objects. We use the objects retrieved from Reason-3D for LayoutVLM.

Object retrieval. For each object in the dataset, we generate a structured textual description using an image captioning model. Each captioning prompt includes two rendered views of the object from different angles and a detailed instruction on what to output (see the extended version for more details). The resulting structured descriptions are embedded using an embedding model to obtain a high-dimensional representation that supports semantic retrieval. In parallel, an LRM processes the scene prompt to extract the set of required objects, with each extracted object similarly described using the same tripartite structure. During retrieval, cosine similarity is used to identify the top five closest object candidates based on their embeddings. These candidates are then evaluated by the LRM, which selects the most semantically appropriate instance or determines if no suitable match is available. This ensures that the retrieved instances faithfully reflect the user's intent, thereby enhancing overall prompt adherence. The retrieved objects are then passed to the placement phase.

Building the scene. Object placement and scene construction are carried out in two phases: initial placement and refinement, where Reason-3D leverages the multi-step reasoning of LRMs to sequentially determine each object's position and rotation in the context of previously placed objects, with full control over all three axes of rotation. The model receives scene constraints, object metadata (name and size), as well as a list of current placements, but no visual feedback (only labeled bounding boxes), requiring it to reason purely from spatial metadata.Before placement, we address two key challenges. First, implicit spatial relations in the scene prompt are made explicit using an LRM extraction step. This involves querying the LRM to imagine the layout and giving as output object-specific constraints, both relative to other objects and absolute. Second, object dependencies are resolved via a placement priority list generated by another LRM, ensuring a logical ordering (e.g., placing a table before placing a plate on top of that table). During placement, the LRM relies on bounding box dimensions to determine spatial relationships. However, in dense scenes, col-

Figure 4: Qualitative comparison of object placement performance when instruction complexity is increased. Every scene is generated from scratch with the entire instruction given up to and including its circled number.

lisions may occur because the model selectively considers only a subset of previously placed objects (typically those it deems locally relevant to the current placement). This limitation stems from constraints on reasoning depth and input length, which prevent the model from exhaustively evaluating all potential spatial conflicts across the entire scene. To mitigate these object collisions, we introduce a refinement step. After initial placement, the LRM receives detected collisions as input and revises object positions and rotations accordingly, processing each object in the same order as during initial placement. These collisions are overlaps of the objects' bounding boxes. The detected collisions comprise a complete list of all overlapping bounding boxes with a small buffer. To alleviate the complexity of the geometric reasoning task, each object is also annotated with a "size after rotation" attribute, which reflects the dimensions of the axis-aligned bounding box after applying the object's rotation. We find that this precomputed property substantially improves the model's ability to reason about spatial constraints in transformed coordinate spaces. Importantly, not all collisions are considered undesirable. Since reasoning is performed over bounding boxes, benign overlaps may be semantically valid (e.g., a trash bin under a table, even if the table and trash bin bounding boxes intersect). Therefore, the refinement step does not indiscriminately eliminate all collisions. Instead, the LRM evaluates each collision involving the current object being refined and determines whether it is contextually appropriate, allowing for fine-grained spatial reasoning and scene-specific tolerance of overlaps. This two-stage approach helps overcome common failure modes (e.g., wrongly ordered placements, unintended occlusions) and reduces reasoning load by injecting task-relevant geometric abstractions into the LRM's input.

4 Results

We conduct a comprehensive evaluation of our system across three core experiments: *object retrieval accuracy*, *object placement precision*, and *overall scene quality*. Additionally, we benchmark various state-of-the-art reasoning language models on a suite of spatial reasoning and planning tasks using the Reason-3D framework. Furthermore, we demonstrate generalization through the synthesis of complex outdoor scenes. We conduct human evaluation studies using an online tool that provides access to trained participants. More information on the human evaluation and the Elo metric calculation can be found in the extended version.

Implementation details. In all experiments, we use the Gemini 2.5 model (Comanici et al. 2025) for object placement, object refinement, and image captioning (VLM). While we tested the pipeline on other LRMs, the results showed that Gemini 2.5 consistently yielded the best output quality. We use Gemini `embedding-004` as the embedding model.

Baselines. We compare our system, Reason-3D, against two recent scene synthesis frameworks: Holodeck (Yang et al. 2024c) and LayoutVLM (Sun et al. 2025). Holodeck is a multi-agent system that uses LLMs to generate domain-specific layout programs, which are later optimized using layout engines. LayoutVLM, on the other hand, uses a VLM to produce visually grounded spatial representations and a self-consistency decoding mechanism to enhance spatial layout planning. We utilize the publicly released Objaverse subset provided by the LayoutVLM source code. This dataset is converted using Objathor (Yang et al. 2024c) for compatibility with Holodeck, while Reason-3D uses its own preprocessing. Since LayoutVLM does not support object

Model	Win-rates [%]			Elo ↑
	HD	LVLM	Reason-3D	
HD	–	26.9	4.8	1500
LVLM	73.1	–	1.6	1650
Reason-3D (ours)	**95.2**	**98.4**	–	**2248**

Table 1: Results of human evaluation comparing our approach to baselines. Win-rates show the percentage of times the row model won against the column model in head-to-head comparisons. Higher Elo scores indicate stronger performance. Results show that participants strongly prefer Reason-3D (R-3D) over both Holodeck (HD) and LayoutVLM (LVLM) in head-to-head comparisons.

Model	Top-1	Top-5	Top-10
Holodeck	7%	8%	8%
Reason-3D (ours)	**75%**	**85%**	**90%**

Table 2: Results of the retrieval experiment. Top-k accuracies indicate the fraction of time the correct object was present in the top-k retrieved objects.

Model	1	2	3	4	5
LayoutVLM	2.8	3.4	3.0	2.5	2.4
Reason-3D (ours)	**4.4**	**3.9**	**4.4**	**4.1**	**4.3**

Table 3: Results of the human evaluation on object placement. Each column indicates the instruction complexity (see Fig. 4). All results are below 0.2 standard deviation and on a scale from 1 (bad) to 5 (excellent). Overall, we observe that participants prefer Reason-3D over previous work, and that as the instruction complexity increases, Reason-3D can place objects more reasonably than previous work.

retrieval, we supply it with objects retrieved using Reason-3D for the full scene synthesis experiments. For experiments focusing on layout design, the same set of objects is predetermined and supplied to both Reason-3D and LayoutVLM, alongside the textual prompts. In summary, we compare Reason-3D against Holodeck on the retrieval task, against LayoutVLM on layout generation, and against both systems in the full scene synthesis evaluation.

Full scene synthesis. We evaluate full scene synthesis using only a text prompt and room dimensions (see Fig. 3) for Reason-3D, LayoutVLM (Sun et al. 2025), and Holodeck (Yang et al. 2024c). LayoutVLM receives additional inputs, including a predefined object list and room configuration via its configuration file. In Table 1, we present the results of the human preference study using 60 participants. We find that Reason-3D produces more semantically aligned and spatially valid scenes compared to both baselines.

Retrieval benchmark. To evaluate retrieval accuracy, we assess whether objects described in a scene prompt can be correctly retrieved from the object database. For controlled testing, we generate scene descriptions from a predefined list

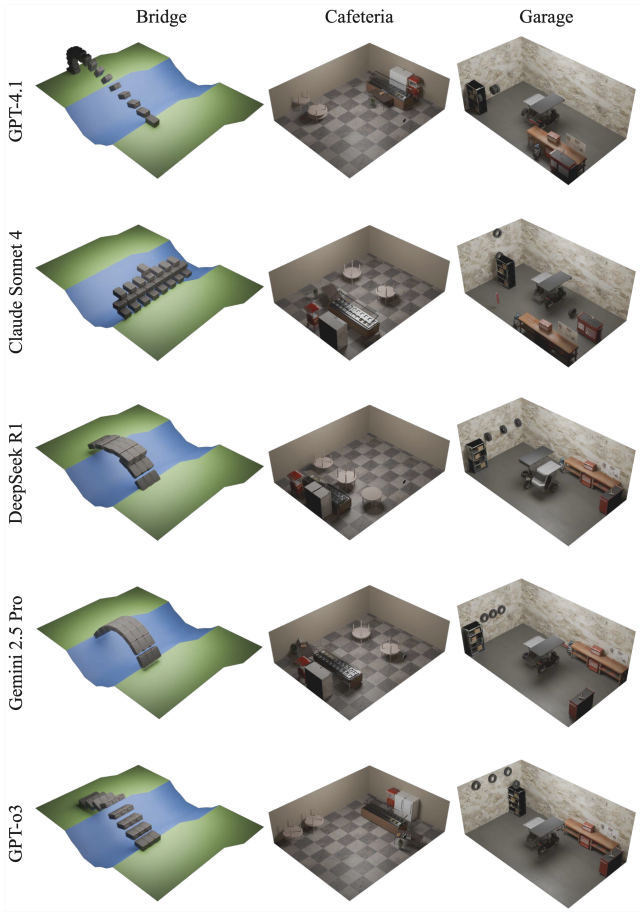

Figure 5: We benchmark various LRMs on three scenes. We find that Gemini 2.5 Pro achieves the best overall performance on spatial reasoning tasks. The bridge-building task also showcases the ability of object composition (building a bridge out of individual stone objects).

of target objects using an LLM, with each object accompanied by a rich semantic description. During retrieval, each target object is compared to all entries in the database via cosine similarity of embedding vectors. The results yield a ranked list of the top 10 most similar objects. We report top-1, top-5, and top-10 accuracy, where top-k indicates the proportion of cases in which the correct object is ranked within the first k retrieved candidates. This evaluation is averaged across multiple runs with a 12-object test set. As observed in Table 2, we find that Reason-3D outperforms Holodeck in retrieval accuracy.

Object placement benchmark. We compare object placement quality between Reason-3D and LayoutVLM. Starting from a simple prompt and a minimal set of objects, we incrementally increase scene complexity by adding objects with layout constraints. We present a qualitative example of this experiment in Fig. 4, illustrating a five-step iteration comparison. In Table 3, we present the results of a human evaluation study with 43 participants asked to rate the quality of the scene and its adherence to the provided

Models	Win-rates [%]					Elo ↑	Tokens	
	GPT-4.1	Sonnet 4	DeepSeek-R1	GPT-o3	Gemini 2.5 Pro		Input	Output
GPT-4.1	–	38.5	18.2	0.0	10.0	1500	41,424	971
Sonnet 4	61.5	–	14.3	11.1	8.3	1566	52,225	1,698
DeepSeek-R1	81.8	85.7	–	40.0	7.7	1809	47,426	114,506
GPT-o3	100.0	88.9	60.0	–	29.4	1938	39,770	35,249
Gemini 2.5 Pro	**90.0**	**91.7**	**92.3**	**70.6**	–	**2091**	48,470	29,312

Table 4: Results of human evaluation on LRM benchmark. Each cell shows the percentage of times the row model won against the column model in head-to-head comparisons. The token statistics for different models were computed on a per-scene basis, assuming an average of 14 objects per scene. While the calculation and accounting of "reasoning tokens" vary across models, they are generally reflected in the output token count. GPT-4.1, which does not support multi-step reasoning, produces the fewest output tokens. In contrast, DeepSeek generates the highest number of output tokens, indicating more extensive reasoning during scene construction. We find that Gemini 2.5 Pro performed the best overall, achieving the highest Elo score.

Figure 6: Qualitative example of an outdoor scene generated with Reason-3D, containing 70 objects. Instruction: *A big circular floating island. On top are 25 trees surrounding the island on its edge. In the middle, on top of the island, is a water fountain. Around this fountain unfolds a lively medieval marketplace, with a fish stand, a meat stand, a weapon stand, and a vegetable stand.*

constraints. We find that Reason-3D consistently produces more spatially coherent placements as the instruction complexity increases. Reason-3D can infer object orientation from bounding box dimensions, successfully aligning objects such as a periodic table, an area in which previous methods like LayoutVLM offer no necessary rotational support across all axes. More details and examples on this can be found in the extended version.

Large Reasoning Model benchmark. To assess spatial reasoning capabilities in isolation, we benchmark the following LRMs: Gemini 2.5 Pro (Comanici et al. 2025), GPT-o3 (OpenAI 2025b), Claude Sonnet 4 (Anthropic 2025), DeepSeek-R1 (DeepSeek 2025). Additionally, we use GPT-4.1 (OpenAI 2025a) as a baseline due to its lack of test-time compute (i.e., no explicit multi-step reasoning). The evaluation covers three dimensions of spatial reasoning. First, we test the ability to reason in a transformed coordinate system, requiring the model to apply correct geometric transformations. Second, we assess whether the model can infer object orientation based on the dimensions of its axis-aligned bounding box. Finally, we combine both challenges in a complex spatial planning task, requiring integrated multi-object reasoning. We test the LRMs on three scenes (more information can be found in the extended version). We conducted a human evaluation with 40 participants, where each participant rated a scene synthesized by two models in a head-to-head comparison. The results are presented in Table 4, with comparisons illustrated in Fig. 5. Furthermore, Table 4 summarizes the average token usage for each model, providing insight into the computational cost associated with different levels of reasoning.

Outdoor scenes. We demonstrate Reason-3D's ability to generalize beyond indoor environments by synthesizing several outdoor scenes. Fig. 6 presents an open-world example composed of 70 objects, illustrating Reason-3D's capability to handle large-scale layouts and diverse spatial contexts. This highlights the flexibility of the system and its applicability across a wide range of scene types. Additional examples in the extended version.

Limitations. Our framework assumes that the object database contains objects that are pre-aligned upright. Misalignment involving other rotational degrees of freedom cannot be resolved currently, as they would require additional processing. Additionally, object scales must be approximately correct between objects. Furthermore, our framework does not generate structural elements such as floors or walls. More generally, Reason-3D depends on LRMs, implicitly inheriting all of their limitations.

5 Conclusion

We presented Reason-3D, a text-driven framework for 3D scene synthesis that leverages the multi-step reasoning capabilities of LRMs. Our results demonstrate that, without requiring task-specific fine-tuning, these off-the-shelf models can retrieve and place objects more desirably than previous baselines. Given the rapid evolution of these models, we anticipate their continued integration and refinement in future research.

References

Aguina-Kang, R.; Gumin, M.; Han, D. H.; Morris, S.; Yoo, S. J.; Ganeshan, A.; Jones, R. K.; Wei, Q. A.; Fu, K.; and Ritchie, D. 2024. Open-Universe Indoor Scene Generation using LLM Program Synthesis and Uncurated Object Databases. arXiv:2403.09675.

Anthropic. 2025. Introducing Claude 4: Claude Opus 4 and Claude Sonnet 4. https://www.anthropic.com/news/claude-4. Accessed: 2025-11-12.

Bu, Q.; Zeng, J.; Chen, L.; Yang, Y.; Zhou, G.; Yan, J.; Luo, P.; Cui, H.; Ma, Y.; and Li, H. 2024. Closed-loop Visuomotor Control with Generative Expectation for Robotic Manipulation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*.

Bucher, M. J.; and Armeni, I. 2025. ReSpace: Text-Driven 3D Scene Synthesis and Editing with Preference Alignment. arXiv:2506.02459.

Çelen, A.; Han, G.; Schindler, K.; Van Gool, L.; Armeni, I.; Obukhov, A.; and Wang, X. 2024. I-Design: Personalized LLM Interior Designer. arXiv:2404.02838.

Cen, J.; Wu, C.; Liu, X.; Yin, S.; Pei, Y.; Yang, J.; Chen, Q.; Duan, N.; and Zhang, J. 2024. Using Left and Right Brains Together: Towards Vision and Language Planning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Comanici, G.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

DeepSeek. 2025. DeepSeek-R1 Release: Fully Open-Source Reasoning Model. https://api-docs.deepseek.com/news/news250120. Accessed: 2025-11-12.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; VanderBilt, E.; Kembhavi, A.; Vondrick, C.; Gkioxari, G.; Ehsani, K.; Schmidt, L.; and Farhadi, A. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. arXiv:2307.05663.

Deitke, M.; VanderBilt, E.; Herrasti, A.; Weihs, L.; Ehsani, K.; Salvador, J.; Han, W.; Kolve, E.; Kembhavi, A.; and Mottaghi, R. 2022. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Epstein, D.; Poole, B.; Mildenhall, B.; Efros, A. A.; and Holynski, A. 2024. Disentangled 3D Scene Generation with Layout Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Feng, W.; Zhu, W.; Fu, T.-J.; Jampani, V.; Akula, A. R.; He, X.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Fu, H.; Cai, B.; Gao, L.; Zhang, L.-X.; Wang, J.; Li, C.; Zeng, Q.; Sun, C.; Jia, R.; Zhao, B.; et al. 2021. 3D-FRONT: 3D Furnished Rooms with Layouts and Semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Gao, R.; Chen, K.; Xie, E.; HONG, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2024. MagicDrive: Street View Generation with Diverse 3D Geometry Control. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hu, S.; Arroyo, D. M.; Debats, S.; Manhardt, F.; Carlone, L.; and Tombari, F. 2024. Mixed Diffusion for 3D Indoor Scene Synthesis. arXiv:2405.21066.

Lee, S.; Park, S.; and Kim, H. 2025. DynScene: Scalable Generation of Dynamic Robotic Manipulation Scenes for Embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, C.; and Yadong, M. 2024. InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mao, J.; Li, B.; Ivanovic, B.; Chen, Y.; Wang, Y.; You, Y.; Xiao, C.; Xu, D.; Pavone, M.; and Wang, Y. 2025. DreamDrive: Generative 4D Scene Modeling from Street View Images. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

OpenAI. 2025a. Introducing GPT-4.1 in the API: GPT-4.1, Mini & Nano. https://openai.com/index/gpt-4-1/. Accessed: 2025-11-12.

OpenAI. 2025b. Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-11-12.

Paschalidou, D.; Kar, A.; Shugrina, M.; Kreis, K.; Geiger, A.; and Fidler, S. 2021. ATISS: Autoregressive Transformers for Indoor Scene Synthesis. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Po, R.; and Wetzstein, G. 2024. Compositional 3D Scene Generation Using Locally Conditioned Diffusion. In *Proceedings of the International Conference on 3D Vision (3DV)*.

Raistrick, A.; Mei, L.; Kayan, K.; Yan, D.; Zuo, Y.; Han, B.; Wen, H.; Parakh, M.; Alexandropoulos, S.; Lipson, L.; et al. 2024. Infinigen Indoors: Photorealistic Indoor Scenes Using Procedural Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ran, X.; Li, Y.; Xu, L.; Yu, M.; and Dai, B. 2025. Direct Numerical Layout Generation for 3D Indoor Scene Synthesis via Spatial Reasoning. arXiv:2506.05341.

Ritchie, D.; Wang, K.; and Lin, Y.-A. 2019. Fast and Flexible Indoor Scene Synthesis via Deep Convolutional Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schult, J.; Tsai, S. S.; Höllein, L.; Wu, B.; Wang, J.; Ma, C.-Y.; Li, K.; Wang, X.; Wimbauer, F.; He, Z.; et al. 2024. ControlRoom3D: Room Generation Using Semantic Proxy Rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Sun, F.-Y.; Liu, W.; Gu, S.; Lim, D.; Bhat, G.; Tombari, F.; Li, M.; Haber, N.; and Wu, J. 2025. LayoutVLM: Differentiable Optimization of 3D Layout via Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tang, J.; Nie, Y.; Markhasin, L.; Dai, A.; Thies, J.; and Nießner, M. 2024. Diffuscene: Denoising Diffusion Models for Generative Indoor Scene Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Unity Technologies. 2025. Unity Asset Store. https://assetstore.unity.com. Accessed: 2025-07-29.

Wen, B.; Xie, H.; Chen, Z.; Hong, F.; and Liu, Z. 2025. 3D Scene Generation: A Survey. arXiv:2505.05474.

Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; et al. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. arXiv:2501.09686.

Yang, Y.; Jia, B.; Zhi, P.; and Huang, S. 2024a. PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Y.; Lu, J.; Zhao, Z.; Luo, Z.; Yu, J. J.; Sanchez, V.; and Zheng, F. 2024b. LLPlace: The 3D Indoor Scene Layout Generation and Editing via Large Language Model. arXiv:2406.03866.

Yang, Y.; Sun, F.-Y.; Weihs, L.; VanderBilt, E.; Herrasti, A.; Han, W.; Wu, J.; Haber, N.; Krishna, R.; Liu, L.; et al. 2024c. Holodeck: Language Guided Generation of 3D Embodied AI Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Z.; Lu, K.; Zhang, C.; Qi, J.; Jiang, H.; Ma, R.; Yin, S.; Xu, Y.; Xing, M.; Xiao, Z.; et al. 2025. MMGDreamer: Mixed-Modality Graph for Geometry-Controllable 3D Indoor Scene Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhai, G.; Örnek, E. P.; Chen, D. Z.; Liao, R.; Di, Y.; Navab, N.; Tombari, F.; and Busam, B. 2024. EchoScene: Indoor Scene Generation via Information Echo over Scene Graph Diffusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhai, G.; Örnek, E. P.; Wu, S.-C.; Di, Y.; Tombari, F.; Navab, N.; and Busam, B. 2023. CommonScenes: Generating Commonsense 3D Indoor Scenes with Scene Graph Diffusion. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Zhou, X.; Ran, X.; Xiong, Y.; He, J.; Lin, Z.; Wang, Y.; Sun, D.; and Yang, M.-H. 2024. GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-Guided Generative Gaussian Splatting. In *Proceedings of the International Conference on Machine Learning (ICML)*.