PARSE: Part-Aware Relational Spatial Modeling

Yinuo Bai[1,2] Peijun Xu[1] Kuixiang Shao[1] Yuyang Jiao[1] Jingxuan Zhang[1]
Kaixin Yao[1,2,†] Jiayuan Gu[1,*] Jingyi Yu[1,*]
[1]ShanghaiTech University [2]Deemos Technology
{baiyn2022, xupj2025, shaokx2025, jiaoyy2022, zhangjx12023,
yaokx2024, gujy1, yujingyi}@shanghaitech.edu.cn
[†]Project Leader [*]Corresponding Author

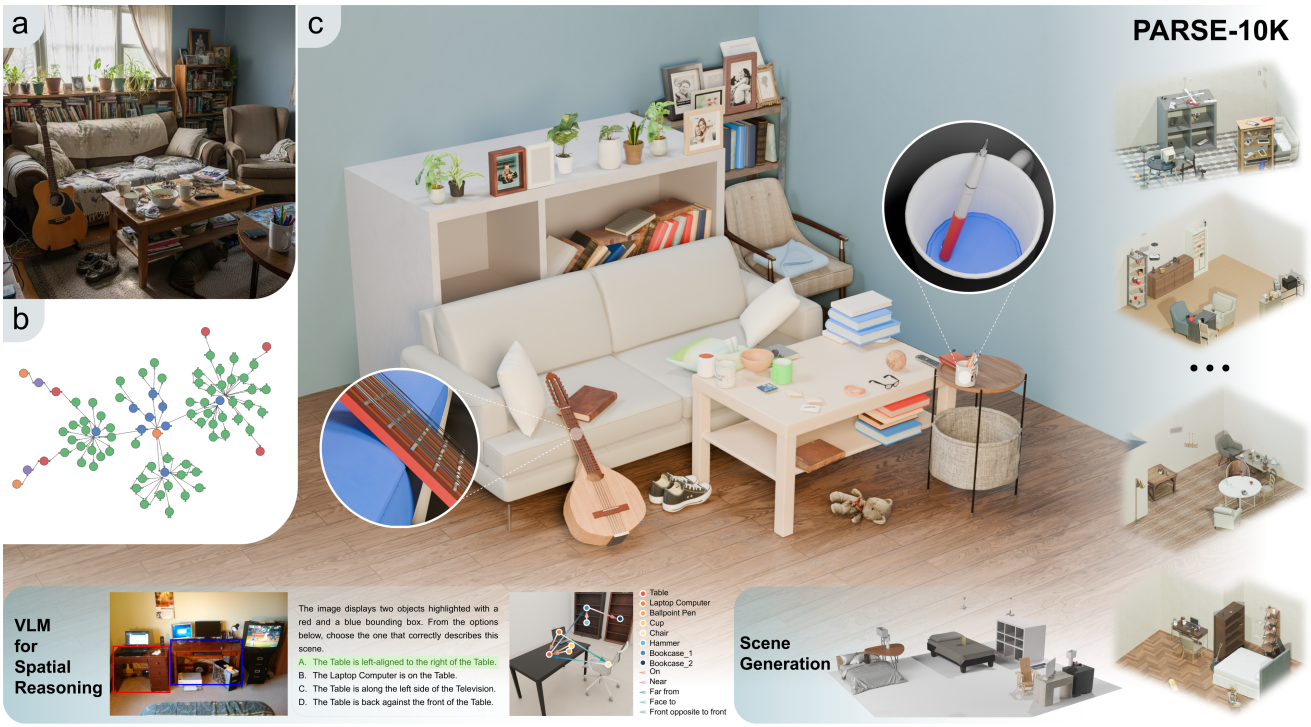

Figure 1. **Overview of the capabilities of PARSE.** Using a real image (a) as reference, we first construct Part-centric Assembly Graphs (PAGs) (b) that capture its spatial organization of objects. Then, by combining PARSE with physics simulation, we generate physically plausible 3D scenes (c) from these PAGs, featuring diverse inter-object relationships and rich part-level contacts. Furthermore, we introduce PARSE-10K, a collection of high-quality 3D indoor scenes with fully part-segmented object instances, which effectively supports downstream tasks such as fine-tuning VLMs for spatial reasoning and enhancing 3D scene generation.

Abstract

Inter-object relations underpin spatial intelligence, yet existing representations—linguistic prepositions or object-level scene graphs—are too coarse to specify which regions actually support, contain, or contact one another, leading to ambiguous and physically inconsistent layouts. To address these ambiguities, a part-level formulation is needed; therefore, we introduce PARSE, a framework that explicitly models how object parts interact to determine feasible and spatially grounded scene configurations. PARSE centers on the Part-centric Assembly Graph (PAG), which encodes geometric relations between specific object parts, and a Part-Aware Spatial Configuration Solver that converts these relations into geometric constraints to assemble collision-free, physically valid scenes. Using PARSE, we

build PARSE-10K, a dataset of 10,000 3D indoor scenes constructed from real-image layout priors and a curated part-annotated shape database, each with dense contact structures and a part-level contact graph. With this structured, spatially grounded supervision, fine-tuning Qwen3-VL on PARSE-10K yields stronger object-level layout reasoning and more accurate part-level relation understanding; furthermore, leveraging PAGs as structural priors in 3D generation models leads to scenes with substantially improved physical realism and structural complexity. Together, these results show that PARSE significantly advances geometry-grounded spatial reasoning and supports the generation of physically consistent 3D scenes.

1. Introduction

Modeling inter-object relations is the next frontier of spatial intelligence because many fundamental tasks—scene generation [29, 32], layout synthesis [12], tidying [52], packing [63], stacking [28], and embodied manipulation [20, 68]—depend more on how objects relate than on their isolated shapes. Relations encode support, containment, attachment, occlusion, and accessibility, which determine stability, affordances, and task feasibility. This view resonates with Latour's actor–network theory (ANT) [27]: objects derive meaning and function from the network of relations they maintain with other objects and agents, not from intrinsic properties alone. The critical question we address is **how to operationalize these rich relations into an effective representation for spatial modeling.**

Vision-language models (VLMs) [1, 34, 36] offer a promising path for understanding inter-object relations expressed through prepositions such as *on*, *in*, or *against*. However, **these linguistic cues are inherently coarse and context-dependent**: "a book on a table" may refer to the spine or the cover contacting the surface, while "a guitar leaning on a bookcase" could involve its head or body. Such expressions are underspecified regarding contact points or supporting regions, making their translation into spatial configurations fundamentally ambiguous. This limitation also exists in prior relational representations, most notably scene graphs [21, 25]. Prior scene graph representations operate at object-level granularity [6, 17, 18, 22, 41], providing insufficient specificity for fine-grained spatial understanding and realistic scene generation.

We argue that **a more powerful and versatile representation emerges from modeling interactions at the part level**. Part-level relations bridge high-level language descriptions and low-level spatial configurations. For instance, a chair stands on the floor *via its feet*, a mug rests on a table *by its base*, and a broom leans against a wall *at its tip*. This part-centric specification transforms ambiguous prepositions into concrete geometric constraints, effectively pruning the vast search space of valid configurations. When integrated into representations such as scene graphs, these fine-grained relations enable a more structured and controllable approach to spatial reasoning and scene synthesis.

In this work, we propose **PARSE**, **PA**rt-aware **R**elational **S**patial mod**E**ling. At its core is the Part-centric Assembly Graph (PAG), a descriptive scene representation where each edge encodes geometric relations between specific parts of connected object nodes. The PAG is organized as a directed acyclic graph, with a hierarchy that guides the assembly of objects into a spatially complex scene. Building on this representation, we introduce a Part-Aware Spatial Configuration Solver, which instantiates PAGs as valid 3D scenes. The solver converts each inter-part relation into geometric constraints, progressively narrowing the feasible pose space of each object and then sampling collision-free solutions efficiently. By traversing the graph from the root, it incrementally generates scenes that adhere to the underlying part-aware structure.

Building on this framework, we construct PARSE-10K, a large and high-quality dataset of 3D indoor scenes with fully part-segmented object instances. We begin by extracting layout priors from real images to obtain a set of semantically plausible and structurally complex PAGs. In parallel, we consolidate multiple public datasets with part annotations [7, 10, 13, 49] and incorporate part-segmented generative assets [62] to build a retrieval database covering 132 object categories for scene assembly. Leveraging these PAGs and the part-level database, we generate 10,000 indoor scenes across 17 room types, each characterized by rich contact structures and accompanied by a corresponding part-level contact graph, which offers an additional source of fine-grained contact information for downstream tasks.

To evaluate the utility of our dataset, we fine-tune Qwen3-VL [36] on PARSE-10K and assess its performance on spatial reasoning tasks. The fine-tuned model shows consistent improvements in both object-level layout reasoning and part-level relational understanding. Furthermore, incorporating PAGs from our dataset as structural priors in 3D generation networks significantly enhances the physical realism and structural complexity of synthesized scenes. These results demonstrate the effectiveness of PAG in advancing geometry-grounded spatial reasoning and physically consistent 3D scene generation.

2. Related Work

Scene graphs [21, 25] provide a structured representation of objects and their relations, powering progress in captioning [55, 65], VQA [3, 47], and image retrieval [40, 50]. Their extension to 3D [4, 24] incorporates geometry and spatial layout, advancing scene understanding [48, 53] and embodied reasoning [30, 51]. However, these methods treat objects as indivisible units, leaving them unable to capture

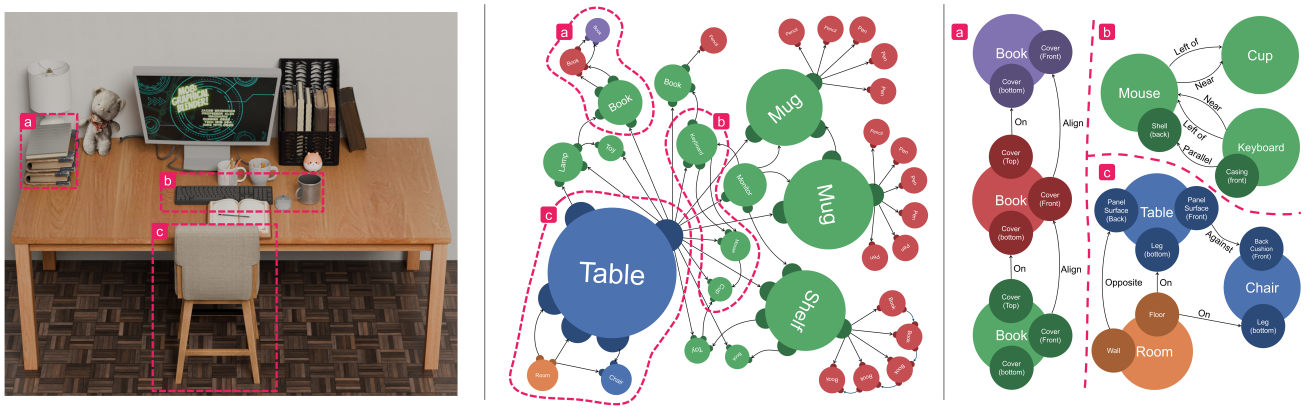

Figure 2. **An illustrative example of our Part-centric Assembly Graph (PAG).** Left: A 3D scene generated by PARSE, with specific local regions highlighted. Middle: The corresponding global PAG, emphasizing the sub-PAGs that match the highlighted regions on the left. In the graph, large labeled circles represent object nodes, while small dark circles attached to them represent part nodes (unrelated part nodes are omitted for clarity). Right: Zoomed-in views of three sub-PAGs. These panels explicitly annotate the specific surfaces used to define relational constraints, alongside the corresponding Object-Level Spatial Edges and Part-Level Geometric Edges.

the part-level interactions that determine physical stability and support. To overcome this limitation, our PAG models fine-grained part–part relations, enabling explicit reasoning about contact, support, and attachment beyond the capabilities of object-level scene graphs.

Recent works [57] in 3D scene generation incorporate inter-object spatial relations to improve structural plausibility. Graph-based approaches condition scene synthesis on semantic or geometric layout structures [15, 16, 32, 60], while multimodal and diffusion-based models further align language with 3D geometry to directly produce structured and coherent environments [14, 19, 39, 44]. Beyond network-based paradigms, procedural generation provides an explicit rule-based mechanism for specifying spatial structure. Early systems such as ProcTHOR [9] rely on rigid placement rules that ensure plausible layouts but limit controllability to coarse factors like room type, preventing users from directly specifying inter-object relations. More recent methods improve flexibility by leveraging large language models to map linguistic spatial cues to object placements [12, 43] or to generate constraint programs that can be solved by geometric optimizers [56, 67]. However, using LLMs as intermediaries introduces semantic ambiguity, often weakening the precision of the resulting geometric constraints. Infinigen [38] mitigates this by adopting human-readable spatial rules, allowing users to author precise and highly controllable layouts. However, existing procedural systems remain object-centric and cannot capture the fine-grained part interactions needed for precise physical arrangements, leading to inefficient search over large solution spaces. In contrast, our PAG-guided solver encodes explicit part–part constraints, sharply reducing the feasible space and enabling far more efficient generation with higher geometric fidelity and physical consistency.

At the data level, existing indoor scene datasets exhibit analogous limitations. Real-world scanned datasets [5, 8, 42, 58] provide high-fidelity spatial information and semantic labels, but their object-level meshes are often noisy, incomplete, or fused due to occlusion and reconstruction artifacts, hindering accurate physical reasoning. Synthetic datasets [9, 13, 31, 59] offer cleaner CAD models and greater diversity. However, many meshes are not cleanly decomposed into distinct object instances and lack part-level granularity, making it difficult to reliably model or detect critical, physics-grounded inter-object relations. To fill this gap, PARSE-10K provides consistent part-level annotations and explicit physical relations, delivering the fine-grained supervision absent from existing indoor scene datasets.

3. Part-Centric Assembly Graph

Previous work on scene graphs [25] often models relationships between whole objects, limiting their precision in capturing fine-grained interactions. To enable a deeper spatial understanding and more precise 3D scene generation, we introduce the Part-centric Assembly Graph (PAG), a representation centered on the expressive power of part-aware relations. As illustrated in Fig. 2, the PAG is a hierarchical graph designed to explicitly model the detailed geometric constraints between object parts, providing a structured foundation for both analyzing and synthesizing complex, physically coherent scenes.

3.1. Nodes: A Two-Level Structure

To effectively model part-aware relations, the nodes ($\mathcal{V}$) in a PAG are organized into a two-level structure.

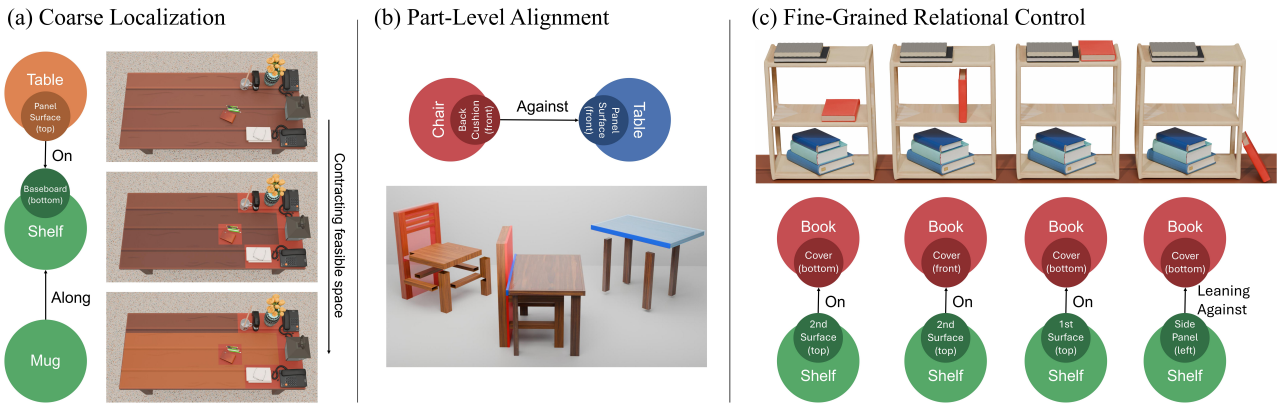

Figure 3. **Controllable Scene Synthesis via Part-Aware Spatial Configuration Solver. (a) Coarse Localization:** The solver first prunes occupied regions (red) from the 2D support surface, then further contracts the feasible space using object-level spatial relations (orange). **(b) Part-Level Alignment:** Precise geometric alignment is achieved by enforcing constraints (*e.g.*, coplanarity) between specific surfaces identified by the solver. This drastically reduces the feasible pose space for final pose sampling. **(c) Fine-Grained Relational Control:** Specifying different part-level geometric relations in the PAG results in distinct and predictable arrangements, showcasing the framework's fine-grained controllability.

Object Nodes ($\mathcal{V}_O$). These nodes form the upper level, representing the primary entities in a scene. Each object node encapsulates a semantic query, not a specific 3D instance. This query can be either a single, concrete category or an explicit set of candidate categories. This design defers the choice of a specific geometric instance to the synthesis stage, greatly enhancing the compositional diversity of generated scenes.

Part Nodes ($\mathcal{V}_P$). These nodes form the lower level and are the fundamental units for our part-aware approach. Each object node serves as a parent to a set of part nodes representing its geometrically meaningful components (*e.g.*, a "chair" connects to its "legs", "seat", and "back cushion"). Each part is further defined by a set of labeled surfaces (*e.g.*, *top, bottom, front, back, left, right*) that are assigned with respect to the asset's canonical pose. These labeled surfaces act as the specific geometric interfaces for alignment and contact, enabling precise constraint definition that govern the scene's assembly structure.

3.2. Edges: Part-Aware Relations

The edges ($\mathcal{E}$) in a PAG primarily model the rich, part-aware relations for scene assembly. Along with the intra-object edges that associate each object node with its own parts, the PAG mainly encodes inter-object relations through edges at two different granularities:

Object-Level Spatial Edges ($\mathcal{E}_{\text{obj}}$). These edges encode coarse spatial relations, such as *left of, behind, near* and *in front of*. Operating at the object level, they connect two object nodes to define macroscopic arrangements. They serve as optional, high-level constraints that guide the overall scene layout.

Part-Level Geometric Edges ($\mathcal{E}_{\text{part}}$). These edges encode fine-grained geometric relations, forming the core of the PAG's expressiveness. Specifically, each edge is associated with a spatial preposition (*e.g.*, *on, in, against,* and *aligned with*). Precise physical interactions are specified by connecting two specific part nodes belonging to different parent objects. This part-level linkage enables the graph to encode highly nuanced arrangements. For instance, describing "a book toppled forward onto a table" simply requires an *on* edge connecting the book's "cover" part node, labeled with the "front" surface, to the table's "surface plane" part node, labeled with the "top" surface.

3.3. Hierarchical Assembly Structure

A static 3D scene can be viewed as a collection of dense, interdependent geometric relations. To manage this complexity, our PAG representation adopts an assembly-centric perspective, viewing a stable scene as the outcome of a sequential construction process. This view allows us to represent scene structure in a more computationally tractable manner.

This assembly-centric perspective is realized through a key structural constraint of PAG: the entire graph must be a Directed Acyclic Graph (DAG). This global acyclic property is the necessary mathematical structure for representing a sequential process without circular dependencies, and it directly ensures physical realizability by enforcing a valid,

Table 1. **Comparison of indoor 3D scene datasets.** Columns from left to right denote: dataset name, number of scenes, number of objects, average objects per scene, layout generation method, whether physics simulation or optimization is applied, whether object parts are annotated, and whether part–part contact annotations are provided.

Dataset	# Scenes	# Objects	# Avg.Objects	Layout Generation	Physical Optimization	Part Annots.	Part-Level Contact Annots.
HSSD-200 [23]	211	18656	329.7	Human-designed	✗	✗	✗
3D-FRONT [13]	18968	13151	6.9	Human-designed	✗	✗	✗
FurniScene [61]	111698	39691	14.4	Human-designed	✗	✗	✗
METASCENES [59]	706	15366	-	Real-world Scanned	✓	✗	✗
PARSE-10K (Ours)	10000	17372	49.9	Real-image Guided	✓	✓	✓

step-by-step construction order. Additionally, we define that each object must have a unique physical supporter, a rule that naturally organizes the scene into a clear hierarchical structure. Ultimately, this overall design makes the scene-wide constraint satisfaction problem computationally tractable by decomposing it into a well-defined sequence of localized subproblems—one for each object in the assembly order.

4. PARSE-10K

We introduce the PARSE framework, our procedural synthesis pipeline that instantiates abstract PAGs into physically plausible and geometrically precise 3D scenes. This framework serves as the engine for building PARSE-10K, a large-scale dataset of diverse, part-aware indoor scenes. At the core of our framework is the Part-Aware Spatial Configuration Solver.

4.1. Part-Aware Spatial Configuration Solver

Given a PAG, the Part-Aware Spatial Configuration Solver instantiates it into a 3D scene by processing its object nodes in a topological sort. This traversal follows the sequential assembly order induced by the PAG's support relations. For each object in this sequence, the solver finds a valid pose through a coarse-to-fine process of progressive refinement. As illustrated by the key steps in Fig. 3, it sequentially applies all relevant constraints, with each new constraint further shrinking the object's feasible pose space until a precise solution is found. The instantiation of each object node unfolds as follows:

Coarse Localization. As each object node in a PAG has a unique supporter, the solving process begins within a 2D candidate region defined on the support surface, from which all previously occupied areas have been excluded. We first apply the high-level, object-level spatial edges. These constraints contract the node's feasible region to a smaller subspace. For instance, a "*left of*" relation imposes a plane that restricts the object's valid translational range to one side of the target object.

Part-Level Alignment. At this stage, a specific 3D asset, complete with per-part segmentation and semantic labels, is instantiated from our asset library based on the node's semantic query. Once an asset is chosen, the solver resolves the part-level geometric constraints. Guided by the spatial preposition of the connecting edge, the resolution strategy diverges based on surface specifications. If specific labeled surfaces of the connected part nodes are explicitly provided, the solver directly uses these identifiers. If exact surfaces are not specified, the solver performs a geometric reasoning step. For example, for an *on* relation, it dynamically identifies the supported part's lowest bottom surface while searching the target part for a suitable upward-facing support plane. The identified surfaces—whether explicitly provided or geometrically inferred—are then used to formulate a new set of geometric constraints. These constraints typically enforce properties such as making the two surfaces parallel and bringing them into contact. Each new constraint, solved in conjunction with existing ones, further contracts the object's feasible pose space towards a minimal valid subspace.

Final Pose Sampling and Validation. Once all constraints have been applied, we randomly sample a final pose from this subspace and validate it for 3D collisions and physical-semantic plausibility (*e.g.*, for an *in* relation, we validate the degree of enclosure via multi-directional ray-casting). Because our solving process is a deterministic accumulation of constraints, any pose sampled from this final subspace is guaranteed a priori to satisfy all non-collision-related geometric and spatial relations. This ensures a high success rate for the final validation step, avoiding costly cycles of blind rejection sampling.

To ensure physical plausibility, the fully instantiated scene undergoes a final refinement step via a brief dynamic simulation in Sapien [54]. This process yields a final 3D scene with an enhanced level of physical realism and stability. From this stable configuration, we additionally generate a detailed part-level contact graph by identifying all part pairs in close proximity (*e.g.*, $\leq$1mm).

Figure 4. **Gallery of PARSE-10K.**

4.2. Dataset Statistics and Analysis

Leveraging the PARSE framework's explicit modeling of part-level spatial relations, we construct PARSE-10K, a large-scale dataset comprising **10,000** unique indoor scenes, each annotated with a corresponding part-level contact graph. The dataset's compositional diversity is rooted in its rich asset library, which contains over **17,372** part-segmented and semantically-labeled assets across **132** object categories. Each scene is densely populated with an average of **49.9** objects, resulting in a high degree of physical plausibility and rich, part-level relational complexity. As showcased in Fig. 4, this explicit modeling enables the generation of intricate arrangements—such as precisely stacked objects, items leaning against surfaces, and complex container-content relationships—that are difficult to synthesize or annotate in existing datasets. Our comparative analysis, detailed in Tab. 1, positions PARSE-10K uniquely within the landscape of 3D scene datasets. PARSE-10K bridges this fundamental gap, providing a large-scale resource of scenes that are simultaneously physically grounded, compositionally diverse, and richly annotated with part-aware geometric relations.

5. Experiments on Spatial Tasks

This section demonstrates the broad utility of PARSE-10K in spatial understanding and generative tasks. First, leveraging its rich spatial relations and fine-grained part-level contacts, we benchmark state-of-the-art VLMs and propose targeted improvements to their spatial grounding and contact reasoning (Sec. 5.1). Second, the dataset's densely annotated, relation-rich scenes serve as a rigorous testbed for controllable and fidelity-preserving scene synthesis (Sec. 5.2).

5.1. VLM for Spatial Reasoning

Dataset construction. We synthesize a large collection of rendered scene images paired with part-level and object-level relation graphs. For each scene, we render multiple camera views and extract the subset of the scene graph corresponding to the objects and parts visible in that view. From these annotations, we construct three evaluation tasks. (1) *Visual Relation Multiple Choice Questions (MCQ)*: for a sampled relation triplet (two objects and their relation), we mark the objects on the image and present a multiple choice question, following protocols from several spatial-understanding benchmarks [11, 66]. Distractors are created by randomly replacing object and relation labels using a predefined object and relation vocabulary. (2) *Part-level Contact MCQ*: for a sampled visible part–part contact pair, we form a multiple choice question of the form "Part M of Object A contacts Part N of Object B"; the two objects are marked on the image, and distractors are generated by randomly replacing objects and parts using a prebuilt object-part mapping. (3) *Scene Graph Generation (SGG)*: given an image and the full set of candidate object names and relation types, the model must both localize all objects (2D bounding boxes and labels) and enumerate the relations among them; object entries contain the label and 2D bbox, and relations are reported as triplets.

Experimental setup. We evaluate several leading VLMs as baselines: GPT-5 [34], Gemini-2.5-Pro [46], Claude-Opus-4 [2], Robobrain2 [45], and Qwen3-VL [36]. These are compared against our model (denoted "Ours"),

Table 2. **Quantitative comparison with baselines.** We evaluate models across three tasks: visual relation MCQ, part-level contact MCQ, and scene graph generation (SGG). For SGG, each metric is reported in the format *With BBox Matching / No BBox Matching*. The *Avg.* column indicates the average number of relations generated per scene by each model.

Models	Visual Relation↑	Part-level Contact↑	Scene Graph Generation			
			Recall↑	Precision↑	F1 Score↑	Avg.
GPT-5	82.1	75.2	13.7/40.9	13.9/41.3	13.8/41.1	15.3
Gemini-2.5-Pro	85.0	75.6	40.5/43.4	48.6/52.0	44.2/47.3	12.9
Claude-Opus-4	80.3	73.2	8.0/33.7	12.7/53.7	9.8/41.4	9.7
Robobrain2.0	60.8	37.2	9.2/11.3	26.7/32.8	13.7/16.9	5.6
Qwen3-VL	86.2	60.4	26.0/29.6	46.0/52.4	33.2/37.9	8.7
Ours	**97.4**	**86.2**	**73.2/74.8**	**80.3/82.0**	**76.6/78.2**	14.1

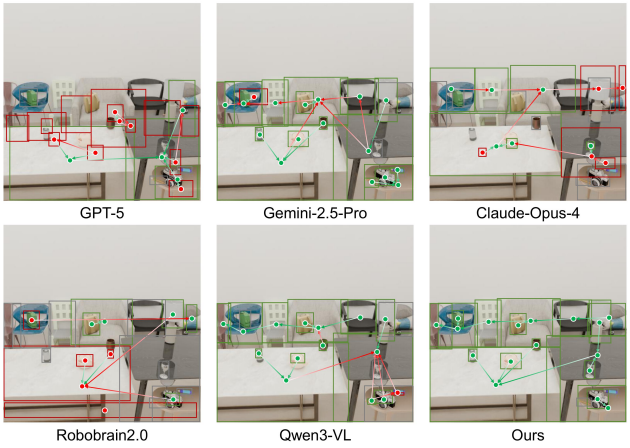

Figure 5. **Visualization of model-predicted graphs.** Green boxes indicate objects correctly matched by both label and grounding; red boxes indicate failed matches; gray boxes denote missed detections. Green arrows denote relations judged correct under the grounding-agnostic metric, red arrows denote incorrect relations.

which is fine-tuned from Qwen3-VL on our constructed dataset to study the efficacy of the targeted data. For the two MCQ tasks, we use accuracy on the selected option as the metric. To verify the model's generalization ability, we also add some manually labeled real images from the COCO [33] dataset to the test set. For Scene Graph Generation, we first normalize synonyms in model outputs, then perform class-wise matching between predicted and ground-truth boxes using an IoU-based Hungarian assignment [26]. A predicted relation is considered correct only if it connects the correctly matched object instances. Since models vary in the volume of relations they generate, we report the average number of predicted relations alongside precision, recall, and F1 score to provide a comprehensive evaluation.

Results and analysis. Tab. 2 summarizes the quantitative results on the MCQ tasks and aggregated part-contact scores. On the Visual Relation MCQ task, our fine-tuned model achieves 97.4%, substantially outperforming the baselines. On the Part-level Contact MCQ task, our model likewise leads with 86.2%. In the Scene Graph Generation task, the fine-tuned model substantially outperforms all baselines: it yields marked gains in object recognition, 2D localization, and the annotation of spatial relations. By explicitly training on PARSE-10K's dense, part-level supervision the model produces more complete and more accurately grounded relation sets compared to generalist VLMs. In contrast, models such as GPT-5 and Claude—while strong at high-level relational reasoning—exhibit weaker visual grounding and therefore suffer during the bbox-matching stage, which degrades their downstream relation scores. As an additional analysis, we also report relation accuracy under a grounding-agnostic metric (i.e., measuring relation correctness without requiring bbox matches) to separate pure relational reasoning from grounding performance. Fig. 5 visualizes the model-predicted graphs.

Our experiments demonstrate that fine-tuning on the constructed PARSE-10K dataset significantly enhances both visual grounding and relational reasoning. The fine-tuned model achieves the highest performance across all tasks, with notable improvements in MCQ accuracy and a substantial lead in Scene Graph Generation metrics. Compared to general-purpose VLMs such as GPT-5, Gemini-2.5-Pro, and Claude-Opus-4, our model produces more complete and accurately grounded scene graphs. The additional grounding-agnostic evaluation further confirms that the observed gains stem not only from improved visual localization but also from stronger relational understanding.

5.2. Scene Generation

Dataset construction. The goal of the scene generation task is to generate rotation, translation, and scale for each given object, with or without scene graph control, and then

Figure 6. **Scene generation comparison**. Left column: scenes generated by InstructScene trained on 3D-FRONT. Middle column: scenes generated by our method trained on PARSE-10K without PAG control. Right column: scenes generated by the model with PAG control.

combine them into a reasonable scene. PARSE-10K poses particular challenges for this task: scenes contain many objects, exhibit complex hierarchical relationships, and require precise object–object contacts. To capture geometric information, we encode each mesh with a Michelangelo [64] encoder and feed the resulting per-object geometry features to the network. We encode the PAG using CLIP [37] and convert its output into a relation embedding matrix. For training targets, we extract object poses from simulated scenes and use them as the denoising targets for the diffusion model.

Experimental setup. We build a graph-transformer-based diffusion network inspired by InstructScene [32]. At each denoising layer, the model fuses the object geometry features with the current noisy pose via cross-attention; the scene-graph control is injected into attention layers using a FiLM-style [35] modulation so that relational constraints can influence the pose refinement. We train and evaluate both conditioned and unconditioned variants (i.e., with and without PAG control) and follow standard diffusion schedules; detailed training hyperparameters and optimization schedules are provided in the Appendix. We present a qualitative comparison between scenes generated by the state-of-the-art method, InstructScene, trained on the 3D-FRONT [13] dataset, and those produced by our proposed method trained on the PARSE-10K dataset, under both PAG-conditioned and unconditioned settings. Furthermore, we conduct a user study to quantitatively evaluate the generated scenes in terms of their complexity, realism, and contact plausibility. A total of 20 participants were involved in the study, each evaluating 12 rendered scenes by selecting the one that best fit the given criterion.

Results and analysis. Qualitative comparisons (Fig. 6) show that training on PARSE-10K produces scenes with a higher object count and richer, more complex contacts than the baseline trained on 3D-FRONT. Conditioning on the scene graph yields scenes whose inter-object relations are more semantically coherent and physically plausible. A user study, summarized in Tab. 3, further quantifies these improvements. Owing to the high complexity and rich contact relationships of our dataset, learning its distribution without PAG conditioning often leads to unrealistic physics and unreasonable layout. Consequently, participants showed limited preference for scenes generated without PAG conditioning. Nevertheless, when conditioned on PAG, the model is able to generate scenes with a larger number of objects and finer contact details. Participants consistently preferred our PAG-conditioned PARSE-10K-trained models on measures of scene complexity, realism, and contact fidelity.

Table 3. **User study.** The table reflects the percentage of user votes for scenes generated from the corresponding model.

Method	Complexity↑	Realism↑	Contact Fidelity↑
InstructScene	7.5%	33.8%	28.8%
Ours(uncond)	45.0%	27.5%	26.3%
Ours(cond)	**47.5%**	**38.8%**	**45.0%**

In summary, our experiments show that the proposed PARSE-10K dataset facilitates the generation of more complex and realistic scenes. Both qualitative comparisons and quantitative user studies confirm that models trained on PARSE-10K produce scenes with higher object counts, richer contact relationships, and greater semantic and physical plausibility than those trained on previous datasets. These results demonstrate the value of our dataset in advancing contact-rich 3D scene-generating techniques.

6. Conclusion

We introduce PARSE, a part-centric framework that encodes geometric interactions between object parts through a Part-centric Assembly Graph and a Part-Aware Spatial Configuration Solver, enabling the synthesis of physically consistent 3D layouts. We also construct PARSE-10K, a large-scale dataset with dense part-level contact annotations that enhance spatial reasoning and 3D scene generation.

While PARSE and PARSE-10K advance part-level spatial modeling, several limitations remain. Relation definitions are complex and require part-specific coordinate reasoning, making PAG construction partially manual and sensitive to canonical poses. Future work will focus on learning part–part relations directly from geometry, developing more flexible contact representations, expanding the diversity of PARSE-10K, and integrating PARSE into embodied tasks for part-level planning and physically grounded manipulation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant W2431046, National Key R&D Program of China 2025YFA1309603, Central Guided Local Science and Technology Foundation of China YDZX20253100001001,and by MoE Key Lab of Intelligent Perceptionand Human-Machine Collaboration (ShanghaiTech University), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence. The experiments of this work were supported by HPC Platform of ShanghaiTech University.

References

[1] 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. 2

[2] Anthropic. Introducing claude (opus 4). https://www.anthropic.com/news/claude-4, 2025. Accessed: 2025-11-12. 6

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[4] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5664–5673, 2019. 2

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 3

[6] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1–26, 2021. 2

[7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3

[9] 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. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 3

[10] 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*, pages 13142–13153, 2023. 2

[11] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024. 6

[12] 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*, 36:18225–18250, 2023. 2, 3

[13] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2, 3, 5, 8

[14] Rao Fu, Zehao Wen, Zichen Liu, and Srinath Sridhar. Anyhome: Open-vocabulary generation of structured and textured 3d homes. In *European Conference on Computer Vision*, pages 52–70. Springer, 2024. 3

[15] 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*, pages 21295–21304, 2024. 3

[16] Lin Gao, Jia-Mu Sun, Kaichun Mo, Yu-Kun Lai, Leonidas J Guibas, and Jie Yang. Scenehgn: Hierarchical graph networks for 3d indoor scene generation with fine-grained geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8902–8919, 2023. 3

[17] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10323–10332, 2019. 2

[18] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1969–1978, 2019. 2

[19] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 3

[20] Ziyuan Jiao, Yida Niu, Zeyu Zhang, Song-Chun Zhu, Yixin Zhu, and Hangxin Liu. Sequential manipulation planning

on scene graph. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8203–8210. IEEE, 2022. 2

[21] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 2

[22] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 2

[23] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024. 5

[24] Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE transactions on cybernetics*, 50(12):4921–4933, 2019. 2

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2, 3

[26] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. 7

[27] Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, 2005. 2

[28] Alex X Lee, Coline Manon Devin, Yuxiang Zhou, Thomas Lampe, Konstantinos Bousmalis, Jost Tobias Springenberg, Arunkumar Byravan, Abbas Abdolmaleki, Nimrod Gileadi, David Khosid, et al. Beyond pick-and-place: Tackling robotic stacking of diverse shapes. In *5th Annual Conference on Robot Learning*, 2021. 2

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

[30] Xinghang Li, Di Guo, Huaping Liu, and Fuchun Sun. Embodied semantic scene graph generation. In *Conference on robot learning*, pages 1585–1594. PMLR, 2022. 2

[31] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7190–7199, 2021. 3

[32] Chenguo Lin and Yadong Mu. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 8

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 7

[34] OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2025-11-12. 2, 6

[35] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 8

[36] QwenLM. Qwen3-vl: Sharper vision, deeper thought, broader action. https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef, 2025. Accessed: 2025-11-12. 2, 6

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 8

[38] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21783–21794, 2024. 3

[39] Xingjian Ran, Yixuan Li, Linning Xu, Mulin Yu, and Bo Dai. Direct numerical layout generation for 3d indoor scene synthesis via spatial reasoning. *arXiv preprint arXiv:2506.05341*, 2025. 3

[40] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 2

[41] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8376–8384, 2019. 2

[42] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[43] Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29469–29478, 2025. 3

[44] 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*, pages 20507–20518, 2024. 3

[45] BAAI RoboBrain Team. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 6

[46] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261, 2025. 6

[47] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017. 2

[48] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 2

[49] Penghao Wang, Yiyang He, Xin Lv, Yukai Zhou, Lan Xu, Jingyi Yu, and Jiayuan Gu. Partnext: A next-generation dataset for fine-grained and hierarchical 3d part understanding. *arXiv preprint arXiv:2510.20155*, 2025. 2

[50] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1508–1517, 2020. 2

[51] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 2

[52] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023. 2

[53] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 2

[54] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020. 5

[55] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 2

[56] 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*, pages 16227–16237, 2024. 3

[57] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Trans. Graph.*, 44(4), 2025. 3

[58] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 3

[59] Huangyue Yu, Baoxiong Jia, Yixin Chen, Yandan Yang, Puhao Li, Rongpeng Su, Jiaxin Li, Qing Li, Wei Liang, Song-Chun Zhu, et al. Metascenes: Towards automated replica creation for real-world 3d scans. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1667–1679, 2025. 3, 5

[60] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems*, 36:30026–30038, 2023. 3

[61] Genghao Zhang, Yuxi Wang, Chuanchen Luo, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Furniscene: A large-scale 3d room dataset with intricate furnishing scenes. *arXiv preprint arXiv:2401.03470*, 2024. 5

[62] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 2

[63] Hang Zhao, Zherong Pan, Yang Yu, and Kai Xu. Learning physically realizable skills for online packing of general 3d shapes. *ACM Transactions on Graphics*, 42(5):1–21, 2023. 2

[64] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 8

[65] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *European conference on computer vision*, pages 211–229. Springer, 2020. 2

[66] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025. 6

[67] Mengqi Zhou, Xipeng Wang, Yuxi Wang, and Zhaoxiang Zhang. Roomcraft: Controllable and complete 3d indoor scene generation. *arXiv preprint arXiv:2506.22291*, 2025. 3

[68] Guoyu Zuo, Jiayuan Tong, Hongxing Liu, Wenbai Chen, and Jianfeng Li. Graph-based visual manipulation relationship reasoning network for robotic grasping. *Frontiers in Neurorobotics*, 15:719731, 2021. 2