Function2Scene: 3D Indoor Scene Layout from Functional Specifications

Ruiqi Wang[1], Qimin Chen[1], Daniel Ritchie[2], Angel X. Chang[1], Manolis Savva[1], Kai Wang[1,3], and Hao Zhang[1]

[1] Simon Fraser University
[2] Brown University
[3] ShanghaiTech University

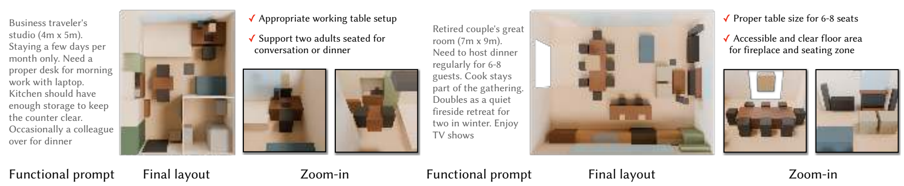

Fig. 1: We present Function2Scene, a framework for generating 3D indoor layouts from functional specifications. Given a detailed functional specification, our method decompose them into functional design constraints, which are then used to iteratively evaluate and refine a generated scene.

Abstract. Most text-driven 3D indoor scene synthesis methods generate rooms from object-centric prompts, asking what furniture should be placed rather than how the space is used. Yet in real interior design, a layout is judged by how well it supports its occupants, e.g., their activities and physical needs. We introduce Function2Scene, a framework for generating 3D indoor layouts from *functional* specifications, i.e., natural-language design briefs describing who will use a room and what they need to do there. Given such a specification, our system parses occupant personas and activities, derives a customized set of functional design constraints from a taxonomy of 17 criteria spanning spatial, ergonomic, activity, and environmental considerations, and uses these constraints to guide layout generation. Rather than relying on an LLM to directly produce a final scene, Function2Scene performs iterative evaluation and refinement through a tool-augmented check-and-repair loop, combining geometric measurements, LLM-based contextual reasoning, and VLM-based visual assessment. Experiments on 30 professionally written interior-design cases show that Function2Scene produces layouts that better satisfy functional requirements than recent LLM-based scene synthesis baselines, with our results preferred in 94.3% of pairwise comparisons. Our work reframes text-driven indoor scene synthesis from placing plausible objects to designing spaces that support human use.

1 Introduction

A furnished room is not only a collection of objects, but a proposal for how a space should be used. Furniture layout determines how people circulate, where they sit and look, what they can reach, and which activities can happen. A room can be visually and semantically plausible, yet still be a poor design: a desk facing direct window glare, a sofa blocking the path to the door, a wardrobe unreachable by a senior who uses it, or a child's play area hidden from their caregiver. Function is not a secondary attribute — it is one of the primary reasons behind any layout decision.

The computer graphics community has long recognized and studied this. Early scene generation works encoded design guidelines, such as clearance, grouping, alignment, accessibility, as explicit cost functions [33,63], but authoring such constraints required significant expertise and effort, making them hard to adapt to different scenarios. To address this limitation, the field pivoted to data-driven approaches, starting with pairwise spatial priors [14,63], and advancing through the deep learning era with increasingly expressive learned priors with convolutional neural networks [53], autoregressive transformers [37], and denoising diffusion [51, 64]. These work capture increasingly rich statistical patterns, but in doing so, make functional design knowledge more and more implicit and uninterpretable.

Entering the foundation model era, LLM-based systems [7, 12, 13, 46, 61] enabled flexible text conditioned scene synthesis. More recently, agentic pipelines also enable iterative scene updates to further improve scene quality [19,29,56,67]. Leveraging foundational priors, these systems support more open-ended scene generation. However, they largely inherit the learning era's "implicit" approach: an LLM directly generates objects, transforms relations, and even in agentic setups, the result is mostly checked only for visual quality and physical plausibility. Consistent with their goals, most LLM-based methods consume detailed object-centric prompts, for example, "a bedroom with a queen bed, two nightstands, and a dresser." They are not targeting more "functionality"-oriented prompts, for example, "a bedroom for a couple where one partner reads late while the other sleeps early." Yet, the latter describes a far more common starting point in real design practices [23, 36]. The irony is that LLMs are well suited to the two tasks that limited classical rule-based layout generation in the first place: parsing open-ended functional descriptions, and optimizing functional criteria that are cost prohibitive to manually specify.

We study 3D indoor scene layout from *functional specifications*: natural-language design briefs that describe who will use a space, what they will do in it, and what constraints their needs impose. Such briefs are closer to an interior design program than to a short text-to-scene command. For the scope of this work, we use professionally written room descriptions adapted from sources such as *Architectural Digest*. This problem formulation introduces unique challenges, as functional specifications are inherently high-level: instead of directly specifying objects and layout, they impose diverse and heterogeneous constraints over the indoor space: spatial constraints, ergonomic rules, activity patterns, environmental contexts, etc. Consequently, LLMs tasked to directly generate scenes from such functional specification not only struggle with producing scenes that meet the functional demands, but also frequently fail to produce meaningful scenes, without the availability of more explicitly worded prompts.

To address these challenges, we introduce *Function2Scene*, a framework that empowers LLMs for such functional scenarios. Given a functional specification, we begin by analyzing the "personas," i.e., who the occupants are and what specific needs they have, and the "activities," i.e., what the occupants do in the space. We then generate a detailed scene description, as well as a set of design guidelines drawn from a taxonomy of 17 criteria organized into four categories, Spatial, Ergonomic, Activity, and Environmental, grounded in interior-design literature [23,36], that conforms to the personas and their activities. This revisits the classical idea of codifying design rules for layout generation [24, 33, 63], but makes the selected rules customized to the specific functional specifications, a customization that is only possible thanks to the foundational knowledge of LLMs. Finally, we generate an initial layout and iteratively refine it through a check-and-repair loop to make the layout conform to the parsed design guidelines.

Our functionality-aware pipeline generates diverse scenes from real-world functional specifications, with results preferred over all baselines and ablations in 94.3% of pairwise comparisons in a crowd-sourced two-alternative forced choice (2AFC) perceptual study.

In summary, our contributions are:

- A functionality-first framing for 3D indoor layout generation, shifting the input from object-centric prompts to functional specifications, and exposing failure modes not addressed by existing scene synthesis methods.
- A design constraint taxonomy (Spatial, Ergonomic, Activity, Environmental) rooted in interior-design literature, together with an LLM-driven method for automatically customizing constraints to specific occupant personas and activities.
- An iterative layout generation framework that combines geometric measurements, LLM-based reasoning, and VLM-based visual assessment in a tool-augmented check-and-repair loop to generate high-quality, functionally valid layouts from functional specifications.

2 Related Works

Scene synthesis pre-LLMs. Indoor scene synthesis has long been studied in computer graphics. Early works leaned heavily on pre-specified design principles [33], simple statistical relationships [63] and hand-written programs [62], which all require heavy manual effort, and is not comprehensive enough to handle open-ended settings, even when massively scaled up [9, 41]. As a result, the field gradually shifted towards data-driven approaches [14, 22, 26], and become dominated by deep learning based approaches [3,25,27,37,43,51–53,66,69]. Such learning-based approaches require minimal manual effort, but generally do not learn explicit design principles, making them hard to adapt for our settings.

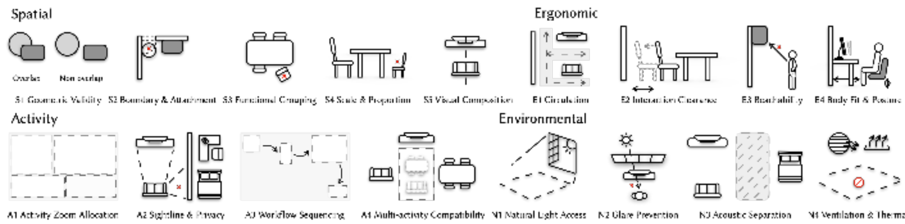

Fig. 2: Constraints Taxonomy. We organize interior design constraints into four categories: Spatial (S1–S5), Ergonomic (E1–E4), Activity (A1–A4), and Environmental (N1–N4), each illustrated with representative examples of how they shape furniture placement in a typical room layout.

LayoutEnhancer [24] attempts to address this limitation by directly injecting ergonomic principles into learned generative models, yet they also involve manual rule authoring that is hard to scale. A parallel line of work has argued that scene plausibility should be grounded in human use. SceneGrok [44] predicted where actions occur from observations, activity-centric synthesis [15] used such predictions to guide scene generation, and PiGraphs [45] learned joint models of human pose and scene geometry from interactions. Subsequent work conditioned layout on activity types [16,30,40] or optimized for human-aware navigation [48]. These methods established that activities can usefully guide object selection and placement, but they map activities directly to arrangements rather than to the underlying design constraints that determine whether a layout actually supports the activity.

LLM-based scene synthesis. While text-to-scene generation long predates [8,31] large language models, the rise of LLMs transformed the scale and flexibility of what language-conditioned systems can produce. LayoutGPT [13] showed that LLMs can directly predict object coordinates from open-vocabulary prompts, Holodeck [61] scaled this to full embodied environments with constraint satisfaction, and I-Design [7] added personalization from user preferences. Since then, the field has expanded rapidly: hierarchical and structured representations [34, 39, 49, 54, 55, 65, 68], deeper spatial reasoning via chain-of-thought [42] or VLM-guided search [4, 10], design-aware placement [6, 12, 18, 60], and agentic pipelines with iterative self-correction [19, 28, 29, 47, 56, 57]. Yet throughout this progression, the input prompt is predominantly about objects, relation and coordinates. Even systems that incorporate richer signals—personalized preferences [7, 59], physical interactability [58], or disentangled semantic-physical refinement [17, 35]—optimize primarily for spatial plausibility and visual coherence rather than for the functional, ergonomic, and environmental criteria that determine whether a room actually supports the activities it was designed for.

Scene evaluation and optimization. How a generated layout is evaluated shapes what it can become. Current evaluation frameworks assess geometric validity, semantic coherence, navigability, and collision avoidance [20, 50], and iterative optimization loops—VLM-based feedback [2, 21], multi-turn RL [67], differentiable VLM optimization [46], and VL-guided editing [5]—have made it possible to progressively refine layouts after initial generation [11]. Recent work has begun extending evaluation toward functional affordance grounding [32], but existing frameworks still lack systematic coverage of the human-centered criteria—ergonomic fit, activity support, environmental comfort—that govern how people actually use indoor spaces. Our framework addresses this by employing typed verification tools—geometric checks for measurable spatial properties, LLM queries for contextual semantic judgments, and VLM assessments for holistic visual quality—organized across four constraint categories to evaluate and iteratively refine layouts against functional design criteria.

3 Design Constraints

Professional interior designers compose layouts by following a rich set of principles and guidelines drawn from both design literature and practical experience [23, 36]. Prior work has formalized such guidelines as optimization criteria for automated layout generation [33], yet these criteria are applied uniformly regardless of who occupies the space or what they do in it.

In practice, layout requirements are not universal. A mobility-limited user and a frequent entertainer sharing the same room type will impose fundamentally different demands on furniture placement, clearance, and zone allocation. We capture this variability through **activity** $\times$ **persona** combinations derived from the user's functional specifications, which together determine which constraints are relevant and how their thresholds are parameterized for a given scene. We detail this process further in Section 4. We ground our constraints taxonomy in established interior design literature and organize constraints into four categories: Spatial, Ergonomic, Activity, and Environmental. We illustrate these constraints in Figure 2 and Table 1:

Spatial Forming the foundation upon which all other constraint categories depend, these constraints govern the placement and relational arrangement of furniture within the room. **Geometry Validity (S1)** requires that every furniture piece fit within the room boundary without overlapping adjacent objects, except where a containment or nesting relationship is explicitly defined between them; **Boundary & Attachment (S2)** specifies that floor-based items rest on the floor plane, wall-mounted objects attach to the correct surface, and large case goods align to the nearest wall; **Spatial Relationships (S3)** captures the grouping logic central to interior design practice, where functionally paired objects must be placed in proximity and with appropriate relative orientation; **Scale & Proportion (S4)** ensures furniture size is commensurate with room dimensions and neighboring objects, preventing pieces from dominating or failing to define the space; **Visual Composition (S5)** captures aesthetic principles such as focal point orientation, visual balance, and alignment, ensuring the arrangement reads as intentional rather than arbitrary.

Table 1: Evaluation tools setup. Each constraint is verified by one or more tools, color-coded by type: numeric/geometric tools compute quantitative measures directly from scene geometry, LLM query tools leverage language model reasoning over structured scene data, and VLM tools interpret rendered images. Tier indicates evaluation priority, where lower-tier constraints are verified first as prerequisites for higher-tier ones.

Category	Constraints	Tools	What it checks (Return)	Tier
Spatial	S1 Geometry Validity	boundary_check() bbox_collison()	within-wall containment (bool); pairwise overlap ratio (%)	T1
	S2 Boundary & Attachment	contact_check() wall_angle_check()	floor/ceiling/wall attachment (bool); object-to-wall angle (degrees)	T1
	S3 Spatial Relationships	object_exist() object_info()	object presence (bool); object size, location and orientation data (l, w, h, x, y, z, facing)	T2
	S4 Scale & Proportion	size_ratio() size_check()	object-to-room size ratio (%); LLM judgement on absolute size plausibility	T2
	S5 Visual Composition	visual_balance_check()	VLM judgement on top-down visual balance assessment	T5
Ergonomic	E1 Circulation	pathfinding() path_width()	the path (list of (x, z) world-coordinate waypoints) or null; minimum clearance (m) and bottleneck position (x, z)	T2
	E2 Interaction Clearance	articulation_zone() chair_clearance()	minimum clearance within swing arc (m); front and rear clearance distances from chair (m)	T4
	E3 Reachability	reach_check()	LLM judgement on reachability given user's attributes and object_info()	T4
	E4 Body Fit & Posture	posture_check()	LLM judgement on posture given user's attributes and object's size	T4
Activity	A1 Activity Zone	free_floor_area() object_in_zone() activity_support_check()	free area within the zone (m^2) ; LLM judgment on if related objects are correctly positioned within the zone; LLM judgment on if related objects can support the required activity well	T3
	A2 Sightlines & Privacy	inbetween_check()	LLM judgement on if objects blocks between two points based on object_info()	T3
	A3 Workflow Sequencing	total_path_length() workflow_check()	total path length across activity sequence (m); LLM judgment on workflow order	T5
	A4 Multi-activity Compat	multi_activity_check()	LLM judgement on if space supports multi-activity simultaneously reusing tools for A1	T5
Environmental	N1 Natural Light Access	window_obs_ratio()	object within proximity radius blocking window ratio (%)	T6
	N2 Glare Prevention	screen_window_info() glare_check()	angle and distance between screen and window; LLM judgment on glare risk	T5
	N3 Acoustic Separation	zone_distance() acoustic_check()	distance between two zones (m); LLM judgment on acoustic risk	T5
	N4 Ventilation & Thermal	vent_obs_ratio() distance_check()	object within proximity radius blocking vent ratio; safe distance guarantee (bool)	T6

Ergonomic Where spatial constraints establish what fits, ergonomic constraints ensure the space can be safely and comfortably navigated and used, with thresholds parameterized to the physical needs and abilities of the target persona. **Circulation (E1)** requires that primary pathways maintain a minimum clear width for unobstructed passage, with thresholds elevated for users relying on mobility aids; **Interaction Clearance (E2)** ensures that all articulated elements have full action zones free of obstruction, and that seating has adequate pull-out space behind it; **Reachability (E3)** constrains frequently used objects and controls to fall within the user's operational height range, accounting for seated versus standing use and any physical limitations; **Body Fit & Posture (E4)** requires work surface heights, seat dimensions, and monitor distances to conform to anthropometric standards for sustained comfortable use.

Activity Beyond physical validity and ergonomic fit, a layout must support the specific tasks the user performs in the space, which is the primary vehicle through

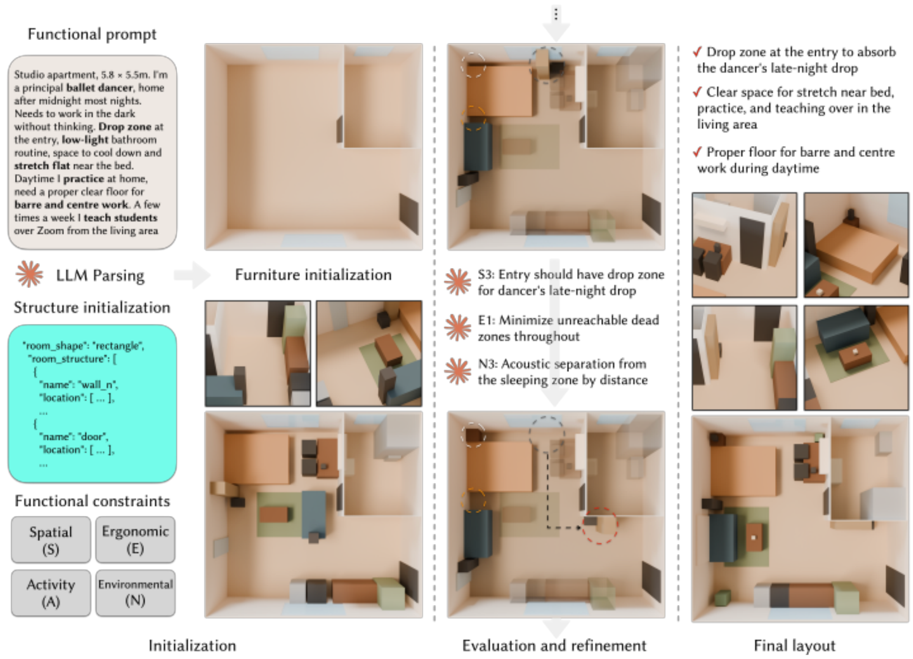

Fig. 3: Overall Pipeline: Given a functional prompt, FUNCTION2SCENE generates 3D indoor scene layout through iteratively evaluation and refinement based on functional constraints.

which the **activity** $\times$ **persona** combination shapes the constraint pool. **Zone Allocation (A1)** ensures each primary activity has a dedicated zone of sufficient size, equipped with the relevant furniture and kept clearable and accessible when needed; **Multi-Activity Compatibility (A2)** demands the layout support zone transformation without full reorganization when a space serves more than one purpose; **Sightlines and Privacy (A3)** encodes directional requirements, where certain tasks demand unobstructed views toward an entry or a child's play area, while others call for visual shielding from the rest of the room; **Workflow Sequencing (A4)** arranges objects to match their order of use, avoiding backtracking and cross-circulation.

Environmental Furniture placement shapes not only how a room is used but how it feels, through its effect on light, sound, and thermal comfort. **Natural Light Access (N1)** asks that furniture arrangement preserve daylight reach into primary activity zones; **Glare Prevention (N2)** guards against screens and resting surfaces receiving direct window light during typical use hours; **Acoustic Separation (N3)** requires for noise-generating activities to be spatially buffered from quiet zones through distance, furniture mass, or deliberate zone boundaries; **Ventilation & Thermal Comfort (N4)** ensures furniture does not block windows, and that temperature-sensitive activities are positioned away from cold drafts or direct heat sources.

Constraints are assigned to one of six priority tiers (T1 to T6) as indicated in Table 1, where lower tiers must be satisfied before higher tiers are considered. Details of how each constraint is evaluated are described in Section 4.2.

4 Method

Our system is a function-driven interior design agent that takes a *functional prompt* as input, which captures the user's living needs as input distilled through multi-turn conversations, and produces a furniture layout tailored to how they actually inhabit the space, guided by design principles derived from those needs.

As illustrate in Figure 3, the pipeline proceed through two main stages: (1) the *Initialization* stage, which parses the user's functional prompt into a parsed scene description and a list of functional constraints, constructs the room structure with user verification, and generates an initial furniture layout; and (2) the *Constraints-based Evaluation and Refinement* stage, which iteratively evaluates the layout against constraints in priority order using specialized tools, and applies targeted adjustments guided by design principles. Following a similar prompting strategy in Holodeck [61], an LLM prompt is designed for each stage throughout the pipeline, consisting of a task description, an ouput format specification, and a one-shot example.

In this section, we first introduce the initialization stage, which parses the user's functional prompt into an initial layout and a set of functional constraints. We then provide a detailed description of the constraints-based layout evaluation and refinement stage, which iteratively optimizes the layout toward functional and spatial coherence.

4.1 Initialization

The initialization stage serves as the foundation of FUNCTION2SCENE, transforming a user-provided functional description of a scene into a structured 3D layout that serves as the starting point for the evaluation and refinement steps. It consists of three steps: parsing, structure generation, and furniture initialization.

Parsing. Given the raw functional description of a scene, the parser extracts: (a) a structured set of constraints grounded in the 4 category constraint taxonomy defined in Section 3; and (b) a parsed scene description that serves as an LLM-friendly reformulation of the original input, analyzed and restructured with functional constraints considerations in mind to produce a more rational and unambiguous description in a style closer to that of current text-to-scene generation models. Together, these two outputs ensure that the original functional intent is interpreted once and propagated consistently throughout the pipeline.

Room Structure Generation. Given the parsed description, Function2Scene initializes the room structure through a set of LLM prompts. The room structure is encoded in a custom JSON-based Domain-Specific Language (DSL), as shown in the cyan box in Figure 3, which provides a well-defined representation of walls, floors, ceilings, doors and windows, with explicit geometric attributes such as dimension and coordinates, and semantic attributes such as orientation encoded as a facing direction. This structural output is visualized for user to verify, reflecting the natural client-designer workflow where the room structure is reviewed and approved before furnishing begins. If corrections are needed, the user can easily refine the geometry through natural language prompting, making the editing process intuitive and accessible.

Furniture Initialization. With the verified empty room and the parsed description, Function2Scene directly generates an initial layout. However, as we demonstrate in our evaluation, LLM-generated layouts at this stage are fundamentally limited in their spatial reasoning: objects may overlap, violate functional adjacency requirements, or produce configurations that are physically plausible but practically unusable, as shown in the initialization stage of Figure 3. The furniture initialization therefore serves as a starting point rather than a final result, motivating the constraints-based evaluation and refinement stage that follows.

4.2 Constraints-based Evaluation and Refinement

Prior works have explored constraint-based layout generation, most notably Holodeck [61], which defines a set of spatial relational constraints (e.g. *in front of*, *near*, *face to*) and optimizes object placements to satisfy them. While effective for basic spatial arrangement, such approaches are restricted to geometric relations between objects, leaving ergonomic, activity, and environmental demands unaddressed.

Function2Scene extends this idea to a richer, 4-category constraint set derived directly from the user's functional input, as detailed in Section 3. Rather than optimizing all constraints simultaneously through a solver, we adopt an LLM-driven iterative evaluation loop that assesses constraints sequentially in priority order, as specified in Table 1.

Constraint Evaluation. For each constraint, specialized tools are invoked to retrieve structured spatial information, which the LLM interprets to determine whether the constraint is satisfied. Specifically, the agent first interprets the constraint description in the context of the current layout state, then selects and invokes the appropriate tool(s) such as `pathfinding()`, `visual_balance_check()`, or `posture_check()` to retrieve varied forms of feedback, ranging from numeric measurements and traversal paths to natural language assessments from numeric algorithm, VLM or LLM-based tools. The LLM then interprets these results against the constraint requirements and produces a justification along with a proposed refinement step if the constraint is not met.

Layout Refinement. For each unsatisfied constraint, the LLM generates a targeted refinement action based on the justification produced in the evaluation step. Refinement guidance is grounded in a set of design principles [36] covering both universal spatial standards and room-specific human factor recommendations. Universal principles include maintaining a minimum `36''` primary circulation path and preserving door swing clearances. Room-specific standards further constrain the refinement: in the bedroom, a minimum `2'0''-3'0''` side clearance around the bed is required for access and bed-making; in the dining room, at least `3'4''` must be preserved behind seated diners for service circulation; and in the living room, conversation groups should be arranged within 8 feet of each other for comfortable interaction. These standards inform how furniture should be repositioned, reoriented, or resized. Each adjustment is applied locally to avoid disrupting already-satisfied constraints, and the affected constraint is re-evaluated before the agent proceeds to the next one.

Termination. Constraints are evaluated sequentially in priority order. If resolving a constraints would introduce violations in higher-priority constraints that have already been satisfied, the constraint is skipped. Once all constraints have been assessed, the Tier 1 spatial constraints are re-evaluated to verify that no adjustments made in later stages have compromised foundational layout quality. The final layout is then returned as output.

5 Results and Evaluation

In this section, we present qualitative and quantitative results demonstrating the ability of our method in generating visually and functionally plausible indoor scenes given functionality-focused natural language prompts. We compare against three representative LLM-based layout generation methods: Holodeck [61], iDesign [7], and LayoutVLM [46], which cover the current landscape of language-driven indoor scene synthesis from persona-aware generation to open-ended spatial instruction following. Our evaluation focuses on how well each approach captures persona-specific functional requirements and lifestyle-driven spatial organization that standard benchmarks often overlook.

5.1 Data

We curated 30 real interior design cases from *Architectural Digest* [1], an internationally recognized magazine and authority on interior design. Each case describes a distinct room designed around a specific occupant persona, resulting in a diverse dataset spanning 10 room types, including bedrooms, kitchens, living rooms, dining rooms, studios/ateliers, a home library, a guestroom, a nursery, a great room, and a mezzanine; 30 unique personas ranging from a retired couple and a chef to a drag queen, a child with autism, and a YouTuber. This breadth ensures coverage of varied functional needs, aesthetic preferences, lifestyle contexts, and demographic profiles.

5.2 Perceptual Study

To evaluate how well different layout generation methods satisfy the functional requirements specified in user scene prompts, we conducted a two-alternative forced-choice (2AFC) perceptual study. We recruited 30 participants through Prolific [38], with diverse backgrounds prior experience with AI evaluation tasks. Each participant completed 30 scene comparisons, where each comparison presented a room brief alongside rendered images of two layouts in randomized order, and selected the more functional layout for the described occupants. We also randomly insert 5 attention checks that compares against randomly generated, implausible layouts, and filter out participants who fail any of these checks.

We have a total of 10 comparison conditions (6 baselines, 4 ablations), resulting in 3 answers per scene and condition. For each comparison, we aggregate responses across all valid participants who saw that pair and report the proportion of selections in favour of our method. To reduce evaluation time and focus judgments on layout quality rather than scene-level details, participants were instructed to prioritize structural validity, such as furniture blocking doorways or objects extending outside boundaries before considering brief-specific criteria.

Table 2 shows the results of this experiment, and Figure 4 shows some qualitative comparisons between generated layouts. Overall, participants preferred layouts generated using our method across all baselines and prompt conditions, with an aggregate preference rate of 94.3%. Against Holodeck, our method was preferred in 92.2% and 88.9% of trials under functional and parsed prompts respectively. Against iDesign, preference rates reached 94.4% and 98.9%, with the parsed condition yielding the highest score across all comparisons. Against LayoutVLM, our method was preferred in 96.7% and 94.4% of trials. Figure 1 and Figure 5 provides a closer look at our generated layouts, highlighting fine-grained constraint satisfaction across spatial and functional requirements. Furthermore, Table 3 presents ablation study results examining the contribution of each component in our pipeline. Notably, retaining iterative updates without evaluation tools performs worst than removing both, indicating that iterative refinement is counterproductive without grounded spatial feedback to guide it. Additionally, prompt format has negligible effect when tools are absent, confirming that richer constraint representations only enforces their benefit when paired with the tool set that measures them. These results establish evaluation tools as the critical enabler of our pipeline.

6 Conclusion

In this paper, we presented Function2Scene, a framework for generating indoor layouts from functional specifications. By focusing on functionality, we take a first step towards designing a LLM-driven scene generation pipeline that better suits real interior design workflows, demonstrating that a well-designed taxonomy of functional design principles, combined with a LLM-driven iterative pipeline, can produce higher quality, more functional scenes than those generated by prior works.

Table 2: 2AFC study results comparison our method with baselines. Each baseline is ran with both the original functional prompt and our parsed scene description and constraints.

Method	Prompt	% Ours preferred
Holodeck [61]	Functional	92.2
	Parsed	88.9
iDesign [7]	Functional	94.4
	Parsed	98.9
LayoutVLM [46]	Functional	96.7
	Parsed	94.4
Overall	—	94.3

Table 3: 2AFC study results comparing against ablations that uses different input and generation strategy.

Prompt format	Iterative update	Evaluation Tools	%Ours preferred
Functional	No	No	83.3
Parsed	No	No	83.3
Functional	Yes	No	78.9
Parsed	Yes	No	80.0
Parsed	Yes	Yes	*Ours*

There exist many opportunities in further extending our method: as stated in the introduction, our method starts from a professionally written, detailed functional specification. Real design workflows, however, usually begin with user demands that are much vaguer and shorter: users need designers' help to discover, articulate, and refine their needs through multiple rounds of conversation and feedback. A conversational interface that helps non-expert users arrive at a detailed functional specification would complete the upstream half of the design workflow that feeds into our method; while our constraint taxonomy provides a general framework for evaluating functionality, our current verification protocol heavily relies on basic numeric checks and LLM queries. They could be made much more powerful with more constraint-specific tools, for example, embodied simulation with articulated models, physically accurate lighting and acoustic estimations, or a domain-specific language for expressing distance and dimension requirements in more semantic manners; more broadly, our framework currently operates within a fixed architectural shell in residential environments, co-optimizing room shape, openings, and partitions alongside furniture placement would better capture the full scope of interior design.

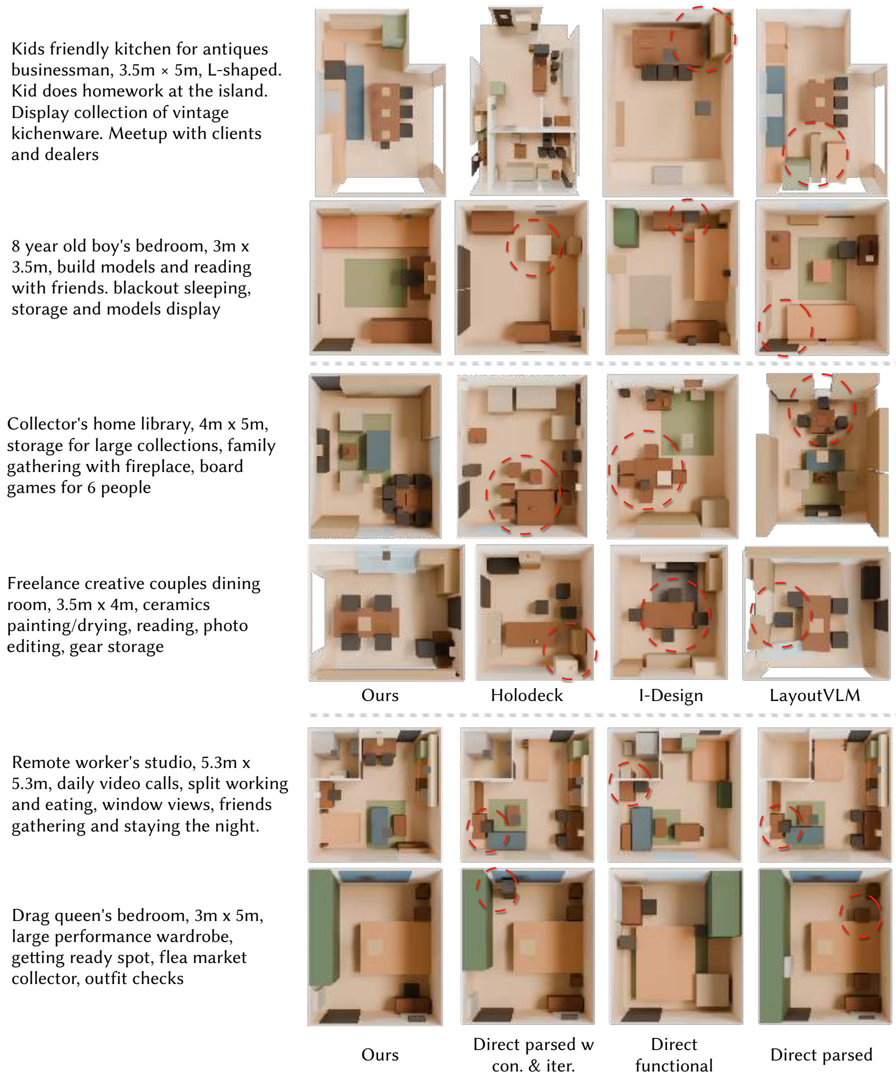

Fig. 4: Qualitative comparisons of our method against various comparison conditions. Top two rows: baselines with original functional prompts; middle two rows: baselines with our parsed specifications; bottom two rows: ablations, from left to right: w/ parsed input and iterative refinement, with original prompt and no iterative refinement, with parsed input and no iterative refinement.

Fig. 5: Functional scenes generated by our method, along with zoomed in highlights. The input prompts are truncated due to space constraints. Please refer to the supplementary materials for all qualitative results, along with visualization of all intermediary optimization steps.

References

1. Architectural Digest: Room tours & interior design case studies. https://www.architecturaldigest.com (2026), accessed: 2026 10
2. Asano, Y., Kondo, N., Fushimi, T., Ochiai, Y.: From geometry to culture: An iterative vlm layout framework for placing objects in complex 3d scene contexts. arXiv preprint arXiv:2503.23707 (2025) 5
3. Bai, T., Bai, W., Chen, D., Wu, T., Li, M., Ma, R.: Freescene: Mixed graph diffusion for 3d scene synthesis from free prompts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5893–5903 (2025) 3
4. Berdoz, F., Lanzendörfer, L.A., Tuninga, N., Wattenhofer, R.: Text-to-scene with large reasoning models. arXiv preprint arXiv:2509.26091 (2025) 4
5. Bian, Z., Ren, R., Yang, Y., Callison-Burch, C.: Holodeck 2.0: Vision-language-guided 3d world generation with editing. arXiv preprint arXiv:2508.05899 (2025) 5
6. Bucher, M.J., Armeni, I.: Respace: Text-driven 3d scene synthesis and editing with preference alignment. arXiv preprint arXiv:2506.02459 (2025) 4
7. Çelen, A., Han, G., Schindler, K., Van Gool, L., Armeni, I., Obukhov, A., Wang, X.: I-design: Personalized llm interior designer. In: European Conference on Computer Vision. pp. 217–234. Springer (2024) 2, 4, 10, 12
8. Chang, A., Monroe, W., Savva, M., Potts, C., Manning, C.D.: Text to 3d scene generation with rich lexical grounding. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 53–62 (2015) 4
9. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., Mottaghi, R.: Procthor: Large-scale embodied ai using procedural generation. Advances in Neural Information Processing Systems **35**, 5982–5994 (2022) 3
10. Deng, W., Qi, M., Ma, H.: Global-local tree search in vlms for 3d indoor scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8975–8984 (2025) 4
11. Feng, H., Niu, Y., Huang, Z., Sun, Y.T., Guo, C., Peng, Y., Sheng, L.: Repurposing 3d generative model for autoregressive layout generation. arXiv preprint arXiv:2604.16299 (2026) 5
12. Feng, W., Zhou, H., Liao, J., Cheng, L., Zhou, W.: Casagpt: cuboid arrangement and scene assembly for interior design. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29173–29182 (2025) 2, 4
13. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. Advances in Neural Information Processing Systems **36**, 18225–18250 (2023) 2, 4
14. Fisher, M., Ritchie, D., Savva, M., Funkhouser, T., Hanrahan, P.: Example-based synthesis of 3d object arrangements. ACM Transactions on Graphics (TOG) **31**(6), 1–11 (2012) 2, 3
15. Fisher, M., Savva, M., Li, Y., Hanrahan, P., Nießner, M.: Activity-centric scene synthesis for functional 3d scene modeling. ACM Transactions on Graphics (TOG) **34**(6), 1–13 (2015) 4
16. Fu, Q., Chen, X., Wang, X., Wen, S., Zhou, B., Fu, H.: Adaptive synthesis of indoor scenes via activity-associated object relation graphs. ACM Transactions on Graphics (TOG) **36**(6), 1–13 (2017) 4

17. Gao, J., Zhou, D., Liang, M., Liu, L., Fu, C.W., Hu, X., Heng, P.A.: Disco-layout: Disentangling and coordinating semantic and physical refinement in a multi-agent framework for 3d indoor layout synthesis. arXiv preprint arXiv:2510.02178 (2025) 4
18. Gupta, K., Mehta, I., Wang, K., Chua, N., Krishna, A., Deng, Y., Ramamoorthi, R., Chandraker, M.: Interioragent: Llm agent for interior design aware 3d layout generation. In: Thirteenth International Conference on 3D Vision (2026) 4
19. He, Y., Yu, K., Zwicker, M.: Sceneorchestra: Efficient agentic 3d scene synthesis via full tool-call trajectory generation. arXiv preprint arXiv:2604.19907 (2026) 2, 4
20. Hwangbo, G., Chae, H., Kang, M., Ju, H., Oh, S., Yeo, J.: Lego-eval: Towards fine-grained evaluation on synthesizing 3d embodied environments with tool augmentation. arXiv preprint arXiv:2511.03001 (2025) 4
21. Jiang, H., Zhang, D., Weng, D., Song, W., Duh, H.B.L.: Hog-layout: Hierarchical 3d scene generation, optimization and editing via vision-language models. arXiv preprint arXiv:2604.10772 (2026) 5
22. Kermani, Z.S., Liao, Z., Tan, P., Zhang, H.: Learning 3d scene synthesis from annotated rgb-d images. In: Computer Graphics Forum. vol. 35, pp. 197–206. Wiley Online Library (2016) 3
23. Kilmer, R., Kilmer, W.O.: Designing interiors. John Wiley & Sons (2024) 2, 3, 5
24. Leimer, K., Guerrero, P., Weiss, T., Musialski, P.: Layoutenhancer: Generating good indoor layouts from imperfect data. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022) 3, 4
25. Li, M., Patil, A.G., Xu, K., Chaudhuri, S., Khan, O., Shamir, A., Tu, C., Chen, B., Cohen-Or, D., Zhang, H.: Grains: Generative recursive autoencoders for indoor scenes. ACM Transactions on Graphics (TOG) **38**(2), 1–16 (2019) 3
26. Liang, Y., Zhang, S.H., Martin, R.R.: Automatic data-driven room design generation. In: International Workshop on Next Generation Computer Animation Techniques. pp. 133–148. Springer (2017) 3
27. Lin, C., Mu, Y.: Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. arXiv preprint arXiv:2402.04717 (2024) 3
28. Liu, X., Tai, Y.W., Tang, C.K.: Agentic 3d scene generation with spatially contextualized vlms. arXiv preprint arXiv:2505.20129 (2025) 4
29. Luo, J., Tang, J., Lu, R., Zeng, G.: Sceneassistant: A visual feedback agent for open-vocabulary 3d scene generation. arXiv preprint arXiv:2603.12238 (2026) 2, 4
30. Ma, R., Li, H., Zou, C., Liao, Z., Tong, X., Zhang, H.: Action-driven 3d indoor scene evolution. ACM Trans. Graph. **35**(6), 173–1 (2016) 4
31. Ma, R., Patil, A.G., Fisher, M., Li, M., Pirk, S., Hua, B.S., Yeung, S.K., Tong, X., Guibas, L., Zhang, H.: Language-driven synthesis of 3d scenes from scene databases. ACM Transactions on Graphics (TOG) **37**(6), 1–16 (2018) 4
32. Maillard, L., Engelmann, F., Durand, T., Pan, B., You, Y., Litany, O., Guibas, L., Ovsjanikov, M.: Sceneteract: Agentic functional affordances and vlm grounding in 3d scenes. arXiv preprint arXiv:2603.29798 (2026) 5
33. Merrell, P., Schkufza, E., Li, Z., Agrawala, M., Koltun, V.: Interactive furniture layout using interior design guidelines. ACM transactions on graphics (TOG) **30**(4), 1–10 (2011) 2, 3, 5
34. Öcal, B.M., Tatarchenko, M., Karaoğlu, S., Gevers, T.: Sceneteller: Language-to-3d scene generation. In: European Conference on Computer Vision. pp. 362–378. Springer (2024) 4
35. Pan, Z., Liu, H.: Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. arXiv preprint arXiv:2503.18470 (2025) 4

36. Panero, J.: Anatomy for interior designers. Whithey Library of Design (1962) 2, 3, 5, 10
37. Paschalidou, D., Kar, A., Shugrina, M., Kreis, K., Geiger, A., Fidler, S.: Atiss: Autoregressive transformers for indoor scene synthesis. Advances in neural information processing systems **34**, 12013–12026 (2021) 2, 3
38. Prolific: Easily collect high quality data from real people. https://www.prolific.com/ (2026), accessed: 2026 11
39. Pun, H.I.D., Tam, H.I.I., Wang, A.T., Huo, X., Chang, A.X., Savva, M.: Hsm: Hierarchical scene motifs for multi-scale indoor scene generation. arXiv preprint arXiv:2503.16848 (2025) 4
40. Qi, S., Zhu, Y., Huang, S., Jiang, C., Zhu, S.C.: Human-centric indoor scene synthesis using stochastic grammar. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5899–5908 (2018) 4
41. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., et al.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21783–21794 (2024) 3
42. Ran, X., Li, Y., Xu, L., Yu, M., Dai, B.: Direct numerical layout generation for 3d indoor scene synthesis via spatial reasoning. arXiv preprint arXiv:2506.05341 (2025) 4
43. Ritchie, D., Wang, K., Lin, Y.a.: Fast and flexible indoor scene synthesis via deep convolutional generative models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6182–6190 (2019) 3
44. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Scenegrok: Inferring action maps in 3d environments. ACM transactions on graphics (TOG) **33**(6), 1–10 (2014) 4
45. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. ACM Transactions On Graphics (TOG) **35**(4), 1–12 (2016) 4
46. Sun, F.Y., Liu, W., Gu, S., Lim, D., Bhat, G., Tombari, F., Li, M., Haber, N., Wu, J.: Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29469–29478 (2025) 2, 5, 10, 12
47. Sun, F.Y., Wu, S., Jacobsen, C., Yim, T., Zou, H., Zook, A., Li, S., Chou, Y.H., Can, E., Wu, X., et al.: 3d-generalist: Self-improving vision-language-action models for crafting 3d worlds. arXiv preprint arXiv:2507.06484 (2025) 4
48. Sun, J.M., Yang, J., Mo, K., Lai, Y.K., Guibas, L., Gao, L.: Haisor: Human-aware indoor scene optimization via deep reinforcement learning. ACM Trans. Graph. (2023), just Accepted 4
49. Tam, H.I.I., Pun, H.I.D., Wang, A.T., Chang, A.X., Savva, M.: SceneMotifCoder: Example-driven Visual Program Learning for Generating 3D Object Arrangements. In: Proceedings of the IEEE Conference on 3D Vision (3DV) (2025) 4
50. Tam, H.I.I., Pun, H.I.D., Wang, A.T., Chang, A.X., Savva, M.: Sceneeval: Evaluating semantic coherence in text-conditioned 3d indoor scene synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7355–7365 (2026) 4
51. Tang, J., Nie, Y., Markhasin, L., Dai, A., Thies, J., Nießner, M.: Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20507–20518 (2024) 2, 3

52. Wang, K., Lin, Y.A., Weissmann, B., Savva, M., Chang, A.X., Ritchie, D.: Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. ACM Transactions on Graphics (TOG) **38**(4), 1–15 (2019) 3
53. Wang, K., Savva, M., Chang, A.X., Ritchie, D.: Deep convolutional priors for indoor scene synthesis. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018) 2, 3
54. Wang, X., Wang, Y., Zhou, M., Fan, J., Zhang, Z.: Hlg: Comprehensive 3d room construction via hierarchical layout generation. arXiv preprint arXiv:2508.17832 (2025) 4
55. Wu, Q., Iliash, D., Ritchie, D., Savva, M., Chang, A.X.: Diorama: Unleashing zero-shot single-view 3d indoor scene modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8896–8907 (October 2025) 4
56. Xia, H., Li, X., Li, Z., Ma, Q., Xu, J., Liu, M.Y., Cui, Y., Lin, T.Y., Ma, W.C., Wang, S., Song, S., Wei, F.: Sage: Scalable agentic 3d scene generation for embodied ai. arXiv preprint arXiv:2602.10116 (2026) 2, 4
57. Yang, Y., Jia, B., Zhang, S., Huang, S.: Sceneweaver: All-in-one 3d scene synthesis with an extensible and self-reflective agent. arXiv preprint arXiv:2509.20414 (2025) 4
58. Yang, Y., Jia, B., Zhi, P., Huang, S.: Physcene: Physically interactable 3d scene synthesis for embodied ai. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 4
59. Yang, Y., Lu, J., Zhao, Z., Luo, Z., Yu, J.J., Sanchez, V., Zheng, F.: Llplace: The 3d indoor scene layout generation and editing via large language model. arXiv preprint arXiv:2406.03866 (2024) 4
60. Yang, Y., Luo, Z., Ding, T., Lu, J., Gao, M., Yang, J., Sanchez, V., Zheng, F.: Optiscene: Llm-driven indoor scene layout generation via scaled human-aligned data synthesis and multi-stage preference optimization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 4
61. Yang, Y., Sun, F.Y., Weihs, L., VanderBilt, E., Herrasti, A., Han, W., Wu, J., Haber, N., Krishna, R., Liu, L., 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 (2024) 2, 4, 8, 9, 10, 12
62. Yeh, Y.T., Yang, L., Watson, M., Goodman, N.D., Hanrahan, P.: Synthesizing open worlds with constraints using locally annealed reversible jump mcmc. ACM Transactions on Graphics (TOG) **31**(4), 1–11 (2012) 3
63. Yu, L.F., Yeung, S.K., Tang, C.K., Terzopoulos, D., Chan, T.F., Osher, S.J.: Make it home: Automatic optimization of furniture arrangement. ACM Trans. Graph. **30**(4), 86 (2011) 2, 3
64. Zhai, G., Örnek, E.P., Wu, S.C., Di, Y., Tombari, F., Navab, N., Busam, B.: Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. Advances in Neural Information Processing Systems **36**, 30026–30038 (2023) 2
65. Zhang, Y., Li, Z., Zhou, M., Wu, S., Wu, J.: The scene language: Representing scenes with programs, words, and embeddings. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24625–24634 (2025) 4
66. Zhang, Z., Yang, Z., Ma, C., Luo, L., Huth, A., Vouga, E., Huang, Q.: Deep generative modeling for scene synthesis via hybrid representations. ACM Transactions on Graphics (TOG) **39**(2), 1–21 (2020) 3

67. Zhao, Y., Sun, S., Zhang, M., Shi, Y., Yang, X., Bian, J.: Scenerevis: A self-reflective vision-grounded framework for 3d indoor scene synthesis via multi-turn rl. arXiv preprint arXiv:2602.09432 (2026) 2, 5
68. Zhou, M., Wang, X., Wang, Y., Zhang, Z.: Roomcraft: Controllable and complete 3d indoor scene generation. arXiv preprint arXiv:2506.22291 (2025) 4
69. Zhou, Y., While, Z., Kalogerakis, E.: Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7384–7392 (2019) 3