LIVED-IN BY DESIGN: BEHAVIOR-GROUNDED 3D SCENE GENERATION VIA FUNCTIONAL-GROUP ASSEMBLY

Anonymous authors
Paper under double-blind review

ABSTRACT

We present *Lived-In by Design*, a framework for behavior-grounded agentic 3D scene generation by carrying everyday-activity context from prompt interpretation to room-scale layout. Unlike existing methods that construct scenes primarily from object inventories and pairwise spatial relations, our method models the functional dependencies and human-use constraints that organize objects in inhabited environments. Specifically, we first design a behavior-to-asset stage to convert plausible activities generated by agent into object requirements, quantities, shared resources, and functional relations. To preserve these relations during construction, we further introduce functional-group assembly, which organizes related objects into behavior-dependent local configurations. Finally, we design a lived-in scene layout stage to place these groups jointly at room scale while considering architectural constraints, accessibility, and human usage space. Evaluations on 179 room-level prompts using automatic metrics, qualitative comparisons, controlled ablations, and human judgments demonstrate improved functional organization and lived-in plausibility.

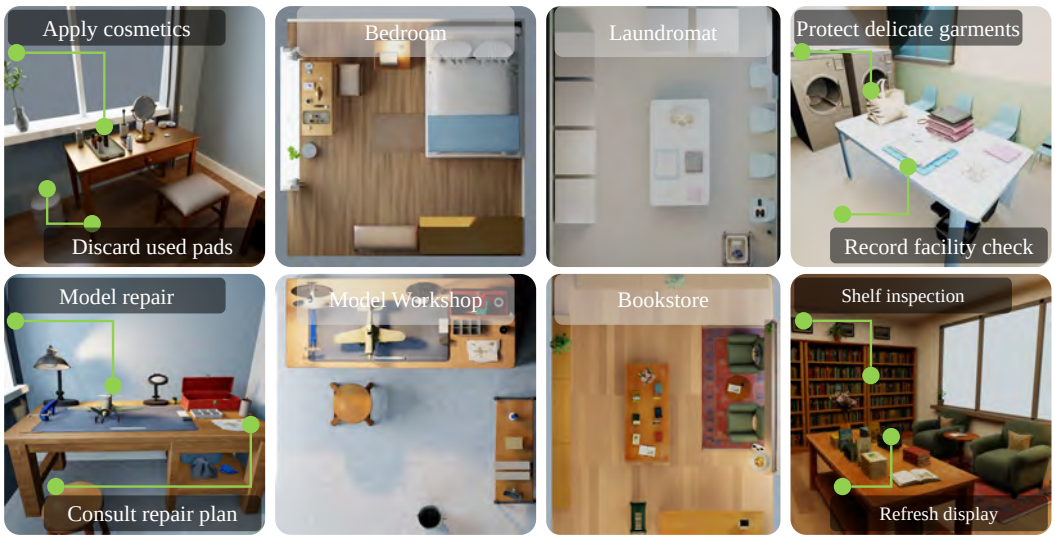

Figure 1: Behavior-grounded everyday context across four indoor scenes. The labeled perspective views identify activity-specific object requirements for applying cosmetics and discarding used pads, protecting delicate garments and recording facility checks, model repair and consulting a repair plan, and shelf inspection and refreshing a display. The paired plan views show the corresponding bedroom, laundromat, model-workshop, and bookstore organizations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Agentic 3D scene generation provides a scalable interface for constructing structured and editable 3D environments from natural-language descriptions, with applications in embodied AI, robot simulation, spatial intelligence, and virtual content creation (Tam et al., 2026b; Yang et al., 2024a). Recent systems coordinate language reasoning, asset retrieval or generation, spatial arrangement, and iterative visual or geometric feedback to construct increasingly complex indoor scenes (Feng et al., 2023; Yang et al., 2024b; Sun et al., 2025; Pfaff et al., 2026; Xia et al., 2026). This setting is inherently challenging: a scene generator must jointly interpret an underspecified prompt, determine appropriate scene contents, ground them into concrete 3D assets, organize heterogeneous objects spatially, and maintain semantic and geometric consistency throughout iterative construction.

Existing agentic systems typically address this problem through a staged construction process: they infer an object inventory and spatial relations from the input prompt, acquire the corresponding assets, and iteratively place and refine them using semantic, visual, and geometric feedback (Yang et al., 2024b; Sun et al., 2025; Pfaff et al., 2026; Xia et al., 2026). This paradigm has substantially improved object diversity, prompt consistency, and physical plausibility, enabling increasingly complete and visually convincing indoor environments. However, the resulting scenes are still predominantly specified and constructed at the level of objects and spatial relations, while the behavioral context that shapes how everyday environments are actually organized remains only weakly represented.

This discrepancy reflects a more fundamental mismatch between scene representation and everyday use. Existing construction pipelines largely answer which objects should appear and where they can be placed, whereas inhabited environments are also structured by the activities those objects jointly support (Fisher et al., 2015; Fu et al., 2017; Qi et al., 2018; Min et al., 2024). Objects participating in the same routine form functional dependencies involving support, orientation, shared resources, and access, and these dependencies extend beyond pairwise object relations (Fu et al., 2017; Min et al., 2024; Chen et al., 2025b). Consequently, individually plausible objects and collision-free placements do not necessarily produce a scene whose contents and spatial organization are coherent with how the environment would be used. We therefore formulate *lived-in scene generation*: constructing 3D environments whose scene content, functional organization, and spatial layout are jointly grounded in plausible everyday activities. Such a formulation provides a common behavioral context for both content generation and layout, allowing latent object requirements, functional dependencies, and human-use constraints to be resolved within the same construction process.

Realizing this objective introduces three coupled technical challenges. First, concise scene prompts under-specify the latent functional requirements of an inhabited environment, requiring the generator to infer plausible activity- conditioned contents while preserving explicit prompt constraints. Second, the inferred requirements form structured relational dependencies rather than an independent object inventory; their support, relative placement, orientation, and shared-use relations must survive asset instantiation and local organization. Third, local functional validity must remain compatible with room-level geometry and accessibility. A furniture arrangement that is coherent in isolation may become difficult to use once neighboring objects obstruct its interaction side or disconnect its usable region from the surrounding free space (Yu et al., 2011; Fu et al., 2020; Yang et al., 2024a; Jin et al., 2026). These challenges require behavioral context to persist across scene specification, local assembly, and global layout rather than being introduced at only one stage of generation.

To this end, we introduce *Lived-In by Design*, a behavior-grounded framework for agentic 3D scene generation that carries everyday-use context throughout the construction process. First, **behavior-to-asset grounding** uses plausible everyday activities to recover latent object requirements, quantities, shared resources, and functional relations from the input prompt, and consolidates them into a behavior–asset graph. Second, **functional-group assembly** organizes objects that jointly support a common use context into local groups, whose internal configuration is determined by object relations, behavior-dependent orientations, and human-referenced usage regions. Finally, **lived-in scene layout** places the assembled groups jointly at room scale while considering physical occupancy, architectural constraints, human usage space, and room-level access. The local organization of each group is preserved whenever possible, while constrained local revision adapts affected groups when required by global constraints. Together, these stages transform behavioral context from an intermediate semantic cue into a persistent constraint on both scene content and spatial organization.

Our contributions are:

- A **behavior-grounded scene specification** that recovers latent object requirements and functional relations from plausible everyday activities, transforming an underspecified prompt into a structured representation for agentic 3D scene construction.
- A **human-centered functional-group construction framework** that preserves behavior-derived organization from local assembly to room-level placement by jointly modeling object relations, behavior-dependent orientation, human usage space, and room-level access.

We evaluate *Lived-In by Design* on the 179 room-level prompts used by SceneSmith (Pfaff et al., 2026). Our evaluation combines SceneEval (Tam et al., 2026a), qualitative comparisons, controlled component ablations, and a blinded human study to assess semantic fidelity, spatial quality, and lived-in plausibility.

2 Related Work

Agentic and text-conditioned 3D scene generation. Indoor scene generation has evolved from learning object composition and spatial layout with autoregressive, graph-based, and diffusion models (Paschalidou et al., 2021; Zhai et al., 2023; Tang et al., 2024) toward increasingly controllable generation from language. Methods such as InstructScene and LayoutGPT introduce semantic or language-based structure into scene planning (Lin & Mu, 2024; Feng et al., 2023), while Holodeck and LayoutVLM translate open-ended descriptions into concrete assets, spatial constraints, and optimized layouts (Yang et al., 2024b; Sun et al., 2025). Recent work further enriches the scene representation itself, including interaction-aware graphs in Graph2Scene (Chen et al., 2025b), semantic-layout-conditioned generation in Layout2Scene (Chen et al., 2025a), and geometry-aware neural scene editing in GeoStyler (Hu et al., 2026). At the system level, HSM, SceneSmith, and SAGE increasingly organize generation hierarchically or agentically, combining structured scene decomposition with iterative semantic, visual, and physical feedback (Pun et al., 2026; Pfaff et al., 2026; Xia et al., 2026). These advances substantially strengthen prompt-conditioned scene specification and spatial construction, but the resulting generation process is still primarily organized around objects, relations, and geometric validity, leaving the everyday-use context of the environment only weakly represented.

Functional and activity-grounded scene organization. A separate line of work addresses this missing context by organizing scenes around what people do in them. Activity-centric synthesis relates object configurations to human activities (Fisher et al., 2015), activity-associated object relation graphs use inferred activities to introduce missing objects and spatial relations (Fu et al., 2017), and human-centric stochastic grammars incorporate human context directly into scene composition (Qi et al., 2018). More recent methods make this functional structure explicit; for example, FuncScene groups objects according to shared function and uses these groups to bridge local arrangements and global scene structure (Min et al., 2024). Together, these works show that scene organization is not determined solely by which objects co-occur, but also by the activities and functional dependencies that connect them. Our work adopts this insight within text-conditioned agentic generation, using plausible activities to recover latent object requirements, shared resources, and functional relations that persist from scene specification into functional-group assembly.

Human-aware scene layout and interaction. Functional organization alone does not determine whether a scene can be comfortably approached and used, motivating a complementary body of work on human-aware spatial reasoning. Early layout methods incorporate ergonomic factors such as visibility, accessibility, and circulation (Yu et al., 2011; Fu et al., 2020), while PiGraphs, POSA, and MIME model human–scene interaction through contact, affordance, or motion (Savva et al., 2016; Hassan et al., 2021; Yi et al., 2023). More recent approaches incorporate physical interactivity, reachability, and anthropometric constraints directly into scene synthesis (Yang et al., 2024a; Jin & Hyun, 2026; Jin et al., 2026). These methods establish that human use introduces spatial requirements beyond collision-free object placement. Building on both this human-aware perspective and the activity-grounded scene organization above, our method attaches behavior-dependent use directions and human-referenced usage regions to functional groups, and carries these constraints into room-level placement and access evaluation.

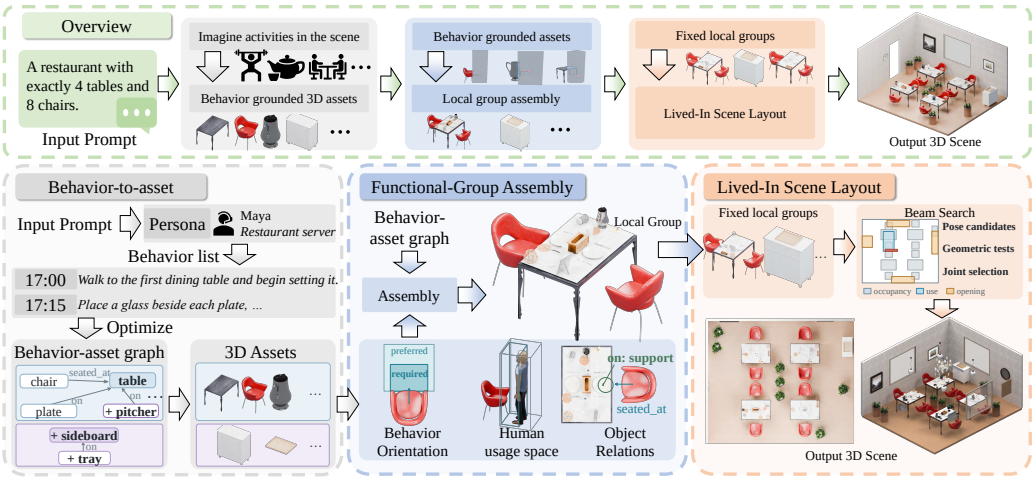

Figure 2: Behavior-grounded lived-in scene generation. Starting from a text prompt, behavior-to-asset grounding imagines plausible everyday activities and converts them into a behavior–asset graph and corresponding 3D assets. Related assets are assembled into local functional groups using object relations, behavior-dependent orientations, and human-use space. The assembled groups are then placed jointly at room scale, where beam search evaluates candidate configurations with respect to physical occupancy, usage space, architectural openings, and placement relations while preserving local organization whenever possible.

3 METHOD

3.1 OVERVIEW

Given a text prompt p and room geometry $\mathcal{R}$, our goal is to generate a furnished 3D scene that satisfies the description while reflecting how the environment may be inhabited and used. We use plausible everyday activities as a *scene-generation scaffold*: they reveal latent object requirements and functional relations, which are instantiated as local groups and carried with their human-use space into room-level layout.

As illustrated in Figure 2, our framework consists of three stages. **Behavior-to-asset grounding** infers plausible activities from the prompt and converts them into behavior-grounded object requirements and relations (Section 3.2). **Functional-group assembly** instantiates related objects in local coordinates while accounting for object relations, behavior-dependent orientation, and human-use space (Section 3.3). Finally, **lived-in scene layout** places the assembled groups at room scale while jointly considering physical occupancy, architectural constraints, and accessibility of their usage regions (Section 3.4). We build on SceneSmith (Pfaff et al., 2026) for room construction, asset acquisition, rendering, and physical post-processing, while restructuring scene construction around behavior-grounded functional organization and soft-space feedback. Figure 3 highlights the two representations that make this restructuring explicit: behavior-grounded object requirements and human-referenced interaction space.

3.2 BEHAVIOR-TO-ASSET GROUNDING

Activities reveal latent scene requirements. A room description usually specifies only part of the content involved in everyday use. We instantiate a lightweight persona whose role is consistent with the prompt and use it to infer plausible activities in the described environment. These activities provide structured hypotheses about how the environment may be used and what the scene may need to support.

We denote the inferred activity records by

$$\mathcal{A} = \{a_m\}_{m=1}^{M}, \tag{1}$$

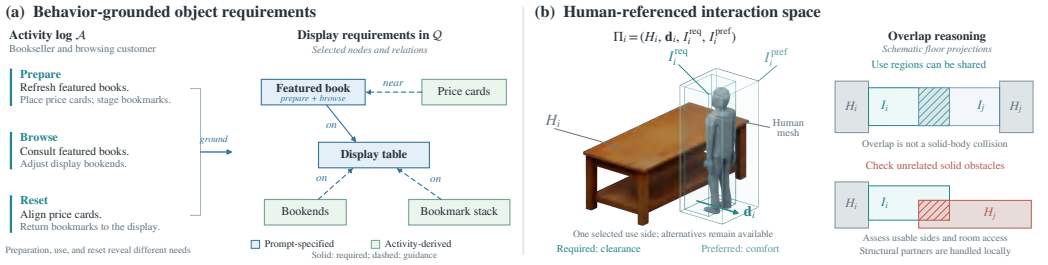

Figure 3: Key mechanisms of lived-in scene generation. (a) Behavior-to-asset grounding converts a bookseller's preparation, browsing, and reset activities into a display-requirement graph $\mathcal{Q}$, combining prompt-specified elements with activity-derived objects and functional relations. (b) A human reference defines an interactive object's selected use direction, required clearance, and preferred comfort region. Use regions may overlap when the associated physical geometries do not collide; unrelated solid obstacles are checked during room-level layout, while structural partners within a functional group are handled locally.

where a_m denotes the m-th activity record and M is the number of inferred activities. Each record describes an activity context, an ordered action sequence, and the objects involved. Reasoning across preparation, execution, and clearing or storage exposes complementary requirements absent from the prompt. For example, setting and serving a dining table may introduce place settings, serving items, trays, and nearby storage.

Behavior–asset graph. We consolidate the activity-derived requirements into a behavior–asset graph

$$\mathcal{Q} = (\mathcal{N}, \mathcal{E}_b), \tag{2}$$

where $\mathcal{N}$ is the set of object-requirement nodes and $\mathcal{E}_b$ is the set of behavior-derived relations between them. A node $n_i \in \mathcal{N}$ specifies an object requirement, including its description, quantity, and source. An edge $e_{ij} \in \mathcal{E}_b$ relates nodes n_i and n_j through a functional or spatial relation, such as *on*, *beside*, *seated-at*, facing, or attachment to the surrounding architecture.

Before asset acquisition, the graph is reconciled with the prompt: explicit objects and quantities are preserved, references to shared resources are merged, and complementary objects are retained when they serve distinct functions. Repeated structures share a group specification and are instantiated at the required count. The graph thus records both the required elements and the relations that should survive scene construction. Figure 3(a) illustrates how activity records add display-supporting objects and relations beyond the initial prompt requirement.

Functional grouping. We organize the behavior–asset graph into a set of functional groups

$$\mathcal{G} = \{G_k\}_{k=1}^{K}, \tag{3}$$

where G_k denotes the k-th functional group and K is the total number of groups. Each group is organized around major furniture anchors and contains objects that jointly support a coherent use context. A dining group, for example, may contain a table, chairs, place settings, and tabletop objects, while shared service storage forms a separate group. Each object belongs to one construction group, although an activity may involve several groups. Intra-group relations guide local assembly; inter-group and architectural relations remain active during room-level placement.

3.3 Functional-Group Assembly

Local assembly around intended use. For each functional group G_k, we acquire its constituent 3D assets and assemble them in an anchor-local coordinate frame. Let $\mathcal{O}_k$ denote the set of object instances belonging to G_k. Floor-supported furniture establishes the group structure; associated furniture and smaller supported objects are then placed according to $\mathcal{Q}$. Geometric checks evaluate collision, support, spacing, and orientation, complemented by rendered visual feedback.

For an object $o_i \in \mathcal{O}_k$, let ΔT_{ki} denote its rigid transform relative to the coordinate frame of group G_k. Let T_k denote the rigid transform that places the complete group in the room. The corresponding world-space pose of object o_i is then

$$T_{ki} = T_k \Delta T_{ki}, \tag{4}$$

where T_{ki} is the world-space transform of o_i. This factorization lets a group move coherently during room-level layout while preserving its local arrangement. Support bindings record which smaller objects follow their supporting furniture during local revision.

Behavior-dependent orientation. Geometric validity alone does not determine intended use: a chair can be close to a table without collision while facing away from it. For interactive furniture, we combine asset observations with the functional context in $\mathcal{Q}$ to estimate behavior-relevant use directions. These orientation constraints distinguish geometrically similar but functionally different configurations during local assembly.

Human-use space. We further associate interactive furniture with geometric regions that approximate the space needed for human use. For an interactive object o_i, we define its usage profile as

$$\Pi_i = \left(H_i,\, \mathbf{d}_i,\, I_i^{\mathrm{req}},\, I_i^{\mathrm{pref}} \right), \qquad I_i^{\mathrm{req}} \subseteq I_i^{\mathrm{pref}}, \tag{5}$$

where H_i denotes the physical geometry occupied by o_i, $\mathbf{d}_i$ is its selected horizontal use direction, I_i^{req} is the required usage region, and I_i^{pref} is a larger preferred usage region. The required region approximates the minimum approach and use space, while the preferred region provides additional clearance. Multiple profiles are retained when several interaction sides are plausible.

The regions are derived from a human reference and represented geometrically. We refer to physical occupancy H_i as *hard space*, and to the interaction regions and their access requirements collectively as *soft space*. Figure 3(b) illustrates the distinction between physical occupancy and interaction space, including the overlap reasoning used during room-level placement.

Local assembly jointly considers the relations in $\mathcal{Q}$, behavior-dependent orientation, and usage regions, producing a group that records both object poses and local conditions for use.

3.4 Lived-In Scene Layout

After local assembly, room-level layout makes the functional groups and their usage space coexist while preserving their local organization.

Group-level candidate search. At room scale, each functional group G_k is assigned a planar pose

$$\mathbf{q}_k = (x_k, y_k, \theta_k), \tag{6}$$

where (x_k, y_k) is the planar position of the group anchor and θ_k is its rotation about the vertical axis. The pose $\mathbf{q}_k$ induces the group transform T_k used in Eq. (4).

Candidate poses follow room geometry $\mathcal{R}$ and architectural or inter-group relations such as wall alignment, relative placement, and repeated formations. Because each group changes the feasible space for the others, bounded beam search explores joint group configurations.

Three test classes evaluate each candidate: *physical tests* for collisions, room bounds, and architectural openings; *relational tests* for architectural and inter-group relations; and *usage-space tests* for obstruction and use-direction violations. Structural partners within a group are distinguished from unrelated blocking furniture because their interactions were established during local assembly.

Candidates are ranked lexicographically: required physical and relational violations take priority, followed by required usage-space obstruction and access failures, then preferred relations and clearance. Appendix A.4 gives the complete ranking.

Room-level access. An unobstructed region beside an object may still be unreachable from the surrounding room. We therefore add a floor-level connectivity diagnostic.

For interactive object o_i, let U_i denote the two-dimensional floor projection of its required usage region I_i^{req}. We construct a free-space region Ω_i by insetting the room boundary and subtracting buffered footprints of surrounding furniture. Furniture in the evaluated object's structural group is excluded because its local interaction was evaluated during assembly. Let Ω_i^{main} denote the largest connected component of Ω_i. We define

$$c_i = \frac{|U_i \cap \Omega_i|}{|U_i|}, \qquad r_i = \mathbf{1}\left[|U_i \cap \Omega_i^{\text{main}}| > \epsilon\right], \tag{7}$$

where $|\cdot|$ denotes area, $c_i \in [0, 1]$ is the free fraction of the required usage region, and $r_i \in \{0, 1\}$ indicates overlap with the principal connected free-space component. The indicator threshold $\epsilon > 0$ rejects negligible connections. For multiple candidate use sides, we retain the alternative with the best valid clearance and connectivity. Appendix A.3 provides the geometry and thresholds.

Joint selection and constrained revision. Beam search retains promising partial layouts using the physical, relational, and usage-space criteria above. A visual critic additionally checks rendered observations against the prompt and behavior-derived organization.

Corrections first move a complete group through T_k. If group-level motion cannot satisfy a required room relation, the affected group is reopened for constrained local revision while the rest of the scene remains fixed. Only eligible furniture poses change; supported objects follow their anchors. The revised group returns to room-level placement after physical, relational, and visual checks.

Thus, plausible activities complement the explicit prompt with latent scene requirements, functional groups preserve their local organization, and room-level layout integrates these groups while maintaining space for human use.

4 Experiments

We evaluate prompt consistency, semantic and relational completeness, and spatial quality using SceneEval (Tam et al., 2026a), a blinded five-way human study, qualitative comparisons, and component ablations.

4.1 Experimental Setup

Prompts and baselines. We use the 179 room-level prompts from SceneSmith (Pfaff et al., 2026), which cover diverse room types, dense arrangements, and themed environments. We compare against HSM (Pun et al., 2026), Holodeck (Yang et al., 2024b), LayoutVLM with Objaverse assets (Sun et al., 2025), and SceneSmith (Pfaff et al., 2026). Each method receives the original scene description, and prompt-based evaluation uses that input. We evaluate the complete scene, including furniture and smaller objects.

Automated evaluation. Following SceneEval (Tam et al., 2026a), we report the average number of generated objects (Obj. #), object-count satisfaction (CNT), attributes (ATR), object–object relations (OOR), object–architecture relations (OAR), navigability (NAV), out-of-bounds objects (OOB), and collisions (COL). The first four scores measure semantic agreement; NAV, OOB, and COL measure spatial organization. We use a common evaluator configuration and report scene-level averages as specified in Appendix A.5.

User study. We conduct a blinded five-way evaluation with 25 participants on five representative scenes. For each scene, participants view the prompt and randomized, anonymized renderings, then select one best result for *Prompt Faithfulness*, *Realism*, and *Lived-in Plausibility*. These criteria assess explicit prompt agreement, credible composition and layout, and consistency with plausible everyday use, respectively. Each participant makes 15 selections, yielding 125 judgments per criterion and 375 in total. We report preference rates with participant-level bootstrap 95% confidence intervals; Appendix A.8 gives the full protocol and analysis.

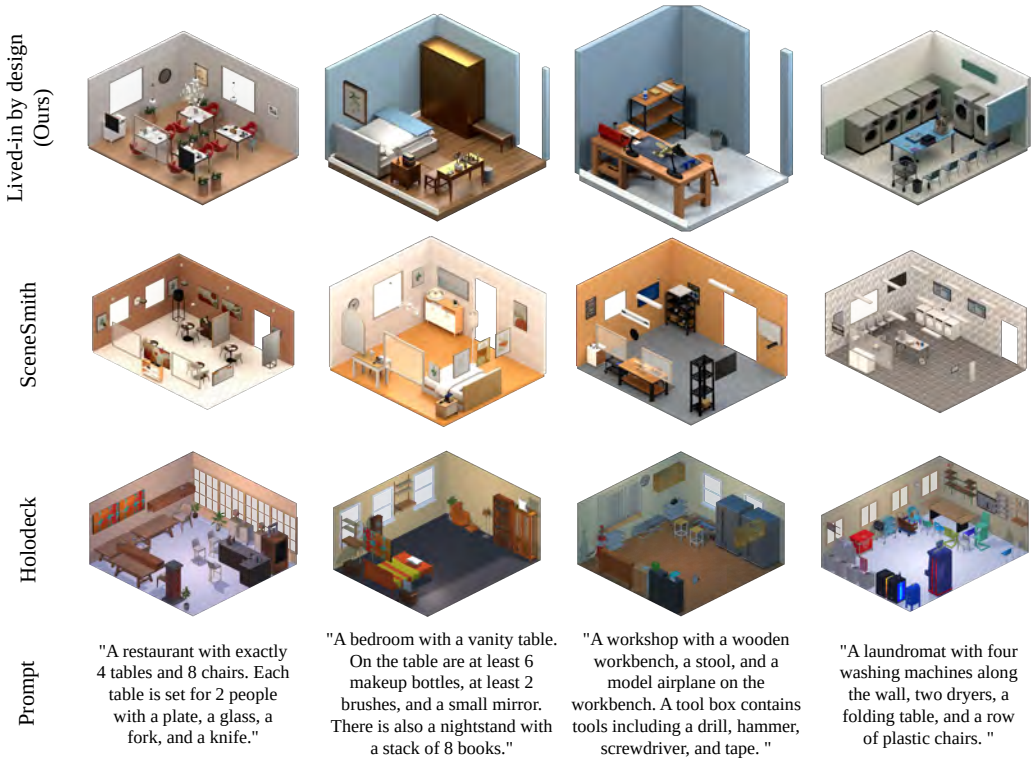

Figure 4: Qualitative comparison with SceneSmith and Holodeck on shared prompts. The rows show Lived-In by Design, SceneSmith, and Holodeck, respectively, using matched views for the same scene descriptions. Whole-room renderings show global organization and circulation, while close-ups highlight behavior-relevant contents, functional object relations, and the space around interactive furniture.

Table 1: Automated comparison. Obj. # is the average object count; CNT, ATR, OOR, and OAR evaluate semantic agreement; NAV, OOB, and COL evaluate spatial quality. All entries report mean scores, and SceneEval scores are in percent. Bold and underlining denote the best and second-best values for each metric.

	Content and semantic alignment					Spatial quality		
Method	Obj. #	CNT↑	ATR↑	OOR↑	OAR↑	NAV↑	OOB↓	COL↓
HSM	15.6	48.0	56.5	21.2	59.2	**99.5**	7.0	30.2
Holodeck	37.5	24.4	42.5	20.8	28.1	<u>99.4</u>	0.8	12.3
LayoutVLM (Obj.)	14.2	55.5	34.3	20.7	17.4	98.7	5.1	28.9
SceneSmith	**71.1**	**82.9**	<u>74.4</u>	<u>67.6</u>	**80.6**	97.6	<u>0.2</u>	**1.2**
Ours	<u>57.1</u>	<u>66.8</u>	**88.9**	**75.2**	<u>72.9</u>	**99.5**	**0.1**	<u>7.0</u>

4.2 Comparison with Existing Methods

Automated comparison. Table 1 shows that the gains of our method are concentrated in the aspects most directly affected by behavior-grounded organization. Ours achieves the highest ATR (88.9) and OOR (75.2), exceeding the next-best scores by 14.5 and 7.6 points, respectively. This pattern is consistent with our design: behavior-to-asset grounding enriches object-level requirements, while functional-group assembly explicitly preserves relations among objects that participate in a common use context. Importantly, these gains are not explained by object count alone. Although SceneSmith generates more objects (71.1 vs. 57.1), it obtains lower ATR and OOR, suggesting that the improvement comes from how scene contents are specified and organized rather than from simply increasing scene density. At the same time, Ours maintains the best-tied NAV (99.5) and the

Table 2: Human preference study. Values are preference rates (%) with participant-level bootstrap 95% confidence intervals from 375 judgments by 25 participants. Five-way chance is 20%; Ours is preferred under an exact case-level label-permutation test ($p < 0.001$) for all criteria.

Method	Prompt Faithfulness↑	Realism↑	Lived-in Plausibility↑
HSM	8.8 [4.0, 14.4]	8.0 [3.2, 13.6]	4.0 [0.8, 7.2]
Holodeck	3.2 [0.0, 7.2]	2.4 [0.0, 6.4]	3.2 [0.8, 6.4]
LayoutVLM (Obj.)	2.4 [0.0, 6.4]	3.2 [0.0, 7.2]	2.4 [0.0, 6.4]
SceneSmith	8.8 [4.0, 14.4]	8.0 [3.2, 13.6]	8.0 [2.4, 15.2]
Ours	**76.8 [63.2, 88.8]**	**78.4 [66.4, 88.8]**	**82.4 [72.8, 91.2]**

lowest OOB rate (0.1), indicating that the additional functional structure can be integrated without degrading room-level navigability or boundary validity. SceneSmith remains stronger on CNT, OAR, and COL, showing that our gains are primarily in object-level semantic and functional organization rather than uniformly across all SceneEval dimensions.

Qualitative comparison. Figure 4 makes the relational differences in Table 1 visually apparent. The main distinction is not the presence of additional objects, but their organization around shared activity contexts: smaller objects are attached to meaningful furniture anchors, repeated functional structures remain internally coherent, and interactive furniture retains usable surrounding space. In contrast, the baseline scenes often recover the correct room type and major furniture while exhibiting weaker local functional organization. The qualitative results therefore suggest that the higher OOR score reflects structured local arrangements rather than isolated pairwise relation improvements.

User-study results. Human judgments show a consistent preference for our scenes across all three criteria, with 76.8% for Prompt Faithfulness, 78.4% for Realism, and 82.4% for Lived-in Plausibility, all well above the 20% five-way chance level ($p = 0.00032$). The strongest preference is for Lived-in Plausibility, which most directly reflects our target and remains high across all five scenes (68–92%). Prompt Faithfulness and Realism are also strongly preferred, suggesting that behavior-derived content and functional organization remain compatible with the original descriptions rather than appearing as extraneous additions. Together with the automated results, this indicates that the method improves functional organization in ways that are both measurable through semantic relations and perceptible at the level of complete-scene plausibility.

Ablation study. Component-ablation definitions and results are reported in Appendix A.6.

5 Conclusion

We introduced *Lived-In by Design*, a behavior-grounded framework for agentic 3D scene generation that incorporates everyday-use context into scene specification and spatial construction. By recovering latent object requirements and functional relations from plausible activities, organizing them into functional groups, and carrying human-use constraints into room-level layout, the framework extends scene construction beyond explicit object inventories and pairwise spatial relations. On 179 room-level prompts, our method achieves the highest SceneEval scores for attributes and object–object relations, while matching the best navigability and obtaining the lowest out-of-bounds rate among the compared methods. In a five-way human study, our scenes receive 82.4% preference for lived-in plausibility, with similarly strong preferences for prompt faithfulness (76.8%) and realism (78.4%). These results suggest that grounding scene construction in plausible everyday use provides an effective intermediate structure for generating environments with coherent content, functional organization, and spatial layout.

Limitations. The inferred activities are plausible usage hypotheses conditioned on the prompt and reasoning model, rather than an exhaustive account of scene use. Human-use space is approximated by static anthropometric regions and floor-level connectivity, without modeling articulated-object operation or full-body interaction. Extending the framework to dynamic human–scene interaction and user-dependent spatial requirements is left for future work.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising parts of the manuscript, searching for and formatting related references, and preparing annotations and layouts for supplementary figures. We did not use generative AI tools to generate or clean datasets, develop the conceptual framework, formulate mathematical claims or proofs, propose or refine hypotheses, design the reported methodology or experiments, implement the reported method, conduct qualitative or thematic analyses, or interpret the reported results. All AI-assisted text, references, and visual annotations were reviewed and edited by the authors. Language and vision-language models are components of the proposed scene-generation pipeline described in Section 3; their outputs are experimental artifacts of the method rather than authoring assistance. The authors verified the reported results and take full responsibility for the final content of this submission, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work generates and evaluates synthetic indoor scenes. Our human study asks 25 participants to compare randomized, anonymized renderings of these synthetic scenes under predefined criteria. The analyses use only the resulting comparative choices and report aggregate statistics; no participant identifiers or individual-level responses are disclosed. The activity inferences used by our method are plausible hypotheses derived from scene prompts rather than observations of real individuals. They can nevertheless reflect cultural or contextual assumptions in the underlying models and assets. Generated scenes should therefore be reviewed for inappropriate, stereotypical, or unsafe content before use in downstream applications.

REFERENCES

Minglin Chen, Longguang Wang, Sheng Ao, Ye Zhang, Kai Xu, and Yulan Guo. Layout2Scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. *arXiv preprint arXiv:2501.02519*, 2025a. URL https://arxiv.org/abs/2501.02519.

Minglin Chen, Rongkun Yang, Qibin Hu, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Graph2Scene: Versatile 3d indoor scene generation with interaction-aware scene graph. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 11313–11320. IEEE, 2025b.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3a7f9e485845dac27423375c934cb4db-Abstract.html.

Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM Transactions on Graphics*, 34(6):179:1–179:13, 2015. doi: 10.1145/2816795.2818057.

Qiang Fu, Xiaowu Chen, Xiaotian Wang, Sijia Wen, Bin Zhou, and Hongbo Fu. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics*, 36(6):201:1–201:13, 2017. doi: 10.1145/3130800.3130805.

Qiang Fu, Hongbo Fu, Hai Yan, Bin Zhou, Xiaowu Chen, and Xueming Li. Human-centric metrics for indoor scene assessment and synthesis. *Graphical Models*, 110:101073, 2020. doi: 10.1016/j.gmod.2020.101073.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. POSA: Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14708–14718, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hassan_Populating_3D_Scenes_by_Learning_Human-Scene_Interaction_CVPR_2021_paper.html.

Qibin Hu, Ye Zhang, Jisheng Dang, Minglin Chen, Longguang Wang, and Yulan Guo. GeoStyler: A generalizable geometry-aware diffusion-based approach for direct 3d gaussian style transfer. *IEEE Transactions on Image Processing*, 35:6730–6743, 2026. doi: 10.1109/TIP.2026.3702842.

Semin Jin and Kyung Hoon Hyun. Human-operational 3D indoor layout generation with LLM-driven anthropometric simulation. *Automation in Construction*, 184:106846, 2026. doi: 10.1016/j.autcon.2026.106846.

Semin Jin, Donghyuk Kim, Jeongmin Ryu, and Kyung Hoon Hyun. Behavior-aware anthropometric scene generation for human-usable 3D layouts. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. Association for Computing Machinery, 2026. doi: 10.1145/3772318.3790341. URL https://doi.org/10.1145/3772318.3790341.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LtuRgL03pI.

Wenjie Min, Wenming Wu, Gaofeng Zhang, and Liping Zheng. FuncScene: Function-centric indoor scene synthesis via a variational autoencoder framework. *Computer Aided Geometric Design*, 111:102319, 2024. doi: 10.1016/j.cagd.2024.102319.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12013–12026, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/64986d86a17424eeac96b08a6d519059-Abstract.html.

Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. SceneSmith: Agentic generation of simulation-ready indoor scenes. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2602.09153.

Hou In Derek Pun, Hou In Ivan Tam, Austin T. Wang, Xiaoliang Huo, Angel X. Chang, and Manolis Savva. HSM: Hierarchical scene motifs for multi-scale indoor scene generation. In *International Conference on 3D Vision (3DV)*, pp. 1356–1367, 2026. doi: 10.1109/3DV69130.2026.00131. URL https://3dlg-hcvc.github.io/hsm/.

Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5899–5908, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Qi_Human-Centric_Indoor_Scene_CVPR_2018_paper.html.

Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning interaction snapshots from observations. *ACM Transactions on Graphics*, 35(4):139:1–139:12, 2016. doi: 10.1145/2897824.2925867.

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*, pp. 29469–29478, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_LayoutVLM_Differentiable_Optimization_of_3D_Layout_via_Vision-Language_Models_CVPR_2025_paper.html.

Hou In Ivan Tam, Hou In Derek Pun, Austin T. Wang, Angel X. Chang, and Manolis Savva. SceneEval: Evaluating semantic coherence in text-conditioned 3D indoor scene synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026a. URL https://arxiv.org/abs/2503.14756.

Hou In Ivan Tam, Hou In Derek Pun, Austin T. Wang, Xiaohao Sun, Qirui Wu, Han-Hung Lee, Angel X. Chang, and Manolis Savva. Survey on compositional 3D indoor scene generation. *Computer Graphics Forum*, pp. e70591, 2026b. doi: 10.1111/cgf.70591.

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*, pp. 20507–20518, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tang_DiffuScene_Denoising_Diffusion_Models_for_Generative_Indoor_Scene_Synthesis_CVPR_2024_paper.html.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, Shuran Song, and Fangyin Wei. SAGE: Scalable agentic 3d scene generation for embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22358–22368, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xia_SAGE_Scalable_Agentic_3D_Scene_Generation_for_Embodied_AI_CVPR_2026_paper.html.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically interactable 3d scene synthesis for embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16262–16272, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_PhyScene_Physically_Interactable_3D_Scene_Synthesis_for_Embodied_AI_CVPR_2024_paper.html.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. 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, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Holodeck_Language_Guided_Generation_of_3D_Embodied_AI_Environments_CVPR_2024_paper.html.

Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12965–12976, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yi_MIME_Human-Aware_3D_Scene_Generation_CVPR_2023_paper.html.

Lap-Fai Yu, Sai-Kit Yeung, Chi-Keung Tang, Demetri Terzopoulos, Tony F. Chan, and Stanley J. Osher. Make it home: Automatic optimization of furniture arrangement. *ACM Transactions on Graphics*, 30(4):86:1–86:11, 2011. doi: 10.1145/1964921.1964981.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30026–30038, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5fba70900a84a8fb755c48ba99420c95-Abstract-Conference.html.

A Appendix

A.1 Generation Configuration

Our implementation follows the three-stage construction described in Section 3: behavior-to-asset grounding, functional-group assembly, and lived-in scene layout. The experimental configuration records the language and vision models used for behavioral reasoning and visual feedback, the asset sources and acquisition settings, the human-reference parameters used to construct usage regions, and the search and refinement budgets used during scene construction.

All methods receive the same room descriptions and retain their standard asset sources and generation pipelines. Their outputs are evaluated using the SceneEval definitions in Section 4.1.

Asset construction and rendering. For generated assets, an image-synthesis backend first produces a reference image, and the SAM3D geometry backend reconstructs the corresponding 3D geometry. To prepare the user-study renderings for our method, we use Blender 4.5.14 with Cycles CUDA on an NVIDIA H100 GPU at 1280×1280 resolution, 64 samples per pixel, and denoising enabled. Camera and lighting settings are specified separately for each scene.

A.2 Behavior-to-Asset and Group Representation

Following Section 3.2, the inferred activities are represented as

$$\mathcal{A} = \{a_m\}_{m=1}^{M},$$

where a_m denotes the m-th activity record and M is the total number of records. Each activity record stores the relevant persona role, an activity context or phase, and an ordered sequence of actions together with references to the objects involved. The persona provides a concrete behavioral context for activity inference.

The activity-derived requirements are consolidated into the behavior–asset graph

$$\mathcal{Q} = (\mathcal{N}, \mathcal{E}_b),$$

where $\mathcal{N}$ is the set of object-requirement nodes and $\mathcal{E}_b$ is the set of behavior-derived relations. A node $n_i \in \mathcal{N}$ stores the object description, required quantity, and provenance of the corresponding requirement. An edge $e_{ij} \in \mathcal{E}_b$ records a functional or spatial relation between nodes n_i and n_j, such as support, relative placement, facing, seated-at relations, or attachment to architectural elements. Where applicable, relations also retain whether they express required or preferred conditions for subsequent placement.

Requirements referring to the same physical resource are merged across activities, while explicit objects and quantities in the input prompt p are preserved. Repeated functional structures share a common group specification and are instantiated according to the required count. Individual instances may nevertheless use different retrieved assets or local details.

For implementation bookkeeping, object requirements are additionally tagged according to their provenance and role. *Prompt-mandated* requirements originate directly from explicit prompt constraints; *activity-essential* requirements are needed by the primary inferred activities; *functional additions* provide complementary resources required by those activities; and *contextual details* describe additional routine-related contents. These tags preserve requirement provenance during planning and consistency checks.

The behavior–asset graph is organized into

$$\mathcal{G} = \{G_k\}_{k=1}^{K},$$

where G_k is the k-th functional group and K is the number of groups. For each group, $\mathcal{O}_k$ denotes its set of instantiated object members. The stored group representation contains its anchor furniture, member set, intra-group relations, group-local object transforms $\{\Delta T_{ki}\}$, and support bindings for attached objects. Here ΔT_{ki} is the rigid transform of object $o_i \in \mathcal{O}_k$ relative to the coordinate frame of G_k.

Each instantiated object is assigned to one construction group, although an activity may involve objects belonging to multiple groups. Relations internal to G_k are primarily resolved during local assembly, whereas inter-group and architectural relations remain active during room-level layout. If supporting furniture is revised, its attached descendants inherit the corresponding transform. Group membership and the object inventory remain fixed during such local pose revision.

A.3 Human Usage Space and Access Diagnostics

Behavior-dependent use directions. For each interactive furniture object o_i, the spatial profiler first estimates candidate horizontal interaction sides from rendered asset views. The behavior–asset graph $\mathcal{Q}$ then provides the functional context used to select or prioritize the direction relevant to the current interaction. The selected horizontal use direction is denoted by $\mathbf{d}_i$. When more than one side remains plausible, multiple alternatives are retained and evaluated during layout.

Interaction regions. As defined in Eq. (5), the usage profile of interactive object o_i is

$$\Pi_i = \left(H_i,\, \mathbf{d}_i,\, I_i^{\mathrm{req}},\, I_i^{\mathrm{pref}} \right),$$

where H_i denotes the physical geometry occupied by the object, $\mathbf{d}_i$ is its behavior-relevant use direction, I_i^{req} is the required usage region, and I_i^{pref} is the preferred usage region, with $I_i^{\mathrm{req}} \subseteq I_i^{\mathrm{pref}}$.

The usage regions are constructed as ground-aligned geometric boxes adjacent to the interaction side of the object and provide a human-referenced spatial proxy for local use. The implementation uses a standing human reference of 1.70 m. The base horizontal extent of the required region is 0.56 m laterally and 0.40 m in depth. The preferred region uses a larger horizontal extent of 0.75 m laterally and 0.70 m in depth, with a reference height of 1.75 m. Horizontal dimensions may be adjusted by a bounded clearance factor to account for asset scale and geometry. Region placement is recomputed from the final asset bounds.

The required region represents the minimum spatial allowance used when testing whether an object remains approachable or operable. The preferred region encodes additional clearance and is therefore treated as a lower-priority layout preference. Physical collision is evaluated independently from these regions using the scene geometry.

Usage-space interference. Usage-space interference is evaluated on the floor projection of each interaction region. Furniture belonging to the same structural functional group is distinguished from unrelated neighboring furniture. This distinction prevents an intended group member—for example, a chair associated with a dining table—from being treated in the same way as an external object that blocks the table's use side.

Intrusion into I_i^{req} contributes to required usage-space violations, whereas intrusion confined to $I_i^{\mathrm{pref}} \setminus I_i^{\mathrm{req}}$ is treated as a preferred-clearance violation. If several candidate use directions are available, the alternatives are evaluated independently and the best valid alternative is retained.

Room-level access diagnostic. Usage-space clearance alone does not establish whether the corresponding region is connected to the surrounding room. We therefore apply the floor-level access diagnostic defined in Eq. (7).

For object o_i, U_i denotes the two-dimensional floor projection of its required usage region I_i^{req}. The free-space region Ω_i is constructed from the room floor by insetting its boundary according to a navigation proxy and subtracting buffered footprints of surrounding furniture. Thin floor coverings and structural partners belonging to the evaluated functional group are excluded from the obstacle set. The largest connected component of Ω_i is denoted by Ω_i^{main}.

The two diagnostics are

$$c_i = \frac{|U_i \cap \Omega_i|}{|U_i|}, \qquad r_i = \mathbf{1}\left[|U_i \cap \Omega_i^{\mathrm{main}}| > \epsilon \right],$$

where $|\cdot|$ denotes two-dimensional area, $c_i \in [0,1]$ measures the fraction of required usage space that remains free, $r_i \in \{0,1\}$ indicates whether the usage region connects to the principal free-space component, $\mathbf{1}[\cdot]$ is the indicator function, and $\epsilon > 0$ is the minimum overlap area used to reject negligible connections.

The implementation uses a navigation-proxy width of 0.30 m. A clearance value $c_i \leq 0.25$ is treated as blocked, whereas $c_i \geq 0.60$ is treated as good clearance. The connectivity threshold is

$$\epsilon = 10^{-3}\ \mathrm{m}^2.$$

For furniture with multiple candidate interaction sides, clearance and connectivity are evaluated for each alternative and the best available valid result is retained.

These quantities provide a group-conditioned geometric approximation of local clearance and room-level access. Excluding structural group members also means that the connectivity diagnostic is intended primarily to detect interference introduced by the surrounding room layout; within-group geometry is handled during local assembly.

A.4 GROUP-LEVEL SEARCH AND CONSTRAINED REVISION

Following Section 3.4, let

$$\Lambda = \{\mathbf{q}_k\}_{k=1}^{K}$$

denote a candidate room-level layout of the K functional groups, where

$$\mathbf{q}_k = (x_k, y_k, \theta_k)$$

is the planar pose of group G_k. Here (x_k, y_k) specifies the planar position of its anchor and θ_k its rotation about the vertical axis. The pose $\mathbf{q}_k$ determines the group transform T_k used in Eq. (4).

Candidate poses are generated from wall segments, room interiors, corners, architectural constraints, inter-group placement relations, and repeated formations. The search expands combinations of group candidates using a bounded beam and subsequently refines their planar translations and orientations.

Lexicographic layout ranking. For a candidate layout Λ, the complete-group search uses the ranking vector

$$\mathbf{v}(\Lambda) = (n_{\text{req}}, e_{\text{req}}, n_{\text{block}}, n_{\text{disc}}, n_{\text{dir}}, n_{\text{pref}}, e_{\text{pref}}, n_{\text{comfort}}, d_{\text{move}}) . \tag{8}$$

All entries are minimized and candidates are ordered lexicographically. The components have the following meanings. n_{req} is the number of required physical or relational violations, including inter-group collision conflicts, blocked architectural openings, and unsatisfied required formations. e_{req} is the aggregate penalty associated with the remaining required layout and formation errors. n_{block} counts blocking intrusions into required human usage regions. n_{disc} counts required usage regions that fail the room-level connectivity test, i.e., cases with $r_i = 0$. n_{dir} counts violated or blocked required use-direction conditions. n_{pref} is the number of unsatisfied preferred placement relations, and e_{pref} is their aggregate penalty. n_{comfort} counts intrusions that affect preferred usage space without violating the corresponding required region. Finally, d_{move} is the movement cost used during pose refinement to discourage unnecessary displacement.

This ordering gives required physical and relational validity the highest priority. Required human-use violations and access failures are considered next, followed by preferred spatial relations and preferred-clearance objectives. Thus, larger comfort margins cannot compensate for a configuration that violates a higher-priority condition. The beam retains a bounded number of alternative partial layouts after each expansion.

Constrained local revision. The local configuration of a functional group is preserved during room-level placement whenever possible. If no group-level transform can satisfy a required room relation, the affected group may be reopened for constrained local revision while the remainder of the scene is held fixed.

Only eligible furniture members within that group are allowed to change their local transforms ΔT_{ki}. Objects attached through support bindings follow their supporting furniture, and the group's membership remains unchanged. The revised group is accepted only after re-evaluating physical validity, required relations, usage-space conditions, and rendered visual consistency. Its updated local transforms are then stored and the group is returned to room-level search.

The selected room layout therefore results from joint group-level search with limited local correction rather than unrestricted independent replacement of individual objects.

A.5 SCENEEVAL REPORTING CONVENTIONS

All methods are evaluated on the same original scene prompts using a common SceneEval configuration. We report the average number of generated objects per scene (Obj. #), together with CNT, ATR, OOR, OAR, NAV, OOB, and COL. Obj. # is computed directly from the generated scene inventory. CNT, ATR, OOR, and OAR follow the SceneEval semantic evaluation of explicit count, attribute, object–object relation, and object–architecture relation requirements. NAV follows the SceneEval spatial evaluation of navigability. OOB and COL are reported from the corresponding object-level out-of-bounds and collision flags.

For each SceneEval score, we first compute a scene-level value and then average over the scenes to which the metric applies. We additionally record metric-specific sample counts and the treatment of generation or evaluation failures.

A.6 Component Ablations

We examine two component removals. *Without human-use feedback* removes local interaction spaces and their orientation and critic feedback while retaining behavior-derived planning and functional-group assembly. *Without room-level access* retains the local checks but disables the connectivity diagnostic in Eq. (7). Table 3 focuses on count and object–object relation satisfaction, the two available measures of whether the resulting arrangements retain the specified contents and functional relations.

Table 3: Selected SceneEval scores (percent) for component removals evaluated on a subset of prompts. Higher is better. The full-method row reproduces the 179-prompt benchmark in Table 1 as context; it is not a matched-subset control.

Variant	CNT↑	OOR↑
Without human-use feedback	20.0	0.0
Without room-level access	40.0	0.0
Full method (Table 1)	66.8	75.2

On the ablation subset, both component removals yield an OOR score of 0.0, with CNT scores of 20.0 and 40.0. These cases expose a failure mode in which the generated scene does not retain the requested object–object relations. The full method obtains 75.2 OOR and 66.8 CNT on the complete 179-prompt benchmark (Table 1). Because that reference uses the full prompt set, the ablation scores characterize behavior on the selected subset rather than a measured gain over a matched control.

A.7 Additional Qualitative Comparisons

Figure 5 complements the main-paper comparison in Figure 4 using the same four prompts and matched camera views. It presents Lived-In by Design alongside HSM and LayoutVLM, so that the qualitative evidence covers all four baselines used in the automated and user-study evaluations.

A.8 Human Evaluation Protocol and Analysis

Protocol. We conduct a blinded human preference study with 25 participants on five representative scenes. For each scene, participants compare five anonymized methods under three criteria: *Prompt Faithfulness*, *Realism*, and *Lived-in Plausibility*. Method order is independently randomized for each scene, and participants select exactly one best result for each criterion.

Prompt Faithfulness evaluates agreement with the input description, including the room type, explicitly required objects, quantities, and stated spatial relations. Realism evaluates whether the generated object composition and spatial arrangement form a visually credible indoor environment. Lived-in Plausibility evaluates whether the scene contents, object organization, usable space, and surrounding spatial arrangement are consistent with plausible everyday use.

Each participant evaluates five scenes under all three criteria, yielding

$$5 \times 3 = 15$$

forced-choice decisions per participant. All 25 submissions are complete, resulting in 125 valid judgments per criterion and 375 judgments in total.

Study interface. The study interface presents the original scene prompt together with five anonymized scene proposals. Participants can inspect the scenes from front, back, left, right, and top orthographic views. A draggable rectangular region can additionally be used to inspect local details; the corresponding crop is synchronized across all five proposals to support direct comparison at the same spatial location. For each criterion, participants select one proposal using radio buttons.

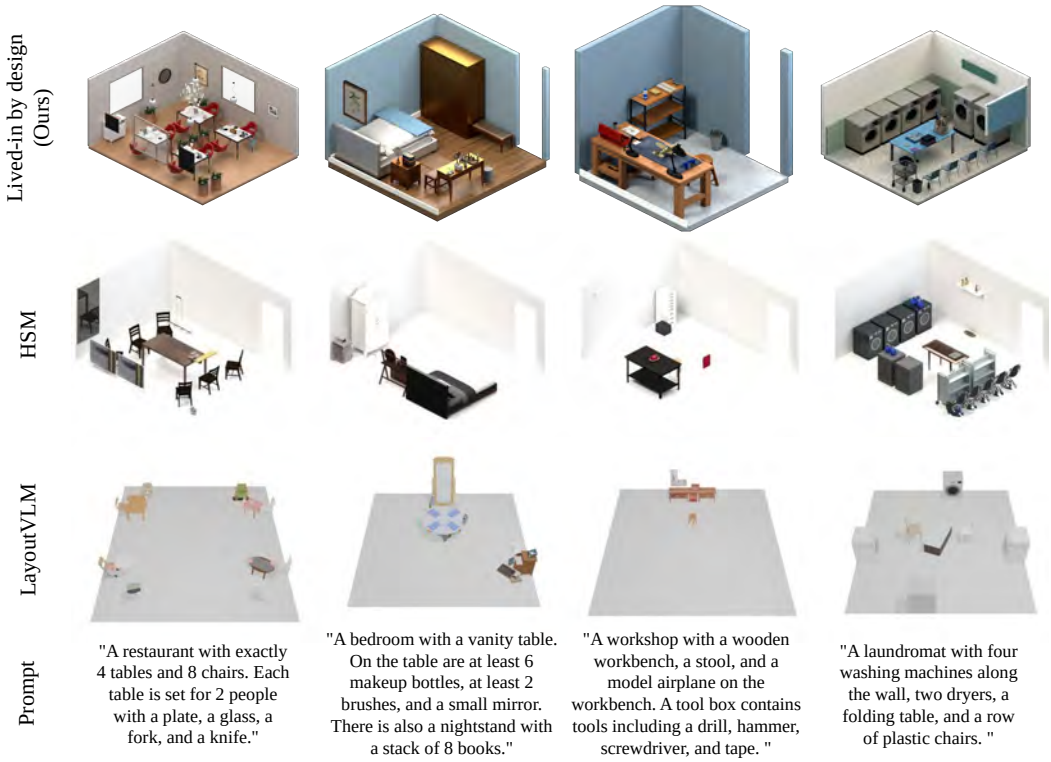

Figure 5: Additional qualitative comparison with HSM and LayoutVLM. The rows show Lived-In by Design, HSM, and LayoutVLM, respectively, for the same restaurant, bedroom, model-workshop, and laundromat prompts shown in Figure 4. Matched views support comparison of prompt coverage, room-scale organization, and activity-relevant contents.

Table 4: Original scene prompts used in the human study.

Case	Prompt
Restaurant (173)	A restaurant with exactly 4 tables and 8 chairs. Each table is set for 2 people with a plate, a glass, a fork, and a knife.
Workshop (133)	A workshop with a wooden workbench, a stool, and a model airplane on the workbench. A tool box contains tools including a drill, hammer, screwdriver, and tape.
Laundromat (143)	A laundromat with four washing machines along the wall, two dryers, a folding table, and a row of plastic chairs.
Bedroom (169)	A bedroom with a vanity table. On the table are at least 6 makeup bottles, at least 2 brushes, and a small mirror. There is also a nightstand with a stack of 8 books.
Bookstore (101)	A bookstore with at least four tall bookshelves, each containing at least five books. A central display table showcases featured titles, and two comfortable reading chairs sit in one corner.

All selections remain editable until final submission. Figures 6 and 7 illustrate the interface and the synchronized local-detail inspection.

Scene prompts. Table 4 lists the five original prompts used in the study. The selected scenes cover distinct functional settings and object compositions, including dining, work, household service, personal use, and retail environments.

Preference aggregation and uncertainty. For method m and criterion d, we define its preference rate as

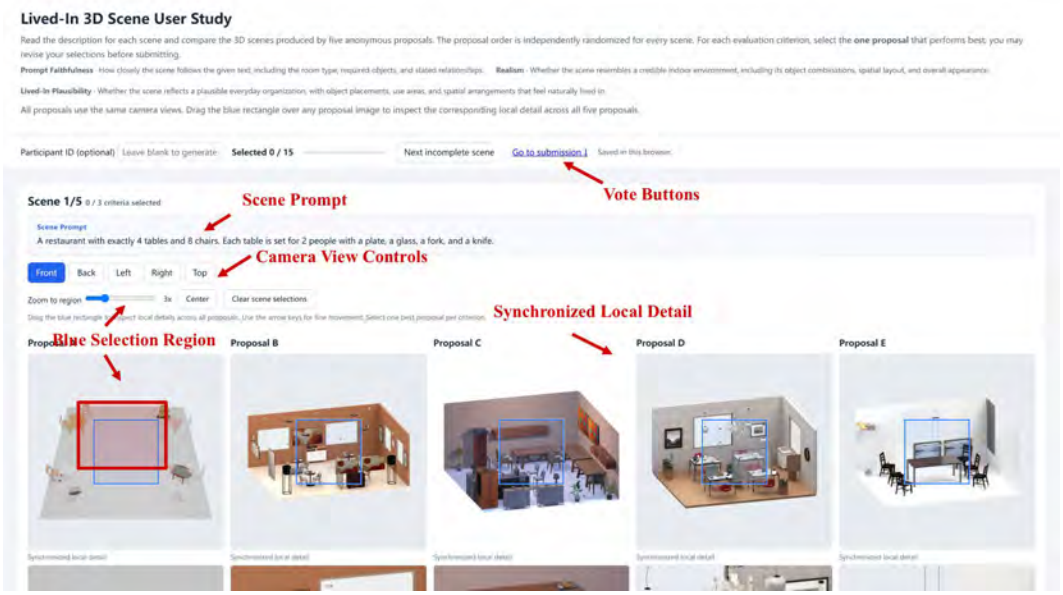

Figure 6: **User-study interface.** Each trial presents the scene prompt, evaluation criteria, five anonymized proposals, camera-view controls, and a draggable local-selection region.

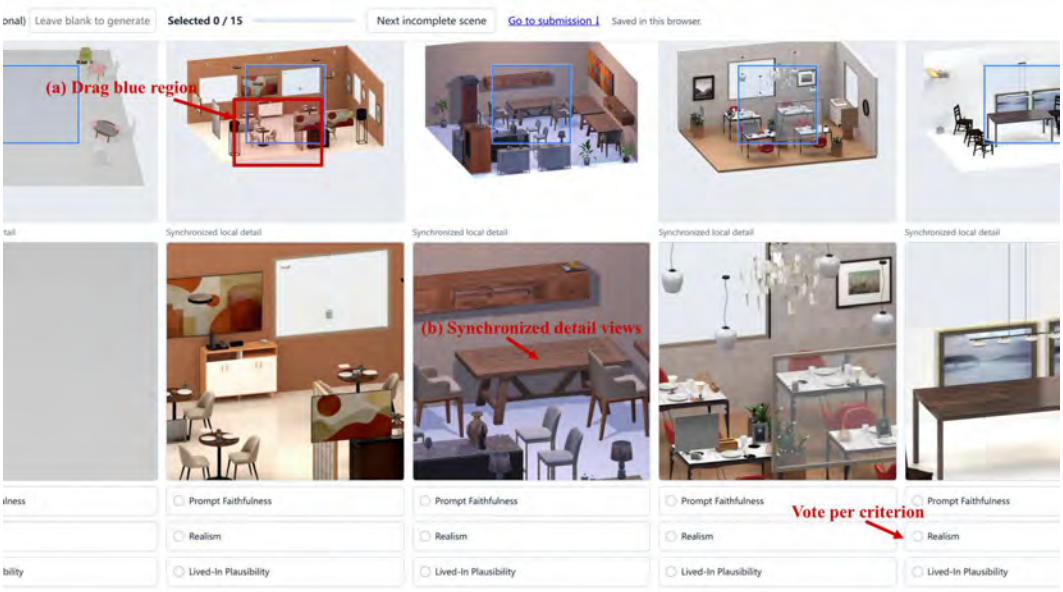

Figure 7: **Synchronized local-detail inspection.** A participant selects a local region in one proposal, and the interface displays the corresponding location across all five anonymized methods for direct comparison.

Table 5: Aggregate human-study results. Values report raw vote counts and preference rates (%) with participant-level bootstrap 95% confidence intervals.

Method	Prompt Faithfulness	Realism	Lived-in Plausibility
HSM	11/125; 8.8 [4.0, 14.4]	10/125; 8.0 [3.2, 13.6]	5/125; 4.0 [0.8, 7.2]
Holodeck	4/125; 3.2 [0.0, 7.2]	3/125; 2.4 [0.0, 6.4]	4/125; 3.2 [0.8, 6.4]
LayoutVLM (Obj.)	3/125; 2.4 [0.0, 6.4]	4/125; 3.2 [0.0, 7.2]	3/125; 2.4 [0.0, 6.4]
SceneSmith	11/125; 8.8 [4.0, 14.4]	10/125; 8.0 [3.2, 13.6]	10/125; 8.0 [2.4, 15.2]
Ours	**96/125; 76.8 [63.2, 88.8]**	**98/125; 78.4 [66.4, 88.8]**	**103/125; 82.4 [72.8, 91.2]**

$$\mathrm{Pref}(m, d) = \frac{N_{m,d}}{N_d}, \tag{9}$$

where $N_{m,d}$ is the number of judgments selecting method m as the best result under criterion d, and $N_d = 125$ is the total number of valid judgments for that criterion. Because each judgment selects exactly one of five methods, the preference rates sum to one within each criterion and the uniform-choice chance level is $1/5 = 0.2$.

We estimate uncertainty using a participant-level percentile bootstrap. Each bootstrap sample draws 25 participants with replacement and retains all five scene judgments associated with each sampled participant. This preserves the dependence among multiple judgments from the same participant. Preference rates are recomputed over 100,000 bootstrap resamples, and the 2.5th and 97.5th percentiles form the reported 95% confidence intervals.

Significance testing. We further evaluate whether the observed preference for Ours can be explained by method-label assignment at the scene level. For each scene, we preserve the observed distribution of votes across the five methods and enumerate permutations of the method labels. The test statistic is the total number of votes assigned to Ours across the five scenes. The exact p-value is the proportion of label permutations whose total preference is at least as large as the observed value.

This case-level exact label-permutation test gives

$$p = 0.00032$$

for Prompt Faithfulness, Realism, and Lived-in Plausibility. As an additional robustness analysis, a participant-cluster label-permutation test yields $p < 10^{-5}$ for all three criteria.

Aggregate preference. Table 5 reports the aggregate vote counts and preference rates. Ours receives 96/125 votes for Prompt Faithfulness, 98/125 for Realism, and 103/125 for Lived-in Plausibility, corresponding to preference rates of 76.8%, 78.4%, and 82.4%, respectively.

Per-scene preference. Table 6 reports the complete per-scene preference distributions. Across all $5 \times 3 = 15$ scene–criterion combinations, Ours obtains the highest preference rate. For Lived-in Plausibility, its per-scene preference ranges from 68% to 92%.

Participant-level consistency. The aggregate preference trends are also reflected at the participant level. For Prompt Faithfulness, 17/25 participants (68%) select Ours in at least four of five scenes, including 14/25 (56%) who select it in all five. For Realism, 17/25 participants (68%) select Ours in at least four scenes, including 13/25 (52%) in all five. For Lived-in Plausibility, 21/25 participants (84%) select Ours in at least four scenes, 12/25 (48%) select it in all five, and all 25 participants select Ours in at least one scene.

Effect size. For completeness, we report Cohen's h relative to the 20% uniform-choice baseline,

$$h = 2 \arcsin \sqrt{p} - 2 \arcsin \sqrt{0.2}, \tag{10}$$

Table 6: Per-scene preference rates (%) for the three human-evaluation criteria. Each scene row sums to 100% within a criterion. Bold denotes the highest preference rate in each row.

(a) Prompt Faithfulness

Scene	HSM	Holodeck	LayoutVLM (O.)	SceneSmith	Ours
Restaurant (173)	8.0	0.0	0.0	4.0	**88.0**
Workshop (133)	0.0	0.0	8.0	12.0	**80.0**
Laundromat (143)	24.0	4.0	4.0	8.0	**60.0**
Bedroom (169)	8.0	0.0	0.0	16.0	**76.0**
Bookstore (101)	4.0	12.0	0.0	4.0	**80.0**

(b) Realism

Scene	HSM	Holodeck	LayoutVLM (O.)	SceneSmith	Ours
Restaurant (173)	4.0	0.0	4.0	0.0	**92.0**
Workshop (133)	8.0	4.0	4.0	0.0	**84.0**
Laundromat (143)	16.0	4.0	8.0	4.0	**68.0**
Bedroom (169)	4.0	0.0	0.0	28.0	**68.0**
Bookstore (101)	8.0	4.0	0.0	8.0	**80.0**

(c) Lived-in Plausibility

Scene	HSM	Holodeck	LayoutVLM (O.)	SceneSmith	Ours
Restaurant (173)	0.0	4.0	0.0	4.0	**92.0**
Workshop (133)	0.0	4.0	8.0	4.0	**84.0**
Laundromat (143)	12.0	8.0	4.0	8.0	**68.0**
Bedroom (169)	4.0	0.0	0.0	20.0	**76.0**
Bookstore (101)	4.0	0.0	0.0	4.0	**92.0**

Table 7: Statistical summary for Ours in the human study. The chance level of the five-way comparison is 20%.

Criterion	Preference (%)	Cohen's h	Exact p
Prompt Faithfulness	76.8	1.209	0.00032
Realism	78.4	1.248	0.00032
Lived-in Plausibility	82.4	1.348	0.00032

where p is the aggregate preference rate of Ours. The resulting effect sizes are 1.209 for Prompt Faithfulness, 1.248 for Realism, and 1.348 for Lived-in Plausibility. Table 7 summarizes these values together with the exact case-level permutation results.

A.9 Representative End-to-End Generation Trace

We provide a trace from a completed Japanese-style living-room generation (case 032), which is separate from the five scenes used in the user study. The input specifies a coffee table next to the window with two floor cushions, a sofa across from the window, and a table with a teapot in front of the sofa. From this input, the behavior planner proposes four activities: preparing a low tea setting, serving tea and conversing from floor seating, resetting the tea setting, and quiet reading beside the window.

The pipeline assembles a coffee-table group with the two cushions and a sofa group containing the serving table and teapot. At room scale, it carries the associated use-space regions during group placement. The final validation records all prompt-hard members as present and the required table-in-front-of-sofa relation as satisfied. Figure 8 shows the local group, the room-level soft-space overlay, and the resulting furnished room.

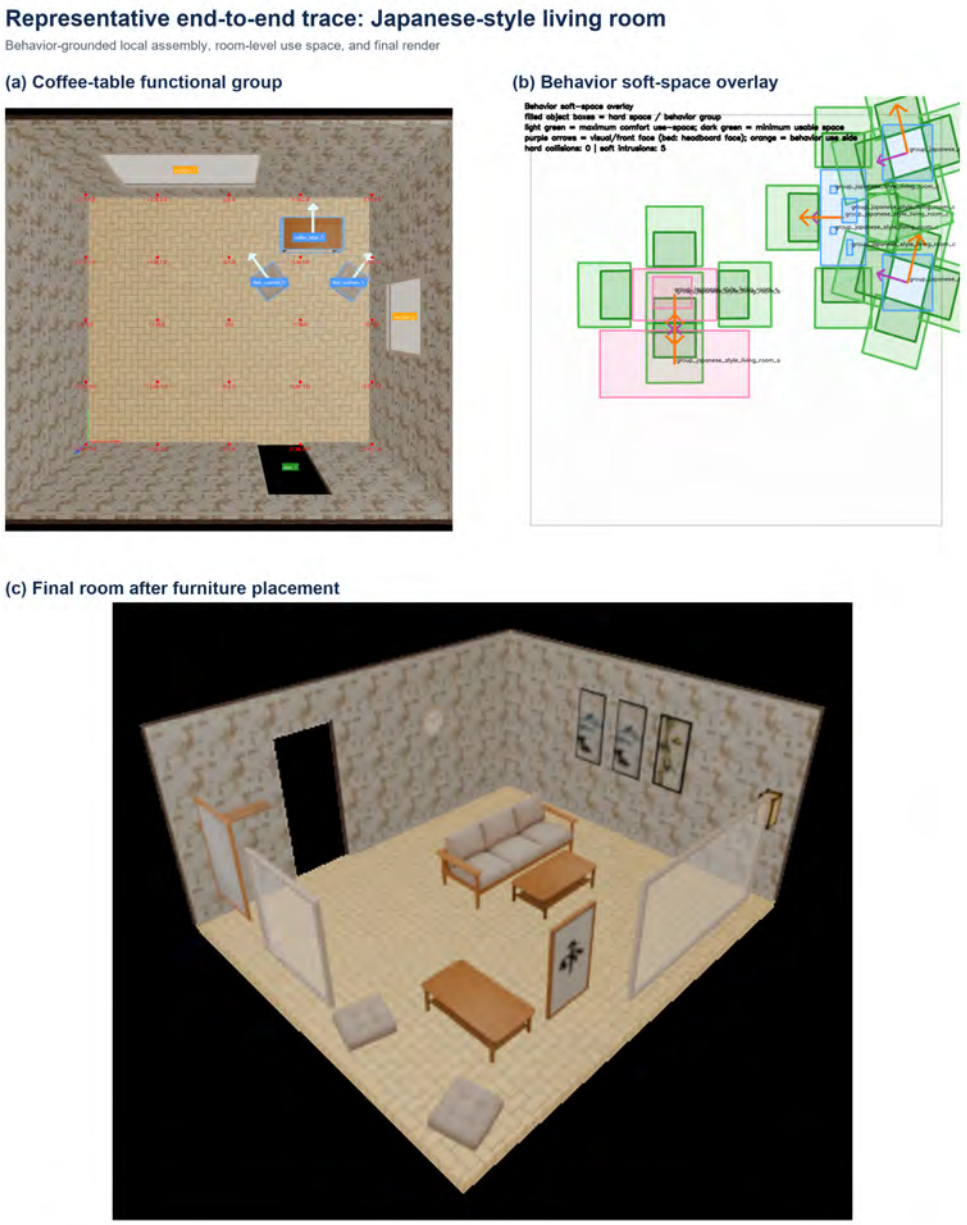

Figure 8: Representative end-to-end trace for a Japanese-style living room (case 032). (a) The coffee-table functional group contains the prompt-required coffee table and two floor cushions. (b) The room-level overlay visualizes hard object boxes, behavior-dependent use spaces, and interaction directions. (c) Final rendering after furniture placement.

A.10 NINE-VIEW VISUALIZATIONS OF USER-STUDY CASES

We provide nine views of our generated scenes for the five representative cases used in the human study. Each contact sheet is arranged from left to right and top to bottom as follows: front, back, left, right, top, diagonal front-left, diagonal front-right, diagonal back-left, and diagonal back-right. The source renders are 2560×2560 pixels.



Figure 9: Nine-view visualization of our generated restaurant scene (case 173).



Figure 10: Nine-view visualization of our generated bedroom scene (case 169).



Figure 11: Nine-view visualization of our generated workshop scene (case 133).

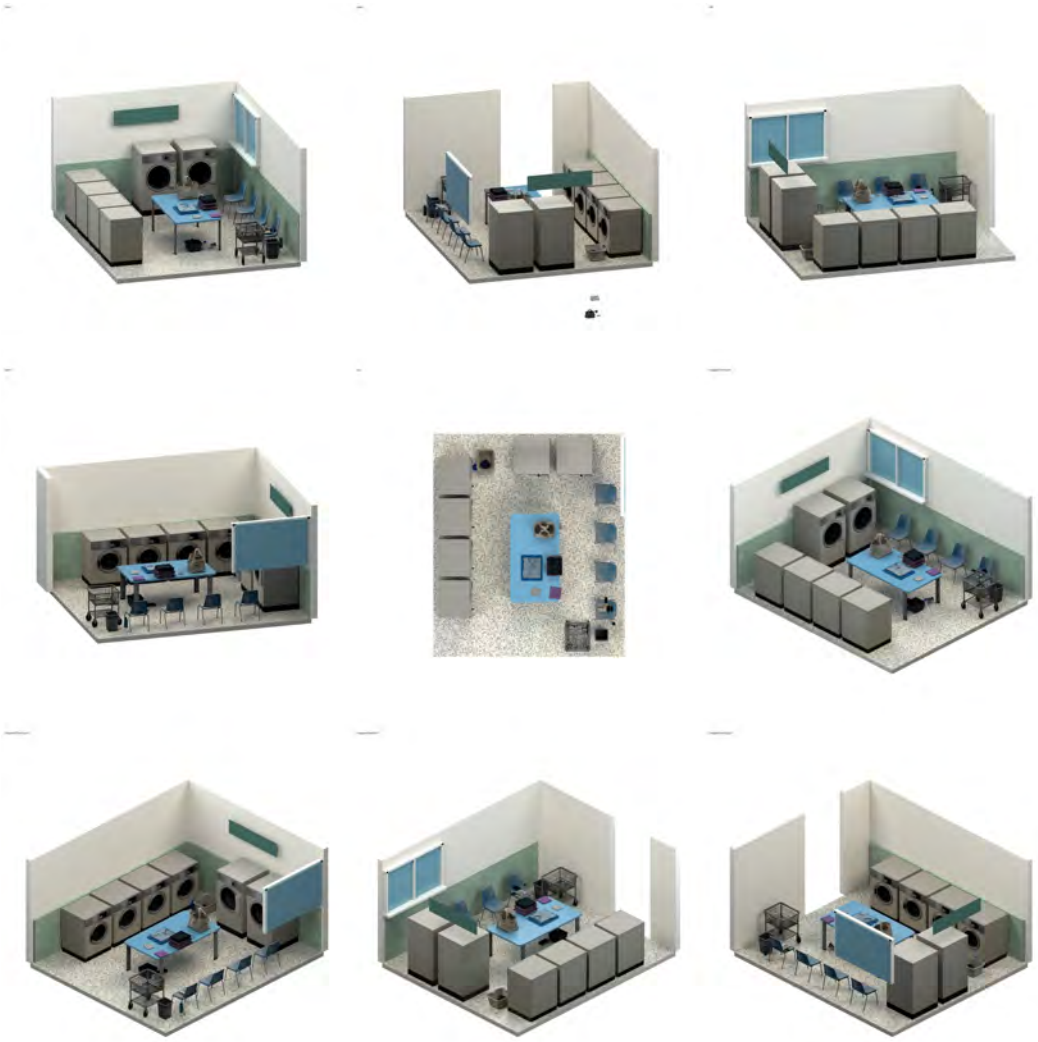

Figure 12: Nine-view visualization of our generated laundromat scene (case 143).



Figure 13: Nine-view visualization of our generated bookstore scene (case 101).