SceneCritic: A Symbolic Evaluator for 3D Indoor Scene Synthesis

Anonymous CVPR submission

Paper ID 20

Abstract

Scaling spatial intelligence in embodied agents demands environments that capture rich compositional structure and precise spatial relationships and can support such diverse task requirements. Large Language Models (LLMs) and Vision-Language Models (VLMs) increasingly generate indoor scenes via intermediate structures like layouts and scene graphs, yet evaluation still relies on LLM or VLM judges whose scores are sensitive to viewpoint, prompt phrasing, and hallucination. This instability makes it hard to disentangle spatial plausibility from evaluation artifacts. We introduce ***SceneCritic****, a symbolic evaluator for floor-plan-level layouts grounded in* ***SceneOnto****, a structured spatial ontology aggregated from 3D-FRONT, ScanNet, and Visual Genome. SceneCritic jointly verifies semantic, orientation, and geometric coherence across object relationships, providing object- and relationship-level assessments that pinpoint specific violations. We further propose an iterative refinement testbed that probes how models revise spatial structure under three critic modalities: a rule-based collision critic, an LLM critic operating on layout text, and a VLM critic operating on rendered observations. Extensive experiments show that (a) SceneCritic aligns substantially better with human judgments than VLM-based evaluators, (b) text-only LLMs can outperform VLMs on semantic layout quality, and (c) image-based VLM refinement is most effective for semantic and orientation correction.*

1. Introduction

The generation of 3D indoor environments has become central to a range of applications, from training embodied agents that must navigate and manipulate objects in realistic spaces [11, 20, 30, 37], to virtual reality and robotics simulation. Recent work has demonstrated that Large Language Models (LLMs) and Vision-Language Models (VLMs) can serve as powerful priors for this task, leveraging their world knowledge to produce object layouts that are both diverse and semantically meaningful [4, 14, 32, 42]. These approaches generate scenes through explicit structured representations (e.g., floor plans, scene graphs, or sets of spatial constraints), often coupled with refinement strategies and constraint-based optimization to improve physical and semantic plausibility before asset placement and rendering. Implicitly, this pipeline frames scene generation as a test of whether a model can construct and maintain a coherent spatial representation that remains consistent across successive placement and refinement steps.

Despite this progress, the evaluation of generated scenes remains surprisingly underexplored. The most common protocol is to ask a VLM to judge a small number of rendered views, but such evaluators are unstable: they miss object relations due to viewpoint and occlusion, are prone to hallucination, and are highly sensitive to prompt design, where minor rephrasing can substantially alter scores [23]. As Figure 1 illustrates, a VLM evaluator assigns scores of 75% and 55% to the same scene depending on the rendered view, highlighting the need for view-independent evaluation. Several works acknowledge these limitations and resort to human user studies [24, 28, 29, 46], which are more reliable but expensive to scale and still produce only single scores or coarse rankings without identifying which spatial constraints were satisfied or violated.

As the field of 3D scene generation continues to grow rapidly, especially in embodied and interactive settings [7, 24, 36, 40], there is a pressing need for a trustworthy and comprehensive evaluation framework. However, designing such an evaluator is far from straightforward, as spatial reasoning involves multiple interrelated concepts including semantic compatibility, geometric property constraints, and inter-object relationships [6]. Existing programmatic metrics have addressed specific dimensions of this problem but remain narrow in scope. Distributional statistics such as FID, KL divergence over furniture categories, and out-of-bound rates [14] capture global similarity but cannot diagnose individual placements. Physics-oriented constraints such as collision avoidance, room-layout IoU, and reachability [40] measure geometric feasibility while leaving semantic coherence unaddressed. Other systems combine collision-free and in-boundary scores with LLM ratings [8, 44], but still outsource semantic judgment to a

CVPR #20

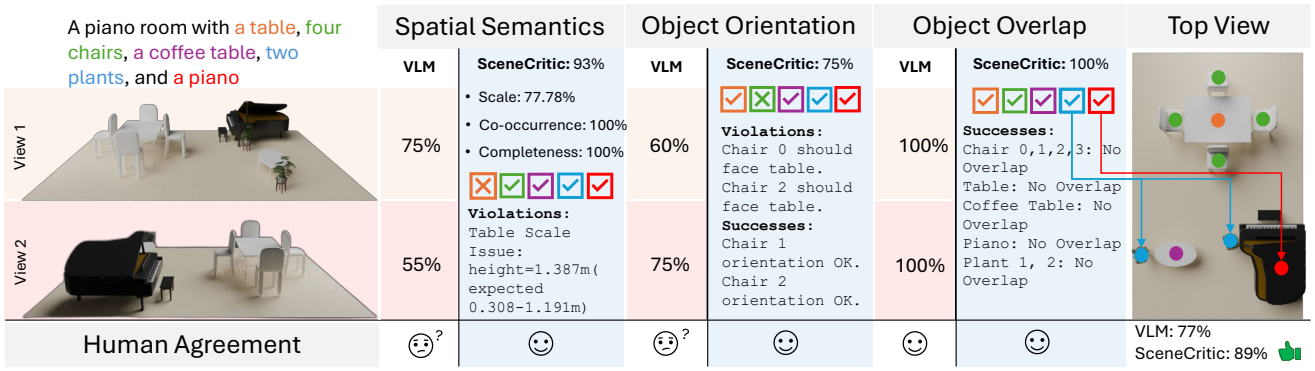

Figure 1. **VLM Evaluation vs. *SceneCritic*.** Given a piano room with nine objects, the VLM assigns different scores to the same scene depending on which view is rendered, while *SceneCritic* evaluates the layout directly by traversing relational constraints from *SceneOnto* (e.g., incorrect table scales, chairs not facing the table they co-occur with). This produces stable, object-level assessments that identify specific violations alongside successful placements. *SceneCritic* scores also align more closely with human preferences.

model-based judge. Among existing frameworks, SceneEval [33] goes furthest by combining text-scene fidelity metrics with physics-based plausibility checks, but does not evaluate semantic layout coherence such as whether object co-occurrences, scales, and orientations reflect plausible indoor configurations. Across this landscape, semantic layout plausibility is consistently either delegated to LLM and VLM judges, or completely unaddressed.

In this paper, we introduce *SceneCritic*, a symbolic evaluator for indoor floor-plan layouts. *SceneCritic*'s constraints are grounded in *SceneOnto*, a structured spatial ontology dataset we construct by aggregating priors from 3D-FRONT, ScanNet, and Visual Genome, encoding object co-occurrence statistics, expected scales, and canonical orientations. While prior work has encoded indoor spatial knowledge implicitly through learned embeddings [47], per-scene graph construction [16], or hand-coded procedural rules [11], these representations are designed for generation or perception, not evaluation. *SceneOnto* instead constructs explicit relational graphs of conditional co-occurrence probabilities, scale distributions, and canonical orientations per room type, aggregated across multiple datasets. This structure is what enables *SceneCritic* to traverse object relationships during evaluation and localize violations to specific object pairs. From these priors, *SceneCritic* derives interpretable constraints that score layouts along three axes: semantic coherence, orientation correctness, and overlap. Because all scene generation pipelines produce a layout before rendering, *SceneCritic* targets this shared intermediate representation directly. By evaluating at the object and relationship level, it explicitly identifies which objects succeed or violate which constraints.

A stable symbolic evaluator also enables a deeper question: *how do models with different post-training objectives build and revise spatial structure over successive refinement steps?* We pair *SceneCritic* with an iterative refinement testbed in which a model repeatedly improves an initial layout under one of several critic modalities: *(a)* a rule-based critic using collision constraints as feedback, *(b)* an LLM critic operating on the layout as text, or *(c)* a VLM critic operating on rendered observations. Because *SceneCritic* scores every intermediate layout, the test bed produces refinement trajectories, making it possible to study which errors different models can correct, where they plateau, and whether fixing one violation introduces new ones elsewhere.

Our key contributions are: *(i)* we construct *SceneOnto*, a dataset-grounded ontology for indoor layout evaluation, including object dimensions, co-occurrence statistics, support surfaces, and orientation preferences; *(ii)* we develop *SceneCritic*, a symbolic evaluator for indoor floor-plan layouts with interpretable constraints over semantic coherence, orientation correctness, and overlap, and validate it against human judgment, showing it aligns substantially better than VLM-based evaluators; *(iii)* we introduce an iterative refinement testbed using heuristic, LLM, and VLM critics to probe how models trained under different post-training objectives build, maintain, and revise spatial structure over multiple correction steps; *(iv)* we use *SceneCritic* to identify comparative strengths and failure modes across critics and post-training regimes, showing that some text-only LLMs outperform VLMs on semantic layout quality, and that image-based VLM refinement is the most effective critic modality.

2. Related Works

LLM and VLM-Driven 3D Scene Generation.

Recent work has increasingly adopted LLMs and VLMs for 3D indoor scene generation, leveraging their spatial reasoning and world knowledge to produce structured layouts from text descriptions. LayoutGPT [14] uses in-context learning to directly predict numerical layouts,

SceneCraft [18] generates Blender code via scene graphs with VLM-based iterative refinement, and I-Design [5] leverages multi-agent LLM collaboration with scene graph mechanisms. Holodeck [42] uses GPT-4 for fully automated environment generation with constraint-based placement, while LLplace [41] and FlairGPT [25] refine LLM-driven layout generation through supervised fine-tuning and detailed object descriptions, respectively. More recently, LayoutVLM [32] combines VLM semantic knowledge with differentiable optimization for physically plausible layouts, and Holodeck 2.0 [4] introduces vision-language-guided generation with interactive editing. A parallel line of work focuses on improving physical plausibility, PhyScene [40], integrates physics-based guidance for collision avoidance and object reachability, PhysGaussian [38] couples physics simulation with 3D Gaussian Splatting, and OptiScene [43] applies Direct Preference Optimization to align layouts with human preferences and physical constraints.

VLM-Based Evaluation and Its Limitations. Several scene generation works [5, 32] adopt VLM-based evaluation, asking a model to score rendered views of generated scenes. However, VLM judges suffer from well-documented limitations: hallucinations are pervasive across large vision-language models [3, 19, 26], with studies showing that even GPT-4V fails on basic visual patterns [34] and that statistical biases and language priors cause systematic object hallucinations [22]. IR3D-Bench [27] further exposes limitations in VLM spatial precision through active 3D reconstruction. On the programmatic side, LayoutGPT [14] reports distributional statistics (FID, KL divergence) and out-of-bound rates, PhyScene [40] introduces collision and reachability constraints, and frameworks like AutoLayout [8] and OptiScene [43] combine geometric scores with LLM ratings. The closest work to ours is SceneEval [33], which introduces a structured evaluation framework with both fidelity metrics (object counts, attributes, spatial relationships relative to the input text) and plausibility metrics (collision, support, navigability, accessibility), grounded in SceneEval-500, a dataset of human-annotated scene descriptions. However, SceneEval still relies on an LLM for object category matching within its pipeline, its plausibility metrics are physics-oriented and do not cover semantic layout coherence such as co-occurrence plausibility, scale correctness, or canonical orientation, and it evaluates only final scenes rather than refinement trajectories. *SceneCritic* addresses these gaps by grounding evaluation in dataset-derived relational priors and producing per-object, per-relationship assessments.

3. *SceneOnto* Construction

SceneCritic derives its constraints from a structured relational representation of indoor scenes. We construct *SceneOnto* to provide this representation by aggregating spatial and semantic priors from three complementary datasets: 3D-FRONT [15], which provides detailed geometric annotations for 6,813 professionally designed indoor scenes; ScanNet [10], which contributes real-world RGB-D reconstructions of 1,511 indoor environments; and Visual Genome [21], which supplies object co-occurrence and support relationships from 61,530 annotated images. From these sources, we extract four types of relational priors: object dimensions, support relations, co-occurrence statistics, and orientation relationships. The resulting ontology is organized as a per-room-type relational graph, where nodes represent object categories and edges encode conditional co-occurrence probabilities, enabling *SceneCritic* to traverse object relationships during evaluation. Figure 2 illustrates a sub-graph of *SceneOnto* for three common room types: bedroom, dining room, and living room.

Dimension Extraction. We extract object dimensions (width, height, and depth in meters) from 3D-FRONT and ScanNet. For 3D-FRONT, we compute bounding boxes from scene annotations with instance-specific scale parameters applied to derive metric estimates. For ScanNet, we identify per-object mesh vertices using segmentation and aggregation annotations and compute bounding boxes from the grouped vertices. For each object category, we record percentile statistics (p5, p25, median, p75, p95) along with the mean, standard deviation, and number of observations. These distributions form the geometric constraints that *SceneCritic* uses for scale verification. Figure 3 (left) visualizes the distribution of object dimensions, showing the expected scale ranges across categories.

Support Relations. Support refers to the surface on which an object rests. We extract support information from 3D-FRONT and ScanNet through geometric reasoning: objects are processed in ascending order of their vertical position, and each is evaluated against previously placed objects whose top surface aligns with the current object's bottom within a 5 cm tolerance. Objects touching the ground are labeled as floor-supported. For Visual Genome, we retain annotated predicates corresponding to physical support interactions (e.g., *on*, *sitting on*, *standing on*) and apply additional filtering to reduce annotation noise, including removing impossible supporting surfaces, rejecting most same-category supports, enforcing that heavier objects cannot be supported by lighter ones, and applying a spatial sanity check based on bounding box positions.

Co-occurrence Statistics. We compute pairwise object co-occurrence across all scenes, both globally and per room type, for all three datasets. For each scene, we extract the set of unique object categories present and count how many scenes contain both in the ordered pair (a, b) with a certain threshold distance. We derive the conditional probability $P(b|a)$, and normalized pointwise mutual information ($nPMI$), defined as $PMI(a, b)/(-\log P(a, b))$, which

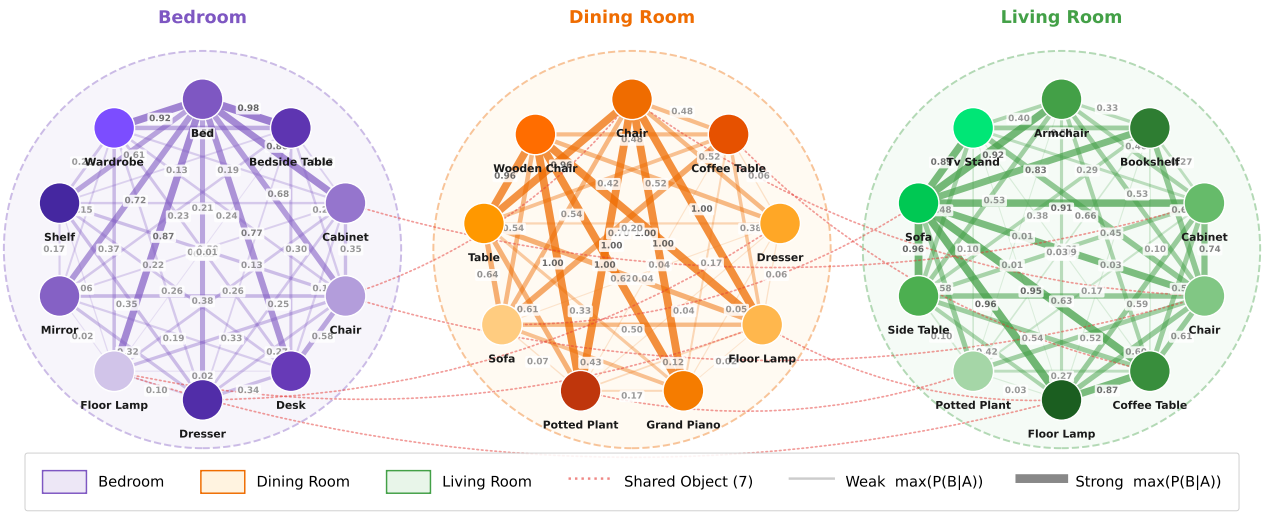

Figure 2. **Relational sub-graph from *SceneOnto* for three common room types.** Nodes represent object categories, with size proportional to frequency. Edges encode conditional co-occurrence probabilities $max(P(B|A))$, with thicker edges indicating stronger associations. Dotted red lines connect shared object categories across room types. *SceneCritic* traverses this structure during evaluation to verify whether object co-occurrences in a generated layout are plausible for the specified room type.

ranges over $[-1, 1]$. These statistics form the edges in *SceneOnto*'s relational graph and are what *SceneCritic* uses to assess co-occurrence plausibility and proximity. Figure 3 (center) plots co-occurrence frequency against semantic association (nPMI), revealing that 73.6% of the 489 object-pair edges have positive association.

Orientation Relationships. We extract orientation statistics exclusively from 3D-FRONT, which provides per-object quaternion rotations and room floor geometry. We compute three types of orientation relationships. **Back-to-wall:** for categories such as sofa, bed, dresser, and bookshelf, we compute the angular difference between the object's back direction and the inward normal of the nearest wall segment. **Faces-center:** for categories such as sofa, chair, and TV, we compute the angular deviation between the object's facing direction and the direction toward the room centroid. **Faces-pair:** for each object, we compute the angular deviation between its facing direction and the direction toward every other object within a 5 m radius. The facing direction is derived from the object's yaw. For each relationship type, the ontology records the fraction of observations satisfying the constraint, along with the mean angular deviation, mean distance, and number of samples. Figure 3 (right) shows the placement strategy landscape, with objects distributed along back-to-wall and faces-center fractions.

4. *SceneCritic* as a Symbolic Evaluator

SceneCritic evaluates generated layouts by traversing the relational structure encoded in *SceneOnto*, verifying whether the spatial arrangement of objects is consistent with the priors derived from real indoor scenes. Following the evaluation axes adopted by prior scene generation work [4, 29, 32], *SceneCritic* assesses layouts along three dimensions: *(a)* spatial semantics, *(b)* orientation correctness, and *(c)* overlap. The key difference is that where prior methods rely on VLM judges to score these criteria from rendered views, *SceneCritic* resolves them symbolically by traversing *SceneOnto*'s constraints at the object and relationship level.

Spatial Semantics. Spatial semantics evaluates whether the objects in a scene are consistent with their expected semantic relationships. This includes whether objects are positioned in ways that align with their functional roles and typical real-world usage patterns. Compositionally, it evaluates whether multiple objects combine to form coherent functional groupings through their joint spatial arrangement, such as a desk and chair forming a workspace or a bed, nightstand, and lamp forming a bedside arrangement. Our spatial semantics verifier evaluates five sub-criteria: *scale*, *co-occurrence*, *plausibility*, *proximity*, and *completeness*.

Orientation Verification. The orientation verifier checks whether each object's orientation aligns with its expected placement as defined by *SceneOnto*. It evaluates three aspects: whether an object has its back to the nearest wall, whether it faces the room center, and whether it faces another object it is expected to be oriented toward. Each check compares the object's yaw-derived facing direction against the target direction, with the ontology specifying which checks apply to which object categories and

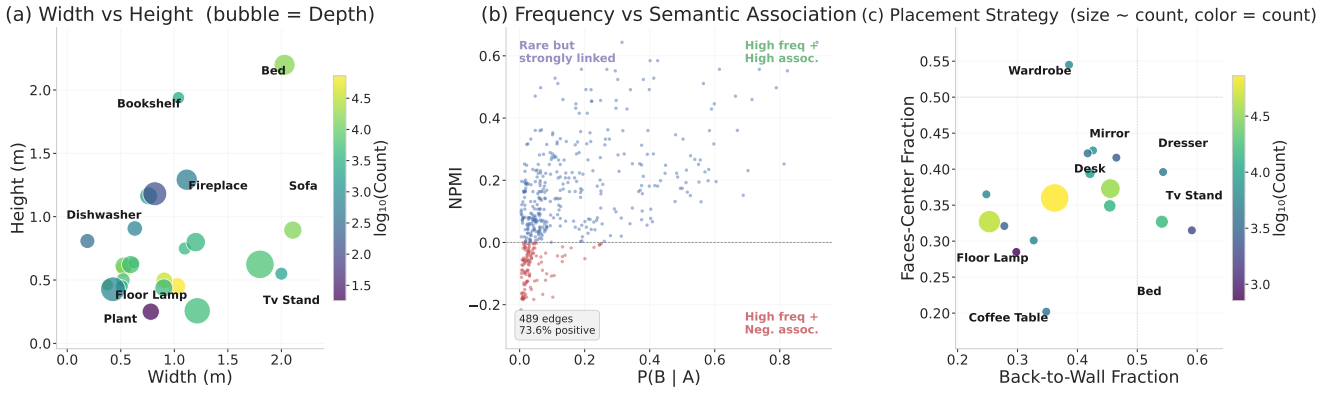

Figure 3. Summary statistics from *SceneOnto*. (Left) Object dimension distributions, with bubble size encoding depth. (Center) Co-occurrence frequency vs. semantic association (nPMI): 73.6% of 489 object-pair edges show positive association. (Right) Object placement strategies, showing the fraction of instances with back-to-wall vs. faces-center orientation (e.g., objects like beds and coffee tables cluster toward back-to-wall, while wardrobes and mirrors favor faces-center).

the expected angular tolerances.

Overlap Detection. The overlap verifier checks whether objects violate basic physical constraints. It detects two types of violations: Proximity Overlap and True Overlap. Proximity Overlap detects when coarse spatial footprints of objects encroach on each other, including cases where objects are merely in close proximity without true geometric contact. True Overlap reports only cases where objects genuinely occupy shared space, filtering out false positives from loose bounding approximations.

5. Experiments Setup

Baseline Methods. We test *SceneCritic* on three representative 3D scene layout generation methods that span different generation strategies: LayoutGPT [14], which uses in-context learning to predict layouts directly; Holodeck [42], which applies constraint-based optimization over LLM-generated placements; and LayoutVLM [32], which jointly optimizes semantic and physical plausibility via a VLM. For each method, we use two generation backbones: Gemini-2.5-Flash and Qwen2.5-VL-72B.

Baseline Evaluator. As the VLM-based baseline evaluator, we select Gemini-2.5-Pro, a stronger proprietary model than the generation backbones. We render the generated layouts in Blender [1] with Objaverse assets [12] scaled to match the object dimensions specified in the layout. From each rendered scene, we capture 2D images from multiple viewpoints (side, front, rear, and top) and provide them to the evaluator along with a prompt describing the evaluation criteria: semantic consistency, orientation correctness, and overlap.

Evaluation Scenes. We evaluate on two categories of rooms. *Pivotal rooms*: bedroom and living room, are the standard room types used across all baselines and prior work, enabling direct comparison. *Extended rooms*: cover less common settings including bookstore, buffet restaurant, children's room, classroom, computer room, deli, dining room, game room, and florist room, which test generalization beyond the typical evaluation scope.

[1] 3D modeling and rendering package: http://www.blender.org

6. Evaluator Reliability Analysis

6.1. VLM Instability

We first examine the reliability of VLM-based evaluation by measuring score variance across viewpoints and repeated evaluations using Gemini-2.5-Pro as the evaluator. Table 1 reports the per-view variance relative to the top-view baseline across all generation methods. Score variance is largest on the metrics where the generated scene contains the most errors. For example, LayoutVLM+Gemini produces frequent overlap errors, which explains its consistently high overlap variance across all views (11.2 Left, 28.1 Right, 17.5 Front). Additionally, left and right views represent mirrored perspectives of the same scene and should receive similar scores, yet orientation variance differs substantially between them. LayoutVLM shows high variance in the right view but low variance in the left, while other methods exhibit the opposite pattern. This asymmetry suggests the evaluator is reacting to 2D visual appearance rather than extracting 3D spatial relationships. Re-evaluating the same top-view image under identical conditions produces substantial variance: the semantic score varies by 12.46 for LayoutGPT+Gemini and 19.32 for LayoutGPT+Qwen72B. If an evaluator cannot produce consistent scores for the same input, downstream comparisons between generation methods become unreliable. We find that method rankings reverse depending on which viewpoint is chosen for evaluation.

Table 1. **Per-view variance relative to the top-view baseline.** Each entry is computed as $\frac{(\text{View}-\text{Top1})^2}{4}$ for Semantic (Sem), Orientation (Ori), and Overlap (Ovlp). The *Re-eval* column measures variance when the same top-view image is evaluated twice under identical conditions. The remaining columns measure variance between the top-view and alternative viewpoints (Left, Right, Front).

		Re-eval			**Left**			**Right**			**Front**		
Method	Backbone	Sem.	Ori.	Ovlp	Sem.	Ori.	Ovlp	Sem.	Ori.	Ovlp	Sem.	Ori.	Ovlp
Holodeck	Qwen72B	0.11	6.58	3.48	**55.06**	**88.27**	1.89	**30.25**	**75.43**	1.44	**47.68**	**37.33**	0.00
LayoutGPT	Qwen72B	**19.32**	1.09	0.18	3.65	25.00	0.41	2.39	9.46	4.08	16.61	0.24	**46.04**
LayoutVLM	Qwen72B	0.49	4.06	0.38	0.44	0.61	1.31	0.05	6.07	2.09	1.90	0.04	0.37
Holodeck	Gemini	0.30	1.97	0.00	4.69	8.56	0.00	0.42	0.09	7.02	1.22	1.97	1.56
LayoutGPT	Gemini	**12.46**	1.90	1.92	0.12	35.64	**11.36**	0.58	5.25	10.30	1.66	0.02	2.43
LayoutVLM	Gemini	2.07	4.95	**19.27**	4.95	1.39	11.16	0.18	9.74	**28.09**	1.19	0.05	17.47

6.2. Comparison with *SceneCritic*

Given these instabilities, we evaluate the same generated layouts using *SceneCritic*, which operates directly on the 3D layout representation without rendered views. Table 2 compares the two evaluators. Interestingly, while VLM evaluation ranks Holodeck+Qwen72B highest (59.46 average), *SceneCritic* ranks LayoutVLM+Qwen72B highest (80.32). Figure 4 provides qualitative results; *SceneCritic* produces per-object, per-constraint assessments that identify the specific violations responsible for each score, such as desks not facing nearby chairs or cabinets not backed against walls. These violations are clearly visible in the 3D layout but are inconsistently detected by the VLM evaluator from 2D renders, which explains the divergence in scores between the two evaluators.

6.3. Human Alignment Verification

To validate which evaluator better reflects human judgment, we conduct a pairwise evaluation study with 16 annotators providing 594 judgments across three criteria (semantic consistency, orientation correctness, and overlap) at two difficulty levels (easy and complex rooms). For reference, prior human evaluations in this space typically use 5 annotators [5, 32, 43], while several prominent methods rely entirely on automated metrics [14, 33, 37].

SceneCritic achieves 94.44% agreement with human judgments for easy scenes and 83.33% for complex scenes. The VLM evaluator achieves 58.82% for easy scenes and 47.06% for complex scenes, only marginally above chance. Critically, when *SceneCritic* disagrees with human judgments, the score difference is small (e.g., less than 7 points in the cases where rankings differ). When the VLM evaluator disagrees, the difference is large: Gemini assigns LayoutGPT a 75 points higher score than Holodeck in easy scenes, while human annotators rank Holodeck higher with strong inter-annotator agreement. This pattern is consistent across both difficulty levels. We also observe that the largest VLM disagreements with humans occur on the overlap metric, whereas *SceneCritic* achieves 100% agreement with human judgments on this criterion across all comparisons. Since overlap is the most visually salient spatial error, the VLM's inability to consistently detect it from renders confirms that it is not reliably extracting geometric relationships from 2D views.

We also compute inter-annotator agreement to assess the trustworthiness of our human evaluation study. Fleiss'

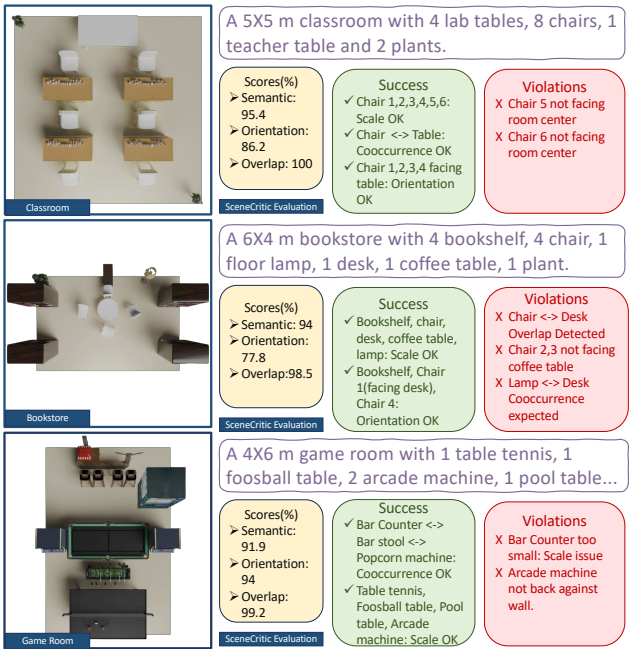

Figure 4. **Qualitative results of *SceneCritic***: *SceneCritic* assessment for extended scenes, reporting semantic, orientation, and overlap scores while identifying specific violations (e.g., desk not backed against wall, cabinets not backed against wall) alongside successful placements across scale, orientation, co-occurrence, proximity, and completeness.

Table 2. Comparison of Baseline methods across VLM-based evaluator and *SceneCritic* on Semantic (Sem), Orientation (Ori), Overlap (Ovlp) and Average (Avg) scores. In *SceneCritic*, Overlap is decomposed into ProxOvlp (Axis-Aligned Bounding Box overlap) and TrueOvlp (Oriented Bounding Box overlap).

Method	Backbone	VLM-Evaluation				*SceneCritic*				
		Sem.	Ori.	Ovlp.	Avg.	Sem.	Ori.	Ovlp.		Avg.
								ProxOvlp	TrueOvlp	
Holodeck	Qwen72B	42.61	**60.61**	**75.17**	**59.46**	76.62	52.75	**95.97**	**95.99**	75.12
LayoutGPT	Qwen72B	**51.09**	56.94	69.02	59.02	75.03	16.83	73.54	73.34	55.10
LayoutVLM	Qwen72B	50.82	45.48	71.06	55.79	**89.46**	**59.38**	91.92	92.31	**80.32**
Holodeck	Gemini	46.88	51.33	**80.91**	**59.71**	74.37	45.05	**97.13**	**97.21**	72.20
LayoutGPT	Gemini	48.18	**53.82**	77.68	59.89	74.70	37.52	76.99	77.79	63.20
LayoutVLM	Gemini	**50.97**	50.09	64.70	55.25	**89.27**	**58.65**	92.45	91.47	**79.96**

kappa values fall within the moderate range for easy rooms (0.30 for semantic, 0.46 for orientation, and 0.39 for overlap) and within the substantial range for complex rooms (0.64 for semantic, 0.59 for orientation, and 0.72 for overlap). In addition, the inter-annotator percent agreement is 81.60% for easy rooms and 91.32% for complex rooms, supporting the reliability of our human evaluation.

7. Probing Spatial Reasoning through Iterative Refinement

To study how models build and revise spatial structure under iterative feedback, we design a three-stage placement pipeline that includes *planning*, *placement*, and *refinement*.

7.1. Our Test Bed

In the **planning stage**, the model receives a placement condition $\mathcal{C}_{\text{env}} = \{\mathcal{C}_{\text{desc}}, \mathcal{C}_{\text{range}}, \mathcal{C}_{\text{objects}}\}$, where $\mathcal{C}_{\text{desc}}$ is a natural language scene description, $\mathcal{C}_{\text{range}}$ defines the spatial boundaries, and $\mathcal{C}_{\text{objects}}$ lists the objects to place. The model outputs a step-by-step placement plan informed by spatial placement principles: hierarchical ordering by object size, semantic asset grouping (e.g., placing bed, nightstand, and lamp together), and navigable space preservation.

In the **placement stage**, the model executes the plan by producing a concrete scene layout. The model is prompted with geometric constraints, including the spatial boundaries of the environment, maintaining realistic object size ratios, and assigning meaningful orientations. The output is a layout specifying each object's scale-aware bounding box, position, orientation, and category label.

In the **refinement stage**, the model iteratively improves the layout under one of several critic modalities. We evaluate three critic types: *(a)* a **heuristic critic** that provides feedback based on three constraint objectives, including whether objects lie within the spatial boundary, whether all required objects are placed, and whether bounding boxes overlap; *(b)* an **LLM critic** that receives the layout as text and provides feedback in natural language; and *(c)* a **VLM critic** that receives rendered observations of the scene. We further divide VLM-based refinement into three input variants: image-only, image+text, and semantics+text. For the heuristic critic, the three objectives are combined into a single reward score with per-object feedback, allowing the model to iteratively update the layout toward a valid configuration. For LLM and VLM critics, feedback is derived from the corresponding model and the layout is modified accordingly. In all cases, *SceneCritic* scores every intermediate layout, producing refinement trajectories that reveal how each model responds to criticism across successive correction steps.

Table 3. Models categorized by post-training strategy and model parameters.

Model	Category	Post-Training Details	Params
Qwen3-14B [39]	General RL	GRPO-style reinforcement learning	14B
Qwen3-235B [39]	General RL	GRPO-style reinforcement learning	235B
UI-Venus-Navi-72B [17]	General RL	GRPO-based reasoning optimization	72B
Gemini-2.5-flash [9]	RLAIF + RLHF	SFT + Reward Model + RL	N/A
Qwen2.5-VL-7B-MM-RLHF [13]	RLHF	PPO-style human feedback alignment	7B
Qwen2.5-72B-VL [2]	RLHF	SFT + DPO (preference optimization)	72B
LLaMA4 Maverick [1]	RLHF	SFT + Online RL + DPO	∼17B (active)
Qwen3-14B-Intuitor-MATH-1EPOCH [45]	RLIF	Iterative feedback RL (Intuitor)	14B
Qwen3-14B-GRPO-MATH-1EPOCH [45]	RLIF	GRPO under RLIF objective	14B
Qwen2.5-14B-GRPO [31]	RLVR	GRPO with verifiable reward	14B
VL-Reasoner-72B [35]	RLVR	GRPO + Self-Selective Revision (SSR)	72B
Holo1.5-72B	RLVR	GRPO-based verifiable reward RL	72B
DeepSeek-3.2V	RLVR	GRPO-style reasoning RL	37B(active)

7.2. Models Evaluated

Table 3 summarizes the 13 models evaluated, spanning four post-training categories: general RL (GRPO-style training), RLHF (PPO or DPO-based preference alignment), RLIF (iterative feedback RL), and RLVR (RL with verifiable rewards). Model sizes range from 7B to 235B parameters.

Table 4. Comparison across refinement variants using MLLM backbones with *SceneCritic*. Overlap is decomposed into ProxOvlp (Axis-Aligned Bounding Box overlap) and TrueOvlp (Oriented Bounding Box overlap).

Method	Semantic Verifier Evaluation				
	Sem.	Ori.	Overlap		Avg.
			ProxOvlp	TrueOvlp	
Heuristic					
Gemini-2.5-flash	76.3	43.7	**99.3**	**98.9**	73.03
Qwen2.5-72B-VL	72.2	48.1	91.3	91.3	70.53
DeepSeek-3.2V	**82.6**	64.4	95.3	95.7	**80.83**
LLaMA4 Maverick	72.6	48.7	95.3	94.1	72.00
Qwen3-235B	75.0	47.1	98.2	97.8	73.37
Qwen3-14B	73.4	44.6	91.4	91.8	69.87
Qwen2.5-VL-7B-MM-RLHF	64.0	47.6	88.3	88.7	66.70
Qwen2.5-14B-GRPO	67.3	46.6	85.2	85.6	66.43
Qwen3-14B-GRPO-MATH-1EPOCH	63.8	**65.1**	86.2	87.4	71.90
Qwen3-14B-Intuitor-MATH-1EPOCH	65.8	34.3	87.4	87.9	62.58
Holo1.5-72B	68.3	48.7	90.0	90.0	69.00
VL-Reasoner-72B	68.9	42.1	95.6	95.4	68.83
UI-Venus-Navi-72B	66.5	36.6	91.7	91.7	64.93
LLM					
Gemini-2.5-flash	74.2	50.7	**98.3**	**98.4**	74.42
Qwen2.5-72B-VL	72.5	48.6	89.5	89.3	70.17
DeepSeek-3.2V	**78.9**	62.8	97.1	96.9	**79.57**
LLaMA4 Maverick	69.9	44.7	89.9	89.5	68.10
Qwen3-235B	76.9	54.6	96.5	95.7	75.87
Qwen3-14B	73.3	49.2	93.4	93.4	71.97
Qwen2.5-VL-7B-MM-RLHF	65.5	56.9	90.4	90.6	70.97
Qwen2.5-14B-GRPO	67.9	50.7	86.2	86.6	68.33
Qwen3-14B-GRPO-MATH-1EPOCH	63.8	**65.5**	68.2	68.9	65.95
Qwen3-14B-Intuitor-MATH-1EPOCH	68.4	43.3	82.6	82.7	64.78
Holo1.5-72B	66.9	42.1	80.9	81.4	63.38
VL-Reasoner-72B	72.8	47.4	91.0	91.0	70.40
UI-Venus-Navi-72B	72.6	45.0	89.6	89.6	69.07
Image					
Gemini-2.5-flash	**78.7**	**55.8**	**97.5**	**96.6**	**77.18**
Qwen2.5-72B-VL	71.2	49.1	89.0	88.2	69.63
Holo1.5-72B	68.5	50.2	84.8	85.1	67.88
VL-Reasoner-72B	71.0	44.0	87.2	87.2	67.40
UI-Venus-Navi-72B	72.2	49.1	87.7	87.1	69.57
Img+Text					
Gemini-2.5-flash	**76.7**	48.4	**95.7**	**96.3**	**73.70**
Qwen2.5-72B-VL	71.0	41.8	89.4	89.0	67.33
Holo1.5-72B	67.9	49.1	83.6	84.3	66.98
VL-Reasoner-72B	72.4	44.1	91.1	90.9	69.17
UI-Venus-Navi-72B	72.0	**51.8**	90.8	90.6	71.50
Sem+Text					
Gemini-2.5-flash	**75.8**	**53.0**	**98.6**	**98.6**	**75.80**
Qwen2.5-72B-VL	72.3	44.6	90.4	89.9	69.02
Holo1.5-72B	67.6	45.8	77.3	77.8	63.65
VL-Reasoner-72B	72.8	50.2	88.2	88.4	70.43
UI-Venus-Navi-72B	70.7	48.6	90.2	90.3	69.85

7.3. Results

Text-only LLMs can outperform VLMs on spatial reasoning: DeepSeek-3.2V achieves the highest average scores under both heuristic refinement (80.83) and LLM-based refinement (79.57), outperforming proprietary VLMs such as Gemini-2.5-Flash. This is driven by strong semantic scores (82.6 heuristic, 78.9 LLM), suggesting that text-based spatial reasoning can be more effective than visual-semantic knowledge for layout generation. **Orientation is the most challenging metric:** Across all models and refinement methods, orientation scores remain consistently low relative to semantic and overlap scores. Notably, Qwen2.5-14B-GRPO, a relatively small 14B model trained with a math-oriented GRPO objective, achieves the best orientation scores (65.1 heuristic, 65.5 LLM) by a significant margin, outperforming much larger models. We hypothesize this is because orientation verification requires precise angular computation, which may benefit from math-oriented training objectives. **Image-based refinement is the most effective critic modality:** Using Gemini-2.5-Flash as a consistent backbone across all refinement methods, image-based VLM refinement yields the strongest performance on semantic and orientation metrics. Orientation scores increase from 43.70 (heuristic) to 50.7 (LLM) to 55.8 (image-based), an 12-point improvement. A similar pattern holds for UI-Venus-Navi-72B, whose orientation score rises from 36.6 (heuristic) to 49.1 (image-based). These results suggest that visual feedback provides richer spatial information for correcting placement and orientation than text-only inputs. Combining text with images can degrade performance, as the additional textual input may cause the model to prioritize textual cues over spatial information available in the rendered views.

8. Conclusion

We introduced *SceneCritic*, a symbolic evaluator for indoor floor-plan layouts, and *SceneOnto*, a structured spatial ontology constructed from 3D-FRONT, ScanNet, and Visual Genome. Together, they provide stable, interpretable, object-level evaluation of semantic coherence, orientation correctness, and overlap, without relying on rendered views or model-based judges. *SceneOnto* is inherently extensible; by providing a simple mapping from a new dataset into our ontology format, novel room types and object categories can be readily integrated, and *SceneCritic* immediately supports these additions. Our experiments show that *SceneCritic* aligns substantially better with human judgments than VLM-based evaluators, particularly on complex scenes where VLM agreement with humans drops to 47.06%. Through our iterative refinement testbed, we find that text-only LLMs can outperform VLMs on semantic layout quality, that math-oriented training objectives improve orientation reasoning even at small model scales, and that image-based VLM refinement is the most effective critic modality for semantic and orientation correction.

References

[1] Meta AI. Introducing llama 4: Advancing multimodal intelligence, 2024. 7

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu,

Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 7

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 3

[4] Zixuan Bian, Ruohan Ren, Yue Yang, and Chris Callison-Burch. Holodeck 2.0: Vision-language-guided 3d world generation with editing. *arXiv preprint arXiv:2508.05899*, 2025. 1, 3, 4

[5] Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, and Xi Wang. I-design: Personalized llm interior designer. In *European Conference on Computer Vision*, 2024. 3, 6

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[7] ChunTeng Chen, YiChen Hsu, YiWen Liu, WeiFang Sun, TsaiChing Ni, ChunYi Lee, Min Sun, and YuanFu Yang. Scenefoundry: Generating interactive infinite 3d worlds. *arXiv preprint arXiv:2601.05810*, 2026. 1

[8] Weixing Chen, Dafeng Chi, Yang Liu, Yuxi Yang, Yexin Zhang, Yuzheng Zhuang, Xingyue Quan, Jianye Hao, Guanbin Li, and Liang Lin. Autolayout: Closed-loop layout synthesis via slow-fast collaborative reasoning. *arXiv preprint arXiv:2507.04293*, 2025. 1, 3

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7

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

[11] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. Outstanding Paper Award. 1, 2

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 5

[13] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024. 7

[14] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 5, 6

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

[16] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 2

[17] Zhangxuan Gu, Zhengwen Zeng, Zhenyu Xu, Xingran Zhou, Shuheng Shen, Yunfei Liu, Beitong Zhou, Changhua Meng, Tianyu Xia, Weizhi Chen, et al. Ui-venus technical report: Building high-performance ui agents with rft. *arXiv preprint arXiv:2508.10833*, 2025. 7

[18] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In *Forty-first International Conference on Machine Learning*, 2024. 3

[19] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 3

[20] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017. 1

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

[22] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[23] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. TopViewRS: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. 1

[24] Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Zhaoshuo Li. Scenethesis: A language and vision

agentic framework for 3d scene generation. *arXiv preprint arXiv:2505.02836*, 2025. 1

[25] Gabrielle Littlefair, Niladri Shekhar Dutt, and Niloy J Mitra. Flairgpt: Repurposing llms for interior designs. In *Computer Graphics Forum*, 2025. 3

[26] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 3

[27] Parker Liu, Chenxin Li, Zhengxin Li, Yipeng Wu, Wuyang Li, Zhiqin Yang, Zhenyuan Zhang, Yunlong Lin, Sirui Han, and Brandon Y Feng. Ir3d-bench: Evaluating vision-language model scene understanding as agentic inverse rendering. *arXiv preprint arXiv:2506.23329*, 2025. 3

[28] Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Agentic 3d scene generation with spatially contextualized vlms. *arXiv preprint arXiv:2505.20129*, 2025. 1

[29] Xingjian Ran, Yixuan Li, Linning Xu, Mulin Yu, and Bo Dai. Direct numerical layout generation for 3d indoor scene synthesis via spatial reasoning. *arXiv preprint arXiv:2506.05341*, 2025. 1, 4

[30] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 1

[31] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 7

[32] 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*, 2025. 1, 3, 4, 5, 6

[33] 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. 2025. 2, 3, 6

[34] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 3

[35] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025. 7

[36] Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. Embodiedgen: Towards a generative 3d world engine for embodied intelligence. *arXiv preprint arXiv:2506.10600*, 2025. 1

[37] Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, et al. Sage: Scalable agentic 3d scene generation for embodied ai. *arXiv preprint arXiv:2602.10116*, 2026. 1, 6

[38] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 7

[40] 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*, 2024. 1, 3

[41] Yixuan Yang, Junru Lu, Zixiang Zhao, Zhen Luo, James JQ Yu, Victor Sanchez, and Feng Zheng. Llplace: The 3d indoor scene layout generation and editing via large language model. *arXiv preprint arXiv:2406.03866*, 2024. 3

[42] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3, 5

[43] Yixuan Yang, Zhen Luo, Tongsheng Ding, Junru Lu, Mingqi Gao, Jinyu Yang, Victor Sanchez, and Feng Zheng. Optiscene: Llm-driven indoor scene layout generation via scaled human-aligned data synthesis and multi-stage preference optimization. *arXiv preprint arXiv:2506.07570*, 2025. 3, 6

[44] Yixuan Yang, Zhen Luo, Tongsheng Ding, Junru Lu, Mingqi Gao, Jinyu Yang, Victor Sanchez, and Feng Zheng. Optiscene: Llm-driven indoor scene layout generation via scaled human-aligned data synthesis and multi-stage preference optimization, 2025. 1

[45] Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025. 7

[46] Kaizhi Zheng, Ruijian Zha, Zishuo Xu, Jing Gu, Jie Yang, and Xin Eric Wang. Constructing a 3d scene from a single image. *arXiv preprint arXiv:2505.15765*, 2025. 1

[47] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2