# How Far Are Vision-Language Models from Constructing the Real World? A Benchmark for Physical Generative Reasoning

Luyu Yang*, Yutong Dai, An Yan, Viraj Prabhu
Ran Xu, Zeyuan Chen

Salesforce AI Research, United States
luyu.yang@salesforce.com

**Abstract.** The physical world is not merely visual; it is governed by rigorous structural and procedural constraints. Yet, the evaluation of vision-language models (VLMs) remains heavily skewed toward perceptual realism, prioritizing the generation of visually plausible 3D layouts, shapes, and appearances. Current benchmarks rarely test whether models grasp the step-by-step processes and physical dependencies required to actually *build* these artifacts—a capability essential for automating design-to-construction pipelines. To address this, we introduce **DreamHouse**, a novel benchmark for *physical generative reasoning*: the capacity to synthesize artifacts that concurrently satisfy geometric, structural, constructability, and code-compliance constraints. We ground this benchmark in residential timber-frame construction, a domain with fully codified engineering standards and objectively verifiable correctness. We curate over 26,000 structures spanning 13 architectural styles—each verified to construction-document standards (LOD 350)—and develop a deterministic 10-test structural validation framework. Unlike static benchmarks that assess only final outputs, DreamHouse supports iterative agentic interaction. Models observe intermediate build states, generate construction actions, and receive structured environmental feedback, enabling a fine-grained evaluation of planning, structural reasoning, and self-correction. Extensive experiments with state-of-the-art VLMs reveal substantial capability gaps that are largely invisible on existing leaderboards. These findings establish physical validity as a critical evaluation axis orthogonal to visual realism, highlighting physical generative reasoning as a distinct and underdeveloped frontier in multimodal intelligence. The benchmark is available at `https://luluyuyuyang.github.io/dreamhouse/`.

**Keywords:** World Models · Agentic Evaluation · Physical Validity

## 1 Introduction

The physical world is not merely a surface; it is a rigorous system of constraints [4, 22, 24, 25, 30, 33, 36, 42]. A floor must bear load, a rafter must respect

* Corresponding author.

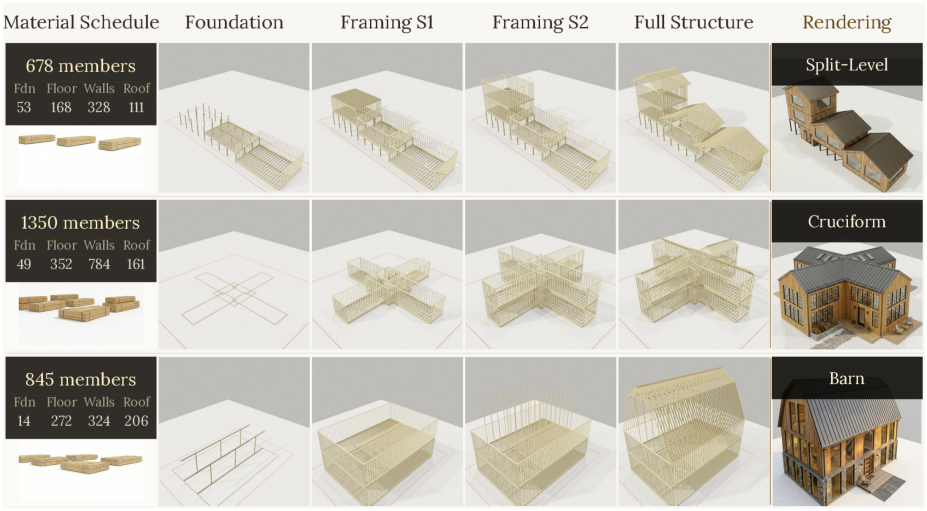


**Fig. 1: DreamHouse benchmark samples.** Each row shows a single structure across five representations: material schedule (member counts by subsystem), foundation, two intermediate framing stages, and complete timber frame, alongside the target exterior rendering from which the model must infer the hidden structural system. Three of the 13 architectural styles are shown: Split-Level (678 members), Cruciform (1,350 members), and Barn (845 members). Member counts span four subsystems: foundation (Fdn), floor, walls, and roof.

its allowable span, and a wall must continuously transfer force from roof to foundation. These are not aesthetic choices, but fundamental physical realities. Yet, the dominant evaluation paradigm for generative vision models treats the world as purely visual, asking only if a model can produce outputs that *look* correct [2, 9, 11, 18, 34, 35]. Whether those outputs could physically *stand* is a question the field has largely left unasked.

While the recent surge in vision-language models (VLMs) has yielded remarkable perceptual capabilities, evaluating their grasp of physical laws remains in its infancy [6, 7, 29, 35, 43]. Recent benchmarks like PhysBench [7] and VSI-Bench [39] probe physical reasoning and spatial recall, but they are fundamentally *comprehension* tasks—the model observes, answers, and is scored. They do not evaluate whether a model can *build*. Generating a physically realizable artifact from scratch, under strict engineering constraints and without visible ground truth, requires moving beyond passive observation. Between perceiving a structure and constructing one lies a critical gap that standard benchmarks fail to measure.

To bridge this gap, we introduce **DreamHouse**, a novel benchmark designed to evaluate **physical generative reasoning**—the capacity to synthesize artifacts that concurrently satisfy geometric, structural, and code-compliance constraints. We ground our benchmark in residential timber-frame construction.

This domain provides fully codified correctness criteria and discrete, verifiable components, while remaining visually complex enough to prove that perceptual plausibility alone is insufficient. DreamHouse comprises over 26,000 structurally verified models spanning 13 architectural styles. Furthermore, we provide a suite of 10 deterministic, physics-based tests covering load paths, span limits, member connectivity, *etc.*. Crucially, these tests operate directly on the scene graph, bypassing the need for computationally expensive simulations.

The DreamHouse task is formulated as an iterative generation process: given rendered views of a target structure, a model must generate Blender Python construction code, process structured validation feedback, and refine its output until all structural tests pass. We evaluate performance across three protocols with varying degrees of external scaffolding (Planner-Atomic, Planner-Reactive and Planner-Managed) and two input conditions (bare framing visible, *Frame*, versus occluded by finished cladding, *Facade*). This framework serves as a controlled ablation study, isolating whether generation failures stem from weak spatial reasoning, deficient planning, or an inability to self-correct.

Our extensive evaluation of three frontier VLMs including GPT-5 [23], Gemini 3 [14] and Claude 4.5 [3] across more than 20,000 independent agentic tasks reveals striking limitations in current state-of-the-art models. Notably, models that excel on standard coding and reasoning leaderboards often struggle with physical generation; highly-ranked generalists frequently underperform compared to their peers when evaluated under our structured protocols. Even the most capable model achieves a joint pass rate of **merely 7.1%**, successfully satisfying both structural validity and visual fidelity simultaneously. These findings demonstrate that physical generative reasoning is not a natural byproduct of general intelligence, but rather a distinct, underdeveloped capability axis demanding dedicated evaluation. Our contributions:

- **Physical generative reasoning** as a novel VLM evaluation axis, distinct from perception, comprehension, and simulation-based benchmarks.
- **The DreamHouse benchmark**, comprising 26,000+ verified timber-frame structures across 13 architectural styles with multi-view renderings and construction phase-wise annotations.
- **A 10-test structural validation suite** that is deterministic and simulation-free, covering International Residential Code (IRC) compliance [8], physics, geometry, and fabrication details.
- **A three-protocol agentic evaluation framework**, showing that scaffold design is as important as model selection for physical generation tasks.

## 2 Related Work

**Physical and Spatial Benchmarking for VLMs.** A growing body of work exposes systematic gaps between VLM semantic fluency and physical-world grounding [6, 7, 17, 27, 29, 43]. PhysBench [7] evaluates 75 VLMs on 10K video-image-text entries spanning object properties, relationships, and dynamics, finding that models excel at static recognition but fail on Newtonian dynamics, a

deficiency attributed to missing physical priors rather than perceptual limits. VSI-Bench [39] measures metric visual-spatial intelligence (distances, sizes, directions) from egocentric video, showing that standard chain-of-thought prompting *degrades* spatial estimation, models must instead build explicit cognitive maps. Think-with-3D [6] endows VLMs with 3D geometric imagination via latent alignment and RL-based spatial rewards, enabling occluded geometry completion from sparse views. These benchmarks evaluate *passive* physical understanding: answering questions about or reconstructing observed scenes. DreamHouse targets the harder *generative* gap, producing structures that are themselves physically valid, without any reference structure to recall.

**3D Scene Generation.** Large-scale 3D generative models [2,11,18,20,34,35,44] optimize photometric or geometric metrics against reference renderings, leaving structural validity unmeasured. TRELLIS [38] achieves state-of-the-art image-conditioned 3D generation via sparse voxel latents, but its representation is derived entirely from visual observations, a generated house mesh may score well on Chamfer distance while failing every structural test. SpatialGen [12] generates photorealistic indoor scenes conditioned on 3D layouts, operating at rendering-grade level-of-detail for design visualization. DreamHouse operates at fabrication-grade detail (LOD 350), specifying member species, cross-sections, and connections sufficient for construction scheduling, complementary positions in the AEC pipeline.

**Code-Driven Structured Generation.** A productive line of work [5, 10, 31, 32, 40, 45] has VLMs write executable programs verified by domain-specific oracles, consistently outperforming direct pixel-level synthesis for structured outputs. DreamHouse adopts this architecture but replaces the visual oracle with deterministic engineering compliance. VIGA [41] frames inverse graphics as a long-horizon agentic loop (WRITE→RUN→RENDER→COMPARE→REVISE), demonstrating that iterative execution feedback recovers performance where single-shot VLMs fail. Its feedback signal is photometric; structural failures (missing members, span violations) are invisible to any rendering oracle. BlenderGym [15] benchmarks VLMs on Blender editing tasks via LPIPS/CLIP-I, sharing our infrastructure but evaluating visual imitation rather than structural correctness. MCP-Universe [21] benchmarks LLM agents across six domains including 3D Design, where even GPT-5 achieves only 43.7% success; DreamHouse provides the domain-specific 10-test physical validation absent from its 3D tasks. AutoPresent [13] and Chart2Code [32] further confirm that programmatic generation dominates pixel-level synthesis for structured outputs; DreamHouse extends this finding to 3D structural engineering where constraints are physical law rather than aesthetic convention.

## 3 DreamHouse Benchmark

We construct **DreamHouse**, a benchmark for physical generative reasoning grounded in residential timber-frame construction. It comprises (1) a large-scale dataset of structurally verified structures paired with multi-view renderings, and

(2) a deterministic validation suite of 10 physics-based tests scoring any generated structure against engineering and code-compliance standards [8, 26]. Together they define both the task and the metric, enabling objective evaluation of whether a model can *construct*, not merely depict, the physical world.

### 3.1 Benchmark Construction

*Testbed choice.* Residential timber framing is an unusually well-suited domain for evaluating physical generative reasoning: its correctness criteria are fully codified (load-path integrity, member sizing, connection geometry, assembly order) and objectively verifiable without human annotation [37]. The domain is also visually rich and stylistically diverse, making perceptual plausibility a meaningful but insufficient criterion. Before this choice, we piloted an alternative **bridge-generation** testbed but found that frontier VLMs could easily satisfy all validation tests by recovering a small set of geometric parameters analytically from the image, bypassing structural understanding entirely; *see Appendix for details.* Timber framing forecloses this shortcut: member spacing, connection hierarchy and load-path topology are not directly legible from exterior renders, and no analytic inversion exists.

*Parametric generation and dataset scale.* Each structure originates from a JSON configuration defining footprint dimensions, story count, roof pitch, overhang depth. A procedural Blender Python generator instantiates this into a fully resolved 3D timber-frame model with human in the loop — individual foundation sills, floor joists, wall studs, headers, ridge beams, rafters, and collar ties, each at the correct position, orientation, and IRC cross-section, with member identities and parent-child relationships preserved in a structured scene graph.

We identify **13 canonical residential styles** [19] whose structural logic is sufficiently distinct to stress different aspects of the reasoning pipeline: *A-Frame*, *Barn*, *Carriage*, *Colonial*, *Courtyard*, *Cruciform*, *Farmhouse*, *Ranch*, *Saltbox*, *Shotgun*, *Split-Level*, *Townhouse*, and *Z-Plan.* These span single- and multi-story footprints, symmetric and asymmetric roofs, and rectangular to complex multi-wing plans (Figure 2 and Figure 1). After filtering through the full validation suite, the released dataset contains **26,543 provably buildable structures**, ranging from 133 to 1,548 members (mean 673, median 656), partitioned into Foundation, Floor, Walls, and Roof categories [1]. All structures are rendered from five canonical viewpoints using Blender [16] Cycles. All structures meet **LoD 350** (fabrication-grade); *see Appendix for details.*

*Structural Validation Suite.* The suite comprises 10 deterministic tests in three pillars operating directly on the scene graph. With no physics simulation required, our evaluation can be fast, deterministic, and interpretable (Figure 3); *full mathematical definitions are provided in Appendix.* Figure 4 illustrates the corresponding agent loop for the Planner-Managed protocol.

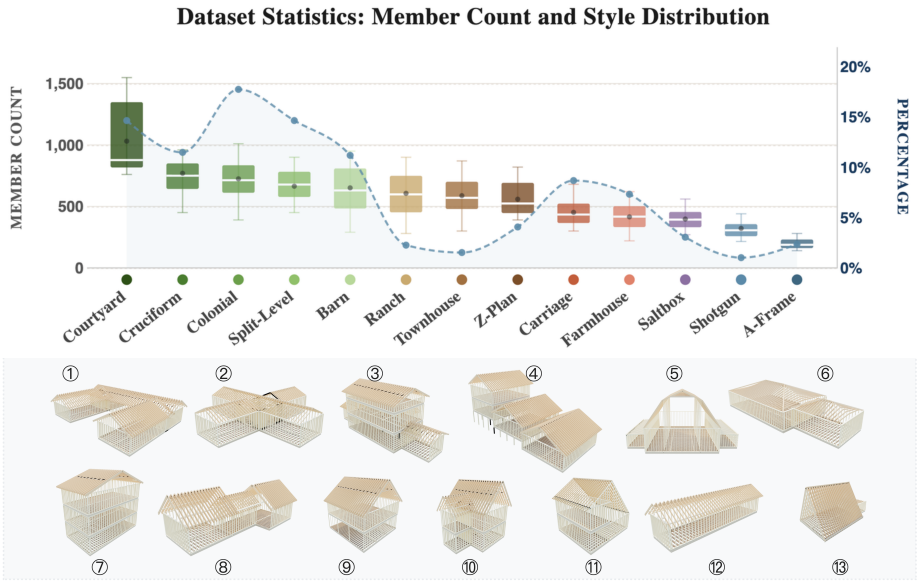


**Fig. 2: DreamHouse Dataset Overview.** *(Top)* Dataset statistics showing member count distributions (box plots, left axis) and style proportions (dashed line, right axis) across all **13** architectural styles, ordered by decreasing structural complexity. Member counts range from 133 to 1,548 (mean 673). *(Bottom)* Representative Cycles-rendered timber frame structures for each style: *Courtyard, Cruciform, Colonial, Split-Level, Barn, Ranch, Townhouse, Z-Plan, Carriage, Farmhouse, Saltbox, Shotgun, A-Frame*, spanning complex multi-wing configurations to compact single-story forms.

### 3.2 Task Formalization

We formalize evaluation as a family of *agentic generation tasks*. Let $\mathcal{A}$ denote the agent (VLM), $\mathcal{E}$ the executor (Blender environment), and $\mathcal{V}$ the structural validator. At each turn $t$, $\mathcal{A}$ receives observation $o_t = (I_0, f_{t-1})$, where $I_0$ is the fixed multi-view task input and $f_{t-1}$ is structured validation feedback from the previous turn, and produces a Blender Python action $a_t$. $\mathcal{E}$ applies $a_t$ to transition the scene graph $s_{t-1} \to s_t$, and $\mathcal{V}$ evaluates $s_t$ to produce feedback $f_t$ reporting per-test pass/fail status and violation counts (e.g., *"BeamPost gap exceeds 3.0 m; detected spacing 7.0 m between BeamPost_01 and rim"*). This feedback closes the loop: $f_t$ becomes part of $o_{t+1}$. On failure, the agent retries from the same scene state $s_t$ without resetting context; the full conversation history $\mathcal{H}_t = (I_0, a_1, f_1, \ldots, a_{t-1}, f_{t-1})$ is preserved across all retries. Before writing any code, $\mathcal{A}$ produces a hierarchical JSON plan specifying member categories, phase ordering, and assembly dependencies. Each task enforces a per-step retry budget $R_{\text{step}}$ and a global budget $R_{\text{global}}$.

We instantiate three protocols that vary the degree of external phase management, forming a controlled ablation over agentic scaffold design.The same

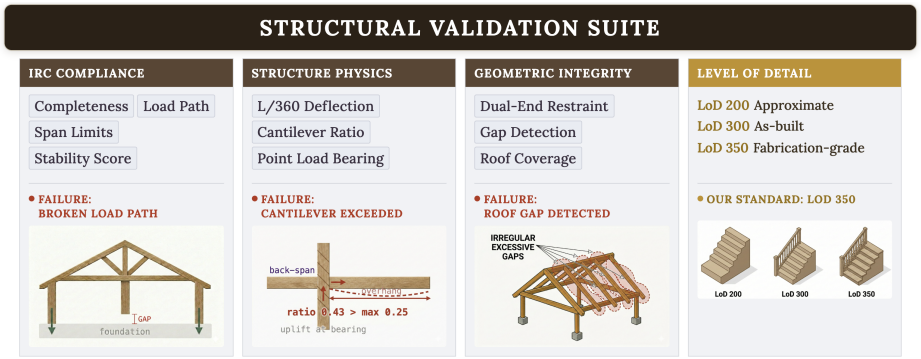


**Fig. 3: Structural Validation Suite.** 10 tests across four pillars. **IRC Compliance** (left): load path, span limits, completeness, stability score. **Structure Physics** (center-left): L/360 deflection, cantilever ratio, point load bearing. **Geometric Integrity** (center-right): dual-end restraint, gap detection, roof coverage. **LoD 350** (right): fabrication-grade member geometry.

deterministic validation suite $\mathcal{V}$ used to certify benchmark structures during dataset construction serves as the evaluation signal during task execution, ensuring ground-truth and assessment criteria are fully aligned.

*Planner-Atomic ($\mathcal{T}_S$).* $\mathcal{A}$ receives $o_1 = (I_0, \varnothing)$ and generates a complete construction script $a_1$ covering all member categories in a single code block. On failure, $a_t$ is regenerated from updated $\mathcal{H}_t$; accepted when $\mathcal{V}(s_t)$ passes all 10 tests or $R_{\text{global}}$ is exhausted. This protocol tests holistic single-pass structural synthesis with no intermediate feedback between phases.

*Planner-Reactive ($\mathcal{T}_Q$).* $\mathcal{A}$ generates a single script covering all $K$ construction phases in a self-determined order. After each phase $k$ is materialized into $s_t$, it is evaluated against $\mathcal{V}_{\text{mid}} \subset \mathcal{V}$ (load path and stability); failure triggers full script regeneration from $\mathcal{H}_t$, invalidating all phases $k{+}1, \ldots, K$. Unlike $\mathcal{T}_S$, the model must re-plan all phase interdependencies in context after each failure — combining the demands of holistic planning and sequential commitment without any external scaffolding.

*Planner-Managed ($\mathcal{T}_W$).* $\mathcal{A}$ generates scripts one phase at a time under external phase management. The scene $s_t$ persists across phases; each phase must pass $\mathcal{V}_{\text{mid}}$ before the next is unlocked, with up to $R_{\text{step}}$ retries per phase. A failure at phase $k$ never invalidates $s_{t-1}$: only $a_t$ is regenerated, leaving all prior scene state intact. This protocol provides the strongest external scaffolding and isolates each phase as an independent sub-task.

*Iterative Refinement with Visual Feedback ($\mathcal{T}_E$).* Beyond the core generation protocols, we introduce an exploratory editing task to test whether models can

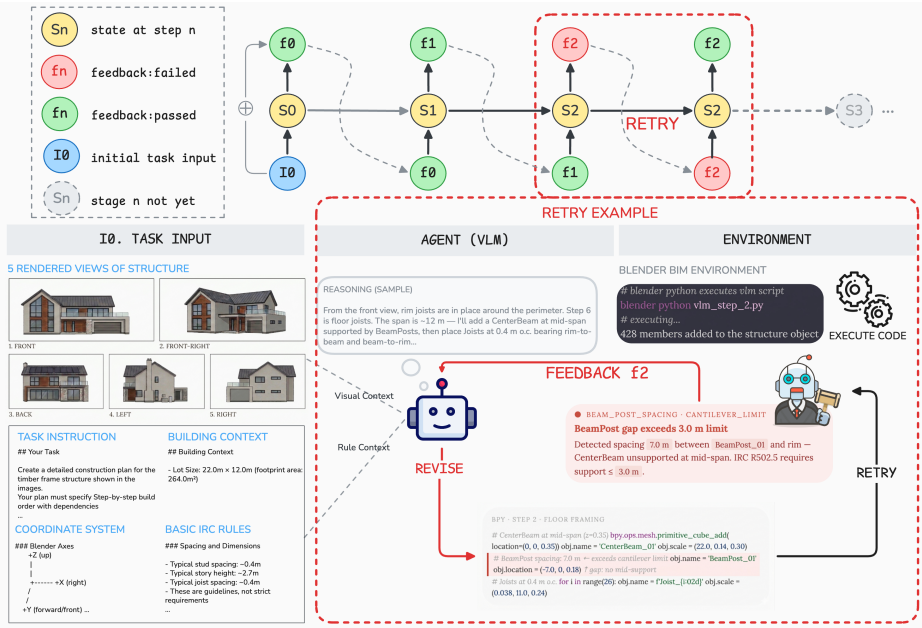


**Fig. 4: Task formalization example Planner-Managed.** *Top (agent loop):* Evaluation is formalized as a recurrent agentic process. At each turn $t$, the agent $\mathcal{A}$ (VLM) receives observation $o_t = (I_0, f_{t-1})$, the original multi-view task image $I_0$ and structured validation feedback $f_{t-1}$ from the previous turn, and generates a Blender Python action $a_t$. The executor $\mathcal{E}$ applies $a_t$ to transition the scene graph $s_{t-1} \to s_t$, and the validator $\mathcal{V}$ produces feedback $f_t$ reporting per-test pass/fail and violation counts. This feedback becomes the input observation for the next turn, closing the loop. On failure (e.g., $f_2$, red), the agent retries from the same scene state $s_2$ without resetting context. $\mathcal{A}$, $\mathcal{E}$, and $\mathcal{V}$ are omitted from the diagram for visual clarity; see Section 3.2 for full formalization. *Bottom (task instantiation):* The task input $I_0$ consists of five rendered views of the target structure paired with building context and rules. The agent reasons over these to produce and iteratively revise construction code; the environment executes the code in Blender and the validator returns structured diagnostic feedback driving the next revision.

refine visual alignment without breaking an already valid structure. Starting from $s_T$ that passed $\mathcal{V}$ under $\mathcal{T}_S$ but achieved visual fidelity score $S < \tau$ (see Section 3.3), $\mathcal{A}$ iteratively refines $s_t$ while preserving structural validity. Feedback $f_t$ is augmented with a rendered side-by-side comparison against the target; $\mathcal{V}$ is re-run at every iteration and any $a_t$ causing structural regression is rejected even if $S$ improves.

### 3.3 Evaluation Metrics

*Structural Validity.* This metric axis is binary: a structure either satisfies all tests in $\mathcal{V}$ or it does not. We deliberately avoid partial credit here. Physical constraints

are discontinuous by nature, a single unrestrained member or one over-spanned joist renders the assembly unsafe regardless of how well the rest is built, and a graded score would obscure this harshness. over all $N$ evaluated structures per condition.

*Visual Fidelity.* We measure pixel-level agreement between multi-view renders of the generated and target structures (Eq.(1)). Standard perceptual embeddings such as DINO or CLIP are ill-suited here: they are trained to be invariant to the geometric and positional differences that matter most in construction (a wall shifted by 0.5 m looks semantically similar but is structurally wrong), and they provide no signal on missing or misplaced members that happen to fall outside the salient region. We instead use alpha-weighted MSE. Let $\hat{R}_v$ and $R_v$ be the rendered and target images of timber frames at view $v$, with union alpha mask $m_v = \alpha_v^g \cup \alpha_v^r$ restricting comparison to non-transparent regions of either structure. The visual fidelity score is:

$$S = \frac{1}{V} \sum_{v=1}^{V} \max\left(0,\ 1 - \lambda \cdot \frac{\sum_p m_v(p)\, \|\hat{R}_v(p) - R_v(p)\|^2}{\sum_p m_v(p)}\right) \tag{1}$$

averaged over $V$ orthographic views. The union mask penalizes missing geometry, a member absent in the generated structure leaves unmatched non-transparent pixels in the target without rewarding empty background agreement. The scale $\lambda$ and threshold $\tau$ are calibrated in Section 4; importantly, $S$ is a visual similarity measure rather than a constructibility proxy, so the joint metric requires both $S \geq \tau$ and structural validity. Mean visual fidelity in the main tables is reported over structurally passed structures, where render comparisons are meaningful.

*Topological Fidelity.* Visual similarity can be gamed: a model that generates the correct silhouette but with wrong member counts, misplaced connections, or incorrect spatial hierarchy will score well visually while producing an unbuildable structure (Eq.(2)). We therefore measure structural correspondence directly on the scene graph, independent of rendering, via three complementary statistics: Census accuracy $C$, Hungarian match rate $M$, and Voxel IoU $V$.

$C$ measures whether the model generates the correct number of members per category. Let $n_k^*$ and $\hat{n}_k$ be the ground-truth and generated member counts for category $k \in \{1, \ldots, K\}$; then $C = \frac{1}{K} \sum_k \min(n_k^*, \hat{n}_k) / \max(n_k^*, \hat{n}_k)$. This catches global over- or under-building even when individual member positions look plausible, a failure mode that positional metrics miss.

$M$ measures whether individual members are placed at the correct spatial positions. Let $\sigma^*$ be the optimal assignment from the Hungarian algorithm between ground-truth and generated member centroids $\{c_i\}$ and $\{\hat{c}_j\}$; then $M = \frac{1}{N} \sum_i \mathbf{1}[\|c_i - \hat{c}_{\sigma^*(i)}\| \leq \delta]$ where $\delta = 0.3\,\text{m}$ is the positional tolerance, chosen to be permissive of minor placement error while rejecting members shifted by more than a stud spacing. $M$ catches misalignment even when counts are correct, a failure mode $C$ cannot detect.

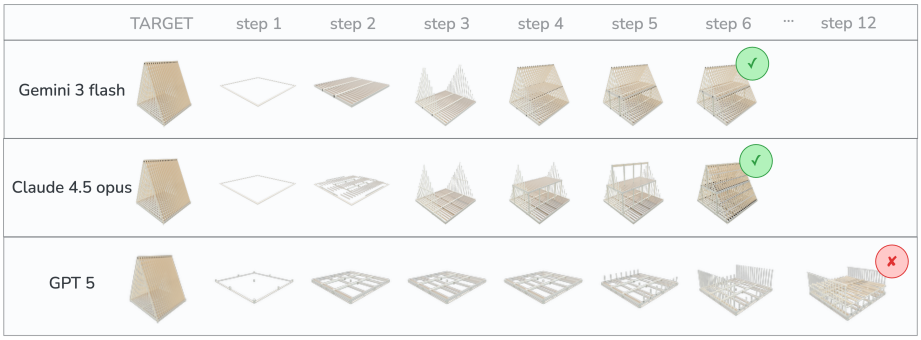


**Fig. 5: Planner-Managed qualitative example AF-01-0060 (A-frame style).** All three models begin from scratch and receive stepwise visual feedback toward the same target structure. Despite reaching a valid result at step 6, **Gemini** and **Claude** employ markedly different construction strategies. Gemini pursues a top-down, shape-first approach: it approximates the overall silhouette early and refines toward it. Claude reasons bottom-up: it first establishes a structurally sound interior frame, then lays the roof rafters over it in the final step – a sequence closer to how a builder would physically construct an A-frame. **GPT-5** fails to recognize the defining constraint of A-frame geometry, that the roof planes double as load-bearing walls, and from step 6 onward enters a false loop of adding conventional wall studs along the perimeter. Unable to escape this structural misconception, it exhausts all attempts without producing a valid result. This example highlights that identical visual feedback can elicit fundamentally different reasoning strategies, and that success depends not just on visual matching ability but on implicit architectural knowledge.

$V$ measures volumetric overlap of the assembled structure. Let $\mathcal{G}$ and $\hat{\mathcal{G}}$ be the voxelized scene graphs at a fixed grid resolution; then $V = |\mathcal{G} \cap \hat{\mathcal{G}}|/|\mathcal{G} \cup \hat{\mathcal{G}}|$. $V$ catches gross shape errors, *e.g.* missing wings, wrong footprint extent, collapsed roof geometry that member-level statistics miss because they operate on individual centroids rather than the assembled volume.

The composite score is:

$$T = w_C\, C + w_M\, M + w_V\, V, \quad w_C + w_M + w_V = 1 \tag{2}$$

where $w_M > w_C = w_V$, reflecting that positional match is the most direct measure of assembly correctness, with count accuracy and volumetric overlap contributing equally as complementary diagnostics. Specific weight values are provided in Section 4. Unlike structural validity, $T$ is continuous and is not used as a pass/fail criterion. We use it to decompose *how* a generated assembly differs from the target; a structure can be topologically close to the target and still fail a discontinuous structural test such as load-path continuity or dual-end restraint.

**Table 1:** Consolidated Performance Metrics for GPT-5, Claude Opus 4.5, and Gemini 3 Flash. *Frame* = bare timber framing; *Facade* = finished exterior. **Bold** = best per column; shaded rows = best model per protocol by average across all conditions. **Structural Pass Rate** in $[0, 1]$: passes only if all 10 structural validation tests clear. **Visual Fidelity**: average visual similarity score across 5 rendered views. **Topological Fidelity**: structural similarity across member counts, spatial alignment, and volumetric overlap. **Joint Pass Rate** in $[0, 1]$: fraction satisfying *both* structural and visual criteria simultaneously.

| Model | Protocol | Structural Pass Rate | | Visual Fidelity | | Topological Fidelity | | Joint Pass Rate | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Frame* | *Facade* | *Frame* | *Facade* | *Frame* | *Facade* | *Frame* | *Facade* |
| GPT-5 | Planner-Atomic | **0.792** | 0.637 | 0.312 | 0.287 | 0.143 | 0.151 | 0.035 | 0.019 |
| | Planner-Reactive | 0.302 | 0.247 | 0.293 | 0.266 | 0.141 | 0.143 | 0.003 | 0.003 |
| | Planner-Managed | 0.333 | 0.157 | 0.179 | 0.175 | 0.178 | 0.157 | 0.008 | 0.003 |
| Claude | Planner-Atomic | 0.716 | **0.779** | **0.406** | 0.377 | 0.164 | **0.168** | **0.071** | **0.064** |
| | Planner-Reactive | 0.428 | 0.494 | 0.239 | 0.245 | 0.133 | 0.141 | 0.003 | 0.013 |
| | Planner-Managed | 0.713 | 0.737 | 0.278 | 0.291 | 0.205 | 0.190 | 0.031 | 0.022 |
| Gemini | Planner-Atomic | 0.454 | 0.539 | 0.376 | **0.394** | 0.171 | 0.170 | 0.031 | 0.036 |
| | Planner-Reactive | 0.507 | 0.376 | 0.345 | 0.342 | 0.160 | **0.152** | 0.019 | 0.013 |
| | Planner-Managed | 0.785 | 0.737 | 0.313 | 0.282 | **0.232** | 0.211 | 0.043 | 0.026 |

**Table 2:** Iterative refinement with visual feedback. Each model starts from its Planner-Atomic structural passes and iterates up to 10 visual-feedback revisions. *Baseline* = oneshot output before revision; *Final* = last attempt; *Best* = highest-scoring attempt across all iterations. $\Delta$ is Final−Baseline mean score. *Improved Tasks* = fraction of tasks that improved their visual fidelity score. *Structures Retained* = fraction retaining structural validity at the final attempt. **Bold** = best per column.

| Model | Mean Visual Score | | | | | |
|---|---|---|---|---|---|---|
| | *Baseline* | *Final* | *Best* | $\Delta$ | *% Improved Tasks* | *Structures Retained* |
| Claude | 0.408 | **0.441** | **0.461** | **+0.033** | **70.3%** | 68.5% |
| GPT-5 | 0.392 | 0.405 | 0.425 | +0.013 | 61.4% | **78.1%** |
| Gemini | 0.433 | 0.443 | 0.471 | +0.010 | 56.1% | 73.8% |

## 4 Experiments

*Models.* We evaluate three frontier vision-language models: **GPT-5** [23], **Gemini 3 Flash** [14], **Claude Opus 4.5** [3]. All models use identical prompting with no fine-tuning. We abbreviate Claude Opus 4.5 and Gemini 3 Flash as **Claude** and **Gemini** respectively for brevity. We additionally evaluate open-source baselines, including **Qwen3.5-397B-A17B** [28], under matching protocols where the models can produce executable Blender code (Table 3).

*Evaluation subset.* We evaluate on a stratified sample of 1,200 structures per model-protocol cell, drawn uniformly across all 13 styles and 4 roof types, each structure represents a full agentic task: a multi-turn conversation of up to ∼50 API calls depending on protocol. Across three models and our three main protocols, this yields **21,600** independent agentic tasks and ∼**200k** API calls. Token

consumption is correspondingly substantial, totalling approximately ∼**5B tokens** across the benchmark. This scale reflects the real computational cost of evaluating frontier models on physically grounded generative tasks.

*Details.* Visual fidelity uses $V = 5$ orthographic views (front, front-right, back, left, right) at 512×512 with scale factor $\lambda = 10$, mapping a 10% mean pixel error to $S = 0$ and passing threshold $\tau = 0.6$, which corresponds empirically to structures visually recognizable as matching the target style and layout. This threshold is not ceiling-bound: ground-truth self-renders score $S = 1.0$, while frontier-model per-structure scores span up to 0.84–0.90, with 11–24% of Planner-Atomic outputs exceeding $\tau$ depending on the model. Topological fidelity is computed with positional tolerance $\delta = 0.3\,\mathrm{m}$ (permissive of minor placement error while rejecting members shifted by more than a stud spacing) and composite weights $w_C = 0.3$, $w_M = 0.4$, $w_V = 0.3$, reflecting that positional match is the most direct continuous diagnostic of assembly correspondence. Visual fidelity is reported over structurally passed structures, while topological fidelity is used diagnostically to compare both passed and failed renderable structures. The joint pass rate $J$ fraction satisfying $\mathcal{V}$ and $S \geq \tau$ simultaneously serves as the primary scalar summary of overall generation quality.

### 4.1 Main Results

We evaluate each model under two input conditions: *Frame*, in which the model receives a bare timber framing image and must generate a structurally valid assembly, and *Facade*, in which the framing is occluded by finished exterior cladding and the model must infer the underlying structure from appearance alone. Intuitively, the *Facade* condition presents a more challenging task, as it requires the model to solve an inverse problem of hallucinating hidden, load-bearing topologies based solely on superficial exterior cues. Table 1 reports structural pass rates, visual and topological fidelity, and joint pass rates under three agentic generation tasks mentioned in 3.2. Results reveal four cross-cutting patterns that together characterize the current frontier of physically grounded generative reasoning. Figure 5 shows a representative Planner-Managed trajectory. We detail each observation $O$ in turn.

The Appendix provides a structural failure breakdown across pipelines and models, showing that Planner-Atomic concentrates on geometric IRC violations while Planner-Managed shifts errors toward load-path and stability constraints.

*O1: Structural validity and visual fidelity are orthogonal.*

GPT-5 leads structurally in Planner-Atomic (79.2%) yet scores lowest visually (0.312); Claude leads visually (0.406) but not structurally; Gemini leads structurally under Planner-Managed (78.5%) while remaining competitive visually (0.313). Models frequently satisfy validation by converging to a generic, physically safe assembly that ignores visual conditioning entirely. Physical validity is not a byproduct of visual imitation, and vice versa.

**Table 3:** Open-source VLM baselines. We report structural pass rate, mean visual fidelity $S$, and joint pass rate for Qwen3.5-397B-A17B. The thinking-mode ablation is only applicable to Qwen3.5-397B-A17B. Other evaluated open-source models—LLaVA-OV-72B, InternVL3, Qwen3-VL-30B/8B, and Kimi K2.5—achieved 0 structural pass rate across all tested protocols.

| Model | Protocol | Thinking | Structural Pass Rate | Visual Fidelity | Joint Pass Rate |
|---|---|---|---|---|---|
| | | | *Frame* | *Frame* | *Frame* |
| Qwen3.5-397B-A17B | Planner-Atomic | yes | 0.231 | 0.395 | 0.029 |
| Qwen3.5-397B-A17B | Planner-Atomic | no | 0.084 | 0.382 | 0.007 |
| Qwen3.5-397B-A17B | Planner-Reactive | yes | 0.014 | 0.353 | 0.005 |
| Qwen3.5-397B-A17B | Planner-Reactive | no | 0.000 | – | – |
| Qwen3.5-397B-A17B | Planner-Managed | yes | 0.206 | 0.261 | 0.016 |
| Qwen3.5-397B-A17B | Planner-Managed | no | 0.000 | – | – |

*O2: Structural reasoning is not a monolithic capability.*

Although GPT-5 leads structurally in Planner-Atomic (79.2%), it collapses under Planner-Managed (33.3%), dropping 46 points; Gemini reverses this entirely (45.4% → 78.5%), gaining 33 points; Claude is comparably protocol-stable (71.6% vs. 71.3%). We interpret this as two distinct reasoning modes: *holistic* (*e.g.*, GPT-5 is strong single-pass synthesis, weak under incremental commitment) and *Planner-Managed* (*e.g.*, Gemini is strongest with phase decomposition and intermediate feedback). Planner-Reactive experiment confirms this: forcing models to self-decompose a structure into sequential steps without any external phase guidance is harder, as it demands both holistic planning and incremental commitment simultaneously.

*O3: Some Models can see structure under the skin, some cannot.*

Claude and Gemini improve structurally under Facade input (Claude 71.6% → 77.9%; Gemini 45.4% → 53.9%); GPT-5 declines sharply (79.2% → 63.7%). Gemini is the only model whose visual score also improves (0.376 → 0.394), suggesting it activates strong architectural priors when reasoning through appearance to infer underlying structure. GPT-5, relying on direct structural mimicry, is disrupted when the framing signal is occluded by cladding.

*O4: Protocol dominates model.*

The performance gap attributable to protocol choice exceeds any cross-model difference within a fixed protocol. Gemini swings from 45.4% to 78.5% structural pass rate, a 33 points gain, simply by changing from Planner-Atomic to Planner-Managed. Crucially, Gemini *overtakes* GPT-5 under Planner-Managed despite trailing it by 34 points under Planner-Atomic, a rank reversal impossible to predict from single-protocol evaluation alone. This suggests that agentic scaffold

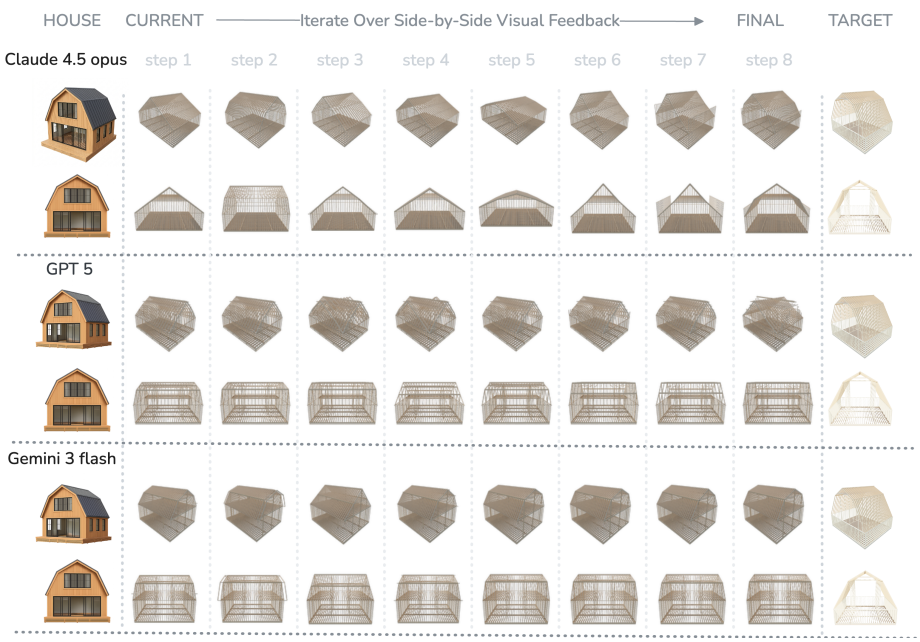


**Fig. 6: Iterative Refinement with Visual Feedback (Section 4.2, qualitative example BN-01-0273 (barn style).** Starting from a structurally valid Planner-Atomic output, each model receives up to 10 rounds of side-by-side visual feedback (current render vs. target) and iteratively revises its structure while maintaining structural validity; steps 9–10 are omitted as changes from step 8 are negligible. **Claude** demonstrates meaningful visual self-correction: it produces a plausible roof (called: gambrel style) by step 2 but with an incorrect ridge orientation, then progressively corrects across subsequent steps, converging to the almost right form by step 8. **GPT-5** commits to an incorrect orientation from step 1 and fails to recover, instead collapsing into a hybrid that blends both orientations without resolving either. **Gemini** makes minimal modifications across all 10 attempts, suggesting it does not meaningfully incorporate the visual feedback signal. This example illustrates the key behavioral divergence in this experiment: models differ not only in final score, but in *whether* they engage with feedback at all.

design is at least as consequential as model selection: the same model can be the weakest or the strongest depending on how the task is structured.

*Other Observations.* Topological proximity does not guarantee structural validity. Composite topological scores are nearly identical between passed and failed structures: in Planner-Atomic, passed vs. failed scores are 0.164 vs. 0.149 (Claude), 0.171 vs. 0.175 (Gemini), 0.143 vs. 0.131 (GPT-5). For Gemini, failed structures score *higher* topologically than passing ones. Physical constraints are *discontinuous*, a single missed connection or an over-span fails the entire structure with no partial credit while topological and visual metrics are smooth and continuous. Because of this fundamental mismatch, standard perceptual bench-

marks cannot be relied upon. They merely measure how accurate a structure *looks* overall, making them insufficient proxies for evaluating whether a model actually understands true physical generative reasoning.

### 4.2 Iterative Refinement with Visual Feedback

We test whether visual feedback can improve similarity without breaking structural validity. Starting from Planner-Atomic structural passes with $S < 0.6$, each model receives up to 10 visual-feedback revisions with validation enforced at every step; Table 2 reports results and Figure 6 shows a representative refinement trajectory. Final structural retention remains high (68.5% Claude, 73.8% Gemini, 78.1% GPT-5), while visual scores improve moderately: Claude gains the most (+0.033, 70.3% of tasks improved), with smaller but positive gains for Gemini and GPT-5.

## 5 Conclusion

We introduced DreamHouse, a benchmark for *physical generative reasoning* in timber-frame construction. Evaluating three frontier VLMs and open-source models reveals that structural validity and visual plausibility are orthogonal signals. Models ranking similarly on standard leaderboards diverge sharply here, yielding structural success rates from 15.7% to 79.2% and exhibiting qualitatively distinct failure modes. Crucially, performance is highly sensitive to the generation protocol. This demonstrates that *how* a model is asked to reason about physical space matters as much as what it inherently knows. These results expose a gap in current evaluations: the ability to maintain global structural consistency, such as load-path continuity and code-compliant spans within complex real-world building systems. We hope DreamHouse and its validation suite serve as a concrete, verifiable target to track progress in this domain.

*Limitations and future work.* High computational costs for executing validation scripts at scale currently limit our evaluation to closed-source frontier models. Future work should extend our physical constraint vocabulary to encompass concrete, steel, or multi-material assemblies. Furthermore, while we primarily evaluate the *output* generation pipeline, exploring the *input* end is vital, identifying visual representations, supervision signals, or training objectives that enable models to proactively internalize structural grammar rather than merely recovering it through iterative feedback.

## References

1. Adel, A., Thoma, A., Helmreich, M., Gramazio, F., Kohler, M.: Design of robotically fabricated timber frame structures. In: 38th Annual Conference of the Association for Computer Aided Design in Architecture: Recalibration on Imprecision and Infidelity, ACADIA 2018. pp. 394–403. ACADIA (2018)

2. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025)
3. Anthropic: Claude Opus 4.5 system card. https://www-cdn.anthropic.com/bf10f64990cfda0ba858290be7b8cc6317685f47.pdf (2025), accessed March 1, 2026
4. Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J.C., Geva, M., He, J., Wu, J., Li, M.: Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. arXiv preprint arXiv:2503.01773 (2025)
5. Chen, Y., Lin, K.Q., Shou, M.Z.: Code2video: A code-centric paradigm for educational video generation. arXiv preprint arXiv:2510.01174 (2025)
6. Chen, Z., Zhang, M., Yu, X., Luo, X., Sun, M., Pan, Z., Feng, Y., Pei, P., Cai, X., Huang, R.: Think with 3d: Geometric imagination grounded spatial reasoning from limited views. arXiv preprint arXiv:2510.18632 (2025)
7. Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., Wang, Y.: Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In: ICLR (2025)
8. Code, B.: International residential code. International Energy: Paris, France (2018)
9. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al.: Understanding world or predicting future? a comprehensive survey of world models. ACM Computing Surveys **58**(3), 1–38 (2025)
10. Doris, A.C., Alam, M.F., Heyrani Nobari, A., Ahmed, F.: Cad-coder: An open-source vision-language model for computer-aided design code generation. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. vol. 89220, p. V03AT03A031. American Society of Mechanical Engineers (2025)
11. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27713–27724 (2025)
12. Fang, C., Li, H., Liang, Y., Zheng, J., Mao, Y., Liu, Y., Tang, R., Zhou, Z., Tan, P.: Spatialgen: Layout-guided 3d indoor scene generation. In: International Conference on 3D Vision (3DV) (2026)
13. Ge, J., Wang, Z.Z., Zhou, X., Peng, Y.H., Subramanian, S., Tan, Q., Sap, M., Suhr, A., Fried, D., Neubig, G., et al.: Autopresent: Designing structured visuals from scratch. In: CVPR. pp. 2902–2911 (2025)
14. Google DeepMind: Gemini 3 Flash system card. https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview (2025), accessed March 1, 2026
15. Gu, Y., Huang, I., Je, J., Yang, G., Guibas, L.: Blendergym: benchmarking foundational model systems for graphics editing. In: CVPR. pp. 18574–18583 (2025)
16. Hess, R.: Blender foundations: The essential guide to learning blender 2.5. Routledge (2013)
17. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. arXiv preprint arXiv:2506.03135 (2025)
18. Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., et al.: 3d and 4d world modeling: A survey. arXiv preprint arXiv:2509.07996 (2025)
19. Kwon, H.J., Lee, H.J., Beamish, J.O.: Us boomers' lifestyle and residential preferences for later life. Journal of Asian Architecture and Building Engineering **15**(2), 255–262 (2016)

20. Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., et al.: Worldmodelbench: Judging video generation models as world models. arXiv preprint arXiv:2502.20694 (2025)
21. Luo, Z., Shen, Z., Yang, W., Zhao, Z., Jwalapuram, P., Saha, A., Sahoo, D., Savarese, S., Xiong, C., Li, J.: Mcp-universe: Benchmarking large language models with real-world model context protocol servers. arXiv preprint arXiv:2508.14704 (2025)
22. Marín, J., Baptiste, T.M., Rodero, C., Williams, S.E., Niederer, S.A., García-Fernández, I.: Sciblend: Advanced data visualization workflows within blender. Computers & Graphics **130**, 104264 (2025)
23. OpenAI: OpenAI GPT-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf` (2025), accessed March 1, 2026
24. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)
25. Pan, Z., Liu, H.: Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. arXiv preprint arXiv:2503.18470 (2025)
26. Peng, Y., Lu, M., Imanpour, A.: Analysis of the rationality of extending lod standards from the design phase to the production and construction phases. In: International Conference on Computing in Civil and Building Engineering. pp. 314–324. Springer (2024)
27. Pun, A., Deng, K., Liu, R., Ramanan, D., Liu, C., Zhu, J.Y.: Generating physically stable and buildable brick structures from text. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14798–14809 (2025)
28. Qwen Team: Qwen3.5-397B-A17B release. `https://huggingface.co/Qwen/Qwen3.5-397B-A17B` (2026), accessed March 1, 2026
29. Rodionov, F., Eldesokey, A., Birsak, M., Femiani, J., Ghanem, B., Wonka, P.: Floorplanqa: A benchmark for spatial reasoning in llms using structured representations. arXiv preprint arXiv:2507.07644 (2025)
30. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Benchmarking spatial reasoning in vision-language models. arXiv preprint arXiv:2503.19707 (2025)
31. Sun, Q., Gong, J., Liu, Y., Chen, Q., Li, L., Chen, K., Guo, Q., Kao, B., Yuan, F.: Januscoder: Towards a foundational visual-programmatic interface for code intelligence. arXiv preprint arXiv:2510.23538 (2025)
32. Tang, J., Zhao, H.H., Wu, L., Tao, Y., Mao, D., Wan, Y., Tan, J., Zeng, M., Li, M., Wang, A.J.: From charts to code: A hierarchical benchmark for multimodal models. arXiv preprint arXiv:2510.17932 (2025)
33. Tang, K., Gao, J., Zeng, Y., Duan, H., Sun, Y., Xing, Z., Liu, W., Lyu, K., Chen, K.: Lego-puzzles: How good are mllms at multi-step spatial reasoning? arXiv preprint arXiv:2503.19990 (2025)
34. Team, H., Wang, Z., Liu, Y., Wu, J., Gu, Z., Wang, H., Zuo, X., Huang, T., Li, W., Zhang, S., et al.: Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. arXiv preprint arXiv:2507.21809 (2025)
35. Wang, X., Liu, L., Cao, Y., Wu, R., Qin, W., Wang, D., Sui, W., Su, Z.: Embodiedgen: Towards a generative 3d world engine for embodied intelligence. arXiv preprint arXiv:2506.10600 (2025)
36. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965 (2025)

37. Xia, B., O'Neill, T., Zuo, J., Skitmore, M., Chen, Q.: Perceived obstacles to multi-storey timber-frame construction: an australian study. Architectural science review **57**(3), 169–176 (2014)
38. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: CVPR. pp. 21469–21480 (2025)
39. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: CVPR. pp. 10632–10643 (2025)
40. Yin, G., Li, Y., Wang, Y., McConachie, D., Shah, P., Hashimoto, K., Zhang, H., Liu, K., Li, Y.: Codediffuser: Attention-enhanced diffusion policy via vlm-generated code for instruction ambiguity. arXiv preprint arXiv:2506.16652 (2025)
41. Yin, S., Ge, J., Wang, Z.Z., Li, X., Black, M.J., Darrell, T., Kanazawa, A., Feng, H.: Vision-as-inverse-graphics agent via interleaved multimodal reasoning. arXiv preprint arXiv:2601.11109 (2026)
42. Zha, J., Fan, Y., Yang, X., Gao, C., Chen, X.: How to enable llm with 3d capacity? a survey of spatial reasoning in llm. arXiv preprint arXiv:2504.05786 (2025)
43. Zhang, W., Zhou, Z., Zeng, X., Liu, X., Fang, J., Gao, C., Li, Y., Cui, J., Chen, X., Zhang, X.P.: Open3d-vqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. arXiv preprint arXiv:2503.11094 (2025)
44. Zhang, Y., Peng, C., Wang, B., Wang, P., Zhu, Q., Kang, F., Jiang, B., Gao, Z., Li, E., Liu, Y., et al.: Matrix-game: Interactive world foundation model. arXiv preprint arXiv:2506.18701 (2025)
45. Zhou, Z., Veeramani, S., Fakhruldeen, H., Uyanik, S., Cooper, A.I.: Genco: A dual vlm generate-correct framework for adaptive peg-in-hole robotics. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 16744–16751. IEEE (2025)