BEYOND A SINGLE DRAWING: MULTIFLOORQA FOR CROSS-VIEW SPATIAL REASONING IN ARCHITECTURAL DRAWINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning in multimodal large language models (MLLMs) remains insufficiently understood, particularly when spatial evidence must be integrated across multiple views. MultiFloorQA is introduced as a benchmark of 608 question–answer pairs over multiple architectural drawings, spanning six spatial task categories and Single-floor, Cross-floor, and Cross-projection evidence. Drawing generation, question construction, and reference-answer computation are grounded in a shared building representation, enabling precise and reproducible evaluation. Evaluation of representative MLLMs reveals substantial variation across models and spatial operations, while precise geometric computation and larger-scale spatial composition remain challenging even for strong systems. Controlled input-condition experiments further indicate that recovering and organizing geometry from the drawings themselves is a major source of difficulty. These findings motivate further advances in visual grounding, geometric computation, and cross-view spatial integration.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made substantial progress in visual understanding and multimodal reasoning (Chen et al., 2024; Cheng et al., 2024). Current models can recognize visual entities and their relations, interpret diagrams and structured visual content, and solve increasingly complex problems that combine visual information with language and domain knowledge. Broad benchmarks such as MMBench (Liu et al., 2024), MMMU (Yue et al., 2024), and MathVista (Lu et al., 2024) reflect this expansion from basic visual perception toward more general multimodal reasoning.

However, strong performance on broad multimodal tasks does not necessarily translate into reliable spatial reasoning. Spatial reasoning requires models to determine how entities are arranged, establish correspondences across different views, and integrate geometric evidence distributed across observations. Dedicated spatial benchmarks continue to expose substantial weaknesses in these abilities. BLINK (Fu et al., 2024) reports difficulties in visual correspondence and multi-view reasoning. VSI-Bench (Yang et al., 2025) shows that models remain limited in acquiring and using spatial information from sequential visual observations. MMSI-Bench (Yang et al., 2026) further identifies errors in grounding, cross-image matching, scene reconstruction, and spatial logic. Therefore, spatial reasoning warrants dedicated evaluation that can identify which capabilities remain unreliable.

Architectural drawings encode spatial structure through explicit boundaries, dimensions, connectivity, and projection conventions, while a single building is typically represented by multiple complementary drawings. Floor plans describe horizontal layouts at different levels, whereas elevations capture vertical structure from different projections. Therefore, reasoning about the building as a whole may require entities to be identified within individual drawings, matched across floors or projections, and combined through geometric and topological relations. At the same time, the explicit spatial structure encoded in architectural drawings allows many such relations to be defined precisely and their reference answers to be computed reproducibly. Recent benchmarks have studied related aspects of this setting. ArchPlanVQA (Li et al., 2026) evaluates architectural floor-plan understanding, DrawingVQA (Jung et al., 2026) studies visual-textual reasoning over construction

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

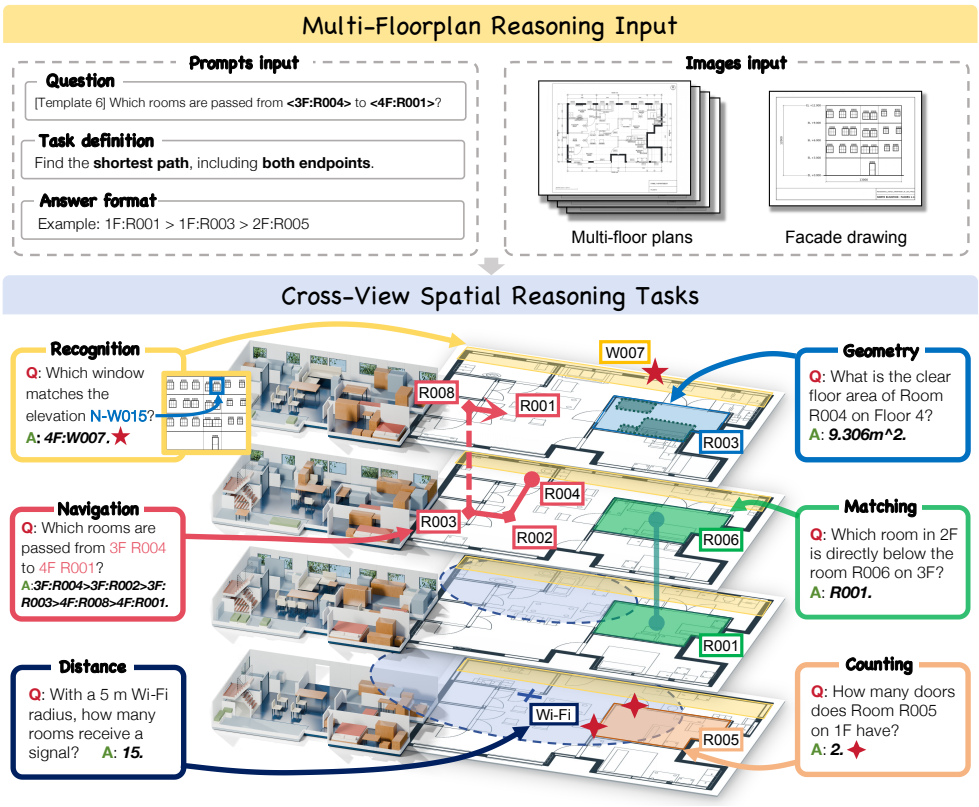


Figure 1: Overview of MultiFloorQA. The benchmark uses multiple floor plans and elevations of the same building to evaluate spatial reasoning within individual drawings, across floors, and across projections. The questions and drawings shown are simplified for illustrative purposes.

drawings, and AEC-Bench (Mankodiya et al., 2026) examines reasoning over architecture, engineering, and construction documents. FloorplanQA (Rodionov et al., 2026) evaluates spatial reasoning from structured floor-plan representations. However, the joint evaluation of precise geometry, physical connectivity, and spatial evidence distributed across multiple floor plans and projections remains underexplored.

To reliably evaluate this type of cross-view spatial reasoning, a benchmark should satisfy several basic requirements. First, the evidence required to answer each question should be present in the supplied drawings. Second, correspondences between related entities across drawings should be well defined within the benchmark, so that cross-view questions have unambiguous reference answers. Finally, reference answers should be computable and verifiable under explicit spatial rules, reducing ambiguity in the answers themselves.

To address these requirements, we construct MultiFloorQA around a shared building representation that supports drawing generation, constrained question instantiation, and deterministic reference-answer computation, with quality checks ensuring consistency between visible evidence and answers. MultiFloorQA contains 608 question–answer pairs from 186 drawings of 25 buildings, organized into MultiFloorQA-Main and MultiFloorQA-Hard, with the latter focusing on taller buildings and more demanding whole-building and cross-floor instances. Its 32 question templates are grouped into six primary spatial operations: Recognition, Counting, Distance, Geometry, Matching, and Navigation. Each question is independently labeled as Single-floor, Cross-floor, or Cross-projection according to where the required evidence is distributed.

We evaluate ten representative MLLMs on MultiFloorQA-Main and five on MultiFloorQA-Hard. On MultiFloorQA-Main, performance varies substantially across models and spatial operations, while precise geometric computations such as net usable area and first-collision distance remain difficult even for strong models. MultiFloorQA-Hard further reveals substantial errors in cross-floor

quantitative reasoning and multi-floor routing. Controlled input-condition experiments provide an additional diagnostic view. Replacing drawings with structured JSON scene representations substantially improves all three tested models, suggesting that extracting and organizing geometric information from drawings is an important source of difficulty in addition to subsequent spatial computation.

The contributions of this work are threefold:

- **A benchmark for cross-view spatial reasoning.** MultiFloorQA contains 608 question–answer pairs over multiple architectural drawings, organized by six spatial task categories and three evidence scopes: Single-floor, Cross-floor, and Cross-projection.
- **A scalable and verifiable benchmark construction pipeline.** A shared building representation supports drawing generation, question instantiation, and reference-answer computation, with staged quality checks maintaining consistency between visual evidence and question–answer pairs.
- **A systematic analysis of spatial reasoning in representative MLLMs.** Evaluations on MultiFloorQA-Main and MultiFloorQA-Hard, together with controlled input-condition comparisons, expose a gap between high aggregate accuracy and reliable spatial reasoning, and identify geometric information extraction from drawings and precise spatial computation as important remaining challenges.

2 Related Work

2.1 Visual and Spatial Reasoning Benchmarks

Multimodal benchmarks have expanded from object- and attribute-level visual understanding to increasingly complex forms of reasoning. MMBench (Liu et al., 2024) evaluates perception and reasoning over objects, attributes, and relations, while MMMU (Yue et al., 2024) and MathVista (Lu et al., 2024) extend evaluation to disciplinary knowledge, mathematical reasoning, and multi-step problem solving. Such broad evaluations characterize general multimodal capability, but strong performance does not by itself establish reliable spatial reasoning.

More targeted benchmarks therefore isolate specific visual-spatial abilities. BLINK (Fu et al., 2024) evaluates depth, direction, correspondence, and other spatial judgments, shifting attention from broad visual understanding toward individual perceptual and spatial operations. Real-3DQA (Ma et al., 2026) further shows that visual question-answering performance can be inflated by textual shortcuts, motivating evaluations in which correct answers depend on the underlying visual-spatial evidence. These studies motivate capability-oriented evaluation in which spatial operations are examined directly rather than inferred from aggregate multimodal performance.

Table 1 compares MultiFloorQA with representative benchmarks on engineering-drawing understanding, structured-layout reasoning, and multi-view spatial reasoning.

2.2 Cross-View Spatial Reasoning

As vision-language models have gained support for multiple images and extended visual contexts, increasing attention has been paid to associating and composing information distributed across observations. ReMI (Kazemi et al., 2024) evaluates joint multi-image reasoning across mathematical, logical, diagrammatic, and spatial tasks. VSI-Bench (Yang et al., 2025) studies how spatial information is acquired, retained, and used from sequential video observations, while MMSI-Bench (Yang et al., 2026) organizes multi-image spatial tasks around relations among cameras, objects, and regions. Disjoint-3DQA (Ravi et al., 2025) further evaluates spatial inference between objects that never co-occur in the same frame.

Beyond measuring final accuracy, recent work has also begun to separate local information recognition from cross-context composition. Compose and Fuse (Wang et al., 2026) controls how information is distributed across inputs or modalities to distinguish failures in extracting local evidence from failures in combining information across contexts. Collectively, these studies show that correctly interpreting individual observations does not necessarily lead to reliable information integration:

Table 1: Comparison of MultiFloorQA with representative benchmarks for spatial reasoning and architectural-drawing understanding. Checkmarks indicate task coverage. Diversity denotes the number of reported task categories or question templates, and Size denotes the number of evaluation instances. Answer formats are MC (multiple-choice), Numeric (tolerance-scored), Free-form (LLM- or human-judged), and Structured (rule-verified IDs, sets, routes, or JSON).

Benchmark	Input	Answer format	Counting	Cross-view matching	Multi-floor	Quant. geometry	Navigation	Diversity	Size
FloorplanQA	JSON/XML	Numeric + Structured	×	×	×	✓	✓	8	16,000
VSI-Bench	Video	MC + Numeric	✓	✓	×	✓	✓	8	5,130
MMSI-Bench	Images	MC	×	✓	×	✓	✓	11	1,000
DrawingVQA	Drawings	MC + Free-form	✓	✓	×	✓	×	7	92
AECV-Bench	Drawings	Numeric + Free-form	✓	×	×	×	✓	4	192
AEC-Bench	Drawings + docs	Structured	×	×	×	×	×	9	196
MultiFloorQA (Ours)	Drawings	Numeric + Structured	✓	✓	✓	✓	✓	**32**	**608**

cross-view correspondence, spatial memory, and relation composition constitute distinct capability requirements. They motivate evaluation settings in which complementary views must be integrated under explicit spatial constraints.

2.3 Architectural and Engineering Drawing Understanding

A building is often represented by multiple related drawings, making architectural question answering a natural setting for evaluating cross-view spatial reasoning. Models may need to identify entities within individual drawings, preserve entity identity across floors, establish correspondences between floor plans and elevations, and further compose precise geometric relations with physical connectivity.

However, existing work covers only parts of this problem. ArchPlanVQA (Li et al., 2026) and AECV-Bench (Kondratenko et al., 2026) mainly evaluate component recognition, instance counting, attribute extraction, and spatial-relation judgments on architectural or engineering drawings. DrawingVQA (Jung et al., 2026) further studies contextual interpretation and professional reasoning over complex construction drawings, while AEC-Bench (Mankodiya et al., 2026) focuses on references, source tracing, and consistency across engineering documents. FloorplanQA (Rodionov et al., 2026) approaches indoor spatial reasoning from structured JSON/XML layouts and evaluates tasks such as distance, visibility, object placement, and path reasoning, but its entities, geometry, and semantics are already provided explicitly. Existing benchmarks therefore do not systematically cover settings in which models must recover precise geometry directly from multiple architectural drawings, maintain correspondence across floors or projections, and compose geometric relations with physical connectivity.

MultiFloorQA targets this gap by providing verifiable cross-view spatial reasoning tasks whose evidence is distributed within individual drawings, across floors, and across projections.

3 The MultiFloorQA Benchmark

MultiFloorQA evaluates whether multimodal large language models can read, associate, and compose spatial information within and across architectural drawings. The benchmark is constructed from a shared building representation that connects drawing generation, question construction, and programmatic answer computation. This section summarizes the benchmark formulation, drawing construction, question design, and quality-control procedure.

3.1 Benchmark Overview

MultiFloorQA is a multi-image question-answering benchmark for cross-view spatial reasoning over architectural drawings. It contains 608 question–answer pairs, comprising 482 instances in MultiFloorQA-Main and 126 in MultiFloorQA-Hard. Each instance provides one or more floor plans or elevations, a question, and the task conventions required to interpret the answer. Models

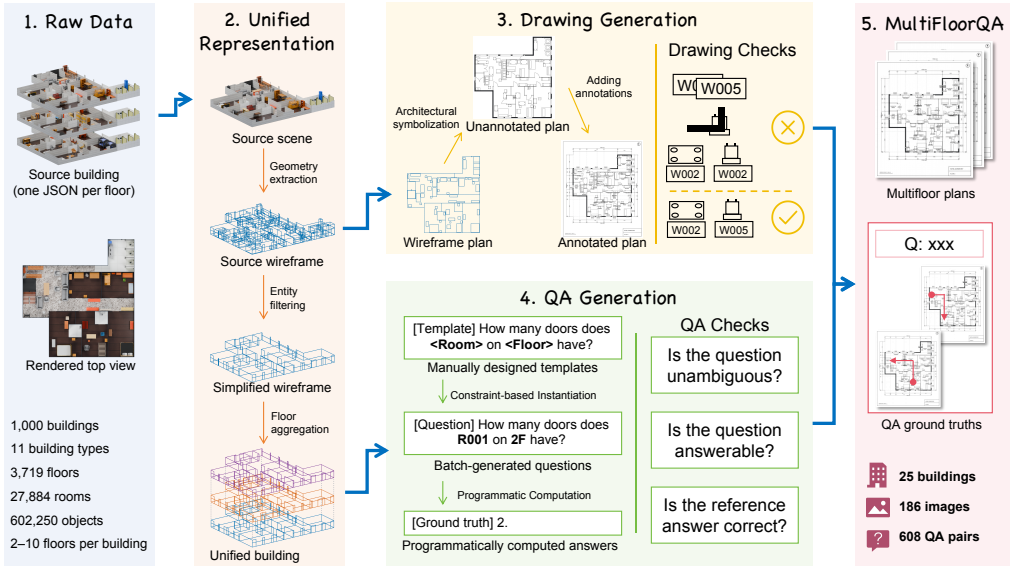


Figure 2: Construction pipeline of MultiFloorQA. Source building data are converted into a unified representation used to generate architectural drawings and instantiate constrained question templates with programmatic reference answers. Drawing and question–answer checks produce the final benchmark.

receive only these inputs, while the underlying building representation, coordinates, spatial graphs, and executable solution programs remain hidden.

Questions are organized into six spatial task categories: Recognition, Counting, Distance, Geometry, Matching, and Navigation. They are also classified into three evidence scopes: Single-floor, Cross-floor, and Cross-projection. The task category specifies the spatial operation required to obtain the answer, while the evidence scope specifies where the necessary information is distributed.

3.2 Source Data and Drawing Construction

MultiFloorQA is constructed from MansionWorld (Che et al., 2026). Rooms, objects, and building components are converted into a unified building representation containing their geometry, semantics, floor identity, identifiers, and connectivity in a common coordinate system. The same representation supports both drawing generation and programmatic reference-answer computation. Floor-level records are kept in a shared metric coordinate system, so positions and distances can be compared directly across floors, and visible objects are restricted to eight common indoor categories.

Floor plans contain room boundaries, walls, doors, windows, selected labeled objects, dimensions, scale, and orientation, while elevations provide complementary vertical and facade information. Architectural symbols are partly sourced from and partly redrawn with reference to the OpenCADLibrary resource (OpenCADLibrary). Only entities represented in the generated drawings are used by the corresponding question and answer programs, maintaining consistency between visible evidence and reference answers.

Figure 3 illustrates the visual information provided in a generated floor plan. Together, the drawing elements provide evidence for entity recognition, cross-view matching, distance and area computation, and connectivity judgments.

3.3 Question Design

MultiFloorQA uses 32 question templates grouped into six categories: Recognition, Counting, Distance, Geometry, Matching, and Navigation. Together, they cover local information extraction,

quantitative geometric reasoning, cross-view correspondence, and spatial connectivity. Questions are additionally assigned to Single-floor, Cross-floor, or Cross-projection according to the drawings required for their solution.

Each template specifies the eligible entities, required drawings, task conventions, parameters, and answer form. Valid instances are instantiated from the unified building representation, and reference answers are computed by deterministic geometric and topological operators. The complete template catalog, computational conventions, and validity conditions are provided in Appendix C.

Main and Hard question sets. MultiFloorQA-Main contains 482 questions from 16 buildings. MultiFloorQA-Hard contains 126 questions from nine taller buildings with eight to ten floors each, drawn from source families disjoint from those in MultiFloorQA-Main. It reuses 14 of the 32 templates, covering Counting, Distance, Geometry, and Navigation, but instantiates them at a larger spatial scale: area aggregation and signal coverage extend over all floors of a building, cross-floor projection spans the full building height, and navigation routes traverse up to ten floors or several simultaneously blocked doors. Recognition and Matching are excluded because they are already near ceiling for the strongest models on MultiFloorQA-Main.

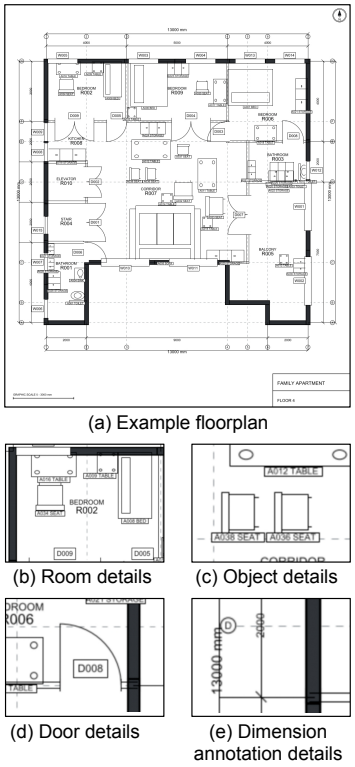


Figure 3: Visual information in a generated floor plan.

3.4 Quality Control

Quality control is applied at both the drawing and question–answer levels. Drawing checks verify that required geometry and entities are present and readable and that correspondences between related views are consistent. Question–answer checks verify that every referenced entity appears in the drawings supplied to the model, remove ambiguous or unsupported instances, and recompute reference answers using the corresponding deterministic spatial operators. Drawing issues are classified as hard errors, which invalidate a drawing, or quality warnings, which are retained for review. The checks cover one-to-one mappings among source entities, rendered symbols, and displayed identifiers, as well as collisions between annotation labels. Templates that a building cannot support are not instantiated for that building.

4 Experiments

4.1 Experimental Setup

We evaluate ten representative MLLMs on all 482 questions in MultiFloorQA-Main: GPT-6 Astra (OpenAI, 2026a) (max), GPT-5.6 Sol (OpenAI, 2026b) (max), Claude Fable 5.1 (Anthropic, 2026) (max), Qwen3.8 Max 0902 (Alibaba Cloud, 2026b) (xhigh), Gemini 3.1 Pro Preview (Google DeepMind, 2026) (high), Qwen3.8 Flash (Alibaba Cloud, 2026a) (xhigh), Kimi K3 (Moonshot AI, 2026) (max), GLM-5.3-Flash (Z.ai, 2026) (max), DeepSeek-V4.1-Flash (DeepSeek, 2026) (max), and MiniMax M3 (MiniMax, 2026) (adaptive). The parenthesized value denotes the reasoning-effort setting used for each model. On MultiFloorQA-Hard, we evaluate five of these models on all 126 questions: GPT-6 Astra, GPT-5.6 Sol, Gemini 3.1 Pro Preview, Qwen3.8 Flash, and DeepSeek-V4.1-Flash, which span a broad range of Main performance and use the same reasoning-effort settings. Evaluations on MultiFloorQA-Main and MultiFloorQA-Hard use a zero-shot protocol. Models receive the drawings required by the question, the question itself, and the corresponding task conventions, and return a final answer in the specified format. Because MultiFloorQA-Main and MultiFloorQA-Hard differ in buildings, template composition, and the number of drawings per question, category-level scores on Main and Hard are compared descriptively rather than as matched measurements.

Table 2: Results on **MultiFloorQA-Main**: accuracy (%), average reported tokens (k), and average request time (s).

Model	Overall	Rec.	Count	Dist.	Geom.	Match	Nav.	Tokens (k)	Time (s)
GPT-6 Astra	**90.25**	**100.00**	**93.51**	**91.24**	**69.35**	**100.00**	**89.81**	8.50	109.61
GPT-5.6 Sol	76.76	92.86	72.73	78.10	51.61	**100.00**	74.07	13.22	201.09
Claude Fable 5.1	74.27	97.62	74.03	75.18	43.55	94.64	71.30	29.17	199.04
Qwen3.8 Max 0902	72.61	95.24	64.94	68.61	51.61	**100.00**	72.22	11.86	172.54
Gemini 3.1 Pro Preview	63.69	95.24	64.94	46.72	45.16	94.64	66.67	14.08	108.28
Qwen3.8 Flash	57.26	73.81	51.95	51.82	38.71	83.93	58.33	11.47	99.34
Kimi K3	69.50	92.86	63.64	62.04	43.55	92.86	76.85	27.29	589.69
GLM-5.3-Flash	62.45	76.19	66.23	54.74	32.26	91.07	66.67	35.68	313.56
DeepSeek-V4.1-Flash	49.59	61.90	49.35	28.47	43.55	75.00	62.04	29.73	80.82
MiniMax M3	35.06	42.86	32.47	29.93	27.42	46.43	38.89	14.26	68.19

Table 3: Results on **MultiFloorQA-Hard**: accuracy (%), average reported tokens (k), and average request time (s).

Model	Overall	Count	Dist.	Geom.	Nav.	Tokens (k)	Time (s)
GPT-6 Astra	**63.49**	**55.56**	**51.85**	**63.89**	**73.33**	23.92	275.34
GPT-5.6 Sol	36.51	0.00	44.44	47.22	37.78	27.71	246.69
Gemini 3.1 Pro Preview	26.19	0.00	11.11	44.44	31.11	24.52	102.82
Qwen3.8 Flash	23.81	5.56	18.52	30.56	28.89	23.86	145.11
DeepSeek-V4.1-Flash	15.08	0.00	7.41	19.44	22.22	33.29	83.16

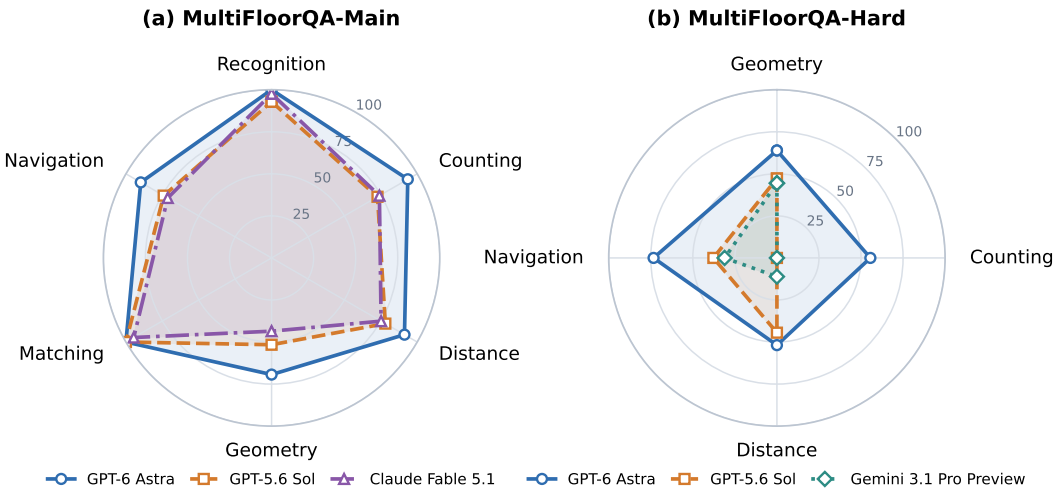


Figure 4: Category-level accuracy for the three highest-overall models on (a) MultiFloorQA-Main and (b) MultiFloorQA-Hard. Values are percentages.

Model responses are normalized before scoring. Categorical, Boolean, count, entity, and set answers use exact matching after normalization, with set order ignored where appropriate. Distance and area answers allow a 5% relative error. Navigation answers are evaluated according to the validity and optimality requirements of the corresponding task. If a response reaches the output limit, one follow-up prompt requests the final answer from the existing response, and this final answer is scored. Empty or unparseable answers are counted as incorrect. Model identifiers, endpoints, evaluation dates, inference settings, output handling, normalization rules, and prompts are given in Appendix E.

4.2 RESULTS ON MULTIFLOORQA-MAIN

GPT-6 Astra achieves the highest overall accuracy on MultiFloorQA-Main at 90.25%, followed by GPT-5.6 Sol at 76.76%, Claude Fable 5.1 at 74.27%, and Qwen3.8 Max 0902 at 72.61%. Overall accuracy ranges from 35.06% to 90.25% across the ten models, indicating substantial variation in architectural spatial reasoning among current MLLMs.

Performance also varies across spatial operations. Recognition and Matching are relatively reliable for several strong models, whereas Geometry remains substantially more difficult. GPT-6 Astra reaches 100% on both Recognition and Matching but 69.35% on Geometry. It correctly answers every question that sums labeled room areas across two floors, but reaches only 40.0% and 46.7% on the two net-usable-area templates and 43.8% on first-collision distance. The net-area templates require reading the dimension annotations, using them or the graphic scale to size each object footprint, and subtracting these footprints from the room area. Reliable entity identification and cross-view correspondence therefore do not necessarily translate into metrically precise spatial computation, and high aggregate accuracy can coexist with localized failures in precise geometric operations. Category profiles also differ across models: Gemini 3.1 Pro Preview reaches 95.24% on Recognition but 46.72% on Distance, and DeepSeek-V4.1-Flash reaches only 28.47% on Distance. Matching, by contrast, is close to ceiling: three models reach 100%, and all models except DeepSeek-V4.1-Flash and MiniMax M3 exceed 83%. Within the Qwen3.8 family, Max outperforms Flash by 15.35 points overall, with the largest gap on Recognition (95.24% vs. 73.81%).

Evidence scope does not impose a simple difficulty ordering. On Main, all ten models achieve higher accuracy on Cross-floor than on Single-floor questions (e.g., 94.2% vs. 87.1% for GPT-6 Astra and 71.3% vs. 56.8% for Gemini 3.1 Pro Preview), as single-floor questions can still require dense visual reading and multi-step geometric computation.

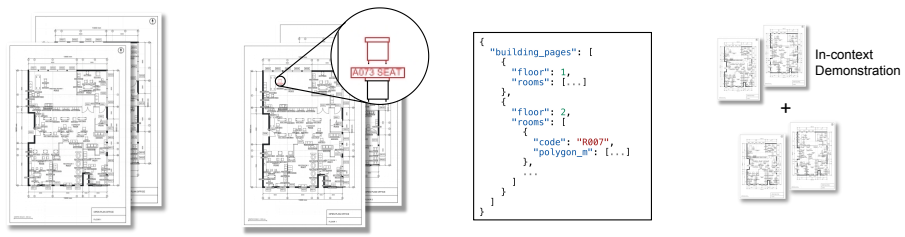
Accuracy does not track reported token usage across models. GPT-6 Astra uses the fewest reported tokens of all models (8.50k on average) while achieving the highest accuracy, whereas Claude Fable 5.1 and GLM-5.3-Flash use 29.17k and 35.68k tokens but reach 74.27% and 62.45%, respectively.

4.3 RESULTS ON MULTIFLOORQA-HARD

MultiFloorQA-Hard is substantially more difficult for all five evaluated models. GPT-6 Astra again achieves the highest overall accuracy, but at 63.49% compared with 90.25% on Main, while the other four models range from 15.08% to 36.51%. For Astra, the largest declines occur in Counting (93.51% to 55.56%) and Distance (91.24% to 51.85%), whereas Geometry changes less (69.35% to 63.89%), and Navigation remains its strongest Hard category at 73.33%. The gap between models also widens: Astra leads GPT-5.6 Sol by 13.5 percentage points on Main but by 27.0 points on Hard. The ranking of the five models is nevertheless identical on the two subsets, suggesting that Hard amplifies existing differences rather than reordering the models. Average reported token usage rises for all five models, from 8.50k to 23.92k for Astra and from 13.22k to 27.71k for Sol, yet accuracy still falls sharply. Templates that require evidence to be combined across the full building height are particularly difficult: full-height projection counting and multi-floor route listing yield zero or near-zero accuracy for most models. Counting is the most severe case, with GPT-5.6 Sol, Gemini 3.1 Pro Preview, and DeepSeek-V4.1-Flash all scoring 0.00%. Full-height projection counting credits an answer only when every projected floor is correct. Only 4 of the 45 final responses (8.9%) meet this criterion, all from GPT-6 Astra. When the relevant rooms, objects, and connections are distributed over many floors, an error in retaining or combining any part of this information can invalidate the final answer, making larger-scale spatial composition markedly less reliable.

4.4 INPUT-CONDITION COMPARISONS

To examine how the input form affects accuracy, we conduct controlled input-condition comparisons with GPT-5.6 Sol (max), DeepSeek-V4.1-Flash (max), and GLM-5.3-Flash (max). The three models are evaluated on the same fixed 64-question subset of MultiFloorQA-Main under five conditions: the original architectural drawings, structured JSON replacing the drawings, red-marker highlighting of question-specified entities, one-shot prompting, and a no-image condition. Across conditions, the target questions, reference answers, scoring rules, and inference settings remain unchanged. Figure 5 illustrates the four conditions with visual or structured input. The no-image condition provides only



(a) Original Drawings (b) Target highlighting (c) Structured JSON (d) One-shot Prompting

Figure 5: Input conditions used in the controlled comparison. (a) Original drawings, (b) red-marker highlighting, (c) structured JSON replacing the drawings, and (d) one-shot prompting. The no-image condition is not shown.

Table 4: Accuracy (%) under five input conditions. Parentheses show percentage-point changes from the original-drawing baseline. Positive changes are green, and negative changes are red. The highest value in each column is boldfaced.

Model	Original	JSON	Red markers	1-shot	No image
GPT-5.6 Sol	**79.69**	**100.00** (+20.31)	**78.12** (-1.56)	**76.56** (-3.12)	23.44 (-56.25)
DeepSeek-V4.1-Flash	45.31	98.44 (+53.12)	53.12 (+7.81)	43.75 (-1.56)	**25.00** (-20.31)
GLM-5.3-Flash	62.50	98.44 (+35.94)	67.19 (+4.69)	59.38 (-3.12)	21.88 (-40.62)

the textual question and task conventions. The subset contains two instances of each of the 32 templates, drawn from three buildings. The JSON condition is an oracle input derived from the source building records, which real projects rarely provide alongside their drawings. It exposes the geometry and explicit relations of the supplied drawing pages but excludes reference answers and solution programs. One-shot examples come from buildings outside all 25 evaluation buildings. Red markers recolor the specified labels and symbols without altering the drawings. Ten questions have no markable target and reuse their original-drawing results.

Replacing the drawings with structured JSON produces the largest improvement for all three models. Accuracy rises from 79.69% to 100.00% for GPT-5.6 Sol, from 45.31% to 98.44% for DeepSeek-V4.1-Flash, and from 62.50% to 98.44% for GLM-5.3-Flash, so that nearly every question is answered correctly once the drawing content is given explicitly. In contrast, red-marker highlighting produces smaller and model-dependent changes, while one-shot prompting does not improve accuracy for any of the three models. Removing the images substantially reduces performance, with accuracy falling to 21.88–25.00% across the three models.

Under the tested setup, neither worked examples nor visual highlighting of question-relevant entities yields consistent improvements. The much larger gains from structured JSON instead suggest that extracting, localizing, and organizing geometric information from drawings is an important source of difficulty. However, JSON changes both the input representation and the explicitness of the supplied spatial information, so the comparison does not provide a complete causal separation between visual information extraction and subsequent spatial reasoning.

5 Conclusion

We introduce MultiFloorQA, a benchmark for evaluating cross-view spatial reasoning over multiple architectural drawings. Results on MultiFloorQA-Main and MultiFloorQA-Hard show that strong aggregate performance can coexist with persistent difficulty in precise geometric computation and larger-scale spatial composition. On the taller buildings of MultiFloorQA-Hard, the gap between models widens, and whole-building counting remains largely unsolved. Controlled input comparisons further indicate that recovering and organizing geometry from the drawings themselves is a major source of difficulty. These findings motivate further research on visual grounding, geometric computation, and cross-view spatial integration in multimodal models.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and language polishing, literature search, parts of the code implementation, aggregation of experimental results, and discussions of experimental design and interpretation. Benchmark questions are instantiated through deterministic templates, and reference answers are computed by deterministic spatial programs. The authors verified AI-assisted content and are responsible for the final manuscript, results, and conclusions.

ETHICS STATEMENT

MultiFloorQA consists of programmatically generated architectural drawings and template-based questions. This work involves no human subjects, crowdsourced annotation, or personally identifiable information, and the drawings do not depict real, identifiable buildings.

REPRODUCIBILITY STATEMENT

Benchmark construction is fully deterministic: questions are instantiated from templates and reference answers are computed by spatial programs (Section 3). Drawing construction, rendering conventions, and entity-selection rules are detailed in Appendix B. The question templates and executable ground truth are given in Appendix C, and evaluation prompts, model settings, and scoring rules are given in Appendix E and Appendix H. The code and benchmark data will be publicly released upon publication.

REFERENCES

Alibaba Cloud. Qwen3.8 Flash. https://www.qwencloud.com/models/qwen3.8-flash, 2026a. Accessed September 23, 2026.

Alibaba Cloud. Qwen3.8 Max 0902. https://www.qwencloud.com/models/qwen3.8-max-0902, 2026b. Accessed September 23, 2026.

Anthropic. Claude Fable 5.1. https://www.anthropic.com/claude/fable, 2026. Accessed September 23, 2026.

Lirong Che, Shuo Wen, Shan Huang, Chuang Wang, Yuzhe Yang, Gregory Dudek, Xueqian Wang, and Jian Su. MANSION: Multi-floor language-to-3D scene generation for long-horizon tasks. *arXiv preprint arXiv:2603.11554*, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian 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*, pp. 14455–14465, 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093, 2024.

DeepSeek. DeepSeek-V4.1-Flash Release Notes. https://api-docs.deepseek.com/news/news260910/, 2026. Accessed September 23, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024.

Google DeepMind. Gemini 3.1 Pro Preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview, 2026. Accessed September 23, 2026.

Yoonhwa Jung, Junryu Fu, and Mani Golparvar-Fard. DrawingVQA: A real-world benchmark for multi-depth visual-textual reasoning on construction drawings. *arXiv preprint arXiv:2607.15418*, 2026. Accepted to CVPR 2026 Findings.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, and Ahmed Qureshi. ReMI: A dataset for reasoning with multiple images. *arXiv preprint arXiv:2406.09175*, 2024.

Aleksei Kondratenko, Mussie Birhane, Houssame E. Hsain, and Guido Maciocci. AECV-Bench: Benchmarking multimodal models on architectural and engineering drawings understanding. *arXiv preprint arXiv:2601.04819*, 2026.

Pengyang Li, Hainan Chen, Zhuang Zheng, Juntao He, and Hai Yang. ArchPlanVQA: Benchmark and comprehensive evaluation of vision-language models for understanding architectural floor plan CAD drawings. *Journal of Computing in Civil Engineering*, 40(5):04026060, 2026. doi: 10.1061/JCCEE5.CPENG-7571.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X. Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3D large language models really understand 3D spatial relationships? In *International Conference on Learning Representations*, 2026.

Harsh Mankodiya, Chase Gallik, Theodoros Galanos, and Andriy Mulyar. AEC-Bench: A multimodal benchmark for agentic systems in architecture, engineering, and construction. *arXiv preprint arXiv:2603.29199*, 2026.

MiniMax. MiniMax M3 Release Notes. `https://www.minimax.io/blog/minimax-m3`, 2026. Accessed September 23, 2026.

Moonshot AI. Kimi K3. `https://www.kimi.ai/blog/kimi-k3`, 2026. Accessed September 23, 2026.

OpenAI. GPT-6 Astra. `https://openai.com/index/gpt-6-astra/`, 2026a. Accessed September 23, 2026.

OpenAI. GPT-5.6 Sol. `https://openai.com/index/gpt-5-6/`, 2026b. Accessed September 23, 2026.

OpenCADLibrary. OpenCADLibrary: Cad blocks and drawing resources. `https://opencadlibrary.com/`. Accessed September 25, 2026.

Sahithya Ravi, Gabriel Herbert Sarch, Vibhav Vineet, Andrew D. Wilson, and Balasaravanan Thoravi Kumaravel. Out of sight, not out of context? egocentric spatial reasoning in VLMs across disjoint frames. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16135–16150. Association for Computational Linguistics, 2025.

Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. FloorplanQA: A benchmark for spatial reasoning in LLMs using structured representations. In *International Conference on Machine Learning*, 2026.

Yucheng Wang, Yifan Hou, Aydin Javadov, Mubashara Akhtar, and Mrinmaya Sachan. Compose and fuse: Revisiting the foundational bottlenecks in multimodal reasoning. In *International Conference on Learning Representations*, 2026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Z.ai. GLM-5.3-Flash Model Card. `https://huggingface.co/zai-org/GLM-5.3-Flash`, 2026. Accessed September 23, 2026.

A Technical Appendix and Supplementary Material

This supplementary material provides reproducibility details for MultiFloorQA and its evaluation. Section B describes the source data, drawing generation, and quality control. Section C gives the 32-template taxonomy, executable reference answers, and question validation procedure. Section D documents MultiFloorQA-Hard, including its building and template selection, dataset statistics, and full template-level and per-building results. Section E reports the general evaluation protocol and full MultiFloorQA-Main results. Section F presents controlled input-condition comparisons, Section G summarizes representative error patterns, Section H gives representative prompts, and Section I records limitations and additional discussion.

B Architectural Drawing Generation and QA Scene Construction

B.1 Data Source

We select MansionWorld as the source of synthetic building data because its structured records support both multi-floor architectural drawing generation and the computation of spatial question-answering ground truth. Each building is organized into explicit floors, with information about rooms, doors, windows, furniture objects, and floor heights. From the same structured source, we generate floor plans for individual levels and an east elevation for the entire building.

B.2 Building-level Representation

MansionWorld provides structured information for each floor of a building. We organize these records into a building-level representation that is shared by the drawing generation and question-answer generation pipelines. Each building contains multiple floors with their source-provided heights. Within each floor, rooms are represented by their semantic types and polygonal boundaries. Doors and windows retain their geometric locations and associations with adjacent rooms, while indoor objects retain their categories, room assignments, positions, and footprints. This representation organizes floor-level records into a building-level structure while preserving the original geometry and semantics.

The geometry of all floors is kept in the same metric coordinate system provided by MansionWorld. We do not independently rotate, translate, or rescale individual floors before spatial relationships are computed. As a result, positions and distances can be compared directly across floors using the underlying building coordinates rather than the pixel coordinates of the rendered drawings. Floor heights are also retained and accumulated to recover the vertical position of each level. For stairs and elevators, cross-floor correspondence is established between adjacent floors only when the regions agree in both semantic type and geometry. This correspondence identifies the same vertical connector across floors but does not by itself imply a traversable connection. These choices provide a consistent geometric basis for questions involving multiple floors.

Entity identities are preserved throughout the generation process so that the same room, opening, or object can be traced from the source data to the generated drawings and their associated questions. The visual labels used in the drawings are therefore separated from the underlying entity identities. For doors and windows that appear in both a floor plan and an elevation, we additionally maintain a cross-view identifier based on their source identity and floor. This allows the same opening to be matched across views without relying on its rendered label, image position, or visual appearance. The resulting correspondence provides the entity-level supervision required for questions that combine evidence from floor plans and elevations.

B.3 Visible Entity Selection and Semantic Mapping

B.3.1 Visible object categories

MansionWorld contains a broad range of indoor objects, including decorative elements, accessories, and long-tail categories with less consistent semantics. To maintain a clear correspondence between the visual drawings and the subsequent spatial questions, we restrict the visible object set to eight

common indoor categories: beds, sofas, tables, seats, storage furniture, appliances, sinks, and toilets. Source object names are mapped to these common categories before being rendered with the corresponding architectural symbols. Classification is based on the object itself rather than inferred from its surrounding room or host context, reducing the possibility of introducing additional semantic cues during preprocessing.

The eight supported categories define the objects shown in the drawings and counted in the benchmark. Decorative and unsupported objects remain in the source records for provenance.

B.3.2 Consistency between visible entities and QA

The QA entity set follows the final drawings. Source objects absent from the drawings are excluded from object counts, nearest-object queries, spatial coverage, area computations, and navigation obstacles. The reference answers therefore use the entities shown to the model.

Each included object has a graphical symbol, entity label, and associated geometry. Crowded annotations and footprint overlaps are recorded as quality warnings; question selection can exclude ambiguous targets.

B.3.3 Eligibility of source objects

The visible entity set includes supported-category objects with an unambiguous identity, valid two-dimensional geometry, and a resolved room assignment. Records missing any of these properties are excluded. Potential geometric or ownership inconsistencies in otherwise eligible records are logged as quality warnings for question selection and review.

B.3.4 Composite objects and counting units

Some MansionWorld object names describe furniture with multiple functions, such as combined seating and sleeping furniture, storage seating, or furniture integrated with appliances or sanitary fixtures. Splitting such names into several benchmark objects would introduce additional entities and geometries that are not present in the source data. We therefore treat each source object record as a single entity and a single counting unit, assigning it one primary semantic category. Additional functions may be retained as attributes, but they do not create new objects.

For example, a sofa bed remains one sofa entity while retaining its sleeping function, a storage ottoman is treated as a seat, a workbench as a table, and a mirror cabinet as storage furniture.

When a compound name does not identify a primary category, the existing classification is retained and flagged for review.

This distinction is particularly important for counting questions. A compound object name is not interpreted as several independently countable components unless MansionWorld itself provides separate source entities. A single source entity therefore corresponds to one drawing symbol, one entity label, and one counting unit in the benchmark.

B.3.5 Entity labels in rendered drawings

To allow questions to refer unambiguously to entities in the drawings, we assign compact labels to the main drawing elements. Rooms, objects, doors, and windows use the prefixes R, A, D, and W, respectively, followed by three-digit identifiers within each floor plan. The assignment is deterministic, so the same source data under the same drawing rules produces a consistent entity mapping, while each visible label retains a one-to-one link to its underlying source entity.

Labels are local to each drawing page; the same numerical label on different floors need not identify the same entity. Cross-floor and cross-view correspondence uses the persistent source identities described in Section B.2.

B.4 Architectural Drawing Generation

Based on the building representation described in Section B.2, we generate one floor plan for each level and one east elevation spanning the complete building.

B.4.1 FLOOR PLAN GENERATION

For each floor, we generate a CAD-style plan from the room polygons, openings, and visible objects defined in Section B.3. Room boundaries are directly inherited from the two-dimensional geometry in the building representation. Indoor objects are rendered using plan-view symbols associated with their semantic categories. Each symbol follows the orientation of its source footprint and is scaled according to the corresponding object extent. Objects of the same category therefore share a consistent graphical representation across buildings, while individual instances preserve their original position, scale, and orientation. Spatial computations continue to use the original object footprints, whereas the symbols provide a clearer visual representation in the drawings.

Openings require an additional geometric step because MansionWorld provides opening segments, while architectural drawings require doors and windows to be aligned with their supporting walls. For each opening, we search the room boundaries for compatible edges, prioritizing the rooms connected by the opening. The opening midpoint is projected onto the selected boundary while preserving its original width. This procedure corrects small offsets between opening segments and room boundaries while maintaining the original opening geometry. When no suitable host boundary is identified, the original opening segment is retained.

After geometric placement, openings are converted into conventional plan symbols. Doors are rendered as single- or double-leaf symbols according to their opening width, while windows use a consistent plan-view representation. Where walls are displayed as filled regions, a local cutout is created at each opening before the corresponding door or window symbol is drawn, keeping openings visually distinguishable from surrounding wall regions.

Entity labels are added after the drawing geometry has been established. Object labels are positioned near their corresponding footprints using candidate-based placement while considering nearby objects, existing labels, and major drawing elements. Door and window labels follow the same strategy around their associated openings. The resulting floor plans combine architectural geometry, semantic symbols, and entity references in a unified visual representation for recognition, counting, and spatial reasoning tasks.

B.4.2 EAST ELEVATION GENERATION

For each building, we additionally generate a single east elevation spanning all floors. For each level, the easternmost extent of the room polygons is first identified, and room edges lying on this boundary are extracted. Under an east-facing projection, these edges define the facade extent of the floor and preserve the divisions between adjacent rooms along the exterior boundary. The corresponding plan coordinate becomes the horizontal coordinate of the elevation. The vertical coordinate is constructed from the source story heights. Story heights are accumulated in floor order to obtain the base and top elevation of each level, allowing all floors to be placed within a common elevation coordinate system. The resulting multi-floor elevation therefore preserves both the plan geometry of the facade and the vertical dimensions provided by the building data.

Doors and windows in the elevation are generated from the same opening records used by the floor plans. An opening is projected only when it belongs to the exterior boundary and its plan segment lies on the east facade. Its horizontal span is obtained directly from the corresponding plan segment, while its vertical position and height are determined from the opening-height information together with the base elevation of its floor. The position and extent of each elevation opening can therefore be traced back to the corresponding architectural record.

During generation, we also check that the room edges form a continuous east-facade range and that each projected opening falls within the facade interval of its associated room. The resulting elevation preserves the geometric relationships among story heights, facade divisions, and exterior openings while remaining tied to the same building data used to generate the floor plans.

B.4.3 PLAN-TO-ELEVATION WINDOW CORRESPONDENCE

Windows in the floor plans and east elevation are associated through their shared source identity. For each plan window, we retain both its floor membership and its original opening identity. When the same window lies on the east facade and is projected into the elevation, the elevation record carries

the same floor and opening identity. The plan and elevation may assign different display labels to the window, but the two representations remain linked through this common identity.

The mapping supports correspondence in both directions. A window identified in a floor plan can be linked directly to its counterpart in the building elevation, while an elevation window can be traced back to its floor and corresponding plan window. Only windows that are present in both views are used for these cross-projection questions, keeping the correspondence aligned with the visual evidence provided to the model.

B.5 Drawing Conventions

To reduce drawing-style variation across the benchmark, all floor plans and east elevations are generated using a common set of visual conventions. We standardize the raster format, wall representation, linework, dimensions, and sheet annotations while keeping the underlying building geometry unchanged. The main drawing settings are summarized in Table 5.

Some elements in the drawings are introduced only for graphical presentation. In particular, the nominal wall thicknesses, inferred grid, north-mark orientation, and displayed door-swing direction are not physical attributes supplied by MansionWorld and are not used to compute QA ground truth. By contrast, room boundaries, object footprints, opening geometry, and story heights are taken directly from the source data. Metric answers are computed from this underlying geometry rather than from displayed dimensions or raster pixel distances.

B.6 Drawing Quality Control

After drawing generation, each floor plan and east elevation is subjected to automatic quality control to verify consistency with the source building representation. The checks cover three aspects: geometric and entity consistency, annotation readability, and east-elevation completeness. Detected issues are categorized into hard errors that invalidate a drawing and quality warnings that require further review.

B.6.1 Geometry and Entity Consistency

We first verify consistency between the rendered drawings and the visible-entity set. For every drawable object, the corresponding source entity record, CAD symbol, and displayed identifier must maintain a one-to-one mapping. After rendering, the generated object set is checked again to ensure that no eligible entity is missing and that no additional entity is introduced. Object category, room assignment, and the footprint used for spatial computation are also verified against the source records.

Doors and windows undergo similar completeness checks. Every visible opening must have a corresponding graphical representation and identifier, and every rendered door must contain the required door-leaf and swing representation. The drawing files are additionally checked for structural validity, and the final raster outputs are verified against the prescribed image dimensions.

A missing drawable entity is treated as a generation error.

Table 5: Drawing conventions used for benchmark rendering. Nominal graphical settings do not alter the source geometry used for QA.

Item	Setting
Raster format	1800×1350 px (4:3), 300 dpi.
Wall representation	Exterior walls: filled, nominal 200 mm; interior walls: double-line, nominal 120 mm.
Linework	Object symbols: 0.25 mm; label frames: 0.05 mm.
Dimensions	Four-sided dimension chains in mm; graphic scale.
Sheet annotations	Inferred grid, north mark, two-row title block.
Elevation layout	Same raster format; dedicated footer and elevation dimensions.

B.6.2 Readability and Overlap Checks

After entity completeness is verified, we evaluate the readability of the annotation layer. Object, opening, and room labels are checked against other labels, object footprints, room text, wall and opening linework, dimension chains, and title-block regions. Object labels are additionally checked to maintain a clear association with their corresponding footprints. For elevation drawings, geometric collision checks are supplemented with pixel-level inspection of the rendered text to identify overlaps introduced during rasterization.

Geometric overlap is checked separately. Footprint intersections are recorded, and substantial overlaps between objects of the same fine-grained category in one room are flagged as quality warnings. The objects remain separate counting units; question selection can exclude visually ambiguous targets.

B.6.3 Elevation Validation and Quality Levels

East elevations require additional validation to ensure the correctness of the plan-to-elevation projection. For each floor, we verify that projected eastern room boundaries form the expected facade range and that each elevation opening remains within the facade region of its associated room. The vertical extent of each opening is also checked against the corresponding story height. Cross-view mappings are required to remain unique so that each elevation opening can be linked to a single plan opening.

Elevation labels are checked using the same procedure as floor-plan annotations. After vector-level placement, the rendered PNG is inspected using text bounding boxes to detect overlaps after rasterization. The final raster dimensions are also verified against the required output format.

Detected issues are divided into two levels. Hard errors indicate failures that break the correspondence between the drawing and source data or prevent the drawing from serving as reliable visual input. Examples include missing source-to-render mappings, missing required opening representations, invalid CAD output, and incorrect raster dimensions. Quality warnings indicate that the underlying entities remain valid but that local presentation may be ambiguous, including crowded annotations, object overlaps, category-review cases, or facade discontinuities. Warning cases are retained for review and do not modify the entity set.

The number of visible windows in the east elevation is additionally used to determine whether a building supports elevation-specific question templates. Buildings with insufficient visible windows are excluded from those specific templates but remain available for other tasks.

B.7 Benchmark Statistics

Table 6 reports the complete statistics of the final benchmark. MultiFloorQA-Main contains 482 instances from 16 buildings, whereas MultiFloorQA-Hard contains 126 instances from nine taller buildings. Together, MultiFloorQA-Main and MultiFloorQA-Hard form the 608-instance MultiFloorQA benchmark.

Table 6: Complete statistics of MultiFloorQA-Main, MultiFloorQA-Hard, and the combined benchmark. Counts refer to final questions and distinct input images. †MultiFloorQA-Hard reuses 14 of the existing templates and four of the task categories rather than introducing new ones.

Statistic	Main	Hard	Total
Buildings	16	9	25
Floors represented	58	88	146
Input images	89	97	186
Standard floor plans	57	88	145
AP-marked floor plans	20	9	29
Elevation images	12	0	12
Question–answer pairs	482	126	608
Question templates	32	14†	32
Task categories	6	4†	6

C Question Taxonomy and Executable Ground Truth

C.1 Question Design and Taxonomy

MultiFloorQA evaluates entity identification and spatial reasoning over architectural drawings, including geometric computation, cross-view correspondence, and navigation. MultiFloorQA-Main contains 482 QA instances from 16 buildings, instantiated from 32 parameterized question templates. We organize the questions into six categories according to the primary spatial operation required to obtain the answer: Recognition, Counting, Distance, Geometry, Matching, and Navigation. The taxonomy therefore reflects the principal reasoning operation rather than merely the answer format or the type of entity appearing in the question.

Recognition concerns identifying entities and their immediate relationships in architectural drawings. A model must use graphical symbols, entity labels, and spatial context to determine an object's category, its containing room, or the spaces directly connected to a target room. These tasks primarily evaluate whether the model can establish the correspondence between visual entities and architectural semantics, which forms the basis for more complex spatial reasoning.

Counting requires identifying the entities belonging to a specified spatial or topological scope and then aggregating or comparing their quantities. Beyond directly counting doors or visible objects, some questions first require the relevant set to be determined from room adjacency, cross-floor projection, or a locally defined region. These tasks therefore combine visual enumeration with correct interpretation of the counting scope.

Distance covers questions whose answers are determined by relative position, distance, or spatial separation. In addition to Manhattan and Euclidean distances, this category includes wireless coverage, operational clearance, collision distance under object translation, center-line intersection, and cardinal direction. Although these questions have different output forms, their answers are primarily determined by the relative geometry of the involved entities.

Geometry focuses on geometric properties and spatial constraints, including room area, usable area after accounting for visible objects, and placement feasibility at a cross-floor projected location. Answering these questions requires room boundaries, visible-object footprints, and geometric relationships between floors to be considered jointly rather than relying on isolated labels or local visual cues.

Matching evaluates spatial correspondence across different floors or architectural views. The questions include correspondence between room positions or layouts across floors, as well as association of the same architectural element between a floor plan and the east elevation. Compared with single-view recognition, these tasks require consistent entity interpretation across multiple spatial contexts.

Navigation treats the building represented by the drawings as a traversable spatial environment. Models must reason over rooms, doors, cross-floor connections, and visible obstacles to determine minimum door transitions, shortest room sequences, metric shortest paths, or evacuation routes.

These tasks therefore combine entity grounding, topological understanding, and multi-step spatial reasoning.

The complete set of 32 parameterized question templates is shown in Table 7. The table reproduces the question forms used by the final QA generator and records the task category and evidence scope required to answer each template.

Table 7: MultiFloorQA question taxonomy. Templates are grouped by category. Evidence scope records where the visual evidence needed to answer a question is distributed. Single-floor uses one floor plan, Cross-floor combines multiple floor plans, and Cross-projection aligns a floor plan with the whole-building east elevation. The Remarks column specifies the rule used to compute each answer.

ID	Evidence scope	Question template	Remarks
Recognition			
R1	Single-floor	What category does the numbered object {object} belong to on Floor {floor}?	—
R2	Single-floor	Which room contains the numbered object {object} on Floor {floor}?	—
R3	Single-floor	On Floor {floor}, which rooms can be entered directly from Room {room} through a door?	—
Counting			
C1	Single-floor	On Floor {floor}, define the construction zone as Room {anchor_room} together with all rooms directly connected to it by one door. How many doors have one side inside the zone and the other side in a non-zone room or outdoors?	Exclude doors entirely within the zone; double-leaf doors count once.
C2	Single-floor	On Floor {floor}, how many labeled rooms have no door directly to the outdoors and exactly one interior door leading to another room?	Count physical doors individually; a double-leaf door counts once.
C3	Cross-floor	How many doors are there in total on Floors {lower_floor} and {upper_floor}, including exterior doors? Count each double door as one door.	A double-leaf door counts as one physical door.
C4	Cross-floor	What is the absolute difference between the numbers of visible A###-labeled objects in Room {room_1} on Floor {lower_floor} and Room {room_2} on Floor {upper_floor}?	—
C5	Cross-floor	Vertically project the full area of Room {upper_room} on Floor {upper_floor} onto Floor {lower_floor} and select the unique room with the largest positive overlap. How many visible A###-labeled objects are in that room?	Largest positive overlap must identify one room.
Distance			
D1	Single-floor	What is the Manhattan distance, in meters, between the centers of Room {room_a} and Room {room_b} on Floor {floor}?	A room center is the area centroid of its region.
D2	Single-floor	What is the straight-line distance, in meters, between the centers of Room {room_a} and Room {room_b} on Floor {floor}?	A room center is the area centroid of its region.

Continued on the next page

Table 7 *(continued)*

ID	Evidence scope	Question template	Remarks
D3	Single-floor	AP-01 on Floor {floor} has a signal radius of {radius} m. How many rooms on this floor can receive the signal? A room counts as covered if any part of it lies within the coverage region.	Coverage is circular on the floor plane; ignore wall and furniture attenuation; boundary contact counts.
D4	Single-floor	In Room {room} on Floor {floor}, does any other object intrude into the {clearance} m operational clearance around {object}? If so, return the nearest object; otherwise, return "none".	Measure between footprint boundaries; contact with the clearance boundary is intrusion.
D5	Single-floor	On Floor {floor}, a rectangular operational clearance extending {clearance} m outward from the due-{direction} side of {object} must remain clear. Does another object intrude into this region? If so, return the nearest object; otherwise, return "none".	The region spans the full corresponding footprint side; boundary contact is intrusion.
D6	Single-floor	On Floor {floor}, how far can the numbered object {object} be translated due {direction} before it first touches another visible object?	Object footprints use oriented bounding rectangles.
D7	Single-floor	On Floor {floor}, which other numbered objects are intersected by or touched by the line segment joining the centers of {object_a} and {object_b}?	Consider only the segment between the two object centers.
D8	Single-floor	On Floor {floor}, in which cardinal direction is Room {room_a} located relative to Room {room_b}?	Use the dominant horizontal or vertical component between room centroids.
D9	Cross-floor	AP-01 on Floor {lower_floor} has a signal radius of {radius} m, and the story height between Floors {lower_floor} and {upper_floor} is {height} m. How many rooms across the two floors can receive the signal?	Coverage is spherical; ignore wall and furniture attenuation; boundary contact counts.
Geometry			
G1	Single-floor	In Room {room} on Floor {floor}, what is the net usable area after subtracting the footprints of all visible A###-labeled objects?	Subtract the union of oriented bounding rectangles; do not double-count overlap.
G2	Single-floor	What is the total net usable area of all labeled rooms on Floor {floor} after subtracting the footprints of their visible A###-labeled objects?	Subtract the union of oriented bounding rectangles within each room.
G3	Cross-floor	What is the total area, in square meters, of all labeled rooms on Floors {lower_floor} and {upper_floor}?	—
G4	Cross-floor	Using the vertical projection of the center of the {room_type} on Floor {upper_floor} as the machine center, can a {machine_size} m × {machine_size} m machine be placed on Floor {lower_floor} without touching a wall or visible object?	The room center is its area centroid; the machine is axis-aligned; contact is a collision.

Continued on the next page

Table 7 *(continued)*

ID	Evidence scope	Question template	Remarks
Matching			
M1	Cross-projection	Which numbered window in the whole-building east elevation corresponds to Window {plan_window} in the Floor {floor} plan?	Match the same window across the floor plan and east elevation.
M2	Cross-projection	Which floor and which plan-window ID correspond to Window {elevation_window} in the whole-building east elevation?	The elevation window has one plan-window correspondence.
M3	Cross-floor	At the room-layout level, are Floors {lower_floor} and {upper_floor} identical?	Compare room boundaries at original positions; ignore names, IDs, furniture, doors, and windows.
M4	Cross-floor	Which room on Floor {lower_floor} corresponds directly below Room {upper_room} on Floor {upper_floor}?	Use the largest positive overlap after vertical projection.
Navigation			
N1	Single-floor	On Floor {floor}, what is the minimum number of doors that must be passed through to travel from Room {room_a} to Room {room_b}?	A double-leaf door counts once; an open passage counts as no door.
N2	Single-floor	On Floor {floor}, what is the shortest door route from Room {room_a} to Room {room_b}? List the rooms in order.	"Shortest" minimizes the number of door connections.
N3	Single-floor	On Floor {floor}, starting from the center of Room {room_a} and traveling only along rectilinear walkable routes to the center of Room {room_b}, what is the shortest distance in meters?	Visible-object footprints are obstacles; boundary-following is allowed.
N4	Single-floor	On Floor {floor}, which rooms are traversed, in order, along the shortest evacuation route from Room {start_room} to the outdoors?	Minimize door connections, including the final exterior door; list the start and final indoor rooms.
N5	Cross-floor	During construction, the unique door connecting Rooms {door_room_a} and {door_room_b} on Floor {door_floor} is impassable. What is the minimum number of passable door connections from Room {start_room} on Floor {lower_floor} to Room {target_room} on Floor {upper_floor}?	Restrict the route to the supplied floors; double-leaf doors count once and open passages as zero.
N6	Cross-floor	What is the shortest route from Room {start_room} on Floor {lower_floor} to Room {target_room} on Floor {upper_floor}? List the rooms in order.	"Shortest" minimizes door connections across the supplied floors.
N7	Cross-floor	Starting from the center of Room {start_room} on Floor {lower_floor} and traveling only along rectilinear walkable routes to the center of Room {target_room} on Floor {upper_floor}, what is the shortest distance in meters? The story height between the two floors is {height} m.	Visible objects are obstacles; the cross-floor component adds the stated story height.

C.2 Evidence Scope

In addition to task category, Table 7 records the evidence scope required to answer each template. Single-floor questions use evidence from one floor plan; Cross-floor questions combine floor plans from multiple levels; and Cross-projection questions align a floor plan with the whole-building east elevation. The 32 templates comprise 19 Single-floor, 11 Cross-floor, and 2 Cross-projection templates. The corresponding MultiFloorQA-Main instance counts are 287, 171, and 24, respectively.

The evidence scope describes where the necessary visual evidence is distributed, rather than imposing a fixed difficulty ordering. A single-floor question may still require dense visual reading or several geometric operations, while a cross-floor question may involve a relatively simple comparison. The scope labels therefore support stratified analysis without treating them as difficulty levels.

C.3 Executable Ground Truth

Reference answers in MultiFloorQA are computed directly from the shared building representation defined in Appendix B. The room boundaries, visible-object geometry, opening records, story heights, and cross-view correspondences used for drawing generation therefore also provide the data basis for executable QA ground truth.

For each question, the reference answer is obtained through a deterministic spatial operation corresponding to its definition. Area and occupancy questions operate directly on room and visible-object geometry; counting questions determine entity sets within a specified spatial or topological scope; distance and clearance questions use relative geometry in the underlying coordinate system; cross-floor and cross-view matching use stored entity correspondences across levels or views; and navigation questions are solved over spatial structures formed by rooms, doors, cross-floor connections, and traversable regions. Numerical results, entity sets, and paths can therefore all be recomputed and verified from the source building data. This construction also preserves consistency between visual evidence and QA ground truth. As specified in Section B, only objects actually rendered in the final drawings are included in the QA scene; source objects that are not displayed do not contribute to counting, distance, area, or navigation computations. Likewise, presentation-only elements such as nominal wall thickness and north-marker orientation are excluded from QA ground-truth computation; metric results are derived from the underlying building geometry rather than raster-pixel distances or displayed dimensions.

Depending on the template, the resulting answer may be a numerical value, category, entity identifier, entity set, or ordered route. For tasks admitting multiple equivalent optima, validation follows the valid-solution criterion defined by the question rather than imposing an arbitrarily selected visual answer. For example, if multiple routes satisfy the same optimality criterion, equivalent valid optimal routes can be accepted. Conversely, if a candidate question cannot yield a stable and verifiable answer under the predefined rules, it is excluded rather than retained by modifying the scene or relaxing the task definition.

C.4 Question Validation

Each question undergoes task-specific validity checks before entering the final evaluation set.

During generation, we first verify that every referenced floor, room, opening, and object is actually present in the drawings supplied to the model and that the spatial relation required to answer the question can be established unambiguously in the current building. Questions involving cross-floor projection, cross-view correspondence, nearest-object reasoning, or navigation undergo additional checks required by their corresponding geometric or topological definitions. If a building cannot support a particular question, an instance is not forced merely to produce the same number of questions for every building.

After generation, the reference answer is recomputed from the structured building records and checked for consistency with the instantiated question and its visual entities. A valid QA instance must therefore preserve the complete correspondence source building representation $\rightarrow$ rendered drawings $\rightarrow$ instantiated question $\rightarrow$ executable answer. Internal conditions used for candidate se-

lection determine whether a candidate can enter the dataset but are not exposed as additional hints in the model-facing question.

Metric questions are also solvable from the supplied drawings alone. The dimension chains and graphic scale on each drawing are generated from the same underlying building geometry used to compute the reference answers, so distances, room dimensions, and object footprint sizes can be recovered from the drawings by reading the dimension annotations or measuring against the graphic scale. The 5% relative tolerance for distance and area answers (Appendix E) absorbs the rounding of displayed dimensions and small reading errors.

We further distinguish QA validity from local drawing quality. Label crowding or local object overlap may be recorded as a drawing-quality warning as long as the identities, geometry, and ground truth of the corresponding entities remain valid; entities are not removed merely to simplify a question. As specified in Section B, objects already admitted to the visible-entity set are not silently removed because of crowded annotations, footprint overlap, or question difficulty. In contrast, an instance that cannot preserve consistency among the source building data, final drawings, question text, and reference answer is excluded from the final valid set.

Following this generation and validation procedure, the main experiments use a fixed set of 482 QA instances. Not every building supports all 32 question templates, so the final set is not constructed by mechanically instantiating every template for every building. All evaluated models are tested on the same questions, drawing evidence, and executable ground truth, providing a common evaluation basis for the model comparisons reported in Section E.

C.5 Worked Example

Figure 6 shows a two-floor drawing pair from a residential duplex townhouse, and Table 8 lists four questions instantiated on it. The questions cover two single-floor and two cross-floor templates from four different categories. In the model-facing prompt, each question is followed by an answer-format instruction (omitted here), and every reference answer is computed from the shared building representation described in Appendix B.

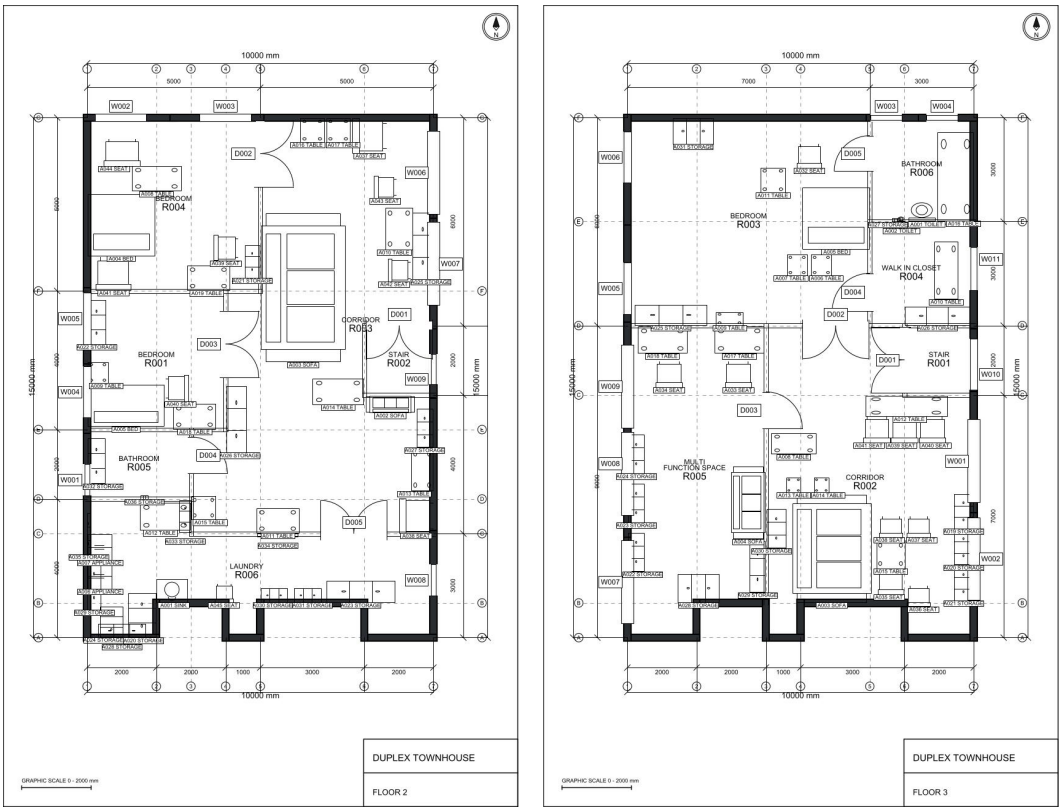


Figure 6: Example two-floor drawing pair from a residential duplex townhouse, showing Floor 2 (left) and Floor 3 (right). Rooms (R###), doors (D###), windows (W###), and visible objects (A###) are labeled with the entity identifiers referenced in the questions. The stairs (Floor 2 R002 and Floor 3 R001) occupy the same footprint and connect the two floors.

Table 8: Example questions and reference answers for the drawing pair in Figure 6. IDs refer to Table 7.

ID	Scope	Question	Answer
D8	Single-floor	On Floor 3, in which cardinal direction is Room R001 located relative to Room R004?	South
C2	Single-floor	On Floor 3, how many labeled rooms have no door directly to the outdoors and exactly one interior door leading to another room?	4
M4	Cross-floor	Which room on Floor 2 corresponds directly below Room R002 on Floor 3?	R003
N6	Cross-floor	What is the shortest route from Room R006 on Floor 2 to Room R005 on Floor 3? List the rooms in order.	2F:R006>2F:R003> 2F:R002>3F:R001> 3F:R002>3F:R005

D MULTIFLOORQA-HARD

This section documents MultiFloorQA-Hard. It is generated by the same pipeline as MultiFloorQA-Main: the shared building representation of Section B, the same drawing conventions and staged quality control, the template-driven question generation and validation of Section C, and the answer-type-specific scoring described in Section E. MultiFloorQA-Hard differs from MultiFloorQA-Main

only in building selection, template selection, and the difficulty parameters of the instantiated questions.

D.1 Building Selection and Construction

The nine hard buildings are drawn from source families disjoint from the 16 MultiFloorQA-Main buildings; an offline audit confirms zero overlap in buildings, question ids, plan images, question text, and normalized question skeletons. Eight buildings have ten floors and one has eight; six are corporate offices and three are residential apartments. Buildings with repeated identical floor plans are excluded, so a tall instance cannot be solved by recognizing one layout. Table 9 gives the per-building composition.

Table 9: Per-building construction statistics for MultiFloorQA-Hard. Eight buildings have ten floors and B09 has eight. Each building has one task-specific access-point (AP) marked plan in addition to its standard per-floor plans; the nine AP variants cover nine distinct floors. Every building contributes exactly one instance of each of the 14 templates (14 questions per building, 126 in total). The nine source families are disjoint from the 16 MultiFloorQA-Main buildings.

ID	Type	Floors	Standard plans	AP plans	Questions
B01	Corporate office	10	10	1	14
B02	Corporate office	10	10	1	14
B03	Corporate office	10	10	1	14
B04	Residential	10	10	1	14
B05	Residential	10	10	1	14
B06	Corporate office	10	10	1	14
B07	Corporate office	10	10	1	14
B08	Residential	10	10	1	14
B09	Corporate office	8	8	1	14
Total (9 buildings)			**88**	**9**	**126**

Each ten-floor building contributes ten standard floor plans and B09 contributes eight, plus exactly one task-specific access-point (AP) marked plan variant per building; the nine AP variants mark nine distinct floors. No elevations are generated for MultiFloorQA-Hard, so MultiFloorQA-Hard contains no Cross-projection questions. In total it comprises 97 input images covering 88 floor identities.

D.2 Template Selection and Difficulty Parameters

MultiFloorQA-Hard reuses 14 of the 32 templates of Table 7, one instantiation per template per building ($14 \times 9 = 126$ questions). The complete Hard template set is **C4, C5, G1–G4, D4, D6, D9, and N3–N7**; these identifiers refer to the full definitions in Appendix C, and Hard keeps these definitions while expanding selected templates to whole-building scope. Recognition and Matching templates are excluded because they are already near ceiling for the strongest models on MultiFloorQA-Main; the retained templates span four categories: Counting C4–C5 (18 questions), Distance D4, D6, D9 (27), Geometry G1–G4 (36), and Navigation N3–N7 (45). Every question supplies an average of 4.7 images (maximum 10), giving 54 single-floor questions (one plan) and 72 cross-floor questions (multiple plans).

Difficulty is raised through instantiation scope and, for C5, a stricter all-floor answer requirement:

- **Whole-building aggregation.** G3 sums labelled areas and G4 compares geometry across every floor of the building rather than two floors, and the cross-floor coverage template D9 is instantiated with coverage over all floors.
- **Full-height projection.** C5 projects the top-floor room set onto every lower floor; an answer is counted correct only when the projected room name and count match on *every* projected floor of that question.

- **Building-scale routes.** N5 requests routes with up to four simultaneously blocked doors, and N6/N7 construct 1-to-10-floor cross-floor routes whose validated sequences span the full building.

The questions and their executable gold answers are frozen for evaluation, including 70 instances inherited verbatim, with identical inputs and gold answers, from the same generator contracts. All 126 questions were completed by every reported model.

D.3 Dataset Statistics

MultiFloorQA-Hard contains 126 question–answer pairs from nine buildings and 97 input images covering 88 floor identities. It includes 18 Counting, 27 Distance, 36 Geometry, and 45 Navigation questions, with 54 single-floor and 72 cross-floor instances; no Cross-projection questions are included. Table 6 reports the overall Main, Hard, and combined benchmark counts, while Table 10 gives the detailed Hard results.

D.4 Full Template-Level Results

General answer normalization and answer-type-specific scoring follow Appendix E. Hard-specific rules are that C5 receives credit only when all projected floors match the gold answer, while N4/N6 routes must satisfy the route contract for connectivity, door semantics, and optimality; the same single follow-up is permitted after an explicitly truncated response. Table 10 reports final post-follow-up accuracy by template, category, and evidence scope for the five evaluated models, using the template identifiers of Table 7. Panel A gives the complete template-level results.

Table 10: MultiFloorQA-Hard accuracy (%) by template, category, evidence scope, and building for the five evaluated models. Scores are final post-follow-up rates under the same format-normalized scoring as the MultiFloorQA-Main; each template row has nine questions, each building row 14, and the scope rows have 54 (single-floor) and 72 (cross-floor) questions. Short template identifiers follow Table 7.

	Astra	Sol	Gemini	Qwen Flash	DeepSeek
A. Accuracy by template and category					
Counting					
C4	66.7	0.0	0.0	11.1	0.0
C5	44.4	0.0	0.0	0.0	0.0
Counting total	55.6	0.0	0.0	5.6	0.0
Distance					
D4	66.7	66.7	22.2	22.2	0.0
D6	88.9	33.3	0.0	11.1	0.0
D9	0.0	33.3	11.1	22.2	22.2
Distance total	51.9	44.4	11.1	18.5	7.4
Geometry					
G1	11.1	0.0	22.2	0.0	0.0
G2	55.6	0.0	22.2	11.1	0.0
G3	100.0	100.0	55.6	33.3	22.2
G4	88.9	88.9	77.8	77.8	55.6
Geometry total	63.9	47.2	44.4	30.6	19.4
Navigation					
N3	100.0	33.3	0.0	22.2	0.0
N4	88.9	55.6	55.6	33.3	33.3
N5	88.9	77.8	66.7	33.3	44.4
N6	0.0	11.1	0.0	22.2	0.0
N7	88.9	11.1	33.3	33.3	33.3

Table 10 *(continued)*

	Astra	Sol	Gemini	Qwen Flash	DeepSeek
Navigation total	73.3	37.8	31.1	28.9	22.2
Overall (126)	63.5	36.5	26.2	23.8	15.1
Single-floor (54)	68.5	31.5	20.4	16.7	5.6
Cross-floor (72)	59.7	40.3	30.6	29.2	22.2
B. Accuracy by building (14 questions each)					
B01	57.1	14.3	14.3	21.4	7.1
B02	42.9	35.7	28.6	14.3	14.3
B03	50.0	50.0	28.6	28.6	7.1
B04	78.6	50.0	50.0	28.6	21.4
B05	78.6	57.1	28.6	14.3	21.4
B06	71.4	35.7	35.7	50.0	21.4
B07	71.4	35.7	28.6	21.4	21.4
B08	71.4	28.6	14.3	14.3	7.1
B09	50.0	21.4	7.1	21.4	14.3

D.5 Per-Building Results

Panel B of Table 10 reports the corresponding final accuracy for each of the nine Hard buildings. These rows expose variation across building instances while preserving the same template-level scoring and denominator conventions. The table also supports the analyses in Section 4.3 of the main paper. The strict C5 projection rule is strongly binding. Only 4 of the 45 final C5 responses (8.9%) are correct as whole questions, and all four come from GPT-6 Astra. No other model answers any C5 question correctly, so partial coverage never earns credit.

Figure 7 compares overall accuracy for the five models evaluated on MultiFloorQA-Hard. Main and Hard contain different buildings and template mixtures, so the comparison is descriptive.

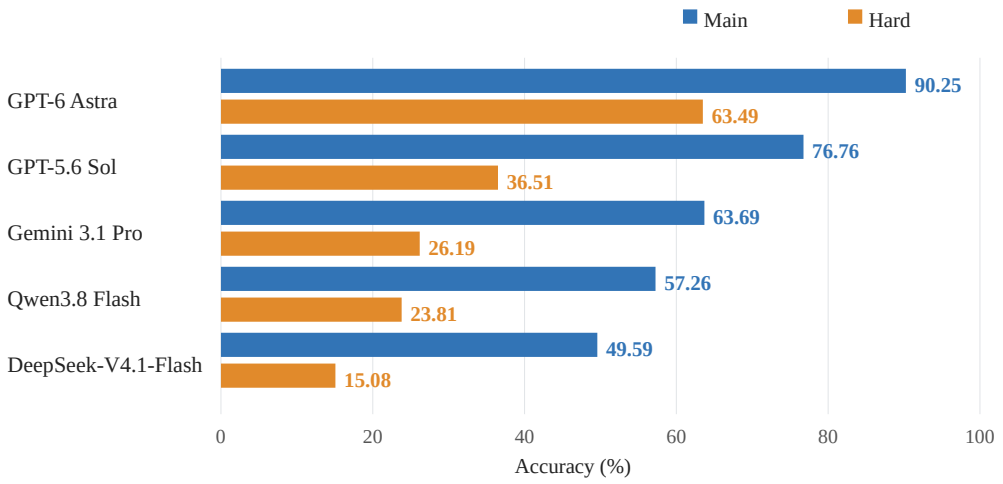


Figure 7: Overall accuracy (%) on MultiFloorQA-Main (482 questions) and MultiFloorQA-Hard (126 questions) for the five models evaluated on MultiFloorQA-Hard.

E Evaluation Protocol and Full Main Results

We evaluate ten multimodal models on all 482 questions of MultiFloorQA-Main, drawn from 16 buildings. The questions cover six categories—Recognition, Counting, Distance, Geometry, Matching, and Navigation—and span Single-floor, Cross-floor, and Cross-projection evidence scopes. All

models use the same questions and reference answers. We report accuracy by template, category, floor relation, evidence scope, and over the full evaluation set for all ten models.

E.1 Models and Inference Settings

We evaluate ten MLLMs in the main experiment. Table 11 lists the request model identifiers, access endpoints, and reasoning-effort settings used in our experiments. All evaluations were conducted between September 19 and September 22, 2026.

All experiments use a maximum output length of 32,768 tokens. Temperature is not explicitly set, and provider defaults are used.

E.2 Output Parsing and Scoring

Model responses are parsed according to the answer format specified by each question and normalized before scoring. Categorical, Boolean, count, entity, and set answers use exact matching after normalization, with set order ignored where appropriate. Distance and area answers allow a 5% relative error, while navigation answers are checked for validity and optimality according to the task requirements.

Cross-floor room identifiers use the form `xF:Rxxx`, such as `1F:R003`, and ordered route answers use > to separate successive rooms.

Transient service failures are retried at most once. If a response reaches the output limit, one additional prompt requests the final answer based on the existing response. The answer returned after this follow-up is used for scoring. Empty, unparseable, unresolved, or otherwise incorrect responses are counted as incorrect.

E.3 Full MultiFloorQA-Main Results

Table 12: Evaluation settings and results. Panels B–E use final answers after the permitted follow-up. Panel F reports explicitly recorded initial token-limit stops. Template, floor-relation, and evidence-scope scores use the same per-question scoring records as the overall and category results. Percentages unless indicated otherwise.

A. Evaluation settings	
QA set	MultiFloorQA-Main: 482 questions; 32 templates; 16 buildings; 299 one-floor and 183 multiple-floor questions (floor relation); 287 Single-floor, 171 Cross-floor, and 24 Cross-projection questions (evidence scope).
Input	Zero-shot; identical question set and reference answers; one, two, or three drawing images (floor plans and, where required, an east elevation).

B. Overall accuracy on all questions

Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
Correct / 482	435	370	350	335	307	358	301	276	239	169
Overall (%)	90.2	76.8	72.6	69.5	63.7	74.3	62.4	57.3	49.6	35.1

Continued on the next page

Table 11: Model identifiers, access endpoints, and inference-effort settings used in the evaluation.

Model	Request model ID	Endpoint	Effort
GPT-6 Astra	`openai/gpt-6-astra`	OpenRouter	max
GPT-5.6 Sol	`openai/gpt-5.6-sol`	OpenRouter	max
Gemini 3.1 Pro Preview	`google/gemini-3.1-pro-preview`	OpenRouter	high
Claude Fable 5.1	`anthropic/claude-fable-5.1`	OpenRouter	max
Qwen3.8 Max 0902	`qwen/qwen3.8-max-0902`	OpenRouter	xhigh
Qwen3.8 Flash	`qwen/qwen3.8-flash`	OpenRouter	xhigh
Kimi K3	`kimi-k3`	Moonshot API	max
DeepSeek-V4.1-Flash	`deepseek-flash`	DeepSeek API	max
GLM-5.3-Flash	`glm-5.3-flash`	Z.ai API	max
MiniMax M3	`MiniMax-M3`	MiniMax API	adaptive

Table 12 *(continued)*

Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
C. Accuracy by question template (%)										
Recognition										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
R1	100.0	100.0	100.0	85.7	100.0	100.0	78.6	85.7	50.0	35.7
R2	100.0	100.0	83.3	91.7	100.0	100.0	58.3	83.3	58.3	50.0
R3	100.0	81.3	100.0	100.0	87.5	93.8	87.5	56.3	75.0	43.8
Recognition	100.0	92.9	95.2	92.9	95.2	97.6	76.2	73.8	61.9	42.9
Counting										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
C1	93.8	93.8	87.5	81.3	75.0	87.5	75.0	56.3	62.5	31.3
C2	100.0	60.0	53.3	53.3	53.3	60.0	80.0	33.3	53.3	40.0
C3	100.0	93.8	100.0	100.0	87.5	100.0	93.8	93.8	75.0	56.3
C4	78.6	50.0	14.3	14.3	35.7	42.9	28.6	35.7	14.3	21.4
C5	93.8	62.5	62.5	62.5	68.8	75.0	50.0	37.5	37.5	12.5
Counting	93.5	72.7	64.9	63.6	64.9	74.0	66.2	51.9	49.4	32.5
Distance										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
D1	100.0	87.5	68.8	68.8	18.8	50.0	50.0	37.5	18.8	12.5
D2	100.0	87.5	87.5	68.8	43.8	56.3	62.5	50.0	31.3	18.8
D3	100.0	75.0	83.3	58.3	58.3	100.0	75.0	66.7	33.3	33.3
D4	87.5	75.0	62.5	56.3	50.0	100.0	43.8	50.0	6.3	31.3
D5	100.0	93.8	68.8	87.5	62.5	100.0	56.3	62.5	12.5	43.8
D6	43.8	25.0	12.5	12.5	0.0	25.0	6.3	0.0	0.0	0.0
D7	93.8	81.3	56.3	50.0	43.8	75.0	43.8	56.3	31.3	25.0
D8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	81.3
D9	100.0	76.9	84.6	53.8	46.2	76.9	61.5	46.2	23.1	23.1
Distance	91.2	78.1	68.6	62.0	46.7	75.2	54.7	51.8	28.5	29.9
Geometry										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
G1	40.0	6.7	13.3	13.3	13.3	20.0	6.7	13.3	13.3	6.7
G2	46.7	6.7	33.3	13.3	20.0	6.7	6.7	20.0	20.0	6.7
G3	100.0	100.0	68.8	56.3	56.3	56.3	25.0	37.5	56.3	6.3
G4	87.5	87.5	87.5	87.5	87.5	87.5	87.5	81.3	81.3	87.5
Geometry	69.4	51.6	51.6	43.5	45.2	43.5	32.3	38.7	43.5	27.4
Matching										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
M1	100.0	100.0	100.0	83.3	100.0	100.0	75.0	75.0	50.0	41.7
M2	100.0	100.0	100.0	100.0	83.3	100.0	91.7	91.7	50.0	8.3
M3	100.0	100.0	100.0	100.0	100.0	87.5	100.0	87.5	100.0	87.5
M4	100.0	100.0	100.0	87.5	93.8	93.8	93.8	81.3	87.5	37.5
Matching	100.0	100.0	100.0	92.9	94.6	94.6	91.1	83.9	75.0	46.4
Navigation										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
N1	100.0	93.3	100.0	100.0	86.7	100.0	93.3	80.0	93.3	66.7
N2	100.0	100.0	100.0	100.0	93.3	93.3	100.0	80.0	86.7	73.3
N3	80.0	73.3	60.0	46.7	13.3	33.3	33.3	20.0	26.7	6.7
N4	73.3	60.0	60.0	60.0	73.3	66.7	53.3	53.3	46.7	33.3
N5	100.0	81.3	87.5	81.3	81.3	93.8	87.5	81.3	75.0	43.8
N6	100.0	81.3	81.3	93.8	87.5	75.0	81.3	75.0	87.5	37.5
N7	75.0	31.3	18.8	56.3	31.3	37.5	18.8	18.8	18.8	12.5
Navigation	89.8	74.1	72.2	76.9	66.7	71.3	66.7	58.3	62.0	38.9
D. Accuracy by floor relation (%)										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
Single-floor	87.6	74.6	70.9	66.2	58.5	72.6	58.9	53.2	43.1	33.8
Cross-floor	94.5	80.3	75.4	74.9	72.1	77.0	68.3	63.9	60.1	37.2

Continued on the next page

Table 12 *(continued)*

Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
E. Accuracy by evidence scope (%)										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
Single-floor	87.1	73.5	69.7	65.5	56.8	71.4	58.2	52.3	42.9	33.4
Cross-floor	94.2	78.9	73.7	73.1	71.3	75.4	66.7	62.0	60.8	39.2
Cross-projection	100.0	100.0	100.0	91.7	91.7	100.0	83.3	83.3	50.0	25.0
F. Recorded initial token-limit stop reason (%)										
Item	Astra	Sol	Q-Max	Kimi	Gemini	Claude	GLM	Q-Flash	DeepSeek	MiniMax
Recognition	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Counting	0.0	0.0	0.0	0.0	0.0	14.3	20.8	0.0	7.8	0.0
Distance	0.0	1.5	0.0	2.2	0.7	9.5	19.0	0.7	16.1	1.5
Geometry	8.1	21.0	0.0	4.8	3.2	29.0	29.0	0.0	35.5	1.6
Matching	0.0	0.0	0.0	0.0	0.0	3.6	1.8	0.0	0.0	0.0
Navigation	0.9	10.2	0.0	3.7	1.9	13.0	12.0	1.9	15.7	6.5
Overall	1.2	5.4	0.0	2.1	1.0	12.0	15.4	0.6	13.9	2.1

Overall performance. In the Overall row of Table 12, GPT-6 Astra achieves the highest accuracy at 90.2% (435/482), followed by GPT-5.6 Sol at 76.8% (370/482), Claude Fable 5.1 at 74.3% (358/482), and Qwen3.8 Max at 72.6% (350/482). Overall gives equal weight to every question. The category sizes range from 42 Recognition questions to 137 Distance questions. The template rows show how performance varies within each category.

Recognition. Recognition approaches ceiling performance for several models. Astra answers every R1–R3 question correctly, while Sol answers every R1 and R2 question correctly but scores 81.3% on R3, which requires identifying rooms directly reachable from a specified room. At the category level, six models exceed 90%, indicating that the basic recognition tasks are comparatively well resolved. Errors cluster in specific templates and models.

Counting. Counting performance varies substantially by the quantity being requested. On C3, which asks for the total number of doors across two floors, Sol scores 93.8% and Qwen3.8 Max scores 100.0%. On C4, which compares object counts in two specified rooms, their accuracies fall to 50.0% and 14.3%, respectively. C4 therefore remains difficult even for models that perform well on cross-floor door aggregation; the Counting total alone would obscure this template-level difference.

Distance. Distance contains both high-accuracy directional judgments and difficult metric calculations. Nine models answer all D8 cardinal-direction questions correctly. On D6, which asks how far an object can move before its first contact with an obstacle, the highest accuracy is only 43.8%, achieved by Astra, and four models score 0.0%. The contrast between D8 and D6 separates directional judgments from clearance calculations.

Geometry. Net-area calculation is the principal difficulty within Geometry. G1 and G2 require subtracting visible-object footprints from room areas. Astra scores 40.0% and 46.7% on these templates, while Sol scores 6.7% on each. On G3, which sums labelled room areas across two floors, both models achieve 100.0%. This gap shows that a model's ability to read and add labelled areas does not reliably transfer to computing usable area from object geometry.

Matching. Matching is the strongest category for the leading models: Astra, Sol, and Qwen3.8 Max answer all 56 questions correctly. Their perfect scores cover the plan–elevation tasks M1 and M2 and the cross-floor plan tasks M3 and M4. The category nonetheless distinguishes models further down the ranking; for example, DeepSeek and MiniMax achieve 75.0% and 46.4% overall, respectively, with markedly different scores across individual matching templates.

Navigation. Navigation accuracy depends on what the answer must contain. Astra, Sol, and Qwen3.8 Max all score 100.0% on N2, which asks for a room sequence minimizing door connections. Their accuracies on the single-floor walkable-distance task N3 are 80.0%, 73.3%, and 60.0%,

respectively; on the cross-floor distance task N7, they are 75.0%, 31.3%, and 18.8%. These results locate the main difficulty in calculating route length under geometric and floor-height constraints, beyond identifying a valid sequence of rooms.

Floor relation and evidence scope. Panel D groups the 299 single-floor and 183 cross-floor MultiFloorQA-Main questions by floor relation. Panel E applies the main paper's evidence scopes to the same 482 questions: 287 Single-floor, 171 Cross-floor, and 24 Cross-projection. M1 and M2 account for the Cross-projection group; Panel D places M1 with one-floor questions and M2 with multiple-floor questions. The groups have different task mixtures, so their accuracies do not isolate the effect of evidence scope.

Figures 8 and 9 display the six category accuracies in Table 12 on common 0–100% scales. The first pair divides the ten models into the two groups shown in the figure; the second pair presents selected model comparisons.

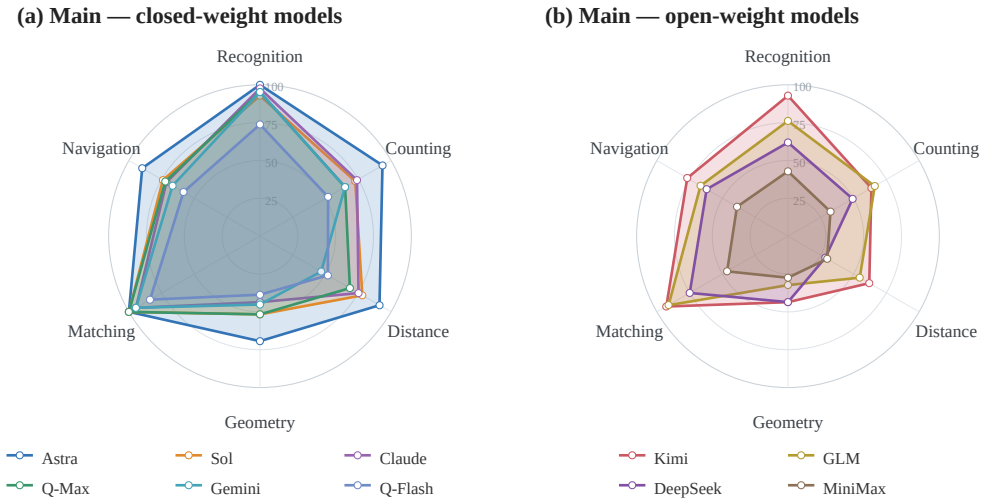


Figure 8: MultiFloorQA-Main accuracy by task category for the two model groups shown in panels (a) and (b). Each radar axis ranges from 0 to 100%.

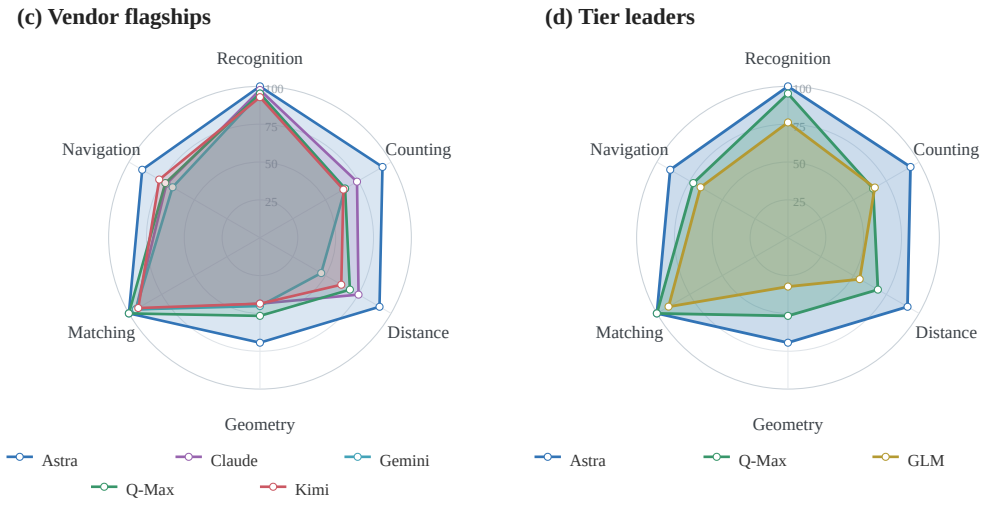


Figure 9: Selected MultiFloorQA-Main model comparisons: five models in panel (c) and three selected models in panel (d). Category axes and scales match Figure 8.

The last two panels of the radar series show category and evidence-scope accuracy across the ten Main models. Figure 10 presents both at the same scale.

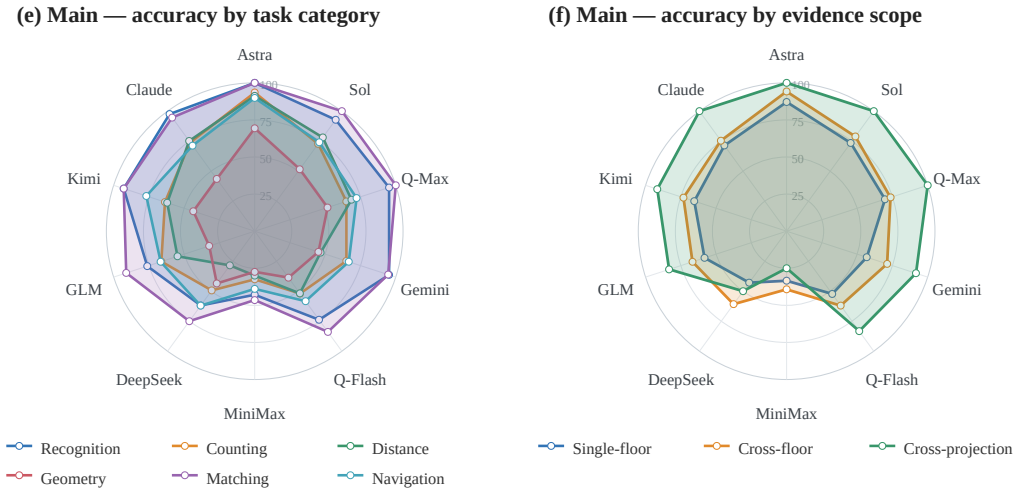


Figure 10: MultiFloorQA-Main accuracy across the ten evaluated models. Panel (e) groups questions by task category, and panel (f) groups them by evidence scope. Spokes identify models; all scores are percentages.

F Diagnostic Experiments

The experiments in this section examine how the information available to a model and the way it is presented affect performance on MultiFloorQA-Main. We use a fixed diagnostic subset of 64 questions from three buildings (21 educational, 21 residential, and 22 office questions), with two instances of each of the 32 templates. The subset is drawn from the 482-question MultiFloorQA-Main evaluation; for each model, its original-drawing result on these same 64 questions provides the reference. GPT-5.6 Sol, GLM-5.3-Flash, and DeepSeek-V4.1-Flash use the same inference settings

as in Appendix E. All scores use the original answers and scoring rules, and failed or missing final answers count as incorrect. As in Section E, an initially truncated answer can receive one follow-up; we report the resulting final score. Table 13 records the inputs, sample, and scoring protocol.

F.1 Input information: drawings, question only, and structural JSON

Table 13: Protocol for the 64-question diagnostic subset. Original-drawing scores are extracted for the same question IDs from the main evaluation. All variants use the same reference answers and scoring rules.

Setting	Specification
Questions	64 questions from three buildings: education 21, residential 21, office 22; two instances per template across 32 templates.
Models	GPT-5.6 Sol; GLM-5.3-Flash; DeepSeek-V4.1-Flash.
Original drawings	Main-evaluation input on the matching 64 question IDs; same question, image evidence, and answer format.
Question only	Question and answer format; no drawing or building data. The prompt requests a best-guess final answer.
Structural JSON	Question and answer format plus public drawing-page geometry and explicit relations; no drawing image, target answer, or private solution program.
Red marks	Original drawing dimensions; specified labels and applicable symbols colored red. For 10 questions without a markable target, the original-drawing result is reused; 54 questions have new red-mark requests.
One shot	One example per template, with example drawings, question, format, and answer, followed by the target input. Example buildings are disjoint from the 25 evaluation buildings (MultiFloorQA-Main and MultiFloorQA-Hard).
Scoring	Final correct answers out of 64, after at most one permitted truncation follow-up. Missing or failed final answers count as incorrect.

We first compare the original drawings with two changes to the information supplied to the model. The question-only condition removes both drawings and building data and instructs the model to give its best answer from the question. The structural-JSON condition replaces the drawings with a public representation of the supplied drawing pages, including relevant geometry and explicit relations; it excludes the target answer and private solution program. Table 14 reports all three conditions on identical questions. With drawings, Sol, GLM, and DeepSeek answer 51, 40, and 29 questions correctly after any permitted follow-up. Without building evidence, the corresponding scores fall to 15, 14, and 16; thus, the question text alone does not reproduce the original results. With structural JSON, they rise to 64, 63, and 63. This large improvement shows that these models can often perform the requested operations when the scene structure is explicit. The JSON exposes coordinates and relations that the drawing condition requires the model to recover visually. Two JSON responses remain incorrect.

Table 14: Input information on the same 64 questions. Cells show correct answers out of 64, with accuracy in parentheses. Final scores include the permitted follow-up after an initially truncated response.

Model	Original drawings	Question only	Structural JSON
GPT-5.6 Sol	51/64 (79.7%)	15/64 (23.4%)	64/64 (100.0%)
GLM-5.3-Flash	40/64 (62.5%)	14/64 (21.9%)	63/64 (98.4%)
DeepSeek-V4.1-Flash	29/64 (45.3%)	16/64 (25.0%)	63/64 (98.4%)

F.2 Visual localization with red marks

The red-mark condition retains the original drawing dimensions and changes the color of the specified labels and applicable symbols. Where no target can be marked, the input is unchanged. We use the original-drawing results for these 10 unchanged questions and make new red-mark requests for the other 54, while retaining all 64 in the denominator. Table 15 compares this condition with the

corresponding original-drawing scores. Sol changes from 51/64 to 50/64, GLM changes from 40/64 to 43/64, and DeepSeek increases from 29/64 to 34/64. The gains are therefore model-dependent, and a highlighted target does not by itself resolve the remaining geometric and relational demands. The conditions were run separately, and 10 unchanged questions reuse their baseline results.

Table 15: Effect of red visual marks on the 64-question subset. Ten unmarkable questions reuse the original-drawing results; 54 marked questions are newly evaluated. Differences are percentage points relative to the original-drawing result on the same question IDs.

Model	Original drawings	Red marks	Difference (pp)
GPT-5.6 Sol	51/64 (79.7%)	50/64 (78.1%)	−1.6
GLM-5.3-Flash	40/64 (62.5%)	43/64 (67.2%)	+4.7
DeepSeek-V4.1-Flash	29/64 (45.3%)	34/64 (53.1%)	+7.8

F.3 One-shot demonstration

In the one-shot condition, each template receives one worked input-output example before the target question. The example includes its own drawing pages, question, answer format, and reference answer, without a reasoning trace. Examples come from buildings outside the 25 evaluation buildings, avoiding example-building overlap with test buildings in either MultiFloorQA-Main or MultiFloorQA-Hard. Table 16 shows that Sol changes from 51/64 to 49/64, GLM from 40/64 to 38/64, and DeepSeek from 29/64 to 28/64. A single example therefore offers no uniform benefit across these models. It also increases the input context, so this comparison combines the effect of a demonstration with the extra visual and textual material.

Table 16: One-shot prompting on the 64-question subset. Each target is preceded by one example of the same template from outside the evaluation buildings. Differences are percentage points relative to the original-drawing result on the same question IDs.

Model	Original drawings	One shot	Difference (pp)
GPT-5.6 Sol	51/64 (79.7%)	49/64 (76.6%)	−3.1
GLM-5.3-Flash	40/64 (62.5%)	38/64 (59.4%)	−3.1
DeepSeek-V4.1-Flash	29/64 (45.3%)	28/64 (43.8%)	−1.6

Taken together, the strongest diagnostic change is the replacement of drawings with explicit scene structure. Removing scene evidence substantially lowers performance, whereas localization marks and demonstrations produce smaller, inconsistent differences across models.

G Examples of Incorrect Final Answers

The main evaluation scores final answers after the permitted follow-up. Three Qwen3.8 Max errors illustrate different answer types. For a net usable area question (template G2), the model reports 89.875 m^2 instead of 99.646 m^2 (a 9.8% error); the response alone does not distinguish a misread dimension from an incorrect calculation.

For a set question (template D7), the model returns `A039` instead of `A008|A017|A020`, omitting the required objects and adding another. For an evacuation route (template N4), it includes an extra room: `R007>R004>R005` instead of `R007>R004`. Route validity is assessed against the benchmark's connectivity rules. Table 17 lists these differences.

Table 17: Examples of observable discrepancies in Qwen3.8 Max final answers. These cases illustrate different answer types; they are not estimates of error-family frequency.

Type	Template	Predicted answer	Reference answer
Area	G2	89.875 m^2	99.646 m^2
Set	D7	`A039`	`A008|A017|A020`
Route	N4	`R007>R004>R005`	`R007>R004`

H Representative Prompts

Each evaluation request pairs a question with the drawing pages needed to answer it. The question specifies the target entities or rooms, while the accompanying text defines any task-specific convention and the required form of the final answer. For multi-image questions, the relevant floor plans and, where needed, the elevation are supplied as separate images in a fixed order. The model is asked for a concise final answer rather than a written reasoning trace.

Prompt 1: Single-floor distance

Input image:
Floor 4 plan.

Question:
On Floor 4, what is the Manhattan distance in meters between the centers of rooms R005 and R003?

Task definition:
Define a room center as the area centroid of its region.

Output format:
Return only a number in meters, with at most three decimal places. Do not explain.

Prompt 2: Cross-floor route

Input images:
Floor 3 plan; Floor 4 plan (in this order).

Question:
Which rooms are traversed, in order, along a shortest route from room R001 on Floor 3 to room R007 on Floor 4?

Task definition:
Minimize the number of door connections. A double door counts as one connection; an open passage without a door does not count. The route is restricted to the supplied floors.

Output format:
Return any valid optimal room sequence, including start and destination. Prefix each room with its floor and separate rooms with > (e.g., 1F:R003>2F:R001). Do not explain.

Prompt 3: Plan-elevation correspondence

Input images:
Floor 1 plan; Floor 2 plan; east elevation (in this order).

Question:
Which window on which floor plan corresponds to window E-W006 in the building's east elevation?

Output format:
Return only a window identifier with a floor prefix, such as 3F:W005. Do not explain.

Figure 11: Representative prompts for single-floor distance, cross-floor routing, and plan–elevation correspondence. The drawing pages in each panel are listed in the order supplied to the model.

Figure 11 shows a single-floor metric question, a cross-floor route, and a plan–elevation match. The metric prompt defines a room center as its area centroid. The route prompt supplies two floor plans,

counts a double door as one connection, excludes open passages from the door count, and requests a floor-prefixed room sequence. The matching prompt supplies two plans and the east elevation and requests the floor-prefixed plan-window ID.

Metric questions request a number with at most three decimal places, route questions request a room sequence separated by $>$, and the matching question requests a floor-prefixed window ID. Answers are parsed using the corresponding task rules and compared with executable reference answers.

For the one-shot diagnostic, an example from another building precedes the target question. It contains drawing pages, a question, its answer format, and a reference answer without a reasoning trace. The target input remains the same as in the baseline condition.

I LIMITATIONS

MultiFloorQA prioritizes controlled evaluation of spatial reasoning over unrestricted visual realism. The drawings use standardized graphical conventions, a selected set of visible objects, and task-relevant abstractions of building structure. These choices reduce variation from drawing style, texture, decorative content, and other visual factors that are not central to the spatial operations being evaluated. Consequently, the benchmark is not intended to measure robustness to the full diversity of real-world drawing styles or document noise. Such variation is complementary to, rather than part of, the controlled spatial reasoning setting studied here.

MultiFloorQA also focuses on questions with well-defined and reproducible reference answers. This makes it possible to evaluate spatial reasoning under explicit geometric and topological conventions without introducing ambiguity from subjective interpretation or open-ended architectural judgment. As a result, tasks involving design preferences, professional recommendations, or other questions without a unique reference answer fall outside the current evaluation scope and would require different evaluation protocols.