ICML 2026 · SEOUL, KOREA

FloorplanQA: A Benchmark for Spatial Reasoning in LLMs using Structured Representations

Fedor Rodionov[1], Abdelrahman Eldesokey[1], Michael Birsak[1], John Femiani[2], Bernard Ghanem[1], Peter Wonka[1] [1]KAUST · [2]Miami University

How well do LLMs reason about floorplans?

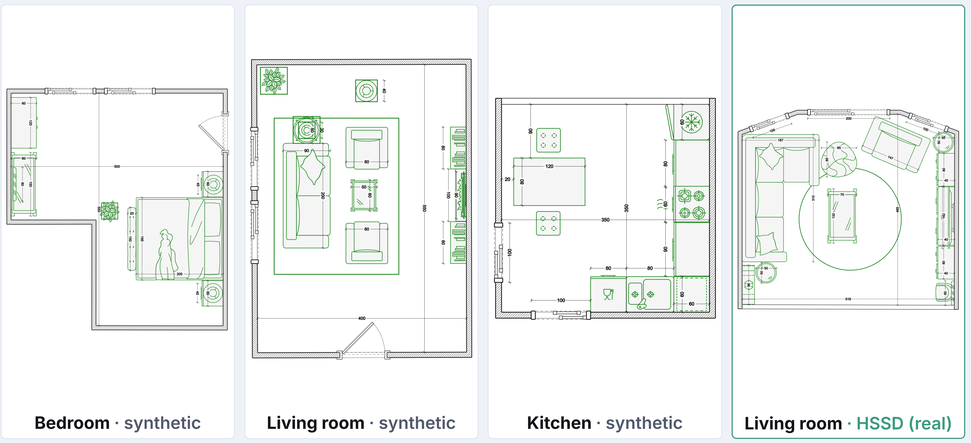
Frontier models handle distance & angle on clean synthetic rooms, yet collapse on free-space, packing and path-finding — and even the basics degrade sharply on real floorplans.

Scan for the paper

1 THE BENCHMARK

Inside FloorplanQA

Indoor scenes in design, robotics and planning are increasingly handled as **structured text**, not pixels. Can LLMs reason *geometrically* over symbolic layouts? **2,000** JSON layouts — **1,800** synthetic (kitchen / living room / bedroom) + **200** real HSSD floorplans — each with **8** questions (**16,000** total), answered from **text alone**: no images, no tools, no code.



Bedroom · synthetic | **Living room** · synthetic | **Kitchen** · synthetic | **Living room** · HSSD (real)

2 THE QUESTIONS

Eight tasks, three modes of reasoning

Metric — NUMERIC COMPUTATION
- **Pair Distance** — centroid-to-centroid
- **View Angle** — angle to reference axis

Dynamic — LAYOUT-CHANGING
- **Repositioning** — slide until contact
- **Shortest Path** — collision-free route

Topology — GEOMETRIC & RELATIONAL
- **Placement** — collision-free feasibility
- **Free Space** — unoccupied floor area
- **Visibility** — line-of-sight blockers
- **Max Box** — largest empty rectangle

Scalars scored by **relative error** (2% tol; 5% for areas); paths by validity & Fréchet distance. **15 LLMs** — 8 standard + 7 reasoning.

How a model sees it

structured input + one question

```
"room": {"polygon":[[0,0],[4,0],[4,6],[0,6]]},
"objects": [
  {"name":"sofa_1",  "bbox":[0.3,4.0,2.8,5.0]},
  {"name":"table_1", "bbox":[1.2,2.0,2.8,3.2]}
],
"place": {"name":"chest", "size":[2.5,1.0]}
```

Placement question

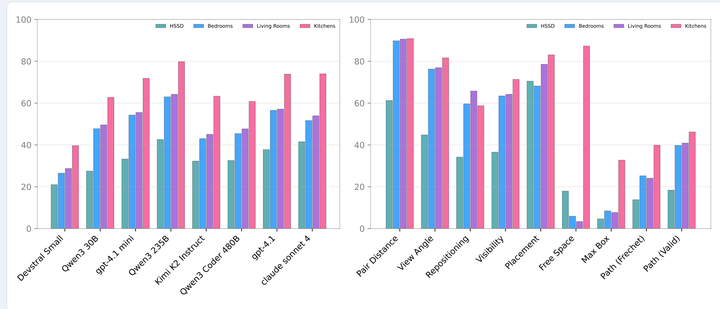
"Can the 2.5 × 1.0 m antique storage chest be placed in the room without overlapping any wall or existing object?"

→ **answer: yes / no**

3 RESULTS

Standard models

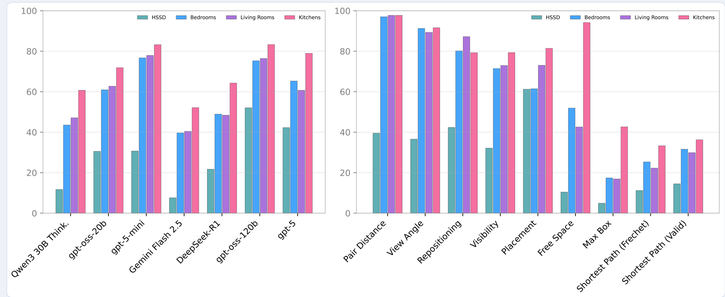
accuracy %, by model & by task, per room type



Strong on Pair Distance / View Angle / Placement / Visibility; weak on **Free Space, Max Box, Shortest Path**. Real **HSSD** layouts (teal) trail synthetic on every task.

Reasoning models

7 models, same axes



Extended inference lifts geometric-union tasks (Free Space, Max Box), but planning gains stay modest.

Accuracy splits into three tiers

mean accuracy · synthetic → real HSSD

89% → 46%	**Metric** — distance & angle
72% → 46%	**Constraint** — placement · visibility · repositioning
35% → 12%	**Planning** — free space · max box · shortest path, *largely unsolved*

Every tier drops sharply on real HSSD layouts. Reasoning models gain +10–40% on geometric unions (Free Space, Max Box). · 15 LLMs · 16,000 questions.

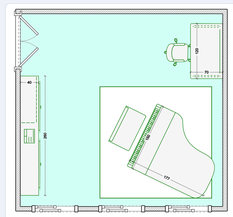
What the results show

Models reliably handle **metric** queries on synthetic rooms, but accuracy roughly **halves on real HSSD** polygons, where centroid / shoelace errors creep in. **Constraint** tasks sit mid-pack, while **optimization & planning** — free space, packing, collision-free paths — stay **largely unsolved**. Reasoning helps most on geometric unions, yet the HSSD gap persists everywhere.

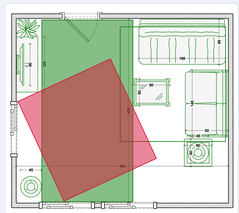
4 WHAT IT REVEALS

Three failure modes & the hardest tasks

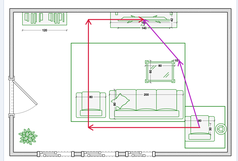
- **Sum, not union** — areas added not unioned → Free Space & Max Box collapse.
- **Centroid errors** — high-vertex HSSD polygons break distance & angle.
- **Planning in clutter** — paths fail where clearance narrows corridors.



Free Space. Union of overlapping objects (mint = unoccupied).



Max Box. Model's rectangle (red) is suboptimal; a larger valid one exists (green).



Shortest Path. Model route (red) collides; optimal in purple.

Tools fix arithmetic, not logic

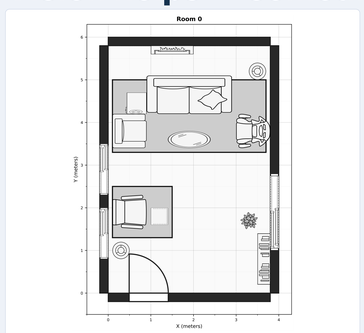
GPT-4.1 on real HSSD · raw vs. + Python interpreter

Metric & geometry ↑
- Pair Distance 56 → 99
- View Angle 55 → 96
- Placement 65 → 95
- Visibility 47 → 87

Logic & planning ↓
- Max Box 7.5 → 3
- Shortest Path 25 → 13
- Repositioning 57 → 48

A code interpreter nails the arithmetic, but **can't supply the missing spatial logic** — constraint & planning tasks get *worse*.

Vision helps — sometimes



Schematic icons — the benchmark rendering.

Photorealistic — AI top-down (Gemini 3.1 Flash Image).

- An image *with* the JSON **helps geometric/path tasks** (Pair Distance +8–10 pp) but **hurts aggregation** (Visibility −15 to −17 pp).
- Image-*only* sits far below any JSON setting — **structured text stays the primary signal**.

CONCLUSION

The bottleneck is **algorithmic reasoning about spatial constraints** — not numerical precision (tools don't fix it) nor computational depth (reasoning models barely help). FloorplanQA isolates and measures this spatial-reasoning blind spot.

FloorplanQA — benchmark, code & data: OldDelorean.github.io/FloorplanQA · fedor.rodionov@kaust.edu.sa

Scan the QR ↑ for the paper