[Code] [Paper] [Lab Homepage]

CVPR JUNE 3-7, 2026 DENVER COLORADO

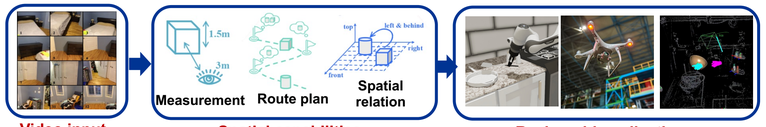
# InfiniBench: Infinite Benchmarking for Visual Spatial Reasoning with Customizable Scene Complexity

Haoming Wang, Qiyao Xue Wei Gao

University of Pittsburgh

## Background and Motivation

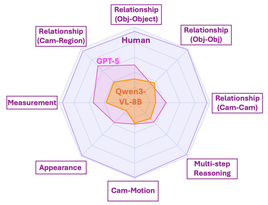
❖ **Background**: Spatial Reasoning for Vision Language Models (VLMs)




Video input — Spatial capabilities — Real-world applications

❖ **Performance of current VLMs:**

❑ **Far Behind humans**




❑ **Struggle in complex scenes**

Simple scenes


Complex scenes

- Need to benchmark VLMs under different levels of scene complexity

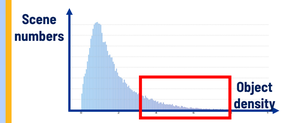
❖ **Customizable Scene complexity in multiple dimensions**


Existing benchmarks: coarse grained definition of scene complexity (e.g. defined by the number of room)

Our goal: Each dimensions of the complexity can be controlled independently

❖ **Existing methods to achieved customizable scene complexity**

❑ **Search in existing datasets**




- Lack of high complexity scenes

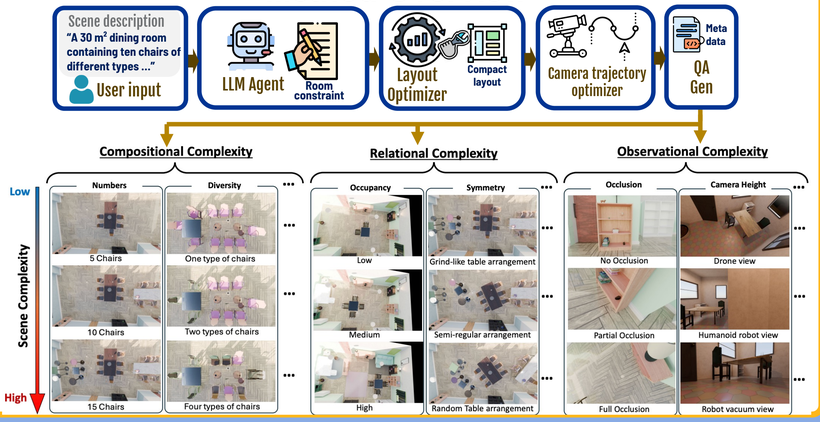
❑ **LLM based layout generation**

- Generated layouts may be physically invalid

❑ **Procedure scene generation**

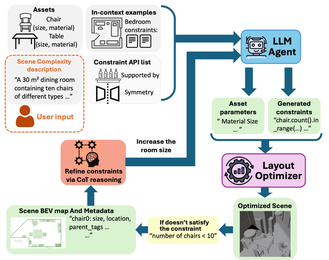
-Difficult to use
-Fail to generate complex scenes

## Method Overview

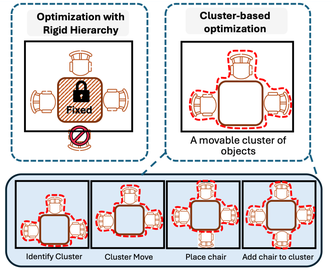



## Method Details

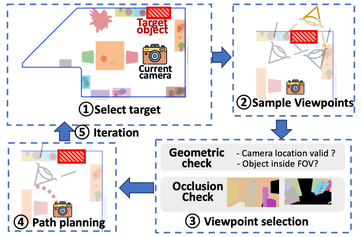
❖ **1 Agentic Generation of Scene Constraints**

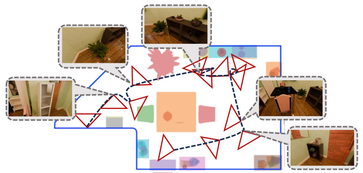



❖ **2 Layout Optimization for Complex Scenes**

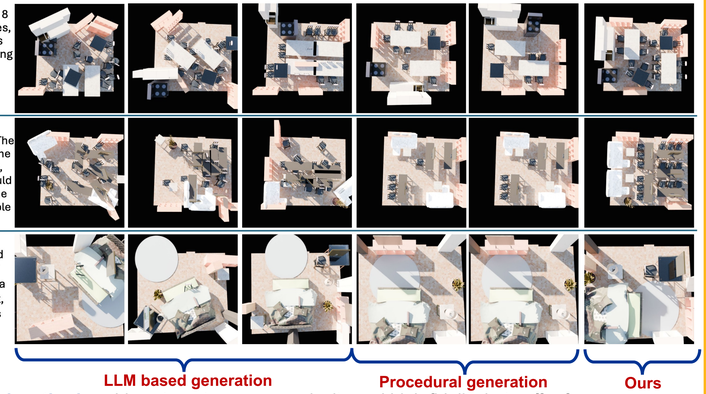



❖ **3 Camera Trajectory Optimization**



## Evaluation of Scene generation

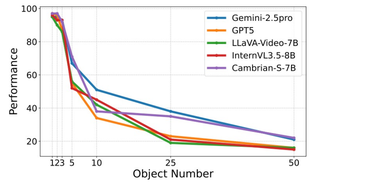



- **LLM-based methods** achieve target scene complexity and high fidelity but suffer from more collisions and physical implausibility as complexity increases.
- **Procedural methods** maintain physical plausibility but struggle to satisfy complexity requirements, leading to lower fidelity in complex scenes

## Using InfiniBench: diagnosing VLM failures

❖ **Object counting under different compositional complexity**

Less objects — More objects




**VLM reasoning performance declines as the number of objects increases**, likely stems from repetitive counting across frames, which often leads to object overestimation.

❖ **Observational complexity (camera height)**


| | | Measure | Perspect | Spatiotemporal |
|---|---|---|---|---|
| Gemini-2.5 | BEV | 68.1 | 77.4 | 73.2 |
| | Ego | 68.9 | 67.2 | 70.1 |
| GPT5 | BEV | 42.0 | 69.1 | 49.0 |
| | Ego | 41.2 | 55.5 | 31.3 |
| LLaVA-Video-7B | BEV | 45.0 | 68.3 | 47.1 |
| | Ego | 45.2 | 56.1 | 36.2 |
| InternVL3.5-8B | BEV | 47.6 | 62.7 | 68.1 |
| | Ego | 48.2 | 52.9 | 64.9 |

**Perspective-taking and spatiotemporal performance improves with higher viewpoints**, likely because bird's-eye views provide a clearer and more complete representation of the scene's spatial layout.