Homepage

CVPR JUNE 3-7, 2026 DENVER COLORADO

Geometrically-Constrained Agent for Spatial Reasoning

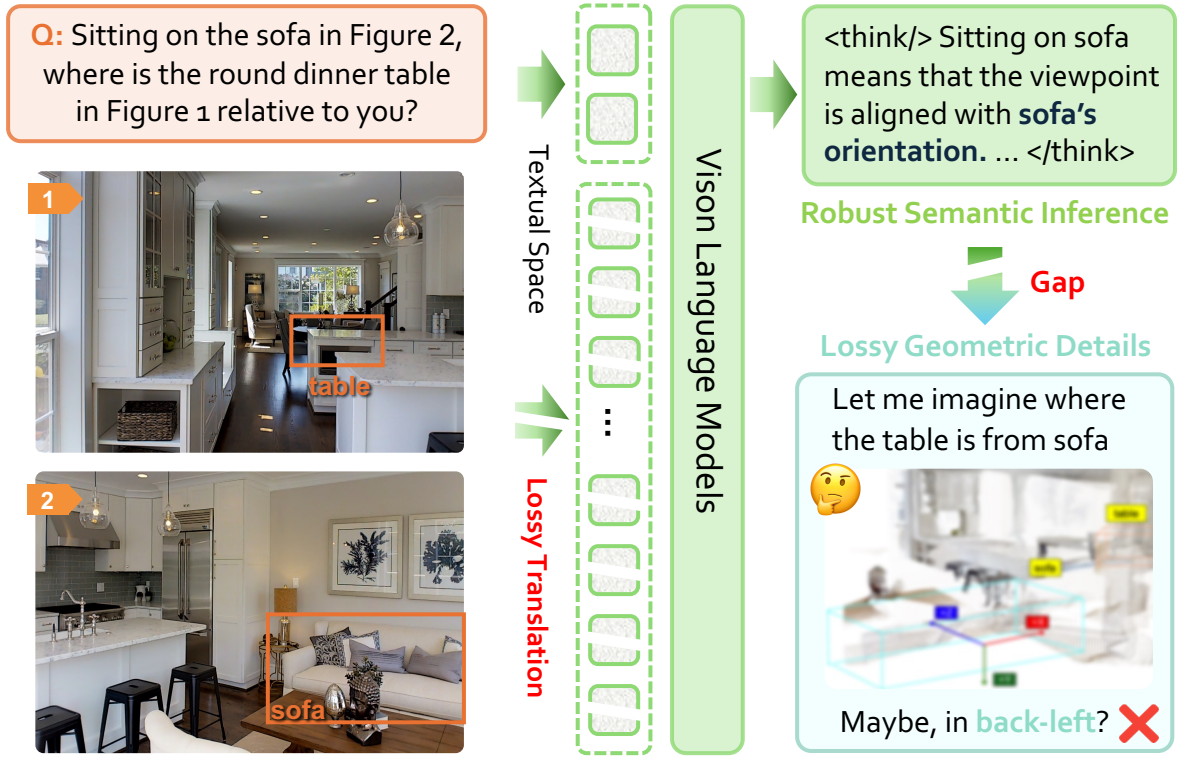
Zeren Chen[1,2*], Xiaoya Lu[2,3*], Zhijie Zheng[1,2], Pengrui Li[1], Lehan He[1,4], Yijin Zhou[2,3,4],
Jing Shao[2], Bohan Zhuang[5†], Lu Sheng[1†]

[1]School of Software, Beihang University, [2]Shanghai AI Laboratory, [3]Shanghai Jiao Tong University,
[4]Shanghai Innovation Institute, [5]ZIP Lab, Zhejiang University

*Equal Contribution, †Corresponding Author

Semantic-to-Geometric Gap

- Spatial reasoning often requires fine-grained geometric details, but a VLM usually compresses visual input into textual semantics, which distorts the geometry.
- Existing methods either inherit flawed supervision from imperfect "oracles", or still plan from unconstrained semantic assumptions.



Formal Task Constraint

- Human spatial reasoning implicitly fixes both where to measure from and what to measure. GCA turns these into explicit constraints.
- We introduce $\mathcal{C}_{\text{task}} = (\mathcal{C}_{\text{R}}, \mathcal{C}_{\text{O}})$ to resolve this ambiguity. $\mathcal{C}_{\text{R}}$ defines the reference frame, and $\mathcal{C}_{\text{O}}$ defines the objective measured within that frame.

Task Constraint

$\mathcal{C}_{\text{task}} = (\mathcal{C}_{\text{R}}, \mathcal{C}_{\text{O}})$

Reference Constraint $\mathcal{C}_{\text{R}}$

Where to measure from

object / camera / direction

Objective Constraint $\mathcal{C}_{\text{O}}$

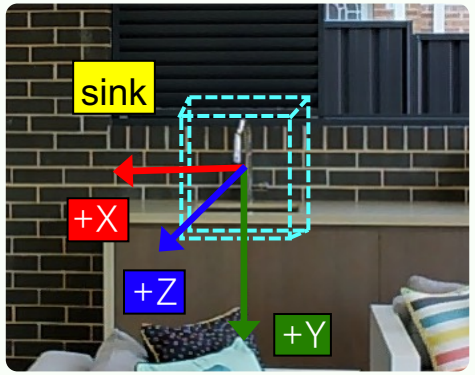
What to measure in that frame

target objects / relation / answer

Reference Constraint $\mathcal{C}_{\text{R}}$

Three ways to anchor the coordinate frame

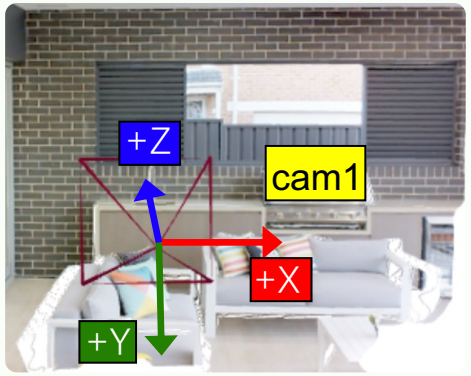
Object-based Frame



"When washing hands…"

$+Z_{\mathcal{R}} = -Z_{\text{sink}} = \text{front}$

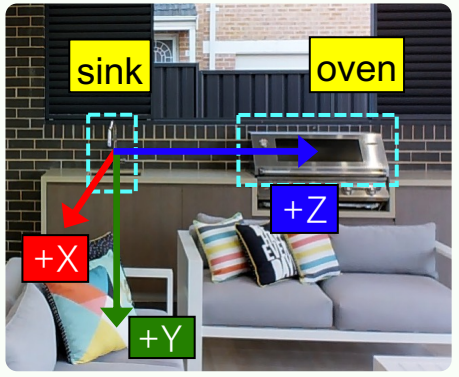
Camera-based Frame



"From viewpoint of Fig.1"

$+Z_{\mathcal{R}} = +Z_{\text{cam1}}$

Direction-based Frame

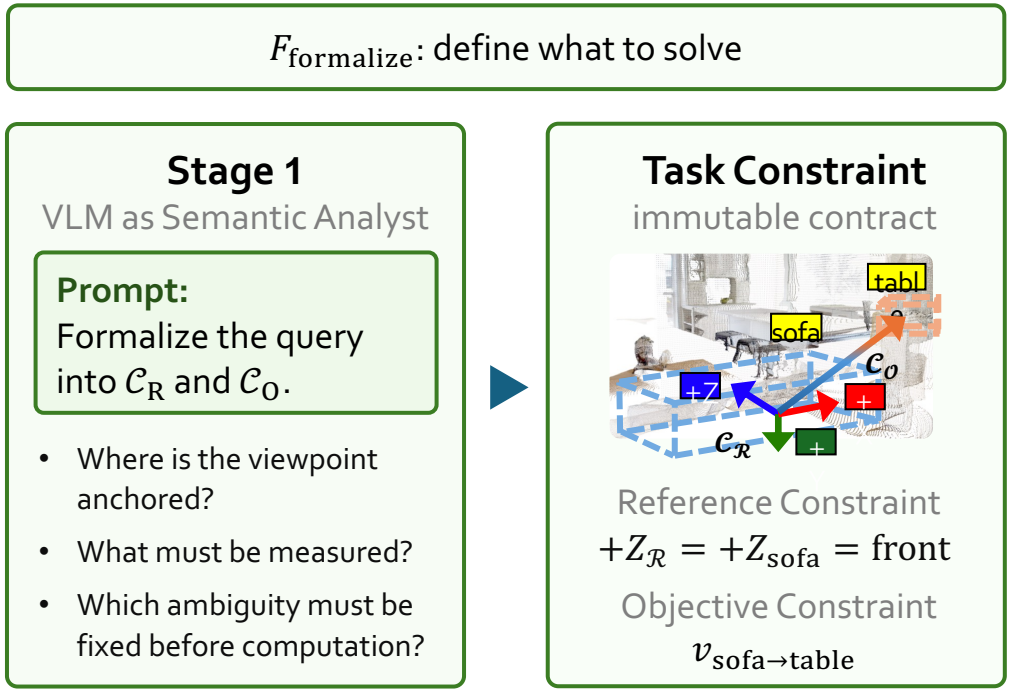


"Oven is north of sink…"

$+Z_{\mathcal{R}} = v_{\text{sink}\to\text{owen}} = \text{north}$

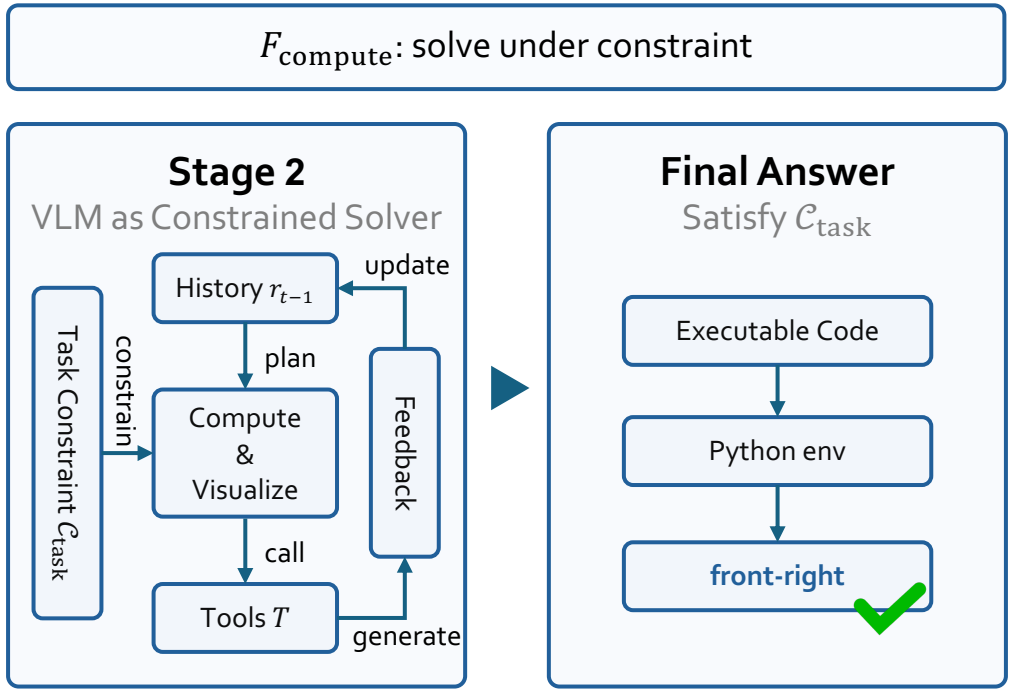
Two-Stage Workflow

- In the Task Formalization stage, VLM translates the ambiguous spatial language into the task constraint $\mathcal{C}_{\text{task}} = (\mathcal{C}_{\text{R}}, \mathcal{C}_{\text{O}})$.
- In the Constrained Computation stage, the VLM no longer has to imagine high-fidelity geometry but binds symbols to measured geometry and computes the answer.



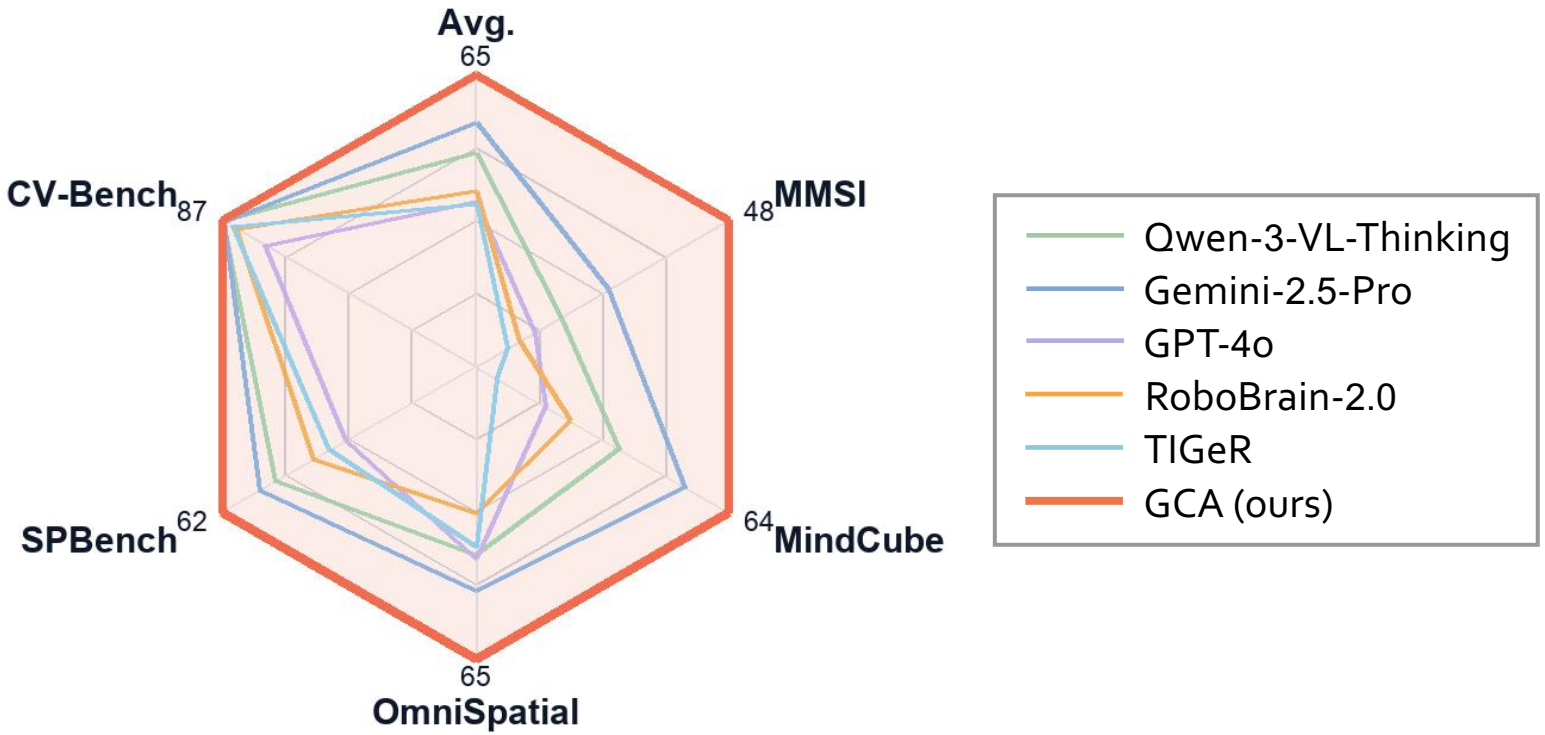
Two-Stage Workflow

- In the Task Formalization stage, VLM translates the ambiguous spatial language into the task constraint $\mathcal{C}_{\text{task}} = (\mathcal{C}_{\text{R}}, \mathcal{C}_{\text{O}})$.
- In the Constrained Computation stage, the VLM no longer has to imagine high-fidelity geometry but binds symbols to measured geometry and computes the answer.



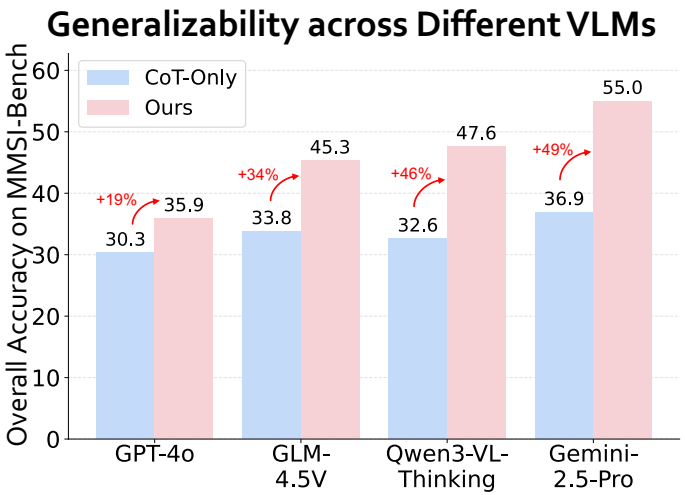
Experiments & Results

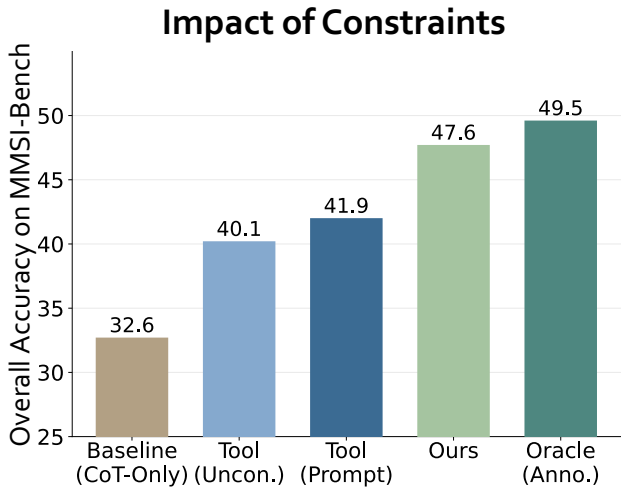
- Across multiple spatial reasoning benchmarks, GCA establishes new state-of-the-art performance. It reaches about 65 percent average accuracy, with clear gains on challenging multi-step settings such as MMSI-Bench and MindCube.

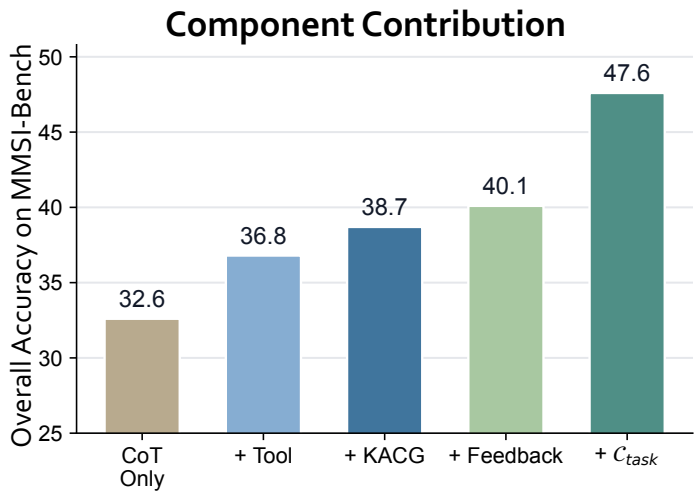


Experiments & Results

- GCA improves different VLM backbones over CoT-only baseline.
- Formal task constraints outperform unconstrained and prompt-guided tool use, and approach human oracle formalization.
- Task constraint gives the largest gain among all components.







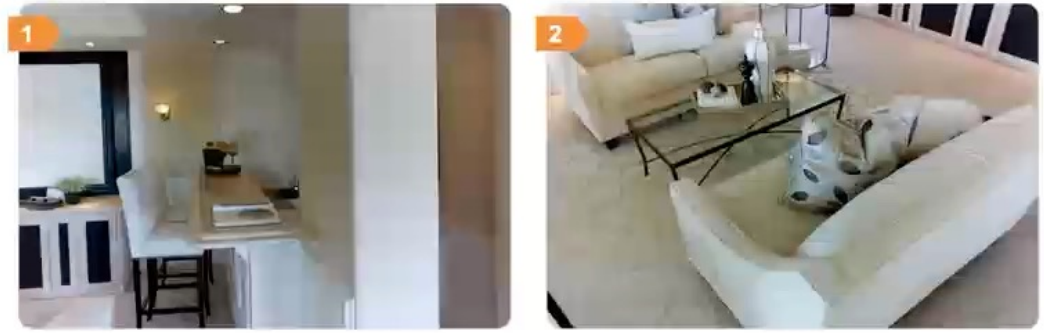
Representative Case Study

Geometrically-Constrained Agent for Spatial Reasoning

Spatial Reasoning Query

Q: What is the direction of the glass coffee table relative to the book on the bar counter (wine glass is north of the book)?

A: Northwest; B: Northeast; C: Southwest; D: Southeast



Case Study #2. Direction-Based Reference Frame

Thank You for Watching