MindCube

MLL LAB
MACHINE LEARNING AND LANGUAGE

Spatial Mental Modeling from Limited Views

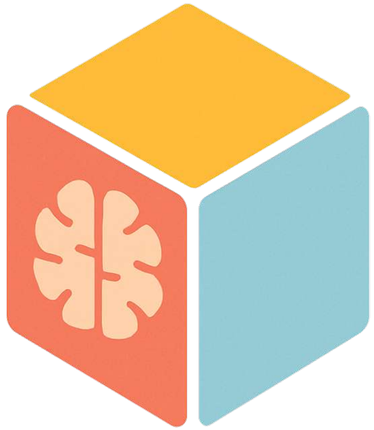
Baiqiao Yin*, Qineng Wang*‡, Pingyue Zhang, Jianshu Zhang,
Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran,
Han Liu, Ranjay Krishna, Saining Xie, Manling Li†, Jiajun Wu†, Li Fei-Fei†




NYU

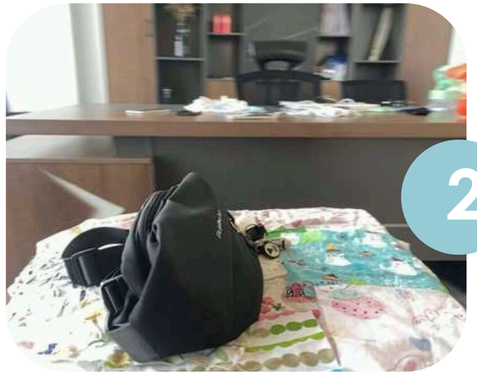

Can VLMs **imagine full scene from limited views?**

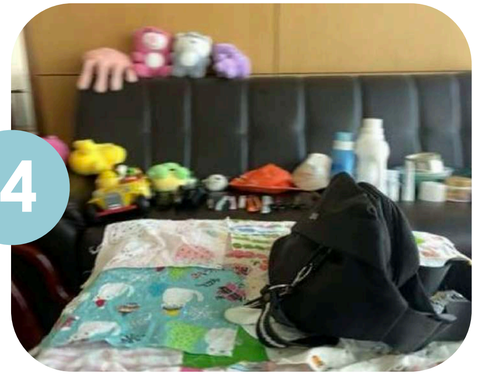
Can VLMs build **spatial mental models** like humans?



spatial mental models


3

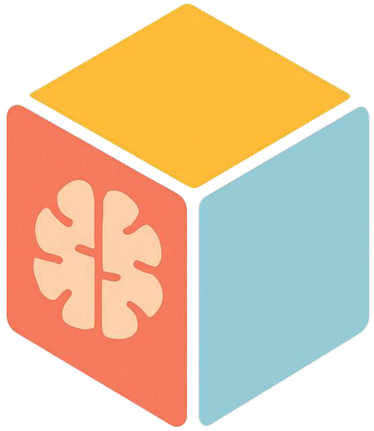

2


4


1

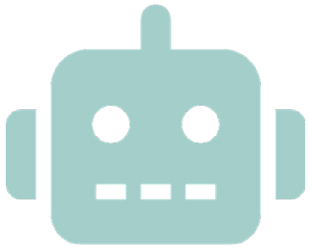
At 4 what is to the **left** of the **black bag** ?

A. Window B. Door C. Desk D. Sofa



At 4 what is to the **left** of the **black bag** ?

A. Window B. Door C. Desk D. Sofa



VLMs

At 4 what is to the **left** of the **black bag** ?

A. Window B. Door C. Desk D. Sofa

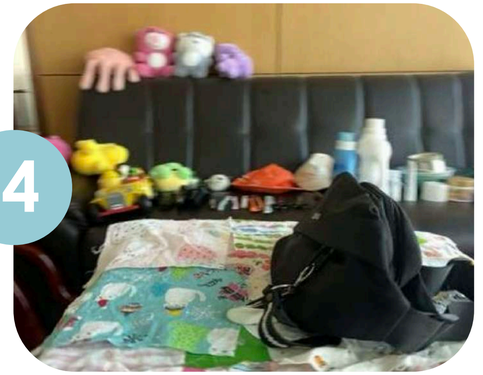
At 4 what is to the **left** of the **black bag** ?

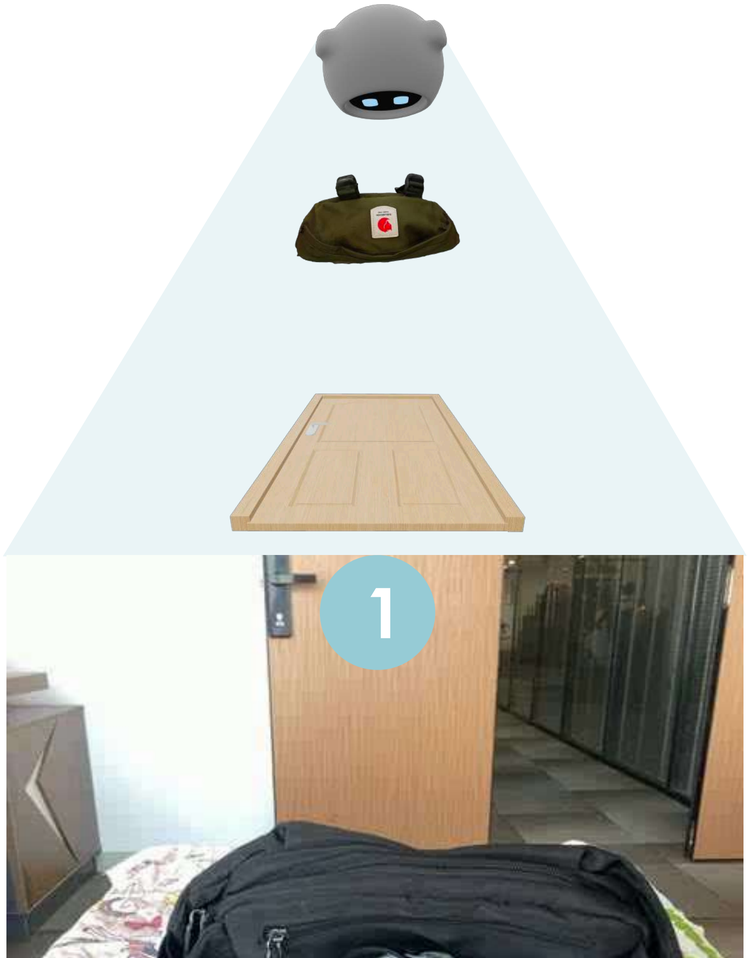
A. Window B. Door C. Desk D. Sofa

?

Cognitive Map

4



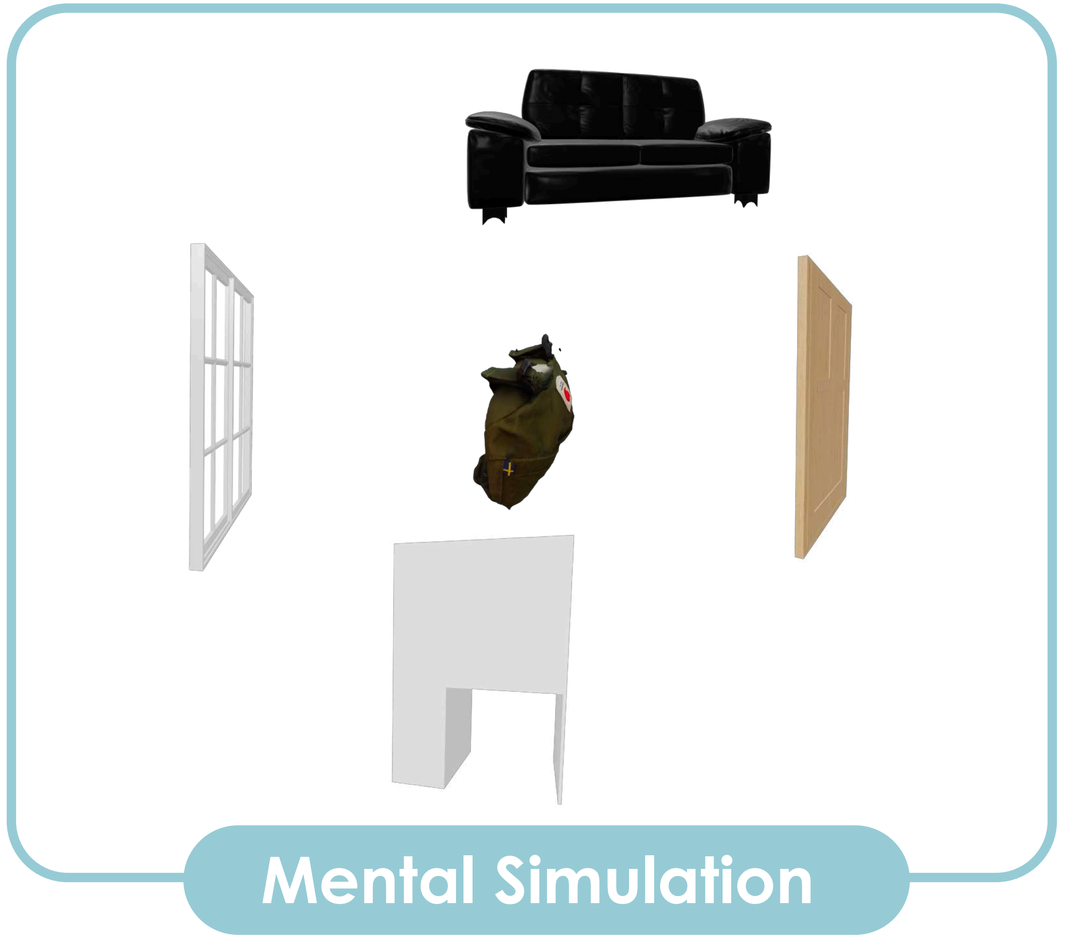

1


2


3


4

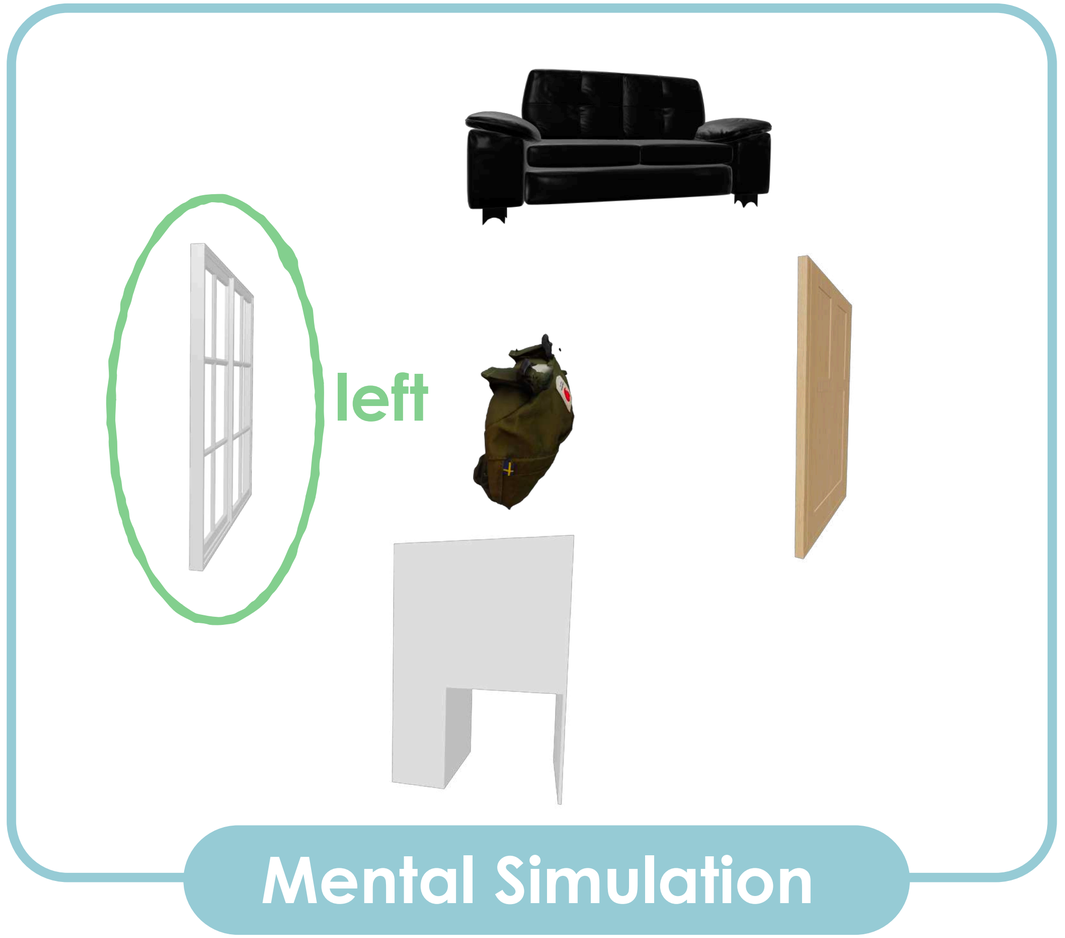
At 4 what is to the **left** of the **black bag** ?



Mental Simulation

At 4 what is to the **left** of the **black bag** ?

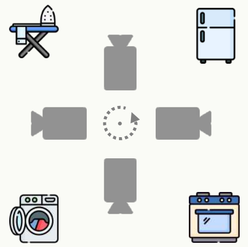
A. Window ✓ B. Door C. Desk D. Sofa



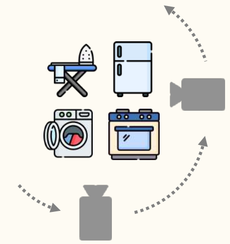

MindCube
21,154 questions
3,268 images

Three Movement Patterns in MindCube

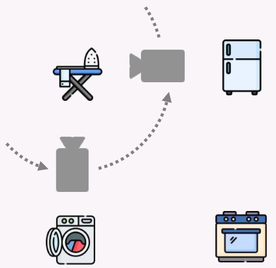
Rotation



Around



Among



Baseline | Open-Weight Multi Image Models | Proprietary Models | Spatial Models

Method	Overall	Rotation	Among	Around
DeepSeek-VL2-Small	**47.62**	37.00	**50.38**	26.91
LLaVA-Onevision-7B	**47.43**	36.45	**48.42**	44.09
Gemma-3-12B-it	46.67	38.39	48.38	34.63
mPLUG-Owl3-7B-241101	44.85	37.84	47.11	26.91
Claude-4-Sonnet-20250514	44.75	**48.42**	44.21	**47.62**
VLM-3R	42.09	36.73	44.22	24.45
LLaVA-Video-Qwen-7B	41.96	35.71	43.55	30.12
Random (frequency)	*33.02*	*38.30*	*32.66*	*35.79*
Random (chance)	*32.35*	*36.36*	*32.29*	*30.66*
Spatial-MLLM	32.06	38.39	20.92	32.82
LongVA-7B	29.46	35.89	29.55	24.88
Qwen2.5-VL-7B-Instruct	29.26	**38.76**	29.50	21.35
SpaceMantis	22.81	37.65	21.26	29.32
InternVL2.5-8B	18.68	36.45	18.20	13.11

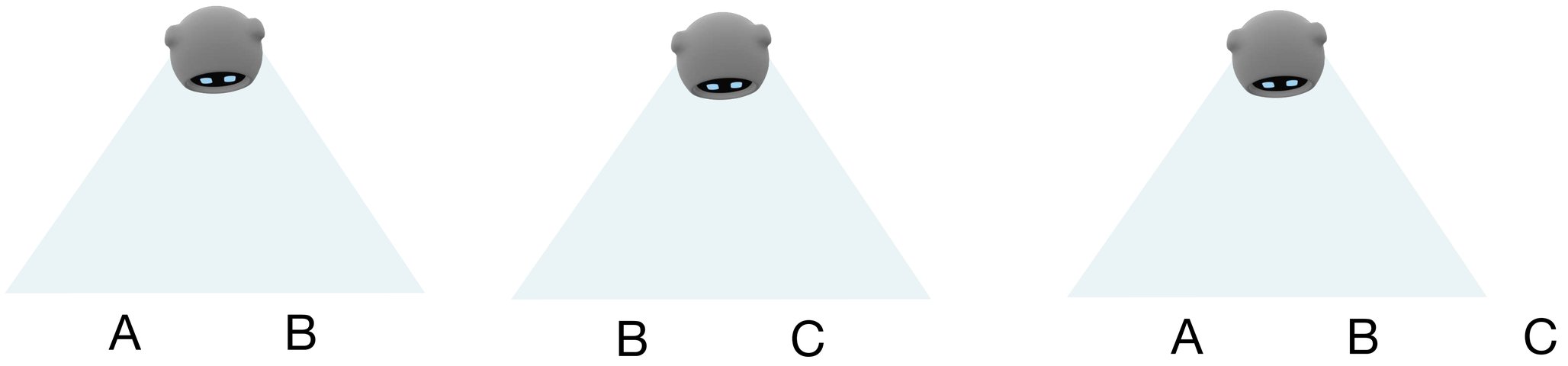
Current VLMs Perform Worse on MindCube

How to teach VLMs to

Approximate Spatial Mental Models?

Spatial Mental Models are not “CONCRETE”

left(A, B) left(B, C) → left(A, C) is True



Approximate Spatial Mental Models in Frozen VLMs

View Interpolation

+0.10%

Reasoning and Cognitive Map are more effective scaffolds

Cognitive Map

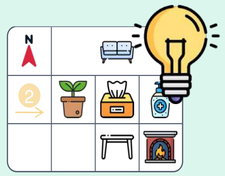

```
{"objects": [{
    "name": "Tissue box",
    "position": [5, 5]
    }, {
    "name": "Hand sanitizer",
    "position": [7, 5]
    }, ...],
 "views": [{
    "name": "View 1",
    "position": [5, 6],
    "facing": "up"
    } ...
```


+3.52%

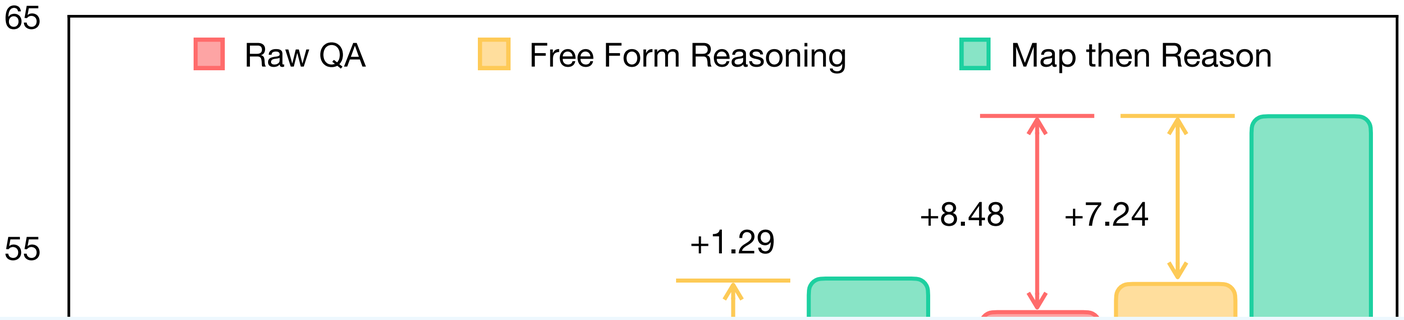
Data Scaffolds

Compare to Raw QA

Approximate Spatial Mental Models

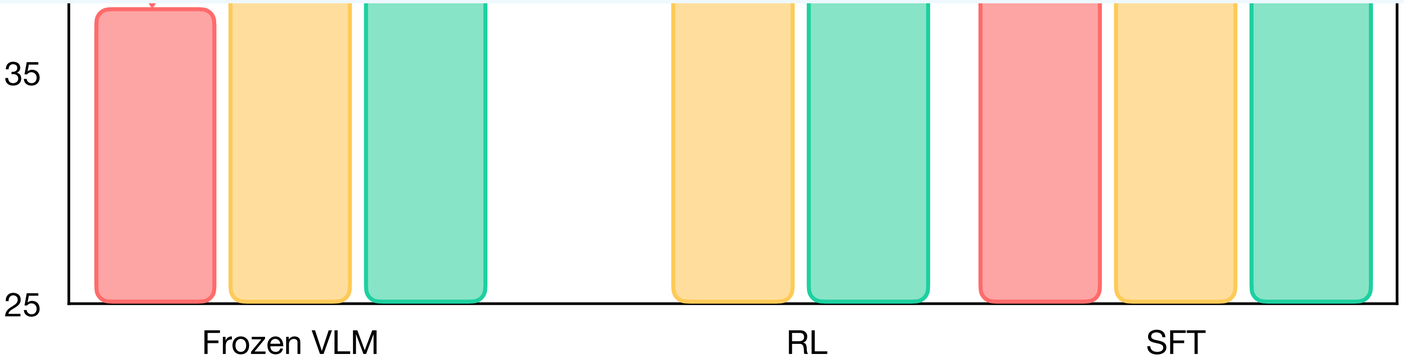


Map then Reason



“Map then Reason” is the most effective approximation

Raw QA



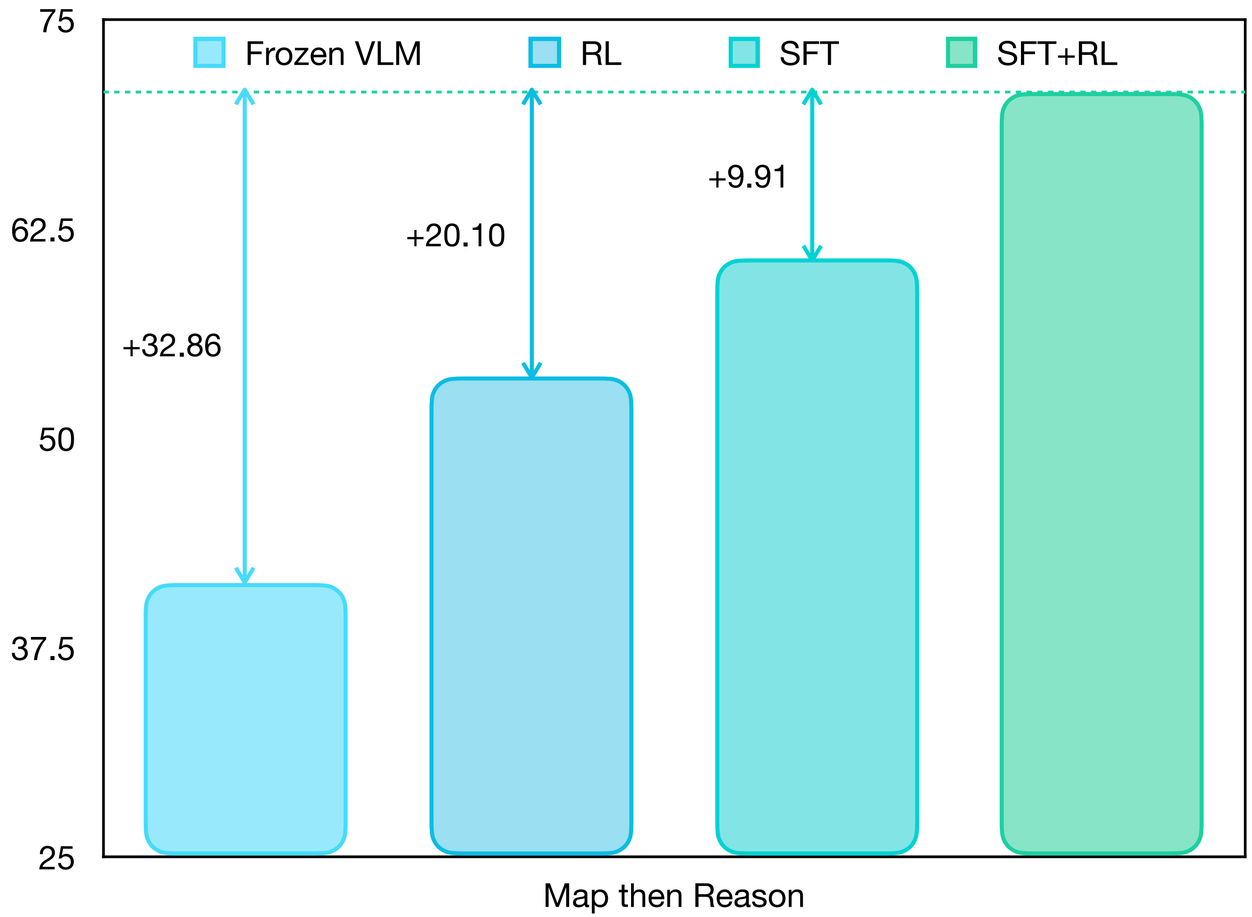
SFT then RL can boost the performance

SFT+RL

Frozen VLM

RL

SFT



Some Interesting Findings (1/3)

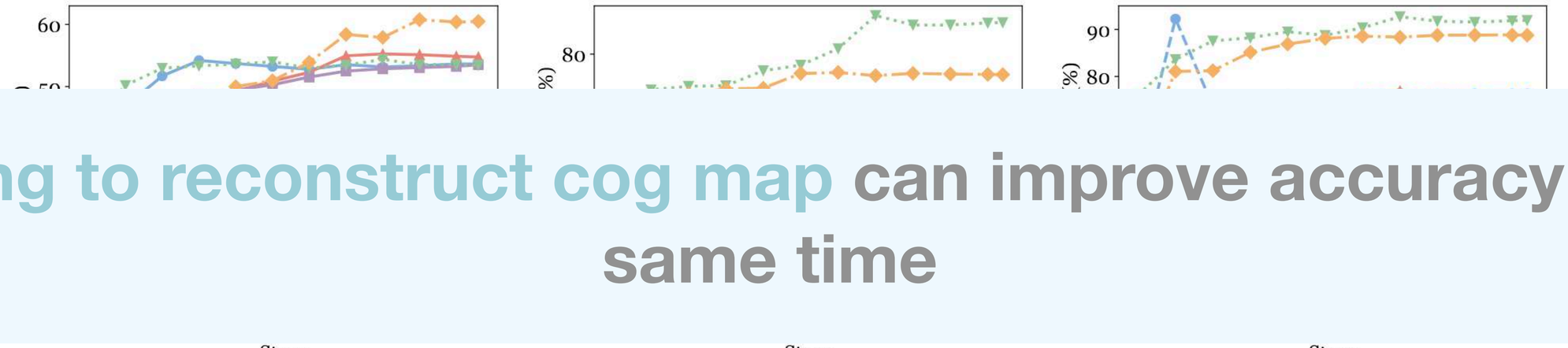
3.1 Scaffolding Spatial Reasoning in Frozen VLMs

💡 **Key Takeaways: Scaffolding Spatial Reasoning in Frozen VLMs**

- *Explicit reasoning is crucial for improving performance.*
- *Cognitive maps can help guide the reasoning process.*
- *Passive structures (like maps as input) alone and visual continuity offer little benefit.*

Performance Analysis: Limited Gains from External Scaffolds +

Some Interesting Findings (2/3)



Learning to reconstruct cog map can improve accuracy at the same time

(a) QA accuracy (%)

(b) Isomorphic Rate (%)

(c) Overall Similarity (%)

Some Interesting Findings (2/3)

3.2 Teaching VLMs to Reason Spatially

💡 **Key Takeaways: Teaching VLMs to Reason Spatially**

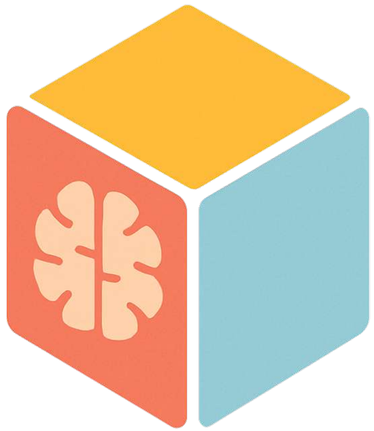
- *Joint cogmap and reasoning setting yields optimal performance through synergistic effects.*
- *Reasoning shapes spatial representations for functional utility, not just structural perfection.*
- *Neither map generation nor reasoning alone largely outperforms the SFT QA baseline.*

Some Interesting Findings (3/3)

3.3 Reinforcement Learning for Spatial Reasoning

💡 **Key Takeaways: Reinforcement Learning for Spatial Reasoning**

- *Combining cognitive maps with reasoning consistently improves all learning outcomes.*
- *Starting from scratch, RL provides only marginal gains for spatial reasoning; its true power is unlocked when building upon a strong SFT foundation.*



MindCube

https://mll-lab-nu.github.io/mind-cube