Google DeepMind
TEXAS The University of Texas at Austin Stanford

PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs

Soroush Nasiriany*, Fei Xia*, Wenhao Yu*, Ted Xiao*, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu , Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, Brian Ichter*

Pitch: How can we enable VLMs to solve robot control and spatial reasoning tasks **without any fine-tuning?**

- Prompting with Iterative Visual Optimization PIVOT: casts problem as **iterative visual question answering**
- Annotate images with **visual markers** representing actions or referrals, query VLM to select best proposals
- **Iteratively refine proposals** by fitting new action distributions and re-querying VLM

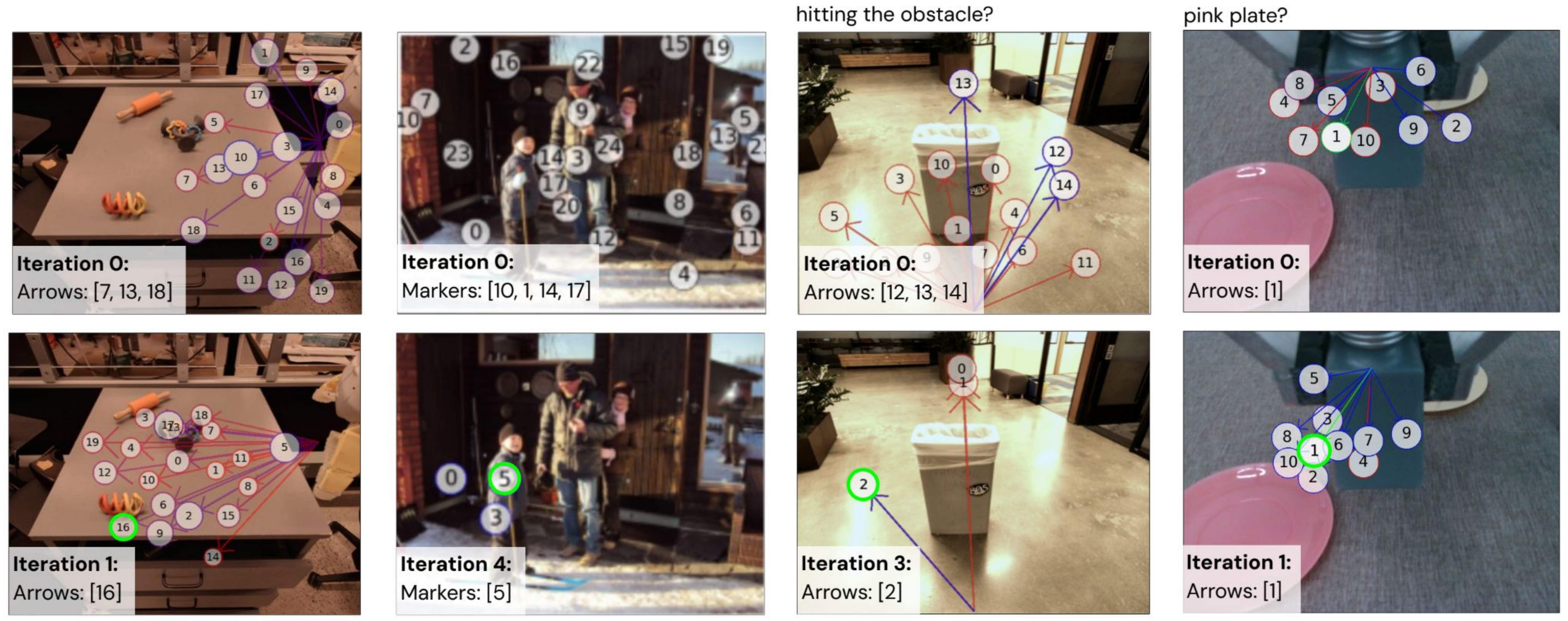
PIVOT can be used across diverse robot and spatial reasoning tasks

Task: What actions should the robot take to pick up the DNA chew toy?

Task: What numbers overlay the "L kid"?

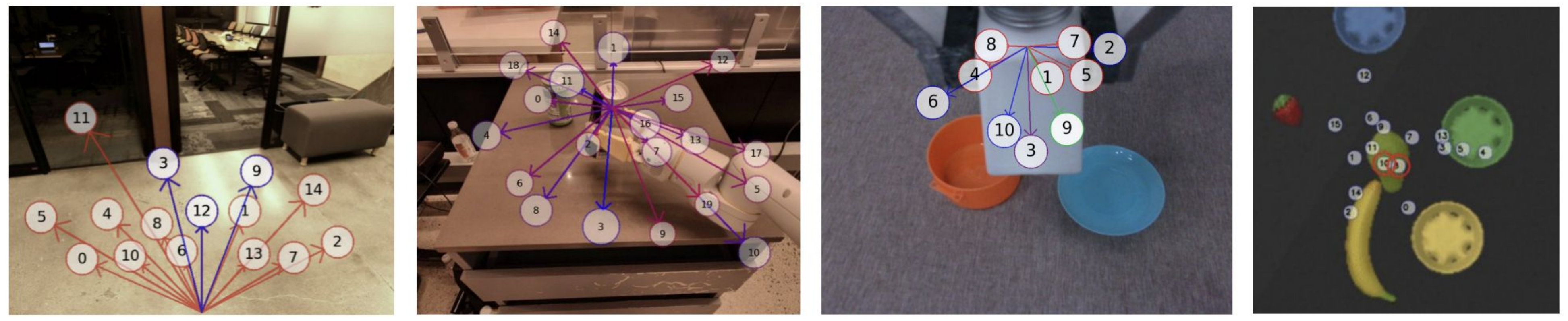
Task: What actions should the robot take to go to wooden bench without hitting the obstacle?

Task: What actions should the robot take to put the pepper shaker on the pink plate?



Robot Experiments

Across three robot manipulation domains and one robot navigation domain, we show that PIVOT can perform a diverse set of tasks zero-shot.

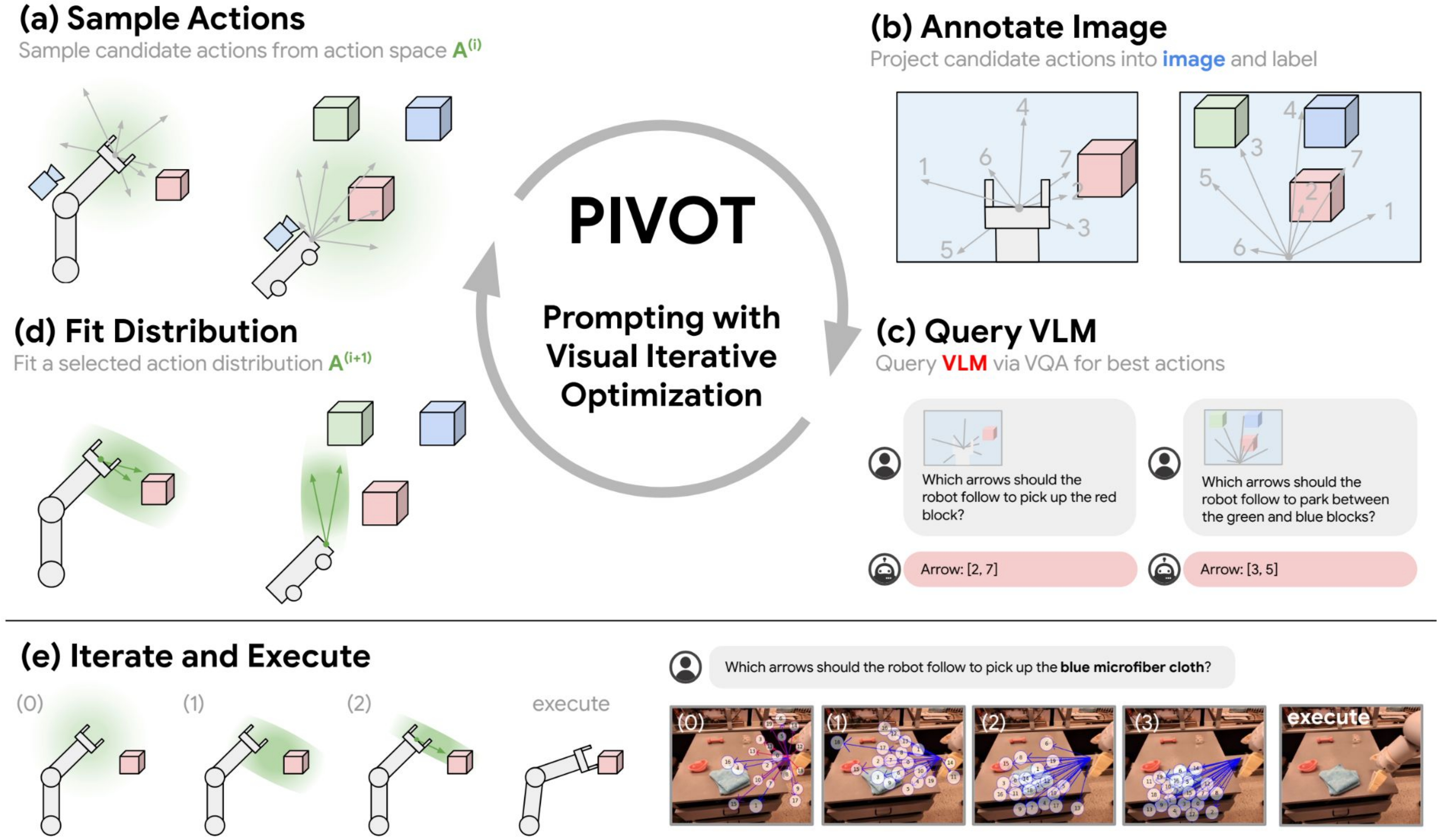


Manipulation

	No Iterations No Parallel			3 Iterations No Parallel			3 Iterations 3 Parallel		
Task	Reach	Steps	Grasp	Reach	Steps	Grasp	Reach	Steps	Grasp
Pick coke can	50%	4.5	0.0%	67%	**3.0**	33%	**100%**	**3.0**	**67%**
Bring the orange to the X	20%	4.0	-	**80%**	**3.5**	-	67%	**3.5**	-
Sort the apple	67%	3.5	-	**100%**	3.25	-	75%	**3.0**	-

Navigation

Task	No Iteration No Parallel	3 Iterations No Parallel	No Iteration 3 Parallel	3 Iterations 3 Parallel
Go to orange table with tissue box	25%	50%	**75%**	**75%**
Go to wooden bench without hitting obstacle	25%	50%	**75%**	50%
Go to the darker room	25%	50%	75%	**100%**
Help me find a place to sit and write	75%	50%	**100%**	75%



(a) Sample Actions
Sample candidate actions from action space $\mathbf{A}^{(i)}$

(b) Annotate Image
Project candidate actions into **image** and label

(c) Query VLM
Query **VLM** via VQA for best actions

Which arrows should the robot follow to pick up the red block?
Arrow: [2, 7]

Which arrows should the robot follow to park between the green and blue blocks?
Arrow: [3, 5]

(d) Fit Distribution
Fit a selected action distribution $\mathbf{A}^{(i+1)}$

PIVOT
Prompting with Visual Iterative Optimization

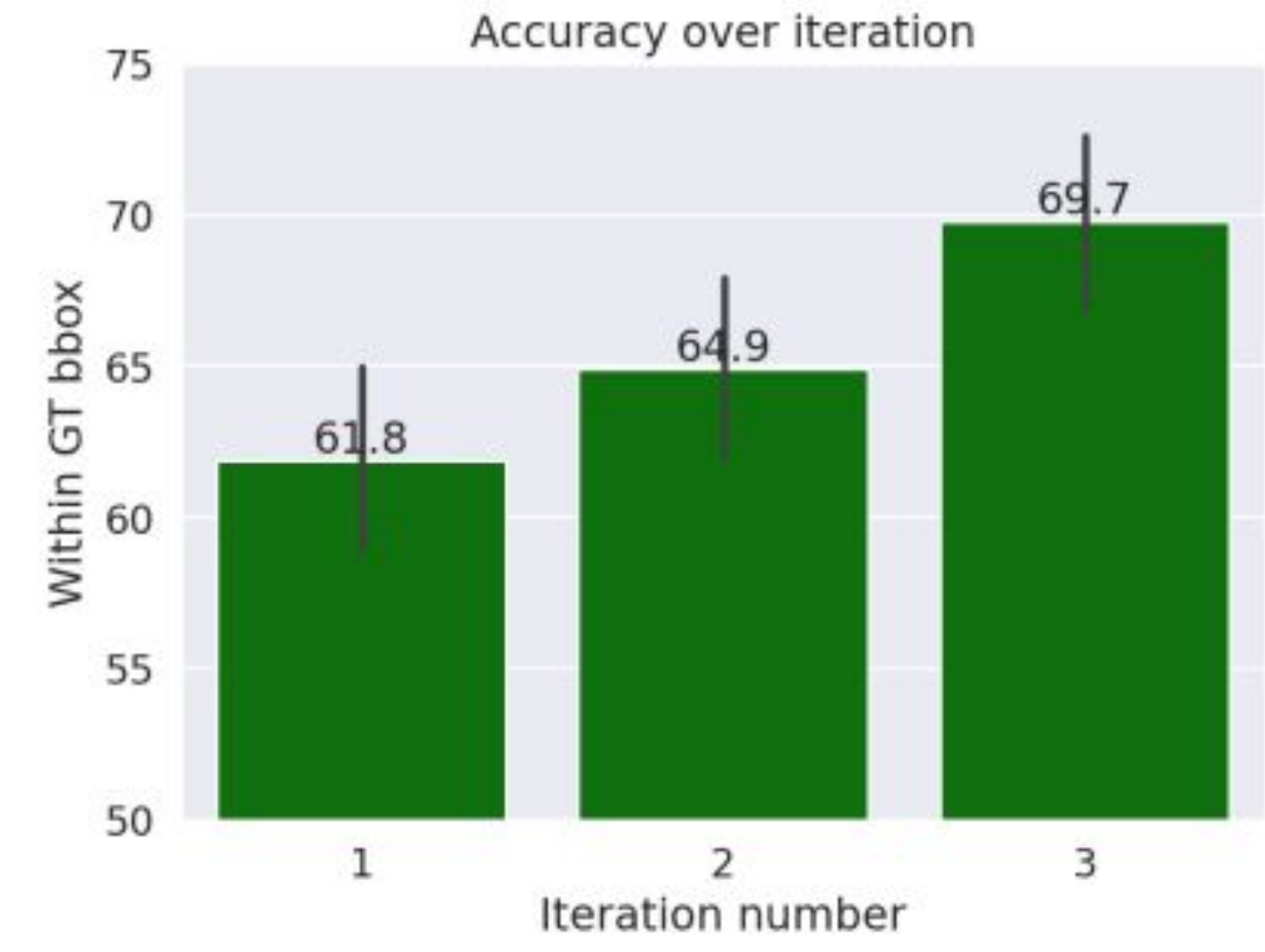
(e) Iterate and Execute

Which arrows should the robot follow to pick up the **blue microfiber cloth**?

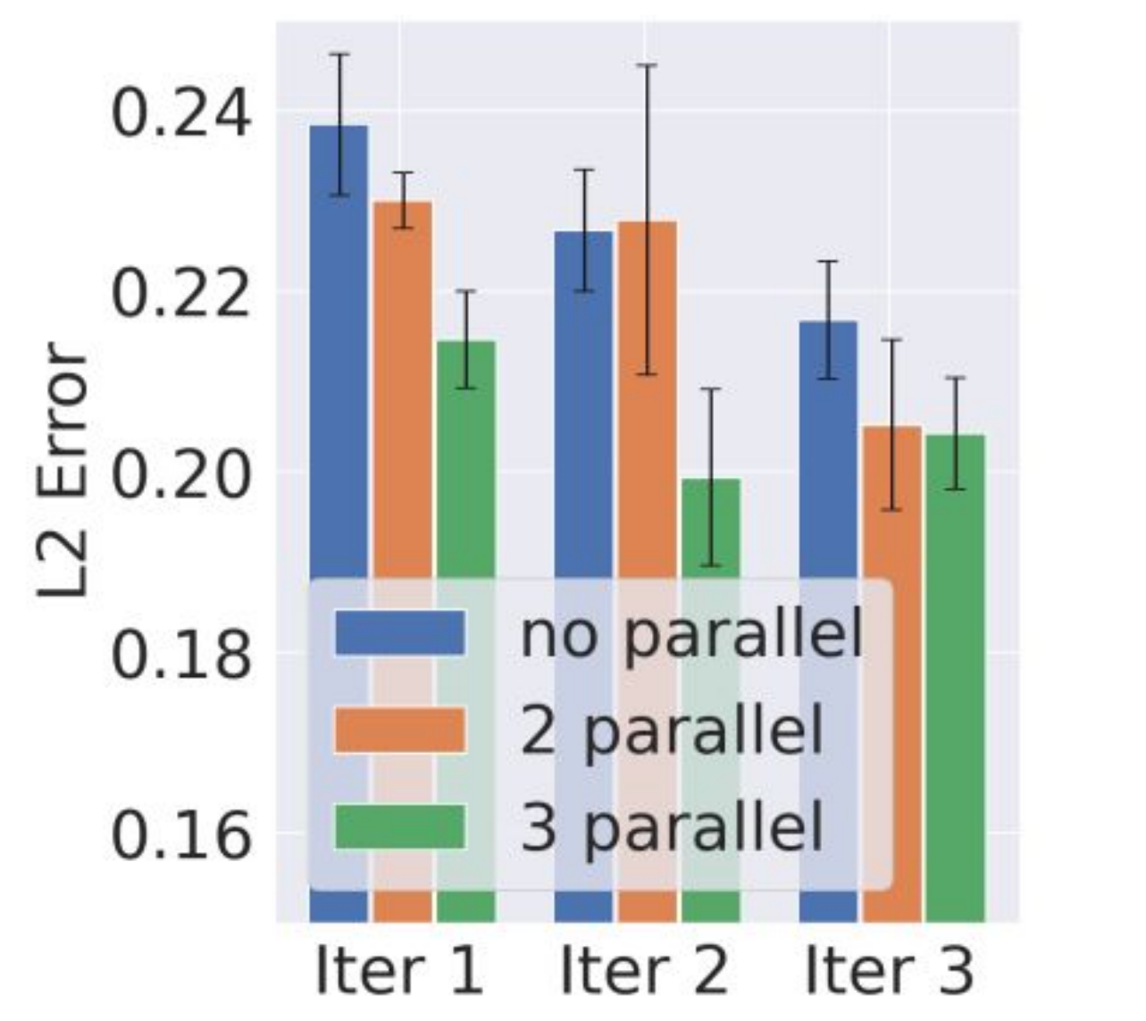
Improving reasoning through iterative refinement

Consistent improvement with iteratively resampling actions

RefCOCO spatial reasoning

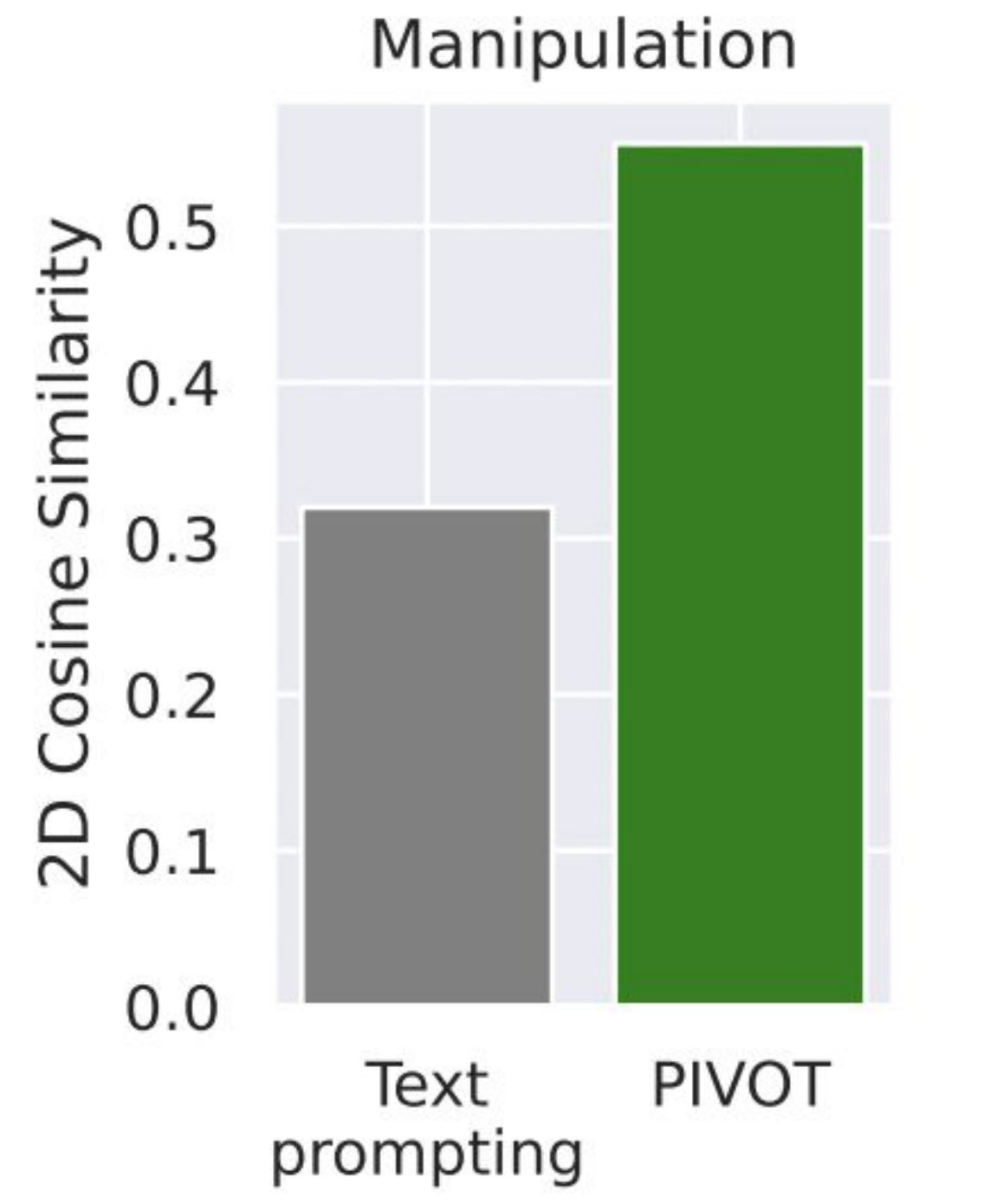


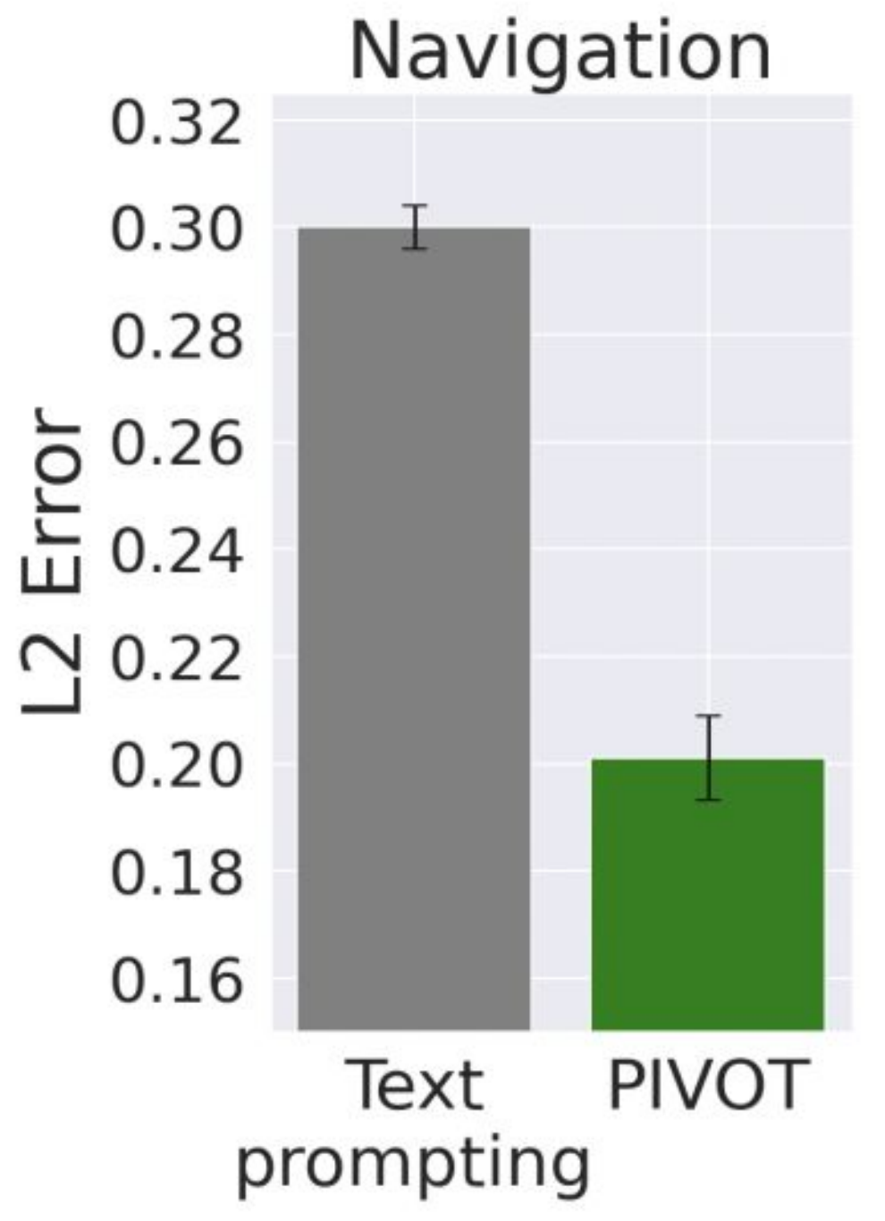
Navigation (offline eval)



Visual vs. text based markers

Representing actions via visual markers is superior to representing as text actions

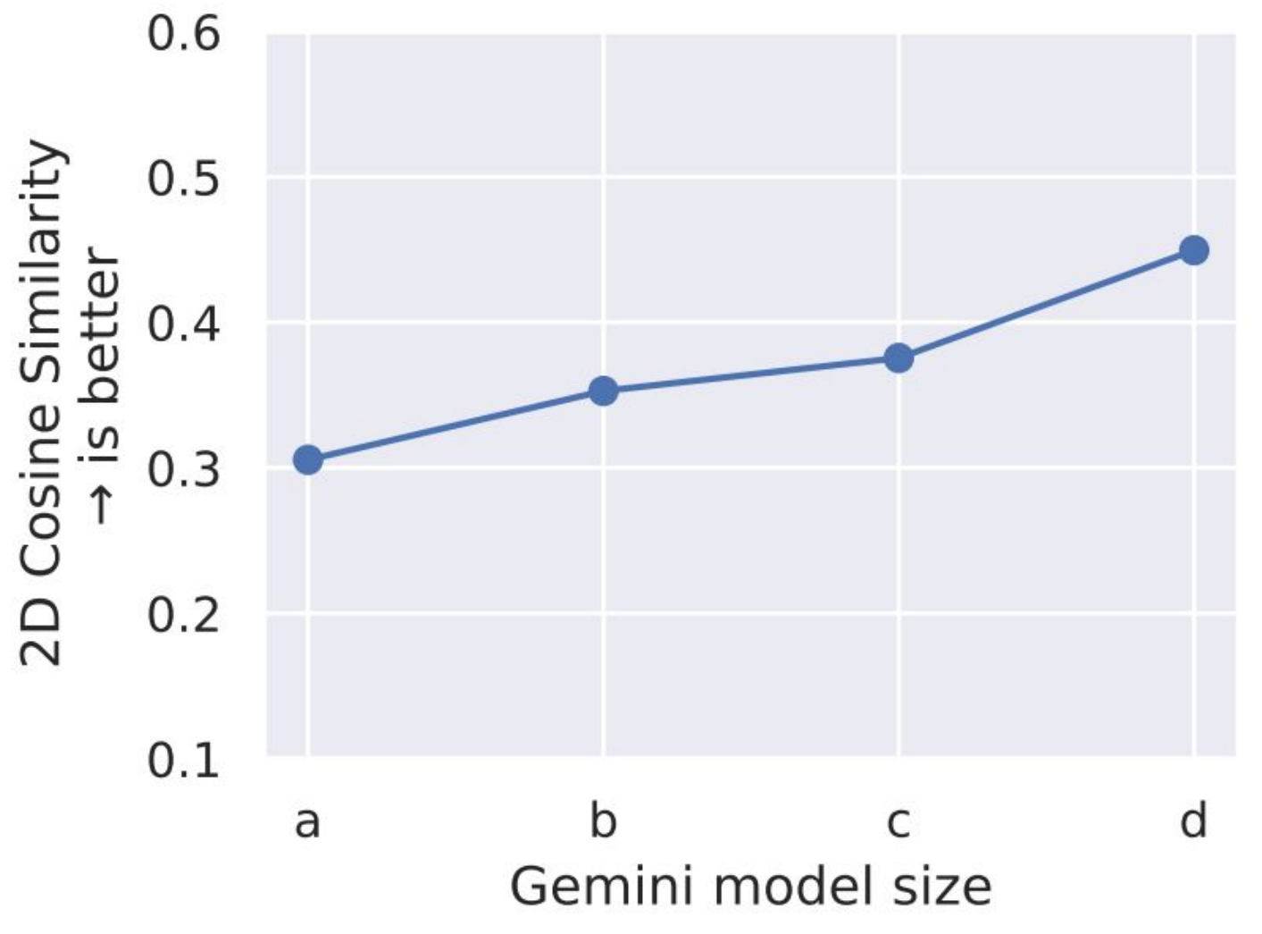




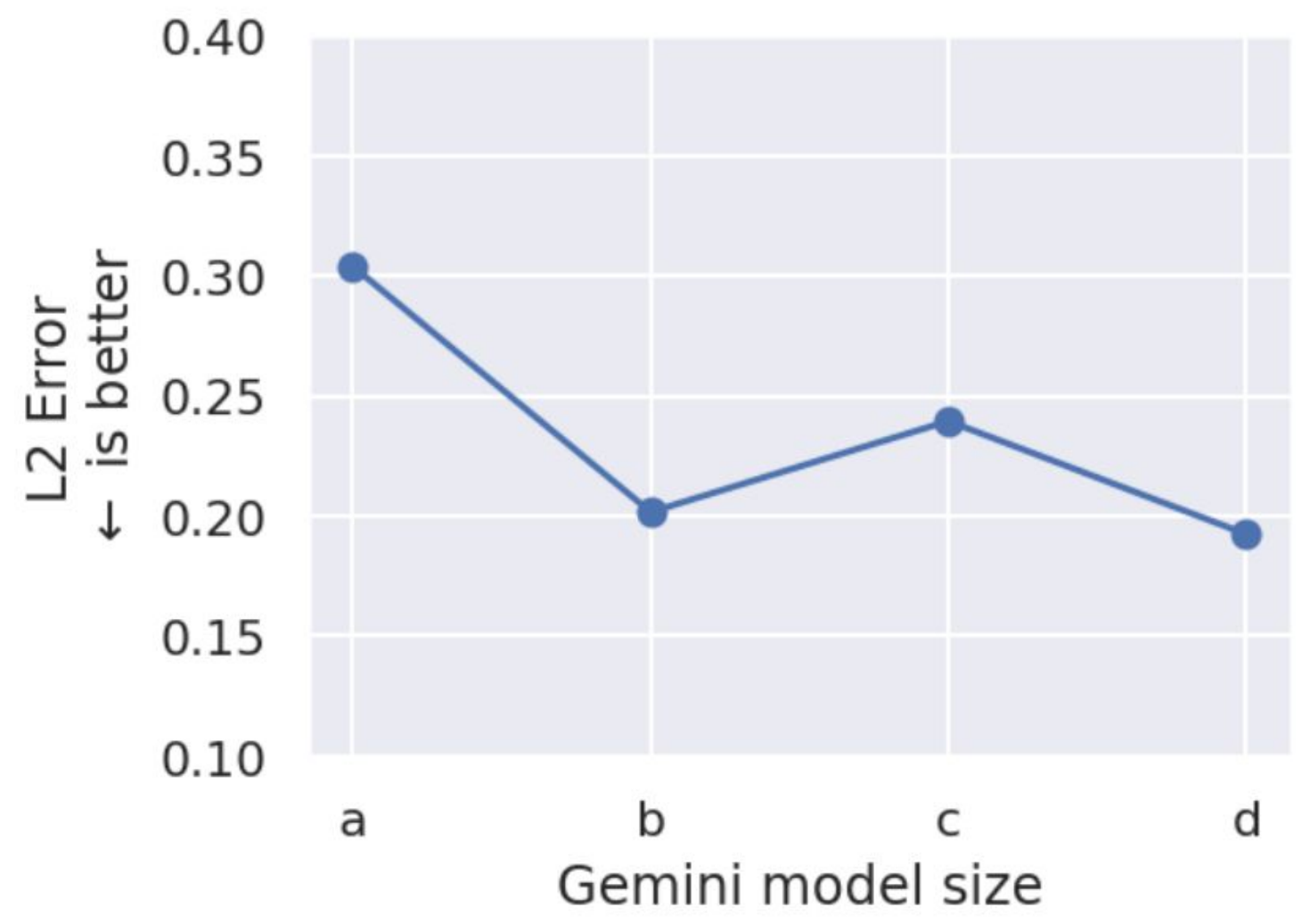
Iterative visual prompting scales with models

Performance scales with increasing Gemini model size

Manipulation



Navigation



Limitations and future steps

- Limited 3D reasoning from VLMs trained on 2D images. Explore VLMs trained to reason about 3D information
- Limited multi-step reasoning. Improve chain-of-thought reasoning capabilities of VLMs
- Limited reasoning for fine-grained manipulation tasks. Explore fine-tuning VLMs on robotic control tasks