Active Open-Vocabulary Recognition: Let Intelligent Moving Mitigate CLIP Limitations

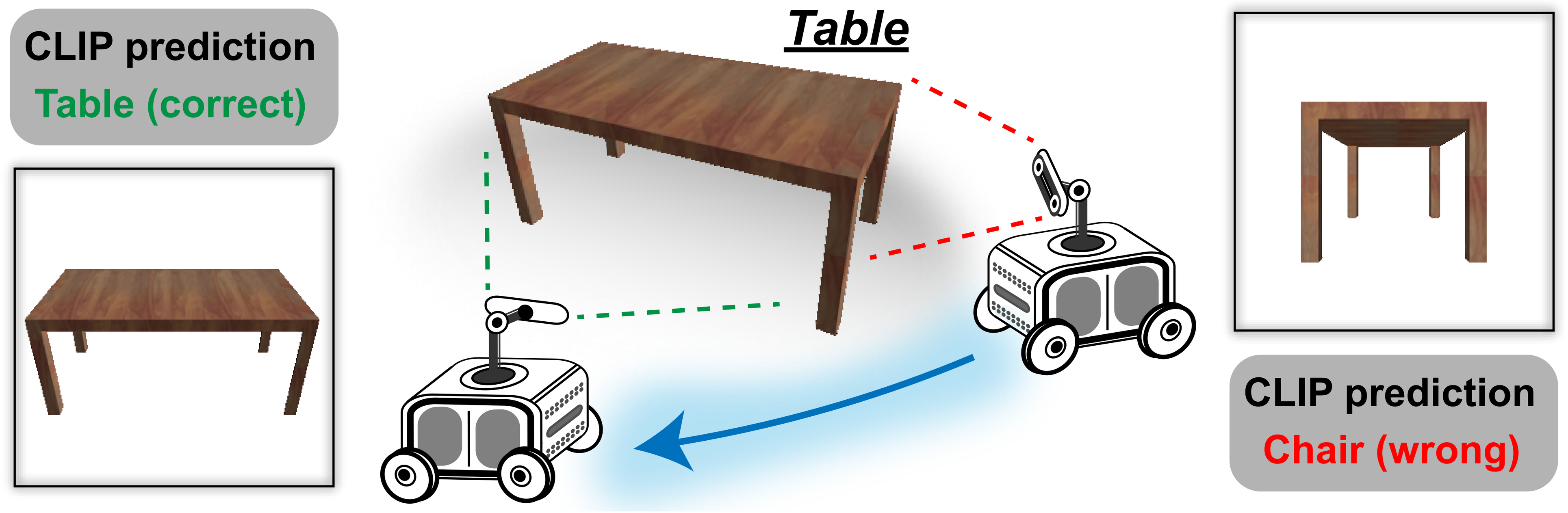
Northwestern University

Lei Fan, Jianxiong Zhou, Xiaoying Xing and Ying Wu

{leifan, jianxiongzhou2026, xiaoyingxing2026}@u.northwestern.edu, yingwu@northwestern.edu

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivations



Active recognition: by making movements, the agent can correct its recognition failure at the starting position.

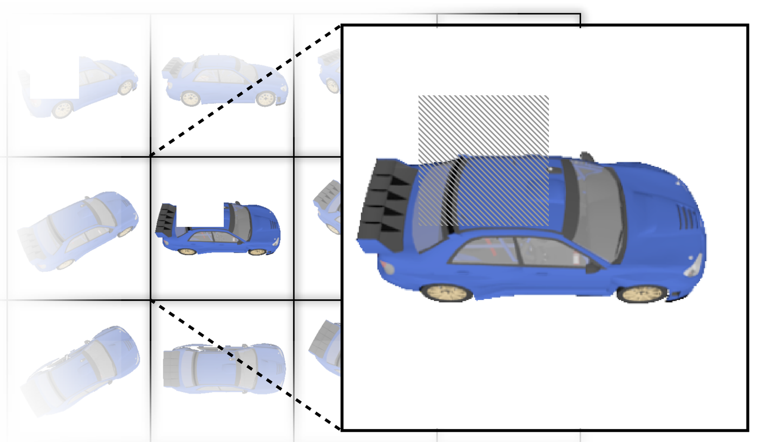
We are driven by dual motivations

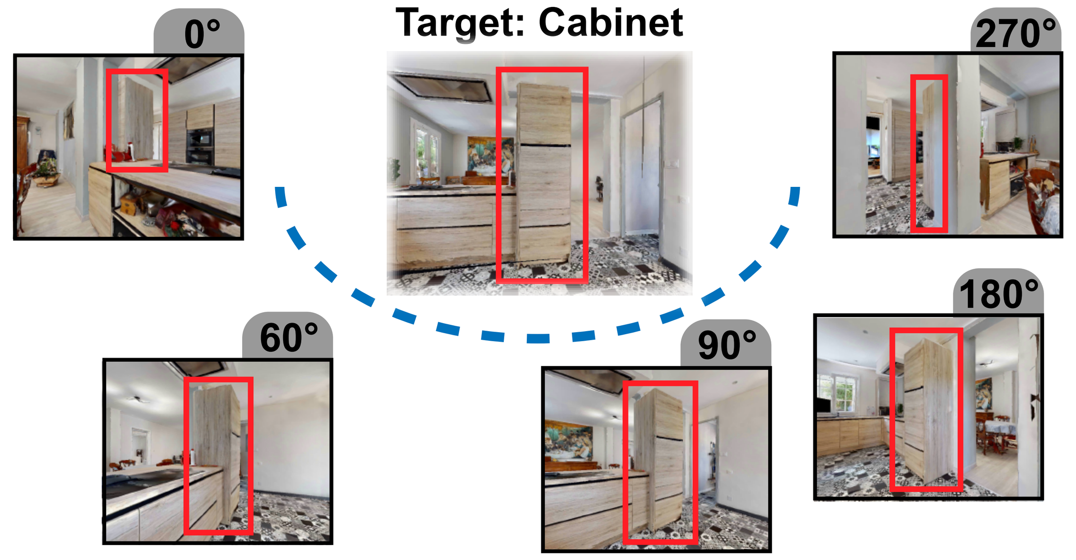
1. Enhance the capabilities of active recognition agents in handling open vocabulary using CLIP.
2. Overcome the inherent limitations of CLIP in unconstrained embodied perception scenarios.

Investigation Dataset

For better investigation of varying viewpoints and occlusions, we collect testing datasets from two widely-adopted platforms.

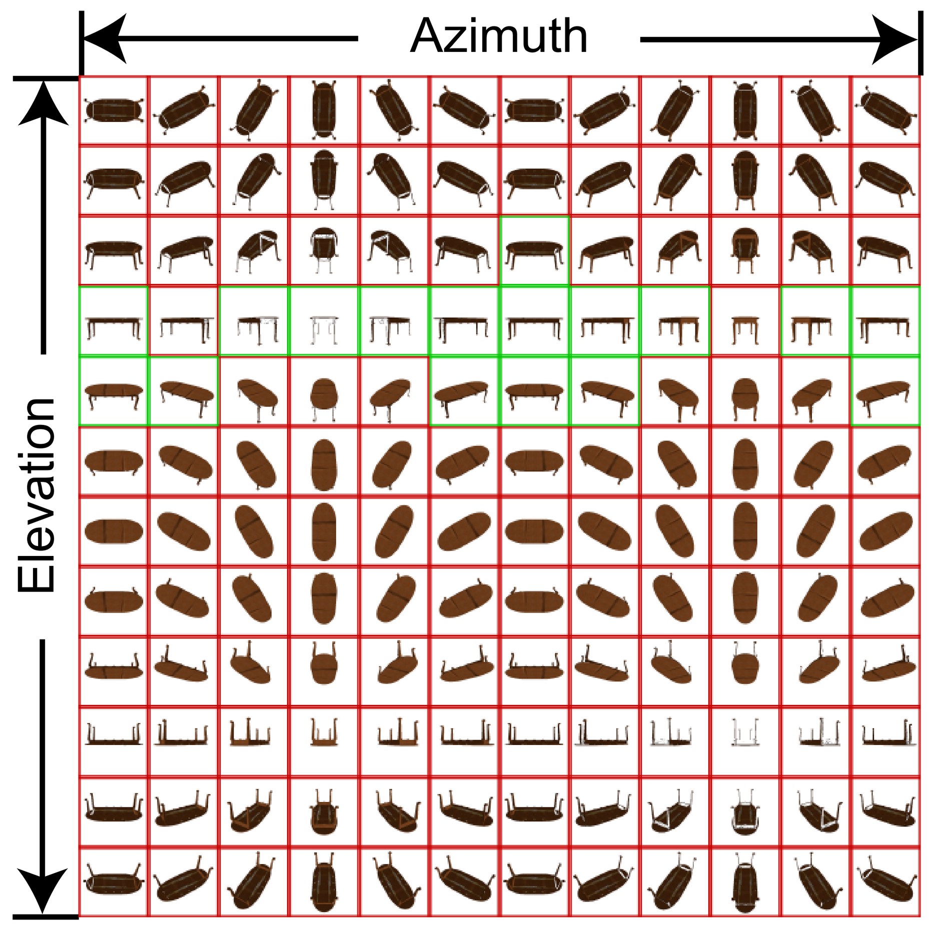
Investigation ShapeNet dataset
a. 12 x 12 different viewpoints.
b. Randomly added occlusions



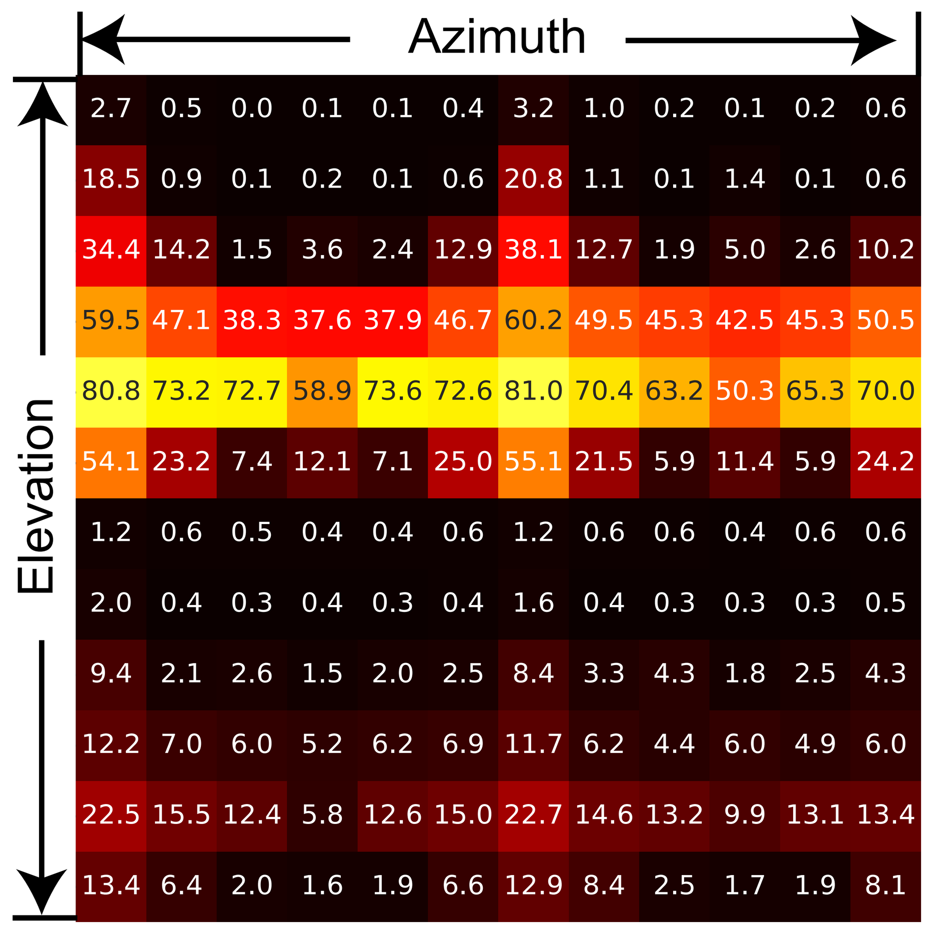


Investigation Habitat dataset: 30° increments around the target

CLIP: Sensitivity to Viewpoints and Occlusions

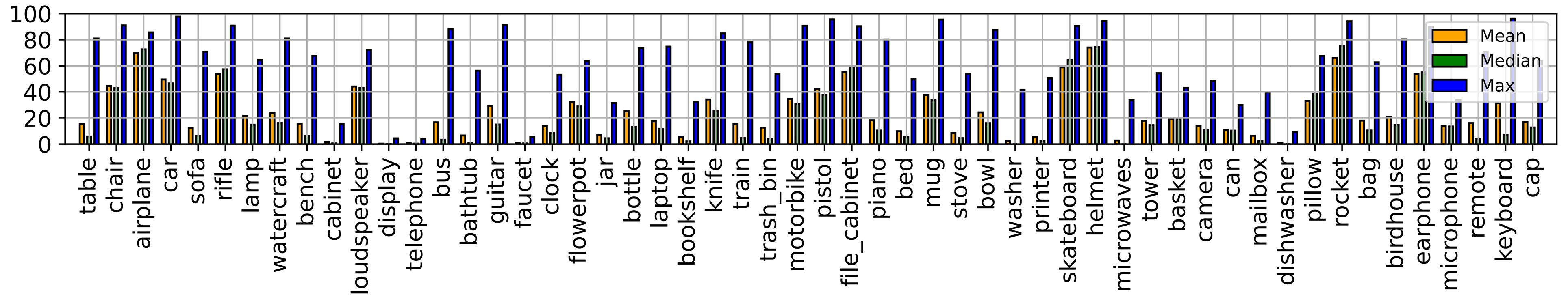


A sample from “table” class. Grid color for correct or wrong prediction.



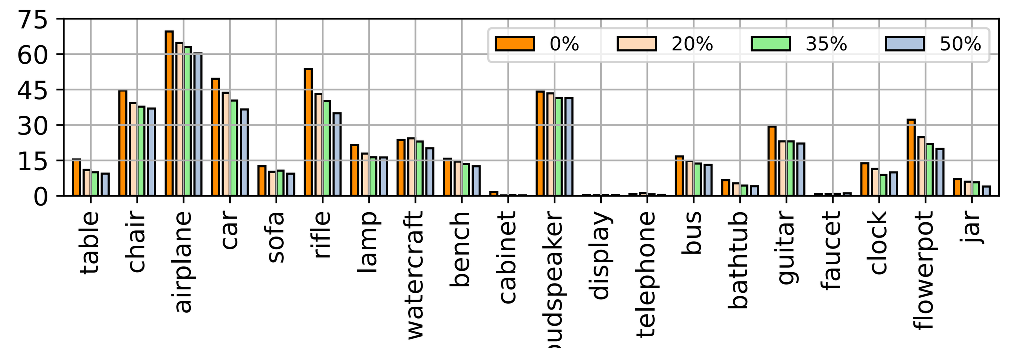
Average accuracy across all samples within the “table” class for each view.

The performance of CLIP on the ”table” class. The heatmap reveals a significant imbalance in accuracy across various viewpoints, underscoring the importance of active observation selection in embodied agents equipped with CLIP.



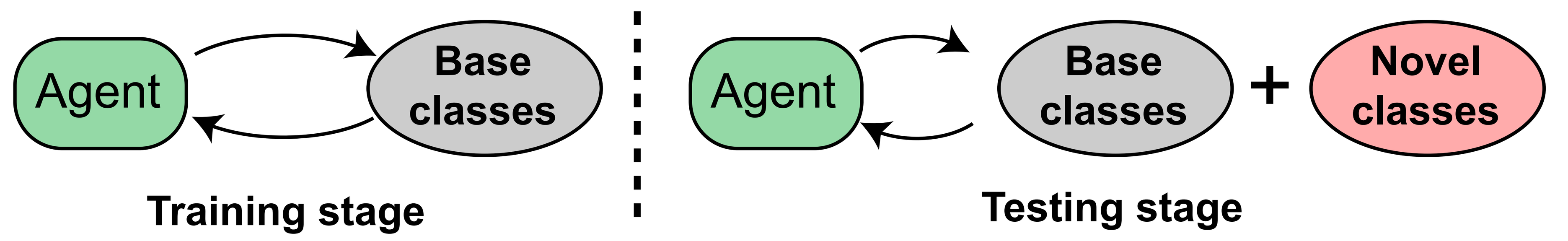
Performance of CLIP across all viewpoints within each category, reporting the mean, median, and maximum accuracy.

For different viewpoints, the discrepancy between the mean and maximum accuracy is an astonishing **40.1%!**



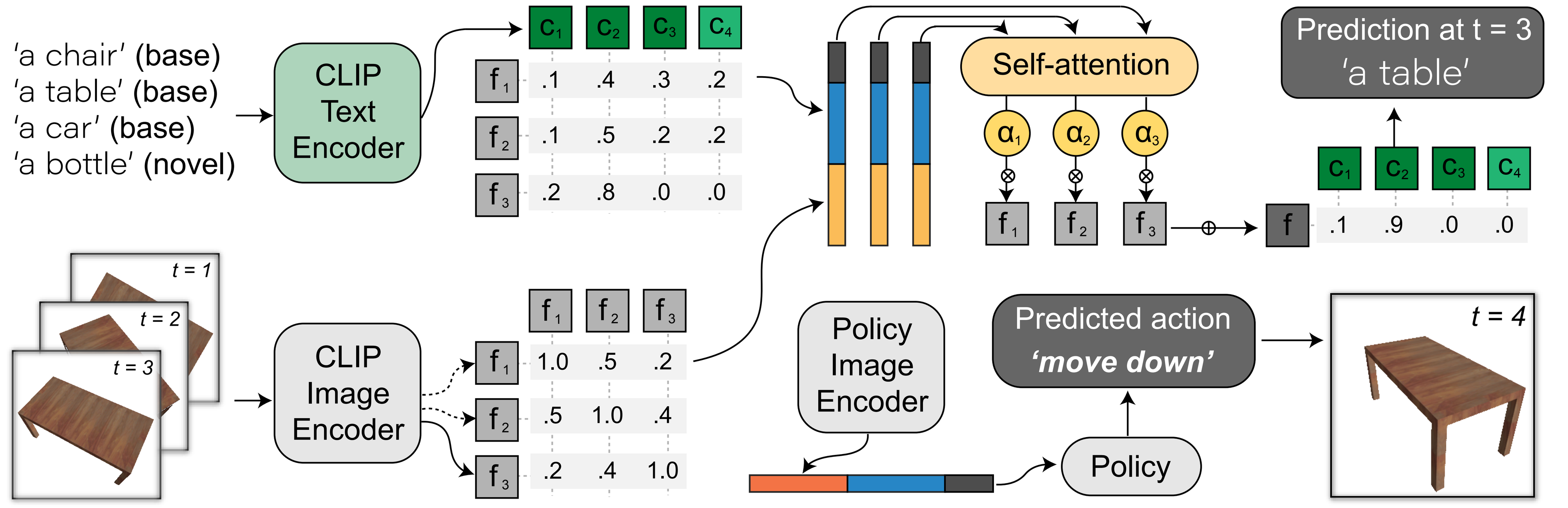
The average accuracy drop at three different occlusion levels are 3.1%, 4.0%, and 5.0%, respectively

Method



The class split setting of active open-vocabulary recognition.

During training: only base classes are presented to the agent.
During testing: the target is sampled from a broader open vocabulary.



Idea: Disentangle semantics from the policy and the fusion modules.
- use prediction confidence instead of semantic feature directly produced by CLIP models.

Our agent is trained with the PPO algorithm using the reward defined as the classification score belonging to the correct class.

Result

Model	Base/novel/open classes split													
	10/45/55						20/35/55						55/0/55	
	Base classes		Novel classes		Open classes		Base classes		Novel classes		Open classes		Base classes	
	top-1	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*
CLIP (ViT-B/32)	33.1	52.2	21.6	34.0	29.6	46.7	30.1	47.4	24.8	39.3	29.6	46.7	29.6	46.7
Ours	**60.6**	**81.3**	**36.6**	**55.1**	**53.3**	**73.4**	**57.9**	**76.8**	**47.8**	**69.0**	**56.6**	**75.7**	**59.2**	**78.8**

For split 10/45/55, the proposed method achieves **53.3%** accuracy for open classes, while the baseline CLIP model has 29.6%.