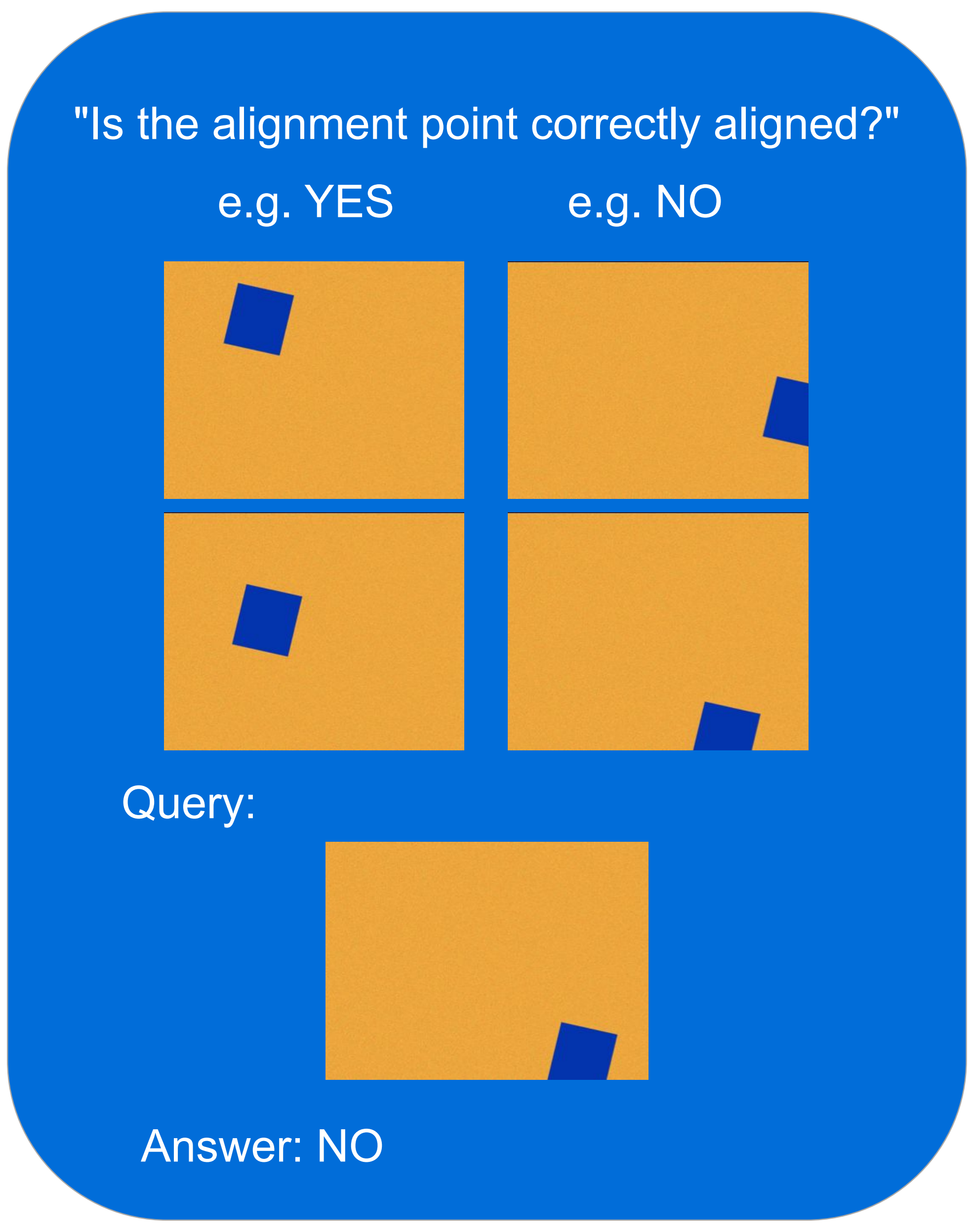
Can Vision Language Models Learn from Visual Demonstrations of Ambiguous Spatial Reasoning?

Groundlight

Bowen Zhao, Leo Parker Dirac, Paulina Varshavskaya

NEURAL INFORMATION PROCESSING SYSTEMS

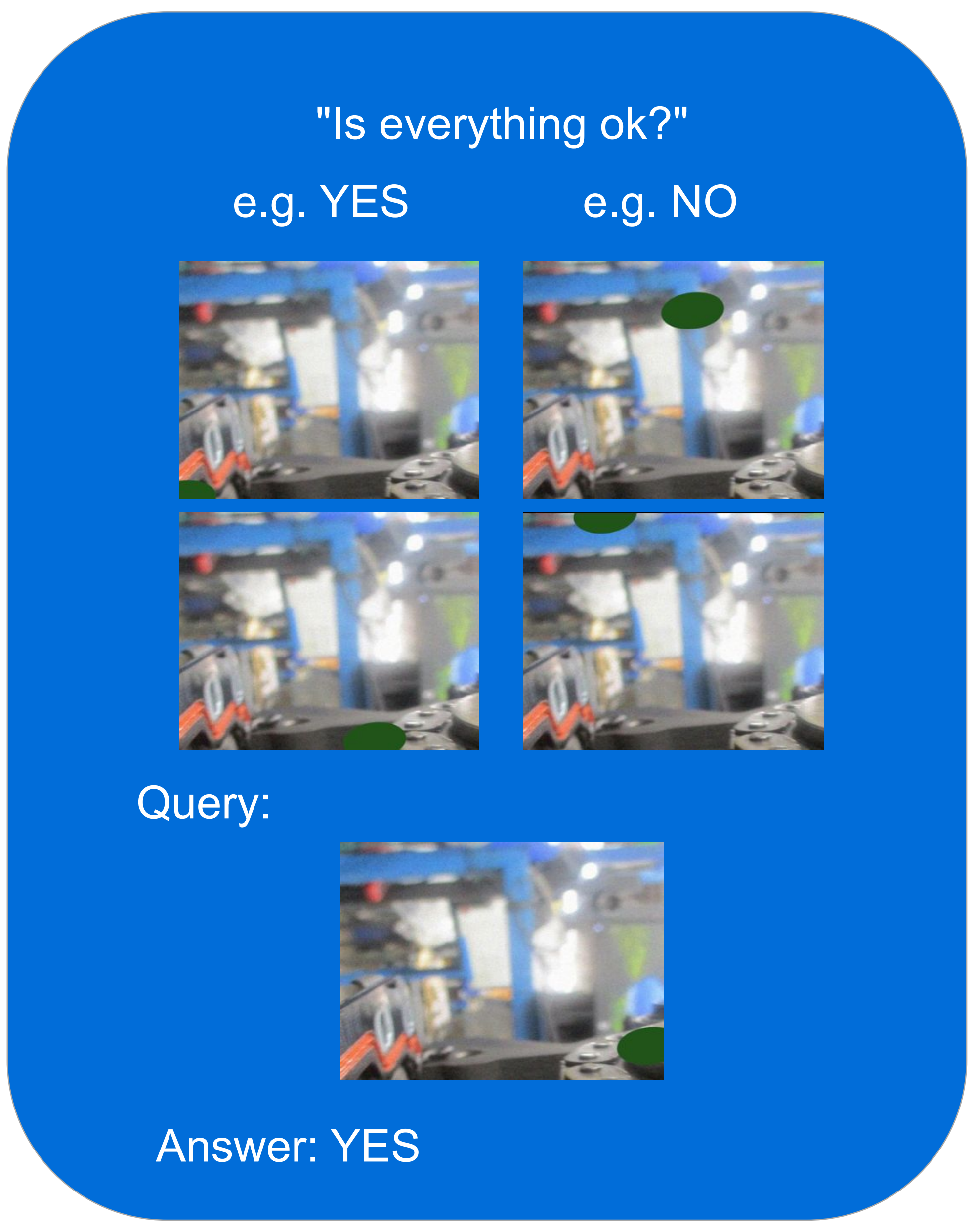


"Is the alignment point correctly aligned?"

e.g. YES e.g. NO

Query:

Answer: NO

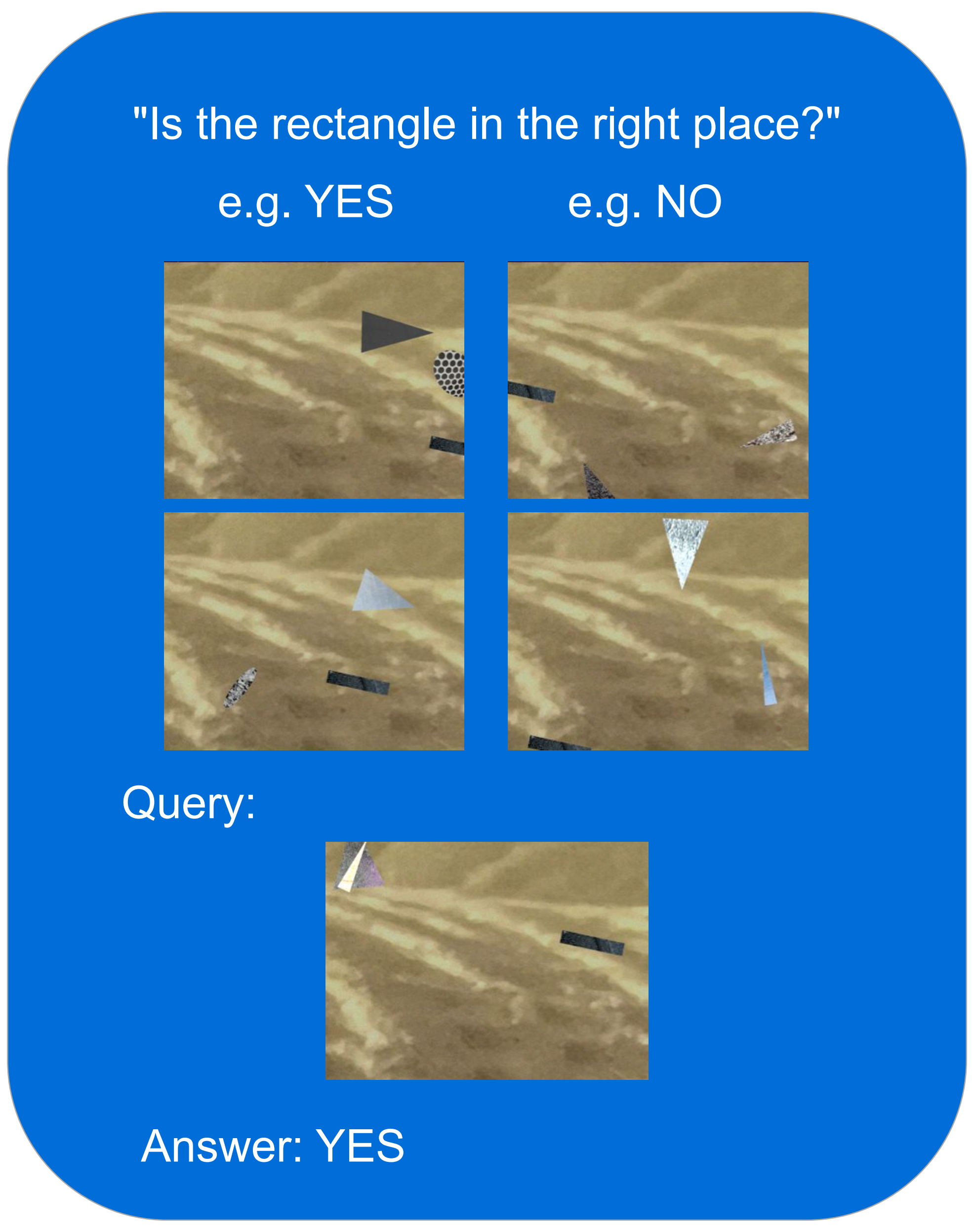


"Is everything ok?"

e.g. YES e.g. NO

Query:

Answer: YES



"Is the rectangle in the right place?"

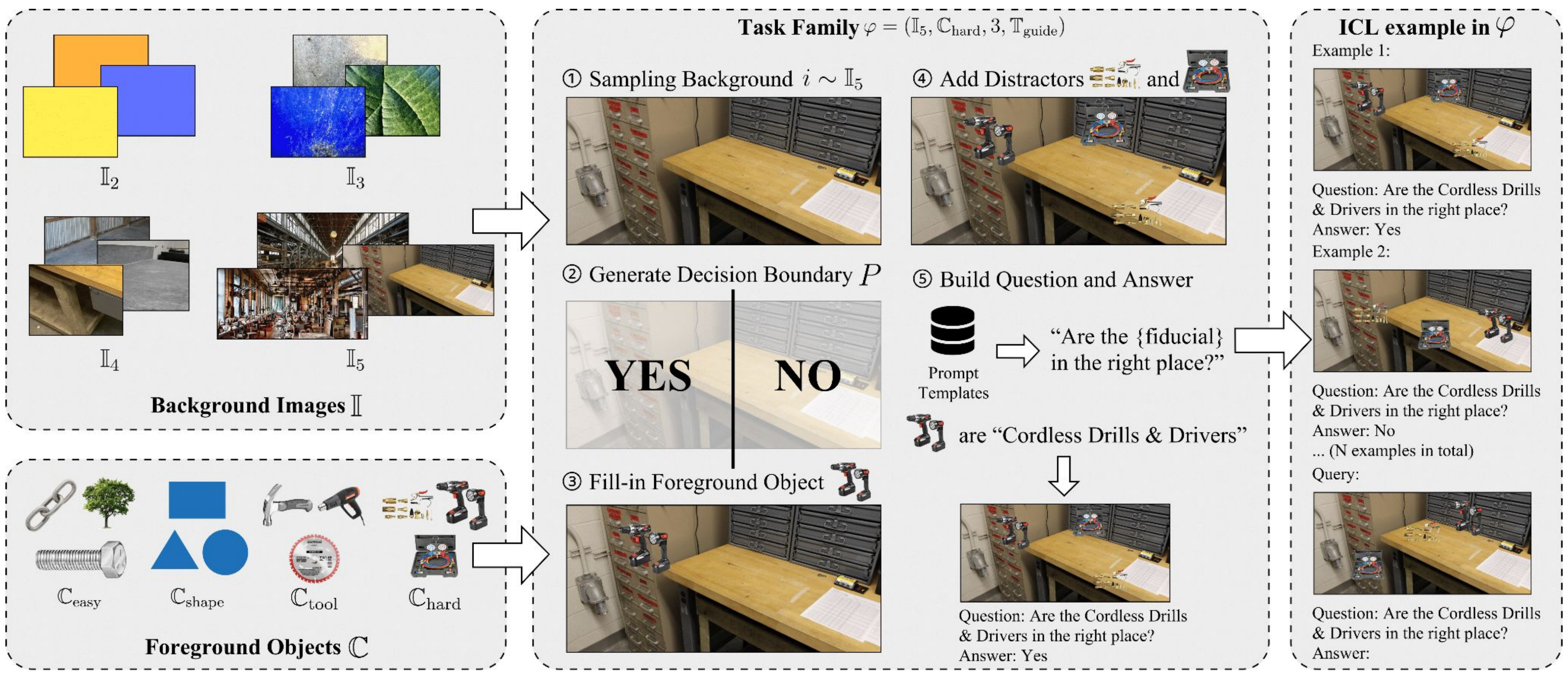
e.g. YES e.g. NO

Query:

Answer: YES

Motivation

- State-of-the-art vision language models (VLMs) are able to do **in-context learning (ICL)**, learning novel tasks at inference time.
- **Visuospatial knowledge is sometimes too ambiguous** to be explicitly described in words.
- AI-naive users in novel domains might assume background **domain-specific knowledge that VLMs are missing**.
- SVAT (Spatial Visual Ambiguity Tasks): synthesized datasets of varying difficulty, using visual demonstrations with ambiguous text to help VLM inference.



Background Images $\mathbb{I}$: $\mathbb{I}_2$, $\mathbb{I}_3$, $\mathbb{I}_4$, $\mathbb{I}_5$

Foreground Objects $\mathbb{C}$: $\mathbb{C}_{\text{easy}}$, $\mathbb{C}_{\text{shape}}$, $\mathbb{C}_{\text{tool}}$, $\mathbb{C}_{\text{hard}}$

Task Family $\varphi = (\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$

① Sampling Background $i \sim \mathbb{I}_5$

② Generate Decision Boundary P

YES NO

③ Fill-in Foreground Object

④ Add Distractors and

⑤ Build Question and Answer

Prompt Templates

"Are the {fiducial} in the right place?"

are "Cordless Drills & Drivers"

Question: Are the Cordless Drills & Drivers in the right place?
Answer: Yes

ICL example in φ

Example 1:

Question: Are the Cordless Drills & Drivers in the right place?
Answer: Yes

Example 2:

Question: Are the Cordless Drills & Drivers in the right place?
Answer: No

... (N examples in total)

Query:

Question: Are the Cordless Drills & Drivers in the right place?
Answer:

Method

Overview

SVAT tasks aim at asking VLMs if a foreground object is at a *correct* location on a background image. What makes SVAT challenging is that the **correct location is not explicitly described in words** but must be inferred by models using the **in-context visual demonstrations**. **Task families in SVAT can be varied** based on the provided texts, complexity of foreground object or background image, and the number of distractors.

Generating SVAT Datasets

Each example in a SVAT dataset contains:

- A natural language question *t*,
- An image *v*, formed by a number of foreground objects *o* and a background image *i*,
- A binary answer label *y*.

Each trian/test instance in SVAT contains four demonstration examples and a querying example, sampled from same distribution. To control the difficulty level of the task families in SVAT, the sampling process is parameterized by the following factors $\varphi = (\mathbb{I}, \mathbb{C}, M, \mathbb{T})$

- $\mathbb{I}$: the background image categories,
- $\mathbb{C}$: the foreground object categories,
- M: the number of distracting foreground objects,
- $\mathbb{T}$: the sampling pool of textual questions.

Curriculum Learning with SVAT

As task families in SVAT naturally forms different difficulty levels, one could use SVAT to do curriculum learning (CL) by training VLMs from easier to harder tasks. We define four different two-step CL strategies along each parameter for any $\varphi_2 = (\mathbb{I}_i, \mathbb{C}_i, M_i, \mathbb{T}_i)$:

$$\mathcal{C}^{\mathbb{I}}(\varphi_2) = (\mathbb{E}_{(\mathbb{I}_1, \mathbb{C}_i, M, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2}), \mathcal{C}^{\mathbb{C}}(\varphi_2) = (\mathbb{E}_{(\mathbb{I}_i, \mathbb{C}_{\text{easy}}, M, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2})$$
$$\mathcal{C}^{M}(\varphi_2) = (\mathbb{E}_{(\mathbb{I}_i, \mathbb{C}_i, 1, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2}), \mathcal{C}^{\text{all}}(\varphi_2) = (\mathbb{E}_{(\mathbb{I}_1, \mathbb{C}_{\text{easy}}, 1, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2})$$

Experiment Results

Table 1: Main results of VLMs' performance on SVAT. M denotes the number of objects per example, and the second row on the header indicates the foreground object category set $\mathbb{C}$ in task family φ. The complexity of the background images is fixed at level 5 ($\mathbb{I}_5$). Accuracy significantly better than random guessing is in green , and each task's best model's result is in **bold**.

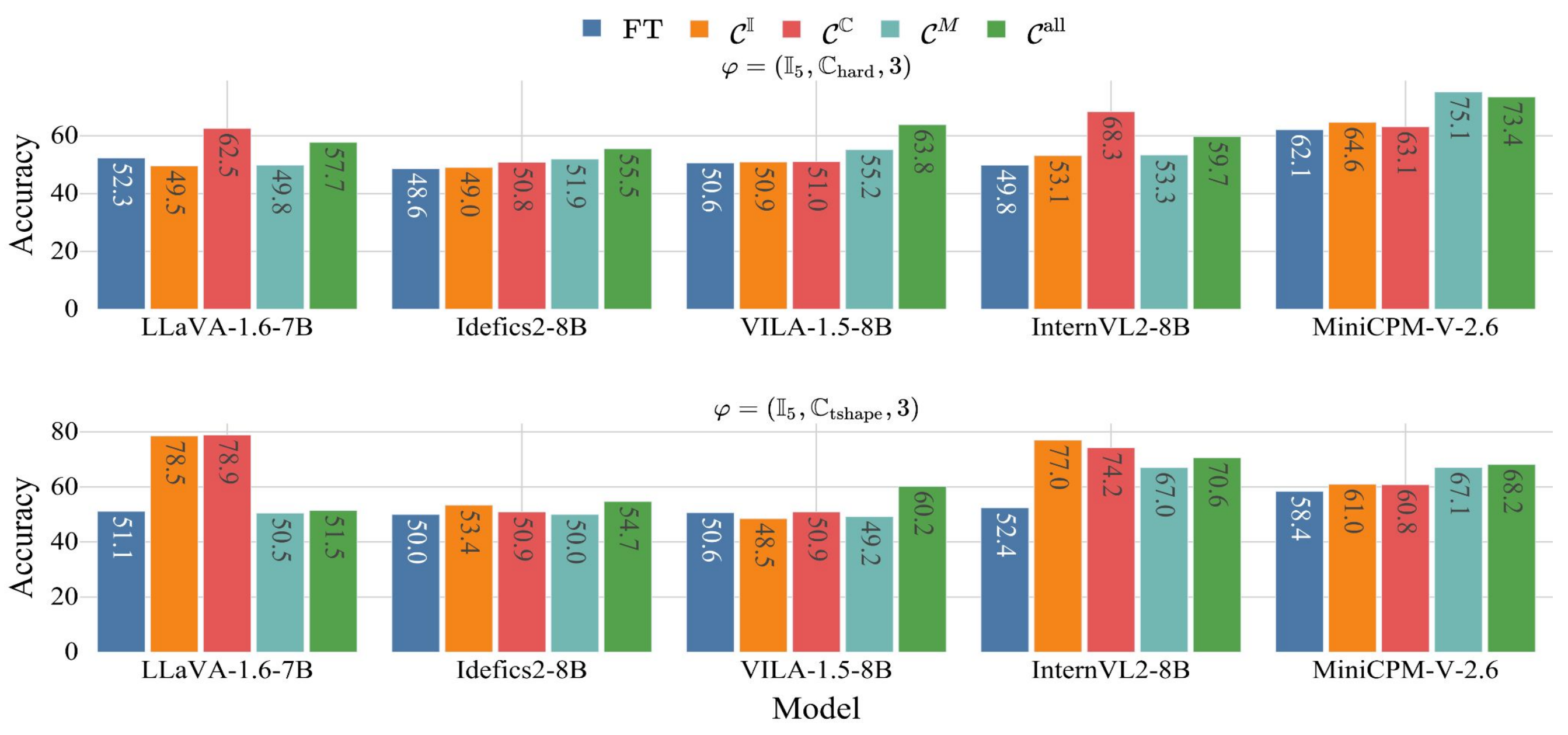
		$M = 1, \mathbb{T} = \mathbb{T}_{\text{none}}$ (no distractors, useless text)					$M = 3, \mathbb{T} = \mathbb{T}_{\text{guide}}$ (distractors, text names objects)				
Category	Model	easy	shape	tshape	tool	hard	easy	shape	tshape	tool	hard
Zero-shot	LLaVA-1.6-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Idefics2-8B	50.4	49.6	49.8	50.7	52.3	51.1	53.7	**52.7**	49.2	49.7
	VILA-1.5-8B	49.3	48.9	49.9	47.6	47.7	49.8	52.4	51.8	**52.3**	48.7
	InternVL2-8B	46.8	49.9	48.2	47.7	46.1	50.2	**54.0**	49.3	49.8	50.1
	MiniCPM-V-2.6	**59.5**	**57.3**	**56.5**	**58.0**	**55.0**	**52.6**	51.9	51.1	50.8	**50.4**
Finetuned (FT)	LLaVA-1.6-7B	52.8	47.9	52.0	49.2	49.3	**80.3**	53.4	51.1	49.3	52.3
	Idefics2-8B	65.6	53.9	51.2	54.6	62.1	49.0	54.1	50.0	49.7	48.6
	VILA-1.5-8B	72.9	49.9	49.9	**77.3**	66.6	49.1	54.5	50.6	49.6	50.6
	InternVL2-8B	70.4	74.7	55.0	52.9	49.8	77.9	**76.9**	52.4	**65.6**	50.9
	MiniCPM-V-2.6	**73.4**	**80.0**	**68.6**	74.2	**71.8**	52.8	72.0	**58.4**	52.2	**62.1**

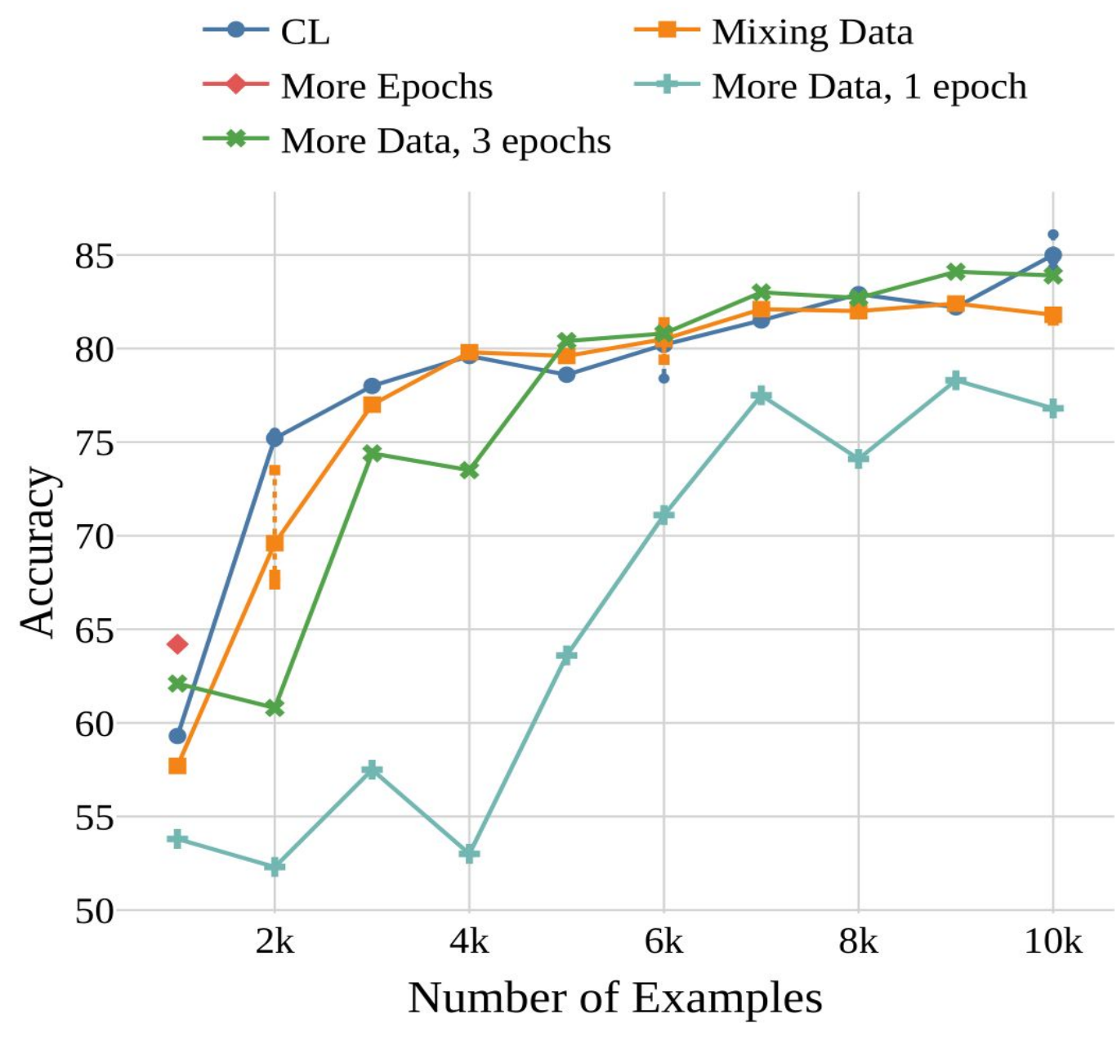
Main findings:

(1) State-of-the-art VLMs struggle at SVAT tasks in zero-shot settings regardless of their pretraining and instruction-tuning recipes. MiniCPM is the only VLM that achieves significantly better performance than random guessing.

(2) Directly finetuning VLMs on SVAT datasets improve their performance, but the gains become minimal when the foreground objects are complex while there are distractors in the images.

(3) Some VLMs (LLaVA, Idefics, InternVL) are better at the $(M = 3, \mathbb{T} = \mathbb{T}_{\text{guide}})$task. These models can be more sensitive to textual queries rather than images at inference time.





Curriculum learning analysis:

(1) CL effectively improves model performance on varied task families in SVAT across different VLMs (figure on left), suggesting that knowledge in task families of SVAT can be transferred.

(2) When training VLMs on SVAT with different number of data (figure on right), CL is the most robust and data-efficient strategy compared to single-task or mixed-data baselines.

Link to paper