ICML International Conference On Machine Learning
POSTECH POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Bring My Cup!

Personalizing Vision-Language-Action Models with Visual Attentive Prompting

Sangoh Lee · Sangwoo Mo · Wook-Shin Han

POSTECH · ICML 2026

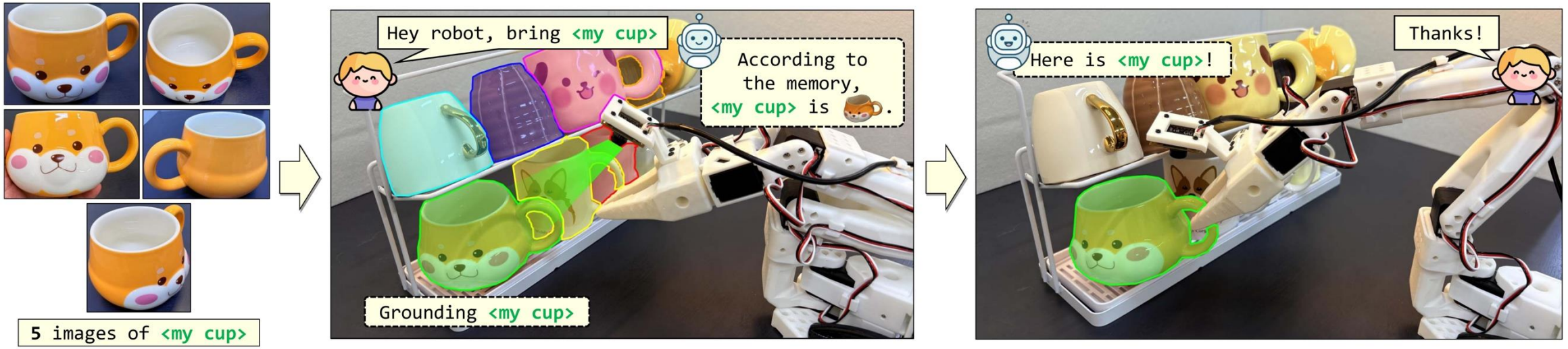
Paper: https://openreview.net/pdf?id=fm6Z3wfTae
Project Page: https://vap-project.github.io
Code: https://github.com/Leesangoh/VAP

Hey robot, bring my cup.

VLAs follow generic instructions reliably, but collapse on personal references.

- Same-category lookalikes look almost identical to the model.
- Language alone (“my cup”) collapses to a category prior.
- Per-object fine-tuning is impractical at deployment.

The robot needs a signal that language cannot carry.



VAP grounds a personal target from a few reference photos, then visually prompts a frozen VLA.

Language alone cannot resolve identity

01

Generic instruction

“Bring a cup.”

Category-level recognition is sufficient. Any cup in view is an acceptable answer, and a frozen VLA already handles this.

02

Personalized instruction

“Bring my cup.”

Two cups of the same category are indistinguishable by text. The discriminative signal lives in pixels, not in words.

Reference images carry the discriminative cues that language cannot.

Visual Attentive Prompting

A training-free, input-side adapter for frozen VLAs.

01

Training-free

Frozen VLA, no fine-tuning.

02

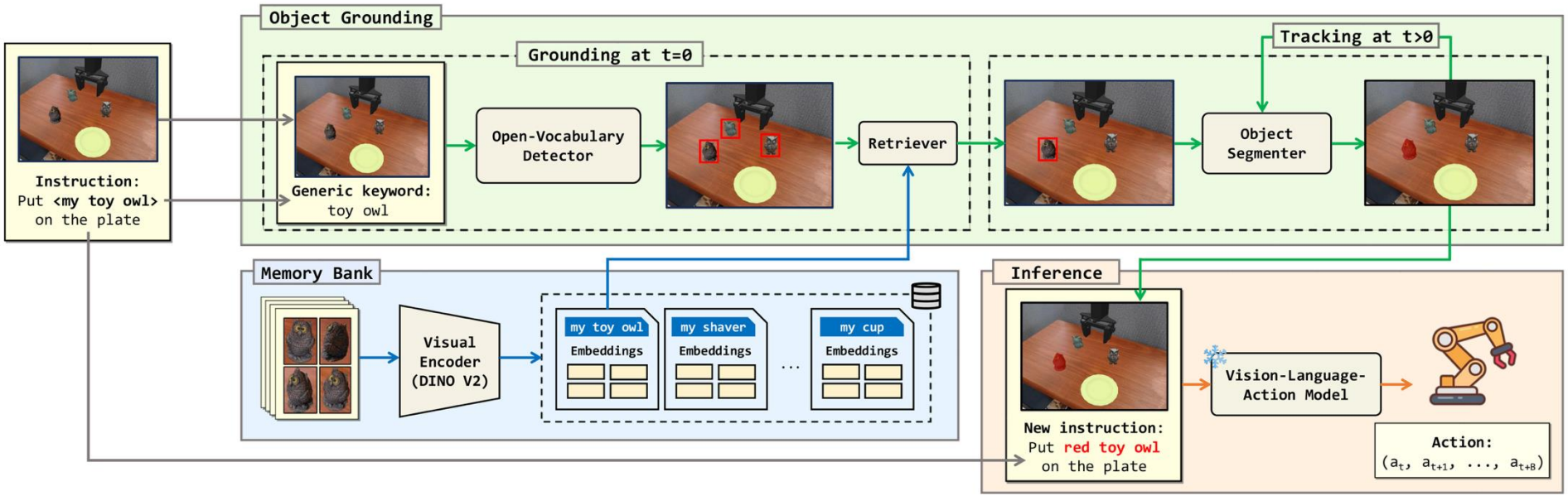
Input-side

Only the (image, instruction) pair.

03

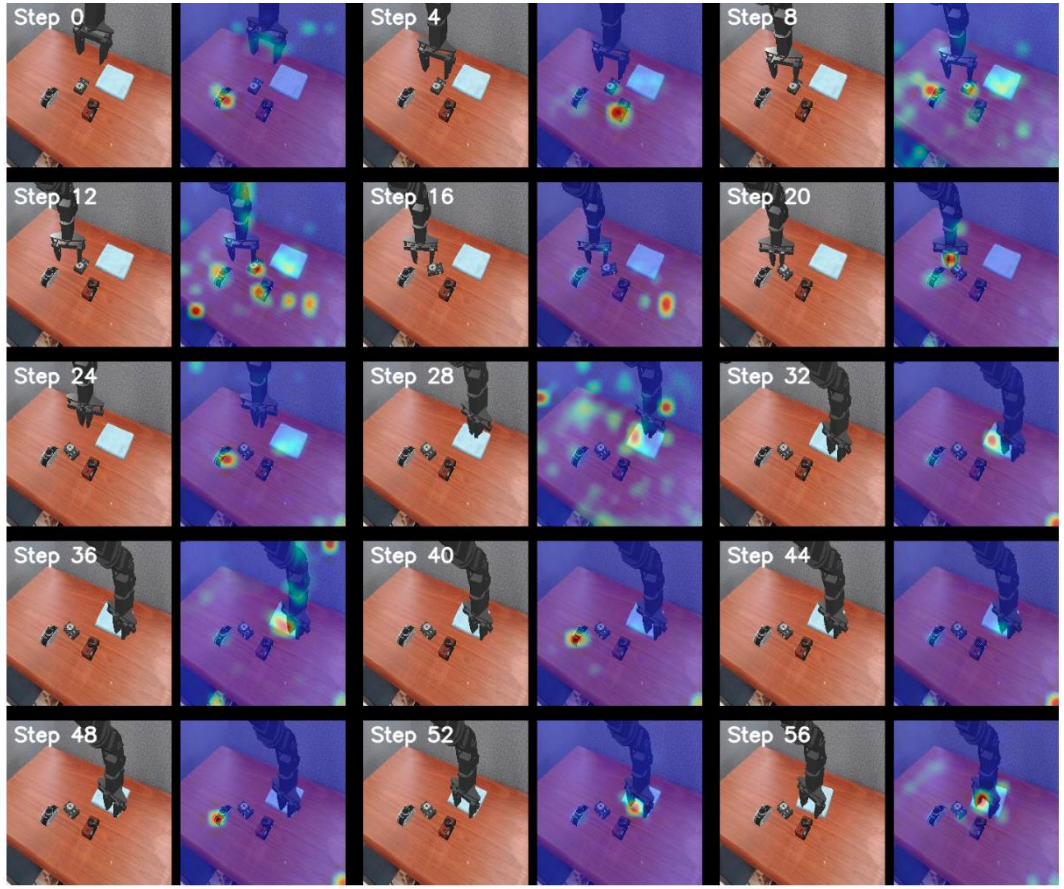
Modular

Swappable perception backbones.



Reference images → Grounding → Visual Prompting → frozen VLA.

Mask and rewrite as a single binding signal



Static personalization (token learning): attention drifts over time, dissolving the instance signal.

Mask · pixel-level anchor

Tells the policy where the personal target is in the image.

Rewrite · language pointer

Tells the policy which anchor to attend to (e.g., “my cup” → “the red cup”).

Remove either signal → policy reverts to its category prior.

Mask-only **7.3%** · Rewrite-only **15.8%** · Both **62.7%**

Success rate on Personalized-SIMPLER (Google Robot).

Personalization benchmarks

Same-category lookalikes as distractors; all test objects are unseen during training.

Experimental results

Same frozen VLA backbone (π_0 / $\pi_{0.5}$). Success rate on representative tasks.

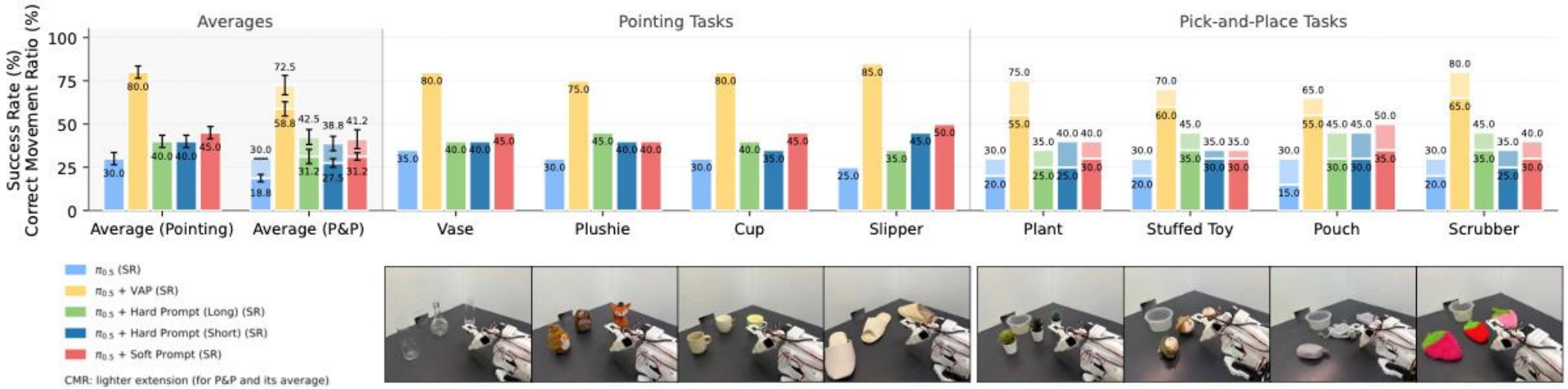
Method	Task 1: Pick my pen holder		Task 2: Move my bottle near coke can	
	CMR (%)	SR (%)	CMR (%)	SR (%)
Track 1: Visual Matching with Unseen Personalization Objects				
π_0	10.5	8.5	50.4	48.6
π_0 + Soft Prompt	33.1	24.0	60.0	51.7
π_0 + Hard Prompt (Short)	11.0	8.9	66.7	57.5
π_0 + Hard Prompt (Long)	11.2	9.1	64.5	58.8
π_0 + VAP	**89.2**	**60.3**	**83.2**	**75.0**
Track 2: Variant Aggregation with Unseen Personalization Objects				
π_0	17.4	10.2	49.2	41.3
π_0 + Soft Prompt	31.8	26.4	57.7	44.7
π_0 + Hard Prompt (Short)	21.6	14.4	71.5	54.2
π_0 + Hard Prompt (Long)	17.3	12.6	67.7	51.3
π_0 + VAP	**87.3**	**58.2**	**72.4**	**62.5**

Method	Task 3: Pick shaver		Task 4: Put camera		Task 5: Put owl		Task 6: Put straw cup	
	CMR (%)	SR (%)	CMR (%)	SR (%)	CMR (%)	SR (%)	CMR (%)	SR (%)
π_0	0.0	0.0	34.2	31.7	35.2	30.3	80.6	27.8
π_0 + Soft Prompt	25.0	8.3	54.2	41.7	79.2	62.5	87.5	29.2
π_0 + Hard Prompt (Short)	0.0	0.0	50.0	25.0	37.5	33.3	83.3	29.2
π_0 + Hard Prompt (Long)	0.0	0.0	53.2	27.9	36.0	31.6	83.4	30.1
π_0 + VAP	**82.9**	**71.3**	**92.1**	**92.1**	**100.0**	**95.0**	**100.0**	**75.6**

Experimental results

Same frozen VLA backbone (π_0 / $\pi_{0.5}$). Success rate on representative tasks.

Method	Task 1: Leather bag	Task 2: Shoe	Task 3: Cat figurine	Task 4: Miniature house	Task 5: Cup
$\pi_{0.5}$	33.6	30.4	35.2	29.6	40.4
$\pi_{0.5}$ + Soft Prompt	51.2	40.8	44.4	39.6	54.8
$\pi_{0.5}$ + Hard Prompt (Short)	52.4	37.6	44.0	38.8	51.6
$\pi_{0.5}$ + Hard Prompt (Long)	52.0	39.6	42.8	39.6	50.0
$\pi_{0.5}$ + VAP	**89.2**	**54.0**	**51.6**	**52.4**	**60.8**



Modular by construction

Each perception block is a swappable foundation model. Improvements drop in without policy retraining.

01

Open-vocab detection

CURRENT

Grounding DINO

DROP-IN UPGRADE

Future open-set detectors

02

Reference matching

CURRENT

DINOv2

DROP-IN UPGRADE

Better visual encoders

03

Spatio-temporal track

CURRENT

SAM2

DROP-IN UPGRADE

Next-gen segmenters

Foundation models improve → VAP inherits the gains directly.

Takeaways

01 Training-free personalization

Register a personal object with a few reference photos. No fine-tuning, no retraining.

02 Mask + rewrite as a single binding signal

The two interventions are not separable boosts; neither works alone.

03 Modular and future-proof

Perception backbones are swappable; foundation-model progress transfers directly.

Future: mobile manipulation · user-specific affordances · long-term personal memory.

Thank you!

Paper: https://openreview.net/pdf?id=fm6Z3wfTae
Project Page: https://vap-project.github.io
Code: https://github.com/Leesangoh/VAP