CVPR JUNE 3-7, 2026 DENVER COLORADO

June 03 – 07, 2026 | Denver, CO, USA

Embodied AI Workshop @ CVPR 2026

ARNOLD CHALLENGE

2nd Place

Learning Ordered Action for Continuous-State Manipulation in Vision-Language-Action Models

Challenge Team: **Kairoba**

Sungyong Park
Soongsil University,

Heewon Kim
Kairoba Inc.

Reality Lab
Soongsil University

kairoba

Introduction

Reality Lab
Soongsil University

- VLAs generalize well to categorical instructions: “pick the object”, “open the drawer”.
- However, many real commands are continuous: “pick up the object 30cm”, “open the drawer 50%”.

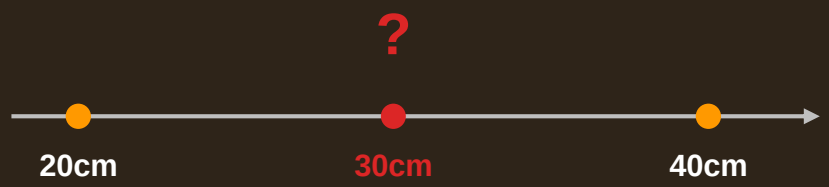
Categorical (what to do)

“Pick the object”

“Open the drawer”

Strong generalization ✓

Continuous (how much)

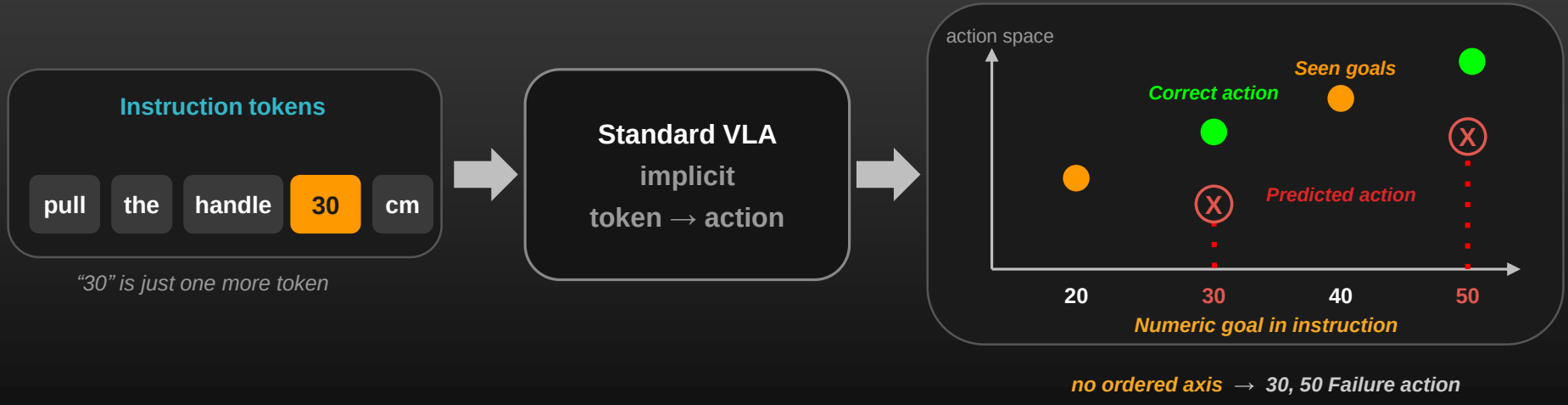


seen 20 & 40 → what about 30 ?

Can VLAs generalize to unseen numeric goals?

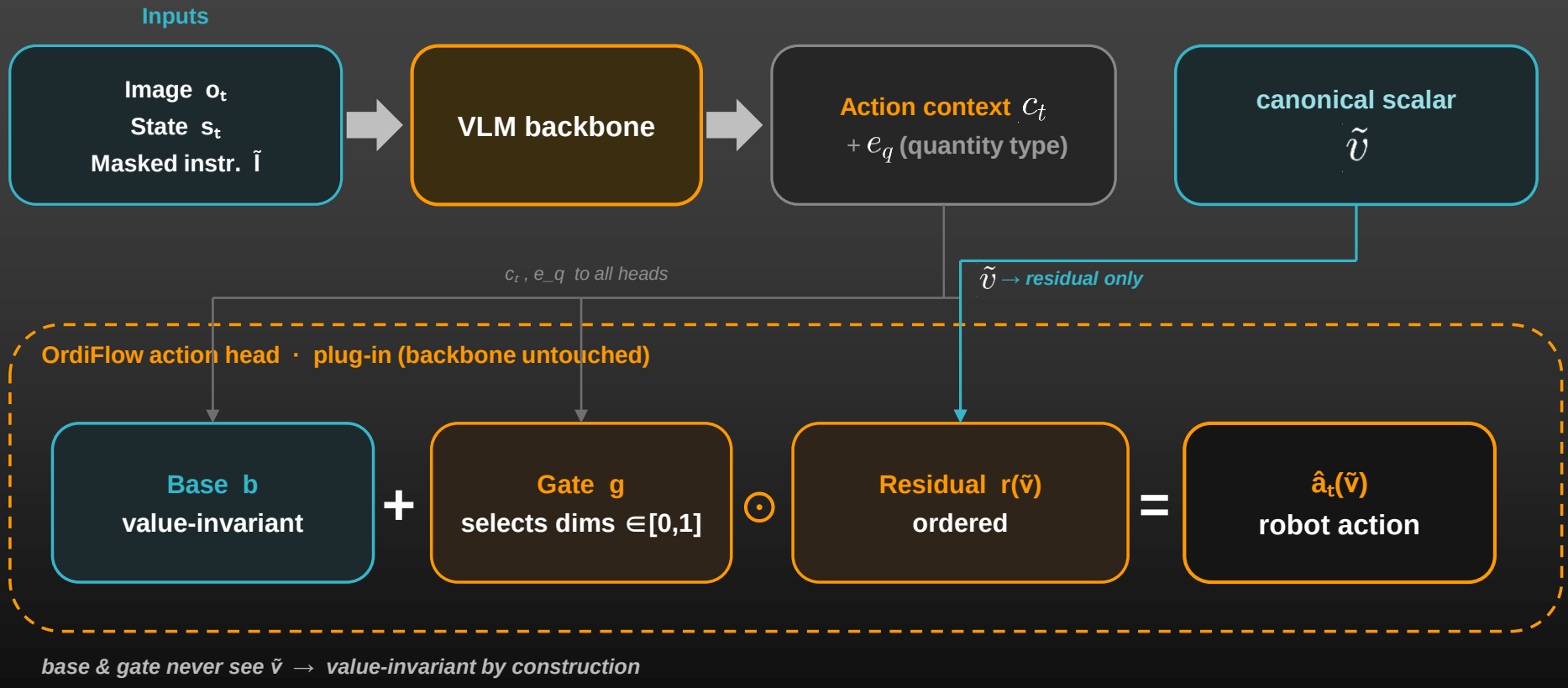
Problem Statement

- Standard VLAs learn numeric goals only through instruction tokens.
- The numeric structure is not explicitly grounded in the action space.
- Limited numeric supervision weakens numeric grounding, limiting generalization to unseen values.



Method: a Plug-in Action Head (OrdiFlow)

Reality Lab
Soongsil University



Method: Canonical Scalar

- Parse the instruction into a **numeric goal, a type, and a masked instruction** with a placeholder.
- This **routes the number through a separate pathway**, instead of fusing it into the token stream.
- Normalize per type; **never clamp** → out-of-range commands stay valid.

"Adjust the top-left drawer 50% open"

Parse (GPT)

$$(r,\ q,\ \tilde{l}) = \mathrm{Parse}(l)$$

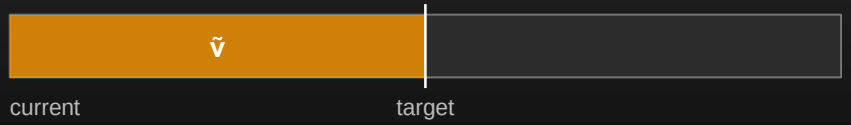
r = 50 q = ratio

$\tilde{l}$ = "… drawer [MASK] open"

canonical scalar $\tilde{v}$

$$v = \max(s^{\mathrm{tgt}} - s^{\mathrm{cur}},\ 0) \qquad \tilde{v} = v \,/\, v_{\mathrm{ref}}^{(q)}$$

remaining motion = v



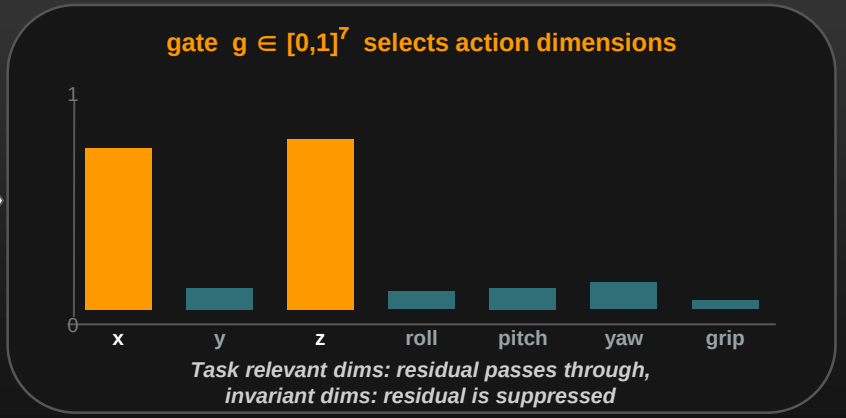
Method: Base + Gated Residual

Reality Lab
Soongsil University

We split the predicted action into a value-invariant base and a value-conditioned residual:

$$\hat{a}_t(\tilde{v}) = b_\phi(c_t, e_q) + g_\phi(c_t, e_q) \odot r_\psi(\tilde{v};\, c_t, e_q)$$

- b_ϕ **Value-invariant base — approach, grasp, contact setup; does not see the numeric value.**
- g_ϕ **Per-dimension gate ∈ [0,1] — selects which action dimensions the command may modify.**
- r_ψ **Ordered, value-conditioned residual — the only term that scales with the number.**



Method: Ordered Value-Conditioned Residual

Reality Lab
Soongsil University

$$r_\psi(\tilde{v};\, c_t, e_q) \;=\; \int_0^{\tilde{v}} \underbrace{\rho_\psi(u;\, c_t, e_q)}_{\text{magnitude}}\; \underbrace{D_\phi(u;\, c_t, e_q)}_{\text{direction}}\; du$$

$$\rho_\psi(u) \;=\; \epsilon_\rho + \mathrm{softplus}(g_\psi([\gamma(u); c_t; e_q])) \;>\; 0$$

positive density ⇒ accumulated magnitude is non-decreasing (ordered)

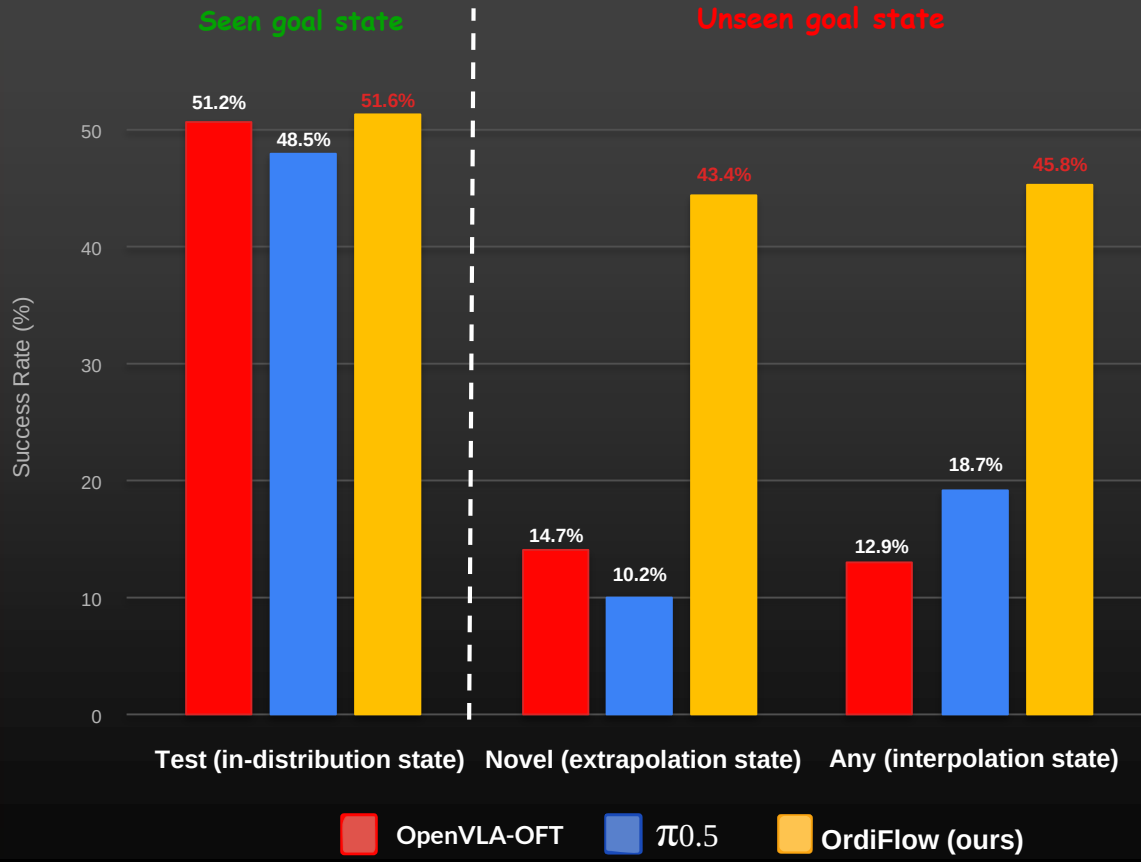
- **Positivity** → ordered response, guaranteed — the structural bias a token head lacks.
- **Controlled extrapolation**: clip queries to the training boundary → residual continues linearly.

Results

Reality Lab
Soongsil University

2nd Place ★

Rank	Participant team	SR (↑)
1	FiveAges (FAM)	0.57
2	Kairoba	0.56
3	RealityLab	0.49
4	EBDAI	0.45
5	Fun Guy (Fusion(SD&PC))	0.39
6	FightOn	0.36
7	larr (final)	0.32
8	Windboy (DT1)	0.25
9	MilkyWay	0.25
10	MCC-EAI	0.25
11	KONGBAI (test1)	0.23
12	Host_31221_Team B	0.22



Conclusion

Reality Lab
Soongsil University

- Standard VLAs treat numeric goals as instruction tokens.
- This makes value-to-action generalization difficult under sparse numeric supervision.
- We separates numeric values from the token stream and models them with an ordered action head.
- This leads to stronger unseen-goal generalization in continuous-state manipulation.

CVPR JUNE 3-7, 2026 DENVER COLORADO

June 03 – 07, 2026 | Denver, CO, USA

Reality Lab
Soongsil University

Thank You

For more details and any questions, please visit our poster session.

12:00PM-1:30PM MDT - Exhibit Hall A Boards 262-276

kairoba