

SCALE: Self-uncertainty Conditioned Adaptive Looking and Execution for Vision-Language-Action Models



Hyeonbeom Choi*



Daechul Ahn*



Youhan Lee



Taewook Kang



Seongwon Cho



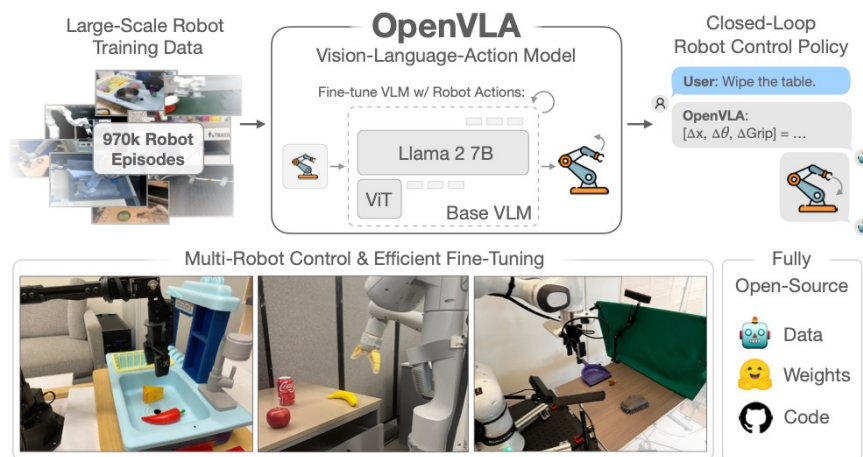
Jonghyun Choi



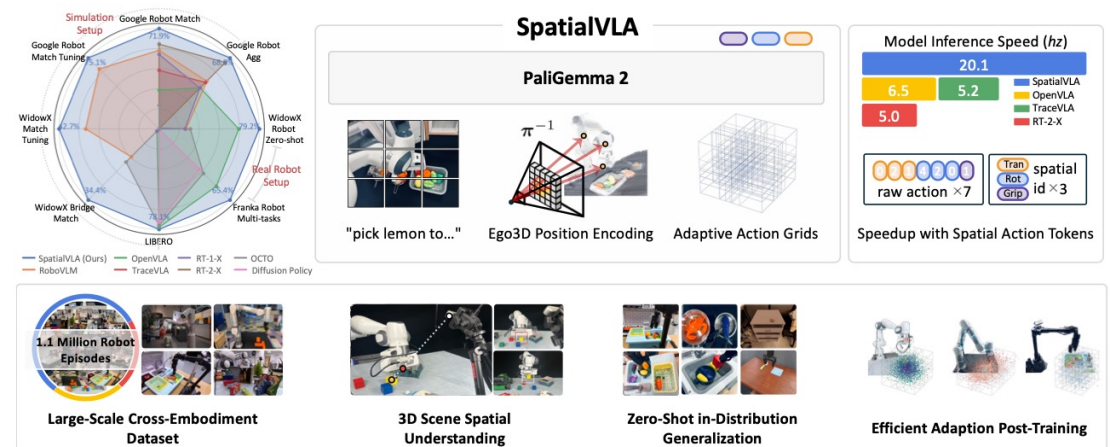
SEOUL
NATIONAL
UNIVERSITY

VLAAs enable general purpose robotic control

- Emerged as a predominant robotic control paradigm, seamlessly leveraging pretrained VLMs with minimal architecture modification^[1,2]



[1]



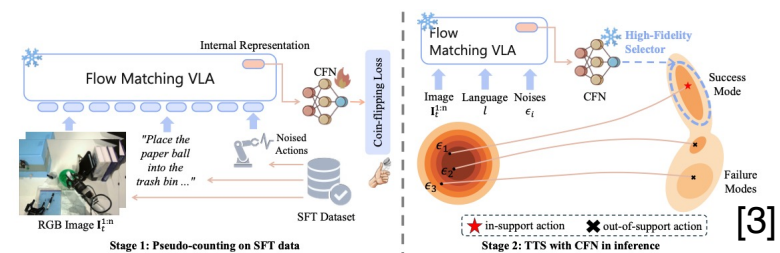
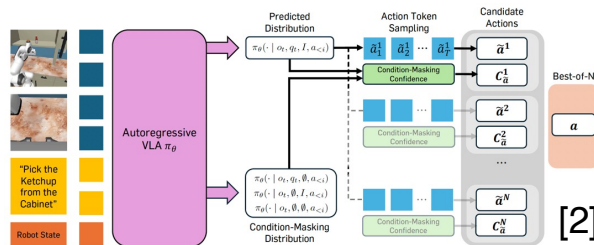
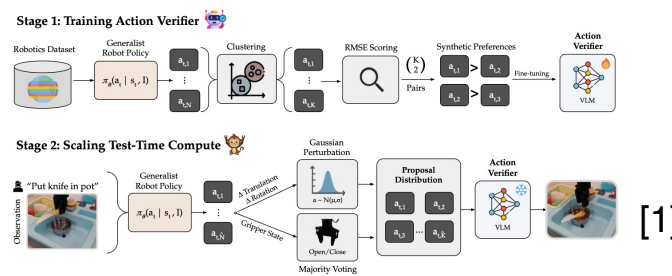
[2]

[1] Kim et al., OpenVLA: An Open-Source Vision-Language-Action Model, CoRL 2024.

[2] Qu et al., SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model, RSS 2025.

Recent research focuses towards enhancing VLA adaptability at *test time*

- A prominent strategy for enhancing test-time adaptability for VLAs is Test-Time Scaling (TTS)^[1,2,3]



[1] Kwok et al., RoboMonkey: Scaling Test-Time Sampling and Verification for Vision-Language-Action Models, CoRL 2025.

[2] Jang et al., Verifier-free Test-Time Sampling for Vision-Language-Action Models, ICLR 2026.

[3] Yang et al., Steering Vision-Language-Action Models as Anti Exploration: A Test-Time Scaling Approach, arXiv'25.12.

Recent research focuses towards enhancing VLA adaptability at *test time*

- A prominent strategy for enhancing test-time adaptability for VLAs is

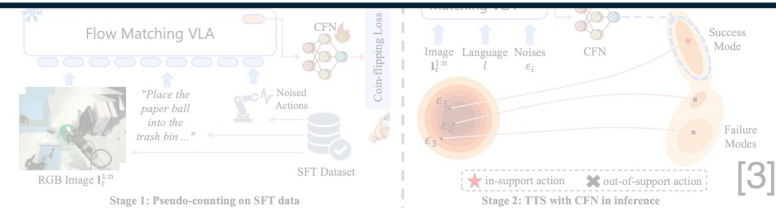
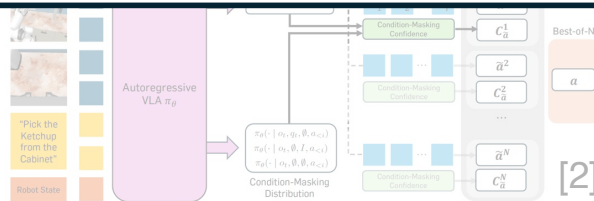
But, TTS methods for VLA have limitations:

(1) High computational cost

- Multiple forward passes, additional training required

(2) Fixed visual representation

- Fail to adapt perception under visual ambiguity



[1] Kwok et al., RoboMonkey: Scaling Test-Time Sampling and Verification for Vision-Language-Action Models, CoRL 2025.

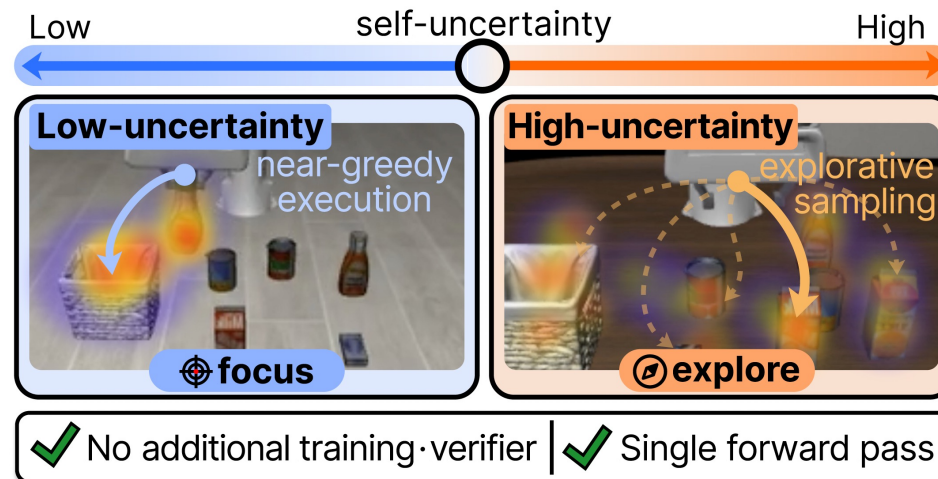
[2] Jang et al., Verifier-free Test-Time Sampling for Vision-Language-Action Models, ICLR 2026.

[3] Yang et al., Steering Vision-Language-Action Models as Anti Exploration: A Test-Time Scaling Approach, arXiv'25.12.

To address these limitations, we propose **SCALE: Uncertainty-Driven Adaptive Inference**

- Inspired by *uncertainty-driven exploration*^[1,2], we propose SCALE that modulates both perception & action via ‘*self-uncertainty*’.

“No additional training / verifier / single forward pass”



[1] Friston et al., Active inference and learning, Neuroscience & Biobehavioral Reviews, 2016.

[2] Schwartenbeck et al., Computational mechanisms of curiosity and goal-directed exploration, eLife, 2019.

Then, how do we define 'self-uncertainty'?

Then, how do we define ‘self-uncertainty’?

- Recent LLM work estimates self-uncertainty by how close the predicted distribution is to uniform, *i.e.*, full ambiguity^[1]

[1] Kang et al., Scalable best-of-n selection for large language models via self-certainty, NeurIPS, 2025.

Then, how do we define ‘self-uncertainty’?

- Recent LLM work estimates self-uncertainty by how close the predicted distribution is to uniform, *i.e.*, full ambiguity^[1]
- But VLAs commit to the top-1 action irreversibly via greedy decoding, so a proper metric needs to quantify top-1 decisiveness, not just distributional uncertainty

[1] Kang et al., Scalable best-of-n selection for large language models via self-certainty, NeurIPS, 2025.

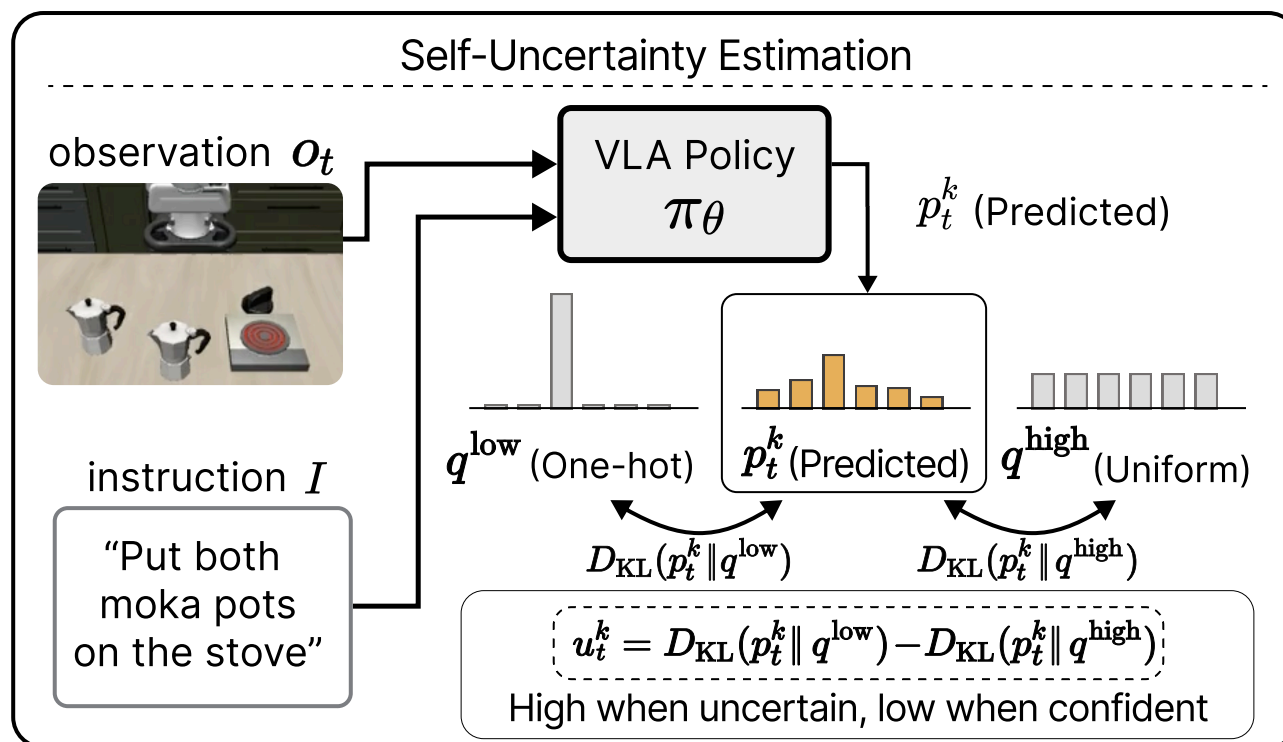
Then, how do we define ‘self-uncertainty’?

- Recent LLM work estimates self-uncertainty by how close the predicted distribution is to uniform, *i.e.*, full ambiguity^[1]
- But VLAs commit to the top-1 action irreversibly via greedy decoding, so a proper metric needs to quantify top-1 decisiveness, not just distributional uncertainty
- To capture both ambiguity and decisiveness, we measure where the predicted action distribution lies between two reference distributions:
 - full certainty (one-hot), q^{low} vs. full ambiguity (uniform), q^{high}

$$u_t^k = D_{\text{KL}}(p_t^k \parallel q^{\text{low}}) - D_{\text{KL}}(p_t^k \parallel q^{\text{high}})$$

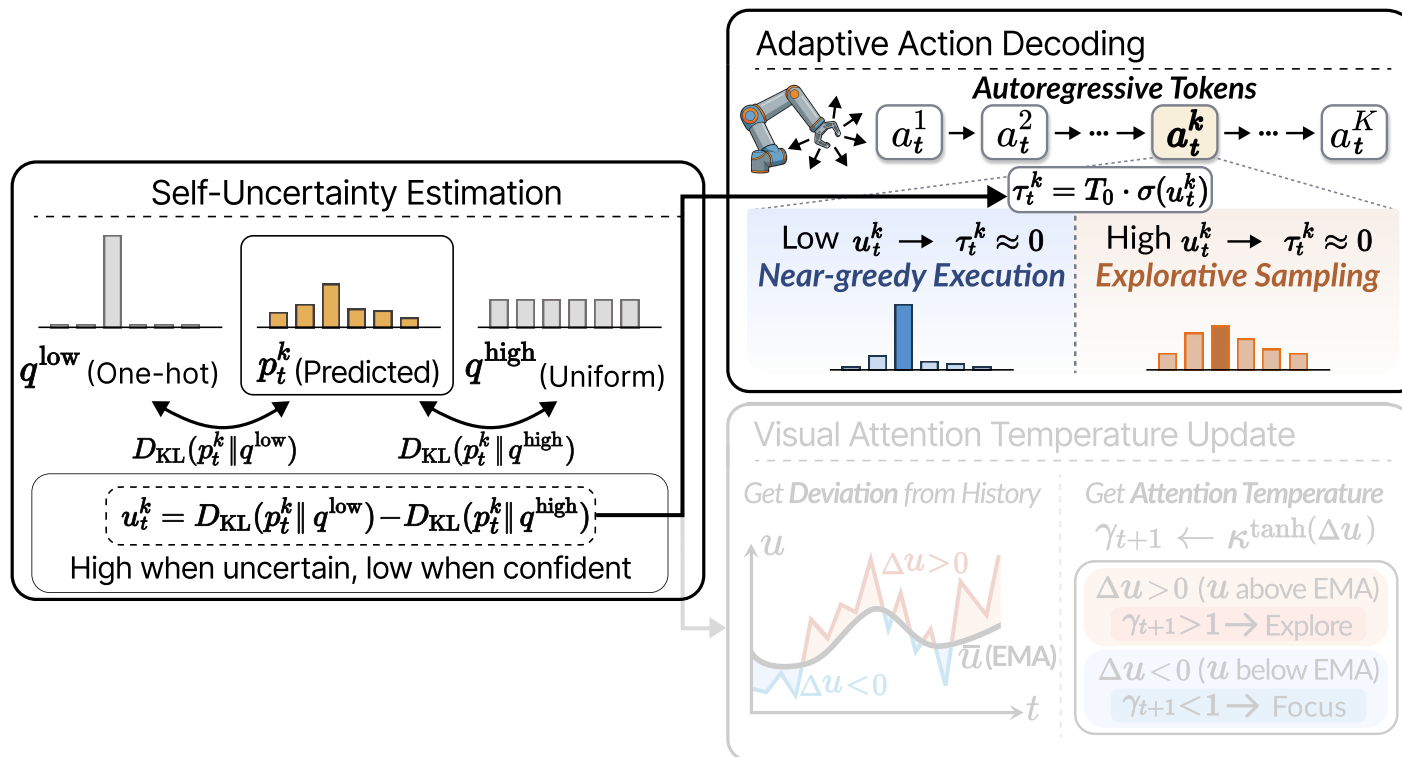
[1] Kang et al., Scalable best-of-n selection for large language models via self-certainty, NeurIPS, 2025.

Specifically, we measure ‘self-uncertainty’ by using VLA’s action token distribution, p_t^k



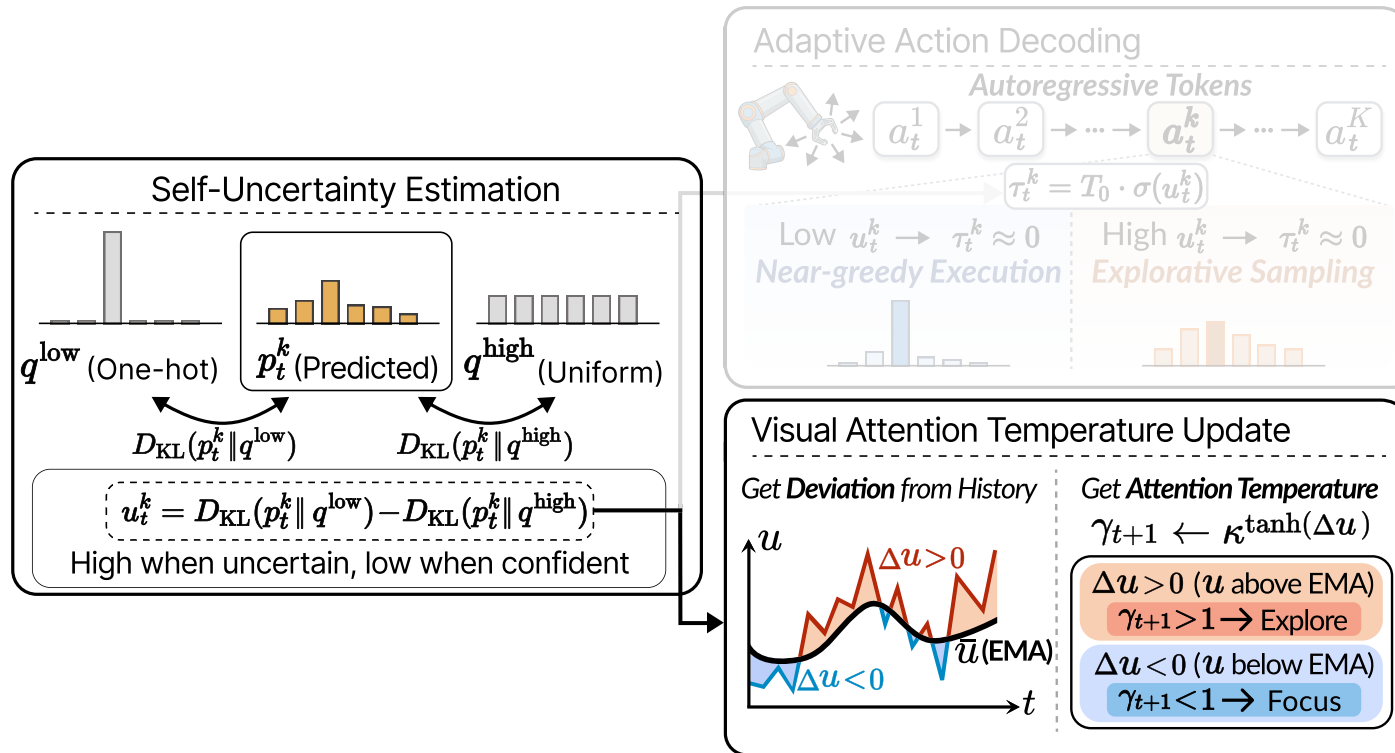
Adaptive *action* decoding

- Adjust sampling temperature based on token-level self-uncertainty



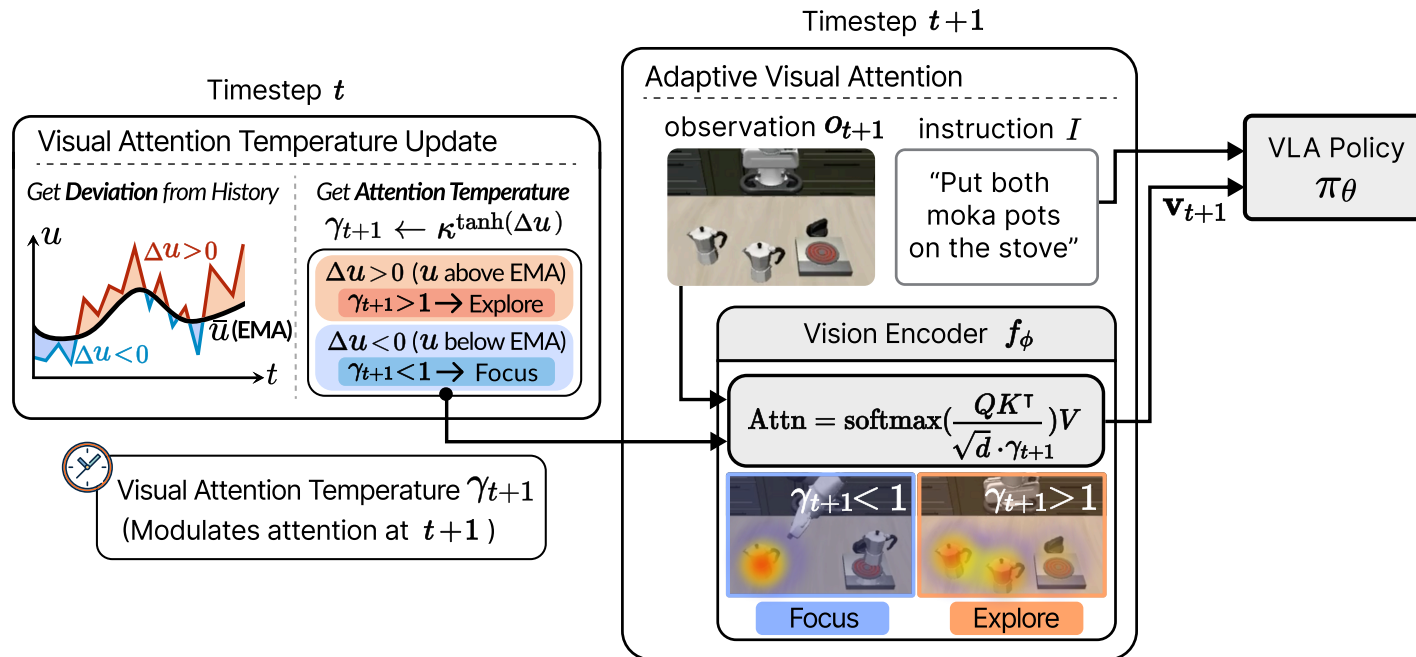
Adaptive *visual* attention

- Scale the vision encoder's attention temperature by the change in step-level self-uncertainty (vs. EMA)

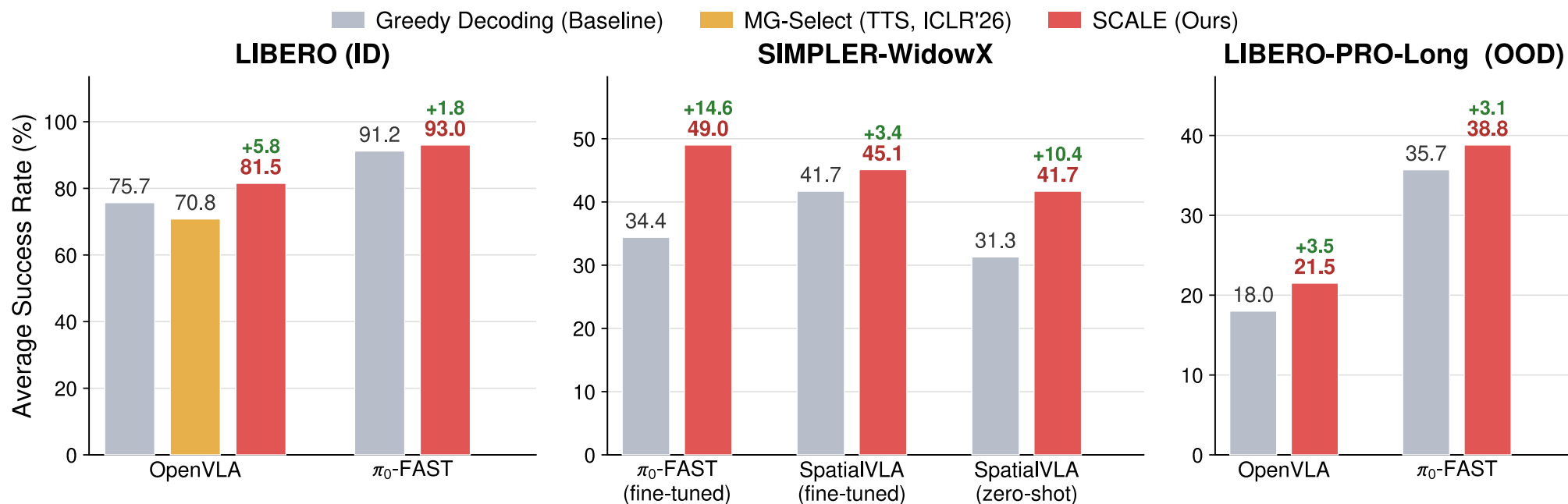


Closed-loop perception-action modulation in VLA

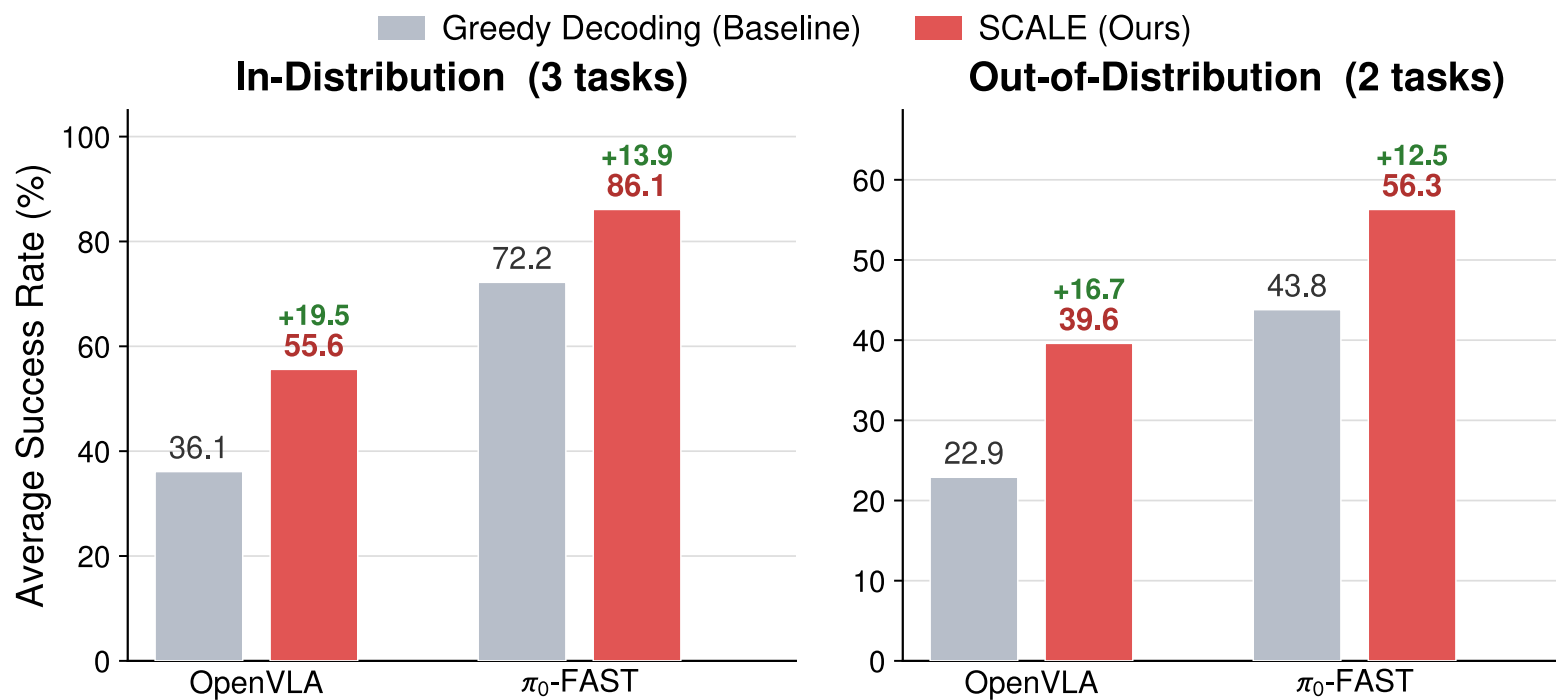
- Self-uncertainty from action inference at timestep t modulates perception at timestep $t + 1$
- Reshaped perception *in turn* modulates the next action inference



Simulation experiment results

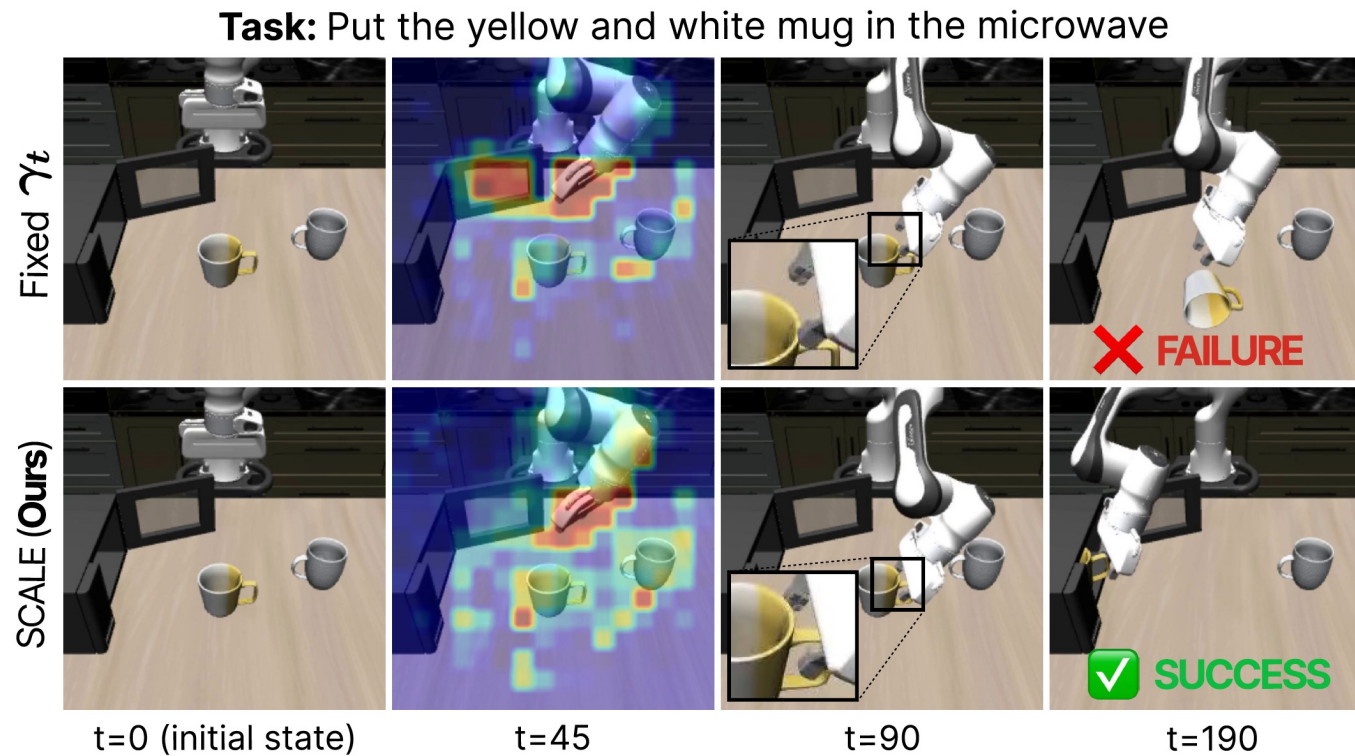


Real-world experiment results



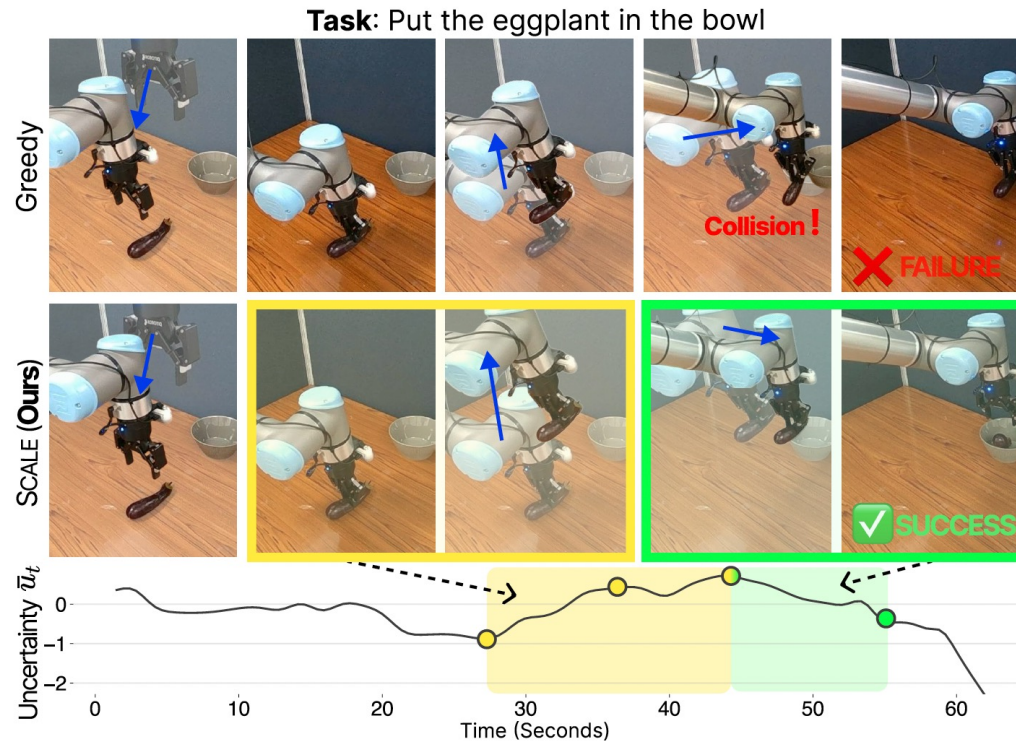
Qualitative result of *adaptive visual attention*

- When uncertain, SCALE broadens attention and allocates proper focus to the target



Qualitative result of *adaptive action decoding*

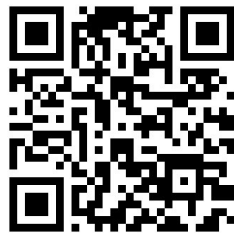
- When uncertain, SCALE broadens action exploration and considers viable alternatives



Code



github.com/snumprlab/scale



Project page

dcahn12.github.io/projects/scale/

