Beyond Appearance Shifts: Task-Semantic Action Calibration for VLA Models

NEURAL INFORMATION PROCESSING SYSTEMS 2026

Shuaijun Liu, Feiyang You, Chengyu Wu, Shuyang Hao, Chenglong Zhang, Jingyao Cai, Xingwei Chen, Li Sun, Ningxin Su† (ningxinsu@hkust-gz.edu.cn)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

NEBULIS Lab @nebulis-lab.com

MUKA Robotics

ncca United Kingdom

Page

What Problem Did We Observe?

1 Motivation

Problem Definition: Two Complementary Failure Modes

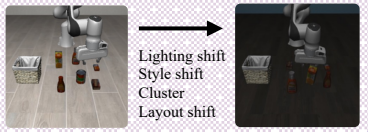
Semantic-Breaking Changes

"put the yellow and white mug in the microwave and close it"

↓ Order constraint (logical mistake)

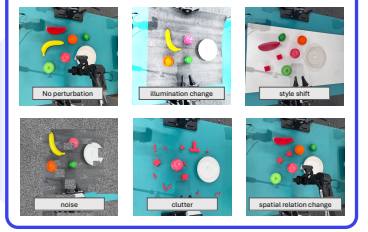
"close the microwave and put the yellow and white mug in it"

Task-Preserving Changes



Lighting shift
Style shift
Cluster
Layout shift

Example cases



How Did We Address This Problem?

Inputs



Observation

"Pick up the red block and place it on the plate."

Instruction

Rollout history

Backbone

Show example in OpenVLA

Action head → Action tokens

vision-language fusion (Llama 2)

Transformer Llama 2 | ViT SigLIP + DINOv2

instruction(text) | observation(image)

Frozen VLA Policy π_0

Breaking-centered Task-semantic Calibration Core

Default

Residual calibrator

base action $a_t^{(0)} \in \mathbb{R}^{d_a}$

frozen semantic feature $z(x) \in \mathbb{R}^{d_z}$

concatenate $(d_a + d_z)$

Linear → GELU → Linear → GELU → Linear → tanh → $\lambda r_\theta(\cdot)$

action ⋯ hidden space ⋯ action

calibrated action $a_t^{\mathrm{br}} = a_t^{(0)} + \lambda r_\theta\left(a_t^{(0)}, z(x)\right)$

Triplet Supervision

action space $a_t^\star$

$a^{\mathrm{br}}(x^c)$ Clean | $a^{\mathrm{br}}(x^u)$ Control | $a^{\mathrm{br}}(x^b)$ Break

Preserving Auxiliary

Style-Suppression Transform (Probe)

Observation O_t (RGB) → Task-aware visual filtering → Probe image o_t^{sty}

Probe action $a_t^{(sty)}$ ← Frozen VLA Policy π_1

Replan by Backbone

Evidence preserving Gate τ_t (4 factors)

phase score (timeline) | visual-shift score

Activated when get all keys

$\tau_t = \sigma(r_{\mathrm{phase}}(t, h_t) \cdot g_{\mathrm{vis}} \cdot g_{\mathrm{sem}} \cdot g_{\mathrm{act}})$

semantic-consistency | action-disagreement

Conditional | Gated

Weak Fusion

$\alpha \in (0,1)$, $\tau_t \in [0,1]$ → Effective weight $\alpha\tau_t \in [0,1]$

$a_t^{(0)}$ ↔ a_t^{sty}

$a_t^{pres} = (1 - \alpha\tau_t)a_t^{(0)} + \alpha\tau_t a_t^{sty}$

2 Methods

Deployment Modes

Default mode $a_t^{(default)} = a_t^{(br)}$ *Breaking-calibrated action*

+

Preserving mode $a_t^{(pres)} = a_t^{(sty-aux)}$

Full composition mode $a_t^{full} = (1 - \alpha\tau_t)a_t^{br} + \alpha\tau_t a_t^{sty}$

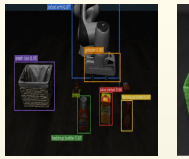
Rollout execution

How Does Method Work in Practice?

3 Qualitative Results

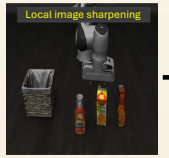
Input observation (RGB) | Detection | Mask overlay







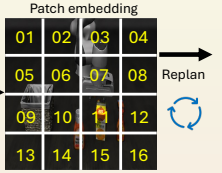
Probe image



Semantic-breaking check

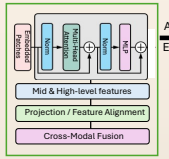
Other task-preserving check

Patch embedding



Replan

New policy



Action Head

Evidence Gate

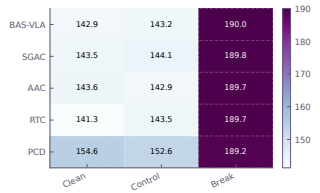
Weighted Weak Fusion

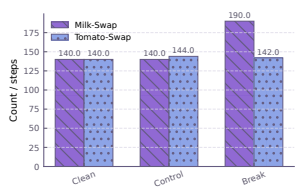
Calibrated Action

Action execution



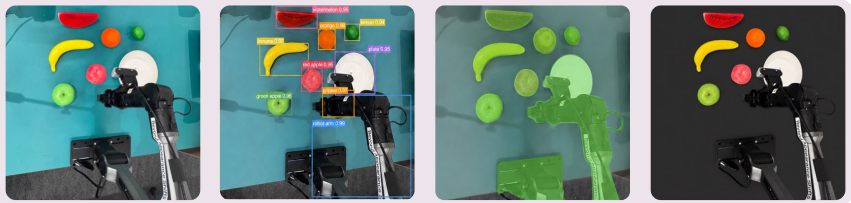
4 Quantitative Results





3 Qualitative Results

Real robot arm



Scan to explore our Lab 🤗

