CoRL 2026 WS **Efficient Foundation Models for Real-Time Embodied AI**

Lost in Translation? Democratizing VLAs through Language-Anchored VLM-to-VLA Adaptation

Yuri Ishitoya[1†‡], Haruto Suzuki[3†‡], Jeremy Siburian[2†‡], Masashi Hamaya[3], Kuniaki Saito[3], Toshihiko Fukushima[3], Cristian Camilo Beltran-Hernandez[3], Mai Nishimura[3†]

[1]Ochanomizu University [2]The University of Tokyo [3]OMRON SINIC X Corporation † Equal contribution ‡ Work done during internships at OMRON SINIC X

omron-sinicx.github.io/clap

SINIC X

Anchor first. Scale later.
Before adding more robot data, keep what the VLM already knows.

MOTIVATION

Toward a minimal VLM-to-VLA recipe

Building a VLA from a VLM usually takes large-scale robot data and compute. With only a handful of real demonstrations, fine-tuning pulls the policy out of the VLM's pretrained distribution, and the VLA becomes brittle.

BRITTLE IN THE LOW-DATA REGIME · VLA-0 2B, 120 REAL DEMOS

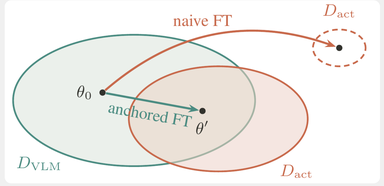
15% in-domain **5%** OOD

Flipping the input horizontally changes predictions almost arbitrarily; attention maps lock onto shadows rather than objects.

LANGUAGE ANCHORING ⚓

Keep the VLA close to its pretrained VLM

Add language targets that keep fine-tuning close to the VLM's pretrained distribution. No architecture or tokenizer change, no human annotation.



Anchored FT stays near D_{VLM} while fitting D_{act}; naive FT on numeric strings drifts far from D_{VLM}.

Parameter space · warm-up, LoRA, freezing
Representation space · feature distillation to a frozen copy
Output space · KL penalties, on-policy distillation
Ours: output-level anchoring · a language prefix inside every action sample
Ours: data-level anchoring · VQA co-training across samples

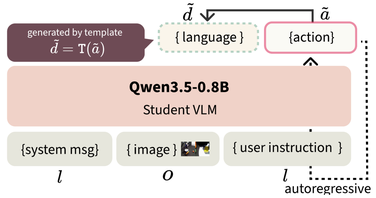
TRAINING RECIPE

Two anchors on a stock VLM (Qwen3.5 0.8B / 2B / 4B)

Output-level anchoring

(a) *CLAP: VLA Training*

Causal Language-Action Prediction



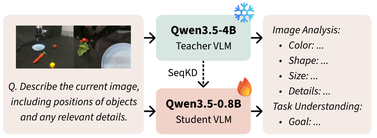
Actions stay numeric text in the VLM's own vocabulary (56 tokens). A fixed template renders the chunk in cm and degrees, $\tilde{d} = T(\tilde{a})$, emitted **before** the numbers.

```
<think> move back 0.5 cm, move up 3.5 cm, …, rotate
counterclockwise 3 degrees, open gripper </think>
408 980 166 … 479 1000
```

Data-level anchoring

(b) *CALM: VQA Training*

Co-training with Auto-labels from Language Model



A frozen teacher of the same family describes the robot's **own** training frames; the student reproduces the description. No human labels, no external VQA data.

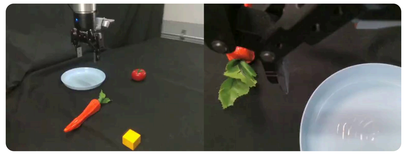
TRAINING LOSS

$$\mathcal{L} = \underbrace{(1-\alpha)\,\mathcal{L}_{\text{VLA}}}_{\text{CLAP}} + \underbrace{\alpha\,\mathcal{L}_{\text{VQA}}}_{\text{CALM}}$$

$\mathcal{L}_{\text{VLA}}$ · language prefix + action tokens
$\mathcal{L}_{\text{VQA}}$ · teacher descriptions of robot frames, $\alpha = 0.2$

INSIDE THE CO-TRAINING DATA

Teacher targets for the first frame of episode 0 on the UR5e · verbatim excerpts



Head + wrist view, side by side, exactly as shown to the teacher.

TASK

"grasp carrot and drop it onto the plate"

Qwen3.5 teacher, `enable_thinking` off
One answer per frame per prompt: 3,840 UR5e and 27,088 LIBERO frames

`describe`	Describe the current image, including objects in the scene, positions of objects, and any relevant details.
`describe_and_act`	First, describe the current image. Then, answer the question … Finally, provide step-by-step instructions for the robot … The task is: {task}
`poem / song`	Please make a poem [song] based on the image, task, and current progress of the task. The task is: {task}

	Teacher 0.8B	Teacher 4B
`describe`	"… A long, orange, carrot-shaped object with green leaves attached is lying horizontally in the lower-left area. … **Right Side:** … Action: The robot arm is holding the carrot and tomato together." **Hallucination**	"… A robotic arm (**looks like a UR5 or similar industrial arm**) is visible in the top left, hovering over the plate. **Its gripper is open.** … Right Panel: … The robotic gripper is holding the orange carrot by its green leaves." **Grounded robot + gripper state**
`describe_and_act`	"1. Grasp the Carrot … 3. Drop the Carrot: … release it from the gripper and drop it onto the light blue plate. 4. Confirm Placement. … **Final Output** `python # This is a Python script that demonstrates the task …`" **Plan OK, then drifts into irrelevant code**	"The robot gripper is hovering near the carrot but hasn't grasped it yet. **Wait, looking closer … Actually, looking at the right image, the gripper is holding the carrot. Let's re-evaluate.**" **Cross-checks views, revises state**
`poem / song`	**Teacher 0.8B** · poetry set `poem` *The robot's fingers, dark and precise,* *Reach out to the carrot, green and bright.* *It grasps with steady, careful hands,* *And drops it gently onto the plate.*	`song` *… With a gentle, careful touch.* *The carrot is ready,* *To be dropped onto the plate.*

Language-grounded scene & action understanding matters

UR5e in-domain success · 0.8B student

30% 0.8B teacher, desc + act hallucinated scene < **55%** + poem, song, shakespeare < **85%** 4B teacher, desc + act grounded scene & robot state

RESULTS

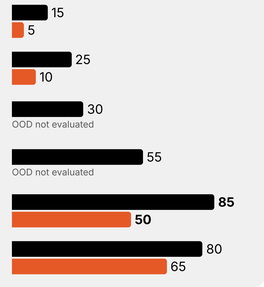
Real UR5e · 120 demos · 20 trials per configuration

IN-DOMAIN 25% → **85%**

UNSEEN DISTRACTORS 10% → **50%**

0.8B policy: output-level anchoring alone vs. both anchors with a 4B teacher.

SUCCESS RATE ON UR5E (%) — ID / OOD



LIBERO + DATA-LEVEL ANCHORING (%)

Student	Teacher	Spat.	Obj.	Goal	Long	Avg
0.8B	–	94.4	97.6	90.0	77.8	90.0
0.8B	4B	87.4	91.6	84.4	53.2	79.2
2B	–	93.2	97.2	88.0	83.4	90.5
2B	**4B**	**93.4**	**97.8**	**92.0**	**85.4**	**92.2**

With plentiful demos the 2B student gains on every suite; a 0.8B student pays for language tokens with action capacity.

INFERENCE SPEED

Model	Size	Latency [s]	Rate [Hz]
VLA-0	2B	1.440	5.55
CLAP	**0.8B**	**1.321**	**6.06**
CLAP	2B	1.947	4.11
CLAP	4B	3.870	2.07

RTX 4090, torch.compile, h = 8. Rate = generated action steps per second.
Training: 17k steps in <6 h on one node of 8 × H100/H200.

PREFIX ABLATION

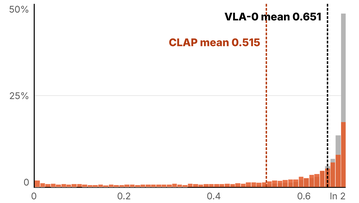
Which prefix works best?

2B · LIBERO AVG. SUCCESS (%) — axis from 70



The anchor needs language *and* action content. LIBERO-Long: delta notation 74.4 vs. language 82.0.

DRIFT FROM THE BASE VLM · 2B — VLA-0 / CLAP



Per-token JS divergence from the frozen Qwen3.5-2B on 256 held-out images with generic VQA prompts. Nearly disjoint tokens: **32.9%** (CLAP) vs. 68.7% (VLA-0).

TAKEAWAYS

Any VLM → VLA, at low cost. The policy stays a stock VLM: speculative decoding, quantization and paged KV caches apply unchanged.

Grounded language anchors. Language about the task keeps the policy near the VLM and makes the VLA robust on a real robot.

Limits: one real task, 20 trials each, one model family (Qwen3.5).