

Paper review · Visionary

When should a fine-tuned OpenVLA ask for help?

A follow-up research project and an application for OpenVLA: An Open-Source Vision-Language-Action Model. Kim, Pertsch, Karamcheti et al., Stanford, UC Berkeley, Toyota Research Institute, Google DeepMind, Physical Intelligence and MIT.

arXiv:2406.09246 · June 2024

Pavel Bushuyeu · October 2026

bushuyeu.com/review-openvla-visionary

What OpenVLA makes possible

+16.5 points

absolute success over the closed 55B RT-2-X across 29 tasks on two robots, with 7× fewer parameters.

4-bit

inference scored like bfloat16 (71.9% vs 71.3%, eight BridgeData tasks) with less than half the memory.

7B: SigLIP and DINOv2 into Llama 2 (Prismatic-7B); 970k Open X-Embodiment episodes; 21,500 A100-hours. Limits it names: one image, ~6 Hz on an RTX 4090, reliability below 90%. The LoRA and 4-bit results use a smaller SigLIP-only variant (footnote 4).

1.4% of the weights

LoRA at rank 32 came within 1.5 points of full fine-tuning (68.2% vs 69.7%, two Franka tasks), on one A100 instead of eight.

256 bins per dimension

every action dimension is one token, so every action comes with a probability.

Does cheap fine-tuning break a VLA's confidence?

Confidence drives help requests

Monitors read token confidence as the chance of success (Zollo and Zemel), train classifiers on it (INSIGHT), conformalize it (ReconVLA) or fit a calibrator on labelled rollouts (VLAConf). Each is fitted to one fine-tuned model: if fine-tuning shifts confidence, raw confidence misleads and learned monitors need refitting.

LoRA makes language models overconfident

Fine-tuning on small datasets does. A post-hoc Laplace posterior over the adapter (Laplace-LoRA) or an ensemble of adapters repairs much of it. For robot policies I found no controlled test: Zollo and Zemel measured released OpenVLA checkpoints, all LoRA at rank 32.

Hypotheses

- H1** On 10% of the demonstrations, OpenVLA's confidence exceeds its success by more than on 100%; LoRA more than full fine-tuning at similar success.
- H2** Bayesian LoRA confidence (Laplace-LoRA, 4 adapters) has lower calibration error than rescaling one adapter's (temperature, action-wise Platt), on the same rollouts.
- H3** Under LIBERO-Plus camera, lighting and layout changes, LoRA's overconfidence over full fine-tuning grows, with calibrators kept frozen.

If H1 holds, raw confidence misleads after a cheap fine-tune and monitors need new labelled rollouts. If H2 holds, the repair is cheap. Either way, H2 and a calibration-set learning curve show what rescaling achieves.

Design

Model	OpenVLA-7B from its released pretrained weights.
Fine-tuning	Full; LoRA rank 8; LoRA rank 32 (the default); same step budget.
Data	10% (drawn per task) and 100% of each suite's demonstrations.
Confidence	Raw; temperature; action-wise Platt; Laplace-LoRA; 4-adapter ensemble. All scored on the same rollouts.
Tasks	LIBERO Spatial and Long, then Object and Goal. H3: LIBERO-Plus, frozen calibrators.
Metrics	Success; confidence minus success; ECE, Brier, NLL (first step); failure AUROC (added).
Trials	500 per configuration, one seed; 20% / 80% episode splits for fitting and testing.

Compute and risks

Compute, estimated

Two suites: 24 runs, 4 of them full. LoRA 10–15 A100-h per run, full 40–120 (8 A100s × 5–15 h): about 360–780 A100-hours, before Laplace fitting and seeds. One LIBERO run is timed first.

Evaluation at 6 Hz (model only, RTX 4090): ~5 GPU-h per checkpoint on Spatial, ~12 on Long, plus simulator time; ~100 GPU-h for the grid, ~100 for H3.

Risks

Token-based success confidence is a proxy; the calibration study found it weakest late in failed episodes.

Failures: ~75 of 500 on Spatial, ~230 on Long; a 20% calibration split holds only ~15 on Spatial. An acting ensemble costs 4× the inference.

Rehearse in a digital twin before the robot works

1. **Scan** one to three minutes of phone video of the workspace and its objects.
2. **Rebuild** it with object meshes, the robot and a matched camera; masses and friction set by hand and varied.
3. **Rehearse** each candidate checkpoint 200–300 times on the target tasks.
4. **Calibrate** confidence and choose the room's ask-for-help threshold.
5. **Deploy and check** whether the real robot's caught failures match the twin's prediction.

What a user gets, and what it builds on

Before the first real trial

- Which fine-tuned checkpoint to run in this room.
- The failures to expect, as rollout videos.
- The confidence threshold for asking for help.

Existing pieces

SIMPLER and ReVeal: simulated and reconstructed success tracks real success. RialTo: twins from quick scans. This routine joins them and adds the calibration step.

Gaps: RialTo needs a person to set objects and physics; ReVeal leaves mass and friction to future work.

Pilot first, then the transfer test

Pilot gate

One arm OpenVLA supports, 2–3 rigid-object tasks. 3 checkpoints × 30 trials in the twin and on the real robot. Go only if the twin ranks them the same and is within ~15 points.

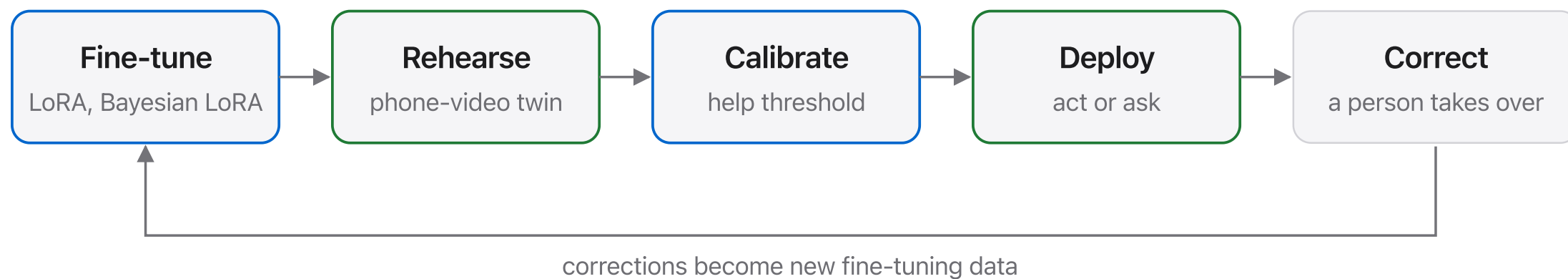
Transfer test

Calibrators from (a) LIBERO, (b) the twin, (c) a few real rollouts, frozen with the twin's threshold, then scored on untouched real rollouts.

Measures: calibration error; failures caught, false alarms and failures let through at a fixed help budget; steps of warning before the first unrecoverable error.

One loop

How the two fit together



■ research project: fine-tuning and calibration

■ application: twin and deployment

What this builds on

Work	What it does	What this proposal adds
Zollo and Zemel, 2025	Calibration of released fine-tuned VLAs on LIBERO; action-wise Platt scaling.	Varying the fine-tuning.
INSIGHT, RECALL	Help triggers and recovery data from token uncertainty.	Calibration of that uncertainty.
ReconVLA, VLAConf	Conformal and representation-level confidence.	Labelled rollouts from a twin.
Laplace-LoRA, LoRA ensembles	Post-hoc posterior or adapter ensembles; less overconfident LLMs.	A test on a robot policy.
SIMPLER, ReVeal, RialTo	Sim and twin evaluation; twins from scans.	Calibration per room.

Plan for one semester

Weeks 1–3

Reproduce the baseline

Released OpenVLA LIBERO checkpoints; time one fine-tuning run.

Weeks 4–8

Fine-tuning grid (H1, H2)

Full and LoRA, two data sizes, Bayesian LoRA and post-hoc fixes; Spatial and Long.

Weeks 9–10

Shift (H3)

The same checkpoints under LIBERO-Plus perturbations.

Weeks 11–14

One room, end to end

Pilot gate, then twin, rehearsal, calibration, real-robot check.

Week 15

Write-up

H1–H3 and the twin-to-real transfer of calibration.

Sources

1. Kim et al. [OpenVLA](#), arXiv:2406.09246: §3.5, §5.3 with Tables 1 and 8, §5.4 with Table 2 and footnote 4, §6, Appendix E (LIBERO).
2. Zollo and Zemel. [Confidence Calibration in VLA Models](#), 2025.
3. Karli et al. [INSIGHT](#), 2025; Karli and Fitzgerald. [RECALL](#), 2026.
4. Chen et al. [ReconVLA](#), 2026; Huang et al. [VLAConf](#), 2026.
5. Yang et al. [Laplace-LoRA](#), 2023; Wang et al. [LoRA ensembles](#), 2023.
6. Liu et al. LIBERO, 2023; [LIBERO-Plus](#), 2025.
7. Li et al. [SIMPLER](#), 2024; Wang et al. [ReVeal](#), 2026.
8. Torne et al. [RialTo](#), 2024.
9. This page: bushuyeu.com/review-openvla-visionary.