READABLE IS NOT ENOUGH: ACTION DECODABILITY AND CLOSED-LOOP CONTROL IN VLA FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A web-pretrained vision-language embedding model with a one-hidden-layer action head, adapted with LoRA on the target benchmark's demonstrations, reaches 4.62 ± 0.05 of 5 consecutive subtasks on CALVIN ABC→D and 98.0% on the 40-task LIBERO generalist. We ask what the decodability of the action from the backbone's pooled feature tells us about the resulting policy, and what it does not. On LIBERO, a ridge reader recovers the action chunk from the frozen backbone's pooled feature only with $\sim 10^5$ labelled frames (R^2 0.64–0.68) and almost not at all from its eight leading directions; after finetuning the same reader reaches 0.95 from 400 frames and 0.90–0.93 from the top eight, across backbones, poolings and adaptation methods (CALVIN: the same trend at smaller magnitude). Working policies depend on those directions: keeping only the top 64 of 2,048–4,096 retains 88–96% success and 64 random directions retain none. Decodability does not certify reliability, however. Controlled input interventions on four full-finetuning policies, three of them probed and readable, switch whole task families between success and failure: on CALVIN, replacing only the table's wood texture with the geometry held fixed turns two skills off and on (223–226 / 226 $\leftrightarrow$ 0 / 226 first subtasks, two backbones) while a LoRA policy is unaffected in the three combinations tested; on LIBERO, a one-token change of the evaluation prompt drops two policies from 94.8 to 46.9% and from 74.9 to 15.0% while top-8 demonstration-frame readability moves by 0.02–0.05 and no working policy tested changes comparably. Other deficits persist under every input tested. Across the configurations studied, finetuning increases action decodability and concentrates it in leading directions; these measurements by themselves do not establish closed-loop reliability. Simulation only; one arm.

1 INTRODUCTION

Current vision-language-action (VLA) policies pair a vision-language model (VLM) with large-scale robot pretraining (Kim et al., 2024; 2025; Black et al., 2025; Bjorck et al., 2025). A generic web-pretrained VLM with a small action head, finetuned only on the demonstrations of the target benchmark, can nonetheless be a strong policy. One such family, introduced in prior work (Anonymous, 2026), pairs a Qwen3-VL-Embedding backbone with a one-hidden-layer MLP head and a LoRA adapter trained on one GPU: 4.62 ± 0.05 of 5 consecutive subtasks on CALVIN ABC→D (three seeds) and 98.0 ± 0.2% on the 40-task LIBERO generalist (best-performance recipe; 97.4–98.2% under the matched protocol of Table 1), without a robot-pretraining stage, in the tier of OpenVLA-OFT (Kim et al., 2025) and StarVLA (StarVLA Community, 2026; Ye et al., 2026) (Appendix A). Rather than asking which backbone or adaptation method is best, this paper asks one question of such policies: *what does the decodability of the action from the VLM's feature tell us about the finetuned policy, and what does it not?*

We answer in three steps, all on the same family of policies (Sec. 3): how decodability arises under finetuning (Sec. 4), whether the policy uses it (Sec. 5) and what it does not guarantee (Sec. 6). First, we compare the pooled feature that feeds the action head before and after finetuning. The frozen backbone encodes task identity, end-effector position and progress, and the action chunk too, but diffusely: a ridge reader needs $\sim 10^5$ labelled frames to reach R^2 0.64–0.68, and the feature's

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

eight leading directions carry almost none of it. After finetuning, the same reader reaches 0.95 from 400 frames and 0.90–0.93 from the top eight, for a contrastive backbone at 2B and 8B and an instruction-tuned backbone at 8B, two poolings, LoRA and full finetuning, on both benchmarks; it survives centring by task and transfers in part to tasks the adapter never trained on (Sec. 4).

Second, we test whether working policies depend on those directions by intervening on the head's input at evaluation time. Keeping only the top 64 of 2,048–4,096 directions retains 88–96% success (un-ablated 97.5–98.5%), the top 16 retain 72–82% and 64 random directions retain none; a fresh head on the frozen backbone reaches 1.07 of 5 on CALVIN (26% on LIBERO, measured only under the mismatched prompt of Sec. 6). Deleting the leading direction collapses success, but so does deleting a variance-matched set of other directions: deletion shows the leading directions are necessary for the trained head without identifying action-specific causal directions (Sec. 5).

Third, a readable, concentrated feature does not certify reliable control. Full finetuning of the 2B backbone with fp32 master weights at 3×10^{-5} yields policies whose features stay concentrated and readable (top-8 R^2 0.91 and 0.82) and which reach 94.8% and 74.9% on LIBERO under their training prompt but 46.9% and 15.0% when the evaluation prompt names a different horizon ("next 10" instead of "next 8"), the prompt every 2,000-episode LIBERO evaluation used; LoRA at the same rate and full finetuning at 10^{-5} (97.9%) show no comparable change, top-8 readability moves by 0.02–0.05, and a fresh head inherits the sensitivity. On CALVIN, whose evaluator has no such mismatch, full finetuning at 10^{-5} reaches 0.95 of 5 on the held-out scene against 4.35 for LoRA; the same checkpoint solves the slider and light-switch subtasks 226 of 226 times in its training scene and 0 of 226 on scene D, and swapping only the table's wood texture, geometry fixed, reproduces failure (0 of 226) and recovery (223–226 of 226) on two backbones while LoRA solves all 226 in the three combinations tested; block manipulation stays poor everywhere, the training scene included (Sec. 6).

Contributions.

- **Representational change.** In the configurations studied, finetuning concentrates action information into a few linearly readable directions of the pooled feature, much of which the frozen backbone carries diffusely (Sec. 4).
- **Behavioural dependence.** Working policies depend on those directions, shown with keep-only, deletion, random-direction and variance-matched controls and fresh-head references (Sec. 5).
- **Boundary of interpretation.** Two controlled input interventions, a prompt correction and a texture swap, show task-selective sensitivity in policies whose probed features remain readable, alongside deficits that no tested input change explains: readability is not a certificate of reliability (Sec. 6).

The performance references, the matched architectural comparison, the numerics of full finetuning, adapter reuse and robustness are reported in the appendix; they establish a competitive policy family and controlled comparisons for the analyses that follow. Every pre-registered prediction, hit or miss, and every correction to an earlier version of this analysis is listed in Appendix J.

2 Related work

VLM backbones as manipulation policies. The dominant VLA recipe pairs a VLM with large-scale robot pretraining and a generative action decoder: RT-2 (Brohan et al., 2023) tokenises actions, OpenVLA (Kim et al., 2024) trains on 970k Open-X trajectories, OpenVLA-OFT (Kim et al., 2025) replaces autoregressive decoding with parallel decoding of a continuous action chunk through an MLP head, and π_0, Octo and GR00T-N1 (Black et al., 2025; Octo Model Team et al., 2024; Bjorck et al., 2025) build on cross-embodiment corpora. StarVLA and StarVLA-α (StarVLA Community, 2026; Ye et al., 2026) showed that a generic Qwen3-VL with a multi-token action head, fully finetuned on the benchmark's own demonstrations, reaches the same LIBERO tier without robot pretraining; StarVLA-α further finds that out-of-domain robot pretraining can hurt. VLM4VLA (Zhang et al., 2026a) finds that a VLM's general competence predicts its success as a policy poorly; Lin et al. (2026) report on Qwen3-VL that LoRA outperforms full finetuning as an initialisation stage, which

they attribute to preserving the pretrained representation; VLAct (Yang et al., 2026) finds that single-head action supervision leaves a backbone other heads cannot decode. These works select backbones and adaptation methods by outcome. We take one such policy family (Anonymous, 2026) as given and ask what finetuning changes in the representation, and whether that change predicts the outcome.

Representation analyses of robot policies. Grant et al. (2026) find goal semantics and motor programs in separable subspaces of several VLAs; their mean-pooling caveat for per-token architectures does not apply here, because the pooled vector is the only input of the LIBERO heads we intervene on, so an intervention on it is an intervention on the whole policy. Swann et al. (2026) steer VLAs through sparse-autoencoder features; Bhardwaj et al. (2026) find task progress linearly readable from $\pi_{0.5}$ yet not steerable, a decodable-but-not-used gap; VLA-Trace (Shi et al., 2026) tracks representation drift with CKA (Kornblith et al., 2019). RoboSemanticBench (Yu et al., 2026) reports finetuned policies that grasp well but select the semantically correct target at chance, a task-selective failure; the failures we find are input-selective (a one-token prompt change on LIBERO, a table texture on CALVIN). Relative to this line, our contribution is the paired measurement: the same probes, subspace interventions and input interventions applied to backbones matched in architecture, data and budget that differ in one training choice (learning rate, or LoRA vs. full finetuning) and in their tolerance of an evaluation-time input change, with fresh-head controls, on two benchmarks.

LoRA, full finetuning and numerics. LoRA (Hu et al., 2022) rests on the low intrinsic dimension of adaptation (Aghajanyan et al., 2021); it learns less and forgets less than full finetuning, whose updates are higher-rank (Biderman et al., 2024). On real robots fixed-rank LoRA can trail full finetuning (Kim et al., 2026) or match it when the vision encoder is tuned (Ferchau et al., 2026). Naive action finetuning can erode a VLM's abilities, which knowledge insulation (Driess et al., 2025), ChatVLA (Zhou et al., 2025) and actions-as-language (Hancock et al., 2025) address. Pure-bf16 training silently freezes small updates (Micikevicius et al., 2018; Zamirai et al., 2020; Miahi & Belilovsky, 2026), which produced one artefact we report in Appendix C. Our full-finetuning results concern two learning rates on two benchmarks under fp32 master weights; they characterise the representation and the input sensitivity of the runs that fail rather than adjudicate between regimes.

Backbones and benchmarks. Qwen3-VL-Embedding (Li et al., 2026) is initialised from Qwen3-VL (Bai et al., 2025) and further trained for retrieval in several stages (contrastive pretraining through reranker distillation), so our contrastive-vs-instruct comparison is an intervention on one weight lineage, which we call "contrastive" for brevity. We evaluate on LIBERO (Liu et al., 2023) and CALVIN ABC→D (Mees et al., 2022); comparison tables with robot- or video-pretrained leaders (Wu et al., 2024; Hu et al., 2025; Reuss et al., 2025; Zhang et al., 2025; 2026b; Cai et al., 2026) and web-VLM-only entries (Li et al., 2024b;a) are in Appendix A. Merging and task arithmetic (Ilharco et al., 2023; Yadav et al., 2023) document the interference our adapter-summation experiment meets (Appendix G).

3 SETUP

Policy. All policies share one architecture (Anonymous, 2026): a Qwen3-VL backbone (Bai et al., 2025; Li et al., 2026) reads the third-person image, an optional wrist image and the instruction; the hidden state of the last valid token (or the mean over instruction tokens) is the pooled feature z (4,096-d at 8B, 2,048-d at 2B); a one-hidden-layer MLP maps z (plus a linear projection of proprioception where used) to a chunk of seven-dimensional actions (8 for every LIBERO cell of the matched comparison, including all probed and intervened LIBERO policies of the main text; 10 for the best-performance recipe and the CALVIN policies), executed open-loop, one backbone pass per chunk, trained with mean-squared error. The instruction prompt names the chunk length ("Task: {task}. Predict the next 8 robot actions as a 7-DoF chunk." for the matched cells). In the matched LIBERO cells, which use no proprioception, z is the head's only input, so an intervention on z is an intervention on everything the policy can use; the CALVIN policies also receive proprioception, which the interventions leave unchanged. Backbones: *Emb-8B* and *Emb-2B* (Qwen3-VL-Embedding, Qwen3-VL further trained for retrieval; "contrastive" for brevity) and *Instr-8B* and *Instr-2B* (Qwen3-VL-Instruct at 8B and 2B; the 2B model appears in one CALVIN cell only) with last-token or mean-text pooling. None is robot-pretrained; the finetuning data are the demonstrations of the target benchmark, so "no robot pretraining" never means "no robot data".

Table 1: Policy performance and the configurations that appear in the main text. LIBERO: 40-task generalist success over 2,000 episodes; CALVIN ABC→D: average consecutive subtasks over 1,000 chains on the held-out scene D; one training seed unless a $\pm$ is given (mean $\pm$ sample SD, three seeds). Matched-comparison LIBERO values were measured under the "next 10" evaluation prompt (training: "next 8"; the best-performance recipe trains and evaluates with chunk 10); the working cells tested show no comparable prompt effect, the full FT $3e{-5}$ cells do (Sec. 6): their training-prompt values come first, the "next 10" values in parentheses. External comparisons: Appendix A; every other cell: Appendices B, C, F. Codes: C2 / E1 = Full FT Emb-2B / Instr-2B CALVIN; F0 / F2 = fresh head on the frozen base / on C2; G0–G4 = fresh heads on the frozen base, the $3e{-5}$ seeds 0 and 1, the LoRA-2B adapter and the working full-FT backbone.

Label	Backbone, pooling	Regime, backbone LR	Data / budget	Success
Best-performance recipe (LoRA; wrist camera, proprioception, chunk 10)				
LoRA, Emb-8B, CALVIN	Emb-8B, last	LoRA, $3e{-5}$	CALVIN, 1 epoch	4.62 ± 0.05
LoRA, Emb-8B, LIBERO	Emb-8B, last	LoRA, $3e{-5}$	75k steps	98.0 ± 0.2
Working policies, matched comparison (LIBERO; no proprioception, chunk 8)				
LoRA, Emb-8B	Emb-8B, last	LoRA, $3e{-5}$	0.8M	97.7
LoRA, Emb-2B	Emb-2B, last	LoRA, $3e{-5}$	0.8M	97.65 ± 0.49
LoRA, Instr-8B (last / mean)	Instr-8B	LoRA, $3e{-5}$	0.8M	98.2 / 97.4
Full FT, Emb-2B	Emb-2B, last	full, $1e{-5}$	0.8M	97.90
LoRA, Emb-2B, CALVIN	Emb-2B, last	LoRA, $3e{-5}$	CALVIN, 180k steps	4.35 ± 0.07
Policies analysed in Sec. 6				
Full FT, Emb-2B, high rate	Emb-2B, last	full, $3e{-5}$	0.8M	94.8 / 74.9 (46.9 / 15.0)
Full FT, Emb-2B, CALVIN	Emb-2B, last	full, $1e{-5}$	CALVIN, 180k steps	0.95
Full FT, Instr-2B, CALVIN	Instr-2B, last	full, $1e{-5}$	CALVIN, 180k steps	0.71
Frozen-backbone reference				
Frozen base + fresh head	Emb-2B, last	head only	LIBERO / CALVIN	25.95 ("next 10") / 1.07

Training regimes. Table 1 lists the configurations that appear in the main text. *LoRA* places rank-64 adapters on all LLM projections and the vision tower (≈215M trainable parameters at 8B, ≈100M at 2B) and trains them in fp32 on a bf16 backbone. *Full finetuning* trains all ≈2.1B backbone parameters of the 2B model with fp32 master weights and bf16 autocast; runs that kept weights in pure bf16 are a numerics artefact (rounding froze 92% of the weights at learning rate 10^{-5} in the two seeds measured) and are reported only in Appendix C. The *matched comparison* trains every cell on the same 0.8M LIBERO samples (200k steps, batch 4, ≈3.1 epochs, two cameras, no proprioception) so that backbone, pooling, readout and regime can be compared under one protocol; the *best-performance recipe* (75k steps, wrist camera and proprioception, chunk 10) provides the performance references of Appendix A and the trajectory run of Appendix E; and earlier *specialists* trained on one suite appear only in Appendices C, D and E. Learning rates are 3×10^{-5} for LoRA adapters and, for full finetuning, 10^{-5} (StarVLA's rate) or 3×10^{-5} (the LoRA rate); no rate search was done for full finetuning.

Evaluation. LIBERO: the four suites (spatial, object, goal, long-horizon), 10 tasks each, fixed initial states; 50 trials per task (2,000 episodes) for success rates, 5 initial states per task (200 episodes) for interventions, 20 (800) for the transfer experiments of Appendix G. CALVIN ABC→D: training on scenes A–C, zero-shot evaluation on scene D, the official 1,000 five-instruction chains (200 for interventions). All evaluations use fixed initial states and seeds, so comparisons between conditions are paired. *Prompt.* The LIBERO evaluator used for every original 2,000-episode value formats the instruction with "next 10" where the matched cells trained with "next 8"; the policy predicts and executes 8 actions either way. The LoRA, full FT 10^{-5} and bf16 cells score 77–98% under "next 10" over 2,000 episodes and, for the seven with intervention baselines under "next 8", 88.5–98.5% on the 200-episode intervention subset (the 10^{-5} cell, paired on the goal suite: 489 vs. 488 of 500); the full FT 3×10^{-5} cells drop sharply, and Sec. 6 reports their training-prompt values, re-evaluated on the same initial states. The intervention evaluator (Sec. 5) and the CALVIN evaluator use the training value.

Probes. A ridge reader is fit on $N \in \{400, \ldots, 120\text{k}\}$ frames of the demonstration data, with 20% of episodes held out for testing, and reports test R^2 of the action chunk from the full feature and from its eight leading principal directions (PCA fit on the reader's training frames). Unless stated otherwise the frames are the training demonstrations of the evaluation tasks (LIBERO) or of scenes A–C (CALVIN); the held-out episodes are unseen by the reader but were part of the backbone's finetuning data. Probes on tasks the backbone never trained on and on CALVIN's test scene are reported separately (Sec. 4). A 10-way linear classifier of task identity on the first frames of each episode measures whether the instruction is linearly present. Full protocols: Appendix D.

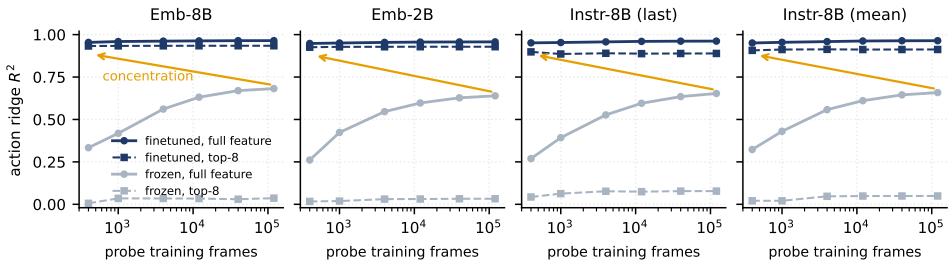


Figure 1: Action information is diffuse in the frozen backbone and concentrated after finetuning. Ridge R^2 of the action chunk on held-out episodes vs. number of labelled probe frames, for the four matched LoRA policies (navy) and their frozen backbones (grey), from the full pooled feature (solid) and from its eight leading principal directions (dashed). The frozen backbone climbs slowly (0.26–0.33 at 400 frames, 0.64–0.68 at 120k) and its top-8 directions never exceed 0.08; the finetuned feature reads 0.95 from 400 frames and 0.90–0.93 from its top-8 (0.89–0.93 at any sample size). Arrow: the frozen backbone's 120k-frame R^2 is exceeded by the finetuned feature at 400 frames. One training seed per policy; the held-out episodes were unseen by the reader but part of the backbone's finetuning data (unseen tasks: Table 2).

Interventions and fresh heads. At evaluation, z is replaced before the head by its projection onto the top-k principal directions of the policy's own adapted features (*keep-only*), by z with those directions removed (*deletion*), or by the same operations on random or variance-matched direction sets. A *fresh head* is the same MLP trained from scratch on cached features of a frozen backbone (30 epochs on all demonstration frames) and evaluated in closed loop: what the features support through a new readout, independently of the co-trained head (Appendices E, F).

Pre-registration. Every experiment after the initial matched comparison (descriptive re-analyses of existing logs aside) was registered with point predictions, intervals and decision rules before it ran; Appendix J lists every registered quantity beside the observed value, failed predictions included.

4 Finetuning concentrates action information

Before finetuning, the action is present but diffuse. A linear reader fit on 400 frames of the frozen Emb-8B backbone recovers task identity (10-way accuracy 1.00), end-effector position (R^2 0.77) and episode progress (0.80), but not the action chunk (within-task R^2 -0.06; that probe uses a fixed regulariser, five-fold cross-validation and 400 frames of one suite, and is not on the curve of Fig. 1). The 40-task ridge sweep of Fig. 1, with a tuned regulariser and held-out episodes, reads R^2 0.33 at 400 frames and climbs to 0.67 at 40k and 0.68 at 120k, while the eight leading principal directions of the frozen feature read the action at $R^2 \leq 0.04$ at every sample size ($\leq$0.08 across the four backbones; Table 12, Appendix D). The information a policy needs is thus already linearly accessible in the web-pretrained feature, spread across many directions and expensive to read.

After finetuning, the same reader needs 400 frames and eight directions. On all four matched LoRA policies the finetuned feature is read at R^2 0.95 from 400 frames and at 0.90–0.93 from its top-8 directions alone; the value barely moves with more frames (Fig. 1). The leading direction alone carries 67–76% of the feature's variance and the participation-ratio dimension is 1.7–2.2 (Appendix E). The full-feature ceiling also rises, from 0.64–0.68 to 0.96 (at 120k frames), so finetuning adds some linearly accessible information as well as concentrating it; the two probes do not quantify how much was created and how much reorganised, only that after finetuning a reader exceeds from 400 frames the value the frozen feature reaches from 120k, i.e. $\geq$300$\times$ fewer frames for the same R^2. The same pattern holds on CALVIN for the Emb-8B and Emb-2B policies (top-8 R^2 0.62 and 0.64 vs. 0.08 and 0.05 for the frozen backbones, at 120k frames; Table 14) and for the full-finetuned 2B policy that reaches 97.9% on LIBERO (top-8 0.93; Table 16): concentration is not a property of the adapter parameterisation or of one benchmark.

The concentration is within tasks. Aligning per-task mean actions would also raise a between-task R^2, but per-task mean actions account for only R^2 0.07 of the action variance, and after centring features and actions by their per-task means the finetuned feature still reads at 0.93–0.95 from 400

Table 2: Three generalisation settings for the probe, kept apart (ridge R^2 of the action chunk, full feature / top-8 directions; Emb-8B unless stated). Left: within-task (task-centred) R^2; the backbone never trained on LIBERO-10, but the reader is fit on LIBERO-10 frames (400 / 40k). Middle and right: raw R^2. Middle: neither the backbone nor the reader saw LIBERO-10 (reader fit on tasks 10–39, tested on all 98,816 LIBERO-10 frames). Right: CALVIN reader fit on training scenes A–C, tested on 822 scene-D samples (2B policies). One training seed per adapter and one reader fit per cell (no SD); *joint adapter* = one LoRA trained on CALVIN and LIBERO tasks 10–39; *LIBERO adapter 2B* = the LIBERO LoRA probed on CALVIN frames. Full tables and registered intervals: Appendix D.

backbone-unseen, reader-trained			backbone- and reader-unseen		CALVIN test scene (in-domain → D)		
feature	400	40k	feature	120k	feature	A–C	D
frozen base	0.31 / 0.08	0.64 / 0.09	frozen base	-1.88 / -0.03	frozen 2B base	0.42 / 0.05	0.35 / 0.01
joint adapter	0.67 / 0.61	0.77 / 0.61	joint adapter	0.49 / 0.48	LoRA 2B	0.79 / 0.64	0.69 / 0.57
LIBERO-30 adapter	0.65 / 0.57	0.75 / 0.59	CALVIN adapter	-0.02 / 0.02	full FT 2B (C2)	0.66 / 0.46	0.47 / 0.35
trained on LIBERO-10	0.96 / 0.94	0.97 / 0.94	trained on LIBERO-10	0.96 / 0.94	LIBERO adapter 2B	0.35 / 0.09	0.23 / 0.02

Table 3: Evaluation-time interventions on the head input, four matched LoRA policies and one full-finetuned policy (LIBERO success %, 4 suites × 10 tasks × 5 initial states = 200 episodes per cell; PCA basis from each policy's own features; feature dimension 4,096 / 2,048 / 4,096 / 4,096 / 2,048). One training seed per column and one evaluation per cell (no SD); the un-ablated row is this 200-episode subset, so it differs slightly from the 2,000-episode values of Table 1; – = not run. Heads trained from scratch through a 64- or 16-dimensional bottleneck (last row) are a different experiment: they show what a head can learn to use, not what these heads use. CALVIN: Table 17.

	Emb-8B	Emb-2B	Instr-8B (last)	Instr-8B (mean)	Full FT 2B (fp32, $1e-5$)
un-ablated	97.5	98.5	98.5	98.0	98.0
keep top-4 / top-16	33.0 / 81.5	37.0 / 80.0	7.5 / 73.5	4.5 / 71.5	– / 79.5
keep top-32 / top-64	89.5 / **96.0**	87.0 / **96.0**	84.5 / **93.0**	77.5 / **88.0**	86.5 / **96.5**
keep random-64	0.0	0.0	0.0	0.0	0.0
delete top-1	2.5	3.5	4.0	9.0	12.0
delete random-1 / random-16	98.0 / 97.5	99.0 / 99.0	99.0 / 99.0	98.0 / 97.0	–
delete PCs 2–17 (variance-matched)	0.0	0.0	0.0	0.0	0.0
head trained through 64 / 16 directions	98.4 / 97.7 (Emb-2B, 2,000 episodes, one seed each)				

frames (top-8 0.88–0.93) against 0.01–0.25 for the frozen backbone; a linear reader given task identity, a progress bin and the proprioceptive state reaches only 0.27–0.29 within task (0.24–0.34 total; Table 13).

It transfers in part beyond the training episodes, tasks and scene. Table 2 separates three settings. When the backbone never trained on the ten LIBERO-10 tasks but the reader is fit on their frames, two adapters trained on the other thirty tasks read the within-task action at 0.65–0.67 (top-8 0.57–0.61) against 0.31 for the frozen backbone and 0.96 for an adapter that trained on them. When neither the backbone nor the reader saw those tasks, the adapter jointly trained on CALVIN and the thirty tasks still reads 0.49 / 0.48, whereas the frozen backbone's reader, fit on other tasks, is worse than the mean predictor (-1.88); an adapter finetuned on CALVIN alone transfers nothing to LIBERO-10 (-0.02), so in this test the concentrated feature transfers within the task family it was finetuned on, not beyond it. On CALVIN, readers fit on the training scenes retain 83–89% of their R^2 on the unseen test scene for the LoRA policies. Held-out episodes guard against reader memorisation only, since every adapter was finetuned on them; the task and scene settings show that the concentrated feature survives, in part, where the adapter never trained (the scene test set is small: 822 samples).

5 Working policies depend on the concentrated directions

Decodability shows what a reader can extract, not what the policy uses. To test dependence we intervene on the pooled feature at evaluation time, in the PCA basis of each policy's own finetuned features, and measure closed-loop success on the four LIBERO suites (Table 3; 200 episodes per condition, fixed initial states).

A few dozen directions carry the policy. Keeping only the top 64 directions (1.6% of the 8B feature, 3% of the 2B) retains 88–96% success against 97.5–98.5% un-ablated; the top 32 retain

78–90% and the top 16 retain 72–82%, so the knee lies between 32 and 64 directions for three of the four heads and just above 64 for Instr-8B (mean); 64 random directions retain nothing, and the full-finetuned 2B policy behaves like the LoRA policies (96.5% through its top 64). On CALVIN, the same intervention on the Emb-8B policy retains 4.53 of an un-ablated 4.70 average chain length (200 chains) through the top 64 directions, 4.25 through the top 32, 2.34 through the top 16 and nothing through 64 random directions (Table 17); there the leading direction holds 24% of the variance rather than 72% and the participation ratio is 14 rather than 2, so how much the first direction carries is benchmark-specific while the few-dozen-direction dependence is not. Heads trained from the start through a 64- or 16-dimensional bottleneck reach 98.4% and 97.7% (Emb-2B, one seed each), so the knee is a property of the trained head, not a floor on the representation (Table 3).

Deletion shows the leading directions are necessary for the trained head, not that they are action-specific. Deleting the single leading direction collapses success to 2.5–9% on the LoRA policies and to 12% on the full-finetuned one, whereas deleting one or sixteen random directions changes nothing. But deleting principal directions 2–17, which carry 14–24% of the variance, also collapses every policy to 0%: every deletion tested that removes more than a tenth of the variance collapses the policy, so the two readings (the leading direction carries the code the head reads; the head tolerates no large input shift) cannot be separated. On CALVIN deleting the less dominant leading direction halves the chain length (2.26) while deleting directions 2–17 still collapses. The pre-registered variance-matched control thus failed on every backbone, and we do not claim that deletion isolates causal action directions. Keep-only interventions also remove variance, so what they establish is that a subspace of a few dozen directions, which also carries most of the variance, suffices for the trained head.

Without the concentrated feature, the same small head supports only a weak policy under the conditions evaluated. The frozen backbone's diffuse information is linearly accessible with enough frames, yet a fresh MLP head trained on all demonstration frames uses it only partly: it reaches 1.07 of 5 subtasks on CALVIN (60% on the first subtask) against 4.58 for the same head on a CALVIN-finetuned backbone (Emb-8B adapter, 200 chains, where the 8B frozen base reaches 0.90; own heads 4.35 and 4.70), and 25.95% on the 40-task LIBERO generalist at 2B (2,000 episodes; 22.1% at 8B on 800) against 97.3% on the LoRA-2B backbone (own head 97.5%; Appendix F). The LIBERO values were measured only under the "next 10" evaluation prompt (Sec. 3), whose effect on the frozen backbone is not known, so they are reference values under that condition, not a bound on what the frozen feature supports, and a weak small-MLP readout does not show that a larger or non-linear readout would be weak.

6 Action decodability does not guarantee reliable control

Secs. 4 and 5 measure how readable the action is on demonstration frames and show that working policies depend on the directions that carry it. This section asks what decodability does not tell us: whether such measurements are enough to judge a policy's reliability when its inputs change. They are not. We study input sensitivity in four full-finetuning policies, two per benchmark; three were probed and carry concentrated, readable features (LIBERO top-8 R^2 0.91 and 0.82, Table 19; CALVIN Emb-2B 0.46, above the frozen base's 0.05 and below LoRA's 0.64, Table 2), and the fourth (Instr-2B on CALVIN) is a behavioural replication only. Controlled input interventions switch some of their tasks between near-complete or complete success and complete failure, and other deficits remain that no tested input change explains. All four are one seed at one learning rate (Table 1); nothing below is a statement about full finetuning in general.

6.1 CALVIN: a texture swap switches task-specific failures on and off

On the held-out scene D the full-finetuning policies score 0.95 (Emb-2B, C2) and 0.71 (Instr-2B, E1) against 4.35 ± 0.07 for LoRA. On the first subtask of each chain (the evaluation list is deterministic, so every policy attempts it on identical states), both policies solve *move slider* 0 of 151 and *turn on / off the lightbulb* 0 of 75, while solving the drawer subtasks 91–94% and the LED button 110 of 110 (LoRA: 100% on all four). Evaluated instead on training scene A with the same 1,000 chains, initial conditions and oracle, the same checkpoints solve the slider 151 of 151 and the lightbulb 75 of 75 and reach 1.79 (C2) and 1.10 (E1); LoRA (seed 0) scores 4.39 on A and 4.40 on D. Tables A

Table 4: A texture swap switches two CALVIN skills off and on. First-subtask successes on *move slider + turn on / off lightbulb* (226 of the 1,000 chains start with one) per geometry × texture; native scenes: A (training), D (held-out). Same checkpoints, chains, initial states and oracle throughout; only the two wood-texture images differ between columns. Average chain length in parentheses. LoRA on native scene D: 226 / 226 (4.40); LoRA was not run on D geometry + A texture. Details: Appendix F.

Policy	Table geometry	A texture	D texture
Full FT Emb-2B (C2)	A (training)	226 / 226 (1.79)	0 / 226 (1.02)
	D (held-out)	223 / 226 (1.27)	0 / 226 (0.95)
Full FT Instr-2B (E1)	A (training)	226 / 226 (1.10)	0 / 226 (0.47)
	D (held-out)	226 / 226 (0.93)	0 / 226 (0.71)
LoRA Emb-2B, seed 0	A (training)	226 / 226 (4.39)	226 / 226 (4.40)

Table 5: LIBERO success of the full-finetuning 3×10^{-5} policies under the paper's evaluation prompt ("next 10") and the training prompt ("next 8"). 2,000 episodes on identical initial states (fresh head, seed 1: goal suite, 500 episodes). Controls: full FT 10^{-5} 97.90 / goal 97.8; the four LoRA cells 97.4–98.2 / 97.5–98.5 (200 episodes).

Backbone	Readout	"next 10"	"next 8" (training)
Full FT 3e−5, seed 0	own head	46.90	94.80
	fresh head	47.10	95.3
Full FT 3e−5, seed 1	own head	15.00	74.9
	fresh head (goal)	11.0	86.2

and D place the slider door and the light switch at identical positions; they differ in the table mesh and its wood texture, block sizes, and the drawer ($+10$ cm) and LED button ($+20$ cm) positions (Appendix F).

A single-factor swap isolates the texture. We replaced only the two wood-texture images of table A by table D's (*A geometry + D texture*) and conversely (*D geometry + A texture*), leaving meshes, joint positions, block sets, initial states, task oracle, weights and heads unchanged, and re-ran the same 1,000 chains under a registered rule (slider + lightbulb first subtasks $\leq 20\%$ under D's texture and $\geq 80\%$ under A's). In both geometries (Table 4), the texture alone moves these 226 first subtasks between near-complete or complete success (223–226 of 226) and complete failure (0 of 226), in both directions, for both backbones; the LoRA policy solves all 226 in the three combinations on which it was run, so the swap is invisible to this working policy (paired per-chain statistics: Appendix F). The scene-D scores remain the benchmark result; the swap establishes that, for these checkpoints, a change of the table's wood texture with geometry held fixed suffices to remove and to restore two skill families, without identifying how the network implements the dependence.

6.2 LIBERO: THE EVALUATION PROMPT, CORRECTED, AND A ONE-TOKEN SENSITIVITY

Every original 2,000-episode LIBERO evaluation formatted the instruction as "Predict the next 10 robot actions as a 7-DoF chunk", a literal inherited from the chunk-10 recipe, whereas the matched cells trained with "next 8"; the policy predicts and executes 8 actions either way, and the intervention evaluator of Sec. 5 and the CALVIN evaluator use the training value. We therefore re-evaluated the two full-finetuning 3×10^{-5} seeds on the same 2,000 initial states with the training prompt (Table 5). Their success rises from 46.9% to 94.8% and from 15.0% to 74.9%; paired per episode, seed 0 gains 410 / 301 / 19 / 263 episodes on the goal / spatial / object / long-horizon suites and loses 1 / 12 / 0 / 22, and each of its twelve tasks at 0 of 50 reaches 42 of 50 or more. A fresh head trained on the seed-0 features follows the prompt the same way (95.3% vs. 47.1%; seed 1, goal suite: 86.2% vs. 11.0%, own head 8.4%): a fresh head reproduces the sensitivity, so it is not a property of the co-trained head, and the earlier reading that a fresh head does not rescue the policy held only under the mismatched prompt. No working cell tested changes comparably: the full-finetuning 10^{-5} policy scores 489 vs. 488 of 500 on the goal suite under the two prompts, and the LoRA and bf16 cells score 88.5–98.5% under "next 8" on the 200-episode subset against 86–98% for the same cells under "next 10" (Sec. 3). The corrected values replace 46.9 / 15.0, and "next 10" is retained as a labelled sensitivity test: one token naming the wrong horizon removes half and four fifths of these two policies' success while no working cell tested moves comparably.

Demonstration-frame readability registers the change only weakly. Encoding the same 73,028 demonstration frames under both prompts, the seed-0 reader fitted under "next 8" and applied un-

changed to the "next 10" features reads the action at top-8 R^2 0.897 (0.917 under "next 8", re-extracted; 0.898 when refit), and the policy's own head at R^2 0.927 (0.949); for seed 1 the same values are 0.780 (0.828; 0.777 refit) and 0.814 (0.877); the 10^{-5} and LoRA features move by less than 10^{-3}. What moves is concentrated: across the 40 tasks, the increase of the own head's open-loop error under "next 10" predicts the per-task success drop (Spearman $\rho = 0.76$, $p = 1.3 \times 10^{-8}$); the nine tasks whose error rises most (1.8–4.4×) all collapse, eight to 0 of 50, while the object suite, whose error barely rises, stays at 96–100%.

6.3 Not every deficit is explained by one input change

The LED button, which moved 20 cm between the tables, is solved 110 of 110 by both full-finetuning policies on scene D, but 0 of 110 by both on D geometry with A texture (on A geometry with D texture C2 107 of 110, E1 0 of 110; E1 57 of 110 on scene A itself): this failure depends on the combination of texture and geometry. Block manipulation on the table is poor under every condition, the training scene included (lift from table: C2 3 / 73 on A, 28 / 73 on D, 36 and 17 of 73 on the hybrids; E1 10 / 73 on A; LoRA 73 / 73), a control deficit that no scene or texture change removes and the reason the chain lengths on scene A stay at 1.79 and 1.10. On LIBERO, seed 1 keeps an unexplained long-horizon residual under the training prompt (169 of 500).

What the measurements establish. Readability of the action on demonstration frames and reliable behaviour under input changes are different properties: the three probed full-finetuning policies keep the first, and all four show task-selective sensitivity to a one-token prompt change (LIBERO) or to a table texture (CALVIN), together with deficits that persist under every input tested. We claim no common mechanism, do not attribute the sensitivity to forgetting of pretrained content (the working fp32 full-finetuning policy loses linearly recoverable image-to-text retrieval through the head's input as completely as the sensitive ones, though not text-to-image; Appendix F), and do not generalise to full finetuning or LoRA (Sec. 7). Readability under the swapped texture, and on the Instr-2B backbone, was not measured.

7 Discussion and limitations

What the results do not say, and limitations. They do not say that full finetuning fails or is fragile in general: at learning rate 10^{-5} it matches LoRA on LIBERO and shows no comparable prompt effect; the sensitive cells are two seeds at one rate on LIBERO and one seed at one rate per backbone on CALVIN, and no rate search was done; nor that LoRA is generally robust; nor that a policy uses only 16 or 64 directions (Sec. 5). They do not identify action-specific causal directions (the variance-matched deletion control failed), nor separate action information finetuning creates from information it reorganises, nor show that the policy uses the task, position and progress information linearly present in the frozen backbone. The adapter-reuse experiments of Appendix G show that a finetuned feature can support a new readout, not that its content is preserved (single seeds, subsampled). Everything is simulation on one arm (a MimicGen (Mandlekar et al., 2023) transfer attempt reaches 3.3%; Appendix G). Most cells have one training seed (three for the performance references, the Emb-2B LoRA cell on each benchmark and the pure-bf16 cell), and one cell's seed spread (± 0.49) bounds no other cell. Interventions use 200 episodes or chains per condition; the CALVIN test-scene probe uses 822 samples; the per-task decompositions are post-hoc and the open-loop-error link is correlational, on training frames; the texture swap tests one factor on two checkpoints, without probing the swapped frames; the LIBERO seed-1 long-horizon residual (33.8%) is unexplained and the frozen backbone's prompt sensitivity is unmeasured; robustness (Appendix H) rests on one contrastive–instruct pair; external numbers are as reported by their authors.

8 Conclusion

Finetuning a web-pretrained VLM into a manipulation policy concentrates action information into a few readable directions of its pooled feature on which working policies depend; the same concentration appears in the probed full-finetuning policies that break under input changes a working policy tolerates. These measurements record what finetuning concentrated, not closed-loop reliability.

AI USE STATEMENT

In this work, we used generative AI tools (Anthropic's Claude, as a coding and writing assistant) for the following tasks with required disclosure: proposing and refining hypotheses, designing and giving feedback on the experimental protocols and the pre-registered predictions, implementing the training, evaluation and probing code, and interpreting results. We have not used generative AI tools to generate synthetic data, and the remaining required-disclosure tasks (theoretical models, mathematical claims and proofs, translation, dataset cleaning, qualitative or thematic analysis) are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing parts of the paper, creating figures, editing code, searching and summarizing the literature, and formatting references. We have reviewed all AI-assisted work: every experiment was specified and approved by the authors before it ran; every number in the paper was checked by the authors against the raw logs and result files through a master results table that records the source of each printed value; every cited reference was verified against its arXiv or publisher page; and AI-drafted text was revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every policy, probe and intervention configuration is specified in Sec. 3 and Appendices D, E and F; training budgets, replan cadences and harness baselines are in Appendix K; the matched comparison and its seeds in Appendix B; the CALVIN protocol and the scene-D verification in Appendix A. Every experiment after the initial matched comparison (descriptive re-analyses of existing logs aside) was pre-registered with point predictions, intervals and decision rules before it ran, and Appendix J lists every registered quantity beside the observed value. Evaluations use fixed initial states and seeds, so all comparisons between conditions are paired. The LIBERO evaluator used for every original 2,000-episode number prompts 'next 10' while all matched cells trained with 'next 8'; the intervention evaluator and the CALVIN evaluator use the training value; the training-prompt re-evaluations of the fp32 3×10^{-5} cells (Appendix F) are paired per episode with the original runs on identical initial states (Appendix K). The per-task and per-subtask decompositions of Appendix F are computed from the evaluation logs; the CALVIN decomposition relies on the deterministic evaluation sequence list, and the scene-A and texture-swap CALVIN evaluations use the same 1,000 chains on modified scene / table assets, whose md5s are recorded in the registration files (Appendix K). Code, configurations, registration files with timestamps, cached features and evaluation logs will be released.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 7319–7328, 2021. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

Anonymous. Reference withheld for double-blind review, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Atiksh Bhardwaj, Edward Weiyi Duan, Prithwish Dan, Wei-Chiu Ma, and Preston Culbertson. Decoding task progress from VLA representations. *arXiv preprint arXiv:2608.13474*, 2026. URL https://arxiv.org/abs/2608.13474.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.09673.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734*, 2025. URL https://arxiv.org/abs/2503.14734.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning (CoRL)*, 2023. URL https://arxiv.org/abs/2307.15818.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, Feng Qiu, Heng Qu, Yifei Su, Qiao Sun, Dong Wang, Donghao Wang, Yunhong Wang, Rujie Wu, Diyun Xiang, Yu Yang, Hangjun Ye, Yuan Zhang, and Quanyun Zhou. Xiaomi-Robotics-0: An Open-Sourced Vision-Language-Action Model with Real-Time Execution. *arXiv preprint arXiv:2602.12684*, 2026. URL https://arxiv.org/abs/2602.12684.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. URL https://arxiv.org/abs/2303.04137. Extended version: IJRR 2024.

Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2505.23705.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models. *arXiv preprint arXiv:2510.13626*, 2025. URL https://arxiv.org/abs/2510.13626.

Finn Ferchau, Daniel Pommer, and Cristian Axenie. On the efficiency of LoRA fine-tuning for vision-language-action models in industrial robotic manipulation. In *International Conference on Artificial Neural Networks (ICANN)*, 2026. URL https://arxiv.org/abs/2607.10172.

Bryce Grant, Xijia Zhao, and Peng Wang. Not all features are created equal: A mechanistic study of vision-language-action models. *arXiv preprint arXiv:2603.19233*, 2026. URL https://arxiv.org/abs/2603.19233.

Asher J. Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning VLMs into VLAs without catastrophic forgetting. *arXiv preprint arXiv:2509.22195*, 2025. URL https://arxiv.org/abs/2509.22195.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2412.14803.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2212.04089.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3D Diffuser Actor: Policy Diffusion with 3D Scene Representations. In *Conference on Robot Learning (CoRL)*, 2024. URL https://arxiv.org/abs/2402.10885.

Donghoon Kim, Minji Bae, Unghui Nam, Gyeonghun Kim, Suyun Lee, Kyuhong Shim, and Byonghyo Shim. Adaptive capacity allocation for vision language action fine-tuning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. URL https://arxiv.org/abs/2603.07404.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. In *Conference on Robot Learning (CoRL)*, 2024. URL https://arxiv.org/abs/2406.09246.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025. URL https://arxiv.org/abs/2502.19645.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3519–3529. PMLR, 2019. URL https://arxiv.org/abs/1905.00414.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

Xinghang Li, Peiyan Li, Long Qian, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Xinlong Wang, Di Guo, Tao Kong, Hanbo Zhang, and Huaping Liu. What matters in building vision-language-action models for generalist robots. *arXiv preprint arXiv:2412.14058*, 2024a. URL https://arxiv.org/abs/2412.14058. v4 (February 2026); v1 titled "Towards Generalist Robot Policies: What Matters in Building Vision-Language-Action Models".

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-Language Foundation Models as Effective Robot Imitators. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://arxiv.org/abs/2311.01378.

Weifeng Lin, Siyuan Huang, Hao Li, Tingwei Chen, Ruichuan An, Xinyu Wei, Jianbo Liu, and Hongsheng Li. Rethinking VLM representation for VLA initialization. *arXiv preprint arXiv:2605.25802*, 2026. URL https://arxiv.org/abs/2605.25802.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.03310.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning (CoRL)*, 2023. URL https://arxiv.org/abs/2310.17596.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108.

Erfan Miahi and Eugene Belilovsky. Understanding and exploiting weight update sparsity for communication-efficient distributed RL. *arXiv preprint arXiv:2602.03839*, 2026. URL https://arxiv.org/abs/2602.03839.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.03740.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024. URL https://arxiv.org/abs/2405.12213.

Moritz Reuss, Jyothish Pari, Pulkit Agrawal, and Rudolf Lioutikov. Efficient Diffusion Transformer Policies with Mixture of Expert Denoisers for Multitask Learning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2412.12953.

Haoyuan Shi, Xiancong Ren, Yingji Zhang, Qinfan Zhang, Jiayu Hu, Haozhe Shan, Han Dong, Jinpeng Lu, Yinda Chen, Yi Zhang, Yong Dai, and Xiaozhu Ju. VLA-Trace: Diagnosing vision-language-action models through representation and behavior tracing. *arXiv preprint arXiv:2605.30117*, 2026. URL https://arxiv.org/abs/2605.30117.

StarVLA Community. StarVLA: A Lego-like Codebase for Vision-Language-Action Model Developing. *arXiv preprint arXiv:2604.05014*, 2026. URL https://arxiv.org/abs/2604.05014.

Aiden Swann, Lachlain McGranahan, Hugo Buurmeijer, Monroe Kennedy III, and Mac Schwager. Sparse autoencoders reveal interpretable and steerable features in VLA models. *arXiv preprint arXiv:2603.19183*, 2026. URL https://arxiv.org/abs/2603.19183.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive Inverse Dynamics Models are Scalable Learners for Robotic Manipulation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2412.15109.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2312.13139.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.01708.

Senqiao Yang, Chengyao Wang, Yuxin Chen, Zixuan Wang, Longxiang Tang, Haokun Gui, Jinhui Ye, Changsheng Lu, Xiaoyang Wu, Mingkang Zhu, Pengguang Chen, Shu Liu, Zhuotao Tian, Hengshuang Zhao, Bei Yu, and Jiaya Jia. Beyond data scaling: Representation-centric continued pre-training for vision-language-action models. *arXiv preprint arXiv:2608.27550*, 2026. URL https://arxiv.org/abs/2608.27550.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. StarVLA-α: Reducing Complexity in Vision-Language-Action Systems. *arXiv preprint arXiv:2604.11757*, 2026. URL https://arxiv.org/abs/2604.11757.

Bin Yu, Yao Zhang, Haishan Liu, Shijie Lian, Yuliang Wei, Xiaopeng Lin, Zhaolong Shen, Changti Wu, Ruina Hu, Bailing Wang, Cong Huang, and Kai Chen. RoboSemanticBench: Diagnosing semantic grounding in action prediction for VLA models. *arXiv preprint arXiv:2606.02277*, 2026. URL https://arxiv.org/abs/2606.02277.

Pedram Zamirai, Jian Zhang, Christopher R. Aberger, and Christopher De Sa. Revisiting BFloat16 training. *arXiv preprint arXiv:2010.06192*, 2020. URL https://arxiv.org/abs/2010.06192.

Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. VLM4VLA: Revisiting vision-language-models in vision-language-action models. *arXiv preprint arXiv:2601.03309*, 2026a. URL https://arxiv.org/abs/2601.03309.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025. URL https://arxiv.org/abs/2507.04447.

Wenyao Zhang, Bozhou Zhang, Zekun Qi, Wenjun Zeng, Xin Jin, and Li Zhang. Disentangled Robot Learning via Separate Forward and Inverse Dynamics Pretraining. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://arxiv.org/abs/2604.16391.

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Ran Cheng, Yaxin Peng, Chaomin Shen, and Feifei Feng. ChatVLA: Unified multimodal understanding and robot control with vision-language-action model. *arXiv preprint arXiv:2502.14420*, 2025. URL https://arxiv.org/abs/2502.14420.

A PERFORMANCE REFERENCES AND EXTERNAL COMPARISONS

Table 6: Best-recipe generalist vs. the strongest robot-pretrained LIBERO policy (500 trials per suite; Ours per-suite values are seed 0, the mean is over three training seeds with the sample SD). OpenVLA-OFT rows are official (Tables XIV and I of Kim et al., 2025); our-harness baselines are in Appendix K.

Model	Robot pretraining	Spatial	Object	Goal	Long	Mean
OFT, combined	970k traj.	97.7	98.0	96.1	95.3	96.8
OFT, per-suite specialists	970k traj.	97.6	98.4	97.9	94.5	97.1
Ours-8B	**none**	**98.8**	**100.0**	97.2	**96.0**	**98.0±0.2**

LIBERO. With no robot pretraining (the robot demonstrations of the target benchmark are the finetuning data), the best-recipe generalist is at least as strong as OpenVLA-OFT's official combined policy (Table 6). Two asymmetries are disclosed: our policy uses proprioception and a chunk-10 open loop where OFT re-plans every 8 actions, and the OFT checkpoint scores 95.5 on our harness (official 96.8; Appendix K). These references establish only that the finetuned models are working policies; the paper's subject is what finetuning changes in the representation.

CALVIN. On CALVIN ABC→D, 4.62 ± 0.05 (three seeds, 1,000 chains) matches the strongest video-pretrained tier (DeFI 4.51) with no robot or video pretraining, above the best web-VLM-only entry (RoboVLMs 4.25), on one GPU in one epoch; Xiaomi-Robotics-0 (robot and video pretraining, 4.75) remains above.

Table 7: CALVIN ABC→D (train scenes A–C, zero-shot scene D; 1,000 five-instruction chains, official protocol). Extra pretraining: robot or video data beyond CALVIN; "VLM only" starts from a web VLM with nothing else; "CALVIN play only" pretrains on CALVIN's own unlabelled play data. External per-position rates are as reported (arXiv pages, 2026-09-16). Ours full-data rows are mean±sample SD over 3 seeds (0.8 ep: 4.59 / 4.48 / 4.62; 1.0 ep: 4.67 / 4.60 / 4.58); per-position rates are from the best 1.0-epoch seed (Emb-2B: three seeds 4.402 / 4.264 / 4.371, sample SD; per-position rates from seed 0); other Ours rows are single seed.

Model	Extra pretrain	1	2	3	4	5	Avg. len
RoboFlamingo (Li et al., 2024b)	VLM only	82.4	61.9	46.6	33.1	23.5	2.48
GR-1 (Wu et al., 2024)	video	85.4	71.2	59.6	49.7	40.1	3.06
3D Diffuser Actor (Ke et al., 2024)	none (3D input)	92.2	78.7	63.9	51.2	41.2	3.27
MoDE (Reuss et al., 2025)	robot data	96.2	88.9	81.1	71.8	63.5	4.01
VLM4VLA (Qwen3-VL-2B) (Zhang et al., 2026a)	VLM only	94.3	88.2	83.1	77.6	71.0	4.14
RoboVLMs (Li et al., 2024a)	VLM only	98.0	93.6	85.4	77.8	70.4	4.25
Seer-Large (Tian et al., 2025)	CALVIN play only	96.3	91.6	86.1	80.3	74.0	4.28
VPP (Hu et al., 2025)	video diffusion	96.5	90.9	86.6	82.0	76.9	4.33
DreamVLA (Zhang et al., 2025)	robot data	98.2	94.6	89.5	83.4	78.1	4.44
DeFI (Zhang et al., 2026b)	robot+video	97.9	94.2	90.7	87.0	81.2	4.51
Xiaomi-Robotics-0 (Cai et al., 2026)	robot+video	100.0	98.3	96.0	92.6	88.1	4.75
Ours-8B (44% data, 0.4 ep)	none	98.0	92.5	86.0	78.6	70.2	4.25
Ours-8B (full data, 0.8 ep)	none	–	–	–	–	–	4.56±0.07
Ours-8B (full data, 1.0 ep)	none	99.4	97.5	94.1	90.4	85.8	4.62±0.05
Ours-8B (full data, 2.0 ep)	none	–	–	–	–	–	4.48
Emb-2B	none	99.0	94.7	88.7	82.2	75.6	4.35±0.07
Instr-8B (mean-text)	none	97.6	94.5	90.3	87.0	82.5	4.52
Instr-8B (last)	none	96.9	93.2	90.2	86.0	80.6	4.47
GLM-4.6V-Flash-10B (last)	none	97.6	93.8	90.1	85.1	78.8	4.45

Scene-D verification. Our harness builds the evaluation environment from the merged configuration, which specifies scene D (an assertion guards this at evaluation time); a fixed-seed re-evaluation on the official task-D configuration reproduced the 0.8-epoch value (4.56 ± 0.07). The scene-A and texture-swap evaluations of Appendix F replace the validation scene section by `calvin_scene_A` (block slots permuted by object name) or by a table whose two wood-texture images are exchanged, keep the same 1,000 chains, and reproduce the original evaluator's per-chain results on scene D; the LoRA policy scores 4.388 on A vs. 4.402 on D and 4.399 on A with D's texture (protocol in Appendix K). The GLM-family cell (GLM-4.6V-Flash, 10.3B, a different lab, architecture and corpus) drops into the pipeline with a one-line LoRA-target change.

Where the contrastive-vs-instruct gap on CALVIN lives. A per-subtask decomposition of all 7,000 evaluated chains localizes the entire gap between the Emb-8B and Instr-8B cells in one task family, pushing blocks across the re-textured tabletop (93.7% vs. 83.6%; every other family within

± 1.9); substituting only the six push-task rates into the instruct cell reproduces the contrastive score (4.63 vs. 4.62). A matched evaluation inside the training scenes (3,000 single-task rollouts) rules out a scene-shift reading: every cell pushes better on unseen scene D than in-domain, and the instruct deficit persists in-domain ($+10.4$ vs. $+11.3$ on D). The push gap is an intrinsic fine-manipulation weakness of the instruct cells, not appearance shift.

B MATCHED COMPARISON: CELLS, SEEDS AND BUDGET CURVES

Table 8: Matched-protocol LIBERO 40-task generalist (0.8M samples, 2,000 trials per entry, one training seed per cell unless marked). †Mean over three seeds: Emb-2B 97.65 ± 0.49 (97.50 / 98.20 / 97.25). Full-FT rows are one seed with fp32 master weights (two at $3e{-5}$); the pure-bf16 runs they replace are a numerics note in Appendix C. All entries were evaluated with the 'next 10' prompt (training: 'next 8'; Sec. 3); no working cell tested changes comparably. ‡training-prompt values first, 'next 10' values in parentheses (Appendix F). Reference rows are self-reported and differ in recipe (4B, full finetuning, LLaVA co-training, CoT prompt, normalized actions).

Cell	Readout	Finetuning	Success
Emb-8B	pooled (last) + 8.5M MLP	LoRA	97.7
Emb-2B	pooled (last)	LoRA	97.65†
Instr-8B (mean)	pooled (mean-text)	LoRA	97.4
Instr-8B (last)	pooled (last)	LoRA	98.2
Emb-8B, OFT port	8 tokens + 168M head	LoRA	98.3
Emb-2B, full FT (fp32)	pooled (last)	full, $1e{-5}$	97.90
same, batch 128	pooled (last)	full, $1e{-5}$, 6.25k upd.	97.50
Emb-2B, full FT (fp32)	pooled (last)	full, $3e{-5}$	94.8 / 74.9 (46.9 / 15.0)‡
StarVLA-OFT (StarVLA Community, 2026), Instr-4B	8 action tokens	full, 3.84M samples	96.6
StarVLA-α (Ye et al., 2026), Instr-4B	tokens + MLP	full, $\leq$12.8M samples	98.8

(i) Interchangeable parts at this resolution. Under one protocol (Table 8), contrastive vs. instruct backbones, 2B vs. 8B, last-token vs. mean-text pooling, and one pooled vector vs. eight action tokens all land within 0.9 points. Retraining one cell at three seeds gives 97.65 ± 0.49, a range of 0.95, so the matched comparison resolves nothing below about a point: these factors are interchangeable at that resolution, not shown equal. The design is a five-cell comparison, not a crossed factorial; one cell's seed SD is not a universal resolution.

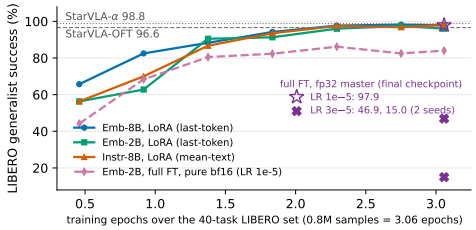


Figure 2: Success vs. training samples under the matched protocol (600-trial curve points: four suites $\times$ 10 tasks $\times$ 15 trials, single seed). The two 8B LoRA cells cross StarVLA-OFT's self-reported 96.6% by 0.6M samples and Emb-2B by 0.72M; with points wobbling 1–2 points these budgets are not separated. The dashed run is the same 2B backbone finetuned fully in pure bf16 (LR $1e{-5}$): rounding froze 92% of its weights (Appendix C); its last two points are 2,000-episode evaluations, on which it ends 13.5 points below its LoRA twin. The fp32-master cells are single final checkpoints, not swept runs: 97.90% at that rate (star) and, at $3e{-5}$, 46.9% / 15.0% (crosses, two seeds) under the 'next 10' prompt used for every curve point; the same two checkpoints reach 94.8% / 74.9% under the training prompt 'next 8' (not plotted; Appendix F), while no working cell tested changes comparably.

(ii) An early failure that was recipe, not backbone. Instr-8B (last) scored 0% in our early runs ($8\times$ fewer samples, frozen vision tower, no wrist view) and 98.2% under the matched protocol; the rescuing ingredient is not isolated (the frozen-geometry correction in Appendix D records the diagnostic error that accompanied the 0% run).

(iii) The LoRA / full-FT comparison. The plain statement of the record is as follows. Full finetuning at $1e{-5}$ with fp32 master weights works on LIBERO (97.90, one seed; 97.50 at batch 128

with 6,250 updates on the same 0.8M samples) and fails on CALVIN (0.95, one seed; Appendix F). At $3e{-}5$ it reaches 94.8 / 74.9 (two seeds) under the training prompt ('next 8') and 46.9 / 15.0 under the 'next 10' evaluation prompt, while LoRA at $3e{-}5$, the rate every LoRA cell uses, works (97.65 ± 0.49) under either prompt (200-episode baselines under 'next 8' in Appendix E). No learning-rate search was done for full finetuning: one rate per benchmark that worked (LIBERO, $1e{-}5$), failed (CALVIN, $1e{-}5$) or is prompt-sensitive (LIBERO, $3e{-}5$), so a one-rate failure is not evidence that full finetuning fails in general. What separates the regimes on LIBERO at StarVLA's rate is cost (99.7M trained parameters against 2.13B) and tolerance of the rate (at $3e{-}5$ full finetuning becomes sensitive to a one-token change of the evaluation prompt; LoRA at the same rate is not), not clean success. Recipe ingredients outside Table 8 move the number far more than any architectural factor: the best-recipe chain runs $83.4 \rightarrow 93.0 \rightarrow 98.0$ (Appendix C).

C Numerics and training recipe

Precision of the full-finetuning cells. The first matched full-FT runs kept backbone weights and AdamW moments in bf16 with no fp32 master copy, whereas the LoRA cells train fp32 adapters on a frozen bf16 backbone. bf16 resolves about 3.9×10^{-3} relative, so an Adam step of size ≈LR on a weight of magnitude 10^{-2} is lost to nearest rounding at $1e{-}5$ (Micikevicius et al., 2018; Zamirai et al., 2020; Miahi & Belilovsky, 2026). An element-wise comparison of each final checkpoint with the base (Table 9) confirms it: at $1e{-}5$ only 8% of backbone elements changed at all over 200k steps, at $3e{-}5$ 25%, word embeddings did not move, and the two $1e{-}5$ seeds agree to 0.1 points, so the freeze is deterministic rounding rather than noise. StarVLA's released training configuration keeps fp32 master weights (DeepSpeed ZeRO-2), as does a real-robot LoRA-vs-full study, which keeps fp32 optimizer states for full finetuning (Ferchau et al., 2026), so copying its learning rate did not reproduce its optimizer conditions. Under pure bf16 the three $1e{-}5$ seeds reached 82.30 ± 5.07 (84.0 / 86.3 / 76.6) against the LoRA twin's 97.65 ± 0.49, and one seed at $3e{-}5$ reached 96.55; the bf16 success at $3e{-}5$ was rounding acting as a damper on a rate that is too high once every update lands. In pure bf16 the decoupled weight decay was also inert: $1 - \mathrm{lr} \cdot \mathrm{wd} = 1 - 10^{-7}$ rounds to 1.

fp32 reruns. A rerun of both rates with fp32 master weights and bf16 autocast compute, registered on 2026-09-15 (Table 28), reaches 97.90% at $1e{-}5$ (registered 95 [88, 98]; training loss 0.0066, below the registered [0.007, 0.016]) and, at $3e{-}5$, 46.90% under the 'next 10' evaluation prompt (per suite 38.0 / 96.2 / 14.8 / 38.6; registered 96.5 [92, 98.5], scored against this value; training loss 0.0076, in the range of the LoRA cells' final losses of 0.0074–0.0111 over three seeds) and 94.80% under the training prompt (95.8 / 100.0 / 96.6 / 86.8; E6c, Appendix F); a second seed at $3e{-}5$ reaches 15.00 (loss 0.033; per suite 12.8 / 38.6 / 8.4 / 0.2) and 74.9 (86.0 / 94.8 / 85.0 / 33.8) under the two prompts. So the bf16 deficit at $1e{-}5$ was a precision artifact, and the fp32 cells are the ones Table 8 carries. The $3e{-}5$ result is a prompt sensitivity at that rate (two seeds, one of which optimized to a loss as low as the working cells'), not a precision effect; the two checkpoints are analysed in Appendix F.

Table 9: Weight change of the full-FT checkpoints relative to the base, compared at bf16 resolution: fraction of backbone elements that changed at all, relative Frobenius norm of the change, and the same fraction for the vision tower and the language model. The batch-128 control moves the weights an order of magnitude less than the 200k-update run at the same rate and reaches the same success, so the size of the weight change does not set the ceiling either. The two $3e{-}5$ seeds are indistinguishable here (99.0% of elements, 0.085 and 0.082) and score 94.8 and 74.9 under the training prompt (46.9 and 15.0 under 'next 10'), so this diagnostic separates the precision regimes, not the outcomes within one. Success column: value under the 'next 10' evaluation prompt (training-prompt value in parentheses where it was measured).

Run	changed	rel. $\|\Delta\|_F$	vision	LM	success
full FT, LR $1e{-}5$, seed 1	7.9%	0.0024	17.0%	7.1%	86.3
full FT, LR $1e{-}5$, seed 2	8.0%	0.0024	17.0%	7.2%	76.6
full FT, LR $3e{-}5$, seed 0	24.5%	0.0096	47.1%	23.5%	96.55
full FT, fp32 master, LR $1e{-}5$, seed 0	98.1%	0.036	98.5%	97.5%	97.90
full FT, fp32 master, LR $3e{-}5$, seed 0	99.0%	0.085	99.1%	98.8%	46.90 (94.80)
full FT, fp32 master, LR $3e{-}5$, seed 1	99.0%	0.082	99.0%	98.8%	15.00 (74.9)
full FT, fp32 master, LR $1e{-}5$, batch 128, 6,250 updates	71.5%	0.0052	91.2%	81.6%	97.50

Learning rate and batch size. The LoRA cells use LoRA LR $3e{-}5$ and head LR $1.5e{-}4$; the full-FT cells use backbone LR $1e{-}5$ (StarVLA's rate for its batch size) or $3e{-}5$ and the same head

LR, batch 4, 200k steps, warmup 1,000, cosine decay. A rerun at StarVLA's batch size and update count (128 by accumulation, 6,250 optimizer steps) on the same 0.8M samples with fp32 master weights at $1e{-5}$ reaches 97.50% (one seed; registered 90 [75, 96.5], a miss above; final loss 0.0091, registered 0.012 [0.008, 0.020]): the sample budget, about three epochs, is what a strong backbone needs on LIBERO, not the update count, and our $4.8\times$ smaller sample budget than StarVLA's is not bought by small-batch updates. The three-seed spread of the LoRA cell (0.95 points) and of the pure-bf16 full-FT cell (9.7 points) fixed the resolution at which the matched comparison is read.

Schedule and recipe chain. The specialist recipe (25k steps at peak LR $1e{-4}$) gives the $5\times$ larger generalist corpus only 0.39 epochs (83.4% mean). Tripling the steps at the same peak LR diverges (loss $0.07 \rightarrow 0.21$ at $\sim$12k steps, at r=64 and 128): the stretched cosine keeps LoRA near peak LR too long. Lowering peak LR to $3e{-5}$ trains stably to 93.0% (single image), $+9.6$ points from schedule alone; rank is not a lever at any budget (r=128: -1.8; r=256 does not converge). Adding the wrist camera and proprioception gives 98.0 (long-horizon $81.0 \rightarrow 96.0$). Two further ingredients were tried and are reported as negative results: an isotropy regularizer (SIGReg) ties the recipe (98.30 vs. 98.0, within noise) and a PALM-style auxiliary progress head lowers CALVIN (4.547 vs. 4.672, worse in the tail).

Table 10: Capacity-placement ablations (LIBERO-spatial specialist, 500 paired trials, McNemar vs. ctrl; *: $p < 0.001$). Readout and rank additions do not help clean success and hurt visual out-of-distribution success; vision-tower LoRA helps everywhere.

Where	Variant	Clean	Colour OOD	Position OOD
—	ctrl (last-token + 1-layer MLP, r=64)	89.0	74.2	51.4
Readout	+ attention pooling	87.6	71.4	52.2
Readout	+ deeper MLP (2-layer)	87.4	63.8*	47.6
Readout	param-matched wide MLP	89.8	64.2*	53.4
LLM LoRA	r=128	88.6	62.0*	50.8
Vision tower	**+ vision LoRA (3 seeds)**	**95.5±0.9***	83.5±9.8	**65.0±1.9***

Table 11: Where the levers are and are not (summary of this appendix and Appendix B).

Intervention	Result	Reading
backbone objective (contrastive vs. instruct)	generalist 97.4–98.2 (matched comparison, Table 8); difference inside one cell's seed range	not resolved at this seed noise
pooling (last-token vs. mean-text)	instruct: 93.0 vs. 86.4 specialist; 98.2 vs. 97.4 generalist (inside seed noise)	a lever for the specialist and for robustness, not resolved on the generalist
readout capacity (attention pool / deeper / wider)	clean ties; depth hurts colour OOD -14	not a lever, harmful OOD
LoRA rank (32/64/128/256)	r=128 hurts colour -12; r=256 diverges	not a lever
finetuning regime (LoRA vs. full)	LoRA 97.65 (3 seeds); full FT with fp32 master weights 97.90 at $1e{-5}$, 94.80 / 74.9 at $3e{-5}$ under the training prompt (46.90 / 15.00 under the 'next 10' evaluation prompt; one and two seeds); pure-bf16 full FT 82.30 / 96.55 (rounding froze 92% / 75% of the weights); full FT 0.95 on CALVIN at $1e{-5}$	not a lever for clean LIBERO success at StarVLA's rate; a lever for cost, for tolerance of the rate (prompt sensitivity at $3e{-5}$), and for whether the policy works on CALVIN
vision-tower LoRA (+31M params)	specialist $+7.6$ clean, $+15.8$ position; long-10 $27 \rightarrow 62$ at matched budget	lever; a real-robot π_0 study needs the vision encoder fully tuned for precision assembly (Ferchau et al., 2026)
schedule (0.39→1.16 ep, LR $1e{-4} \rightarrow 3e{-5}$)	generalist $83.4 \rightarrow 93.0$; naive $3\times$ steps diverges	lever
wrist camera + proprioception	$93.0 \rightarrow 98.0$; long-10 $81 \rightarrow 96$	lever

D PROBE PROTOCOLS AND FULL SWEEP TABLES

Probe protocols. The 400-frame within-task probe draws 40 frames from each of the 10 tasks of a suite, fits a ridge reader per target variable, and reports within-task R^2 (the target is centred per task, so task identity cannot leak into the score); the two action values quoted from it are 0.49 for the LIBERO-spatial specialist's full pooled feature with state and 0.38 for the vision-LoRA generalist's backbone-only feature. The ridge sample-size sweep (Table 12) is a different reader on different data (all frames of the 40-task training set, episode-split test, float64 solve, regularizer chosen on a validation split relative to $\mathrm{tr}(G)/D$); its numbers are never mixed with the 400-frame values. All sweeps use frames of the training data with held-out episodes: probe frames are not the policy's training frames in the sense of a held-out set, and the probe test split is not backbone-unseen data. Probes on test-scene and unseen-task frames, which are the two settings where the data is unseen by the finetuned backbone, are in Table 14.

Table 12: Ridge sample-size sweep (episode-split test, λ tuned on a validation split; "top-8" = reader restricted to the training split's eight leading principal directions). LIBERO rows: the four matched generalists and their frozen bases (single training seed each). CALVIN rows and the last two rows: the Emb-8B line on an earlier subsample with the same reader.

	full feature			top-8 directions		
Feature (probe frames →)	400	40k	120k	400	40k	120k
Emb-8B, frozen base	0.33	0.67	0.68	0.01	0.03	0.04
Emb-8B, finetuned	0.95	0.96	0.96	0.93	0.93	0.93
Emb-2B, frozen base	0.26	0.63	0.64	0.02	0.03	0.03
Emb-2B, finetuned	0.95	0.96	0.96	0.93	0.93	0.93
Instr-8B (last), frozen base	0.27	0.63	0.65	0.04	0.08	0.08
Instr-8B (last), finetuned	0.95	0.96	0.96	0.90	0.89	0.89
Instr-8B (mean), frozen base	0.32	0.65	0.66	0.02	0.05	0.05
Instr-8B (mean), finetuned	0.95	0.96	0.96	0.91	0.91	0.91
Emb-8B, frozen base, CALVIN	0.16	0.44	0.47	–	0.08	–
Emb-8B, CALVIN-finetuned, CALVIN	0.75	0.81	0.82	–	0.62	–
Joint LIBERO-30+CALVIN adapter, LIBERO	0.82	0.88	0.88	–	0.78	–
CALVIN-finetuned backbone, read on LIBERO	0.24	0.60	0.62	–	0.13	–

Table 13: Task-centred sweep and grounding oracles (pre-registered 2026-09-12T16:03Z, before the run; same cached features and reader as Table 12). "within" = R^2 against the per-task mean-action predictor, with per-task means of z and of the action taken from the training draw; oracles use the same draws: *task* = per-task mean action, *template* = per-(task, 20 progress bins) mean action, *grounding* = ridge on [task one-hot, progress-bin one-hot, 8-d proprioceptive state]. The grounding oracle's within-task R^2 is 0.27 at 400 and 0.29 at 120k frames for every backbone. Bottom block: LIBERO-10 frames only (tasks the joint and LIBERO-30-only adapters never trained on).

	within, full feature		within, top-8		oracles, total (400 / 120k)	
Feature (probe frames →)	400	120k	400	120k	template	grounding
Emb-8B base / finetuned	0.16 / 0.94	0.66 / 0.96	0.18 / 0.93	0.18 / 0.93	0.14 / 0.74	0.24 / 0.34
Emb-2B base / finetuned	0.01 / 0.93	0.61 / 0.95	0.12 / 0.92	0.14 / 0.93	0.12 / 0.74	0.24 / 0.34
Instr-8B (last) base / finetuned	0.10 / 0.94	0.63 / 0.96	0.12 / 0.88	0.13 / 0.89	0.12 / 0.74	0.24 / 0.34
Instr-8B (mean) base / finetuned	0.25 / 0.95	0.64 / 0.96	0.15 / 0.90	0.17 / 0.91	0.12 / 0.74	0.24 / 0.34
per-task mean action (total R^2)	-0.04 at 400 frames, 0.07 at 120k (between-task share)					
LIBERO-10 frames only (400 / 40k probe frames): within, full feature | within, top-8						
frozen base	0.31 / 0.64 | 0.08 / 0.09				template 0.38 / 0.63; grounding 0.29 / 0.34	
joint adapter (LIBERO-10 unseen)	0.67 / 0.77 | 0.61 / 0.61					
LIBERO-30-only adapter (unseen)	0.65 / 0.75 | 0.57 / 0.59					
LIBERO adapter (trained on LIBERO-10)	0.96 / 0.97 | 0.94 / 0.94					

Table 14: Probes on test-scene and unseen-task frames (ridge R^2 of the action chunk at N=120k reader frames, full feature / top-8 PCs of the reader's frames; registered intervals in the scorecard tables of Appendix J). *CALVIN scene D:* reader fit on training-scene (ABC) features, in-domain reference on held-out ABC episodes, tested on 822 scene-D samples from 17 annotated windows and 6 task names of the CALVIN debug set; this test set is small. *LIBERO unseen tasks:* reader fit on tasks 10–39, in-domain reference on their held-out episodes, tested on all 98,816 LIBERO-10 frames. Every other ridge sweep, task-centred sweep and PCA basis in this paper uses training-scene frames (LIBERO: demonstrations of the evaluation tasks, held-out episodes; CALVIN: scenes A–C, held-out episodes).

	in-domain		test-scene / unseen-task	
Feature	full	top-8	full	top-8
CALVIN: reader fit on scenes A–C, tested on scene D				
Emb-2B LoRA CALVIN adapter (the 4.40 policy)	0.79	0.64	0.69	0.57
Emb-2B full FT, fp32 master, LR $1e{-5}$ (C2)	0.66	0.46	0.47	0.35
Emb-2B frozen base	0.42	0.05	0.35	0.01
Emb-2B LIBERO adapter	0.35	0.09	0.23	0.02
Emb-8B CALVIN adapter	0.82	0.62	0.68	0.54
Emb-8B frozen base	0.47	0.08	0.37	0.03
Emb-8B LIBERO adapter	0.43	0.17	0.34	0.19
LIBERO: reader fit on tasks 10–39, tested on LIBERO-10 (Emb-8B)				
L_L: LIBERO adapter, trained on LIBERO-10 (reference)	0.96	0.93	0.96	0.94
J_L: joint adapter, never trained on LIBERO-10	0.95	0.91	0.49	0.48
L_C: CALVIN adapter	0.70	0.16	-0.02	0.02
L_0: frozen base	0.74	0.07	-1.88	-0.03

Effective dimension is setup-dependent. The participation-ratio effective dimension of the finetuned pooled feature is 2.4 for the single-image vision-LoRA generalist (11.0 at initialization, falling progressively over the second half of training), 11.5 for the LIBERO-spatial specialist, 12.5 for the CALVIN policy and 7.4 ± 1.1 for the full-stack recipe over three seeds. We therefore report the concentration descriptively and make no universal "compresses to k dimensions" claim; the claims about what the policy depends on are made with eval-time interventions instead (Appendix E).

CALVIN reproduces the concentration. On 408 CALVIN frames the within-task action R^2 of the pooled feature moves from -0.55 (base) to 0.40 after finetuning, comparable to the 0.38 (generalist, backbone-only) and 0.49 (specialist, with state) LIBERO values.

A correction on frozen-geometry diagnostics. In early experiments we attributed the instruct backbone's 0% (last-token pooling, weak recipe) to a "collapsed manifold" measured by a tiny raw within/between-task cosine gap (0.007 vs. the contrastive backbone's 0.066). The raw gap is dominated by the feature mean: after de-meaning it inflates 51–139×, functional probes show instruct pooled features carry scene and proprioceptive information at parity with the contrastive model under every pooling, and trainability refutes the inference outright (the same cell trains to 98.15%). Frozen-geometry summaries do not predict trainability, and cosine diagnostics must be reported de-meaned.

E Intervention protocols and full tables

Protocol. For each matched generalist, a PCA basis (float64 SVD) is fit on an episode-stratified 60k-frame subsample of its own finetuned pooled z over the training frames, together with a seeded random orthonormal 128-basis. At evaluation the head input is replaced by $z' = z - UU^\top(z-\mu)$ (delete), $z' = \mu + UU^\top(z-\mu)$ (keep-only), or the same with U spanning principal directions 2–17 (variance-matched deletion control), in float32, with the evaluation prompt horizon equal to the training chunk ('next 8'; `eval_ours_libero_ablate2`), unlike the 2,000-episode evaluations, which used a literal 'next 10' (Sec. 3; Appendix K). Each condition is scored on 4 suites × 10 tasks × 5 initial states (200 episodes); extra initial-state blocks (5–9, 10–14) were run for the contrastive 8B and instruct last-token models. Decision rules (registered): necessity holds on a backbone iff delete top-1 $<$ 10 and delete random-1 $\geq$ baseline $-$ 10; sufficiency holds iff the smallest $k \in \{4, 16, 32, 64\}$ with keep-only top-$k \geq 90\%$ of baseline is ≤ 16 and keep-only random-16 $< 30\%$ of baseline; both on $\geq$ 3 of 4 backbones would have licensed "necessary and sufficient" in the abstract; sufficiency needing $k > 64$ on any backbone licenses necessity only. The registered verdict

is "necessity only"; because the variance-matched deletion control also collapses (below), deletion does not identify action-specific causal directions, and the main text states dependence through the keep-only results (Table 3).

Table 15: Explained variance of the finetuned feature's principal directions, and per-suite keep-only results (four matched generalists; 200 episodes per condition).

	Emb-8B	Emb-2B	Instr-8B (last)	Instr-8B (m-text)
variance in PC1	71.5%	76.0%	66.5%	75.2%
variance in PCs 2–17	18.6%	17.3%	24.3%	13.6%
cumulative, top-16 / top-64	89.9 / 93.5%	93.1 / 96.1%	90.6 / 93.9%	88.6 / 92.0%
participation ratio	1.9	1.7	2.2	1.8
keep top-16 (sp / ob / go / long)	88 / 100 / 100 / 38	86 / 100 / 82 / 52	84 / 100 / 72 / 38	88 / 98 / 70 / 30
keep top-64 (sp / ob / go / long)	94 / 100 / 100 / 90	96 / 100 / 96 / 92	94 / 100 / 98 / 80	92 / 100 / 88 / 72
delete top-1 (sp / ob / go / long)	4 / 0 / 6 / 0	6 / 0 / 8 / 0	6 / 0 / 10 / 0	8 / 6 / 22 / 0
init-state blocks 5–9 / 10–14: baseline	97.0 / 97.5	–	98.5 / 99.0	–
keep top-16	85.0 / 86.0	–	77.5 / 76.0	–
delete top-1; delete random-1	3.5, 1.0; 97.5, 98.0	–	4.5, 4.0; 97.5, 97.5	–

Table 16: Full-finetuning checkpoints probed with the intervention pipeline (E3, registered 2026-09-15; 200 episodes per condition, init block 0–4; ridge readers as in Table 12). The reference is the Emb-2B LoRA cell. The fp32 column is the fp32-master $1e{-5}$ cell (E3b, registered 2026-09-16; 98% of elements moved); the bf16 columns are pure-bf16 runs (25% and 8% moved). The bf16 $1e{-5}$ seed-2 checkpoint (76.6%) was swept but not intervened on: full / top-8 R^2 at 400 frames 0.83 / 0.75. The 2,000-episode success row is under the 'next 10' evaluation prompt and the interventions are under the training prompt 'next 8'; for these working cells no prompt effect of the kind seen in the $3e{-5}$ cells was found (the $1e{-5}$ own head: 489 vs. 488 of 500 on the goal suite, paired; the LoRA and bf16 cells: 88.5–98.5% on the 200-episode intervention subset under 'next 8' against 86–98% for the same cells over 2,000 episodes under 'next 10'; Appendix F).

	LoRA	fp32 $1e{-5}$	bf16 $3e{-5}$	bf16 $1e{-5}$, seed 1
2,000-episode success	97.5	97.90	96.55	86.30
full R^2 at 400 / 120k frames	0.95 / 0.96	0.95 / 0.96	0.91 / 0.92	0.83 / 0.88
top-8 R^2 at 400 / 120k frames	0.93 / 0.93	0.93 / 0.93	0.89 / 0.89	0.73 / 0.74
within-task (centred) full / top-8 at 400	0.93 / 0.92	0.94 / 0.92	0.87 / 0.89	0.74 / 0.76
un-ablated	98.5	98.0	96.0	88.5
delete top-1	3.5	12.0	1.0	0.0
delete PCs 2–17	0.0	0.0	0.0	0.0
keep only top-16 / 32 / 64	80.0 / 87.0 / 96.0	79.5 / 86.5 / 96.5	74.0 / 88.5 / 91.5	12.0 / 33.0 / 60.0
keep only random-64	0.0	0.0	0.0	0.0

Table 17: Eval-time interventions on the Emb-8B CALVIN adapter (C1, registered 2026-09-16; the first 200 standard chains, identical across conditions; PCA basis from 53,610 CALVIN training frames of the same adapter; the intervention acts on z before the proprioceptive term is added; misses are above their interval). Registered values are fractions of the un-ablated chain length. Spectrum of the finetuned z: leading direction 24.3% of the variance, top 16 / 64 = 59.8 / 74.4%, participation ratio 14.1 (LIBERO Emb-8B: 71.5%, 89.9 / 93.5%, 1.9).

Condition	1	2	3	4	5	Avg. len	fraction	registered
un-ablated	99.5	97.0	94.5	92.5	86.0	4.695	1.000	4.65 [4.40, 4.85] ✓
keep top-64	99.0	96.5	91.0	86.0	80.5	4.530	0.965	0.85 [0.70, 0.95] ×
keep top-32	98.5	93.5	86.0	79.0	68.0	4.250	0.905	0.72 [0.50, 0.88] ×
keep top-16	82.0	61.0	41.0	30.0	20.0	2.340	0.498	0.55 [0.35, 0.80] ✓
keep random-64	0.0	0.0	0.0	0.0	0.0	0.000	0.000	0.05 [0, 0.25] ✓
delete top-1	73.5	58.0	42.0	30.0	22.0	2.255	0.480	0.20 [0.05, 0.50] ✓
delete PCs 2–17	0.5	0.0	0.0	0.0	0.0	0.005	0.001	0.10 [0, 0.40] ✓

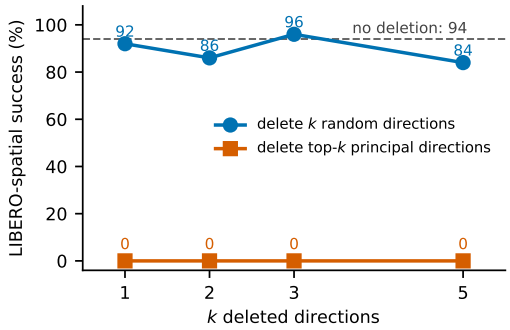


Figure 3: Deletion on the LIBERO-spatial specialist (25k steps, 50 trials per condition, single seed). Projecting out the top-k principal directions of the pooled feature collapses success to 0% for all $k = 1 \ldots 5$; the top direction alone (24.7% of variance) takes 94% to 0%. Deleting k random directions leaves 84–96% (92 / 86 / 96 / 84 for $k = 1, 2, 3, 5$).

Why the deletion control matters. On the contrastive 8B model the success after deletion tracks the variance removed almost monotonically: random-16 ($\sim$0.4% of variance) 97.5; the tail beyond PC 64 (6.5%) 96.0; beyond PC 32 (8.2%) 89.5; beyond PC 16 (10.1%) 81.5; PCs 2–17 (18.6%) 0.0; PC 1 (71.5%) 2.5. A post-hoc comparison at roughly matched variance (keep-only top-4 removes 17.6% and retains 33.0; deleting PCs 2–17 removes 18.6% and retains 0.0; 37.0 vs. 0.0 on the 2B model) hints that which directions are removed matters beyond how much variance they carry, but this comparison was not registered and we do not build on it. Deletion therefore shows that the leading directions are load-bearing and non-redundant; it does not identify action-specific causal directions. The dependence claim rests on keep-only and on the frozen-base reference (CALVIN; on LIBERO measured only under 'next 10').

Bottleneck heads (a separate experiment). Heads trained from the start through a 64- or 16-dimensional bottleneck (registered 2026-09-13; LoRA-2B cell, 2,000 episodes) reach 98.40% and 97.70% (registered 97 [93, 99], a hit, and 90 [70, 97], a miss above). A head trained to use 16 directions finds them, so the eval-time knee at 32–64 directions measures the coordinates the jointly trained head happens to use, not a floor on the action information; PCA of a trained policy and bottleneck-trained heads are different measurements and are not merged.

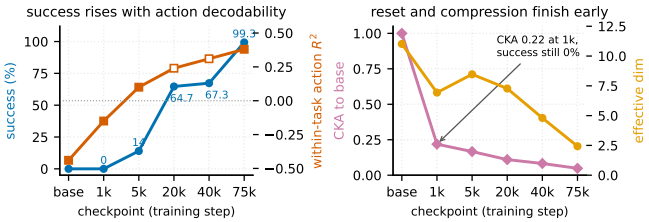


Figure 4: One training run, checkpoint by checkpoint (single-image vision-LoRA generalist, 75k steps; success on 3 tasks $\times$ 50 trials; open squares are interpolated R^2 values). Left: success rises with within-task action decodability, which crosses zero at 2–5k steps. Right: CKA to the base and the effective dimension move early; CKA is 0.22 by step 1k while success is still 0%. One run; correlational.

Trajectory (correlational). In one run probed at 13 checkpoints (1k spacing up to 3k, then 5k–75k) (Fig. 4), CKA to the base falls from 1.0 to 0.22 by step 1k, most of the geometric displacement, yet success is 0% there. Success instead rises with linear action decodability (R^2 $-0.44 \rightarrow 0.38$), which crosses zero at 2–5k steps just as success begins to climb, reaching 99% as R^2 saturates (checkpoints base / 1k / 5k / 20k / 40k / 75k: success 0 / 0 / 14.0 / 64.7 / 67.3 / 99.3; R^2 -0.44 / -0.15 / 0.10 / 0.24 / 0.31 / 0.38; effective dimension 11.0 / 6.9 / 8.5 / 7.3 / 4.8 / 2.4). This is one run and a correlation; it orders the two changes in time and does not establish that decodability causes success.

F FAILED AND PROMPT-SENSITIVE POLICIES: FRESH HEADS, PROBES, THE PROMPT MISMATCH AND THE CALVIN SCENE TESTS

Where to find what. Part (a) holds the cells with their fresh-head controls on CALVIN (Table 18) and on LIBERO under both prompts (Table 19), the feature statistics (Table 20), the probes, and the E4 / E4b retrieval measurements. Part (b) holds the per-task and per-subtask decompositions: CALVIN position-1 successes per subtask (Table 21), the scene-A and texture-swap results per task group (Table 22; chain lengths in Table 18), and the LIBERO per-task counts under both prompts with own and fresh heads (Table 23). Part (c) holds the historical analyses of the $3e{-}5$ cells under the mismatched prompt and their resolution (Table 24); part (d) the registered rules and their outcomes (with Tables 31 and 32); part (e) the corrections to earlier versions.

This appendix collects the evidence on the matched policies that fail on CALVIN (C2, E1: full finetuning of the 2B backbones at $1e{-}5$) and on the two LIBERO $3e{-}5$ seeds of Emb-2B that fail under the 'next 10' evaluation prompt and work under the training prompt 'next 8' (E6b, E6c). In each case the training recipe, budget and evaluation protocol are those of a working LoRA cell; only the regime and, on LIBERO, the rate differ. On LIBERO the evaluation protocol of the working LoRA cell carries the same prompt mismatch, under which the working cells tested show no comparable change; the CALVIN protocol has none (Appendix K). Every full-finetuning number is one seed and one learning rate per cell, and no learning-rate search was run for full finetuning. Parts (a)–(d) keep the record in the order it was built; part (e) lists what later measurements corrected.

(A) CELLS, FRESH HEADS AND PROBES

Table 18: Finetuning regime and fresh-head controls on CALVIN ABC→D, 2B backbones (1,000 chains each; one seed and one learning rate per cell). Registered values and rules: Appendix J. Full FT: fp32 master weights, bf16 autocast, backbone LR $1e{-}5$, head LR $1.5e{-}4$, 180k steps × batch 4, matched training budget and identical evaluation protocol to the LoRA cells (LoRA cells: LoRA LR $3e{-}5$, head LR $1.5e{-}4$); fresh heads are trained on cached ABC features of the frozen backbone (fresh-head protocol, Appendix G). Lower block: the same checkpoints on training scene A with the same 1,000 chains (CS, registered 2026-09-25) and on the texture-swapped tables (CT, registered 2026-09-25): A+texD = scene-A geometry, blocks and initial states with scene D's two wood-texture images, D+texA = the converse (protocol in Appendix K; per task group in Table 22).

Cell	1	2	3	4	5	Avg. len
Emb-2B LoRA, seed 0	99.0	94.7	88.7	82.2	75.6	4.402
Emb-2B LoRA, seed 1	98.7	93.9	86.5	78.5	68.7	4.264
Emb-2B LoRA, seed 2	98.5	94.5	88.5	81.1	74.5	4.371
Emb-2B LoRA, three-seed mean ± sample SD	–	–	–	–	–	4.35±0.07
Emb-2B full FT, fp32 master, LR $1e{-}5$ (C2)	53.4	25.4	10.6	3.8	1.8	0.950
Instr-2B full FT, same recipe (E1)	50.5	16.3	3.1	1.0	0.2	0.711
frozen Emb-2B base + fresh head (F0)	59.9	29.0	12.4	3.7	1.8	1.068
Emb-2B LIBERO adapter + fresh head (F1)	51.4	24.0	10.1	4.4	1.8	0.917
C2 backbone + fresh head (F2)	54.6	26.5	11.6	5.0	2.6	1.003
Emb-2B full FT (C2) on training scene A	73.6	49.0	29.7	17.6	9.4	1.793
Instr-2B full FT (E1) on training scene A	61.8	31.3	11.9	4.0	1.2	1.102
Emb-2B LoRA, seed 0, on training scene A	98.2	94.1	89.3	82.1	75.1	4.388
C2 on A+texD (A geometry, D texture)	55.4	27.1	12.2	4.9	2.2	1.018
C2 on D+texA (D geometry, A texture)	62.3	35.2	17.7	8.4	3.7	1.273
E1 on A+texD	36.9	8.1	1.0	0.6	0.0	0.466
E1 on D+texA	57.1	24.3	8.3	2.5	0.6	0.928
Emb-2B LoRA, seed 0, on A+texD	98.2	94.0	89.2	82.6	75.9	4.399

Final training losses: C2 0.0612, E1 0.0934, LoRA seeds 0.0478–0.0555; the C2 loss is within 10–28% of the LoRA cells', so the difference in training loss does not by itself account for the closed-loop gap. Validation MSEs of the fresh heads: F0 0.076, F1 0.080, F2 0.057. The 8B fresh-head references on 200 chains (C_C 4.575, C_L 1.175, C_0 0.90) are in the fresh-head transfer row of Table 27 (rounded in Table 25). The C2 feature's readability on training-scene and scene-D frames

is in Table 14 (top-8 R^2 0.46 / 0.35 against 0.64 / 0.57 for the LoRA cell and 0.05 / 0.01 for the frozen base).

Table 19: Fresh heads on the LIBERO backbones (experiment G, registered 2026-09-18 before any G job ran). Each backbone is frozen and read by a fresh MLP head trained on its cached features (all 261,614 training samples, 30 epochs, batch 512, AdamW $3e{-}4$, 2% held-out episodes); evaluation is the standard 2,000-episode protocol on the same fixed initial states as the own-head reference, so each pair is paired. Per-suite values are spatial / object / goal / long-10. Probe: ridge R^2 at 120k frames, full feature / top-8 directions. Head MSE: fresh / the checkpoint's own head, both on the same 20k training samples (the deciding pair of rule R-Gf; the frozen base has no own head). The LoRA-2B adapter's own-head value is its 2,000-episode endpoint; the 96.0 quoted as its reference at registration was the 600-trial final curve point of the same checkpoint. The standard protocol prompts 'next 10'; rows marked 'next 8' are the training-prompt re-evaluations (E6b fresh heads, E6c own heads; registered 2026-09-25), paired per episode with the 'next 10' runs on identical initial states (exact McNemar; rescued / lost per suite, spatial / object / goal / long-10: G1 own 301 / 19 / 410 / 263 rescued and 12 / 0 / 1 / 22 lost; G1 fresh 287 / 13 / 407 / 286 and 7 / 1 / 0 / 21; G2 own 371 / 291 / 385 / 169 and 5 / 10 / 2 / 1; G2 fresh, goal only, 377 and 1; G4 own, goal only, 10 and 9; G4 fresh, goal only, 7 and 10). The seed-1 fresh head and the G4 control were re-evaluated on the goal suite only.

Cell (backbone)	Head	Spatial	Object	Goal	Long-10	Mean	Probe full / top-8	Head MSE
G0 (frozen Emb-2B base)	fresh	37.0	22.2	39.0	5.6	25.95	0.639 / 0.033	–
G1 (full FT fp32, $3e{-}5$, seed 0)	own	38.0	96.2	14.8	38.6	46.90	0.945 / 0.914	0.01057
	fresh	42.0	97.4	14.2	34.8	47.10		0.00968
	own, 'next 8'	95.8	100.0	96.6	86.8	94.80		
	fresh, 'next 8'	98.0	99.8	95.6	87.8	95.3		
G2 (full FT fp32, $3e{-}5$, seed 1)	own	12.8	38.6	8.4	0.2	15.00	0.869 / 0.821	0.02745
	fresh	11.0	40.6	11.0	0.8	15.85		0.02567
	own, 'next 8'	86.0	94.8	85.0	33.8	74.9		
	fresh, 'next 8'	–	–	86.2	–	–		
G3 (LoRA-2B adapter, seed 0)	own	98.4	99.8	99.4	92.4	97.50	0.957 / 0.929	0.00776
	fresh	98.4	99.4	98.4	93.0	97.30		0.00641
G4 (full FT fp32, $1e{-}5$, seed 0)	own	99.4	100.0	97.6	94.6	97.90	0.961 / 0.925	0.00698
	fresh	98.8	99.4	98.4	94.6	97.80		0.00622
	own, 'next 8'	–	–	97.8	–	–		
	fresh, 'next 8'	–	–	97.8	–	–		

Table 20: Feature statistics of the LIBERO backbones (pooled 2048-d feature over the training frames). Mean norm and absolute maximum activation over all 261,614 cached samples (master row G-F01); PC1 share, participation ratio and the variance share of the four largest-magnitude coordinates on a 40k-frame sample (probe_results/G/G_feat_stats.json; not registered, descriptive). Success is under the 'next 10' prompt (training-prompt value in parentheses); the feature statistics are computed on the cached training-prompt features and are unchanged.

Backbone	success	mean norm	absmax	PC1 share	particip. ratio	var. in 4 largest coords
frozen Emb-2B base	25.95 (fresh head)	108.2	14.7	0.238	10.0	0.3%
LoRA-2B adapter	97.50	39.8	7.25	0.769	1.7	1.7%
full FT fp32 $1e{-}5$	97.90	50.7	9.56	0.785	1.6	2.4%
full FT fp32 $3e{-}5$, seed 0	46.90 (94.80)	90.4	51.5	0.712	1.9	12.5%
full FT fp32 $3e{-}5$, seed 1	15.00 (74.9)	100.3	59.0	0.752	1.7	18.0%

Reading of (a). On LIBERO a fresh head reproduces the backbone's prompt sensitivity: seed 0 scores 95.3 under the training prompt and 47.10 under 'next 10' (own head 94.80 / 46.90); seed 1 scores 86.2 vs. 11.0 on the goal suite (own head 85.0 / 8.4); the same pipeline reproduces a working full-FT policy (97.80 vs. 97.90) and reaches the top of the LoRA range (97.30). On CALVIN a fresh head changes nothing (1.003 vs. 0.950). The $3e{-}5$ backbones carry pooled features as concentrated as the working ones (PC1 0.71–0.75, participation ratio 1.7–1.9, against 0.24 / 10.0 for the base) and readable (top-8 R^2 0.914 for the seed-0 cell against 0.925 for the $1e{-}5$ cell; 0.821 for seed 1, 0.10 below the working fp32 cell and above the 0.743–0.764 of pure-bf16 cells that score 86.3 and 76.6). Under 'next 10' the per-suite shape survives the head swap (object 97.4, goal 14.2); under the training prompt the fresh head reaches spatial 98.0, object 99.8, goal 95.6 and long-10 87.8. Unlike CALVIN, where full finetuning landed at the frozen-base level (1.003 vs. 1.068), the LIBERO frozen-base comparison (25.95 overall; goal 39.0, object 22.2) exists only under 'next 10', the only prompt the frozen base was evaluated with. The two $3e{-}5$ cells also carry a few outlier coordinates (four coordinates hold 12.5–18% of the variance, absolute maxima 3.5–8$\times$ every other

cell's, 51.5 and 59.0 against 7.3–14.7); zeroing them at evaluation under the mismatched prompt moves success by $+0.4$ points (T2 below). The mean feature norm stayed inside the registered numerical-pathology window [10, 200], so rule R-Gk did not fire; the absolute maximum was not part of that screen and is recorded as an observation.

Pretrained image–text retrieval through the head's input (E4, E4b). Each backbone was run through the official Qwen3-VL-Embedding inference (chat template, retrieval instruction, L2-normalised final-layer last-token state), so that the measured vector is the vector the action head reads; retrieval is scored zero-shot on the Flickr30k 1k test set (Recall@1) and as linearly recoverable content on the MS-COCO 5k test set (a ridge map from the cell's vector into the frozen base's embedding space with image-level folds, Recall@1 in that space), with retention = cell R@1 / base R@1 (registered 2026-09-25 in `H2_prereg_amend1.json` and `amend2.json`). The frozen base scores 95.4 Flickr i2t R@1 zero-shot; the linear-recoverable COCO ceiling is 65.5 (i2t) / 62.7 (t2i). Zero-shot, recoverable i2t and recoverable t2i R@1 (retention in brackets): LoRA 24.1 (0.25), 43.5 (0.66), 57.5 (0.92); working fp32 $1e{-5}$ (G4) 0.1 (0.001), 1.0 (0.016), 11.9 (0.19); fp32 $3e{-5}$ seed 0 (G1) 0.0, 0.7 (0.011), 1.8 (0.03); seed 1 (G2) 0.2, 0.3 (0.005), 1.2 (0.02). Pure-bf16 cells (dose-response, descriptive): $3e{-5}$ 39.2 zero-shot / 0.51 recoverable; $1e{-5}$ seeds 1 and 2: 92.5 / 0.96 and 91.7 / 0.99. E4b repeats the linear-recoverable measurement at six readouts (last token at layers 28 and 14; mean over the modality tokens at layers 28, 21, 14 and 7): best image-to-text retention LoRA 0.88, G4 0.11, G1 0.03, G2 0.008 (all at mean-L7); best text-to-image retention LoRA 1.01, G4 0.92, G1 0.57, G2 0.12. On the image side every fp32 full-finetuning cell, the working one included, loses the linearly recoverable caption content at every readout, so loss of pretrained image–text alignment does not separate the prompt-sensitive cells from the working one; on the text side the mid-layer residue orders with the collapse (0.92 / 0.57 / 0.12) and is reported without interpretation; LIBERO success is not monotone in retention (bf16 $3e{-5}$ at 96.55% retains less than bf16 $1e{-5}$ at 76–86%; fp32 $1e{-5}$ at 97.9% retains nothing).

(B) PER-TASK AND PER-SUBTASK DECOMPOSITIONS

The per-subtask and per-task breakdowns below are reconstructions from evaluation logs already on disk (zero GPU); they were not pre-registered and are reported as descriptive. For CALVIN, the sequence list of the 1,000 chains is deterministic and each log records how many consecutive subtasks succeeded, so every attempted subtask and its outcome can be recovered; position 1 is unconditional (the same 1,000 first subtasks for every cell), so the position-1 rates are directly comparable across cells. For LIBERO, every cell attempts all 50 trials of all 40 tasks. The LIBERO logs are 'next 10' evaluations; training-prompt per-task values from E6b / E6c are given where they were run. The CALVIN scene-A and texture-swap evaluations (CS, CT) were registered before they ran (Appendix J); their per-task-group decomposition is descriptive.

Table 21: CALVIN position-1 successes per subtask (successes / attempts; the attempt count is identical across cells because the first subtask of chain i is the same for every cell). Cells: Emb-2B LoRA seed 1 (4.264); frozen Emb-2B base + fresh head (F0, 1.068); Emb-2B full FT (C2, 0.950); C2 backbone + fresh head (F2, 1.003); Instr-2B full FT (E1, 0.711). All cells on scene D; the same checkpoints on training scene A and on the texture-swapped tables are grouped by task family in Table 22. Source: probe_results/CALVIN_verb_breakdown.json, by_task_pos1.

Subtask	n	LoRA	base + fresh head	full FT	full FT + fresh head	Instr-2B full FT
open_drawer	115	115	115	115	114	115
close_drawer	41	41	41	41	41	41
turn_on_led	56	56	34	56	53	56
turn_off_led	54	54	26	54	54	54
move_slider_left	64	64	62	0	0	0
move_slider_right	87	87	81	0	0	0
turn_on_lightbulb	41	41	0	0	0	0
turn_off_lightbulb	34	34	1	0	0	0
lift_blue_block_slider	30	30	8	12	10	0
lift_blue_block_table	24	24	7	4	2	7
lift_pink_block_slider	25	25	8	11	10	0
lift_pink_block_table	25	25	24	14	16	19
lift_red_block_slider	26	26	8	10	9	0
lift_red_block_table	24	24	12	10	5	8
push_blue_block_left	35	32	12	15	26	25
push_blue_block_right	23	20	9	13	12	20
push_into_drawer	15	15	2	5	12	0
push_pink_block_left	31	29	24	23	29	30
push_pink_block_right	29	28	22	20	20	23
push_red_block_left	35	35	15	18	25	35
push_red_block_right	29	29	10	14	15	16
rotate_blue_block_left	24	24	13	15	17	0
rotate_blue_block_right	28	27	9	11	7	9
rotate_pink_block_left	23	23	18	18	16	12
rotate_pink_block_right	31	31	9	22	22	15
rotate_red_block_left	22	22	16	16	16	13
rotate_red_block_right	29	26	13	17	15	7
all subtasks	1000	987	599	534	546	505

Table 22: CALVIN first-subtask successes per task group across scenes (successes of n; the same 1,000 chains in every column, so every cell of a row is paired with every other). A = training scene A; D = scene D (the benchmark result); A+texD = scene-A geometry, blocks and initial states with scene D's two wood-texture images; D+texA = the converse. Groups: slider door and lightbulb switch (identical joint origins in tables A and D); LED button (moved 20 cm between the tables); drawer (moved 10 cm); lift a block from the slider / from the table; rotate; push. Sources: experiments/h2/CS_task_groups.json and CT_task_groups.json (master rows H2-CS-C2, H2-CS-E1, H2-CS-LORA, H2-CT-C2-A_texD, H2-CT-C2-D_texA, H2-CT-E1-A_texD, H2-CT-E1-D_texA, H2-CT-LORA-A_texD).

		Emb-2B full FT (C2)				Instr-2B full FT (E1)				Emb-2B LoRA		
Task group	n	A	D	A+texD	D+texA	A	D	A+texD	D+texA	A	D	A+texD
slider door + lightbulb switch	226	226	0	0	223	226	0	0	226	226	226	226
LED button	110	110	110	107	0	57	110	0	0	110	110	110
drawer	171	163	161	162	162	165	156	158	165	171	171	171
lift block from slider	81	75	33	56	69	9	0	0	14	81	81	81
lift block from table	73	3	28	36	17	10	34	28	19	73	73	73
rotate block	157	72	99	83	69	59	56	55	61	146	154	146
push block	182	87	103	110	83	92	149	128	86	175	175	175
avg. chain length		1.793	0.950	1.018	1.273	1.102	0.711	0.466	0.928	4.388	4.402	4.399

Two subtask families sit at exactly zero on scene D in all three full-FT variants (`move_slider` 0/151; `turn_lightbulb` 0/75). Neither is absent from the policy: on training scene A with the same 1,000 chains C2 solves them 151/151 and 75/75, as does E1 (LoRA 151/151 and 75/75 on both scenes; paired first-subtask changes on A vs. D over the 1,000 chains: C2 $+305$ / -103, $p = 2.7 \times 10^{-24}$; E1 $+276$ / -163, $p = 8 \times 10^{-8}$). The two elements have identical joint origins in tables A and D; the tables differ in mesh and texture, block sizes, and the drawer ($+10$ cm) and LED-button ($+20$ cm) positions. A registered single-factor swap of the two wood-texture images (CT; Table 22) isolates the trigger: on A's geometry with D's texture C2 solves slider + switch 0/226, on D's geometry with A's texture 223/226 (rule R-CT-texture fired; LoRA 226/226 in the three combinations on which it was run, scene A, scene D, and A geometry with D texture; A+texD vs. D paired $+86$ / -66, $p = 0.12$, i.e. indistinguishable from scene D; D+texA vs. D $+284$ / -195,

$p = 5.5 \times 10^{-5}$). The LED button (110/110 on A and on D) is 0/110 on D's geometry with A's texture and 107/110 on A's geometry with D's texture: its success depends on the combination of geometry and texture (E1, below, is 0/110 on both hybrids), and the cue the network uses is not identified. Block manipulation on the table is poor on both scenes and worst on A (lift from table 3/73 on A, 28/73 on D, 36/73 and 17/73 on the hybrids; LoRA 73/73): a control deficit that no scene or texture removes. E1 replicates the texture result (slider + switch 0/226 on A+texD, 226/226 on D+texA; R-CT-texture fired for E1 as well) with a larger block deficit (scene A 1.102, R-CS-fails; lift from slider 9/81 on A, 0/81 on A+texD, 14/81 on D+texA) and an LED button whose success depends on the geometry–texture combination (57/110 on A, where `turn_on_led` is 56/56 and `turn_off_led` 1/54; 110/110 on D; 0/110 on both hybrids); its chains are shorter on A+texD (0.466; paired first subtask vs. D $+24$ / -160, $p = 7 \times 10^{-26}$) and longer on D+texA (0.928; $+270$ / -204, $p = 0.003$) than on scene D (0.711). Excluding `move_slider`, the C2 cell is better than the frozen base at position 1 on scene D (534/849 vs. 456/849).

Table 23: LIBERO successes per task (of 50) for the two fp32 $3e{-5}$ cells and the controls under the 'next 10' evaluation prompt (own head 46.90 / 15.00; fresh head 47.10 / 15.85) and, in the three right-most columns, under the training prompt 'next 8' (E6c own heads and the E6b fresh head, 2,000 episodes each, same initial states). Columns: seed 0 with its own head (46.90) and with a fresh head (G1, 47.10); seed 1 with its own head (15.00); frozen Emb-2B base + fresh head (G0, 25.95); LoRA-2B adapter + fresh head (G3, 97.30); the working fp32 $1e{-5}$ cell with its own head (97.90) and with a fresh head (G4, 97.80); then, under 'next 8', seed 0 with its own head (94.80), seed 0 with the fresh head (95.3) and seed 1 with its own head (74.9). Spatial tasks are "pick up the black bowl ... and place it on the plate"; object tasks are "pick up the ... and place it in the basket". Sources: probe_results/LIBERO_task_breakdown.json (its `working_ft` cell is the fresh-head evaluation G4 for three suites), the per-task lines of the evaluation logs (logs/reva/REVA_fullft2b_fp32_1e-5_eval_libero_*.log, lad/logs/g_eval_G4_ft1e5_libero_10.log), and experiments/h2/E6b.json and E6c.json for the 'next 8' columns.

Task	$3e{-5}$ s0		$3e{-5}$ s1	base	LoRA	$1e{-5}$		'next 8'		
	own	fresh	own	fresh	fresh	own	fresh	s0 own	s0 fresh	s1 own
Spatial										
between the plate and the ramekin	34	32	5	8	50	50	50	50	50	47
from table center	0	0	6	46	50	50	50	50	49	49
in the top drawer of the wooden cabinet	42	38	9	13	48	50	48	43	49	39
next to the cookie box	19	19	0	35	50	50	50	47	46	46
next to the plate	4	5	0	4	50	50	50	50	50	40
next to the ramekin	2	3	0	6	50	50	50	50	50	41
on the cookie box	36	38	17	27	50	48	48	46	49	48
on the ramekin	10	29	1	20	47	49	48	48	50	44
on the stove	0	2	0	21	50	50	50	47	50	37
on the wooden cabinet	43	44	26	5	47	50	50	48	47	39
Object										
alphabet soup	43	49	35	1	49	50	50	50	50	48
bbq sauce	43	43	13	0	50	50	48	50	50	49
butter	50	50	48	7	50	50	50	50	50	50
chocolate pudding	50	50	4	8	48	50	49	50	50	50
cream cheese	50	50	0	30	50	50	50	50	50	46
ketchup	45	46	0	20	50	50	50	50	50	47
milk	50	50	13	15	50	50	50	50	50	47
orange juice	50	50	8	13	50	50	50	50	50	47
salad dressing	50	49	30	6	50	50	50	50	50	48
tomato sauce	50	50	42	11	50	50	50	50	49	42
Goal										
open the middle drawer of the cabinet	0	0	0	8	50	49	48	49	46	45
open the top drawer and put the bowl inside	0	0	0	5	48	47	47	42	43	37
push the plate to the front of the stove	0	0	9	31	48	49	50	50	50	46
put the bowl on the plate	0	0	0	5	50	49	50	49	49	41
put the bowl on the stove	0	0	0	35	50	50	50	50	47	48
put the bowl on top of the cabinet	49	50	17	30	48	48	49	49	50	43
put the cream cheese in the bowl	16	10	0	9	50	50	50	46	46	35
put the wine bottle on the rack	0	7	0	0	49	46	48	50	47	36
put the wine bottle on top of the cabinet	0	0	0	24	49	50	50	48	50	44
turn on the stove	9	4	16	48	50	50	50	50	50	50
Long-horizon (LIBERO-10)										
pick up the book and place it in the back compartment of the caddy	0	0	0	8	50	50	49	47	47	37
put both moka pots on the stove	17	10	0	0	36	46	46	33	37	2
put both the alphabet soup and the cream cheese box in the basket	22	5	0	1	50	48	48	50	50	8
put both the alphabet soup and the tomato sauce in the basket	7	5	0	0	49	48	48	47	46	10
put both the cream cheese box and the butter in the basket	43	37	0	0	49	50	50	48	49	29
put the black bowl in the bottom drawer of the cabinet and close it	0	0	0	0	47	49	49	42	43	18
put the white mug on the left plate and the yellow-and-white mug on the right plate	0	3	0	1	47	47	48	48	45	7
put the white mug on the plate and the chocolate pudding to the right of the plate	39	42	0	1	41	43	42	39	38	13
put the yellow-and-white mug in the microwave and close it	22	26	0	5	48	45	45	35	36	14
turn on the stove and put the moka pot on it	43	46	1	12	48	47	48	45	48	31
total of 2,000	938	942	300	519	1946	1958	1956	1896	1906	1498
tasks at exactly 0/50	12	9	22	6	0	0	0	0	0	0

Under the 'next 10' prompt twelve of the 40 tasks are at exactly 0/50 in the seed-0 cell (seven goal-suite, two spatial, three long-horizon tasks), nine after the head swap, largely the same tasks; 22 in the seed-1 cell; none in any working cell. Under the training prompt every one of the twelve is at 42/50 or more with the own head (the lowest of its 40 tasks is 33/50) and none of the nine is below

43/50 with the fresh head; the seed-1 own head recovers every zero task of the goal, spatial and object suites and reaches 169/500 on long-10 (33.8%; seven of its nine long-horizon zero tasks stay below 20/50), the one residual. The frozen-base comparison (five of the zero tasks solved 21–46 times out of 50 by the base read by a fresh head, "from table center" 46, "put the bowl on the stove" 35, "push the plate to the front of the stove" 31, "put the wine bottle on top of the cabinet" 24, "on the stove" 21; LoRA 48–50 on all five) holds only under 'next 10', the only prompt the frozen base was evaluated with. As on CALVIN (0.66 / 0.46), a probe averaged over all tasks reads the action chunk at R^2 0.94 / 0.91 from features of a policy that is at zero on twelve tasks under a one-token change of the evaluation prompt.

(C) HISTORICAL ANALYSES UNDER THE MISMATCHED PROMPT (EXPERIMENT H) AND THEIR RESOLUTION (E6, E6B, E6C, E7)

Experiment H (registered 2026-09-18 after the G headline, before any test ran) asked what is missing when the pooled feature is readable but the policy fails, with three hypotheses: H-A, the instruction is eroded from the feature; H-B, the feature is readable on demonstration states but the policy leaves that manifold (open-loop divergence or off-manifold features); H-C, the outlier coordinates are causal. It was registered and run with the 'next 10' evaluator before the prompt mismatch was known (E6, 2026-09-25), so its four tests are reported as behaviour under that prompt. Table 24 gives every value with the rule that fired; H-A, H-B and H-C failed as stated, and the cause identified afterwards (E6) is the prompt.

Table 24: The four H tests. T2, T3 and T4 are rollouts under the 'next 10' prompt; T1 and T1b are demonstration-frame properties under the training prompt. T1: 10-way task classifier (multinomial logistic regression, episode split, standardised features) on frames with offset ≤ 5, i.e. before the arm has moved; within the goal suite the ten tasks share one scene and differ only by instruction. T1b: task-centred ridge sweep, within-task R^2 at 120k frames. T2: the G1 assembled checkpoint ($3e{-}5$ backbone + fresh head) with the four largest-magnitude feature coordinates zeroed before the head, vs. four random coordinates; goal and spatial suites, 500 episodes each. T3: the goal suite with re-planning every step (execute horizon 1) for G1 and the working control G4, 500 episodes each. T4: pooled z of every forward pass of a horizon-8 goal-suite evaluation (17,387 rollout frames for G1, 7,241 for G4); early = model step ≤ 5 (benchmark initial states, which are new samples of the demonstration placement distribution: none within 5 mm of a demonstration layout, per-state minimum 8–13 mm, median 19 mm, the same as the demonstration-to-demonstration nearest-neighbour distance of 17–21 mm; master row H2-C1), late = step > 40; the instruction reader is the T1 goal classifier trained on demonstration first frames; NN distance = mean nearest-neighbour cosine distance of rollout frames to 20k of the cell's own goal-suite demonstration features. Sources: probe_results/H/H_instr_probe.json, H_centred_all.json, H_offmanifold_G1.json, H_offmanifold_G4.json; master rows H-T1, H-T1b, H-T2, H-T3, H-T4.

Test and quantity	Values	Rule fired
T1 (demonstration-frame property, training prompt): instruction accuracy at offset ≤ 5, full / top-8 (chance 0.10; 40-way 0.025)	goal: G1 1.000 / 0.989; G2 1.000 / 0.977; G4 1.000 / 0.986; LoRA-2B 1.000 / 0.986; base 1.000 / 1.000. object: 1.000 / 0.993; 0.999 / 0.985; 1.000 / 0.882; 1.000 / 0.930; 1.000 / 1.000. spatial: 1.000 / 1.000; 1.000 / 1.000; 1.000 / 1.000; 1.000 / 0.997; 1.000 / 1.000. long-10: 1.000 / 1.000; 1.000 / 1.000; 1.000 / 0.949; 1.000 / 0.997; 1.000 / 1.000. all 40 tasks (40-way): 0.999 / 0.985; 1.000 / 0.971; 1.000 / 0.881; 1.000 / 0.926; 1.000 / 0.980.	R-H1b (instruction present; H-A fails as stated). Registered G1 0.55 [0.2, 0.95] and G2 0.45 [0.15, 0.9]: misses above.
T1b (demonstration frames, training prompt): within-task R^2 at 120k, full / top-8	G1 0.941 / 0.912; G2 0.859 / 0.814; G4 0.958 / 0.923; LoRA-2B 0.954 / 0.926; frozen base 0.614 / 0.140. Grounding oracle within-task 0.29–0.30 for every cell. Under the evaluation prompt the aggregate top-8 reads 0.897 with the training-prompt reader and 0.898 refit (E7).	R-H1e (within-task action mapping intact). Registered 0.85 [0.6, 0.95]: hit.
T2 (behaviour under the 'next 10' prompt): G1 with the four outlier coordinates zeroed, success % goal / spatial (mean)	zero-4: 21.6 (108/500) / 35.4 (177/500), mean 28.5; random-4: 16.6 (83/500) / 39.8 (199/500), mean 28.2; un-ablated: 14.2 / 42.0, mean 28.1. Net change of the mean $+0.4$ (outliers), $+0.1$ (random); the registered clip condition was not triggered. Under 'next 8' the same checkpoint is at 95.6 / 98.0 un-ablated (E6b).	R-H2b (outliers are a symptom, not the cause; H-C fails). Registered 16 [8, 40] / 42 [25, 60] / random 14 [8, 25]: hits.
T3 (under the 'next 10' prompt): goal-suite success at execute horizon 1 (horizon 8 in brackets)	G1 13.8 (69/500) [14.2]; G4 89.2 (446/500) [98.4]. Per task at horizon 1, G1: "put the bowl on top of the cabinet" 49/50, "put the cream cheese in the bowl" 10/50, "turn on the stove" 8/50, every other goal task 0–1/50: the same dead tasks as at horizon 8. Those seven tasks are at 43–50/50 under 'next 8' (E6b).	R-H3b (re-planning does not help; not open-loop divergence); R-H3c does not fire (the protocol is valid). Registered 18 [8, 40] / 90 [70, 98]: hits.
T4 (rollouts under 'next 10'): absmax ratio, rollout-late / demonstration	G1 0.97; G4 0.96. Mean norm G1 89.3 (demo) $\rightarrow$ 83.4 (rollout late), G4 50.6 $\rightarrow$ 51.4; outlier-4 variance share G1 0.133 $\rightarrow$ 0.119, G4 0.011 $\rightarrow$ 0.012.	R-H4a does not fire. Registered G1 1.5 [1.0, 3.0]: miss below; G4 1.05 [0.9, 1.4]: hit.
T4: instruction accuracy on rollout frames, early / late	G1 0.764 / 0.163; G4 1.000 / 0.107 (training fit 1.000 for both). G1's early value becomes 1.000 when the same 500 first observations are re-encoded under 'next 8' (E6; inside the 'next 8' evaluator also 1.000); G4 is 1.000 under both prompts. The late values are not interpretable: the reader saw only first-frame states, so both cells fall to chance.	R-H4b does not fire (confounded). Registered late$-$early -0.15 [-0.5, 0.1]: formally a miss below, uninterpretable.
T4: first-observation distance to own demonstrations relative to the leave-episode-out reference (E1-R0 re-measurement of the T4 dump; the held-out baseline of an earlier version, 0.0013 / 0.0022, was computed with query frames drawn from the reference set and is withdrawn)	G1 28.7 in cosine (per task 11.9–105.1), 4.9 in standardised Euclidean distance (3.5–8.4); G4 0.94 / 0.94. Under 'next 8' (E6): G1 1.1 / 1.05, G4 0.9 / 0.93; prompt shift 5.27 vs. 0.26 LEO.	not registered in H; E6 registered separately (Appendix J).

Reading of (c). Under the 'next 10' prompt every registered mechanism failed as stated (T1–T4). The instruction is linearly present in the $3e{-}5$ backbone's pooled demonstration feature, at 1.000 for every cell including the frozen base (T1; at that level the test shows presence, not correct use). The within-task action mapping on demonstration frames is intact (T1b). Zeroing the outlier coordinates moves the mean by $+0.4$ points (T2). Re-planning every step leaves the same tasks dead (T3). Feature norms and outlier shares do not change between demonstration and rollout frames (T4), so by the registered rules off-manifold instability is not supported. The first-observation gap seen in T4 was re-measured against a leave-episode-out reference (E1-R0: G1

$\approx 29\times$ in cosine and $\approx 5\times$ in standardised Euclidean distance, G4 $\approx 1\times$; the same in two independent runs, 5.42 vs. 5.46 LEO in L2), and its image-side candidates were tested under registered rules (`probe_results/H2_prereg.json`, 2026-09-24): benchmark layouts are new samples of the demonstration placement distribution, and rolling out from the demonstrations' own layouts gives 80/500 = 16.0% under the mismatched prompt against 71 / 76 of 500 on benchmark layouts, with the feature gap at 32.7 vs. 28.7 (working backbone 0.85; E1-D); JPEG-q95 re-encoding of the live frame gives 83/500 = 16.6% ($p = 0.065$ / 0.34 against the two baseline runs), moves the feature by 0.19 LEO and does not revive the dead tasks (E1-J). The source is the prompt (E6, registered 2026-09-25 before any E6 number): recomputing the same 500 first observations with 'next 8' brings the gap to $1.1\times$ / $1.05\times$ and instruction accuracy from 0.768 to 1.000, while the working backbone moves 0.26 LEO against 5.27 for G1; the behavioural consequence is in Tables 19 and 23 (E6b, E6c). The hypothesis H-D of an earlier version (recorded demonstration video vs. live rendering) is withdrawn: the appearance difference between regenerated training frames and live rendering exists (pixel MSE ≈ 200 for the same physical state), but the $3\mathrm{e}{-5}$ backbone's feature does not react to it measurably once the prompt matches. Under the evaluation prompt the same demonstration frames still read the action at top-8 R^2 0.897 with the training-prompt reader (0.917 under the training prompt; refit 0.898; full 0.925 / 0.949, refit 0.945) and the policy's own head at 0.927 (from 0.949), while the per-task increase of the own head's open-loop error predicts the per-task success drop (Spearman $\rho = 0.76$, $p = 1.3 \times 10^{-8}$ over 40 tasks; seed 1 $\rho = 0.33$); the $1\mathrm{e}{-5}$ and LoRA features move by less than 0.001 (E7). The per-task link is correlational and the frames are training frames.

(D) The registered rules and their outcomes

Experiment G was registered as `probe_results/G_prereg.json` (v1, 2026-09-18 16:48 local time by file time) and amended before any G job ran as `G_prereg_v2.json` (16:56 local time; the first G job started 16:57; v1 is kept unedited). The amendment followed a pre-launch review that found v1's rules left part of the outcome space unassigned (G1 between 38.9 and 55 had no rule), gated the probe rule on the behavioural rule, compared a held-out MSE against an in-sample one, and had no positive control showing that the fresh-head pipeline preserves a working full-FT policy at 2B; v2 added the G3/G4 controls, the validity gate R-Gj, the unreadable branch R-Gi and the numerics screen R-Gk, re-keyed R-Gf on the matched training-MSE pair, and recalibrated R-Ge's threshold from 0.60 to 0.74 (the lowest top-8 R^2 on record for a LIBERO cell above 82% success). The v2 rules, quoted with the outcome in brackets:

> **R-Gj (validity gate, evaluated FIRST):** "IF G4 $<$ 80% THEN the fresh-head + assemble + eval path does not preserve a working full-finetuned policy, and NOTHING is concluded from G1/G2: the experiment has failed, not the backbones. Report the failure and stop. (G3 $<$ 85% is the same verdict for the non-assembled path used by G0/G3.)" [G4 97.80, G3 97.30: gate passed.] **R-Ga (rescue):** "IF G1 $\geq$ 85% THEN the 3e-5 failure is recoverable by refitting the readout alone, without touching the backbone: the collapsed backbone still carries a usable code and only the jointly trained head was broken. Sec 4.2 gains that sentence. No claim is made about WHY LoRA tolerates the rate, since LoRA's head is not what this tests." [did not fire.] **R-Gb (no rescue):** "IF G1 $\leq$ 54.9% THEN the head swap does not rescue and the backbone carries the failure, as on CALVIN (F2). Recorded as a LIBERO replication of the CALVIN result; whether Sec 7's takeaway widens beyond CALVIN is then put to the author[s] rather than assumed. If additionally G1 $<$ 38.9% the fresh head is WORSE than the collapsed run's own head – report that as a further loss, not as a rescue of any size." [fired at 47.10 under the 'next 10' prompt; the second clause did not. Under the training prompt G1 = 95.3, which satisfies R-Ga's $\geq$ 85 (E6b, registered 2026-09-25).] **R-Gc (partial):** "IF 55 $\leq$ G1 $<$ 85 THEN report the number as a partial recovery and make no claim change." [did not fire.] **R-Gd (floor):** "IF G1 $\leq$ G0 + 5 points THEN the collapsed backbone is behaviourally no better than the untouched base, which is the CALVIN F2/F0 relation (1.003 vs 1.068)." [did not fire: 47.10 vs. 25.95, both under 'next 10'.] **R-Ge (readable and unusable):** "IF G1 top-8 R2 $\geq$ 0.74 AND R-Gb fires THEN the code is readable at a level no LIBERO cell below 82% success has ever shown, yet is not usable: 'readable but not usable' is demonstrated on LIBERO as well as CALVIN. 0.74 is the lowest top-8 value on record for a LIBERO cell that reached 82%." [fired on the 'next 10' value (0.914); superseded: the code is usable under the training prompt (E6b).] **R-Gi (unreadable):** "IF G1 top-8 R2 $<$ 0.60 AND G1 $<$ 85 THEN the LIBERO collapse LOSES linear decodability, unlike CALVIN's C2 which kept it (0.46

> against a 0.05 base): the two collapses are then NOT the same phenomenon, Sec 7 stays CALVIN-scoped, and the LIBERO cell is reported as a different failure." [did not fire.] **R-Gf (head fit, matched pair):** "Compare final_train_mse with orig_head_train_mse, both on tr[:20000], samples both heads saw. IF fresh $\leq 0.8 \times$ original THEN the original head was underfit on its own features (a head-side explanation gains support). IF fresh $\geq 0.95 \times$ original THEN the original head was already near-optimal for these features and the failure is not a head-fit failure. The held-out pair (final_val_mse vs orig_head_val_mse) is reported too but is NOT the deciding statistic: the fresh head chose that split, the original head did not." [neither branch: ratio 0.916 (G2 0.935; the working G4 0.891).] **R-Gg (seed 1):** "IF G2 $\leq$ 23% (its own 15.00 + 8) THEN the no-rescue result replicates on the second seed and the LIBERO statement rests on two seeds, matching the two-seed collapse in Sec 4.2. IF G2 $\geq$ 60% THEN the more collapsed seed is rescued while the milder one may not be: report the discrepancy and make no mechanism claim from one seed each. Between 23 and 60: partial, report the number and stop." [replication branch fired: 15.85 under 'next 10'; goal 86.2 under 'next 8' (E6b).] **R-Gh (language):** "IF the goal suite recovers above 50% while the overall stays below 85% THEN report a partial, language-conditioned recovery; do not generalise it." [did not fire: goal 14.2 under 'next 10'; 95.6 under 'next 8'.] **R-Gk (numerics):** "IF G1_feat_mean_norm falls outside [10, 200] THEN the collapse has a candidate numerical explanation (activation blow-up or collapse) and the representational reading is suspended until activation statistics are checked." [did not fire: 90.4.]

Registered point estimates and 80% intervals, with the observed values: G1 50 [35, 68] (47.10, hit); G2 25 [10, 48] (15.85, hit); G0 18 [8, 32] (25.95, hit); G3 95 [85, 98] (97.30, hit); G4 95 [80, 98] (97.80, hit); G1 top-8 R^2 0.85 [0.55, 0.95] (0.914, hit); G1 full R^2 0.95 [0.85, 0.97] (0.945, hit); G0 top-8 0.05 [0.01, 0.15] (0.033, hit); fresh/original training MSE 0.90 [0.60, 1.05] (0.916, hit); G1 goal suite 25 [8, 60] (14.2, hit); G1 mean feature norm 45 [15, 120] (90.4, hit). All G outcomes were scored on 'next 10' evaluations. The training-prompt re-evaluation was registered separately (E6b, 2026-09-25 02:14 local time) and scored on its own intervals: G1 goal 60 [25, 95] $\rightarrow$ 95.6 (0.6 above the interval), G4 goal 98 [94, 100] $\rightarrow$ 97.8 (hit), G2 goal 40 [10, 85] $\rightarrow$ 86.2 (1.2 above), dead tasks revived 4 [1, 6] $\rightarrow$ 6 of 6 (hit), G1 all suites 80 [45, 97] $\rightarrow$ 95.3 (hit); these are registered 80% intervals, not confidence intervals, and a hit is not evidence of mechanism (Table 32). One registration error is on record: the reasoning field of the G1 prediction (in both v1 and v2) states that the $3\mathrm{e}{-5}$ cell "has the LOWEST training loss of any cell (0.0076)"; this is false (the fp32 $1\mathrm{e}{-5}$ cell ended at 0.0066 and four LoRA cells at 0.0059–0.0074), and no rule uses the training loss, so the error affected no outcome.

Experiment H (`probe_results/H_prereg.json`, 2026-09-18 19:21 local time by file time (T1 and T1b written at 19:22 and 19:24), before any test ran; predictions in Table 24). Rules quoted with outcomes:

> **R-H1a:** "IF G1 goal accuracy $\leq$ 0.70 AND every working cell (LoRA-2B, G4) $\geq$ 0.90 AND the base $\geq$ 0.80 THEN the instruction is linearly present in the base and in working finetunes and eroded in the collapsed one: H-A supported, and 'grounding is inherited' becomes 'inherited and destructible'." [did not fire.] **R-H1b:** "IF G1 goal accuracy $\geq$ 0.90 THEN the instruction IS in the collapsed feature; H-A fails as stated and the failure is downstream (H-B or H-C)." [fired: 1.000.] **R-H1c:** "IF the frozen base scores $<$ 0.6 THEN the pooled last token does not carry the instruction even before finetuning, the test cannot speak to erosion, and T1 is reported as inconclusive." [did not fire: base 1.000.] **R-H1d:** "IF G1 and G2 disagree by more than 0.3 THEN report per seed and make no H-A claim from one seed." [did not fire.] **R-H1e:** "IF G1 within-task top-8 $\geq$ 0.80 THEN the within-task motor code is intact and whatever is broken is not the action mapping itself." [fired: 0.912.] **R-H2a:** "IF zero-4 raises goal+spatial mean success by $\geq$ 20 points over the un-ablated 28.1 (goal 14.2, spatial 42.0) AND random-4 does not ($<$ 5 points) THEN the outlier coordinates are causal for the collapse: H-C supported." [did not fire.] **R-H2b:** "IF zero-4 changes the mean by $<$ 5 points THEN the outliers are a symptom, not the cause: H-C fails." [fired: $+0.4$.] **R-H2c:** "IF zero-4 LOWERS success by $\geq$ 10 points THEN the coordinates carry usable signal; report and stop." [did not fire.] **R-H3a:** "IF G1 goal at horizon 1 $\geq$ 50% (from 14.2 at horizon 8) AND G4 stays $\geq$ 80 THEN the failure is chunk-level open-loop divergence: the backbone's code is usable when re-read every step. H-B (divergence form) supported." [did not fire.] **R-H3b:** "IF G1 goal at horizon 1 $\leq$ 25% THEN re-planning does not help and the failure is not open-loop divergence." [fired: 13.8.] **R-H3c:** "IF G4 at horizon 1 $<$ 70 THEN horizon 1 is itself a damaging protocol change for this policy family and T3 is inconclusive." [did not fire: 89.2.] **R-H4a:** "IF G1's rollout absmax ratio $\geq$ 2.0 AND G4's $\leq$ 1.3 THEN the collapsed backbone's features destabilise

> off the demonstration manifold: H-B (off-manifold form) supported." [did not fire: 0.97 / 0.96.] **R-H4b:** "IF the instruction classifier's accuracy on G1's late rollout frames falls by ≥ 0.3 relative to early frames while G4's falls by < 0.1 THEN the instruction is lost as the episode proceeds: a second signature of H-B/H-A interaction." [did not fire: both cells fall, because the reader knows only first-frame states.] **R-H4c:** "IF neither fires THEN off-manifold instability is not the mechanism." [applies.]

The registration's stated consequence for "nothing fires" was that the sufficiency gap stays open and is stated as such; that was the paper's position before E6. The gap is closed by the prompt mismatch (E6, E6b, E6c; Table 31) and H-D is withdrawn. The first pass of the T4 instruction reader labelled rollout frames with a fixed task-id offset, which does not match the evaluator's task order, and scored chance for both cells; the corrected pass maps evaluator task ids to dataset tasks through the instruction text, and only the corrected values are reported.

Registrations of 2026-09-25 (E6, E6b, E6c, E7, CS, CT). Each was written before the corresponding measurement (`probe_results/H2_prereg_amend3.json` to `amend8.json`, mirrored in `experiments/h2/`); registration times, hypotheses, rules, outcomes and the numbers are indexed in Tables 31 and 32. In summary: R-E6-prompt fired (gap under 'next 8' $1.1 \leq 2.0$), which made the behavioural run mandatory; R-E6b-v passed (G4 97.8) and R-E6b-rescue fired (G1 goal 95.6; 6 of 6 dead tasks alive); R-E6c-headline fired (seed-0 own head $94.80 \geq 85$), R-E6c-seed1 did not (74.9, not < 40) and the control hit (97.8); R-E7-fixed-breaks did not fire (fixed readout 0.897 against 0.917), R-E7-info-kept, R-E7-subspace, R-E7-tracks and R-E7-lang-kept fired, R-E7-translation and R-E7-ownhead did not, and seed 1 was 'partial' by the letter (relearned top-8 0.777 against the threshold 0.778); CS: C2 partial (1.793), E1 R-CS-fails (1.102), gate V-CS-lora passed (4.388); CT: R-CT-texture fired for C2 (0/226 and 223/226) and for E1 (0/226 and 226/226), gate V-CT-lora passed (4.399; 226/226). Intervals are registered 80% intervals, not confidence intervals; a hit is not evidence of mechanism.

(E) Corrections to earlier versions

Each item gives the condition under which the earlier number or reading was obtained, the later finding, the corrected condition, and what is replaced.

(1) LIBERO evaluation prompt. *Original condition:* every 2,000-episode LIBERO evaluation in the paper, including those of the fresh-head (G) and H analyses, formatted the instruction as 'Predict the next 10 robot actions', whereas every matched cell was trained with 'next 8' (Appendix K). *New finding:* E6 (feature), E6b (fresh heads) and E6c (own heads) show that the two fp32 $3\mathrm{e}{-5}$ checkpoints are sensitive to this one token while the working cells tested show no comparable change. *Corrected condition:* the $3\mathrm{e}{-5}$ checkpoints re-evaluated with the training prompt on the same initial states. *Replaced:* 46.9 / 15.0 (own heads) and 47.1 / 15.85 (fresh heads) are 'next 10' values; under 'next 8' they are 94.8 / 74.9 and 95.3 / 86.2 (seed 1 on the goal suite only); the working cells tested show no comparable change (the $1\mathrm{e}{-5}$ own head 97.8 vs. 97.6 on the goal suite, 489 vs. 488 of 500 paired; its fresh head 97.8; LoRA and bf16 cells 88.5–98.5% on the 200-episode subset under 'next 8' against 86–98% for the same cells over 2,000 episodes under 'next 10'). The earlier statement that the $3\mathrm{e}{-5}$ cells fail in closed loop is replaced by: they fail under the mismatched prompt and work under the training prompt, with the seed-1 long-10 residual (33.8%) unexplained.

(2) H-D and the T4 nearest-neighbour baseline. *Original condition:* the T4 held-out baseline (0.0013 for G1, 0.0022 for G4) was computed with 2,000 query frames drawn from the reference set, and the tenfold gap of the first rollout observations was read as a recorded-versus-live input difference (hypothesis H-D). *New finding:* with a leave-episode-out reference the gap is $\approx 29\times$ (cosine) / $\approx 5\times$ (standardised Euclidean) for G1 and $\approx 1\times$ for G4; the image-side candidates were tested (E1-D layouts, E1-J compression) and the gap closes under the training prompt (E6). *Corrected condition:* leave-episode-out reference, both prompts. *Replaced:* the 0.0013 / 0.0022 baseline and the "ten times farther" sentence are withdrawn; H-D is withdrawn; the first-observation gap is attributed to the prompt.

(3) The T1–T4 tests. *Original condition:* reported as four candidate explanations of a closed-loop failure, each excluded by a registered test. *New finding:* the failure they probed is specific to the mismatched prompt. *Corrected condition:* T1 and T1b are demonstration-frame properties under the

training prompt; T2, T3 and T4 are behaviour under the 'next 10' prompt. *Replaced:* the "excluded" reading; the numbers stand as measured and the rules as registered.

(4) CALVIN skills. *Original condition:* `move_slider` (0/151) was described as a capability the frozen backbone has (143/151) that finetuning had "removed", and `turn_lightbulb` (0/75) as one that full finetuning had "never acquired", both from scene-D logs only. *New finding:* on training scene A with the same chains both skills are 151/151 and 75/75 for C2 and for E1 (CS); a single-factor texture swap switches them off (0/226) and on (223/226; E1 0/226 and 226/226) with geometry, blocks and initial states fixed (CT); block manipulation on the table is poor in both scenes (lift from table 3/73 on A). *Corrected condition:* scene A, scene D and the two hybrid tables, same 1,000 chains. *Replaced:* the "removed", "never acquired" and "fails skill by skill" wording and the inference from the frozen-base 143/151 (the base number itself stands); the scene-D values remain the benchmark result.

(5) F2. *Original condition:* rule R-F2b fired ($1.003 \leq 1.3$) and was read as "readable but not usable, with a fresh head as with its own". *New finding:* CS and CT locate the failure in a scene component (slider door and lightbulb switch) that the table texture alone switches off and on, and a block-manipulation component present in the training scene. *Corrected condition:* unchanged for F2 itself. *Replaced:* the outcome is kept as "a fresh head does not rescue C2 (1.003 vs. 0.950)"; the interpretation becomes that demonstration-frame readability (top-8 0.46) records neither the texture sensitivity nor the control deficit.

G ADAPTER SUMMATION, TRANSFER AND JOINT TRAINING

The experiments in this appendix ask what an adapter finetuned for one benchmark is worth on another. They are thinner than the rest of the paper: single training seeds, 800-episode and 200-chain subsamples, one benchmark pair on one arm (Table 25); their negative results and controls travel with them. The main text makes no contribution from them.

Table 25: Reuse experiments (single training seed). Summation and fresh-head transfer: Emb-8B, 800 LIBERO episodes / 200 CALVIN chains. Joint adapter: 600 episodes on the trained suites, 200 chains, 200 on held-out LIBERO-10. Intervals fixed in advance and the Open-X initialization: Appendix J.

Experiment	LIBERO (%)	CALVIN (avg. length)
Summed LIBERO and CALVIN adapters		
own adapter, own head	98.4	4.70
summed adapters, own heads	**2.0**	**0.09**
summed adapters, fresh heads	82.3 (89 / 93 / 88 / 59)	4.0
Fresh head on a frozen backbone, other benchmark's data		
CALVIN-finetuned / frozen base / LIBERO-finetuned backbone	**62.4** / 22.1 / 97.8	–
LIBERO-finetuned / frozen base / CALVIN-finetuned backbone	–	1.18 / 0.90 / 4.58
Joint adapter on CALVIN and 30 LIBERO tasks; LIBERO-10 held out		
trained suites (single-benchmark ref.)	98.8 (99.3)	4.36 (4.58)
fresh head on held-out LIBERO-10: joint adapter	**76.0**	–
CALVIN-only adapter / frozen base / adapter trained on LIBERO-10	24.0 / 6.5 / 95.0	–
control: 30 LIBERO tasks alone, same budget share	**79.0**	–

Two adapters do not compose by weight summation. Summing the Emb-8B LIBERO and CALVIN adapters, each read by its own head, we expected mild interference: the deltas' write subspaces are near-orthogonal (energy-weighted alignment 0.05 over 341 modules). Instead both collapsed: 2.0% and 0.09, the level of controls carrying only the *other* task's delta (0.0 / 0.00). Weight-space orthogonality is a poor proxy for composability, in line with the interference documented for naive merging (Yadav et al., 2023). Fresh heads retrained on the summed backbone's features recover 82% on LIBERO (89 / 93 / 88 / 59 per suite) and 4.0 on CALVIN (registered 70

[40, 90] and 3.0 [1.5, 4.3], both hits). This shows that the summed features still support a policy through a new readout; it does not show that the content of either adapter was preserved unchanged, and the reading that the sum "scrambled the coordinates" was formed after the fact. The costs are lopsided: the two adapters took about 40 GPU-hours to train (16.3 h and 22.6 h), the addition that destroys them is free, and caching the summed backbone's features and refitting the two heads costs 3.0 GPU-hours.

A single-benchmark adapter transfers partly and asymmetrically through a fresh head. Head-only training on cached frozen features recovers the end-to-end policies on the matching backbone (97.8 and 4.58 against 98.4 and 4.70), so the test is sensitive. Read on LIBERO, the CALVIN-finetuned backbone reaches 62.4% against 22.1% for the frozen base, a transfer ratio of 0.53 graded from 0.81 on the spatial suite to 0.23 on the long-horizon suite; read on CALVIN, the LIBERO-finetuned backbone reaches 1.18 against 0.90 (ratio 0.075; at 2B on 1,000 chains, 0.92 against 1.07, ratio -0.05, so the LIBERO-to-CALVIN direction is zero to slightly negative at 2B and near zero at 8B). Three readings follow. The summation collapse was head co-adaptation: the same CALVIN-finetuned backbone scores 0% under the LIBERO-trained head and 62% under a fresh one. What transfers is a short-horizon reach–grasp–place primitive, not chains. And the sharing is asymmetric, plausibly a matter of skill coverage. Both benchmarks drive a Franka arm in different simulators, so this is reuse across visual domain and task vocabulary within one embodiment. Open-loop error is not a proxy for closed-loop competence here: the two fresh heads on LIBERO had validation MSE 0.053 (CALVIN-finetuned backbone) and 0.050 (frozen base) and succeeded at 62 vs. 22%.

One adapter hosts both benchmarks, and a fresh head reads related held-out tasks from it. One LoRA finetuned jointly on CALVIN and the 30 LIBERO tasks of the spatial, object and goal suites carries both at a modest cost (98.8 vs. 99.3; 4.36 vs. 4.58 on 200 chains, at half the per-task data), so the summation collapse was a failure of composition, not of capacity. On the ten held-out LIBERO-10 tasks the joint head scores 0% zero-shot; a fresh head trained on their demonstrations with no backbone update reads 76.0% from the frozen joint backbone, above the interval we had registered (40 [15, 70]; a reading formed after the fact), against 24.0% from the CALVIN-only adapter and 6.5% from the base. The control fixed in advance sets the reading: an adapter trained on the same 30 LIBERO tasks *alone* gives 79.0%, so the 76% is exposure to thirty related tasks in one simulator, not cross-benchmark diversity (CALVIN co-training: -3 points, inside the ± 10-point margin fixed in advance), and it sits 19 points below an adapter that trained on them.

Stacking details. Both adapters were trained with seed 0, so their LoRA-A initializations coincide and the full-rank read-space overlap (0.80) is an initialization artifact; the energy-weighted alignments (write 0.049, read 0.077, i.e. $2.8\times$ and $4.6\times$ the random-subspace floor; top-8 singular directions hold 78–84% of each delta's energy) are the meaningful values. The summed adapter is a rank-128 concatenation with per-half scale preserved; delta-equivalence to sequential application was verified to 1.1×10^{-8}. Per suite the summed LIBERO policy scores 0 / 0.5 / 7.5 / 0 (spatial / object / goal / long). Budget: the LIBERO adapter trained for 16.3 h and the CALVIN adapter for 22.6 h on one H100 (200k and 228k steps at their logged rates); the merge is a CPU state-dict operation; the fresh-head recovery cost 1.93 h and 1.06 h of feature caching plus a fit of seconds on the cached vectors.

Fresh-head protocol. Backbone frozen (base alone, base+Δ_L, base+Δ_C, or base+Δ_L+Δ_C); pooled z cached for all frames of the target data (LIBERO stride 1, CALVIN stride 4); head $\mathrm{Linear}(D, 2048)$–SiLU–$\mathrm{Linear}(2048, \mathrm{chunk} \times 7)$ with D = 4096 (8B) or 2048 (2B) (plus a fresh state projector for CALVIN) trained 30 epochs, batch 512, AdamW $3\mathrm{e}{-4}$ cosine; positive controls must land within $[\mathrm{ref} - 6, \mathrm{ref} + 2]$ of the own-head references or the transfer ratio is not interpreted. CALVIN references in Table 25 are 200-chain subsamples (own head 4.695, fresh head 4.575), not the 1,000-chain headline. The same protocol produced the CALVIN 2B controls F0–F2 and the LIBERO fresh heads of experiment G (Appendix F).

Joint adapter: geometry and per-task pattern. Inside the joint model the action subspaces (top-k_{90} principal directions, k_{90} = directions needed for 90% of the full ridge R^2) have $k_{90} = 12$ for LIBERO and 16 for CALVIN (single-benchmark references 4 and 24; the base never reaches 90% within 256 directions); no principal angle between the two is below 30° (smallest 34.8°), yet cross-decodability is high (ratios 0.94 / 0.90 against random-k baselines of about 0.65–0.73). Held-out LIBERO-10 successes per task (of 20) for the two-object basket tasks: H_J 19 / 20 / 16, H_C 0 / 4

/ 0, H_0 0 / 4 / 0, H_L 19 / 20 / 19; for the articulated tasks (drawer, microwave, stove, caddy) the CALVIN-only adapter alone reaches 10 / 8 / 6 / 13.

Open-X initialization: result. An adapter pretrained on Open-X data was a liability early and no help later. We trained one adapter on 14 Open-X datasets (up to 300 trajectories each, all if fewer; 8 embodiments; sampler balanced over embodiments) and a control on Bridge alone, then used each to initialize the 2B matched-protocol recipe with nothing else changed. Against a registered head start of $+12$ points, at 10k steps the OXE-initialized runs reach less than half the success of scratch (17.5 / 15.8 vs. 40.0) despite a *lower* training loss (0.0576 / 0.0578 vs. 0.0672); at 20k and 30k the differences sit inside a non-monotonic band that all conditions traverse with flat loss (scratch 63.0 / 56.2; uniform 65.7 / 48.7; Bridge 68.0 / 59.7; Table 27). Eight-embodiment uniformity bought nothing over Bridge alone, and the pretrained adapter's weight delta is no more aligned with the LIBERO delta (0.036 / 0.041 at 8B / 2B) than two mutually colliding task deltas are with each other (0.049 / 0.044). The sign agrees with StarVLA-α's observation that out-of-domain robot pretraining can hurt (single-view data, 3.5k trajectories, one seed); we claim no mechanism, since the geometric one we registered was refuted.

Open-X initialization: design and caveats. 14 Open-X datasets $\times$ up to 300 trajectories (all if fewer) $=$ 3,510 trajectories, 710k frames, 8 embodiments, converted to a common format; since episode lengths differ $\sim$20$\times$ across datasets the natural mix is 78% Franka by frames, so the training sampler balances the eight embodiment groups at 12.5% of gradient samples each. Pretrained adapters: contrastive 2B LoRA r=64 including the vision tower, 0.8M samples, single view, no state, chunk 8; the Bridge-only control has the same trajectory count and budget but $6\times$ fewer unique frames. Transfer: the 2B matched-protocol recipe with the pretrained adapter as initialization of the LoRA and a fresh head; 4 suites $\times$ 15 trials $=$ 600 episodes per point. Caveats recorded at registration: single view vs. LIBERO's two views (a live alternative explanation for the magnitude of the early deficit); control frequencies of 2–30 Hz so a chunk of 8 spans different durations; one dataset's raw actions in non-metric units; constant action dimensions in five datasets. A rerun of the scratch 30k point reproduced 56.2 vs. 56.3, which bounds pipeline reproducibility, not seed variance; all conditions show a non-monotonic reshuffle between 20k and 30k with flat loss.

H Robustness under perturbation

Table 26: LIBERO-Plus zero-shot success (%) per perturbation axis. Mean$\pm$SD over three subsamples (100 variants per axis), one model per cell (variant sampling, not seeds). Emb-2B and the fp32 full-FT cell: text below.

Axis	Emb-8B	Instr-8B (mean)	Instr-8B (last)
Lighting	92.7$\pm$1.7	89.7$\pm$1.7	91.3$\pm$0.9
Background	89.0$\pm$3.6	90.3$\pm$1.2	92.0$\pm$1.4
Object layout	78.3$\pm$5.0	71.3$\pm$2.9	76.0$\pm$2.9
Language	**82.0$\pm$2.4**	58.0$\pm$1.4	80.3$\pm$3.1
Initial states	**71.7$\pm$3.3**	45.0$\pm$2.2	64.3$\pm$3.8
Camera	**59.0$\pm$1.6**	35.0$\pm$5.7	45.0$\pm$4.5
Sensor noise	**68.0$\pm$2.9**	40.7$\pm$2.1	60.0$\pm$4.1
Macro	**77.2$\pm$0.3**	61.4$\pm$0.9	72.7$\pm$2.0
Clean	97.7	97.4	98.2

Clean success does not separate the backbones; perturbation does, and only in part. On LIBERO-Plus (Fei et al., 2025) (Table 26; matched-protocol generalists, zero-shot), the largest apparent gap, language ($+24$: 82.0 vs. 58.0), is a pooling artifact: switching the instruct model to last-token pooling recovers $+22$ (80.3), leaving $+1.7$, inside noise. With pooling equalized the residual is modest and consistent in sign on the hardest visual axes: camera $+14.0$, sensor noise $+8.0$, initial states $+7.4$, macro 4.5 ± 2.0. The two backbones differ in the contrastive post-training stage; the edge is attributable to it, on one pair. The fp32 full-finetuning cell of Table 8, under the same protocol, lands in the LoRA-2B tier (macro 68.6 vs. 69.9; language 68 vs. 66): the regime does not separate robustness either, on one pair and one subsample seed.

Feature drift under perturbation. For 200 frames under proxy versions of four axes (camera shift, sensor noise, lighting, language paraphrase), drift relative to the clean set's spread is smaller

for the contrastive policy on every axis (visual axes 1.09–$1.15\times$, language $3.35\times$ for the mean-text cell, whose pooling averages instruction tokens into z); per-axis drift predicts the per-axis behavioral drop (Spearman $\rho = 0.86$, $p = 0.007$; de-meaned cosine version $\rho = 0.88$), and within each model the visual-axis ordering of drift matches the ordering of behavioral damage. Drift is a directional predictor, not a proportional one (10–16% extra drift maps to $\sim$30-point behavioral gaps, since closed-loop rollouts compound errors); linear-probe retention under perturbation stays at 98–102% for both models, so the discriminative signal is drift magnitude, not decodability loss; and the contrastive policy's residual language failures (-19.5 points) occur with near-zero drift. The single-subsample Emb-2B column of Table 26 is: light 95, background 96, layout 72, language 66, robot init 63, camera 40, sensor noise 57; macro 69.9 (clean 97.5). The fp32 full-finetuning cell (97.90 clean), same protocol and subsample seed: light 88, background 88, layout 72, language 68, robot init 65, camera 36, sensor noise 63; macro 68.6 (registered 64 [52, 74]; all eight axis intervals hit). The 8B instruct model at its best pooling (72.7) lands in the same tier as the 2B contrastive model.

I EFFICIENCY

Inference cost is not part of the paper's argument; it is reported here for completeness. Three latency regimes are never mixed. **Regime A, pure policy forward, single sample (RTX 5090; H100 in parentheses):** 2B bf16 2.48 (2.51) ms, 8B bf16 3.01 (3.05) ms; 8B 8-bit 16 ms, 8B 4-bit 4.5 ms; full-model VRAM 4.3 / 18.4 / 10.9 / 7.2 GB (2B bf16 / 8B bf16 / 8-bit / 4-bit). A consumer flagship matches the data-centre card because the single-vector, small-batch workload is bandwidth-bound, and quantization does not speed this workload up (de-quantization overhead); its only gain is memory. **Regime B, end-to-end per action including image preprocessing (RTX 5060 Ti 16 GB, chunk-amortized):** Ours-8B 8-bit 49.1 ms ($\sim$20 Hz, 9.0 GB); OpenVLA-OFT-7B 8-bit 92.5 ms ($\sim$11 Hz, 8.9 GB), bf16 1768 ms (VRAM thrash, 16.1 GB), 4-bit 54.6 ms ($\sim$14 Hz, 5.3 GB); vanilla OpenVLA-7B bf16 / 8-bit / 4-bit 750 / 1646 / 1131 ms; Ours-2B bf16 8.5 ms single view (8.9 ms with the wrist camera), 4.3 GB. The wrist camera costs $+4$ ms per action on the 8B model. **Regime C, quantization vs. success (H100 evaluation):** relative to bf16, 8-bit / 4-bit change LIBERO-spatial by -0.8 / -0.2, LIBERO-10 by -0.6 / -4.0, and CALVIN (500-chain subsample of one checkpoint) by -0.09 / -0.13: quantization costs appear only on long horizons.

J PRE-REGISTRATION SCORECARD

Table 27: Every pre-registered quantity in this paper, with its registered point estimate and interval and the observed value. Registered, all before the corresponding evaluations: stacking and the Open-X initialization 2026-08-31; fresh-head transfer 09-02; joint adapter 09-03; interventions, sweep, summed-backbone heads and the LIBERO-30 control 09-09; task-centred sweep 09-12; training-level registrations in Table 28; the G and H registrations of 09-18 in Table 30; the registrations of 09-25 (E6, E6b, E6c, E7, CS, CT) in Tables 31 and 32.

Experiment	Quantity	Registered	Observed
Stacking	LIBERO summed	88 [70, 96]	2.0 ×
	CALVIN summed	3.8 [2.7, 4.5]	0.09 ×
	own references	96.5–98.5 / 4.4–4.85	98.38 / 4.695 ✓
	wrong-adapter controls	$\leq$10 / $\leq$0.5	0.0 / 0.00 ✓
	both sides drop beyond noise	yes	yes ✓
Open-X init.	P1 transfer, smallest budget	$+12$ [$+3$, $+25$]	-22.5 / -24.2 ×
	P2 uniform − Bridge	$+4$ [-2, $+10$]	$+1.7$ / -2.3 / -11.0 ×
	P4 weight alignment	> 0.15	0.036 / 0.041 ×
	P3 capability retention, P5 composability	–	not run
Fresh-head transfer	L_L / L_C / L_0	97 [93, 99] / 50 [20, 85] / 20 [3, 50]	97.75 / 62.38 / 22.12 ✓ / ✓ / ✓
	C_C / C_L / C_0	4.6 [4.3, 4.85] / 1.8 [0.5, 3.5] / 0.4 [0.05, 1.2]	4.575 / 1.175 / 0.90 ✓ / ✓ / ✓
Joint adapter	trained LIBERO / CALVIN / zero-shot LIBERO-10	95 [88, 99] / 4.3 [3.6, 4.7] / 8 [0, 25]	98.83 / 4.36 / 0.0 ✓ / ✓ / ✓
	H_J / H_C / H_0 / H_L	40 [15, 70] / 20 [8, 35] / 3 [0, 10] / 85 [70, 95]	76.0 × (above) / 24.0 / 6.5 / 95.0 ✓ / ✓ / ✓
	shared subspace fraction; cross-decode ratio	0.6 [0.3, 0.9]; 0.7 [0.4, 0.95]	0.0 ×; 0.94 / 0.90 ✓
LIBERO-30 control	H_{L30}	60 [35, 85]	79.0 ✓
Interventions	delete top-1 <10 and random-1 $\geq$ base$-$10, per backbone	4 of 4	4 of 4 ✓
	delete PCs 2–17 $\geq$ base$-$20	4 of 4	0 of 4 ×
	delete top-4/16 $\leq$5; random-16 $\geq$ base$-$15	4 of 4	4 of 4 ✓
	keep top-16 $\geq$ 90% of base (knee $\leq$16)	4 of 4	0 of 4 (knee 32 / 64 / 64 / >64) ×
	keep random-16 <30%, random-64 <60% of base	4 of 4	4 of 4 ✓
	init-state blocks within $\pm$8 (base, keep-16)	yes	yes ✓
Ridge sweep, 3 new cells	base top-8 $\leq$0.15 at 40k; finetuned top-8 $\geq$0.80 at 400	3 of 3	3 of 3 ✓
	base full feature at 120k in [0.55, 0.75]	3 of 3	0.64 / 0.65 / 0.66 ✓
Summed backbone + fresh heads	LIBERO / CALVIN	70 [40, 90] / 3.0 [1.5, 4.3]	82.25 / 4.0 ✓ / ✓
Task-centred sweep	finetuned within-task R^2 at 400 $\geq$0.60, top-8 $\geq$0.45	4 of 4	4 of 4 (0.93–0.95; 0.88–0.93) ✓
	base within-task at 400 $\leq$0.20; at 120k $\leq$0.50	4 of 4	3 of 4; 0 of 4 (0.61–0.66) ×
	finetuned − grounding oracle (within, 400) $\geq$0.20	$\geq$3 of 4	4 of 4 (margin $\approx$0.66) ✓
	between-task share (per-task mean)	[0.25, 0.55]	0.07 × (smaller)
	unseen LIBERO-10: joint / LIBERO-30 adapter within at 400 $\geq$0.45 and $\geq$ base $+0.25$	yes	0.67 / 0.65 vs. base 0.31 ✓

Table 28: Scorecard, continued: training-level registrations (seeds, learning rate and bottleneck heads registered 2026-09-13; full-FT probe and fp32-master control 09-15; fp32-checkpoint probe, LIBERO-Plus and CALVIN interventions 09-16; CALVIN full finetuning and LoRA seeds (C2, C3) 09-16; C2 feature probe, fresh-head controls and Instruct-2B full FT (D1, F0, F1, E1) 09-17; test-scene / unseen-task probes (DP) and the C2 fresh head (F2) 09-18. C2/C3 rows: 1,000 chains (C1: 200 chains), one seed and one learning rate per cell).

Experiment	Quantity	Registered	Observed
Seeds and LR	LoRA-2B, seeds 1 and 2	97.5 [96.3, 98.5] each	98.20 / 97.25 ✓
	LoRA-2B, 3-seed SD	$\leq$0.6	0.49 ✓
	full-FT-2B at $1e{-5}$, seeds 1 and 2	84 [77, 90] each	86.30 ✓; 76.60 × (0.4 below)
	full-FT-2B, 3-seed SD	$\leq$4	5.07 × (less stable)
	full-FT-2B at $3e{-5}$	88 [60, 96]	**96.55** (1 seed) × (above; deficit $15.4 \rightarrow 1.1$)
Bottleneck head	rank-64 head trained end to end	97 [93, 99]	98.40 ✓
	rank-16 head trained end to end	90 [70, 97]	**97.70** × (above)
Full-FT probe (E3)	LR $3e{-5}$: top-8 R^2 at 400 / keep-64 / delete top-1	0.70 [0.30, 0.93] / 85 [60, 97] / 15 [0, 60]	0.89 / 91.5 / 1.0 ✓ / ✓ / ✓
	LR $3e{-5}$: full R^2 at 400 / keep-16 / keep-32 / random-64 / PCs 2–17	0.90 [0.75, 0.96] / 55 [20, 85] / 70 [35, 92] / 0 [0, 15] / 5 [0, 40]	0.91 / 74.0 / 88.5 / 0.0 / 0.0 ✓ (all)
	LR $1e{-5}$, seed 1: top-8 R^2 at 400 / keep-64	0.45 [0.10, 0.85] / 70 [35, 92]	0.73 / 60.0 ✓ / ✓
fp32-checkpoint probe (E3b)	LR $1e{-5}$, fp32: top-8 R^2 at 400 / keep-64 / delete top-1	0.90 [0.80, 0.94] / 94 [85, 98] / 3 [0, 25]	0.93 / 96.5 / 12.0 ✓ / ✓ / ✓
	full R^2 at 400 / keep-16 / keep-32 / random-64 / PCs 2–17	0.94 [0.88, 0.96] / 78 [55, 90] / 88 [70, 95] / 0 [0, 15] / 2 [0, 40]	0.95 / 79.5 / 86.5 / 0.0 / 0.0 ✓ (all)
LIBERO-Plus, fp32 cell (E6-Plus)	macro / language / camera / noise / init / light / background / layout	64 [52, 74] / 56 [35, 76] / 36 [20, 55] / 50 [30, 65] / 58 [40, 72] / 93 [85, 98] / 94 [85, 98] / 68 [50, 80]	68.6 / 68 / 36 / 63 / 65 / 88 / 88 / 72 ✓ (all)
fp32-master full FT (E4)	LR $1e{-5}$ / $3e{-5}$, seed 0	95 [88, 98] / 96.5 [92, 98.5]	**97.90** ✓ / **46.90** × (below; 'next 10' prompt; 94.80 under the training prompt, E6c)
	fp32 $1e{-5}$ training loss	0.010 [0.007, 0.016]	0.0066 × (below)
	E4b: $3e{-5}$, seed 1	55 [20, 97]	**15.00** × (below; 'next 10'; 74.9 under the training prompt, E6c)
Batch-128 control (E5)	fp32, LR $1e{-5}$, batch 128, 6,250 updates, 0.8M samples	90 [75, 96.5]	**97.50** × (above)
	final training loss	0.012 [0.008, 0.020]	0.0091 ✓
CALVIN interventions (C1)	Emb-8B CALVIN adapter, 200 chains, fraction of the un-ablated 4.695 (registered 4.65 [4.40, 4.85] ✓): keep-64 / delete top-1 / random-64	0.85 [0.70, 0.95] / 0.20 [0.05, 0.50] / 0.05 [0, 0.25]	0.965 × (above) / 0.480 ✓ / 0.000 ✓
	keep-16 / keep-32 / delete PCs 2–17	0.55 [0.35, 0.80] / 0.72 [0.50, 0.88] / 0.10 [0, 0.40]	0.498 ✓ / 0.905 × (above) / 0.001 ✓
CALVIN full FT (C2)	Emb-2B, fp32 master, LR $1e{-5}$: avg. length / training loss	4.40 [4.20, 4.55] / 0.055 [0.045, 0.070]	**0.950** × (below) / 0.0612 ✓
CALVIN LoRA seeds (C3)	Emb-2B LoRA, seeds 1 and 2 / 3-seed SD	4.40 [4.25, 4.55] each / 0.06 [0.02, 0.15]	4.264 / 4.371 ✓ / ✓; 0.072 ✓

Table 29: Scorecard, continued: C2 feature probe, fresh-head controls, Instruct-2B full FT and the test-scene / unseen-task probes (D1, F0, F1, E1 registered 2026-09-17; F2 and DP 09-18; CALVIN rollouts 1,000 chains, one seed and one learning rate per cell. DP: ridge R^2 at N=120k reader frames, full feature / top-8; values in Table 14. Scene-D probes: reader fit on scenes A–C, tested on 822 scene-D samples from 17 windows / 6 task names; unseen-task probes: reader fit on LIBERO tasks 10–39, tested on all LIBERO-10 frames.)

Experiment	Quantity	Registered	Observed
C2 feature probe (D1)	C2 backbone on CALVIN frames: full / top-8 R^2 at 400; at 120k; PC1 share; participation ratio	0.25 [0.10, 0.60] / 0.10 [0.00, 0.45]; 0.55 [0.35, 0.80] / 0.20 [0.02, 0.55]; 0.20 [0.05, 0.60]; 12 [3, 40]	0.558 / 0.441; 0.662 / 0.456; 0.199; 14.34 ✓ (all six)
	LoRA-2B top-8 / full at 400; frozen base full at 400 / top-8 at 120k	0.55 [0.35, 0.75] / 0.70 [0.50, 0.85]; 0.15 [0.05, 0.35] / 0.08 [0.00, 0.25]	0.630 / 0.728; 0.127 / 0.054 ✓ (all)
Fresh-head controls (F0, F1)	frozen Emb-2B base + fresh head: avg. length / subtask 1	0.80 [0.40, 1.30] / 0.48 [0.30, 0.65]	1.068 / 0.599 ✓ / ✓
	Emb-2B LIBERO adapter + fresh head: avg. length / 2B transfer ratio	1.00 [0.50, 1.80] / 0.06 [0.00, 0.20]	0.917 ✓ / −0.046 × (below)
Fresh head on C2 (F2)	avg. length / subtask 1	2.0 [0.8, 4.0] / 0.80 [0.50, 0.97]	1.003 ✓ (low end) / 0.546 ✓
Instruct-2B full FT (E1)	same recipe as C2: avg. length / training loss	1.2 [0.5, 4.3] / 0.07 [0.05, 0.10]	0.711 ✓ (low end) / 0.0934 ✓ (upper end)
Scene-D probes (DP)	Emb-2B LoRA CALVIN adapter: full / top-8 (retention)	0.50 [0.25, 0.75] / 0.40 [0.15, 0.65] (0.65 [0.35, 0.95])	0.69 / 0.57 ✓ / ✓ (0.87 ✓)
	Emb-2B full FT, C2 (retention)	0.15 [−0.10, 0.45] / 0.08 [−0.10, 0.35] (0.25 [−0.15, 0.70])	0.47 × (above) / 0.35 ✓ (at bound) (0.71 × above)
	Emb-2B frozen base	0.05 [−0.15, 0.30] / 0.02 [−0.10, 0.15]	0.35 × (above) / 0.01 ✓
	Emb-2B LIBERO adapter	0.00 [−0.20, 0.20] / 0.02 [−0.10, 0.15]	0.23 × (above) / 0.02 ✓
	Emb-8B CALVIN adapter (retention)	0.55 [0.30, 0.78] / 0.45 [0.20, 0.68] (0.70 [0.40, 0.95])	0.68 / 0.54 ✓ / ✓ (0.83 ✓)
	Emb-8B frozen base	0.05 [−0.15, 0.30] / 0.02 [−0.10, 0.15]	0.37 × (above) / 0.03 ✓
	Emb-8B LIBERO adapter	0.00 [−0.20, 0.20] / 0.02 [−0.10, 0.15]	0.34 / 0.19 × / × (above)
Unseen-task probes (DP)	reader fit on LIBERO tasks 10–39, tested on LIBERO-10, full / top-8: L_L (upper reference)	0.70 [0.45, 0.90] / 0.60 [0.35, 0.85]	0.96 / 0.94 × / × (above)
	J_L (joint adapter, never trained on LIBERO-10)	0.45 [0.20, 0.70] / 0.35 [0.10, 0.60]	0.49 / 0.48 ✓ / ✓
	L_C (CALVIN adapter)	0.15 [0.00, 0.40] / 0.08 [−0.05, 0.30]	−0.02 × (below) / 0.02 ✓
	L_0 (frozen base)	0.10 [−0.10, 0.35] / 0.03 [−0.10, 0.15]	−1.88 × (below) / −0.03 ✓

Table 30: Scorecard, continued: fresh heads on the LIBERO backbones (G) and the four tests on the LIBERO $3e{-}5$ backbone (H). G registered 2026-09-18 16:48 local time (v1, file time) and amended before any G job ran (v2, 16:56 local time; first G job 16:57); H registered 19:21 local time, after the G headline and before any H test ran. All G and H rollouts: 2,000 episodes (G) or 500 episodes per suite and condition (T2, T3, T4), all under the 'next 10' prompt; the 2026-09-25 registrations that re-evaluated them under the training prompt are in Tables 31 and 32. Rules and outcomes: Appendix F(d).

Experiment	Quantity	Registered	Observed
Fresh heads (G, v2)	G1 / G2 / G0 / G3 / G4 success	50 [35, 68] / 25 [10, 48] / 18 [8, 32] / 95 [85, 98] / 95 [80, 98]	47.10 / 15.85 / 25.95 / 97.30 / 97.80 ✓ (all five; 'next 10' prompt; under 'next 8': 95.3 / 86.2 goal / not run / not run / 97.8 goal)
	G1 probe top-8 / full R^2; G0 top-8	0.85 [0.55, 0.95] / 0.95 [0.85, 0.97]; 0.05 [0.01, 0.15]	0.914 / 0.945; 0.033 ✓ (all)
	fresh / original head training MSE; G1 goal suite; G1 mean feature norm	0.90 [0.60, 1.05]; 25 [8, 60]; 45 [15, 120]	0.916; 14.2; 90.4 ✓ (all)
H, T1 (demonstration frames, 'next 8')	goal-suite instruction accuracy: G1 / G2 / G4 / LoRA-2B / base	0.55 [0.2, 0.95] / 0.45 [0.15, 0.9] / 0.97 [0.85, 1.0] / 0.97 [0.85, 1.0] / 0.9 [0.6, 1.0]	1.000 × (above) / 1.000 × (above) / 1.000 / 1.000 / 1.000 ✓ / ✓ / ✓
H, T1b (demonstration frames, 'next 8')	G1 within-task top-8 R^2 at 120k	0.85 [0.6, 0.95]	0.912 ✓
H, T2 ('next 10')	G1 zero-4: goal / spatial; random-4 goal	16 [8, 40] / 42 [25, 60]; 14 [8, 25]	21.6 / 35.4; 16.6 ✓ (all)
H, T3 ('next 10')	goal suite at horizon 1: G1 / G4	18 [8, 40] / 90 [70, 98]	13.8 / 89.2 ✓ / ✓
H, T4 ('next 10')	absmax ratio rollout / demo: G1 / G4; G1 instruction accuracy late − early	1.5 [1.0, 3.0] / 1.05 [0.9, 1.4]; -0.15 [-0.5, 0.1]	0.97 × (below) / 0.96 ✓; -0.60 × (below; confounded, see Appendix F; the early value 0.764 becomes 1.000 under 'next 8', E6)

Table 31: Index of the registrations of 2026-09-25 (`probe_results/H2_prereg_amend3.json` to `amend8.json`, mirrored in `experiments/h2/`; each written before the corresponding measurement, times are the `registered_at` stamps in the files, local time). Registered numbers and observed values: Table 32. Intervals are registered 80% intervals, not confidence intervals; a hit is not evidence of mechanism.

Registration	Hypothesis / manipulation	Registered rules	Outcome	Current interpretation
E6 (amend 3), 02:10:09	Is the $3e{-5}$ backbone's first-observation feature gap (T4, E1-R0) the 'next 10' vs. 'next 8' prompt? The evaluator's 500 goal-suite first observations encoded under both prompts, G1 and G4	R-E6-prompt (gap under 'next 8' ≤ 2); R-E6-excluded (≥ 10 cosine and ≥ 3 standardised); R-E6-partial (between)	R-E6-prompt fired (1.1); the registered consequence, a behavioural run under 'next 8', became mandatory	the gap was the prompt; every LIBERO number in the paper was measured under 'next 10'; the working backbone moves 0.26 LEO for the same change, the $3e{-5}$ backbone 5.27
E6b (amend 4), 02:14:17	Do the fresh-head checkpoints G1 / G2 recover under the training prompt? Same evaluator with `-prompt_h 8`, same initial states, paired McNemar; G4 as control	R-E6b-v (G4 goal ≥ 94, first); R-E6b-rescue (G1 goal ≥ 60 and ≥ 4 of 6 dead tasks $\geq 20/50$); R-E6b-partial (25–60); R-E6b-null (< 25)	gate passed (97.8); R-E6b-rescue fired (95.6; 6 of 6)	47.10 and 15.85 are values under the mismatched prompt; 95.3 (all suites) and 86.2 (goal) under 'next 8'; the fresh head inherits the sensitivity
E6c (amend 5), 02:44:49	Do the own-head headline numbers (46.90, 15.00) survive under the training prompt? 2,000 episodes per seed, paired; fp32 $1e{-5}$ goal suite as control	R-E6c-headline (seed 0 ≥ 85); R-E6c-residual (60–85); R-E6c-seed1 (seed 1 < 40)	R-E6c-headline fired (94.80); R-E6c-seed1 did not (74.9); control 97.8	46.90 / 15.00 are cited as 'next 10' values with 94.80 / 74.9 beside them; the seed-1 long-10 residual (33.8) is unexplained; not a property of full finetuning in general ($1e{-5}$ insensitive)
E7 (amend 6), 14:02:37	Is the expert action still linearly decodable from the same checkpoint's demonstration-frame features under 'next 10'? Readouts fit under 'next 8' applied fixed, recentred, subspace-refit and relearned; own and fresh heads; displacement; per-task link to closed-loop drop	R-E7-fixed-breaks; R-E7-info-kept / -lost; R-E7-translation; R-E7-subspace; R-E7-ownhead; R-E7-tracks; R-E7-lang	fixed-breaks did not fire (0.897 vs. 0.917); info-kept, subspace, tracks and lang-kept fired; translation and ownhead did not; seed 1 partial by 0.001	aggregate decodability moves by 0.02 while success falls from 94.80 to 46.90 (seed 0); the loss is concentrated in the tasks that die (ρ 0.76); correlational, on training frames; no generalisation to full finetuning
CS (amend 7), 18:10:03	Do the CALVIN full-FT policies (C2, E1) work in a training scene? Scene A with the same 1,000 chains, block slots permuted by name; LoRA control; gates V-CS-scene, V-CS-build, V-CS-lora	R-CS-works (avg ≥ 3.0 and success@1 ≥ 0.75); R-CS-fails (avg ≤ 1.5); partial (between); E1 as replication	gates passed (LoRA 4.388; build 5/5 on the first four chains); C2 partial (1.793); E1 R-CS-fails (1.102)	two-part failure: slider / switch 226/226 on A vs. 0/226 on D for both backbones, and block manipulation poor in both scenes (lift from table 3/73 on A); no single-mechanism wording
CT (amend 8), 23:51:06	Is the slider / switch component caused by the table texture alone? Single-factor swap of the two wood-texture images between tables A and D (geometry, blocks, initial states fixed); LoRA control; E1 as replication	R-CT-texture (group $\leq 20\%$ on A+texD and $\geq 80\%$ on D+texA); R-CT-not-texture (the converse); partial; gates V-CT-frames, V-CT-lora	R-CT-texture fired for C2 (0.0%, 98.7%) and for E1 (0.0%, 100%); V-CT-lora passed (4.399; 100%)	for these two checkpoints the texture image alone switches the two skills off and on; the LED button's success depends on the geometry $\times$ texture combination (0/110 on both hybrids for E1; 0/110 and 107/110 for C2); the block-manipulation deficit is not addressed by this test

Table 32: Scorecard, continued: the registrations of 2026-09-25 (E6, E6b, E6c, E7, CS, CT; rules and outcomes in Table 31 and Appendix F(d)). LEO = the cell's leave-episode-out nearest-neighbour distance between demonstration first frames; A1–A3 = the E7 readout arms (A1 fit and tested under 'next 8'; A2 the A1 readout applied unchanged to 'next 10' features; A2c recentred; A2s top-8 subspace kept, weights refit; A3 refit under 'next 10'), top-8 / full R^2 at 40k frames. Intervals are registered 80% intervals, not confidence intervals; a hit is not evidence of mechanism.

Experiment	Quantity	Registered	Observed
E6 (amend 3)	G1 gap under 'next 8': cosine / standardised Euclidean; G1 instruction accuracy; G4 cosine; prompt shift G1 / G4 (LEO)	12 [2, 28] / 3.0 [1.0, 5.0]; 0.88 [0.7, 1.0]; 1.0 [0.7, 1.5]; 1.5 [0.3, 5.0] / 0.4 [0.1, 1.5]	1.1 × (below) / 1.05 ✓ (at the bound); 1.000 ✓ (at the bound); 0.9 ✓; 5.27 × (above) / 0.26 ✓
E6b (amend 4)	goal suite under 'next 8': G1 / G4 / G2; G1 dead tasks revived (of 6); G1 all suites	60 [25, 95] / 98 [94, 100] / 40 [10, 85]; 4 [1, 6]; 80 [45, 97]	95.6 × (0.6 above) / 97.8 ✓ / 86.2 × (1.2 above); 6 ✓; 95.3 ✓
E6c (amend 5)	own heads under 'next 8', 2,000 episodes: seed 0 / seed 1; $1e{-5}$ control, goal suite	93 [78, 98] / 75 [35, 96]; 98 [94, 100]	94.80 / 74.9 ✓ / ✓; 97.8 ✓
E7 (amend 6), G1	A1 top-8 / full; A2 top-8 / full; A2c top-8; A2s top-8; A3 top-8 / full	0.91 [0.89, 0.93] / 0.94 [0.92, 0.955]; 0.0 [−10, 0.6] / 0.0 [−10, 0.7]; 0.5 [−1.0, 0.9]; 0.80 [0.40, 0.92]; 0.88 [0.75, 0.93] / 0.935 [0.88, 0.955]	0.917 / 0.949 ✓ / ✓; 0.897 / 0.925 × / × (above); 0.897 ✓; 0.898 ✓; 0.898 / 0.945 ✓ / ✓
	top-8 subspace overlap; displacement (LEO, median); translation share	0.6 [0.2, 0.95]; 5 [2, 12]; 0.5 [0.2, 0.9]	0.997 × (above); 1.47 × (below); 0.005 × (below)
	own head R^2 under 'next 8' / 'next 10'; fresh head under 'next 10'	0.95 [0.90, 0.98] / 0.2 [−3, 0.8]; 0.2 [−3, 0.8]	0.949 ✓ / 0.927 × (above); 0.928 × (above)
	goal-suite task identity: matched / fixed / relearned; Spearman ρ (per-task error increase vs. success drop)	0.98 [0.90, 1.00] / 0.6 [0.1, 1.0] / 0.95 [0.6, 1.0]; 0.4 [0.0, 0.7]	1.000 / 0.687 / 1.000 ✓ (all three); 0.76 × (above; rule R-E7-tracks fired)
E7, G2 / G4 / G3	G2 A1 / A2 / A3 top-8; G4 A1 / A2 / A3 top-8, own head 'next 8' / 'next 10', displacement; G3 A1 / A2 / A3 top-8	0.82 [0.79, 0.85] / −0.5 [−10, 0.6] / 0.78 [0.60, 0.85]; 0.925 [0.91, 0.94] / 0.91 [0.85, 0.935] / 0.925 [0.91, 0.94], 0.96 [0.92, 0.99] / 0.95 [0.88, 0.99], 0.3 [0.1, 1.0]; 0.93 [0.91, 0.94] / 0.91 [0.80, 0.935] / 0.93 [0.91, 0.94]	0.828 ✓ / 0.780 × (above) / 0.777 ✓; 0.928 / 0.928 / 0.928, 0.964 / 0.963, 0.18 ✓ (all); 0.931 / 0.931 / 0.931 ✓ (all)
CS (amend 7)	scene A, avg. length / success@1: LoRA (gate); C2; E1	4.5 [3.8, 4.9] / 0.97 [0.90, 1.00]; 2.0 [0.7, 4.0] / 0.65 [0.45, 0.92]; 1.5 [0.5, 3.5] / 0.55 [0.35, 0.90]	4.388 / 0.982 ✓ / ✓; 1.793 / 0.736 ✓ / ✓; 1.102 / 0.618 ✓ / ✓
CT (amend 8)	slider + switch first-subtask success (%) / avg. length: C2 on A+texD; C2 on D+texA; LoRA on A+texD (gate); E1 on A+texD; E1 on D+texA	15 [0, 60] / 1.0 [0.6, 1.9]; 80 [30, 100] / 1.6 [0.9, 2.5]; 100 [95, 100] / 4.4 [3.8, 4.9]; 15 [0, 60]; 75 [20, 100]	0.0 / 1.018 ✓ / ✓; 98.7 / 1.273 ✓ / ✓; 100.0 / 4.399 ✓ / ✓; 0.0 ✓; 100.0 ✓

Decision rules registered with the seeds-and-learning-rate block. The finetuning-regime reading of the matched comparison (Appendix B) survives if the LoRA mean exceeds the full-FT mean at $1e{-5}$ by at least 5 points, *and* full-FT at $3e{-5}$ lands at least 5 points below the LoRA mean. If full-FT at $3e{-5}$ comes within 2 points of the LoRA mean, the gap was a learning-rate artifact and the matched-comparison text and the abstract are rewritten accordingly. If either regime's seeds spread by more than 5 points, the comparison's sub-point differences are demoted to seed noise; they did (9.7 points), and they were. Both consequences fired: full-FT at $3e{-5}$ (pure bf16) reached 96.55, so the survival condition failed and the artifact condition held, and the matched-comparison text, the abstract and the introduction were rewritten as the rule prescribes.

Rules registered on 2026-09-15 (E3, E4). E3 reads the three bf16 full-FT checkpoints with the ridge sweep, the task-centred sweep and the eval-time interventions of Appendices D and E. Concentration is declared a property of finetuning rather than of the adapter if, on the $3\mathrm{e}{-5}$ checkpoint, top-8 R^2 at 400 frames is at least 0.80, keep-64 retains at least 90% of the un-ablated success and deleting the top direction leaves at most 25% of it; full finetuning is declared to spread the action information more widely if top-8 R^2 is at most 0.50 while the full feature reads at least 0.85, or keep-64 retains less than 75%; anything between is reported as partial concentration. The $1\mathrm{e}{-5}$ checkpoints carry the bf16 caveat and enter no rule. E4 trains the two full-FT rates with fp32 master weights, seed 0, all else identical. With $L = 97.65$ the LoRA mean: fp32 at $1\mathrm{e}{-5} \geq L - 3$ makes the bf16 deficit a precision artifact and the matched-comparison text is rewritten around the fp32 value; ≤ 88 keeps the learning-rate reading; between, both are reported and no single cause is claimed. For $3\mathrm{e}{-5}$, a value $\geq L - 0.5$ closes the residual 1.1 points; below 96.55 the residual is not attributable to precision. *Outcome of E3:* rule C fired on the $3\mathrm{e}{-5}$ checkpoint (top-8 $0.89 \geq 0.80$; keep-64 at 95% of the un-ablated 96.0; delete top-1 at 1.0%); every registered interval was hit, most near its upper end, so we under-predicted how much of the action subspace survives pure-bf16 full finetuning. *Outcome of E4:* fp32 at $1\mathrm{e}{-5}$ reached $97.90 \geq L - 3$, so rule (a) fired: the bf16 deficit was a precision artifact, and the matched-comparison text, Table 8, the abstract and the introduction were rewritten as the rule prescribes; the training loss came in below its interval, a miss in the favourable direction. At $3\mathrm{e}{-5}$ the fp32 rerun reached 46.90 under the 'next 10' evaluation prompt, below its interval and below the 60-point divergence threshold, and 94.80 under the training prompt (E6c, registered 2026-09-25: 93 [78, 98], hit); it is reported as prompt sensitivity at that rate, not as a precision effect. *E4b (registered 2026-09-16):* a second seed at $3\mathrm{e}{-5}$ reached 15.00 under 'next 10' (registered 55 [20, 97], a miss below) and 74.9 under the training prompt (E6c: 75 [35, 96], hit); by E4b's rule (below 70) the 'next 10' result reproduced, and the text now reads "prompt-sensitive at this rate, where LoRA at the same rate is not", with two seeds. *E3b and E6-Plus (registered 2026-09-16):* the same probe on the fp32 $1\mathrm{e}{-5}$ checkpoint hit all eleven intervals and fired rule C (top-8 0.93; keep-64 at 98% of the un-ablated 98.0; delete top-1 12%); LIBERO-Plus on that cell hit all eight axis intervals with macro 68.6 against the LoRA cell's 69.9, so by the registered rule full finetuning at StarVLA's rate keeps the inherited grounding as well as LoRA, on one pair and one subsample seed. *E5 (registered 2026-09-16):* full finetuning with fp32 master weights at batch 128 and 6,250 updates on the same 0.8M samples reached 97.50, above its interval, so the registered *budget* reading fired: 0.8M samples suffice regardless of update count, and our $4.8\times$ smaller budget than StarVLA's is not bought by small-batch updates. *C1 (registered 2026-09-16):* the interventions of Appendix E on the Emb-8B CALVIN adapter. Concentration was declared to replicate if keep-64 retained at least 70% of the un-ablated chain length, deleting the top direction left at most 50%, and 64 random directions sat at least 30 points below the top 64; all three held (96.5%, 48.0%, 0%), the second at its edge, so the text says the leading direction *halves* the CALVIN policy; two keep-only intervals were missed above (Table 17). *C2 and C3 (registered 2026-09-16):* full finetuning of Emb-2B with fp32 master weights on CALVIN reached 0.950 against 4.40 [4.20, 4.55], so rule R2 (< 3.80) fired: a failure at this rate and budget, one seed and one rate. Two further LoRA seeds landed at 4.264 and 4.371 (three-seed SD 0.072), and the 8B-vs-2B CALVIN gap of 0.27 exceeds twice the pooled SD (0.12), so rule R1 calls it real. *D1 (registered 2026-09-17):* the C2 backbone's CALVIN features hit all six intervals and rule R-D1c fired (partial concentration: top-8 R^2 0.44 at 400 frames, between the 0.20 and 0.50 thresholds). *DP (registered 2026-09-18):* rule R-DPb fired (C2's top-8 R^2 on scene D $0.35 \geq 0.30$: the full-FT feature's action information transfers to the evaluation scene) and rule R-DPf fired (the joint adapter's top-8 R^2 on never-trained LIBERO-10 $0.48 \geq 0.25$: its feature is readable on those tasks). *F0 (registered 2026-09-17):* $|\mathrm{C2} - \mathrm{F0}| = 0.12 \leq 0.30$, so rule R-Fa fired: at matched scale, full finetuning landed at the level of the frozen backbone read by a fresh head (one seed each). *F1 (registered 2026-09-17):* the 2B transfer ratio -0.046 missed below its interval, so rule R-Fd fired and the LIBERO-to-CALVIN transfer wording of Appendix G was re-examined (one seed each). *F2 (registered 2026-09-18):* $1.003 \leq 1.3$, so rule R-F2b fired: a fresh head does not rescue C2 (1.003 vs. 0.950; one seed each). CS and CT (registered 2026-09-25) then locate the failure: on the training scene C2 reaches 1.793 (partial by R-CS; E1 1.102, R-CS-fails), its slider and light-switch skills are 226/226 on scene A and 0/226 on scene D, switched off by scene D's wood texture alone (0/226) and on by scene A's (223/226; R-CT-texture fired, replicated on E1), and its block manipulation on the table fails under every scene and texture (lift from table 3–36 of 73; LoRA 73/73). *E1 (registered 2026-09-17):* Instruct-2B full finetuning reached $0.711 \leq 2.0$, so rule R-E1b fired: the failure reproduces across backbone objectives, one seed each.

Rules registered on 2026-09-18 (G, H). The G rules (v1 and the pre-launch amendment v2) and the H rules for T1, T1b, T2, T3 and T4 are quoted in full with their outcomes in Appendix F(d). In summary: the validity gate R-Gj passed (G4 97.80, G3 97.30); R-Gb and R-Ge fired on G1 and R-Gg's replication branch on G2, all scored on 'next 10' evaluations, and R-Ga, R-Gc, R-Gd, R-Gh, R-Gi and R-Gk did not fire; R-Gf fired on neither branch. Re-evaluated with the training prompt (E6b, registered 2026-09-25) G1 reaches 95.3 and G2 86.2 on the goal suite, so the G verdicts describe behaviour under the mismatched prompt. For H, R-H1b (instruction present), R-H1e (within-task mapping intact), R-H2b (outliers not causal) and R-H3b (not open-loop divergence) fired; R-H4a and R-H4b did not, so R-H4c applies; the gap it left open is closed by E6, the prompt (Tables 31 and 32). Two G-related registration notes: the reasoning field of the G1 prediction contains a false statement about training loss that no rule uses (Appendix F(d)), and the G3 reference value quoted at registration (96.0) is the 600-trial curve point of the LoRA-2B checkpoint, whereas the matched 2,000-episode own-head value is 97.50; the G3 interval was registered on success, not on the difference, so the hit is unaffected.

K REPRODUCIBILITY

Table 33: Training budgets. External rows are as reported by the respective papers.

Policy	LIBERO finetuning budget	Robot pretraining
Ours best-recipe generalist	75k steps $\times$ batch 4 $=$ 300k sample-gradients (1.16 ep); 5.7 h on one H100	none
Ours matched comparison (each cell)	200k steps $\times$ batch 4 $=$ 0.8M samples (3.1 ep)	none
Ours batch-128 control (Appendix C)	6,250 steps $\times$ batch 128 (4 $\times$ 32 accumulated) $=$ 0.8M samples (3.1 ep)	none
StarVLA-OFT (StarVLA Community, 2026)	30k steps $\times$ batch 128 $=$ 3.84M samples	none
StarVLA-α (Ye et al., 2026)	up to 100k steps, $\leq$12.8M samples	none
OpenVLA-OFT (Kim et al., 2025)	50–150k gradient steps at batch 64–128 on 8 GPUs (as reported)	970k Open-X trajectories

Replan cadence. Ours executes a chunk of 10 (best recipe) or 8 (matched comparison) actions open-loop per backbone forward; OpenVLA-OFT parallel-decodes a chunk of 8; vanilla OpenVLA re-plans every step. **Evaluation prompt.** The instruction prompt names the chunk length ('Task: {task}. Predict the next h robot actions as a 7-DoF chunk.'). `dryrun_scripts/eval_ours_libero.py`, written in May 2026 for chunk-10 models, formats the prompt with a literal 'next 10' (line 144 of the 2026-05-24 backup; line 386 today) although the model reads `chunk_size` from the checkpoint (line 154); since the 2026-08-21 switch to chunk 8, every LIBERO checkpoint of the matched comparison (LoRA, bf16, fp32) was trained with 'next 8' and evaluated with 'next 10' by this script and its copies. `eval_ours_libero_ablate2.py` (2026-09-10) and the CALVIN evaluator (`scripts/eval_ours_calvin.py`, line 108) use h = chunk size, so the interventions of Appendix E and all CALVIN numbers are matched; Table 10 and the specialist of Fig. 3 are chunk-10-era models evaluated with 10. The training-prompt re-evaluations (E6b, E6c) add a `-prompt_h` argument to the same evaluator and change nothing else. **CALVIN scene-A and texture-swap evaluations (CS, CT).** `calvin_table_A_texD` is a copy of `calvin_table_A` in which `textures/dark_wood.png` and `textures/dark_wood__gray_handle.png` are replaced by table D's images (same file names, different content; the `.mtl` files reference the same texture names in both tables); meshes, joint origins, URDF and block set are unchanged. `calvin_table_D_texA` is the converse. The scene files `calvin_scene_A_texD` and `calvin_scene_D_texA` differ from `calvin_scene_A` and `calvin_scene_D` only in the table file line. The evaluator (`lad/eval_ours_calvin_scene.py`, the standard evaluator plus a `-scene` argument) replaces the validation folder's scene section, which equals `calvin_scene_D.yaml` plus a name key, permutes the three block slots of the initial-condition `scene_obs` by object name because `PlayTableScene.reset` assigns them in configuration order (A lists pink, blue, red; D lists red, blue, pink), and keeps the same `get_sequences(1000)` chains, task oracle, annotations, episode length 360 and execute

horizon 10; the renderer stays on its default device, as in every CALVIN run on record, and a black-frame guard aborts if a camera returns an all-black frame. Scenes B and C were not used because their slider sits elsewhere, so the multistep initial block positions would be invalid there. Asset md5s (textures, meshes, URDFs, scene files) are recorded in `probe_results/H2_prereg_amend8.json` (mirrored in `experiments/h2/`). **Our-harness generalist baselines.** Trained and evaluated on our harness under the same 40-task generalist protocol, OpenVLA reaches 16.6 (16.6 / 14.2 / 28.0 / 7.6 per suite, spatial / object / goal / long) and Diffusion Policy (Chi et al., 2023) 10.3 (10.0 / 16.0 / 10.8 / 4.2); no official generalist number exists for either, and their official *per-suite finetuned* averages are 76.5 (OpenVLA) and 72.4 (Diffusion Policy as reproduced in the OpenVLA paper). The OpenVLA-OFT combined checkpoint re-evaluated on our harness scores 95.5 (official 96.8). **Reproducibility checks.** An independent rerun of the 2B scratch 30k point gives 56.2 vs. 56.3; the summed adapter matches sequential application to 1.1×10^{-8}; a fixed-seed re-evaluation on the official CALVIN task-D configuration reproduces the 0.8-epoch value; every registration file precedes its evaluations, by an ISO/UTC stamp inside the file or, for the G and H registrations, which carry only local-time estimates, by file modification time (Tables 27–30); the G0 fresh-head run reproduced the cached frozen-base probe row exactly (0.639 / 0.033), the registered equivalence check for reusing cached features; the G4 control reproduced a working full-FT policy through the fresh-head path to within 0.1 point; the E7 cached-feature control reproduces the paper's probe values (top-8 0.917 vs. 0.914 at 40k frames for the seed-0 $3\mathrm{e}{-5}$ cell; 0.928 vs. 0.925 for the $1\mathrm{e}{-5}$ cell); the training-prompt re-evaluations (E6b, E6c) are paired per episode with the original logs on identical initial states (exact McNemar); and the CS scene-A build reproduces the original evaluator's per-chain results on scene D (5/5 on each of the first four chains) and its first-observation action exactly. Matched-protocol runs use raw (unnormalized) actions uniformly, differing from StarVLA's normalized actions. The compositional breakdowns of Appendix F(b) are recovered from evaluation logs already on disk and required no new rollouts.