WHAT BEHAVIOR CLONING GIVES AND TAKES: ACTION LEARNING AND VISUAL REORGANIZATION IN VLAS

Anonymous authors
Paper under double-blind review

ABSTRACT

A VLA can forget an answer without forgetting what it saw. During $\pi_{0.5}$ adaptation on LIBERO, a base-fitted localization readout degrades from 1.55 to 6.28 cm, yet refitting the same readout family on the adapted checkpoint reduces error to 1.13 cm. The visual information remains linearly recoverable; the pre-adaptation reader no longer matches its coordinates. We follow two behavior-cloning runs and align closed-loop control with native heads, fixed and refit probes, six-site representations, and matched interventions. The trajectory has three stages. Pre-adaptation readouts fail by 500 updates while success is still 0–5%. Control then rises to 85–88% by 4–8k while spatial information remains recoverable. As success reaches 94–97%, part of the checkpoint-local recoverability declines. A rank-64 orthogonal alignment closes over 95% of the destination-readout gap, the native language head fails on a separate schedule, and the action expert builds another destination code behind a frozen backbone. Gradient conflict arrives after the first remapping, and freezing its strongest layers offers no advantage over freezing random layers. Freezing and replay preserve selected readouts at an action cost; an auxiliary action-token objective accelerates the earliest action endpoint. Behavior cloning reorganizes the relations among representations, readers, and actions. Training should maintain or deliberately relearn those relations instead of preserving parameters indiscriminately.

1 INTRODUCTION

After behavior cloning, $\pi_{0.5}$ can fail to report a target location even though the same checkpoint supports a more accurate location decoder than the pretrained model. This is the central observation of the paper. A score that looks like visual forgetting can arise because the representation lost information, because its reader no longer matches the code, or because a downstream module stopped using the code. These cases produce the same failed answer and require different repairs.

Consider a cup-location query. The image first produces an internal code. A language head or probe translates that code into an answer. The action expert, the policy module that maps fused vision-language features and robot state to action chunks, consumes a related code to move the robot. We call each learned mapping from an internal code to one of these consumers an *interface*. Capability belongs to this relation among representation, reader, and behavior. Adaptation can preserve the content of a code while changing the coordinate system expected by its reader.

We separate the relation into three measurements. *Native compatibility* tests pretrained heads and base-fitted probes. *Checkpoint-local recoverability* refits the same probe family on each checkpoint. *Control utility* measures closed-loop success. Their order in training reveals which part of the relation changed.

We apply this measurement to two full-parameter $\pi_{0.5}$ runs on LIBERO (Physical Intelligence et al., 2025; Liu et al., 2023). At matched checkpoints, we align success and action dispersion with fixed/refit probes, the native language head, and representation similarity. Six capture sites trace the change from the vision tower to the action expert. Low-rank alignment describes its geometry; same-observation gradients and matched freezing test a proposed cause.

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.



Figure 1: Three stages of behavior-cloning adaptation. Base-fixed spatial readouts fail before control forms; control then rises to 85–88% as the native LM head fails and repeated gradient conflict appears; finally, success reaches 94–97% while action dispersion contracts and checkpoint-local target recoverability declines.

Figure 1 shows a three-stage trajectory. By 500 updates, old spatial readers have failed while success remains at 0–5% and checkpoint refits are stronger than base. From 500 to 8k, success rises to 85–88% and the native language head fails, although matched spatial readouts remain strong. From 8k to 30k, success reaches 94–97%, actions concentrate, and part of the checkpoint-local recoverability declines. Behavior cloning changes how vision is read before it learns reliable control, then compresses part of the recoverable code after control has nearly saturated.

This order predicts which interventions should work. Freezing the vision-language model retains old readers but loses 16–19 action-success points. Teacher anchoring uses a feature-matching loss to preserve selected base activations, yet leaves the native language head at 0%. Semantic replay interleaves labeled vision-language examples with robot training; it improves object readout while reducing early action success from 15/100 to 1/100. A FAST-token auxiliary head predicts discretized actions during training and raises 500-step success from 5/200 to 20/200, with heterogeneous effects at 2k.

Our contributions are:

- A three-stage law of adaptation: reader remapping precedes control acquisition, and reduced recoverability appears after control approaches saturation.
- A structural account of the remapping: a low-rank orthogonal map closes over 95% of the destination gap, and the action expert creates a second code behind a frozen VLM.
- A training consequence: parameter-preserving interventions trade against control, whereas changing the action-learning objective accelerates the earliest stage.

2 Related Work

2.1 VLA Pretraining and Adaptation

$\pi_{0.5}$ combines web-scale vision-language objectives with robot data and an action expert, making it a natural setting for asking what survives downstream adaptation (Physical Intelligence et al., 2025). OpenVLA provides complementary evidence that large pretrained vision-language backbones can be adapted into robot policies (Kim et al., 2024). VLM4VLA shows that generic VLM and embodied-task scores do not consistently predict control quality and identifies visual processing as a major bottleneck (Zhang et al., 2026). Our question is temporal and diagnostic: for one architecture and adaptation recipe, which measured capabilities become incompatible with old readouts, which remain recoverable, and which support control?

2.2 Grounding Degradation and Representation Tracing

ST4VLA studies competition between spatial grounding and control and introduces spatial supervision to reduce grounding degradation (Ye et al., 2026). VLA-Trace relates representation changes

to behavioral interventions across VLA stages (Shi et al., 2026). Both motivate looking inside the adapted policy. We follow dense, matched checkpoints and compare fixed with refit readouts to distinguish deleted information from changed coordinates.

2.3 Optimization Interference and Retention

GAP attributes failures in vision-proprioception policies to modality-gradient competition during motion transitions and uses phase-guided adjustment to mitigate it (Lu et al., 2026). Critical learning periods have also been observed in multisensory integration (Kleinman et al., 2023). Language-action pretraining and flatness-aware fine-tuning offer alternative ways to preserve or redirect capabilities (Lin et al., 2026; Zhang & Bisk, 2026). We test a related but distinct hypothesis: whether the action and visual objectives conflict on the *same observation*, at the same site, before visual readout degradation. This temporal and paired requirement prevents generic high-dimensional gradient disagreement from being interpreted as a mechanism.

2.4 What Probes Establish

Linear probes establish decodability under a function class rather than model use (Alain & Bengio, 2016; Hewitt & Liang, 2019). A base-fitted probe tests compatibility with an old coordinate system; a checkpoint refit tests recoverability under the same function class. We report both and keep closed-loop behavior as a separate outcome.

3 What Does BC Adaptation Change? A Measurement Framework

Our framework tests whether adaptation changes the available code, its compatibility with a reader, or its contribution to behavior. These are different properties of a capability rather than alternative metrics for one scalar. For checkpoint t, closed-loop control utility is the mean episode success

$$A(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\text{episode } i \text{ succeeds}], \tag{1}$$

and $D(t)$ is the mean pairwise squared distance among action chunks under a fixed sampling protocol. $A(t)$ measures successful control; $D(t)$ measures sampler-conditioned action concentration.

Let $f_t^s(x)$ be the activation for observation x at site s and checkpoint t, with $t = 0$ denoting the pretrained model. We transfer a readout fitted at the base checkpoint and refit the same family at every checkpoint:

$$R_{\text{fixed}}(t, s, c) = \mathcal{M}_c\big(q_{0,c}^s(f_t^s(x)), y_c\big)\,, \tag{2}$$

$$R_{\text{refit}}(t, s, c) = \mathcal{M}_c\big(q_{t,c}^s(f_t^s(x)), y_c\big)\,. \tag{3}$$

The fixed score measures compatibility with the base coordinate system; the refit score measures information recoverable at the current checkpoint. Fixed decline with stable refit identifies a changed code–reader relation. Joint decline identifies reduced recoverability within the probe family. Closed-loop control forms a separate axis because probe scores measure decoding rather than behavioral use.

We test a proposed optimization mechanism on the same robot observation. For action loss L_{act} and target-position loss L_{vis}, we compare their gradient cosine at each parameter group against a deterministically shuffled-target null:

$$C_s(t) = \frac{\langle \nabla_{\theta_s} L_{\text{act}}, \nabla_{\theta_s} L_{\text{vis}} \rangle}{\|\nabla_{\theta_s} L_{\text{act}}\|_2 \|\nabla_{\theta_s} L_{\text{vis}}\|_2}. \tag{4}$$

A gradient mechanism must precede the readout change and predict a matched intervention. We therefore compare the paired cosine with a shuffled-target null and freeze the repeatedly implicated layers against an equal-width random-layer control. The same logic governs all interventions: each diagnosis must predict which reader or behavior changes when its proposed component is constrained.

We adapt the LeRobot implementation of $\pi_{0.5}$ to LIBERO with full-parameter behavior cloning. Two independent runs contribute 15 checkpoints each; evaluated seed–step cells use 100 closed-loop episodes. At each checkpoint we measure the vision tower, visual projector, VLM fusion input, action-expert input, action-expert middle, and pre-action state. Fixed and refit probes measure task content. Centered kernel alignment, abbreviated CKA, measures overall representation similarity to the base model. Target and destination position, object identity, image-plane relations, and native question answering provide complementary readouts. Development and confirmation examples are grouped by simulator state, probe choices are frozen before confirmation, and dataset construction is reported in Appendix B.

The training run is the unit of replication; checkpoints and episodes are repeated measurements within a run. We attribute a mechanism when its signal exceeds a matched null, precedes the measured change, and predicts the corresponding intervention (Appendix A).

4 The Coupled Dynamics of Action Learning and Visual Reorganization

4.1 Does Action Acquisition Consume Visual Information?

Control grows after old readers fail and before the measured spatial code becomes difficult to recover. The two full-fine-tuning runs achieve only 0/100 and 5/100 successes at 500 updates, rise to 85–88/100 by 4–8k, and reach 97/100 and 94/100 at 30k. Under a fixed sampler, pairwise action MSE contracts from 0.588 at 500 to 0.085 at 30k, and gripper-sign entropy reaches zero after 8k. Reader compatibility, control acquisition, and checkpoint-local recoverability therefore occupy three successive phases.

The 4k vision tower isolates compatibility from recoverability. A localization readout fitted at the base model has 1.55 cm error on base features and 6.28 cm on adapted features. Refitting the same ridge family at 4k reduces the error to 1.13 cm. The transferred reader incurs four times the error even though the current representation supports a better matched linear readout than the base representation.

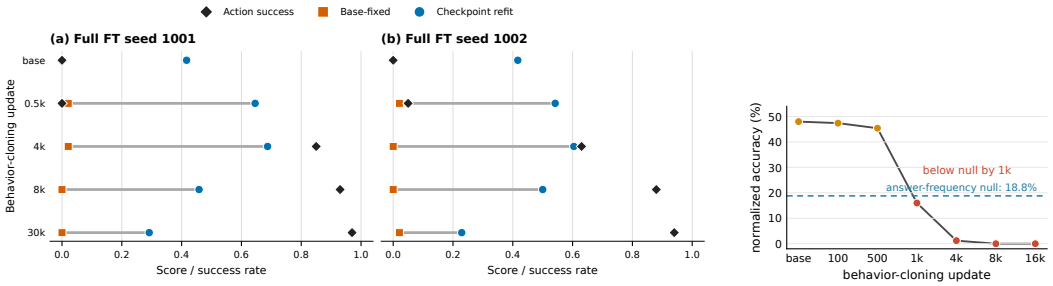
Destination decoding shows that this separation is a training pattern rather than a single target-localization example. The base representation reaches $R^2 = 0.983$. By 8k, the two fixed readouts fall to 0.212 and 0.341, while checkpoint refits reach 0.975 and 0.979. The destination code has moved away from the pretrained coordinates while retaining nearly all matched linear recoverability.

Figure 2 exposes the order of events at matched milestones. We define hit@2 cm as the fraction of predictions within two centimeters of the target. At 500 updates, this score has already fallen from 0.417 to 0.021 under the fixed readout in both runs, although action success is only 0.00 and 0.05. The refit scores instead rise to 0.646 and 0.542. Action success reaches 0.85–0.88 by 4–8k, and refit localization declines later. Compatibility loss, control acquisition, and reduced recoverability are separated in time.

The native LM head follows a separate trajectory. The native language head shows the same loss of compatibility through the model's own reader. Grounded question answering, abbreviated GQA, measures whether that head can answer visual questions about the scene. Normalized accuracy falls from 48.0% at base to 45.4% at 500, 16.0% at 1k, 1.2% at 4k, and 0% at 8k and 16k in Figure 2. The answer-frequency baseline is 18.8%. By 1k, the pretrained language head has failed while matched target and destination decoders remain strong. Behavior cloning preserves a spatially informative code after the native path that expressed it has become incompatible.

Different visual contents then follow different trajectories. Object identity remains at ceiling on the current ROI bank, destination R^2 remains near base through 8k after refitting, and four-way relation accuracy shows an early fixed/refit gap followed by a smaller decline in refit performance. Adaptation changes the accessibility of particular contents rather than lowering every visual capability together; Appendix C reports the complete values and shortcut controls.

Finding 1. Behavior cloning first breaks backward compatibility, then builds control, and only later reduces checkpoint-local spatial recoverability.



(a) Action success and target-position readouts (b) Native LM-head readout

Figure 2: Action learning and visual readout change at matched checkpoints. (a) Gray segments connect base-fixed and checkpoint-refit target-position scores at the same checkpoint; diamonds show closed-loop success. (b) Native question-answering accuracy falls below the 18.8% answer-frequency baseline by 1k. Spatial information remains recoverable after both old geometric and native language readouts become unusable.

4.2 Where Does Reorganization Occur, and What Structure Does It Have?

Reorganization is structured and repeated: a low-rank change rewrites VLM coordinates, then the action expert builds another task-specific code. We project base and checkpoint features into a shared rank-64 subspace fitted on development data and solve the orthogonal Procrustes map that aligns paired observations. This map closes 95.6–99.1% of the destination fixed–refit gap across four checkpoints in one run and 96.6–97.2% at 8k in the second. A permutation test gives $p = 0.00498$. The early mismatch is largely an orderly coordinate change.

Reorganization extends through the policy. The six-site profile shows that compatibility and representation similarity answer different questions (Figure 3). Fixed-readout loss is largest and persistent at the vision tower and VLM fusion site, whereas refit scores remain near or above base through early and middle training at most sites. CKA gives another ordering: vision and projector activations remain highly similar to base, while VLM fusion and action-expert input move much farther. A representation can remain globally similar to base while invalidating a task-specific reader.

The first changed site is therefore not automatically the causal source. The vision tower is the earliest location where the old readout fails, yet its 4k refit still recovers target coordinates. An upstream coordinate change can propagate through every later capture point without identifying a uniquely damaged layer.

The action-only run demonstrates that downstream reorganization does not require VLM drift. Freezing the VLM holds vision-tower, projector, and fusion readouts at their base values. Yet destination fixed R^2 at action-expert middle and pre-action falls to -7.725 and -5.944, while refitting recovers 0.430 and 0.363. The action expert constructs a new destination code even when the upstream visual code cannot move.

Finding 2. Adaptation reorganizes code–reader relations twice: a low-rank VLM coordinate change invalidates old readers, and the action expert creates a second task code behind a frozen VLM.

4.3 Is Gradient Conflict the Cause?

Gradient conflict marks the later control-learning phase. Its timing and matched freeze intervention separate this signal from the earlier reader remapping.

The timeline rules out conflict as the trigger of the first compatibility break (Figure 4). The only groups significant at both 4k and 8k are VLM layers 4, 9, and 10. Their matched-minus-shuffled mean cosines at 4k are -0.0299, -0.0111, and -0.0129; at 8k they are -0.0816, -0.0574, and -0.0603, with all six batch-bootstrap intervals below zero. At 500 updates, when the fixed vision-tower readout has already collapsed, none of these layers differs reliably from its shuffled null. The repeated signal begins after the transition it would need to cause.

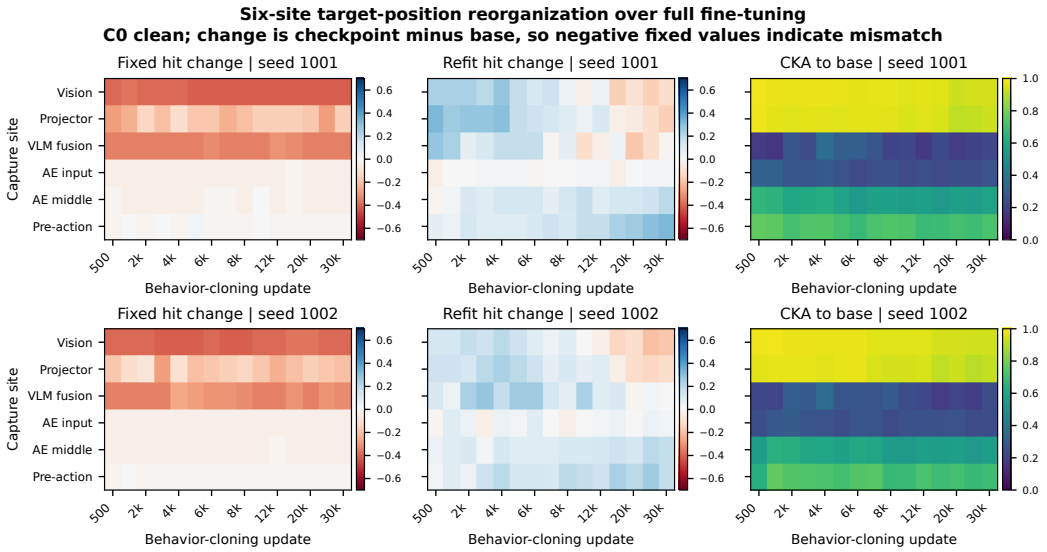


Figure 3: Six-site target-position profiles. Fixed and refit panels show checkpoint hit@2 cm minus the base score; negative values denote old-readout incompatibility and reduced matched-probe recoverability, respectively. CKA compares each checkpoint with base on the same observations. Rows are independent training seeds; columns are repeated checkpoints.

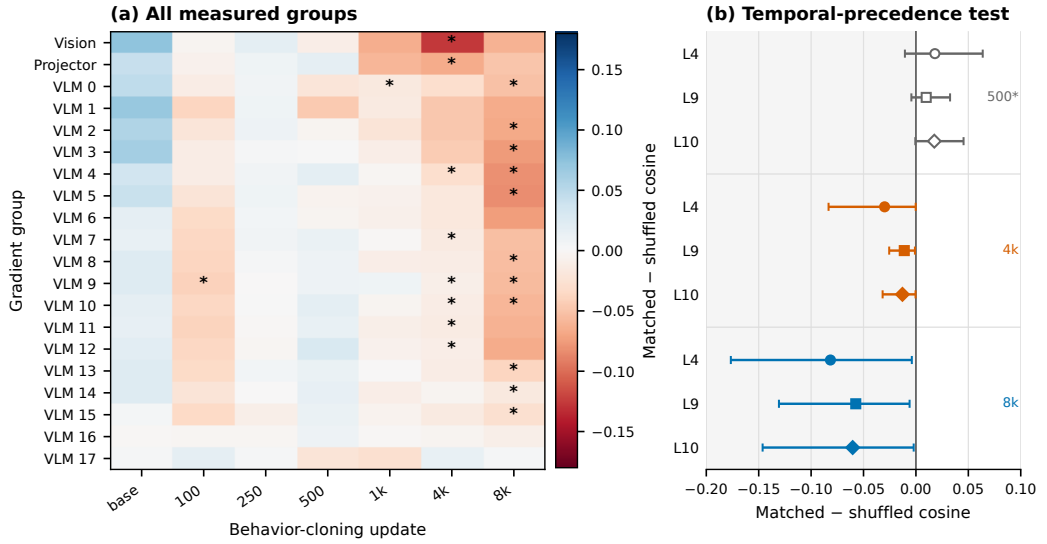


Figure 4: Same-observation action versus target-XY gradients. Heat-map cells are paired matched-minus-shuffled cosines; negative values indicate greater conflict for the matched target, and stars mark bootstrap 95% intervals below zero. The coefficient plot follows the three layers repeated at 4k and 8k. Their intervals cross zero at 500*, after the fixed readout had already collapsed.

We then branch the same 8k checkpoint to 16k with identical 800 batch indices. The target arm freezes layers 4/9/10, an equal-width control freezes randomly selected layers 2/12/17, and a restart control freezes neither. Clean target-XY refit scores are 35.42%, 47.92%, and 35.42%; nine-condition means are 28.70%, 37.96%, and 26.16%; closed-loop success is 98/100, 97/100, and 96/100. The random-layer control exceeds the diagnosed-layer arm on both recoverability summaries. The repeated gradient signal identifies a phase of adaptation rather than a privileged storage location.

Finding 3. Repeated gradient conflict follows the first compatibility break, and the layers it selects have no specific effect under matched freezing.

5 Can Diagnosis Guide Intervention?

The measurements above distinguish three training strategies: immobilize the visual representation, retain selected visual targets, or improve the action-learning signal. Their outcomes show which part of the representation–reader–behavior relation each strategy actually controls.

5.1 Can Freezing the VLM Preserve Knowledge Without Hurting Action?

We call this condition action-only training: the vision-language model is frozen while the action expert is optimized. It preserves old readers but prevents the policy from reaching the full-fine-tuning action endpoint. Native GQA remains at 48.0% and VQAv2 at 59.6%, while closed-loop success reaches 78/100 at 30k, compared with 97/100 and 94/100 for full fine-tuning. Fisher's exact test gives $p = 0.004$. The frozen VLM also permits substantial re-encoding inside the action expert, as shown in Section 4.2. Upstream compatibility is therefore insufficient for full control and costs 16–19 success points.

5.2 Do Anchoring and Replay Protect the Native Readout?

Anchoring and replay preserve what they supervise rather than the full route from vision to language and action. Local teacher anchoring matches selected activations to a frozen base-model teacher. It reaches 90/100 versus 93/100 at 8k but leaves GQA at 0% and preserves only selected early features. Semantic replay interleaves labeled visual examples with robot batches. Relative to a compute-matched replay arm with shuffled labels, it raises object-choice accuracy from 43.2% to 84.4%, lowers GQA from 49.2% to 37.2%, and reduces action success from 15/100 to 1/100. McNemar's test on action success gives $p = 0.0005$. Protecting one endpoint can reallocate capacity away from another reader or from early control acquisition.

5.3 Can an Auxiliary Objective Accelerate Action Acquisition?

FAST changes the action-learning signal while leaving rollout architecture unchanged. The auxiliary head autoregressively predicts discretized action tokens during training; rollout still uses the original flow expert. At 500 updates, paired success rises from 5/200 to 20/200. The paired McNemar test gives $p = 0.00260$. At 2k, one seed is unchanged at 66/100 in both conditions, while the other improves from 57/100 to 73/100. The pooled comparison is 139/200 versus 123/200, with $p = 0.05189$. FAST reliably moves the earliest action endpoint; by 2k, its effect is seed-dependent and the pooled test is above 0.05.

Table 1: Completed interventions separate compatibility, retention, and action acquisition.

Arm	Comparison	Result
Action-only 30k	Native retention vs. action	GQA 48.0%; action 78/100
Teacher anchor 8k	Retention without action loss	90/100; GQA 0%
Vision/random replay 500	Content specificity vs. action	Object 84.4/43.2%; action 1/15
Freeze 4/9/10	Diagnosed vs. random layers	Refit 35.42/47.92%; action 98/97
FAST auxiliary	Token head plus flow action	500: 20/200 vs. 5/200; 2k: 139/200 vs. 123/200

The interventions expose a common constraint. Freezing, anchoring, and replay preserve exactly the endpoint they supervise while leaving other readers or action acquisition behind. FAST improves early action acquisition by changing the learning signal rather than fixing the representation in place. The useful object of preservation is the mapping that connects a changing code to a reader or action consumer.

Finding 4. Retention methods move the cost among action ceiling, native readout, and learning speed. FAST is the only tested intervention that improves the earliest action endpoint, with heterogeneous effects by 2k.

6 Discussion

The central conceptual result is that VLA capability is relational. A representation does not by itself answer a question or generate an action; it must remain compatible with a reader, and the resulting signal must affect behavior. Full fine-tuning separates these conditions: it can raise control success, invalidate a pretrained head, preserve checkpoint-local spatial information, and later reduce that recoverability. A knowledge-retention claim must therefore name the reader and the behavior that remain functional.

This view changes how forgetting should be measured. A fixed probe is a test of backward compatibility, not a direct inventory of information. A checkpoint refit tests recoverability, not policy use. High closed-loop success tests control on the adapted distribution, not preservation of the pretrained semantics. The rank-64 alignment makes the distinction concrete: a large fixed–refit gap can be explained largely by a change of coordinates. Reporting all three quantities prevents readout mismatch from being mislabeled as deletion and prevents decodability from being mislabeled as use.

The same view changes the training target. Freezing the whole VLM preserves native readers but lowers action success; anchoring and replay preserve selected endpoints; the action expert re-encodes even behind a frozen backbone; and FAST shows that an acquisition objective can move the early action trajectory. The practical goal is functional compatibility: either maintain the mappings from a changing visual code to the readers that matter, or relearn those mappings deliberately. Parameter proximity is useful only when it preserves those relations.

7 Conclusion

Behavior cloning reorganizes visual capability before and during action learning. In $\pi_{0.5}$, it first changes the relation between visual codes and their readers, then builds reliable control, and later reduces part of checkpoint-local recoverability. Old geometric and language readers fail early, much of the spatial code remains accessible under a new reader or low-rank alignment, and the action expert creates another code behind a frozen VLM. This sequence explains why parameter-preserving interventions protect selected measurements yet constrain control: they preserve a component while adaptation is changing the relation among components. VLA training should target functional compatibility by maintaining or deliberately relearning the mappings that connect representations to answers and actions.

AI Use Statement

Generative AI tools were used to assist with code inspection, experiment bookkeeping, manuscript organization, literature discovery, and language editing. All generated claims, citations, code, and numerical results were checked against source artifacts by the authors, who take responsibility for the final content.

Reproducibility Statement

We retain frozen data manifests, checkpoint and input hashes, evaluation scripts, development/confirmation split rules, replacement logs, and an evidence ledger linking each reported result to its source artifact.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods for multisensory integration in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24296–24305, 2023. doi: 10.1109/CVPR52729.2023.02327.

Tao Lin, Yuxin Du, Yiran Mao, Zewei Ye, Yilei Zhong, Bing Cheng, Yiming Wang, et al. LA4VLA: Learning to act without seeing via language-action pretraining. *arXiv preprint arXiv:2606.27295*, 2026. URL https://arxiv.org/abs/2606.27295.

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*, 2023. URL https://arxiv.org/abs/2306.03310.

Jingxian Lu, Wenke Xia, Yuxuan Wu, Zhiwu Lu, and Di Hu. When would vision-proprioception policies fail in robotic manipulation? In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.12032.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

Haoyuan Shi, Xiancong Ren, Yingji Zhang, Qinfan Zhang, Jiayu Hu, Haozhe Shan, Han Dong, et al. 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.

Jinhui Ye, Fangjing Wang, Ning Gao, Junqiu Yu, Yangkun Zhu, Bin Wang, Jinyu Zhang, et al. ST4VLA: Spatially guided training for vision-language-action models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.10109.

Haochen Zhang and Yonatan Bisk. Flatness preserves instruction following in vision-language-action models. *arXiv preprint arXiv:2606.23641*, 2026. URL https://arxiv.org/abs/2606.23641.

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. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.03309.

A Attribution Criteria

We attributed action–visual gradient conflict when the paired statistic differed from the shuffled-target null, preceded the readout change at the same site, and predicted a matched freeze intervention. Teacher anchoring and visual replay were evaluated jointly on their supervised retention endpoint and closed-loop action success. Random replay matched semantic replay in samples, tokens, update times, and compute.

B Dataset and Probe Details

The geometry bank contains 1,400 accepted records split by layout into 560 development and 840 confirmation examples. Six-site profiles contain 5,832 rows; the closed-loop matrix contains 4,800 episodes; and the same-observation bank contains 40 development and 40 confirmation states. Base-fixed probes share one frozen development fit, while checkpoint refits use the same development examples at every checkpoint. Dataset manifests, fit hashes, checkpoint hashes, and replacement chains are retained with the released artifacts.

Replay objective. Visual replay uses MS COCO 2014, VQAv2, and RefCOCO-family public data for external visual retention and a separately generated LIBERO object bank for robot-domain retention. Answer and location tokens receive autoregressive cross-entropy; prompt tokens are masked. Random replay uses the same inputs and target lengths with deterministic within-bucket label derangement.

C Additional Results

Checkpoint-local semantic profiles. Object identity remains at 100% under fixed and refit probes at 500 and 1k, and checkpoint-refit identity remains 100% in both 8k runs. Four-way relation accuracy changes from a 36.66% base score to 26.10/36.74% fixed/refit at 500, 24.81/31.31% at 1k, and 24.89/29.07% and 25.00/29.26% in the two 8k runs. Destination R^2 changes from 0.983 at base to 0.649/0.989 at 500 and 0.402/0.984 at 1k. A location-only object baseline scores 37.29%.