PRESENT BUT NOT REMEMBERED: AUDITING HOW FROZEN VLAS ENCODE, DEPLOY, AND STEER VISUAL HISTORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Across Octo-Small, Octo-Base, and CronusVLA, linear probes recover previous-frame content, while predicting the residual left by a current-frame predictor remains difficult. The reported probe sweeps yield residual-target R^2 scores of at most 0.019, alongside raw-target scores of 0.142–0.389. We examine how this shared pattern relates to action outputs using representation interventions on BridgeData V2 observation pairs. Blacking out the current image increases the action displacement caused by replacing history in Octo and decreases it in CronusVLA. In the detailed Octo-Small analysis, the occlusion-related increase is concentrated in early layers. A separate Octo-Small injection study finds a small shift toward the action associated with the injected history ($+0.069$, 95% CI $[0.043, 0.094]$); estimates for clean-action recovery and an ambiguous-observation test remain uncertain. Cross-model evidence thus covers residual prediction and occlusion responses, while recovery is an Octo-Small case study. These open-loop results motivate evaluating visual memory through its added predictive value, the action changes it produces, and recovery toward a specified action target.

1 INTRODUCTION

History gives a vision-language-action (VLA) policy access to earlier observations when interpreting the present. To understand what this access contributes, an evaluation must connect the content of a history representation to the policy's actions. Adjacent camera frames share much of the same scene. A decoder can therefore recover a previous frame by exploiting content also visible in the current frame. Comparing raw reconstruction with prediction of the residual left by a current-frame predictor helps identify which part of that result merits closer investigation.

Recent robot policies provide temporal context through memory modules, retrieval, longer observation windows, and visual traces (Shi et al., 2025; Sridhar et al., 2026; Torne et al., 2025; Jang et al., 2025; Zheng et al., 2024; Li et al., 2026). Their task evaluations measure the performance of the resulting systems. Our question concerns the history representations inside such a policy: what frame content can be recovered, how replacing that content changes the action, and whether a targeted replacement helps recover the action under visual disruption. These measurements connect representation analysis to a concrete use of memory: supporting action selection as the available visual input changes. Figure 1 locates the three measurements in the policy and makes their different targets explicit.

We study Octo-Small, Octo-Base (Ghosh et al., 2024), and CronusVLA (Li et al., 2026) using observation pairs from BridgeData V2. Three findings structure the analysis. **Low residual reconstruction recurs across all three checkpoints**, alongside positive raw reconstruction scores. **The action response to replacing history changes in opposite directions under occlusion**: it increases in Octo and decreases in CronusVLA. Thus, policies with a shared probe pattern can respond differently to the same type of visual disruption. **Octo-Small supports a small directional action shift, while recovery remains uncertain**. Injecting a different history shifts the action toward the output associated with that history; supplying the sample's own history under occlusion yields a broad interval around a near-zero recovery estimate.

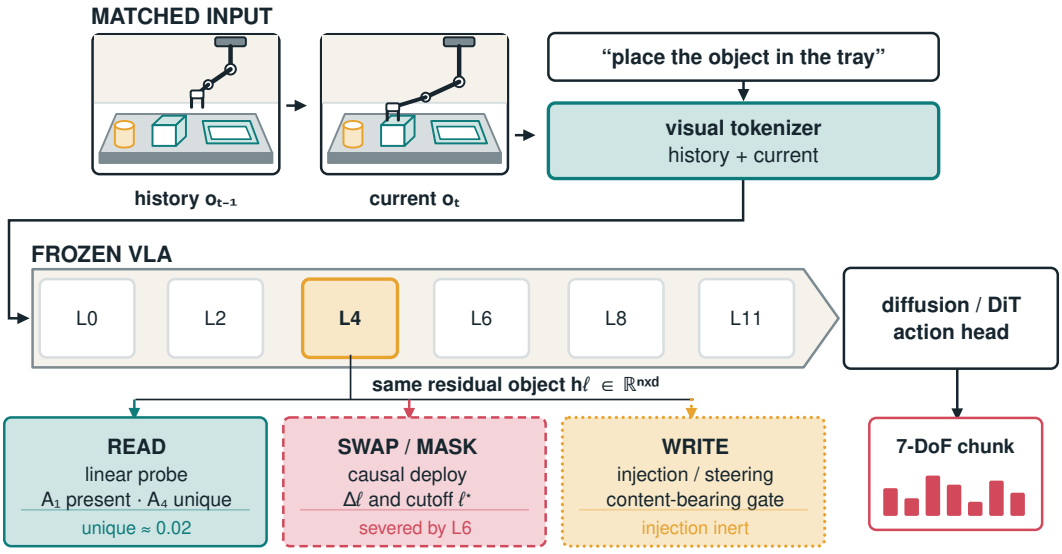


Figure 1: **Connecting history representations to action outputs.** The overview locates three measurements within a multi-frame policy. Read: reconstruct the previous frame and the residual left by a current-frame predictor. Swap/mask: compare action displacement after replacing history or masking attention, separately for clean and occluded inputs. Write: measure donor-directed steering and recovery toward the recipient's clean action. †The original in-panel "severed by L6" and "injection inert" labels require correction; the reported results show stronger early occlusion gaps and a small positive Octo-Small steering contrast.

The first two findings have cross-model evidence. We examine their depth dependence in Octo-Small, and use Octo-Small for recovery tests and tests with ambiguous current observations (state aliasing). CronusVLA provides an additional directional-steering comparison. All action measurements are open-loop: we compare policy outputs for fixed observations. This scope makes the findings directly relevant to interpreting probes and activation interventions, with closed-loop task evaluation as the next step for a proposed memory method.

For a researcher evaluating memory, the practical consequence is to keep three comparisons explicit. A residual probe asks what can be predicted after accounting for a current-frame predictor. A history intervention asks how much the action changes within a specified observation condition. A recovery test asks whether that change moves the action toward a chosen target. Reporting these together makes it possible to identify a directional effect and the uncertainty about its intended benefit within the same study.

2 Related Work

Representations and action control. Existing VLA analyses connect internal features and input pathways to action outputs. Grant et al. (2026) find strong visual control and task-dependent language sensitivity across models. Häon et al. (2025) identify internal directions for steering speed and motion, with simulation and physical-robot evaluation. Buurmeijer et al. (2026) study feature observability and controllability, while Molinari et al. (2025) probe state-transition representations in OpenVLA. Our comparison focuses on previous-frame reconstruction, its residual after current-frame prediction, and the action response to replacing history under different observation conditions.

Temporal context and evaluation. MemoryVLA, MemER, long-context diffusion policies, ContextVLA, and TraceVLA provide different mechanisms for using earlier observations (Shi et al., 2025; Sridhar et al., 2026; Torne et al., 2025; Jang et al., 2025; Zheng et al., 2024). LIBERO studies lifelong knowledge transfer (Liu et al., 2023), and SimplerEnv evaluates real-world manipulation policies in simulation (Li et al., 2025). We complement this task-level perspective with controlled observation tests on BridgeData V2. Action memorization in diffusion policies (He et al., 2026)

concerns retrieval of training actions; the present study examines responses to observations supplied at inference time.

Measurement tools. Linear probes test which target properties can be predicted from representations (Alain & Bengio, 2017; Belinkov, 2022). Interchange and activation patching measure output changes after replacing activations at selected internal locations (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Wang et al., 2023). Activation steering measures directed output changes (Turner et al., 2023). We use these tools to compare reconstructible content, action displacement, and recovery. Appendix I gives the extended literature comparison.

3 Study design and measurements

3.1 Comparisons, observations, and action outputs

Octo-Small (27M) provides the detailed layer analysis. Octo-Base (93M) tests whether the probe pattern and occlusion response recur at a second size in the same family. Both use two-frame inputs and a diffusion action head. CronusVLA-0.5B supplies a comparison with a different history representation: a per-frame feature, called a cognition feature, that conditions a diffusion-transformer action head. This comparison tests the recurrence of the observed patterns across the selected checkpoints. Architecture, training, size, and intervention location vary together across families.

We use a fixed set of 250 observation pairs (o_{t-1}, o_t, c) from BridgeData V2 (Walke et al., 2023), where o_{t-1} and o_t are consecutive images and c is the instruction. For each intervention, the instruction stays fixed. The main Octo-Small layer comparison uses 200 pairs, and the image-perturbation comparison uses 160. Octo-Small injection results are reported over 250 pairs and three seeds. The additional seed runs and their sample counts are listed in Appendix A. Policy weights remain fixed; linear probes are fitted to the extracted activations.

Observation conditions. The *clean* condition supplies the complete current image. The *occluded* condition blacks out that image and retains the previous image. Partial masking, heavy masking, and Gaussian blur test how the response changes across other forms of visual disruption. These matched conditions let us compare action responses while controlling the supplied history. Figure 2 illustrates the inputs, and Appendix H maps the original condition identifiers to these names.

Action outputs. We compare sampled policy outputs for fixed observations, an *open-loop* evaluation. Each action chunk is a sequence of seven-dimensional actions. Outputs are de-normalized using the checkpoint's dataset statistics and averaged over diffusion samples; $a(x)$ denotes the resulting action summary for observation pair x. The reported distances below apply to this summary. Environment specifications are collected in Appendix C.

3.2 What can be reconstructed from history?

We compare two prediction targets to separate raw reconstruction from residual reconstruction. The raw target y_{t-1} is the previous RGB frame, principally downsampled to 16×16 pixels (RGB16). A linear ridge probe g_ℓ predicts it from the history representation h_ℓ at layer ℓ:

$$A_1(\ell) = R^2\big(y_{t-1}, g_\ell(h_\ell)\big). \tag{1}$$

For Octo, h_ℓ pools activations at the token positions corresponding to the previous image (Figure 3). We call A_1 the *raw reconstruction score*.

A predictor using the current frame first estimates the previous frame as $\hat{y}_{t-1|t}$. Its error, $r_{t-1} = y_{t-1} - \hat{y}_{t-1|t}$, is the target for a second probe q_ℓ:

$$A_4(\ell) = R^2\big(r_{t-1}, q_\ell(h_\ell)\big). \tag{2}$$

We call A_4 the *residual reconstruction score*. A higher score indicates better linear prediction of the part left unexplained by this current-frame predictor. The reported evaluation uses five folds and searches layer, pooling, target, and regularization choices. The notation A_1, A_4 is retained to match the original outputs; probe configurations appear in Appendices B, E, and H.

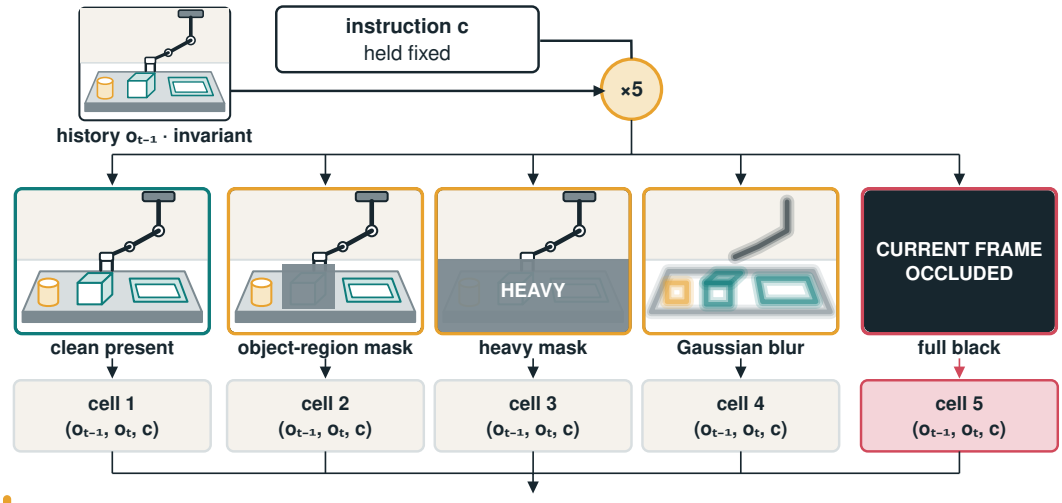


Figure 2: **Matched image-perturbation conditions.** History and instruction stay fixed while the current image is complete, partially masked, heavily masked, blurred, or fully blacked out. Compare each condition with the clean input for the same observation pair. This design tests how the current image changes the action response to a history intervention. The corresponding Octo-Small results are in Figure 5a and Table 7.

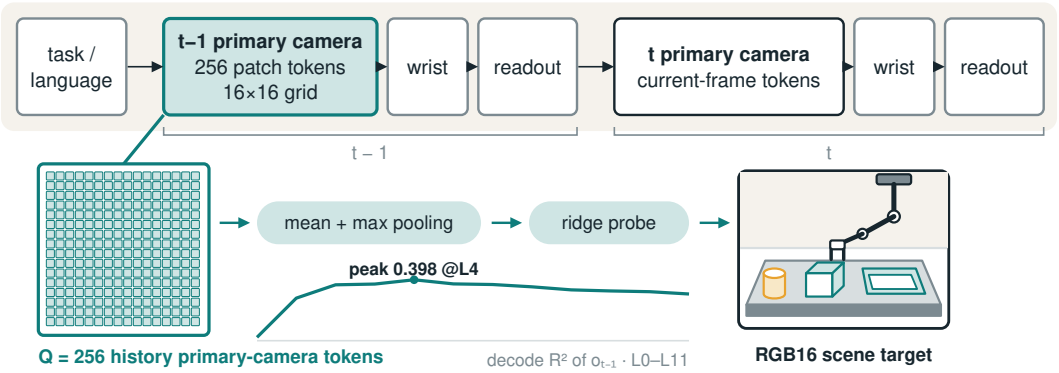


Figure 3: **Reconstructing the previous frame from history tokens.** The probe pools the history primary-camera tokens to predict an RGB16 image. The displayed layer sweep reaches $R^2 = 0.398$ at L4. This schematic identifies the sampled representation and raw reconstruction target; the residual comparison in Eq. 2 tests prediction of the error left by a current-frame predictor.

3.3 How much does replacing history change the action?

We use *interchange* to replace the selected history representation of one observation pair with that of another. The pair receiving the replacement is the *recipient* x; the pair providing it is the *donor* d. The recipient's current image and instruction stay fixed. With D denoting the reported action-distance calculation, the raw displacement is

$$C_\ell^{\text{raw}}(x, d) = D\big(a(\text{patch}_\ell(x; d)), a(x)\big). \tag{3}$$

A *self-swap* replaces the history with itself and measures variation from stochastic action sampling. Subtracting this reference gives the *adjusted intervention displacement*:

$$C_\ell(x, d) = C_\ell^{\text{raw}}(x, d) - C_\ell^{\text{raw}}(x, x). \tag{4}$$

We report C separately for clean and occluded inputs. Their paired difference, the *occlusion gap*, measures how occlusion changes the intervention displacement:

$$\Delta_\ell = C_\ell^{\text{occ_black}} - C_\ell^{\text{clean}}. \tag{5}$$

A positive gap means a larger displacement under occlusion; a negative gap means a smaller one. When the gap is near zero, the two displacements are similar. Reporting both C values therefore shows the action response in each condition as well as its change across conditions.

Attention knockout provides a second intervention. In Octo, readout tokens collect information used by the action head. Knockout masks their attention connections to history tokens at a selected layer. Interchange replaces activations, whereas knockout masks connections; comparing their depth profiles tests whether the early-to-late trend recurs across these two perturbations. Self-swap calibration is detailed in Appendix B.

3.4 Where does the action move, and does it recover?

The injection analysis compares direction and recovery. For *donor-directed steering*, we measure the cosine between the intervention-induced displacement and the direction toward the donor's clean-image policy output:

$$S(x, d) = \cos\big(a(\text{patch}_\ell(x; d)) - a(x),\ a_{\text{clean}}(d) - a(x)\big). \tag{6}$$

An unrelated-donor control supplies the reference for the reported paired contrast,

$$\bar{S} = \text{mean}(S_{\text{intended}} - S_{\text{control}}). \tag{7}$$

This contrast measures directional alignment relative to the control. Its exact donor assignment remains to be verified.

For *clean-action recovery*, the current image is occluded and we inject the recipient's own history representation from its clean run. We measure the reduction in distance to the same policy's clean-image output:

$$r_{\text{repair}} = D(a_{\text{occ}}, a_{\text{full}}) - D(a_{\text{inj}}, a_{\text{full}}). \tag{8}$$

Positive recovery means movement toward this reference action. We also examine *state aliasing*: different histories accompany an identical or ambiguous current observation. The paired test measures the change in a designated action branch. Appendix G gives the additional temporal-position intervention, which uses a fitted direction distinguishing the two frame positions. The estimates use the reported bootstrap intervals and seed summaries.

4 Results

Low residual reconstruction is shared across the three checkpoints, while the action response to occlusion differs by model. In Octo-Small, directional steering is measurable and recovery remains uncertain.

4.1 Low residual reconstruction recurs across three checkpoints

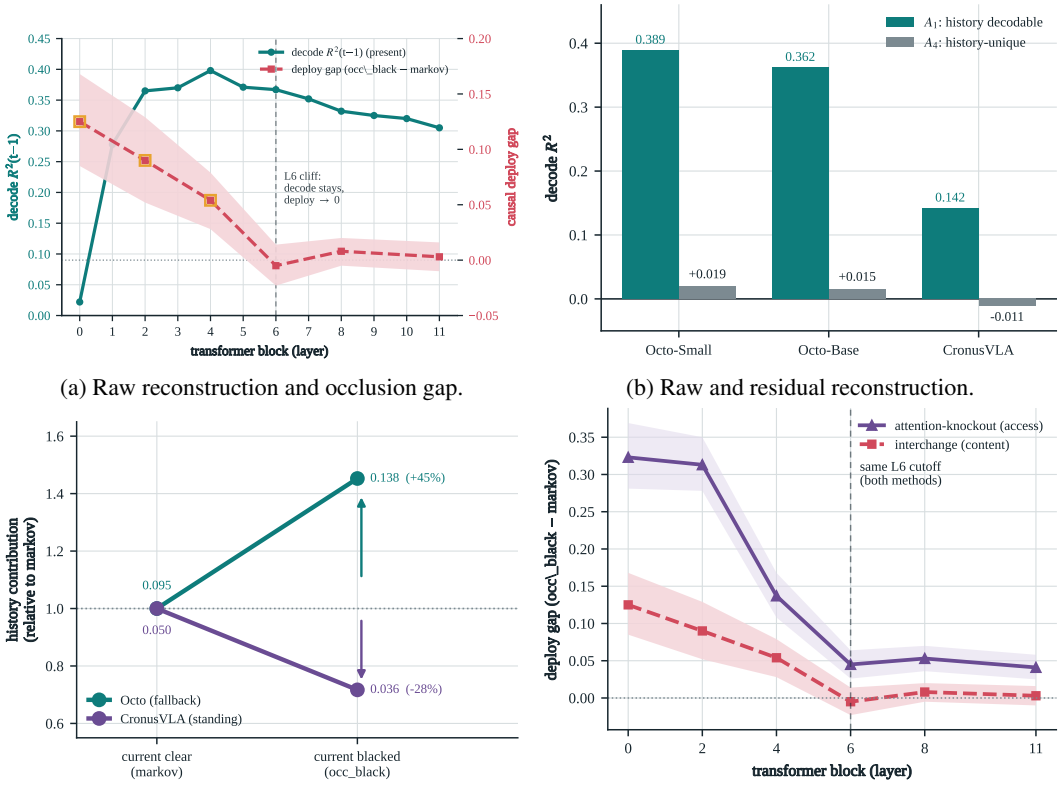
Across all three checkpoints, the residual reconstruction score remains close to zero while raw reconstruction is positive (Table 1). The reported best residual scores are $+0.019$ for Octo-Small, $+0.015$ for Octo-Base, and -0.011 for CronusVLA. Their raw-target sweep scores are 0.389, 0.362, and 0.142. Figure 4b places the two targets side by side: the common pattern is weak prediction of the residual left by the current-frame predictor.

Within Octo-Small, raw reconstruction remains positive across depth. Its main layer sweep rises from 0.022 at L0 to 0.398 at L4 and ends at 0.305 at L11 (Figure 4a). Thus, the low residual scores coexist with a history representation from which the raw previous frame remains reconstructible at late layers. The layer sweep and configuration sweep are separately reported analyses; their scores and configuration mapping are retained in Appendix A.

The residual result also depends on how the probe is fitted. Octo-Small's best listed residual configuration uses L2, mean-plus-max pooling, and ridge penalty $\alpha = 100$; its raw score is 0.248.

Table 1: **A shared reconstruction pattern accompanies opposite occlusion responses.** Raw and residual R^2 are reported configuration-sweep scores and can come from different probe configurations. The final column records the direction of the within-model change in adjusted intervention displacement under full occlusion.

Model	Raw A_1	Residual A_4	Displacement under occlusion
Octo-Small	0.389	$+0.019$	Increases
Octo-Base	0.362	$+0.015$	Increases (reported)
CronusVLA	0.142	-0.011	Decreases



(a) Raw reconstruction and occlusion gap.

(b) Raw and residual reconstruction.

(c) Opposite occlusion responses.

(d) Interchange and knockout profiles.

Figure 4: **Shared residual reconstruction, different action responses.** (a) Octo-Small's main layer comparison ($n = 200$) places raw reconstruction alongside the occlusion gap. (b) Separate model-wise probe sweeps show low residual scores. (c) Compare clean with occluded inputs within each model: intervention displacement increases in Octo and decreases in CronusVLA. (d) Interchange and knockout have stronger early-layer gaps; knockout retains positive late-layer gaps (Table 3). †Original in-panel "not deployed" and "cutoff" labels require correction; the plotted gap is defined by Eq. 5.

CronusVLA uses standardized cognition features and a different regularization range. The recurrence is therefore qualitative across the tested probe setups. Raw and residual R^2 use different target variances and are interpreted as separate prediction scores. Full configurations appear in Tables 12–13 and Appendix E.

4.2 Occlusion increases displacement in Octo and decreases it in CronusVLA

The shared probe pattern accompanies opposite action responses to current-image loss. Octo-Small's adjusted intervention displacement at L4 rises from 0.087 with a clean image to 0.141 under occlusion, giving a paired gap of $+0.054$ (95% CI $[0.028, 0.079]$). CronusVLA's three-seed mean

Table 2: **Occlusion-related increases are largest at early Octo-Small layers** ($n = 200$). Compare clean and occluded displacements within each row; the gap and its interval describe their paired difference. †The L11 displayed values have an unresolved subtraction discrepancy.

Layer	Clean C	Occluded C	Gap	Reported 95% CI	Excludes 0
0	0.356	0.481	$+0.125$	$[0.085, 0.168]$	Yes
2	0.271	0.361	$+0.090$	$[0.052, 0.129]$	Yes
4	0.087	0.141	$+0.054$	$[0.028, 0.079]$	Yes
6	0.045	0.040	-0.005	$[-0.023, 0.014]$	No
8	0.019	0.027	$+0.008$	$[-0.005, 0.020]$	No
11†	0.000	0.000	$+0.003$	$[-0.010, 0.016]$	No

Table 3: **Knockout gaps decrease early and remain positive at late layers.** Octo-Small, readout-to-history masking, all heads. Each gap is the occluded-minus-clean knockout effect. The reported intervals remain above zero after the early-layer decline. †The analysis sample count requires verification.

Layer	L0	L2	L4	L6	L8	L11
Gap	.323	.313	.137	.045	.053	.041
95% CI	[.28,.37]	[.28,.35]	[.11,.17]	[.03,.06]	[.04,.07]	[.03,.06]

instead falls from 0.0498 to 0.0357. Figure 4c compares these within-model changes. The direction of change remains informative even though the models use different intervention locations and reported distance scales.

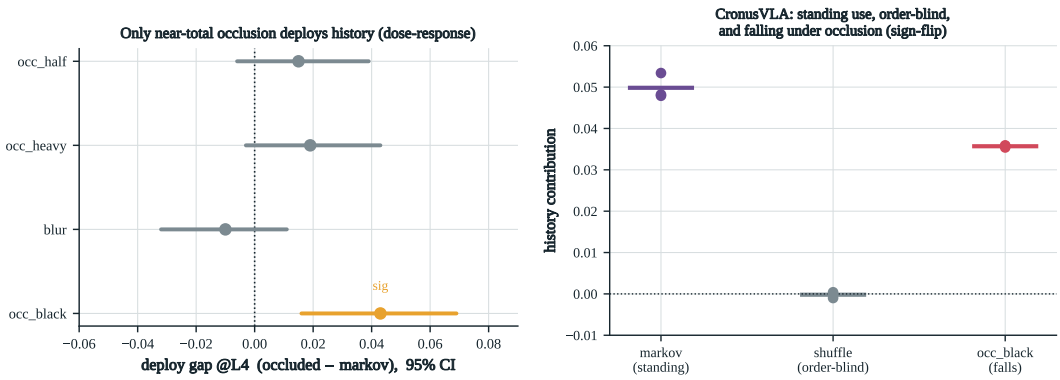
In Octo-Small, the occlusion-related increase is strongest early. Table 2 shows gaps of $+0.125$, $+0.090$, and $+0.054$ at L0, L2, and L4. At L6 and L8, the gaps are close to zero, with reported intervals spanning both signs. At L6, the clean and occluded displacements are 0.045 and 0.040, respectively: the small gap describes two similar action responses. Reading these values alongside the reconstruction curve in Figure 4a distinguishes persistent reconstructible content from the depth profile of the occlusion-related change.

Attention knockout provides a complementary view of the same depth trend (Figure 4d, Table 3). Its gaps fall from 0.323 at L0 and 0.313 at L2 to 0.137 at L4, followed by 0.045, 0.053, and 0.041 at L6, L8, and L11. All six reported intervals are above zero. The two interventions therefore support stronger early effects, with positive late-layer gaps retained by knockout. The layer pattern is a graded change whose late-layer detail depends on the intervention.

The clearest positive gap occurs under complete image loss. The matched image-perturbation comparison asks how this early response depends on the supplied current image. At L4 ($n = 160$), partial masking, heavy masking, and blur yield gaps of $+0.015$, $+0.019$, and -0.010, each with an interval spanning zero. Full blacking yields $+0.043$ $[0.016, 0.069]$. Figure 5a and Table 7 compare these conditions together. The clearest evidence for an increase thus comes from the most complete removal of current-image content among these tested conditions.

The direction recurs across the reported runs. Octo-Small's L0 and L4 gaps are positive in all five seed runs; CronusVLA's intervention displacement decreases under occlusion in all three. At Octo-Small L4, the seed gaps range from 0.009 to 0.054, with mean 0.034 and SD 0.016 (Table 8). This repeated direction supports the early-layer increase while showing variation in its size. CronusVLA's clean and occluded seed values are paired in Table 16.

Model size changes displacement magnitude within Octo. Octo-Base also shows attenuation of the interchange gap with depth. In the separate clean L0 size comparison, its displacement is 0.029, compared with 0.339 for Octo-Small. Normalizing by each checkpoint's full-history or full-current swap gives Base/Small ratios of about 0.31 (Table 9). The smaller displacement therefore persists relative to these within-model references. This two-checkpoint comparison complements the shared low residual reconstruction scores by showing another difference in action response.

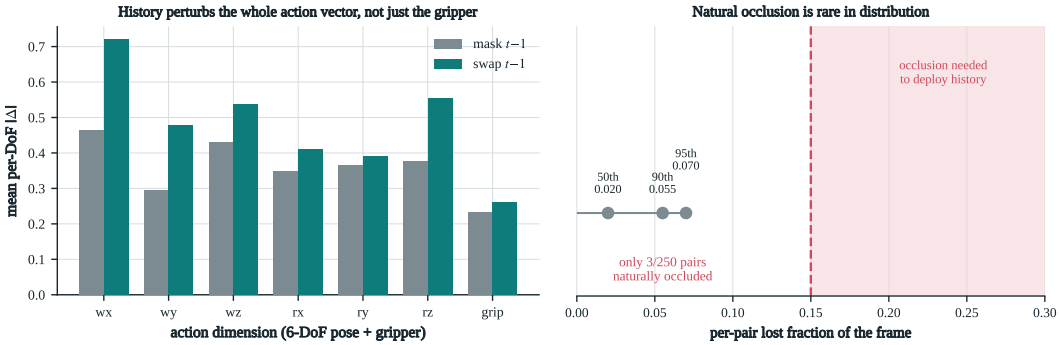


(a) Octo-Small image perturbations. (b) CronusVLA observation conditions.

Figure 5: **History interventions respond differently to visual disruption.** (a) Compare each Octo-Small perturbation with the clean condition at L4 ($n = 160$). The full-black gap is positive with a reported interval above zero; partial-mask, heavy-mask, and blur intervals span zero. (b) CronusVLA's mean intervention displacement decreases under occlusion in each of three seeds. †The original "fallback," "standing," and "order-blind" labels require correction. Shuffle construction and CronusVLA interval construction remain verification items.

The image conditions define the scope of the occlusion finding. A descriptive scan flags $3/250$ natural-occlusion candidates at the loosest threshold, and the 90th-percentile lost fraction is 0.055 (Figure 6b). Full-black inputs therefore probe a more complete loss of current visual content than is typical of this observation set. The result is most directly useful as a controlled robustness comparison. The scan measures visible image loss; tasks requiring velocity, contact history, or past events would need corresponding temporal targets.

History masking and swapping also displace several action coordinates in the clean condition (Figure 6a). The reported changes span translation, rotation, and gripper outputs. These measurements describe action responsiveness under ordinary observations, providing context for the additional displacement measured by the occlusion gap.



(a) Action-coordinate perturbations. (b) Image-loss statistic in the pool.

Figure 6: **Behavioral and observation-set context.** (a) History masking and swapping affect several action coordinates; the gripper is one component. (b) The image heuristic flags three candidates in 250 pairs at its loosest threshold. The plot characterizes this heuristic and pool, not the prevalence of every history-dependent state. †Scalar and per-coordinate behavioral summaries differ and need an aggregation map.

4.3 Octo-Small shows directional steering, with uncertain recovery

The Octo-Small injection results distinguish movement toward another sample's action from recovery of the recipient's clean-image action. The donor-directed cosine contrast is $+0.069$

Table 4: **Directional steering, recovery, and supporting injection checks.** Octo-Small results use a reported 250-pair set and three seeds; CronusVLA reports 250 observation pairs, three seeds, and three donors. The first three Octo-Small rows compare the intended outcomes. The remaining Octo-Small rows characterize layer, donor, and temporal-position effects. Units differ across rows. †Control definitions and CronusVLA's rounded lower endpoint require verification.

Model	Contrast	Estimate [reported 95% CI]
Octo-Small	Donor-directed cosine minus control	$+0.069$ [.043,.094]
	Clean-action recovery (own history)	-0.010 $[-.42, +.41]$
	Paired state-aliasing effect	$+0.004$ $[-.12, +.11]$
	L4/L8 intervention magnitude	7–9$\times$
	Donor separation minus null	$+11.0$ [9.6,12.5]
	Temporal-position direction minus random	-0.040 $[-.071, -.009]$
CronusVLA	Donor-directed cosine minus control	$+0.016$ $[-.000, .034]^{\dagger}$

$[0.043, 0.094]$, indicating a small positive directional effect relative to the reported control. Clean-action recovery is -0.010 $[-0.42, 0.41]$, while the state-aliasing contrast is $+0.004$ $[-0.12, 0.11]$. The latter intervals cover appreciable effects of either sign. Table 4 groups the measurements by their target.

The targets explain how these results fit together. Donor-directed steering rewards alignment with the action associated with the injected history. Clean-action recovery rewards movement toward the recipient's own unoccluded output. The positive directional result demonstrates a controlled change in the reported action-direction measure; the recovery estimate leaves its restorative benefit uncertain under the tested occlusion. This is why evaluating a memory intervention benefits from specifying the desired action target before interpreting its displacement.

The additional injection checks support the distinction between action response and target outcome. L4 injection produces a 7–9$\times$ larger displacement than L8, and the donor-separation contrast is positive. An intervention along a fitted temporal-position direction yields a contrast of -0.040 $[-0.071, -0.009]$ against a norm-matched random direction. Appendix G details these controls and their separate units. CronusVLA contributes a directional contrast of $+0.016$ $[-0.000, 0.034]$; its rounded lower endpoint leaves that comparison uncertain. Recovery and state-aliasing evidence in this study comes from Octo-Small.

5 Implications and Conclusion

The central pattern is a separation between reconstructible frame content and the action consequences of history interventions. All three checkpoints have low residual reconstruction scores. Their occlusion responses differ in direction, and the detailed Octo-Small analysis places the largest occlusion-related increases early in the network. Octo-Small also supports a small donor-directed action shift, while its recovery estimates remain uncertain.

These findings give memory evaluation a concrete set of comparisons. Measure the prediction remaining after a current-frame predictor, retain the intervention displacement within each observation condition, and evaluate recovery toward a specified target. For a proposed memory augmentation, applying these comparisons to the same task-relevant examples would connect its predictive value to its action effects. Closed-loop evaluation could then test whether those effects improve task success. Earlier object locations and past task events are candidate targets for such an extension.

Scope and limitations. The evidence covers two policy families, two-frame observations, and open-loop outputs on BridgeData V2. RGB reconstruction measures a specific kind of visual content; its residual depends on the current-frame predictor, probe family, and fitting procedure. Local interventions measure the selected representation or attention connections, with earlier transmission to other tokens remaining possible. The recovery target is the policy's own clean action. Broader claims about task success require closed-loop evaluation.

Reproducibility Statement

Appendices A–H record checkpoints, the observation-set hash, probe configurations, sample counts, and reported results. Completing the reproducibility package requires the experiment scripts, split and run manifests, and per-example outputs. This revision retains the supplied results without re-running the experiments.

AI Use Statement

Generative AI tools were used to revise manuscript structure and wording, review the interpretation and consistency of reported results, check bibliographic metadata, and translate the manuscript into Chinese. †This statement records the verified revision workflow; the authors must complete and verify disclosure of any AI use in the original research, implementation, data processing, or figure creation before submission.

Ethics Statement

This study analyzes publicly released robot-learning checkpoints and BridgeData V2 observations through open-loop tests. Physical deployment and real-robot safety require separate evaluation. Any release of derived observations must follow the dataset's terms; the reproducibility package should identify permitted reconstruction steps.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop)*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112v1*, 2023. URL https://arxiv.org/abs/2303.08112v1.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. URL https://arxiv.org/abs/2410.24164.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025. URL https://proceedings.mlr.press/v305/black25a.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and

Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/.

Hugo Buurmeijer, Carmen Amo Alonso, Aiden Swann, and Marco Pavone. Observing and controlling features in vision-language-action models. *arXiv preprint arXiv:2603.05487*, 2026. URL https://arxiv.org/abs/2603.05487.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://www.roboticsproceedings.org/rss19/p026.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.02997.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024. URL https://www.roboticsproceedings.org/rss20/p090.html.

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.

Bear Häon, Kaylene Caswell Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models. In *Conference on Robot Learning (CoRL)*, 2025. URL https://proceedings.mlr.press/v305/haon25a.html. arXiv:2509.00328.

Chengyang He, Xu Liu, Gadiel Mark Sznaier Camps, Joseph Bruno, Guillaume Adrien Sartoretti, and Mac Schwager. Demystifying robot diffusion policies: Action memorization and a simple lookup table alternative. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=PL0tJOfm7I.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Huiwon Jang, Sihyun Yu, Heeseung Kwon, Hojin Jeon, Younggyo Seo, and Jinwoo Shin. ContextVLA: Vision-language-action model with amortized multi-frame context. *arXiv preprint arXiv:2510.04246*, 2025. URL https://arxiv.org/abs/2510.04246.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu,

Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.120. URL https://www.roboticsproceedings.org/rss20/p120.html.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025a. URL https://arxiv.org/abs/2502.19645.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R 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 *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

Hao Li, Shuai Yang, Yilun Chen, Xinyi Chen, Xiaoda Yang, Yang Tian, Hanqing Wang, Tai Wang, Dahua Lin, Feng Zhao, and Jiangmiao Pang. Towards efficient and robust manipulation via multi-frame vision-language-action modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18388–18396, 2026. doi: 10.1609/aaai.v40i22.38903. URL https://ojs.aaai.org/index.php/AAAI/article/view/38903.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. CogACT: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. URL https://arxiv.org/abs/2411.19650.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3705–3728. PMLR, 2025. URL https://proceedings.mlr.press/v270/li25c.html.

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)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Marco Molinari, Leonardo Nevali, Saharsha Navani, and Omar G. Younis. Emergent world representations in OpenVLA. *arXiv preprint arXiv:2509.24559*, 2025. URL https://arxiv.org/abs/2509.24559.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. URL https://arxiv.org/abs/2503.14734.

Open X-Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477. URL https://doi.org/10.1109/ICRA57147.2024.10611477. arXiv:2310.08864.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.012. URL https://www.roboticsproceedings.org/rss21/p012.html.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*, 2025. URL https://arxiv.org/abs/2508.19236.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025. URL https://arxiv.org/abs/2506.01844.

Ajay Sridhar, Jennifer Pan, Satvik Sharma, and Chelsea Finn. MemER: Scaling up memory for robot control via experience retrieval. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=1dH4ARGdwD. arXiv:2510.20328.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Marcel Torne, Andy Tang, Yuejiang Liu, and Chelsea Finn. Learning long-context diffusion policies via past-token prediction. *arXiv preprint arXiv:2505.09561*, 2025. URL https://arxiv.org/abs/2505.09561.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248v1*, 2023. URL https://arxiv.org/abs/2308.10248v1.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 4D-VLA: Spatiotemporal vision-language-action pretraining with cross-scene calibration. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38158–38181, 2025. doi: 10.52202/085713-1139. URL https://arxiv.org/abs/2506.22242.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023. URL https://www.roboticsproceedings.org/rss19/p016.html.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D vision-language-action generative world model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61229–61245. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhen24a.html.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=kt51kZH4aG. arXiv:2510.10274.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. URL https://arxiv.org/abs/2412.10345.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

A REPORTED NUMBERS AND ANALYSIS PROVENANCE

The original source reports several analysis arms on a fixed observation-pair set. The main Octo-Small layer sweep has raw reconstruction scores (L0–L11) $0.022, 0.278, 0.365, 0.370, 0.398, 0.371, 0.367, 0.352, 0.332, 0.325, 0.320, 0.305$. Table 2 reports its interchange profile. The validation and design-sweep scores below are retained as separate results; a manifest is needed to map their configurations, donors, and splits.

Observation set and analysis subsets. The fixed set contains 250 observation pairs and 213 unique instructions, with content hash `bb4992ac8bbc6803`. The main Octo-Small layer comparison uses 200 pairs; the image-perturbation comparison uses 160. Five Octo-Small seed runs use 60 pairs each, producing 300 pair-run observations whose distinct-pair count depends on overlap. CronusVLA's intervention-displacement check uses three runs of 80 pairs. Injection analyses report 250 observation pairs and three seeds. Subset membership and overlap require the run manifest.

Checkpoint-size comparison. In the separately reported clean L0 comparison, Octo-Small and Octo-Base have adjusted displacements of 0.339 and 0.029. Normalizing by each model's full-history or full-current swap gives Base/Small ratios of about 0.31 (Table 9). These ratios describe the two selected checkpoints and their specified normalizers.

Table 5: **Decode validation arm**, Octo-Small, $n = 200$, five-fold ridge probe. Scores differ from the main layer sweep. †The source labels the baseline $R^2(t)$; its exact predictor/target definition must be verified before treating the margin as extra predictive value for one common target.

Layer	$R^2(t-1)$	Source $R^2(t)$ baseline	Margin
0	.022	.027	$-.005$
2	.362	.347	$+.015$
4	.388	.374	$+.014$
6	.338	.342	$-.004$

Table 6: **Interchange validation arm**, Octo-Small, $n = 200$. At L4, the reported partial-mask gap is $+0.010$ $[-0.008, 0.028]$, and the full-black gap is $+0.051$ $[0.026, 0.074]$. These are a separate arm from Table 2.

Condition	L0	L2	L4	L6	L8	L11
Clean	.344	.254	.088	.051	.015	.002
Partial mask	.385	.279	.099	.043	.012	.004
Full black	.472	.367	.139	.049	.015	.000

Table 7: **Image perturbations**, Octo-Small, $n = 160$. Compare each L4 gap with the clean baseline. Only full blacking has a positive gap whose reported interval excludes zero. †The shuffle row's gap differs from subtraction of its displayed shifts, and the shuffle intervention needs definition.

Condition	C at L2	C at L4	Gap at L4	Reported 95% CI
Clean	.257	.095	–	–
Partial mask	.334	.110	$+.015$	$[-.006, .039]$
Heavy mask	.324	.114	$+.019$	$[-.003, .043]$
Blur	.310	.085	$-.010$	$[-.032, .011]$
Full black	.376	.138	$+.043$	[.016,.069]
Shuffle†	.000	.000	$-.098$	$[-.116, -.080]$

Natural-occlusion scan. The reported model-free scan ($n = 250$) finds $3/250$ candidates at the loosest threshold and zero at stricter thresholds. Frame-difference percentiles (50/90/95) are $5.08/9.99/12.52$; lost-fraction percentiles are $0.020/0.055/0.070$. Thresholds and image-processing code are required to reproduce this candidate count.

Table 8: **Five-seed Octo-Small gaps**, five runs of 60 pairs. The source pools 300 pair-run observations for 5000 sign-flip resamples. †Pair overlap, correction of the displayed zero permutation value, and the unrounded sign of L8 seed 1 must be checked. Values are preserved as reported.

Layer	Seed gaps (0–4)	Mean	SD	Positive	p
0	.142/.119/.040/.096/.141	.108	.038	5/5	.0000†
4	.037/.009/.023/.047/.054	.034	.016	5/5	.0004
6	−.001/.005/.013/−.008/−.008	.000	.008	2/5	.4882
8	.018/.000/.024/−.012/−.016	.003	.016	3/5†	.3376

Table 9: **Two-checkpoint scaling comparison**, clean L0 interchange arm. The source reports normalization by within-model full-history and full-current swap effects. Denominators and aggregation need implementation records. Ratios describe these two checkpoints and this measurement.

Model	Clean L0 C	History-normalized	Current-normalized
Octo-Small	.339	.637	.555
Octo-Base	.029	.199	.172
Base/small	.09	.31	.31

Noise calibration. At layers 0/4/6/8/11, raw clean donor-swap displacements are $0.666/0.299/0.221/0.177/0.162$, and self-swap values are $0.151/0.150/0.166/0.161/0.158$. These are calibration-arm values, not an alternative representation of Table 2.†For example, $0.299 - 0.150 = 0.149$ at L4, whereas Table 2 gives 0.087; the arm, normalization, and aggregation map must explain the difference.

B PROBE AND INTERVENTION DETAILS

Octo probes. The principal reported probe pools history-token activations by concatenating their mean and maximum and reconstructs a 16×16 RGB frame. The source specifies ridge regression with five-fold evaluation. The design sweep reports 81 configurations over layers, pooling, targets, and regularization. The best residual scores are 0.019 for Octo-Small and 0.015 for Octo-Base. These are maxima within the reported search. The shuffled-label control measures reconstruction after breaking the label assignment. Probe power and leakage depend additionally on the fitting and split procedures.

Intervention locus. The reported Octo pre-hook at `encoder_blocks.`ℓ replaces 256 history primary-camera tokens, arranged as a 16×16 patch grid. The source describes the layout as task-language tokens followed by t_0 primary, wrist, and readout tokens, then the corresponding t_1 tokens; it reports 690 tokens at width 384 for Octo-Small. The hook leaves the recipient's other token positions untouched. Earlier propagation of history into those positions remains possible.

This algorithm returns reconstruction, intervention displacement, occlusion gap, steering, and clean-action recovery as separate measurements.

Table 10: **Behavioral action-coordinate shifts**, Octo-Small, clean condition, source sample count 48×32 action samples. The listed values are coordinate-wise mean absolute changes. †Their reduction differs from the separately reported scalar summary in Table 14; the source reports a $1.22\times$ occluded/clean scalar ratio without supplying the component values here.

Intervention	wx	wy	wz	rx	ry	rz	Gripper
Mask history	.463	.294	.429	.348	.365	.376	.233
Swap history	.720	.477	.538	.409	.392	.553	.260

Algorithm 1 Separate measurements for a frozen-policy history audit

Require: Frozen policy, observation pairs, donor assignments, layers, split and RNG records

1: **for** each layer ℓ **do**
2: Fit and evaluate raw and residual probes on their recorded folds: $A_1(\ell), A_4(\ell)$
3: Compute donor-swap and self-swap shifts within each observation condition
4: Retain C_ℓ^{clean}, C_ℓ^{occ}, and their paired gap Δ_ℓ
5: Record attention-knockout effects separately where available
6: Measure donor-directed cosine contrasts at the selected injection locus
7: Measure clean-action recovery and the reported state-aliasing and temporal-position contrasts
8: Return the separate estimates, uncertainty records, and analysis-specific sample counts

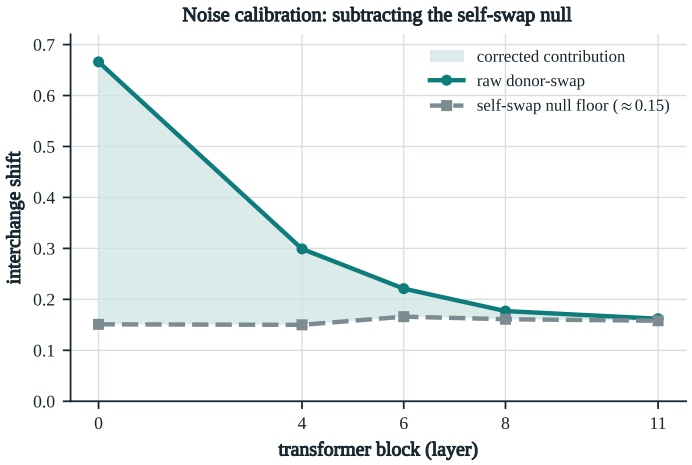


Figure 7: **Self-swap calibration.** Compare raw donor-swap displacements with self-swap variation across the displayed layers. Their difference is the reported noise adjustment. The calibration-arm values require a normalization/provenance map before direct comparison with the main intervention-displacement table.

C Environment and reproducibility records

The manuscript reports two pinned environments. The Octo path uses torch 2.6.0+cu124, transformers 4.34.1, CPU JAX 0.4.20 for weight loading, tensorflow-cpu 2.15.0, and numpy 1.26.4, with a PyTorch inference port. The CronusVLA path uses torch 2.2.0+cu121, transformers 4.40.1, numpy 1.26.4, tokenizers 0.19.1, timm 0.9.10, and `draccus`; a scikit-learn pin is mentioned but not specified. These are reported settings, not supplied environment files.

Reported checkpoints are Octo-Small/Base 1.5 and the CronusVLA-0.5B Bridge/RT-1 post-trained checkpoint `step-042500-epoch-07-loss=0.0587.pt`. The source reports a 24GB GPU, no quantization, and a 387GB Bridge corpus subsampled to the hashed pool. Policy weights are frozen; ridge, preprocessing, and temporal-position estimators involve data fitting. Code commits, downloadable checkpoint identifiers, environment specifications, and an observation-set builder script must accompany a reproducible release.

D Model selection and coverage

The study selects two open policy families with history inputs compatible with the available RGB observation pairs. The two Octo sizes test recurrence within a family; CronusVLA supplies a different history representation and action-conditioning path.

The original survey considered OpenVLA-OFT, π_0, $\pi_{0.5}$, X-VLA, ACT, and 3D-/4D-VLA (Kim et al., 2025a; Black et al., 2024; 2025; Zheng et al., 2026; Zhao et al., 2023; Zhen et al., 2024; Zhang et al., 2025). For that survey, multi-view inputs, default single-frame interfaces, checkpoint availability, and missing geometric inputs were practical selection concerns. ACT is CVAE-based.

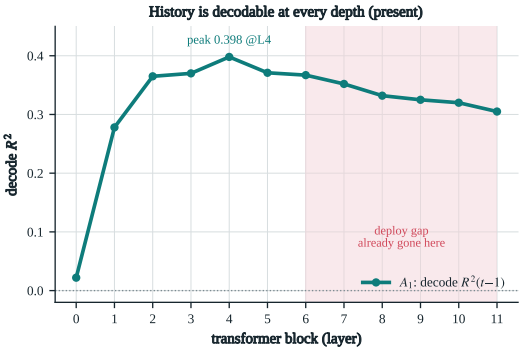
4D-VLA uses sequential RGB-D inputs; the survey's exact checkpoint and input requirements for every excluded model remain to be documented. An unnamed predictive-model candidate in the original notes cannot be independently identified. The survey records the practical choices behind the evaluated set.

E CronusVLA probe standardization

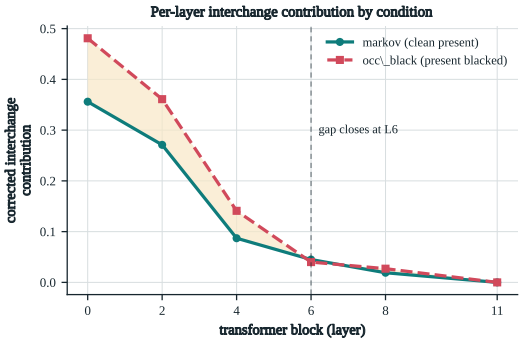
The source identifies CronusVLA history as an 896-dimensional per-frame cognition feature, taken from the VLM's last-layer last token. An initial probe produced negative scores, reported as real-label -0.004 versus shuffled-label -0.463. The revised configuration z-scores features, searches regularization 10^2–10^5, and includes a PCA-50 variant. The reported scores are $A_1 = 0.142$, $A_4 = -0.011$, and shuffled-label -0.032, with similar conclusions for full and PCA-reduced representations. Numerical scaling is a plausible contributor to the initial behavior, but the source does not establish its cause. Preprocessing must be fitted within training folds; the code is needed to verify that requirement.

F Supporting layer and observation-condition panels

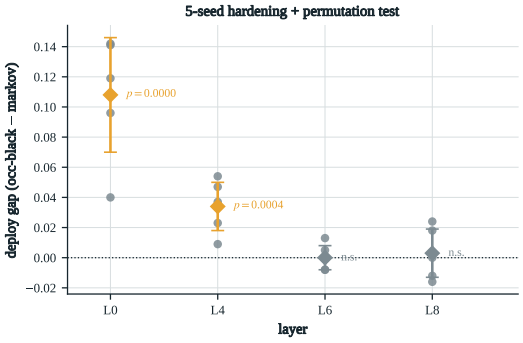
The panels below retain all original data artwork. Their captions identify the quantities to compare and mark legacy interpretations that must be revised in the editable artwork before submission. Interchange was sampled at layers $\{0, 2, 4, 6, 8, 11\}$, whereas the decode curve covers all 12 Octo-Small layers.



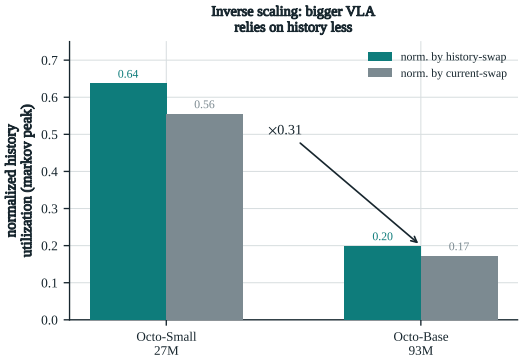
(a) Raw reconstruction by layer.



(b) Condition-specific displacements and their gap.



(c) Five-seed gap summaries.



(d) Two Octo checkpoint sizes.

Figure 8: **Depth and checkpoint comparisons.** (a) Raw frame reconstruction remains positive through L11. (b) Inspect each condition as well as their difference: a small gap can coexist with nonzero shifts. (c) Compare seed-level signs; the pooled permutation implementation requires verification. (d) The normalized clean L0 displacement ratio is about 0.31 across these two checkpoints, not a general scaling law. †Original "cutoff," zero-p, and deployment labels require correction.

G INJECTION CHECKS AND RECOVERY TARGETS

Table 4 separates intervention displacement, donor-directed steering, and clean-action recovery. The source reports Octo-Small checks on 250 pairs with three seeds. CronusVLA is described both as using the same fixed pool and as using a held-out 250-pair set; the membership must be resolved before claiming either shared observation pairs or independent replication.

Donor information and direction. Octo-Small L4 injection has a 7–9$\times$ larger action effect than L8. The separation-minus-null estimate is $+11.0$ $[9.6, 12.5]$, and the directional contrast is $+0.069$ $[0.043, 0.094]$. These contrasts quantify the dependence of action displacement on the donor and its direction. The separation metric's units, donor matching, and null construction are missing from the source package.

Clean-action recovery. Injection supplies the recipient's clean-run history representation while its current image is blacked out. Equation 8 measures movement toward that policy's unoccluded action. Recovery is -0.010 $[-0.42, 0.41]$, while an own-minus-unrelated contrast is $+4.9$ $[3.1, 6.9]$ in separately reported units. The own-minus-unrelated contrast compares two interventions, whereas clean-action recovery compares the own-history intervention with the occluded baseline. Their scaling must be documented before comparing magnitudes.

State aliasing and temporal position. The reported aliasing procedure uses a identical or ambiguous current image with different histories and a paired contrast intended to remove common drift. Its estimate is $+0.004$ $[-0.12, 0.11]$. Exact pair construction, branch targets, and pairing are required for replication. The temporal-position check pushes current tokens along a fitted temporal-position direction and compares with a norm-matched random push. Its classifier accuracy is reported as 1.0, and the intervention contrast is -0.040 $[-0.071, -0.009]$. The classification split and whether its target encodes position rather than useful past content need to be specified.

CronusVLA. Intended and unrelated-direction cosine summaries are reported as 0.583 and 0.566, with paired contrast 0.016 $[-0.000, 0.034]$. The marginal subtraction 0.017 can differ at the last decimal from a rounded paired result; unrounded paired outputs are required to verify this and the lower interval endpoint. The source descriptions disagree on which part of the donor control is changed. This result is treated as inconclusive. The CronusVLA injection analysis covers directional steering.

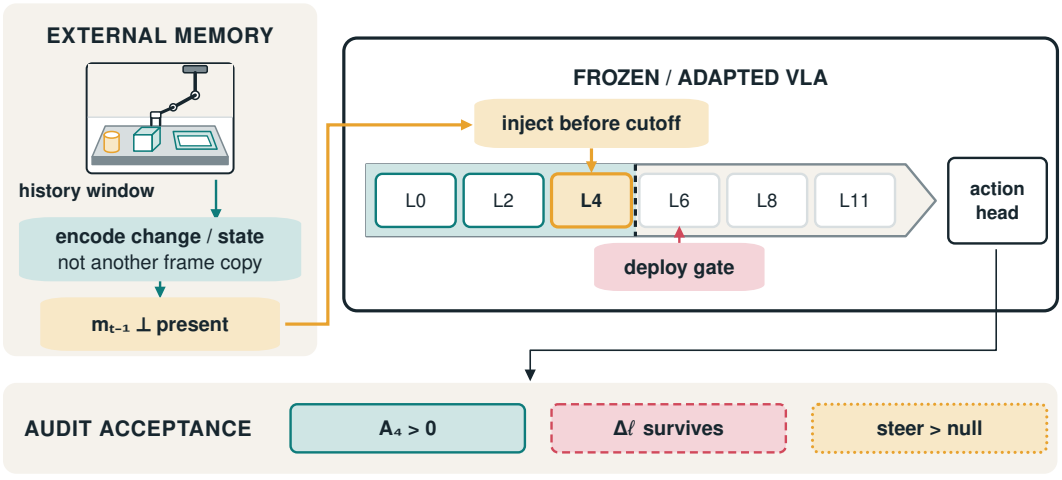


Figure 9: **Original memory-audit schematic retained for revision.**†Its pass/fail criteria are not conclusions of the revised paper. A zero occlusion gap can coexist with useful history effects in both conditions, and a positive residual RGB score is not necessary for every useful memory system. The revised evaluation separates prediction, intervention sensitivity, and task-relevant recovery without imposing these universal gates.

H NOTATION, PROBE SWEEPS, AND REPORTED MODULES

Tables 12 and 13 list the top eight residual-probe configurations for each Octo model. Table 14 preserves a separately reported behavioral summary. Table 15 maps source module names to reported outputs; these names do not establish that code is included. Table 16 retains the CronusVLA seed values and reported intervals.

Table 11: **Terminology, notation, and retained source identifiers.**

Term or symbol	Meaning
History representation	Activations associated with the previous observation
Donor / recipient	Pair providing / receiving the replacement representation
Interchange	Replacement of a selected history representation
Self-swap	Replacement with the recipient's own representation
Readout tokens	Tokens that collect information for the action head
Raw reconstruction	Prediction of the previous RGB frame
Residual reconstruction	Prediction of the error left by a current-frame predictor
Clean-action recovery	Movement toward the recipient's clean-image policy output
$t, t-1$	Current and previous observation positions
ℓ	Intervention/probe layer index
A_1	Raw reconstruction: previous-frame R^2
A_4	Residual reconstruction: R^2 of the predictor error
A_2	Source name for shuffled-label probe control
C_ℓ	Self-swap-adjusted intervention displacement in reported units
Δ_ℓ	Occluded minus clean intervention displacement
$\bar{S}$	Paired intended-donor versus control cosine contrast
r_{repair}	Reduction in distance to the policy's clean action
`markov`	Source identifier for clean current image
`occ_black`	Fully blacked current image
`shuffle`	Source temporal perturbation; exact operation pending

Table 12: **Octo-Small design sweep**, top eight configurations ranked by residual A_4 among 81 reported configurations. Best listed $A_4 = 0.019$; the source's best raw score 0.389 is from meanmax/RGB16/L2/$\alpha = 10$, outside this A_4-ranked view.

Layer	Pool	Target	α	A_1	A_4
L2	meanmax	rgb16	100	.248	+.019
L4	meanmax	rgb16	100	.250	+.017
L6	meanmax	rgb16	100	.223	+.014
L2	max	rgb16	100	.179	+.012
L4	max	rgb16	100	.178	+.012
L4	mean	rgb16	100	.174	+.009
L2	mean	rgb16	100	.165	+.009
L6	max	rgb16	100	.155	+.008

I EXTENDED RELATED WORK

Representations. Dictionary learning and sparse autoencoders can expose features within learned representations (Bricken et al., 2023; Huben et al., 2024; Templeton et al., 2024). Linear representation geometry (Park et al., 2024) and intermediate prediction lenses (Belrose et al., 2023) motivate further ways to inspect the content of activations. These tools offer possible extensions to the present measurements. Recovering task events rather than RGB pixels would change the target of the residual comparison and requires a new evaluation.

Table 13: **Octo-Base design sweep**, top eight configurations ranked by A_4. Best listed residual score is 0.015; the separately reported best raw score is 0.362.

Layer	Pool	Target	α	A_1	A_4
L4	meanmax	rgb16	100	.285	+.015
L4	max	rgb16	100	.243	+.014
L6	meanmax	rgb16	100	.272	+.011
L4	mean	rgb16	100	.202	+.008
L6	max	rgb16	100	.233	+.007
L6	mean	rgb16	100	.189	+.004
L2	mean	rgb16	100	.200	+.003
L2	max	rgb16	100	.254	+.003

Table 14: **Separately reported scalar action displacement**, Octo-Small. †The source labels this a mean absolute change per action coordinate over the fixed observation-pair set, but does not reconcile it with Table 10; neither summary has been substituted for the other.

Intervention	Mean absolute change	SD
Mask history	.425	.20
Swap history	.527	.27

Table 15: **Reported module-to-result map.** These module names come from the manuscript; their implementations are absent from the supplied package. This table is a checklist for the experiment release, not evidence of reproduction.

Analysis	Source module	Reported output
Observation pairs	`build_stimuli`	250 pairs; hash `bb4992ac8bbc6803`
Decode	`decode`	Octo-Small sweep $A_1 = .389$, $A_4 = .019$
Interchange	`deploy_gate`	Main L4 gap $+.054$ [.028,.079]
Image perturbations	`deploy_gate`	Full-black L4 gap $+.043$
Knockout	`knockout`	Positive reported late-layer gaps
Seed check	`stats`	L4 $p = .0004$; positive in 5/5 seeds
Size	`cross_scale_octobase`	$A_4 = .015$; normalized ratio .31
Family	`cronus_decode/deploy`	$A_1 = .142$, $A_4 = -.011$; clean .0498, occluded .0357
Image scan	`natural_occlusion`	3/250 candidates

Table 16: **CronusVLA intervention displacement across three seeds**, 80 pairs per run. Each seed decreases under occlusion. †The reported 95% intervals equal the ranges of the displayed seed means; resampling units and interval construction must be verified. Shuffle interpretation requires its actual intervention definition.

Seed	Clean	Shuffle	Occluded
0	+.0479	-.0000	+.0354
1	+.0482	+.0004	+.0358
2	+.0534	-.0010	+.0359
Mean ± SD	$.0498 \pm .0025$	$-.0002 \pm .0006$	$.0357 \pm .0002$
Reported 95% CI†	[.0479,.0534]	–	[.0354,.0359]

Policy landscape. RT-2, OpenVLA, and FAST illustrate action-token modeling and tokenization (Zitkovich et al., 2023; Kim et al., 2025b; Pertsch et al., 2025); diffusion and flow approaches include Diffusion Policy and the π models (Chi et al., 2023; Black et al., 2024; 2025). CogACT, OpenVLA-OFT, SmolVLA, and GR00T provide other action-modeling or adaptation designs (Li et al., 2024; Kim et al., 2025a; Shukor et al., 2025; NVIDIA et al., 2025). Open X-Embodiment and DROID expand robot-learning data resources (Open X-Embodiment Collaboration, 2024; Khazatsky et al., 2024). These models and datasets provide context for the three evaluated checkpoints and the BridgeData V2 observation set.

Table 17: **Related-work comparison, revised as study focus.** The original checkmark table is replaced by explicit descriptions of what the cited work evaluates; layer-resolved steering is acknowledged.

Work	Relevant focus
Grant et al.	Visual/language pathways and action control across VLA models
Häon et al.	Internal feature interpretation and activation steering
Buurmeijer et al.	Feature observability and controllability in VLAs
Memory-oriented VLAs	Temporal context mechanisms and manipulation performance
This study	Residual frame prediction, occlusion gaps, and selected injection outcomes

J Reported harness checks

The source reports an NHWC-to-NCHW pipeline correction, a backbone-isolating hook map, and readout-parity mean absolute difference 2.4×10^{-4} against a reference forward pass. It also mentions deterministic parity tightening, full/mask/swap checks, and reuse of a self-swap-adjusted interchange routine. CronusVLA checks are described as analogous. The package lacks scripts, commits, and parity outputs, so these are reported checks rather than independent certification. Each production run should identify its hook mapping and reference checkpoint.

K Interpretive scope

Probe sensitivity. Residual reconstruction depends on the target, features, sample size, and decoder search. Raw reconstruction and shuffled-label controls address different comparisons. Nested fitting and trajectory-level splits are needed to assess leakage and selection effects; nonlinear content requires a different decoder evaluation.

Intervention localization. The measurements concern selected history tokens or attention connections. History can propagate to other tokens before those locations. Self-swap adjustment depends on sampling and pairing. The positive late knockout gaps support a graded depth profile alongside the smaller late interchange gaps.

Transfer and generalization. Steering measures action direction; recovery measures distance to the policy's clean action. Task-success claims require closed-loop rollouts. The observation set and two families define the study's coverage. Across families, architecture, training, and measurement choices vary together.