mmCLAW: SEPARATING SPATIAL READABILITY FROM ACTION-HEAD GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Imitation policies can expose readable spatial coordinates without using the corresponding probe directions to act. We study this distinction in mmClaw, an auditable, path-level radar testbed. Two teachers label identical requests using a public estimate or a return centroid; paired students differ only in supervision. Frozen policies and probes preserve a positive teacher-by-variable interaction on 48 new layouts: $D = 0.156\,[0.125, 0.236]$. We then decompose native evidence responses through the final action head. Coordinate-related editing can mix originally output-null changes into the selected response: covariance operators produce signed null-component projections of 0.207 and 0.357 for the two teachers. LEACE exhibits stronger fit-dependent cancellation, while coordinate increments become small after teacher-direction erasure. Nonlinear action-and-pose references and independently varied eraser/probe supports further distinguish task-label predictability from stable semantic use. These results support a supervision effect on relative readability, but not a unique coordinate-mediated action mechanism. We provide an output-consistency check, retained finite-action components, and explicit failure controls for interpreting representation edits.

1 INTRODUCTION

Learning to act from partial observations is central to active perception and embodied decision making (Bajcsy, 1988; Kaelbling et al., 1998). Radar-guided search makes this dependence explicit because returns are noisy and reflected. A policy may summarize measurements into a location estimate, keep detailed spatial evidence, or map observations directly to movement. If a coordinate can be recovered from the hidden state, it is accessible. That alone does not show that the coordinate drives the policy's actions.

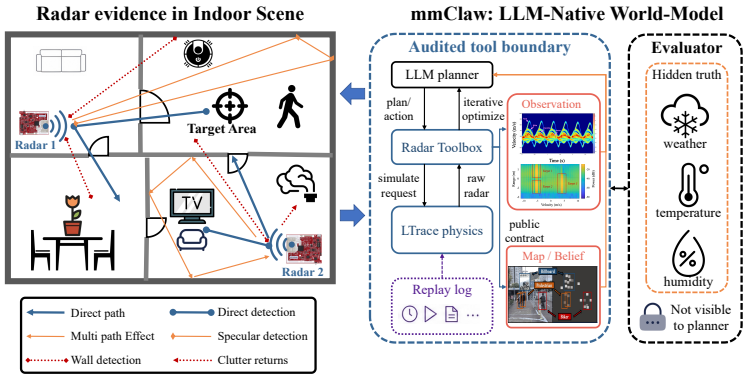


Figure 1: The mmClaw sensing boundary. The policy receives public measurements and reconstructed state, while hidden scene variables stay with the evaluator. The architecture supports several clients; this study evaluates the learning configuration in Section 3.

Existing tools answer different questions. Probes measure which variables a representation exposes, and control tasks separate accessibility from readout properties (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022). Concept erasure and activation interventions examine behavior after

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.

hidden components are removed or replaced (Ravfogel et al., 2020; Elazar et al., 2021; Belrose et al., 2023), and counterfactual alignment tests whether distributed activations support an interpretable intervention (Geiger et al., 2024). However, probes can exploit action correlations, and edits can create pathways that the unedited computation lacks (Makelov et al., 2024), so neither alone identifies a natural spatial mechanism. This matters for sensing policies, where movement direction, current pose and a candidate coordinate are geometrically related. A policy trained to approach a coordinate may make it readable without computing it in a reusable way. An edit may also mix hidden changes that the action head ignores with changes that already affect its output, which credits the intervention's own effect to the policy. A convincing audit must therefore connect supervision, genuine evidence changes and frozen-policy geometry, and test whether its conclusions hold beyond the samples used to fit it.

We introduce mmClaw, an auditable path-level radar testbed and a controlled study of spatial readability in imitation policies. We address three challenges. *First, isolating supervision:* how can we compare spatial objectives while keeping the training requests fixed? *Second, interpreting interventions:* how can we tell output responses carried by the original hidden change apart from influence created by an edit? *Third, checking reliability:* how can we tell a fitted erasure guarantee apart from stability across independent eraser supports, readout supports and unseen layouts? This design tests spatial representations without presupposing a world model (Li et al., 2023; Vafa et al., 2024).

Our key observation is that evidence changes and representation edits expose different dependencies. Changing the measured returns produces a native hidden response. Splitting that response with a coordinate-related operator adds a separate dependency on the operator. Because the action head is linear, we can track both operations exactly before movement decoding. This shows that a hidden component with no direct output effect can gain one under the edit. The result is an output-consistency check for whether the selected response depends only on the original action change. Passing this check does not establish semantic use.

Specifically, two teachers label the same requests using either a public location estimate or a weighted recent-return centroid. Paired Set–CNN students share initialization and minibatch order, and neither coordinate is a model input. With policies and readouts frozen, we keep a controlled fraction of acquired returns at fixed acquisition poses and rebuild each public request. The primary condition keeps 50%, with 25% and 75% used to test sensitivity. Neither teachers nor students receive hidden target truth (Figure 1). We keep selected, complementary, output-row-space and output-null responses, including cancellation. We test competing explanations with direction erasure, geometric-strength controls, nonlinear action-and-pose readouts and independent support resampling. Readouts and edits are fitted on training data and selected on validation data, and finite movements use the original normalization and boundary decoder. Annotated simulation supports controlled comparison, not physical radar validation.

We evaluate forty policies on 48 newly generated layouts, with five sensing-noise seeds and all 1,200 scored requests. The primary policies and probes were fixed before these layouts were generated. The layouts test generalization within the same simulator family and teacher-acquisition distribution. Supervision changes relative coordinate readability, with a pooled teacher-by-variable contrast of $D = 0.156\,[0.125, 0.236]$. Yet covariance-selected edits yield signed output-null projections of 0.207 and 0.357, showing that the edits create their own influence. Coordinate increments shrink after direction erasure, action-and-pose references explain much of own-coordinate readability, and both fitting supports affect held-out erasure. So supervision affects accessibility, but a unique coordinate-mediated action mechanism remains unresolved. Because imitation errors differ, this is a total training effect and cannot be separated from fitting quality. The retained pilot also fails its evidence-necessity criterion, which limits claims about adaptive closed-loop sensing.

Our contributions are:

- A matched-supervision study with native evidence interventions and frozen new-layout evaluation that separates readable public statistics from claims about hidden target state.
- An action-head audit and output-consistency criterion that expose edit-induced output-null mixing and cancellation, and limit how far selected responses can be read as semantic contributions.
- Direction, action-reference, fitting-support and task-validity controls that make the scope of spatial representation claims explicit and reproducible.

2 An Auditable Sensing Testbed

The testbed is a controlled source of public evidence, not the learned representation under investigation. LTrace denotes the bundled custom ray/path backend, rather than a commonly adopted external radar benchmark. Before analyzing students, we regenerated its default and stress fixtures from the supplied implementation. Figure 2 checks error distributions and return composition; Table 1 reproduces the legacy stress table. These measurements establish what this backend generates, not sensor realism or a learning advantage.

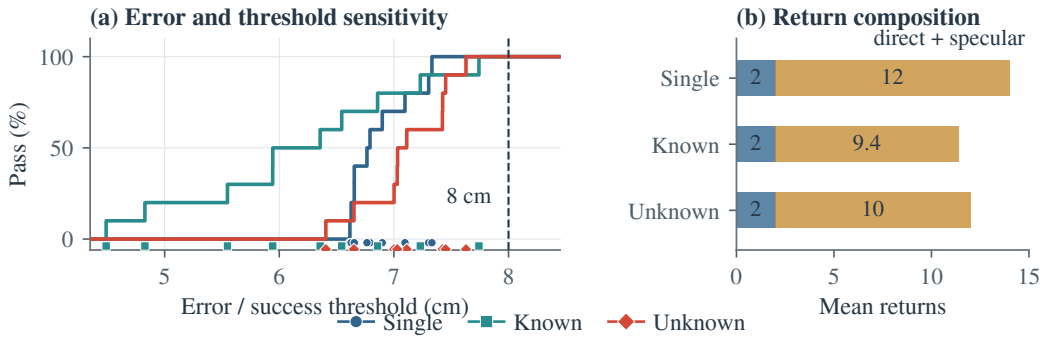


Figure 2: Backend calibration from 30 episodes, 10 seeds per mode. (a) Empirical pass curves and individual errors (rug marks), all below 8 cm. (b) Mean direct (blue) and specular (gold) returns. These heuristic-policy fixtures contain no clutter.

Table 1: Regenerated legacy stress fixtures. Ten seeds per column; error is mean ± population SD. The complete table matches the original displayed precision.

	Adverse	Rain+occ.	Fog+occ.	Three radars
Radars / steps	1/2	1/3	2/5	3/5
Error (cm)	6.87±0.32	197.57±3.57	7.11±0.37	6.91±0.24
Returns	17.8	15.2	26.3	51.0
Clutter returns	3.8	7.1	9.3	0.0
Mean SNR (dB)	52.65	26.68	58.41	90.92
Mean precision (cm)	42.16	82.88	58.91	4.80
Success (%)	100	0	100	100

The rain–occlusion failure is retained. High simulated SNR or a favorable default threshold does not certify evidence-dependent behavior: both the trajectory and the separate scoring estimator determine success. Consequently, our learning questions concern paired requests and hidden-state responses, with endpoint success used only to audit task validity.

3 Matched Supervision and Native Evidence Changes

A request contains current pose and bounds, annotated scan history, a two-channel count map, and a public estimate $\hat{x}$. Estimate moves toward $\hat{x}$, computed from clustered returns with preference for direct paths. Evidence moves toward a weighted recent-return centroid c. Both use the same map-frontier fallback when needed. Source and path annotations are authorized simulator outputs, a privilege that limits transfer to physical sensing.

Set–CNN combines pooled return features, a map convolution and a pose branch into 80 coordinates, followed by two 64-unit tanh layers and a linear two-vector head. Neither public coordinate is an explicit input. Paired teachers relabel the same 3,072 training requests from 48 layouts and 384 validation requests from 12 other layouts. Ten matched seeds per teacher share initialization and minibatch order; ten additional seeds per teacher provide independent training-variation controls. Validation selects checkpoints and probes (Appendix A).

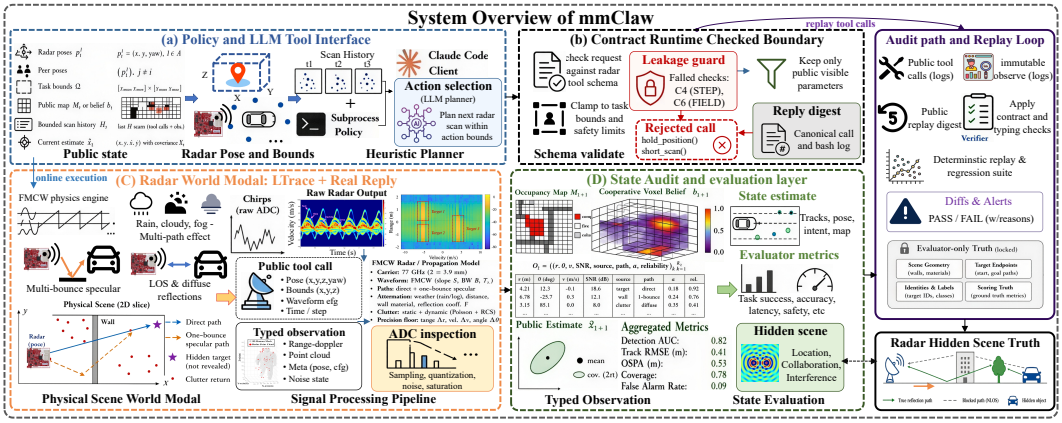


Figure 3: System overview of the sensing testbed. Learned policies occupy the policy-client slot; request validation, public state reconstruction and evaluator separation support the experiments. Language clients, ADC inspection and miniature metric displays show optional architecture context.

Each episode has six scans and five scored moves. The decoder normalizes a raw action to a 0.65 m step and clamps it inside room bounds. For a native intervention, we retain 25%, 50% or 75% of each scan's returns along the recorded acquisition path, then rebuild history, map, estimate and metadata before a forward pass. Acquisition poses stay fixed. The 50% condition is primary; no new rollout or physical measurement is implied by a frozen-request action change.

4 Supervision Changes Relative Spatial Readability

Changing teacher labels changes the relative accessibility of two correlated public statistics. It does not make either statistic exclusive to its teacher. Let $R_{T,L}$ denote held-out coordinate R^2 for teacher $T \in \{E, C\}$ and decoded label $L \in \{E, C\}$. We report the paired contrast

$$D = (R_{E,E} - R_{E,C}) - (R_{C,E} - R_{C,C}). \tag{1}$$

All coordinate readouts use training-only fitting and validation-only selection. Negative R^2 values remain visible; the denominator is recomputed for each resample.

The original pooled contrast is $0.189\,[0.136, 0.253]$, whereas independent same-teacher cohort differences are close to zero. On the new layouts, pooled linear $D = 0.156\,[0.125, 0.236]$; the primary cohort's frozen nonlinear probes give $0.103\,[0.051, 0.199]$. This supports a total effect of the supervision treatment beyond a particular initialization or the original request cohort.

The new cohort comprises layouts 14000–14047, five sensing-noise seeds and all 1,200 scored requests. The plan and existing weights/readouts were fixed before these layouts were generated. It is distinct from the prior 48-layout cohort and from the reused 24-layout development set. Its trajectories still come from Full Estimate acquisition, so this is a same-family generalization test, not policy-independent state coverage.

Fitting quality remains a competing explanation for how supervision acts. Development action errors span 4.22–6.53 cm for Estimate and 1.87–2.86 cm for Evidence; validation matching retains only five pairs. Adjusting a post-training error does not identify a quality-independent direct effect. We therefore retain the supervision total effect and test a more specific action-encoding explanation in Section 8.

Unless stated otherwise, intervals use 10,000 crossed fit/layout/noise bootstrap draws with seed 20270917, preserving teacher and request pairing. Independent training cohorts have independent fit resamples. Ratios, centered coordinate sums of squares and conditional denominators are recomputed within each draw.

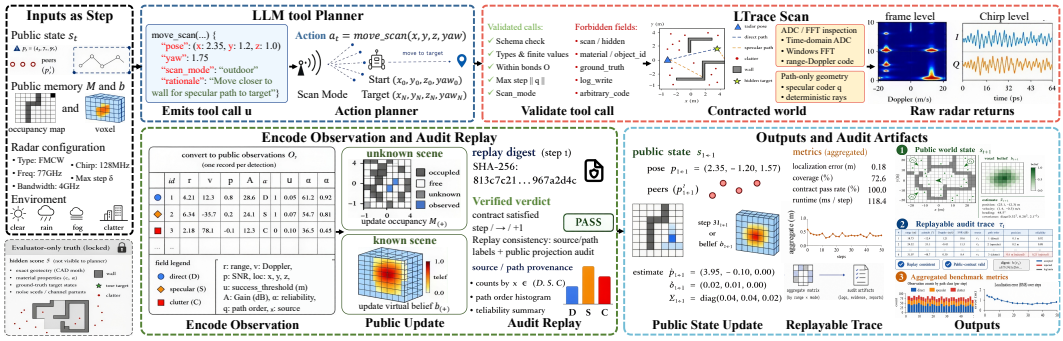


Figure 4: Execution pipeline. The evaluated path-only branch scans, rebuilds public state, applies an intervention, and validates a policy proposal before movement. Target truth is excluded from policy inputs and teacher actions; waveform settings and example tool rationales are outside the scope.

5 When an Edit Creates Output Influence

An activation change from a genuine input intervention need not remain a natural semantic pathway after it is split by an edit. Write the frozen action head as $a = Ah + b$, and let Q denote the selected update operator. Native thinning yields $\delta h = h(r_{\text{thin}}) - h(r)$, so

$$\delta a = AQ\delta h + A(I - Q)\delta h, \qquad f(v) = \frac{\sum_i v_i^\top \delta a_i}{\sum_i \|\delta a_i\|^2}. \tag{2}$$

This exact identity permits negative or greater-than-one signed projections under cancellation. It does not identify a unique causal allocation.

Let $R = A^+A$ and $N = I - R$ in the native hidden coordinates. Although $AN\delta h = 0$, the selected response obeys

$$AQ\delta h = \underbrace{AQR\delta h}_{v_{\text{row}}} + \underbrace{AQN\delta h}_{v_{\text{null}}}. \tag{3}$$

The second term measures output influence introduced by mixing an originally output-null component through Q.

Output-consistency check. A selected response is determined solely by $A\delta h$ if and only if $AQN = 0$, equivalently $AQ = TA$ for some output-space map T. Indeed, equal-output hidden changes differ by an element of $\ker A$; the condition makes their selected outputs equal. Conversely, equality for every such pair forces $AQ \ker A = 0$. This elementary criterion detects one failure mode of subspace interpretation, related to disconnected-pathway illusions in activation patching (Makelov et al., 2024); passing it alone does not establish a semantic mechanism.

Covariance operators give null-component projections of $0.207\,[0.161, 0.266]$ for Set–E and $0.357\,[0.303, 0.410]$ for Set–C. LEACE's corresponding means are positive but have wide intervals spanning zero because components cancel across fits. Thus the clearest replicated mixing evidence comes from covariance projection, while LEACE exposes instability in assigning the selected effect to an output-visible path.

6 Direction Controls Change the Interpretation

Coordinate-related edits overlap strongly with the supervised action function. We fit LEACE (Belrose et al., 2023) to the actual teacher unit-direction loss labels, then fit a coordinate eraser on the surviving training activations. With erasers P_d and $P_{z|d}$, the incremental coordinate update is

$$Q_{z|d} = (I - P_{z|d})P_d. \tag{4}$$

We also reverse the order. All maps are training-fitted and frozen before new-layout evaluation. Sequential effects are order-dependent diagnostics, not an identified direct effect after conditioning on a mediator.

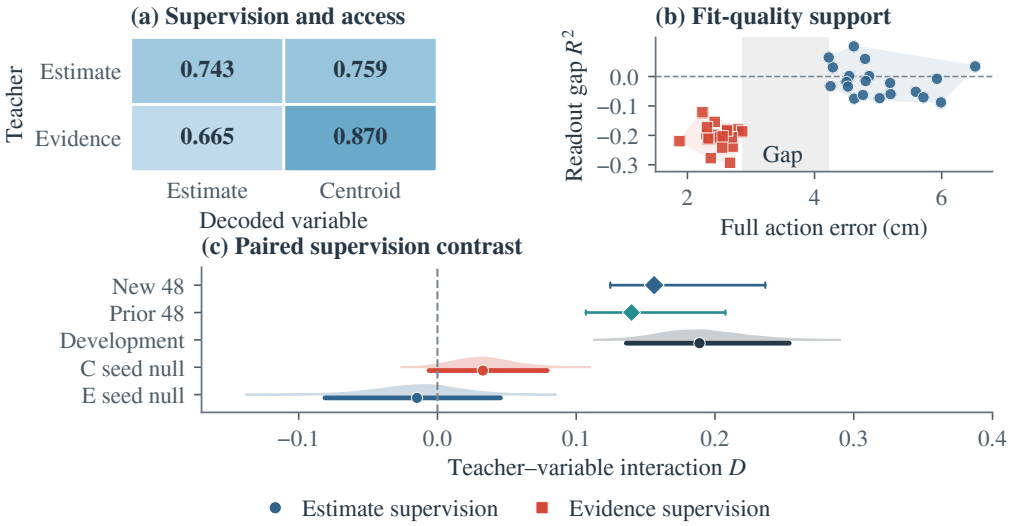


Figure 5: Relative spatial readability. (a,b) Forty fitted policies on the development cohort. (c) Same-teacher seed controls and frozen evaluation on development, prior and new layouts. Intervals jointly resample fits, layouts and noise; all cases and the fit-quality gap are retained.

After direction erasure, coordinate increments have bounded signed projections $-0.006\,[-0.013, -0.001]$ for Set–E and $-0.001\,[-0.004, 0.001]$ for Set–C. Reversing the order leaves larger direction increments. This weakens a coordinate-specific reading of the unconditioned edit. It does not exclude an upstream coordinate computation whose behavioral effect is mediated by direction, nor prove that sequential erasure stays on the natural activation distribution.

Two controls separate algebra from interpretation. Erasing the head's own raw output recovers a selected response of one, as expected for this LEACE control on covariance support; it is a positive implementation check, not evidence for semantic specificity. Random orthogonal conjugates preserve the complete operator singular spectrum and can produce large positive or negative projections. They expose orientation-dependent amplification without pretending to match every input-dependent displacement. Equal-rank orthogonal and per-request norm-matched perturbations remain separate controls in Appendix C.

Finite actions. Every component passes through the original normalization and bounds decoder F_r. We retain $d = F_r(a + \delta a) - F_r(a)$, selected/complement responses and their nonadditive residual. We additionally retain the row-plus-null residual relative to the selected response. Raw vector additivity never licenses adding movement magnitudes. On the original cohort, relative-direction ridge projection differs sharply between MLP and Set–CNN; the corresponding action-head row-energy fractions, approximately 0.850 versus $0.117/0.130$, provide a direct geometric explanation for this architecture-dependent diagnostic.

7 Erasure Reliability Depends on Both Fitting Supports

Empirical covariance erasure is exact on its fitted sample, but independent-layout readouts can remain unstable. For centered training covariance $C = \mathrm{Cov}(h, z)$, a rank-two covariance projection $P_C = I - UU^\top$ satisfies $P_C C = 0$. LEACE enforces the same constraint with a minimum-distortion oblique map. Neither property guarantees disappearance under a separately fitted held-out readout.

The original forward/reverse audit changed both eraser and probe supports: Set–E's erased R^2 changed from -0.499 to 0.523, whereas identity readouts were similar. To localize this sensitivity, we now vary one support at a time. Each arm uses sixteen disjoint layouts for the eraser and sixteen for the probe, with twenty recorded resamples. The fixed side retains its layouts and fitting rule; probe coefficients are refitted for the current transformed activations.

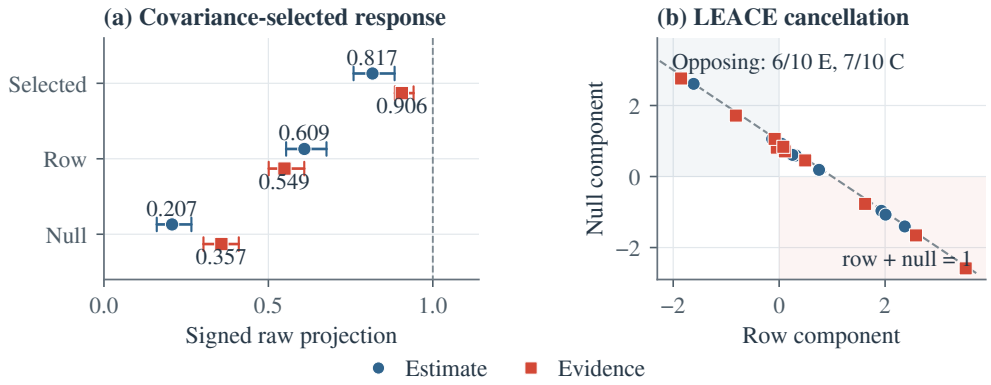


Figure 6: Action-head audit on new layouts at 50% retention. (a) Covariance-selected, row and output-null responses: means and joint intervals. (b) All ten LEACE fits per teacher; shaded quadrants mark opposing components, and the dashed line sums to one. Cancellation precludes an additive semantic interpretation.

Both supports matter. With a fixed eraser, its held-out covariance and orientation stay fixed while readout scores change. Varying eraser support also changes its row-space orientation: median largest angles to the reference are about $57°$ and $46°$ in the two groups. The new sixteen-layout protocol differs from the original half-split, so it diagnoses separate sources of sensitivity rather than assigning a variance percentage to the earlier sign reversal.

Negative R^2 means prediction is worse than a test-mean constant, not that information is absent. We retain squared mean prediction bias, centered residual energy, held-out covariance, conditioning diagnostics and perturbation norms. Their loss decomposition separates mean bias from residual error without calling the latter irreducible noise. These checks support reporting erasure guarantees, operator stability and readout generalization as three different quantities.

8 What Is Readable Beyond Action and Pose?

Nonlinear action-and-pose references explain much of each teacher's own-coordinate readability, but not all cross-statistic information in H2. We compare three readout inputs: teacher plus student unit directions, those directions with pose and room bounds, and H2. Each uses 256 training-landmark Gaussian features and identical validation-selected bandwidth/ridge grids. This matches the fitting budget and selection rule, not every hypothesis class. Including the teacher label makes the action reference a diagnostic, not a deployable student-only baseline.

For the own coordinate, the H2 advantage over action and pose is $-0.059\,[-0.150, -0.028]$ under Estimate supervision and $0.018\,[-0.025, 0.050]$ under Evidence supervision. By contrast, the cross-coordinate advantages are $0.138\,[0.086, 0.276]$ and $0.200\,[0.067, 0.330]$. Thus the students are not simply coordinate-exclusive encoders, yet own-coordinate decodability alone poorly distinguishes a spatial representation from supervised action geometry. This reviewer-motivated readout comparison was specified after other new-layout diagnostic.

Hidden target state. New-layout landmark H2 target readouts remain weak, and their improvements over the action-and-pose reference are unresolved. In the development cohort, the previous two-layer nonlinear probes likewise showed no resolved H2 advantage over raw summaries. A larger complete-history probe achieved $R^2 = 0.810$, but had richer inputs and greater capacity. It establishes an achievable predictive reference, not a capacity-matched information-loss claim or an information-theoretic ceiling. Acquisition-step differences are retained in Appendix G.

Counterfactual alignment. Earlier-layer distributed interchange search (DAS) (Geiger et al., 2024) improves alignment on its recorded donor distribution. Set–E encoder cosine rises from 0.911 to 0.987, and the 30–$60°$ stratum also improves. However, only ten distinct requests support its 60–$90°$ stratum; larger turns are absent, and Set–C has no requests above $60°$. These results estab-

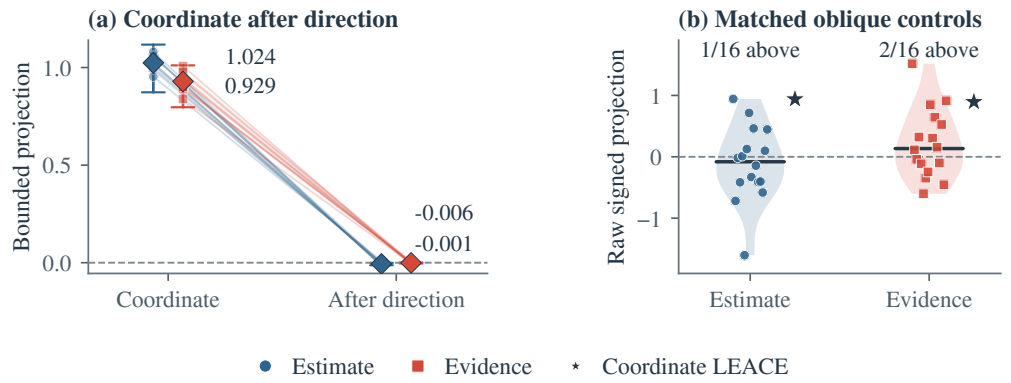


Figure 7: Direction and geometric-strength controls on new layouts. (a) All paired fits, means and joint intervals before/after direction erasure. (b) Sixteen fixed random conjugates per teacher; stars mark coordinate LEACE and counts denote controls above it. Operator spectra match; per-request displacement norms need not.

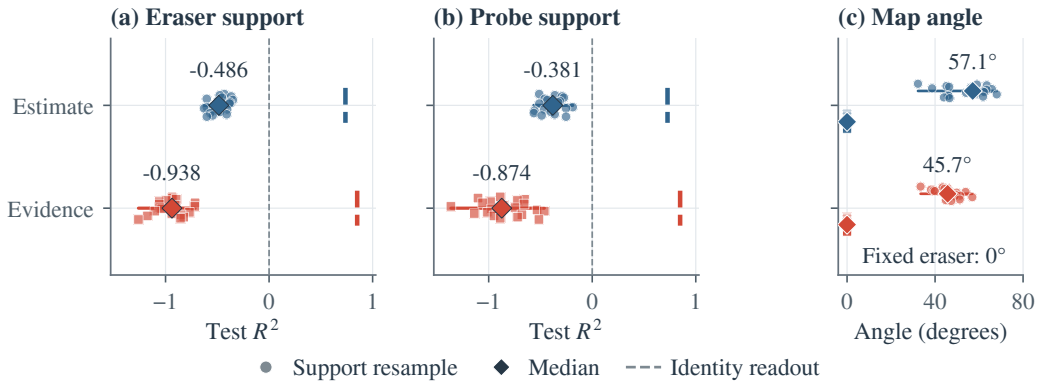


Figure 8: Independent support sensitivity on the development cohort. (a,b) Twenty resample means per arm and teacher, each over ten fits; diamonds show medians, lines full ranges, and dashed marks identity readouts. (c) Row-space angles vary only with eraser support. Overlapping resamples describe sensitivity, not independent replications.

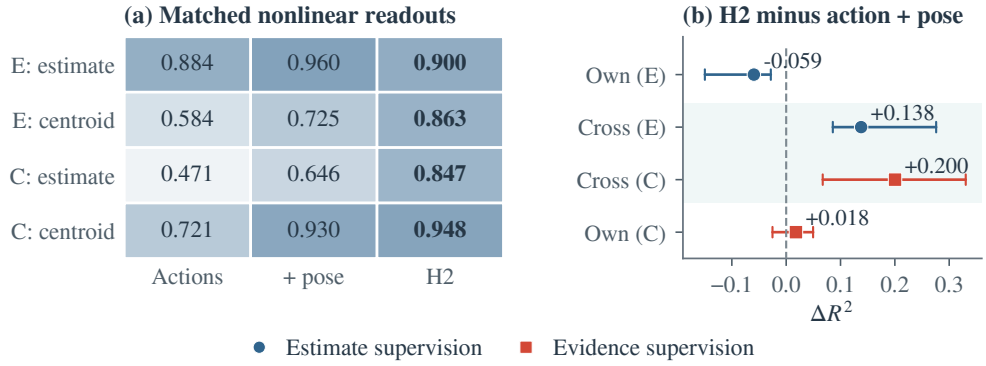


Figure 9: Nonlinear readouts on new layouts. (a) Test R^2 under the same 256-landmark budget. (b) Paired H2 minus action-and-pose R^2: means and joint intervals. E/C denote supervision; shaded rows decode the other public statistic. Predictive advantage does not identify a direct causal effect.

lish counterfactual alignability on the tested support. Without a recipient-pose by donor-coordinate compositional test, they do not identify a reusable natural coordinate computation. We retain DAS as complementary evidence rather than using it to override the action-head audit.

9 Task Validity, Related Work and Implications

Closed-loop success does not yet validate evidence necessity in this task. A retained pilot on 32 separate validation layouts found that the best of four fixed routes succeeds in 72.5% of episodes, failing the planned below-30% criterion. Full teachers outperform their Blind fallback, but the scoring estimator pools every acquired scan, including scans hidden from the policy. An effective open-loop route can therefore support endpoint localization without adaptive evidence use.

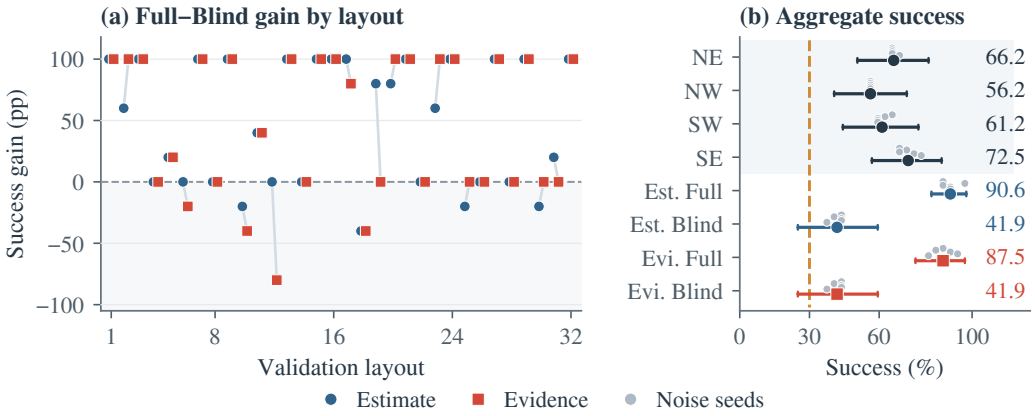


Figure 10: Task pilot: all 32 layouts and five noise seeds. (a) Full–Blind success gain by layout, averaged over seeds. (b) Eight conditions: global means, paired layout/noise 95% intervals and individual seed means. Fixed routes exceed the 30% criterion.

Relation to prior work. Probing controls separate accessibility from task relevance (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022). INLP, amnesic probing and LEACE offer distinct operations and guarantees (Ravfogel et al., 2020; Elazar et al., 2021; Belrose et al., 2023). Our contribution is not a replacement eraser: it integrates matched imitation supervision, native evidence deltas, output-null mixing, and independent readout supports into the same frozen policies. Causal confusion in imitation and activation-patching illusions motivates these competing explanations (de Haan et al., 2019; Makelov et al., 2024). A spatial probe falls short of the coherent understanding of the environment sought in evaluations (Li et al., 2023; Vafa et al., 2024).

10 Conclusion

This paper presents mmClaw, an auditable path-level radar testbed and a controlled study of spatial readability in imitation policies. The key insight is that evidence changes and representation edits expose different dependencies. Because the action head is linear, we can separate the native hidden response to changed returns from the influence an edit adds. Across forty policies and 48 new layouts, supervision reproducibly changes relative coordinate readability ($D = 0.156\,[0.125, 0.236]$). However, covariance-selected edits yield signed output-null projections of 0.207 and 0.357, so head geometry limits how coordinate-edit effects can be interpreted. These results cover only modest policies in one annotated synthetic domain, with no LLM evaluation, physical radar validation or closed-loop results for edited policies. Reporting the variable, operator, fitting support and finite decoder makes this distinction testable without claiming a complete world model. Future studies can check whether a conclusion survives changed supervision, an independent readout support and an output-consistent intervention. Agreement would strengthen a mechanistic interpretation, and disagreement would show whether the evidence concerns accessible statistics, the edit, or the sensing task itself.

REPRODUCIBILITY STATEMENT

The accompanying artifact retains the supplied v11 source corpora, selected policy/probe weights, erasers and analysis scripts. This presentation revision uses the supplied v16 numerical results without retraining policies or rerunning the new-layout experiments. Exact plotted values are embedded in the standalone MATLAB script and exported figure ledgers; the vFinal renderer regenerates the delivered experimental graphics from these values. The underlying v16 experiment scripts and ten referenced v16 source files were not included in the available source package, so the released bundle supports figure reproduction and the inherited v11 checks, but not a complete rerun of every v16 experiment. Appendix K documents the reported audit protocol and Appendices A–J retain the historical specifications. The LaTeX project compiles from the supplied vector assets. MATLAB was statically checked; no MATLAB runtime execution is claimed.

AI USE STATEMENT

AI assistants supported code review, mathematical analysis, visualization, manuscript drafting and layout verification. The experimental values in this presentation revision are retained from the supplied study outputs, not newly generated measurements. The analytic negative control is distinguished from radar experiments. The authors remain responsible for reviewing the evidence, supplying the missing experiment sources, verifying the interpretation and approving the final submission. No independent replication is claimed by this revision.

REFERENCES

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

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, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.11979.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

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. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26941–26975, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

A Policy, Observation and Probe Specifications

A.1 Observation and action interfaces

Training, validation and test layouts are disjoint: 1000–1047, 2000–2011 and 3000–3023. The common corpus contains 3,072 training requests and 384 validation requests. Frozen evaluation retains five nonterminal requests from each of 24 occluded layouts and five acquisition-noise seeds, 500–504. Both teachers relabel this same corpus with unit directions before the action clamp. The target truth is available only to the evaluator and diagnostic probes. It is excluded from policy inputs, imitation labels, counterfactual goals and learned interchange targets.

The path-level backend enumerates direct and specular paths and adds clutter returns in static rooms. Each return carries range, azimuth, Doppler, SNR, precision, path order and source type. Source/path annotations come from the simulator. Public state is rebuilt from an immutable scan history; direct returns increment free-path and endpoint map counts, while non-direct returns remain available in history. The source-aware estimate prefers a cluster with radius 0.45 m and multiplies direct-return weights by four. Evidence instead uses a weighted centroid of at most eight recent scans. Teacher fallbacks select an unknown-map frontier. Full and intervened requests share current feasibility information.

The decoder normalizes a raw two-vector, proposes a 0.65 m displacement, and clamps the endpoint 0.1 m inside the bounds. A six-scan episode has five movements followed by acquired evidence. Scoring pools all acquired scans, including scans a blind policy could not inspect at decision time. Endpoint success uses the public scoring estimator's distance from target truth with an 8 cm threshold. The fixed world-coordinate registration vector has magnitude 6.72 cm. These conventions make success sensitive to sensing trajectories and the separate scoring estimator, rather than solely to policy evidence use.

A.2 Architectures and policy training

Set–CNN uses 14 features per return: scaled range, sine/cosine azimuth, scaled Doppler, SNR, precision, path order, three source indicators, two acquisition-pose displacements and sine/cosine acquisition yaw. A shared 14–24–24 tanh encoder followed by mean/max pooling yields 48 coordinates. The two-channel 8×8 map passes through a 2-to-4 channel convolution with a 3×3 kernel, tanh, 2×2 average pooling and a 36-to-16 tanh layer. Seven feasibility features pass through a 7-to-16 tanh branch. Their 80-dimensional concatenation feeds two 64-unit tanh layers and a linear two-vector head. Neither centroid is supplied as an explicit coordinate. The engineered MLP has 66 input features and two 64-unit tanh layers.

The primary seeds are $11, 22, 33, 44, 55, 66, 77, 88, 99, 110$. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 128 and 300 epochs. Full-validation unit-direction mean squared error selects the checkpoint. Inputs, initialization and minibatch order are paired across teacher groups. Additional same-teacher null cohorts use seeds $111, 121, 131, 141, 151, 161, 171, 181, 191, 200$ under the identical protocol. Selected NumPy weights are checked against the training implementation; the complete curves and chosen epochs remain in the artifact.

The GRU consumes the 66 engineered features, has hidden dimension 64 and a linear two-dimensional output head. Ten seeds train on 384 eight-step demonstration episodes, with 48 validation episodes, for 300 epochs. Episode states begin at zero. Frozen analysis replays a Full prefix and then compares the current request under identical incoming memory. Sustained closed-loop interventions instead carry state generated by their own preceding inputs. Post-GRU edits directly precede the linear head; the recurrent nonlinear computation occurs before the edited state.

A.3 Readout fitting and nuisance controls

Raw summaries concatenate return-feature means/maxima, the flattened map and feasibility features into 163 coordinates. This is a fixed summary, not a lossless observation representation. Frozen random H2 uses the matched architecture and initialization before policy training. Coordinate labels are normalized by room width and height. Linear readouts use training-only standardization and select ridge penalty from $\{0.01, 0.1, 1, 10, 100\}$ on validation requests. The native coefficient conversion

Table 2: Teacher–variable selectivity under independent-seed and fitting-quality controls. Except for the final log-scale row, entries are differences of R^2 differences. Joint intervals resample fits, layouts and noise. The fitting-quality regression is descriptive; validation matching retains five pairs and leaves test accuracy unmatched.

Contrast	Value	Joint 95% interval
Initial teacher pairs	0.165	[0.113, 0.228]
Additional teacher pairs	0.213	[0.143, 0.305]
All teacher pairs	0.189	[0.136, 0.253]
Same Estimate teacher: cohort contrast	-0.015	[-0.081, 0.045]
Same Evidence teacher: cohort contrast	0.033	[-0.006, 0.079]
Initial D minus Estimate cohort contrast	0.180	[0.118, 0.258]
Initial D minus Evidence cohort contrast	0.132	[0.058, 0.209]
Full-action-error-adjusted D	0.242	[0.125, 0.473]
Validation-matched D (five pairs)	0.216	[0.137, 0.333]
Log unexplained-variance interaction	0.876	[0.627, 1.148]

is $B_{\mathrm{raw}} = \mathrm{diag}(1/s)B$, with intercept $b_{\mathrm{raw}} = b - \mu B_{\mathrm{raw}}$. All subspace geometry and policy edits use these native coordinates.

The nonlinear control uses two 64-unit tanh layers and a two-coordinate output, training-only feature/label standardization, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 256 and 150 epochs. Validation selects the checkpoint. Conditional probes first fit affine nuisance regressions for both activations and labels on training requests. The nuisance variables are teacher and student unit directions; the action-plus-pose control additionally includes world pose and room bounds. Fixed nuisance regressions produce validation/test residuals, then a ridge probe predicts residual labels. These scores address conditional linear accessibility, and permit nonlinear action information to remain.

B Selectivity Controls and Their Support

For teacher and label indices E, C, the only selectivity symbol is $D = (R_{E,E} - R_{E,C}) - (R_{C,E} - R_{C,C})$. Each R^2 pools two-coordinate SSE and centered SST over the same test requests. Bootstrap draws recompute SST, preserve teacher/label pairing, and independently resample the original and additional seed cohorts. Same-teacher cohort differences provide a training-variation null; zero interactions from identical raw/random representations are algebraic controls only.

Quality adjustment regresses each fit's difference between the two variable readouts on teacher identity and its Full action error. Action error is an outcome of training, so the adjusted coefficient is a descriptive sensitivity analysis rather than an identified direct teacher effect. Minimum-cost matching uses validation action error without replacement and a fixed 0.5 cm caliper. Test accuracy is never used to choose matched pairs. Sparse support and remaining test-quality imbalance are retained in the table.

The unexplained-variance analysis uses $\log(\mathrm{SSE}/\mathrm{SST})$, allowing negative R^2 rather than clipping it to a probability. Its interaction is oriented so positive values indicate the same teacher-matched selectivity as Eq. 1. A floor of 10^{-8} is an explicitly recorded numerical guard. Coordinate-wise test Pearson correlations are descriptive; estimate-to-centroid predictive R^2 uses a training-fitted, validation-selected ridge decoder.

C Closed-form Erasure and Independent Readout Audits

The numerical tables in Appendices A–J retain the historical development and prior-layout evidence. The new action-head audit changes their interpretation, not these source measurements.

Let $S = \mathrm{Cov}(h)$ and $C = \mathrm{Cov}(h, z)$ use centered training activations and two-dimensional labels. Orthogonal covariance projection uses $P_C = I - UU^\top$, where $U = \mathrm{orth}(C)$. For LEACE, set

Table 3: Dependence between the public estimate $\hat{x}$ and Evidence centroid c. Coordinates are normalized to room bounds. The estimate-to-centroid readout is fitted on training layouts and selected on validation layouts; all entries below use held-out test layouts.

Measure	Value	Joint 95% interval
Pearson correlation, x coordinate	0.787	[0.656, 0.869]
Pearson correlation, y coordinate	0.880	[0.836, 0.923]
Held-out R^2: $\hat{x} \to c$	0.600	[0.394, 0.741]

$W = (S^{1/2})^+$, $U_W = \mathrm{orth}(WC)$ and

$$P_L = I - W^+ U_W U_W^\top W. \tag{5}$$

Row activations transform as $(H - \mu)P^\top + \mu$. The transpose is essential because LEACE is generally oblique. Covariance eigenvalue and cross-covariance singular-value cutoffs are fixed relative thresholds of 10^{-10}. The audit checks update rank, idempotence, training covariance residual, mean restoration, frozen-head parity and LEACE's training distortion inequality. Rank refers to $I - P$, not the dimension of the surviving representation.

Both erasers satisfy a finite-sample covariance constraint of rank at most two. This gives affine least-squares uninformative readout on the fitted empirical distribution. It establishes neither nonlinear forgetting nor distribution-independent erasure. A fresh affine probe trained on that same erased sample must be constant; its test failure is consequently an algebra check. We retain train/validation/test scores and held-out covariance diagnostics, then use the disjoint-layout protocol for an informative generalization test.

The forward split fits erasure on layouts 1000–1023 and the audit probe on 1024–1047. The reverse split exchanges these roles. Both were fixed before their outcomes. Validation layouts choose ridge penalties; test layouts fit neither map. The comparison includes identity, covariance projection, LEACE, equal-rank orthogonal random removal and per-request norm-matched random perturbations. The latter have no fixed removed rank. Empirical covariance projection and LEACE can share the same update image on covariance support but differ in how input activations map to that image; "operator" is therefore more precise than calling each a distinct orthogonal subspace.

Table 4: Public-estimate erasure across four frozen policies. Cells show Full action error in cm with joint 95% intervals. E/C denote Estimate/Evidence supervision; Set denotes Set–CNN. Norm-matched controls match each LEACE perturbation, while random rank-two controls match removed rank.

Operation	MLP / E	Set / E	Set / C	GRU / E
Unedited	0.2 [0.1, 0.2]	4.9 [3.8, 6.2]	2.5 [2.0, 3.0]	0.4 [0.3, 0.5]
Covariance	22.9 [20.9, 24.7]	21.3 [19.0, 23.5]	16.7 [14.5, 18.9]	22.2 [19.9, 24.5]
LEACE	17.7 [16.1, 19.2]	17.4 [15.8, 19.1]	13.8 [12.6, 15.0]	17.8 [16.1, 19.5]
Random rank 2	0.9 [0.6, 1.1]	4.9 [3.8, 6.2]	2.6 [2.1, 3.1]	1.6 [1.2, 2.0]
LEACE norm control	3.3 [3.0, 3.6]	5.6 [4.5, 6.8]	3.3 [2.9, 3.8]	5.1 [4.6, 5.6]

C.1 Paired identity and repeated layout splits

Identity uses exactly the same audit-probe training half and validation rule as its paired erasure map. Twenty recorded random half-layout partitions retain all forty fitted policies. Each partition fits the eraser and audit readout on disjoint halves; the resulting distribution measures the whole protocol's sensitivity. Overlapping partitions do not supply twenty independent environment replications. Covariates for the original halves include target/pose coordinates, estimate availability, acquisition step and map occupancy.

Table 5: Independent eraser/probe training layouts. Public-estimate test R^2 [joint 95% interval]. Both split directions were specified before these outcomes.

Policy	Eraser	Forward split	Reverse split
MLP / E	Covariance	−0.254 [−0.603, 0.019]	0.632 [0.512, 0.726]
MLP / E	LEACE	−0.252 [−0.601, 0.025]	0.635 [0.516, 0.729]
Set / E	Covariance	−0.474 [−0.822, −0.258]	0.538 [0.428, 0.627]
Set / E	LEACE	−0.499 [−0.860, −0.269]	0.523 [0.410, 0.617]
Set / C	Covariance	−0.386 [−0.751, −0.139]	0.519 [0.404, 0.617]
Set / C	LEACE	−0.393 [−0.760, −0.152]	0.474 [0.367, 0.562]
GRU / E	Covariance	−0.429 [−0.791, −0.158]	0.527 [0.401, 0.623]
GRU / E	LEACE	−0.428 [−0.793, −0.146]	0.524 [0.397, 0.617]

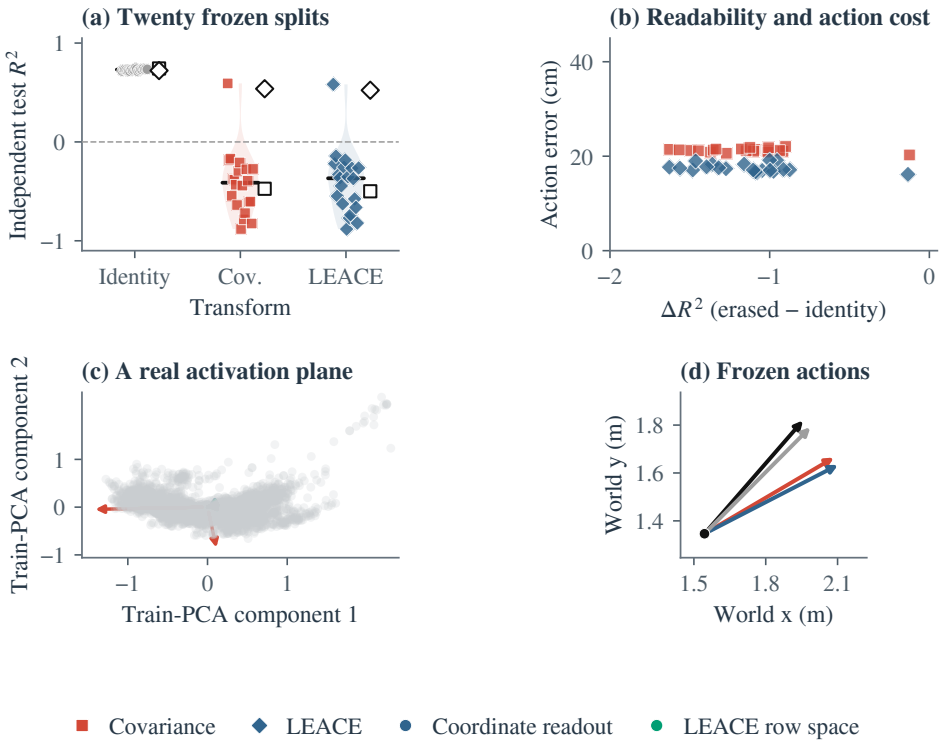


Figure 11: Historical erasure diagnostics on the development cohort. Split means, paired action costs, a training-PCA projection and one frozen request are retained. Projected angles are not full-space angles; frozen actions are not rollout trajectories.

C.2 Variable and operator controls

We erase each teacher coordinate, the actual teacher unit direction, pose and step, with matched random perturbations. Pose and step are context variables that may affect fallback behavior; they are not assumed null concepts. Pairwise principal angles and action-head energy are reported in the artifact. The covariance and LEACE maps can have the same update image while producing different finite changes, so dividing an action error by image energy does not identify a causal share.

Iterative regression-direction deletion repeatedly fits a readout in the surviving state and removes its residual coefficient directions. The rank-32 cap and the validation-threshold extension are retained as distinct procedures. Validation $R^2 < 0.05$ requires 54–62 dimensions for MLP, 62–64 for Estimate Set–CNN and 64 for Evidence Set–CNN; GRU ranges from 34 to 60. These costs do not

Table 6: Identity distinguishes baseline readout support from the full erasure audit in Set–E.

Split	Identity R^2	LEACE R^2	Paired change (95% CI)
Forward	0.743	−0.499	−1.242 [−1.617, −0.972]
Reverse	0.722	0.523	−0.199 [−0.294, −0.111]

Identity and erased probes use the same audit half and validation layouts within each row. Forward/reverse exchanges both eraser-fit and probe-fit support, so the difference cannot be assigned exclusively to either one.

Table 7: Twenty random half-layout audits retain the erasure penalty while revealing split sensitivity.

Model	Identity	LEACE median [range]	Δ median
MLP–E	0.975	−0.374 [−0.997, 0.691]	−1.349
Set–E	0.731	−0.368 [−0.882, 0.582]	−1.096
Set–C	0.854	−1.055 [−1.415, 0.629]	−1.899
GRU–E	0.949	−0.379 [−1.216, 0.677]	−1.327

Pooled test R^2 is summarized across 20 fixed splits, each with 24 eraser-fit and 24 disjoint probe-fit training layouts. Δ is LEACE minus same-split identity. Ranges describe split sensitivity, not confidence intervals; splits reuse models and test layouts. The label is the public estimate for E and the evidence centroid for C. Individual split estimates and joint fit/layout/noise intervals are released.

Table 8: LEACE label controls: action error and native response. Coordinate, direction and pose updates have rank two; step updates have rank one. Each mean uses ten fits. A dash marks a coordinate statistic reported with intervals in Table 9.

Model	Own coord.	Direction	Pose	Step
Frozen Full-request action error (cm)				
MLP–E	17.7	22.1	8.0	5.7
Set–E	17.4	21.3	8.8	6.4
Set–C	12.8	16.7	6.0	5.0
GRU–E	17.8	22.1	8.3	5.7
Native raw-response projection at 50% retention				
MLP–E	–	1.000	−0.008	−0.016
Set–E	–	0.987	−0.007	−0.011
Set–C	–	1.004	0.009	−0.001
GRU–E	–	0.995	0.012	−0.017

Own coordinate is the estimate for E and centroid for C. Direction is the pre-clamp teacher loss label. Pose and step are context variables, not semantic nulls. Signed projections can lie outside $[0, 1]$. Direction intervals: Set–E 0.987 $[0.941, 1.026]$; Set–C 1.004 $[0.969, 1.039]$.

constrain the rank of other erasure maps. All threshold failures and held-out residual information remain in the results. Closed-form action comparisons here are frozen-request tests; no additional closed-loop success under LEACE is implied.

D Native Responses, Layerwise Edits and Learned Alignments

Coherent thinning keeps a fixed, ceiling-rounded subset of each scan's returns at fractions 0.25, 0.50 and 0.75. The permutation depends on acquisition index and decision step, independently of outcome and source. Acquisition poses stay fixed. History, count map, estimate, recent summaries and scan metadata are reconstructed, so the intervention acts through the ordinary feature encoders. For GRU, current-step Full and thinned requests start with identical incoming states from the recorded Full prefix. State changes caused by earlier thinning are a separate untested intervention.

At the fixed linear action head, the vector identity in Eq. 2 holds exactly. For a covariance eraser or LEACE, $Q = I - P$ defines removed and preserved components, with oblique components allowed. Signed contribution is a ratio of aggregated inner products and total raw-response energy. Its complement sums to one; negative values and values above one reflect cancellation and are retained. The artifact also stores component vectors, the cross term, reconstruction residual, decoded endpoint changes and decoder nonadditivity. Raw action vectors, normalized directions and clamped positions are different estimands.

The primary native comparison uses the public-estimate regression-row space, teacher-relative-direction space and equal-rank random spaces. A timestamped follow-up adds the absolute Evidence-centroid space to compare the same variable in absolute and relative form. A second outcome-motivated follow-up applies the completed closed-form erasers to the already retained native deltas. The final exploratory contrast compares teacher-by-erased-variable signed contributions under each operator at every retention level; it was specified after inspecting marginal contributions. It is consequently reported as a follow-up rather than an independent confirmation.

Table 9: Teacher-matched coordinate maps across frozen policies at 50% retention. Signed native raw-response projections are percentages with joint 95% intervals. The coordinate is the estimate for E and the centroid for C. LEACE components are oblique.

Map	MLP / E	Set / E	Set / C	GRU / E
Covariance	38.2 [28.5, 59.8]	86.0 [80.0, 90.1]	90.4 [86.8, 94.3]	21.7 [15.8, 35.7]
LEACE	34.1 [23.9, 59.6]	82.8 [75.1, 90.7]	81.2 [74.1, 89.9]	25.4 [15.5, 51.1]
Random rank 2	3.4 [1.3, 6.0]	2.6 [1.3, 4.2]	5.4 [1.3, 10.3]	2.4 [1.4, 3.8]

Table 10: Exploratory native teacher-by-variable interaction. The four signed fractions are recomputed within each paired bootstrap draw. Per-fit ranges remain visible where signs differ.

Map	Retention	D_{native} [95% CI]	Fit range
Covariance	25%	0.064 [0.029, 0.103]	[−0.010, 0.160]
Covariance	50%	0.077 [0.039, 0.113]	[−0.010, 0.158]
Covariance	75%	0.069 [0.031, 0.108]	[−0.003, 0.132]
LEACE	25%	0.733 [0.492, 0.943]	[−0.053, 1.104]
LEACE	50%	0.628 [0.445, 0.833]	[0.182, 0.863]
LEACE	75%	0.589 [0.424, 0.809]	[0.399, 0.806]

D.1 Finite direction and bounded-position responses

All component-only edits pass through the original normalization and bounds clamp. The residual is $\eta = d_Q + d_{\text{pres}} - d$; plotted closing corrections use $-\eta$. Signed projections retain cancellation, values above one and negative values. Spatial examples use layout 3000, noise seed 500, fit 11 and all five scored decisions. These are frozen-request action changes, not edited closed-loop trajectories.

Uniform-goal steering samples room-normalized goals with seed 20270917. Minimum-norm changes use a training-fitted native-coordinate probe. Random changes match the hidden Euclidean norm per request. Relative directions use the recipient pose; distance is a separate label. Earlier-layer edits continue through the frozen nonlinear downstream network. At H2, $\delta a = AW^{\dagger}\delta z$ for probe W, while head-energy fraction $\|AQ\|_F^2/\|A\|_F^2$ also depends on the selected projector. These quantities are related but not identical: steering additionally depends on probe singular values and the requested label change.

The action-row positive control includes the original output and bias. It targets a raw unit direction and verifies raw output recovery; the position decoder is applied afterward. An analytic coordinate-minus-pose representation checks that the measurement can detect a known coordinate effect. Normalized gains divide improvement over the unedited baseline by the positive control's improvement. They may exceed one, and they are not a unique mediation fraction. Input estimate replacement re-

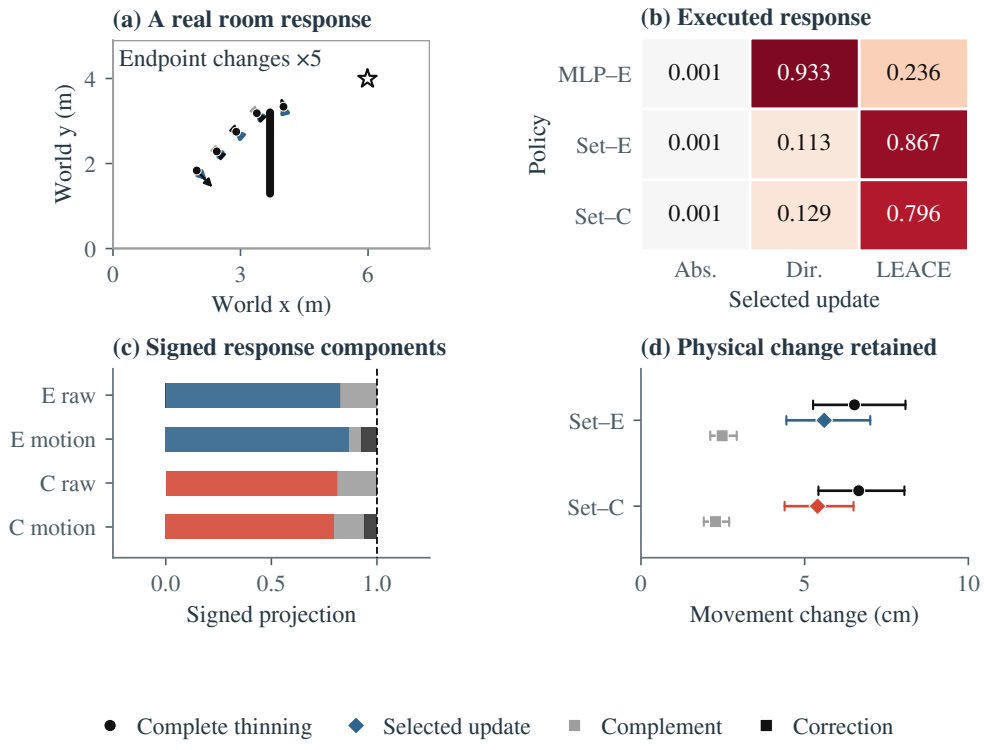


Figure 12: Historical native 50% retention responses. Selected, complementary and decoder correction terms remain visible. The new row/null and sequential-direction audits qualify the interpretation of these operator-dependent effects.

builds the engineered MLP's derived feature vector. Set–CNN has no estimate-coordinate field, so its native thinning response is the available coherent input-level comparison.

Distributed interchange learns an orthonormal basis $U \in \mathbb{R}^{d\times k}$ and applies $h'_R = h_R + UU^\top(h_D - h_R)$. Candidate ranks are 2, 4 and 8. With frozen policy weights, 100 epochs of Adam optimize the raw direction MSE using learning rate 0.01 and batch size 128. QR supplies the orthonormal basis. Fixed donor permutations are generated separately for training, validation and test. Validation jointly selects rank and epoch. Donor teacher coordinates are transported through normalized room coordinates before computing the desired direction from the recipient pose. No target truth is used. Equal-rank random subspaces and unedited outputs share the same donor pairs.

Interchange accuracy is the proportion with executed direction cosine above 0.95; continuous cosine, baseline accuracy, random controls, hidden displacement and probe–alignment principal angles are also retained. High baseline cosine is expected when teacher coordinates are correlated across recorded requests. Learned alignment can recover an output-driving subspace without uniquely identifying the natural computation of an absolute coordinate, which is why the native response experiment is reported separately.

D.2 Donor difficulty

Angle strata use the required change from the original action to the transported donor goal: 0–30°, 30–60°, 60–90°, and above 90°. Original, random, learned and positive outputs retain the same recipient–donor pairs. Empty strata remain visible. The displayed difficult example uses the maximum available angle within the fixed fit and condition when the originally requested large-angle case is absent.

Table 11: Finite selected-component projections at 50% measurement retention. Each entry is a point estimate with its joint 95% interval below. The signed statistic is $\sum_i v_i^\top d_i / \sum_i \|d_i\|^2$, evaluated after the named transformation; it is not a causal share. The residual replaces v by $v_Q + v_C - d$. Component effects after normalization are nonadditive. Absolute ridge and LEACE use the teacher-matched public coordinate; relative ridge uses the teacher-matched unit direction. Every group retains 10 fits and 600 requests per fit. Bounds never activate on this frozen sample, explaining the equality of direction and bounded projections.

Policy / operator	Raw	Direction	Bounded	Bounded residual
MLP-E	0.003	0.001	0.001	-0.002
Absolute ridge	$[-0.004, 0.011]$	$[-0.003, 0.005]$	$[-0.003, 0.005]$	$[-0.009, 0.004]$
MLP-E	0.893	0.933	0.933	-0.026
Relative ridge	$[0.858, 0.929]$	$[0.901, 0.960]$	$[0.901, 0.960]$	$[-0.049, -0.006]$
MLP-E	0.341	0.236	0.236	-0.024
LEACE	$[0.239, 0.596]$	$[0.128, 0.518]$	$[0.128, 0.518]$	$[-0.092, 0.047]$
Set-E	0.828	0.867	0.867	-0.075
LEACE	$[0.751, 0.907]$	$[0.802, 0.935]$	$[0.802, 0.935]$	$[-0.140, -0.026]$
Set-C	0.812	0.796	0.796	-0.059
LEACE	$[0.741, 0.899]$	$[0.740, 0.856]$	$[0.740, 0.856]$	$[-0.099, -0.026]$
GRU-E	1.070	1.030	1.030	0.010
Relative ridge	$[1.052, 1.086]$	$[1.019, 1.044]$	$[1.019, 1.044]$	$[0.001, 0.017]$
GRU-E	0.254	0.194	0.194	-0.055
LEACE	$[0.155, 0.511]$	$[0.094, 0.475]$	$[0.094, 0.475]$	$[-0.118, 0.012]$

Table 12: Finite movement changes at 50% measurement retention (cm). Changes are Euclidean distances from the unmodified decoded endpoint on the same request. The last column is $\|v_Q + v_C - d\|$, after the actual normalization and bounds decoder. The three displacement norms must not be added as contributions. Intervals jointly resample fit, layout, and sensor noise. These are frozen-request edits, not edited-policy closed-loop trajectories.

Policy / operator	Full thinning	Selected only	Complement only	Nonadditive residual
MLP-E	9.08	0.12	9.06	0.12
Absolute ridge	$[4.88, 14.48]$	$[0.06, 0.20]$	$[4.86, 14.43]$	$[0.05, 0.23]$
MLP-E	9.08	8.40	0.68	0.76
Relative ridge	$[4.88, 14.48]$	$[4.45, 13.49]$	$[0.35, 1.14]$	$[0.28, 1.49]$
MLP-E	9.08	5.82	5.84	1.34
LEACE	$[4.88, 14.48]$	$[4.21, 7.71]$	$[2.63, 10.10]$	$[0.61, 2.33]$
Set-E	6.52	5.60	2.48	0.61
LEACE	$[5.26, 8.08]$	$[4.44, 7.00]$	$[2.12, 2.93]$	$[0.38, 0.94]$
Set-C	6.65	5.39	2.28	0.54
LEACE	$[5.42, 8.05]$	$[4.39, 6.49]$	$[1.92, 2.69]$	$[0.33, 0.79]$
GRU-E	9.46	9.84	0.48	0.46
Relative ridge	$[5.22, 14.88]$	$[5.45, 15.43]$	$[0.32, 0.67]$	$[0.20, 0.84]$
GRU-E	9.46	6.21	6.33	1.35
LEACE	$[5.22, 14.88]$	$[4.47, 8.33]$	$[3.10, 10.67]$	$[0.62, 2.42]$

E Label Semantics, Function Proximity and Decoder Transfer

Label equivalence determines what an unseen input combination tests. History, map and estimate deletions are denoted H, M, X; strict blindness deletes HMX while retaining feasibility. For decoded policy position $g_P(r)$, action error is $E_{\mathrm{act}} = \mathbb{E}\|g_B(r) - g_T(r)\|$ and response error is $E_{\mathrm{resp,k}} = \mathbb{E}\|[g_B(I_k r) - g_B(r)] - [g_T(I_k r) - g_T(r)]\|$. If $E_{\mathrm{mask,k}}$ is imitation error on the intervened request, then $|E_{\mathrm{resp,k}} - E_{\mathrm{mask,k}}| \le E_{\mathrm{act}}$. Accurate Full imitation consequently makes response error close to masked-input imitation error; the representation experiments above supply the additional explanatory comparisons.

Table 13: Finite direction changes and raw radial energy at 50% measurement retention. Angles are exact finite absolute changes from the baseline raw direction. The radial ratio is the summed squared component of the full raw response parallel to the baseline direction, divided by the summed squared full raw response; it is not a motion-direction contribution. No compared raw vector is degenerate. The GRU comparison holds its incoming Full hidden state fixed.

Policy / operator	Full (degrees)	Selected (degrees)	Complement (degrees)	Full raw radial ratio
MLP-E	8.86	0.11	8.83	0.606
Absolute ridge	[4.55, 14.43]	[0.05, 0.18]	[4.53, 14.40]	[0.359, 0.711]
MLP-E	8.86	8.09	0.60	0.606
Relative ridge	[4.55, 14.43]	[4.12, 13.25]	[0.31, 1.00]	[0.359, 0.711]
MLP-E	8.86	5.17	5.90	0.606
LEACE	[4.55, 14.43]	[3.73, 6.86]	[2.50, 10.51]	[0.359, 0.711]
Set-E	5.78	4.96	2.19	0.217
LEACE	[4.65, 7.19]	[3.93, 6.23]	[1.87, 2.58]	[0.143, 0.311]
Set-C	5.89	4.77	2.01	0.107
LEACE	[4.80, 7.14]	[3.88, 5.75]	[1.69, 2.38]	[0.061, 0.168]
GRU-E	9.32	9.77	0.42	0.629
Relative ridge	[4.89, 15.07]	[5.13, 15.77]	[0.28, 0.59]	[0.379, 0.735]
GRU-E	9.32	5.51	6.36	0.629
LEACE	[4.89, 15.07]	[3.97, 7.41]	[2.91, 10.97]	[0.379, 0.735]

Table 14: Validation-selected donor interchange. Normalized gain and accuracy (cosine > 0.95) are percentages [joint 95% interval].

Policy / layer	Normalized gain (%)	Accuracy (%)
MLP / E / H1	90.6 [88.5, 92.4]	97.8 [96.0, 99.3]
MLP / E / H2	86.8 [83.6, 89.3]	95.0 [92.1, 97.5]
Set / E / Encoder	85.6 [79.4, 89.7]	92.4 [88.5, 95.6]
Set / E / H1	80.7 [74.4, 85.0]	89.1 [84.9, 92.8]
Set / E / H2	76.6 [70.5, 80.9]	85.7 [81.3, 89.9]
Set / C / Encoder	90.2 [87.7, 92.2]	99.6 [98.9, 100.0]
Set / C / H1	84.0 [80.3, 87.1]	98.7 [97.2, 99.7]
Set / C / H2	80.1 [76.0, 83.7]	97.9 [96.1, 99.3]

We audit both the pre-clamp unit direction used for supervision and the bounded action used in evaluation, with tolerance 10^{-6}. On each base request, a withheld label overlaps if it equals any H, M or X teacher label. Every HX test request overlaps for both teachers. For bounded HMX actions, 436/600 Estimate requests and 462/600 Evidence requests overlap, leaving 164 and 138 for conditional analysis. Pre-clamp and post-clamp equivalence agree on every audited test comparison. The strata differ sharply in acquisition stage and map occupancy, so their conditional errors describe different request populations. Same-request nonoverlap is a local criterion; it does not imply that the fallback function is absent from other training poses.

Filtering matching single-deletion augmentations redistributes error. Estimate's nonoverlap response error falls from 48.8 cm under Same-label supervision to 33.1 cm under the strict filter, while its aggregate error rises from 46.5 to 84.2 cm. The corresponding Evidence values are 35.3 to 34.5 cm and 23.5 to 41.1 cm. Appendix E reports pre/post-clamp equivalence, retained variant counts, actual sampling probabilities, Full imitation error and masked error for these engineered MLPs. Nearest-neighbor labels, fallback convex-hull fits and stratum covariates describe the remaining alternatives. Without matched filtering-frequency controls, these contrasts characterize the complete training treatment rather than a uniquely semantic filtering effect.

The function-proximity audit further limits a novelty claim. After strict filtering, all 600 test blind labels per teacher have a retained training neighbor within 0.10 in normalized pose-plus-unit-direction distance. The strata also occur at different acquisition stages: Estimate nonoverlap requests have mean step 3.67, versus 1.37 for overlap. Filtering same-request equalities therefore leaves nearby

Table 15: Recorded-donor interchange at the Set–CNN encoder. Means and joint 95% intervals use 10 fits, 600 requests and 24 layouts per teacher. Accuracy is the fraction of executed directions with cosine > 0.95. Raw and executed directions are shown separately; random subspaces match the selected rank.

Teacher / edit	Raw cosine	Executed cosine	Accuracy (%)	Pairs
Set–E Unedited	0.911 [0.894, 0.927]	0.911 [0.894, 0.927]	56.1 [49.2, 63.0]	6,000
Set–E Random	0.915 [0.893, 0.933]	0.915 [0.893, 0.933]	57.7 [51.4, 63.8]	6,000
Set–E DAS	0.987 [0.983, 0.991]	0.987 [0.983, 0.991]	92.4 [88.5, 95.6]	6,000
Set–E Positive	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	100.0 [100.0, 100.0]	6,000
Set–C Unedited	0.959 [0.950, 0.966]	0.959 [0.950, 0.966]	73.0 [67.7, 77.9]	6,000
Set–C Random	0.961 [0.949, 0.971]	0.961 [0.949, 0.971]	74.4 [67.4, 81.2]	6,000
Set–C DAS	0.996 [0.995, 0.997]	0.996 [0.995, 0.997]	99.6 [98.9, 100.0]	6,000
Set–C Positive	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	100.0 [100.0, 100.0]	6,000

Table 16: Donor difficulty at the Set–CNN encoder. Bins use the unedited raw recipient direction versus donor target; all edits share membership. Cosine is raw; accuracy uses executed direction. $N/R/L$ counts fit–request pairs, distinct (layout, noise, step) tuples, and layouts. Intervals jointly resample fits, layouts and noise. Empty bins remain dashes and zero counts, without extrapolation.

Angle	Unedited cosine	DAS cosine	Unedited accuracy (%)	DAS accuracy (%)	$N/R/L$
Set–E					
$[0, 30)^\circ$	0.972 [0.965, 0.979]	0.989 [0.984, 0.992]	76.4 [69.0, 83.5]	93.0 [88.4, 96.6]	4,406/478/24
$[30, 60)^\circ$	0.752 [0.727, 0.774]	0.984 [0.977, 0.991]	0.0 [0.0, 0.0]	91.1 [85.1, 96.9]	1,545/195/24
$[60, 90)^\circ$	0.465 [0.449, 0.492]	0.962 [0.898, 0.988]	0.0 [0.0, 0.0]	75.5 [34.8, 100.0]	49/10/6
$[90, 180]^\circ$	—	—	—	—	0/0/0
Set–C					
$[0, 30)^\circ$	0.973 [0.968, 0.977]	0.997 [0.996, 0.997]	79.1 [74.1, 83.6]	99.9 [99.7, 100.0]	5,534/568/24
$[30, 60)^\circ$	0.794 [0.745, 0.827]	0.988 [0.981, 0.993]	0.0 [0.0, 0.0]	95.5 [87.7, 100.0]	466/62/22
$[60, 90)^\circ$	—	—	—	—	0/0/0
$[90, 180]^\circ$	—	—	—	—	0/0/0

supervision of the pose/step-dependent fallback function. Conversely, the proposed interpolation explanation is incomplete: strict-filtered Estimate actions remain 81.2 cm [75.1, 86.2] from the convex hull of the four candidate teacher actions. These diagnostics identify specific remaining alternatives without assigning the behavior a compositional mechanism.

Cross-condition probes test the portability of a readout. Full- and Blind-trained readouts are each fitted only on training layouts and evaluated in both conditions, with pose/step and constant controls. Their targets remain the reference Full-request coordinates. Separately trained Blind readouts give public-estimate R^2 of -0.056 [-0.278, 0.074], -0.050 [-0.259, 0.071] and -0.067 [-0.274, 0.053] for MLP, Estimate Set–CNN and Evidence Set–CNN. These unresolved readouts do not val-

idate a common default coordinate. A decoded point's relation to the student's action endpoint and the teacher fallback is evaluated separately. This distinction prevents an out-of-support Full decoder from being interpreted as a calibrated internal belief. The filter comparison uses the engineered MLP and each teacher's own demonstration corpus, separately from the common-request Set–CNN supervision experiment. The augmentation set contains Full and the single deletions H, M, X. Full-label supervision retains the Full teacher direction; Same-label recomputes the direction after intervention. The empty-map filter drops matching fallbacks only when the original map is empty. The strict filter drops every matching single-deletion label on that request. Full examples remain. Each minibatch samples a base request uniformly and then samples uniformly among its retained variants. The treatment therefore changes both the set of labels and their marginal sampling frequencies.

Both actual loss labels and decoded bounded positions are compared at tolerance 10^{-6}. Equivalence is checked per training, validation and test request. A same-request label-nonoverlap set is not a claim of function-level novelty. We therefore compute nearest-neighbor distance from each test blind label to retained training augmentations in the concatenation of room-normalized pose and unit direction. The prespecified bins have edges 0, 0.05, 0.10, 0.20 and infinity. All strict-filtered test labels remain close to some retained training example, limiting a claim of globally new target functions.

Stratum descriptions retain step index, map count occupancy, normalized pose and distance to room bounds. Covariate differences are reported rather than removed by selecting favorable requests. Convex-hull projections use teacher Full/H/M/X decoded endpoints; a timestamped follow-up applies the same analysis to the pre-clamp unit labels. These projections test a geometric mixture explanation. Nonunique weights are explicitly flagged, and a large residual does not establish a unique alternative mechanism.

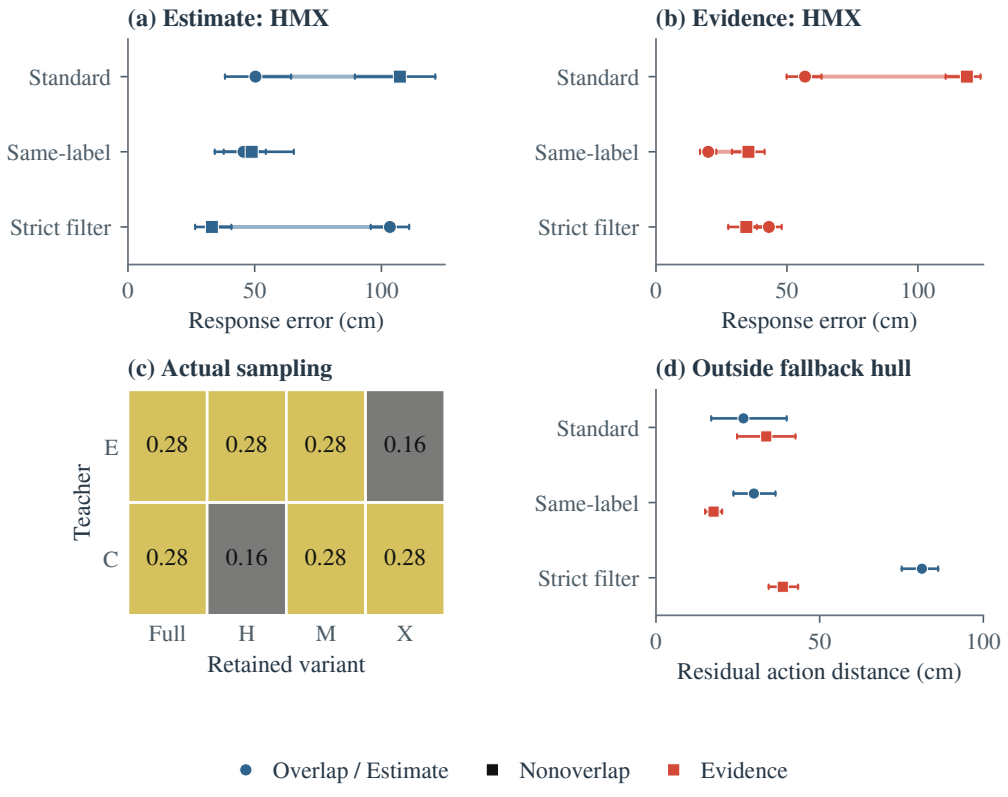


Figure 13: HMX label strata and the filtering treatment. Circles denote overlap, squares nonoverlap; connecting lines pair strata within a treatment, not individual requests. Joint 95% intervals retain all audited denominators. Sampling heatmap reports actual strict-filter marginal probabilities. Hull residuals do not identify unique mixing weights.

Full/Blind readouts are fitted separately on training activations and validated in their respective training condition. Each is evaluated in both conditions. Labels remain the reference Full public variable or target truth. Pose/step and training-constant predictions supply baselines. Decoder-to-action endpoint and decoder-to-frontier distances use room-normalized coordinates and metres as explicitly labeled; Pearson correlations and directional cosines are descriptive and can share pose dependence. A reference-coordinate readout under blindness is not automatically a current belief variable.

The full-history target probe is a train-fitted achievable prediction reference. It encodes every recorded return through a 128–256 shared network, pools the complete history, and uses a three-layer 256-unit head with feasibility inputs. Five seeds train for 150 epochs with validation selection. No target information is added to the policy. An achieved score depends on model and optimization choices and cannot be interpreted as an information-theoretic ceiling. Ratios to a negative reference R^2 are therefore omitted.

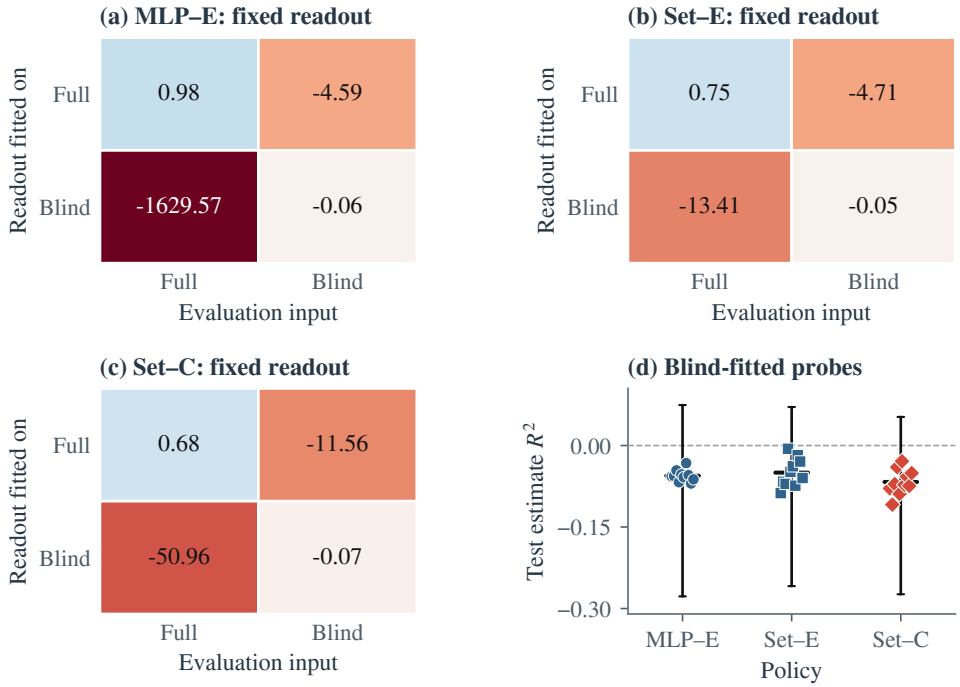


Figure 14: Fixed readout transfer retains untruncated negative R^2 values. Color uses zero-centered $\mathrm{sign}(x)\log_{10}(1+|x|/0.1)$, shared bounds ± 4.3. Separate probes fit on training/validation layouts only. Labels remain the original Full public estimate, not a current internal belief. Blind-fitted fits and full joint intervals are shown in (d).

Table 17: Equivalence to at least one single-channel deletion. Unit labels enter the training loss; positions follow normalization and bounds clamping. The Euclidean tolerance is 10^{-6} for each object. Every comparison agrees in the recorded train, validation, and test sets.

Teacher	Split	Mask	n	Unit	Position	Differ
Estimate	Train	HX	3072	3072	3072	0
Estimate	Train	HMX	3072	1088	1088	0
Estimate	Val	HX	384	384	384	0
Estimate	Val	HMX	384	138	138	0
Estimate	Test	HX	600	600	600	0
Estimate	Test	HMX	600	436	436	0
Evidence	Train	HX	3072	3072	3072	0
Evidence	Train	HMX	3072	1082	1082	0
Evidence	Val	HX	384	384	384	0
Evidence	Val	HMX	384	133	133	0
Evidence	Test	HX	600	600	600	0
Evidence	Test	HMX	600	462	462	0

Table 18: MLP supervision and actual variant sampling by teacher (E: Estimate; C: Evidence). Every treatment uses 3,072 base requests, the 66–64–64–2 architecture and 3,072 variant draws per epoch. Columns Full/H/M/X give percentage sampling probability under uniform base requests and uniform retained variants within each base.

Teacher	Treatment	Variants	Full	H	M	X
E	Standard	3072	100.0	0.0	0.0	0.0
E	Full-label	12288	25.0	25.0	25.0	25.0
E	Same-label	12288	25.0	25.0	25.0	25.0
E	Empty-map filter	11529	27.1	27.1	27.1	18.8
E	Strict filter	11200	28.0	28.0	28.0	16.1
C	Standard	3072	100.0	0.0	0.0	0.0
C	Full-label	12288	25.0	25.0	25.0	25.0
C	Same-label	12288	25.0	25.0	25.0	25.0
C	Empty-map filter	11511	27.1	18.7	27.1	27.1
C	Strict filter	11206	27.9	16.2	27.9	27.9

Table 19: MLP imitation errors under supervision treatments on all 600 frozen requests. Cells are cm with joint fit/layout/noise 95% intervals. Action error uses Full inputs; masked imitation and response errors use HMX. E/C denote Estimate/Evidence teachers.

Teacher	Treatment	E_{act}	$E_{\mathrm{mask,HMX}}$	$E_{\mathrm{resp,HMX}}$
E	Standard	0.16 [0.13, 0.19]	65.86 [60.96, 72.04]	65.88 [60.97, 72.06]
E	Full-label	1.06 [0.88, 1.28]	59.99 [57.28, 62.82]	59.93 [57.21, 62.78]
E	Same-label	2.54 [2.24, 2.86]	46.74 [41.37, 52.56]	46.48 [41.15, 52.21]
E	Empty-map filter	2.07 [1.80, 2.36]	56.39 [47.57, 66.03]	56.15 [47.37, 65.70]
E	Strict filter	1.98 [1.71, 2.26]	84.49 [79.37, 88.92]	84.15 [79.16, 88.61]
C	Standard	3.81 [3.15, 4.61]	71.37 [67.18, 76.01]	71.12 [67.18, 75.48]
C	Full-label	3.55 [2.90, 4.36]	59.68 [56.97, 62.42]	60.02 [57.22, 62.82]
C	Same-label	4.07 [3.41, 4.83]	23.11 [20.67, 25.42]	23.48 [20.89, 25.90]
C	Empty-map filter	3.94 [3.27, 4.75]	29.20 [24.02, 34.98]	28.73 [23.54, 34.53]
C	Strict filter	3.86 [3.13, 4.72]	41.56 [37.63, 45.88]	41.14 [37.24, 45.34]

Table 20: HMX-stratum covariates and room-normalized position. Cells show means and layout/noise 95% intervals. Step is zero-based, map cells count nonzero free/occupied cells, and boundary is distance to room bounds. Strata differ in support and are not exchangeable randomized groups.

Teacher / stratum	Step	Map cells	Boundary (m)	x	y
Estimate Overlap	1.37 [1.23, 1.51]	0.21 [0.00, 0.66]	1.52 [1.44, 1.60]	0.283 [0.265, 0.300]	0.374 [0.353, 0.396]
Estimate Non-overlap	3.66 [3.58, 3.77]	12.99 [11.55, 14.35]	1.80 [1.67, 1.91]	0.441 [0.424, 0.458]	0.596 [0.573, 0.622]
Evidence Overlap	1.49 [1.34, 1.63]	0.00 [0.00, 0.00]	1.55 [1.48, 1.63]	0.292 [0.276, 0.307]	0.385 [0.362, 0.408]
Evidence Non-overlap	3.71 [3.62, 3.84]	12.51 [11.12, 13.82]	1.82 [1.67, 1.95]	0.460 [0.443, 0.478]	0.593 [0.568, 0.622]

Table 21: Nearest-neighbour audit of both teachers. Distances use room-normalized position and unit label against retained training variants. Each teacher reports request count n and HMX response error (cm, joint 95% interval). Fixed bins, empty bins and all outcomes are retained. Proximity is descriptive and does not define an unseen function.

Treatment	Distance	n_E	Error E (cm)	n_C	Error C (cm)
Standard	$[0, 0.05)$	230	32.2 [17.4, 52.2]	250	43.0 [33.1, 52.5]
Standard	$[0.05, 0.10)$	20	33.8 [15.8, 56.2]	60	31.5 [24.4, 38.7]
Standard	$[0.10, 0.20)$	15	40.6 [21.9, 63.6]	50	32.2 [22.4, 42.7]
Standard	$[0.20, \infty)$	335	92.0 [84.5, 99.8]	240	118.4 [111.2, 123.3]
Full-label	$[0, 0.05)$	230	10.7 [8.4, 13.0]	250	12.4 [9.8, 14.9]
Full-label	$[0.05, 0.10)$	20	13.0 [5.9, 23.2]	60	29.3 [22.8, 35.9]
Full-label	$[0.10, 0.20)$	15	21.4 [12.8, 32.5]	50	36.9 [26.8, 47.1]
Full-label	$[0.20, \infty)$	335	98.2 [92.9, 104.7]	240	122.2 [117.4, 126.1]
Same-label	$[0, 0.05)$	522	46.7 [41.0, 53.1]	533	22.4 [19.7, 25.3]
Same-label	$[0.05, 0.10)$	78	45.2 [28.8, 63.7]	67	31.8 [24.7, 40.2]
Same-label	$[0.10, 0.20)$	0	–	0	–
Same-label	$[0.20, \infty)$	0	–	0	–
Empty-map filter	$[0, 0.05)$	496	61.3 [51.1, 73.2]	511	27.5 [22.6, 32.5]
Empty-map filter	$[0.05, 0.10)$	104	31.3 [24.4, 38.7]	89	36.0 [25.2, 50.5]
Empty-map filter	$[0.10, 0.20)$	0	–	0	–
Empty-map filter	$[0.20, \infty)$	0	–	0	–
Strict filter	$[0, 0.05)$	489	93.5 [86.1, 100.3]	511	43.0 [38.5, 47.7]
Strict filter	$[0.05, 0.10)$	111	42.8 [29.5, 57.7]	89	30.6 [24.7, 38.4]
Strict filter	$[0.10, 0.20)$	0	–	0	–
Strict filter	$[0.20, \infty)$	0	–	0	–

Table 22: Convex-hull distance of each student HMX output from the teacher Full/H/M/X outputs. E/C denote Estimate/Evidence supervision. The primary analysis uses decoded bounded positions; the pre-clamp unit-label analysis is a registered follow-up. Weights are nonunique whenever multiple optimal simplex representations differ. Averaged representative weights in the result files are descriptive; this projection does not establish a learned mixture mechanism.

Teacher / treatment	Position (cm)	Unit label	Nonunique (%)
E / Standard	26.7 [16.9, 39.9]	0.411 [0.260, 0.614]	80.4
E / Full-label	12.5 [10.9, 14.1]	0.192 [0.167, 0.216]	83.9
E / Same-label	29.9 [23.7, 36.5]	0.460 [0.364, 0.562]	78.0
E / Empty-map filter	51.2 [42.1, 61.2]	0.787 [0.647, 0.941]	58.0
E / Strict filter	81.2 [75.1, 86.2]	1.250 [1.155, 1.326]	65.2
C / Standard	33.7 [24.8, 42.6]	0.518 [0.381, 0.656]	57.1
C / Full-label	8.2 [6.8, 9.8]	0.127 [0.105, 0.151]	83.7
C / Same-label	17.6 [15.0, 20.2]	0.271 [0.231, 0.311]	57.5
C / Empty-map filter	24.2 [20.0, 28.3]	0.372 [0.308, 0.436]	44.4
C / Strict filter	38.7 [34.4, 43.4]	0.596 [0.530, 0.668]	45.2

Table 23: Achieved target-coordinate prediction from public inputs. Ridge penalties and the large set-probe checkpoint are selected on validation layouts. The latter uses every recorded return, a shared 128–256 encoder, mean/max pooling, and a 256–256–256 head, with five fits. These are measured predictive references, not information-theoretic ceilings; negative values remain visible.

Public input	Readout	Fits	Target R^2
Public estimate	Ridge	1	-0.027 [-0.526, 0.169]
Evidence centroid	Ridge	1	0.139 [-0.211, 0.268]
Complete return history + pose	Large set probe	5	0.810 [0.663, 0.881]

Table 24: Ridge train–test matrix for all three coordinate labels. Cells are test R^2 with joint 95% intervals. E/C denote Estimate/Evidence teachers. Readouts fit training layouts with validation selection and predict the label of the original Full record. Pose–step and constant controls are masking-invariant. Negative scores are retained as extrapolation failures, not calibrated beliefs.

Policy / teacher	Readout train → test	Estimate	Centroid	Target
MLP / E	Full→Full	0.98 [0.97, 0.98]	0.70 [0.57, 0.79]	0.37 [0.09, 0.49]
MLP / E	Full→Blind	-4.59 [-6.68, -3.12]	-12.06 [-19.02, -7.35]	-27.67 [-43.21, -20.12]
MLP / E	Blind→Full	-1629.57 [-2633.41, -874.96]	-8270.27 [-21424.96, -1722.27]	-186.00 [-310.80, -112.23]
MLP / E	Blind→Blind	-0.06 [-0.28, 0.07]	-0.30 [-0.71, -0.05]	-0.00 [-0.31, 0.07]
MLP / E	Pose + step	-0.27 [-0.53, -0.12]	-0.61 [-1.08, -0.33]	-0.04 [-0.38, 0.03]
MLP / E	Constant	-0.58 [-0.94, -0.36]	-1.18 [-1.83, -0.78]	-0.05 [-0.40, -0.00]
Set / E	Full→Full	0.75 [0.62, 0.85]	0.77 [0.71, 0.82]	0.18 [-0.12, 0.29]
Set / E	Full→Blind	-4.71 [-7.71, -2.64]	-15.36 [-26.00, -8.10]	-38.48 [-76.33, -19.21]
Set / E	Blind→Full	-13.41 [-21.22, -7.23]	-41.89 [-91.53, -13.92]	-6.84 [-15.65, -2.03]
Set / E	Blind→Blind	-0.05 [-0.26, 0.07]	-0.32 [-0.71, -0.08]	-0.02 [-0.37, 0.06]
Set / E	Pose + step	-0.27 [-0.53, -0.12]	-0.61 [-1.08, -0.33]	-0.04 [-0.38, 0.03]
Set / E	Constant	-0.58 [-0.94, -0.36]	-1.18 [-1.83, -0.78]	-0.05 [-0.40, -0.00]
Set / C	Full→Full	0.68 [0.54, 0.79]	0.87 [0.82, 0.90]	0.20 [-0.10, 0.30]
Set / C	Full→Blind	-11.56 [-24.02, -2.92]	-10.07 [-17.04, -4.86]	-26.10 [-52.48, -9.24]
Set / C	Blind→Full	-50.96 [-98.02, -16.26]	-190.91 [-421.23, -56.60]	-10.88 [-19.46, -6.47]
Set / C	Blind→Blind	-0.07 [-0.27, 0.05]	-0.31 [-0.68, -0.08]	-0.01 [-0.35, 0.07]
Set / C	Pose + step	-0.27 [-0.53, -0.12]	-0.61 [-1.08, -0.33]	-0.04 [-0.38, 0.03]
Set / C	Constant	-0.58 [-0.94, -0.36]	-1.18 [-1.83, -0.78]	-0.05 [-0.40, -0.00]

Table 25: Behavioral distances of the Full-trained coordinate readout evaluated on Blind activations. Entries are metres with joint intervals. The three comparators are the student bounded action endpoint, the teacher bounded Blind endpoint, and the actual empty-map frontier waypoint. None is equated with an internal belief.

Policy / decoded label	Student endpoint	Teacher endpoint	Frontier waypoint
MLP / E / estimate	1.40 [1.11, 1.74]	1.59 [1.20, 2.02]	3.27 [3.11, 3.44]
MLP / E / centroid	1.69 [1.32, 2.01]	1.77 [1.40, 2.13]	3.24 [3.10, 3.40]
Set / E / estimate	2.42 [1.79, 3.25]	2.62 [1.97, 3.52]	3.55 [3.06, 4.15]
Set / E / centroid	3.16 [2.33, 4.06]	3.28 [2.35, 4.28]	3.87 [3.24, 4.58]
Set / C / estimate	3.83 [3.05, 4.85]	3.90 [3.11, 4.86]	4.04 [3.27, 5.08]
Set / C / centroid	2.29 [1.56, 3.09]	2.41 [1.52, 3.36]	3.21 [3.01, 3.46]

Table 26: Association of the extrapolated public-estimate readout with Blind behavior. Coordinate columns report Pearson correlation; direction reports cosine from the current pose. Joint intervals resample fit, layout and noise. The shared pose and step can explain associations; these statistics do not identify a causal state variable. A dash denotes an unreported frontier-direction comparison.

Policy / teacher	Comparator	x correlation	y correlation	Direction cosine
MLP / E	Student endpoint	0.623 [0.417, 0.857]	0.314 [0.153, 0.600]	0.521 [0.207, 0.795]
MLP / E	Teacher endpoint	0.330 [0.252, 0.462]	0.307 [0.200, 0.584]	0.253 [0.196, 0.324]
MLP / E	Frontier waypoint	-0.345 [-0.463, -0.287]	0.027 [-0.036, 0.120]	–
Set / E	Student endpoint	0.106 [-0.083, 0.203]	-0.139 [-0.323, 0.035]	0.612 [0.441, 0.759]
Set / E	Teacher endpoint	0.016 [-0.018, 0.056]	-0.106 [-0.308, 0.041]	0.308 [0.195, 0.408]
Set / E	Frontier waypoint	0.075 [0.043, 0.141]	-0.017 [-0.072, 0.027]	–
Set / C	Student endpoint	0.104 [-0.033, 0.228]	0.007 [-0.064, 0.077]	0.303 [-0.134, 0.723]
Set / C	Teacher endpoint	0.039 [0.010, 0.142]	0.025 [-0.009, 0.082]	0.197 [0.114, 0.272]
Set / C	Frontier waypoint	0.007 [-0.053, 0.036]	0.006 [-0.008, 0.028]	–

F Response, Memory and Feature-Sufficiency Controls

For student and teacher decoded responses $b = g_B(Ir) - g_B(r)$ and $t = g_T(Ir) - g_T(r)$, sensitivity is $S_B = \mathbb{E}\|b\|$ or $S_T = \mathbb{E}\|t\|$. Response error is $\mathbb{E}\|b - t\|$. We report E_{resp}/S_T as an aggregate ratio and $E_{\text{resp}} - E_{\text{act}}$ as a diagnostic difference, not a debiased estimand. Direction cosines require both vector norms above 10^{-8} m and retain eligible counts. Ratios are recomputed inside each bootstrap replicate.

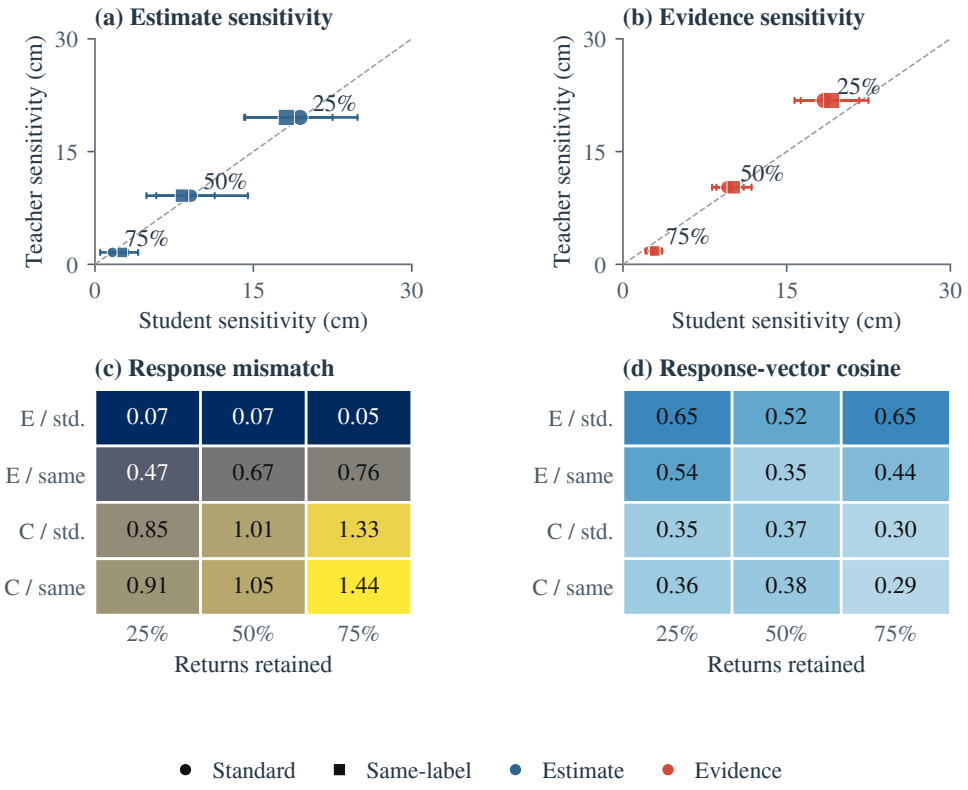


Figure 15: Coherent thinning rebuilds history, map and estimate along the same acquisition path. Glyph size identifies retention; circles are Standard, squares Same-label. Horizontal intervals quantify student sensitivity. Magnitude agreement on the diagonal does not imply directional agreement; response-error/teacher-sensitivity ratios and cosine are shown separately. Complete intervals and denominator rules remain in tables.

The GRU frozen comparison first replays Full observations to obtain an incoming state, then deletes current evidence with either retained or reset memory. Sustained blind episodes instead evolve on their own blind requests, with no Full-prefix information imported. These definitions yield different estimands. The stateless teacher's fallback is a reference action, not an optimal target for a memory-bearing student. Sustained retained-minus-reset blind success has a wide interval and substantial fit variation; it does not identify a robust benefit caused by recurrent architecture.

The feature-sufficiency control relabels the common corpus with Evidence actions. A compressed-feature MLP and a matched MLP augmented with the public centroid's relative direction and presence flag share seeds, optimizer and checkpoint rule. The added variable comes from authorized scans, not evaluator truth. Its improved fit is an achievable reference with more sufficient features, not an oracle error bound.

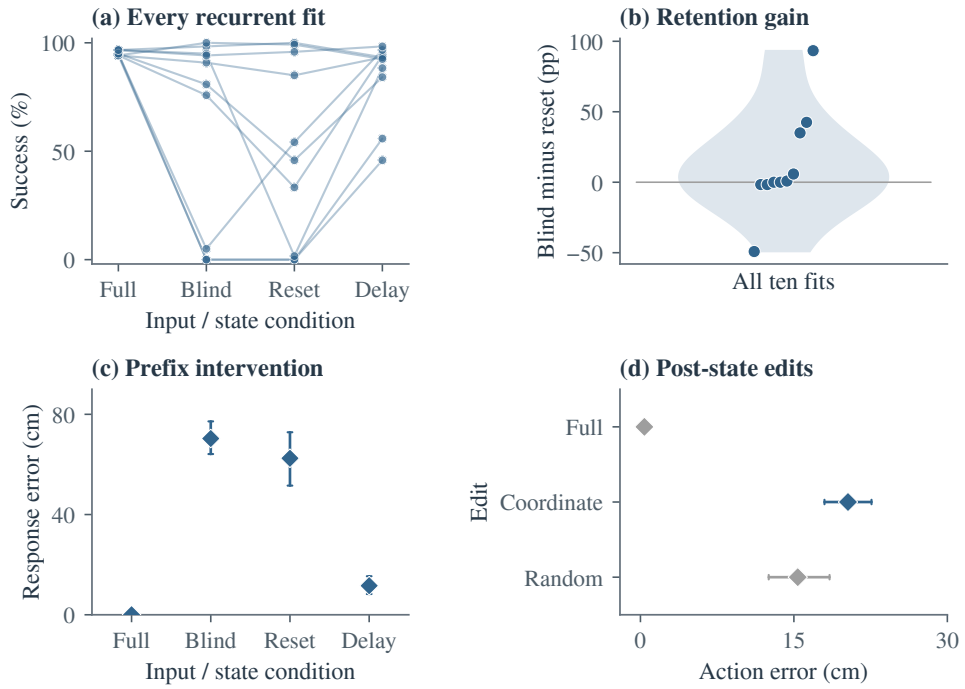


Figure 16: Recurrent-policy heterogeneity retains all ten fits and negative paired gains. The KDE is descriptive. Frozen-prefix input interventions and edited post-GRU states differ from sustained Blind rollouts. The edited post-state directly feeds a linear head; unmatched architecture comparisons do not isolate an effect of memory.

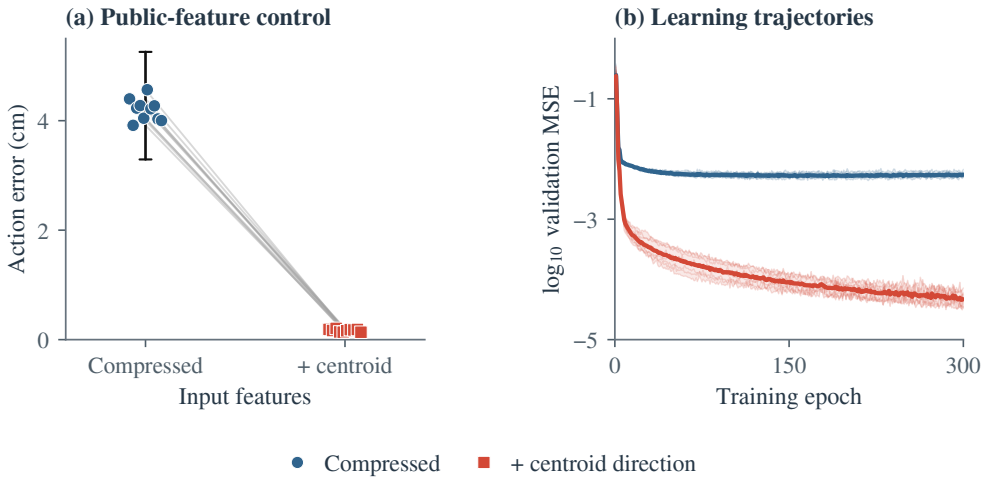


Figure 17: The added centroid direction and presence derive from authorized public observations. Paired fit errors, joint intervals, all validation curves, and full fit ranges are retained. This is an achievable feature-information control, not evaluator-truth access or an irreducible-error bound.

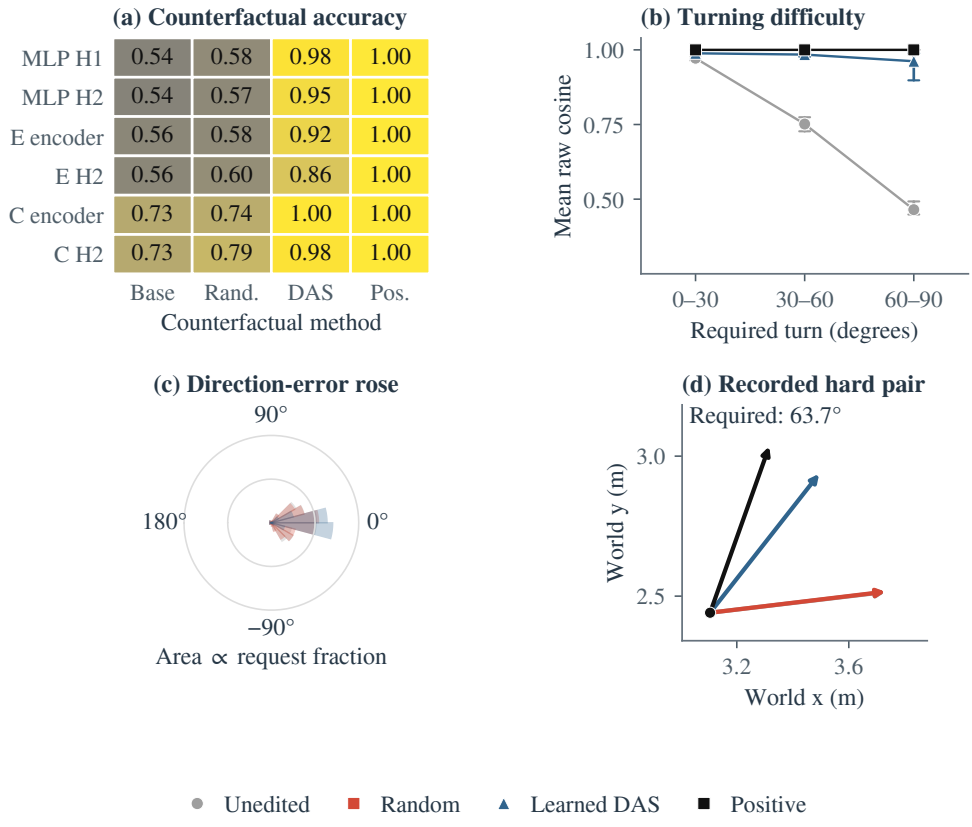


Figure 18: Interchange on its recorded donor distribution. (a) Pass rates at cosine > 0.95. (b) Raw cosine by required turn angle for the Estimate Set–CNN encoder. (c) Signed direction-error distributions; sector area represents frequency. (d) The maximum-angle eligible pair for fit 11, used because the frozen $> 90°$ criterion had no eligible pair. Bounded endpoints share the recipient pose; the pair is not a newly sampled goal.

G Spatial Readouts Across Acquisition Steps

The projection uses a training-fitted H2 PCA basis and all scored requests for the fixed displayed fit. Color represents the measured teacher-relative angle or public estimate coordinate; no ring or other geometric structure is imposed. Existing frozen probe predictions are stratified by acquisition step without fitting to test labels. Complete-history and H2 readouts have different capacities and inputs and are therefore achievable references, not matched information-theoretic bounds.

H Geometry and a Prospective Evidence-Necessity Pilot

The D4 audit pairs 24 original layouts with 24 transformed layouts and 96 confirmation layouts with 96 transformed partners. Each base layout receives one prescribed transformation; this supports an aggregate paired comparison, not a separate causal effect for each rotation or reflection. Start–target distances, walls and occluder counts agree under transformation. A world-fixed registration vector is transformed only in the explicitly named calibration control.

A random controller keyed to five acquisition seeds reuses five world-direction streams across all layouts. Their scored direction sums all point south, so this finite route set need not have rotation-invariant performance. Separating policy randomness by base layout yields original success 30.2% and D4 success 28.8%, a paired change of -1.5 percentage points $[-13.3, 10.6]$. Rotating routes with scenes checks path equivariance; jointly rotating registration separates calibration convention.

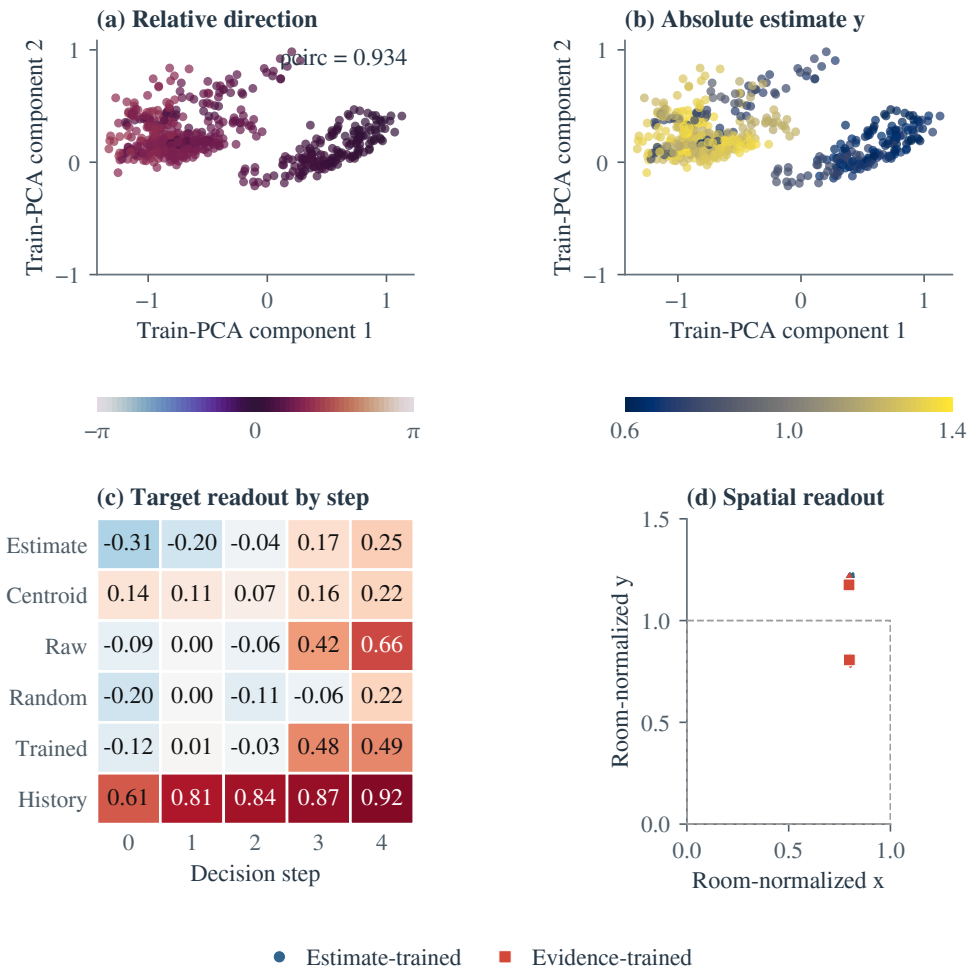


Figure 19: One fixed Set–Estimate fit (seed 11), all test requests, and PCA fitted only on training activations. Panel (a) color is the relative direction angle in radians; panel (b) color is room-normalized absolute estimate y. Circular-linear correlation is descriptive. Target R^2 by step compares raw/random/trained 64×64 probes with a richer, larger full-history reference, not an isolated capacity-matched information-loss test. Spatial arrows show true public estimate to decoded position for all five decisions of layout 3000/noise 500/seed 11; out-of-room estimates are retained.

This audit explains the random baseline discrepancy without selecting a geometric subset from outcomes.

The evidence-necessity pilot freezes an unchanged randomized-endpoint generator on 32 fresh validation layouts, 12000–12031, with five noise seeds. Four fixed world routes (northeast, northwest, southwest and southeast) and both teachers under Full and strict blindness produce 1,280 episodes. Success criteria were fixed before any pilot outcome: the best fixed route must be below 30%, and both teachers must improve by at least 20 percentage points over their Blind fallback with positive paired interval lower bounds. No alternative generator was selected after observing performance.

See Figure 10 for all retained pilot outcomes.

Both teacher contrasts satisfy their improvement criterion, but the fixed-route criterion fails. The result demonstrates the usefulness of evidence relative to these fallbacks while leaving evidence necessity against open-loop routes unestablished. Endpoint randomization alone is therefore insufficient to certify the stronger task claim. Adding observation range gates or a success-on-approach objective would change the task and require an independently validated protocol.

Table 27: Public-estimate diagnostics by stored, zero-based decision step. All 120 requests per step are retained. Every present estimate is finite; tested pose, room-center and origin defaults are absent. Direct denotes a direct-path return in the public history. At step 4, 109 requests with direct returns have mean error 0.067 m; the remaining 11 have mean error 1.855 m. Out-of-room estimates and failure tails remain in all means.

Step	Present (%)	Outside room (%)	Direct in history (%)	Mean estimate error (m)
0	100.0	58.3	0.0	1.808
1	100.0	81.7	0.0	2.277
2	100.0	91.7	4.2	2.486
3	100.0	54.2	45.8	1.431
4	100.0	9.2	90.8	0.230

Table 28: Target prediction by stored, zero-based decision step using the same frozen pooled readouts. Each step has 120 requests and its own recomputed R^2 denominator. Raw and random H2 use the Estimate group. Their nonlinear probes and both trained-H2 probes use two 64-unit layers; the estimate-only probe uses ridge. Full history has richer observations and larger readout capacity, so it is an achievable reference, not a matched information ceiling. The machine-readable summary retains joint intervals and all negative scores.

Step	Estimate only	Raw	Random H2	Set–E H2	Set–C H2	Full history
0	-0.305	-0.086	-0.204	-0.115	-0.162	0.614
1	-0.201	0.005	0.005	0.014	0.209	0.810
2	-0.041	-0.057	-0.111	-0.026	0.182	0.842
3	0.168	0.423	-0.056	0.479	0.490	0.870
4	0.245	0.660	0.223	0.486	0.485	0.916

Table 29: Unchanged validation-only evidence-necessity pilot. All 32 layouts and five noise seeds are retained for each of eight conditions. Success and paired layout/noise 95% confidence bounds are percentages.

Policy / condition	Episodes	Success	CI lower	CI upper
Fixed northeast	160	66.2	50.6	81.2
Fixed northwest	160	56.2	40.6	71.9
Fixed southwest	160	61.2	44.4	76.9
Fixed southeast	160	72.5	56.9	86.9
Estimate Full	160	90.6	82.5	97.5
Estimate Blind	160	41.9	25.0	59.4
Evidence Full	160	87.5	75.6	96.9
Evidence Blind	160	41.9	25.0	59.4

I Reproduction, Reporting and Execution Boundaries

The artifact provides scripts, corpora, selected policy/probe weights, fitted erasure and alignment maps, per-request outputs and bootstrap summaries. The publication builder selects rows from retained CSVs and generates numerical macros and tables. The manifest records source files, exact selectors, generating commands, code content hashes and the local source revision. Hashes distinguish the code snapshot used for an execution from later reporting edits. The artifact audit checks numerical consistency, coverage and mathematical invariants, rather than claiming an independent replication of all training.

The local analysis plan is timestamped before execution. Separate timestamped addenda distinguish the algebra-motivated independent-readout audit, the matched-centroid native comparison,

Table 30: Frozen-policy confirmation on 48 additional layouts (1,200 requests). Native contrasts use 50% measurement retention. Joint intervals resample fits, layouts and sensor noise.

Contrast	Value	Joint 95% interval
Linear D: original cohort	0.134	[0.099, 0.184]
Linear D: additional cohort	0.146	[0.102, 0.242]
Linear D: pooled cohort	0.140	[0.107, 0.208]
Nonlinear D: original cohort	0.071	[0.049, 0.120]
Native D: covariance, raw	0.056	[0.026, 0.095]
Native D: covariance, direction	0.062	[0.024, 0.101]
Native D: covariance, bounded position	0.062	[0.023, 0.101]
Native D: LEACE, raw	0.582	[0.458, 0.741]
Native D: LEACE, direction	0.541	[0.417, 0.696]
Native D: LEACE, bounded position	0.542	[0.418, 0.697]

the eraser/native bridge, the unit-label hull and the exploratory native selectivity contrast. Their chronology and outcome-motivated origins remain in the artifact. These records are local protocol freezes, not external preregistration. Numerical reporting rounds after aggregation using decimal round-half-even; paired differences are calculated before display rounding.

All requested policy seeds, candidate interchange ranks and failed episodes are retained. The test set does not select checkpoints, erasers, probe penalties or alignment rank. Core policy representations and the added null cohort are evaluated on a reused development cohort. The fresh pilot is a validation diagnostic and did not establish a task family requiring evidence. A second domain, stronger matched memory teacher, nonlinear forgetting guarantee, physical radar and signal-level sensing remain unexecuted research directions.

Figures 1, 3 and 4 preserve the legacy vector artwork. Their embedded client and waveform examples are not learning experiments, and their dense labels remain smaller than the body text. Experimental plots use vector exports at their native printed width with 9-point ordinary text. The figure-data specification and plotting source preserve the plotted values and units. The AI disclosure and authors' responsibility apply to the complete manuscript and artifact.

J Frozen Confirmation on Additional Layouts

The confirmation uses original-family occluded layouts 13000–13047 and noise seeds 500–504. Full Estimate teacher episodes supply five scored decisions each, yielding 240 episodes and 1,200 requests. Policies, probes, erasers, retention rules and statistical contrasts remain fixed. The comparison retains all requests and outcomes, including boundary clamps. Main statistics are pooled linear selectivity, the primary-cohort nonlinear interaction and the native teacher–variable interaction at 50% retention; 25% and 75% are reported alongside them.

The artifact preserves the original design and the chronology of deterministic re-execution. Repeating an inspected analysis is not counted as another independent confirmation, and the later timestamps do not establish new blinding or external preregistration. Predictions and statistics are regenerated from the source requests and frozen models. These layouts extend the same-family evaluation beyond policy/probe fitting; they do not constitute a second domain or an evidence-necessary task.

Table 31: Native teacher–variable interaction point estimates at every retained-measurement rate on the additional layouts. Full intervals are in the released CSV. Signed projection contrasts are not causal shares.

Operator	Retained	Raw	Direction	Bounded
Covariance	25%	0.054	0.056	0.059
Covariance	50%	0.056	0.062	0.062
Covariance	75%	0.016	0.020	0.020
LEACE	25%	0.691	0.645	0.660
LEACE	50%	0.582	0.541	0.542
LEACE	75%	0.442	0.507	0.507

Table 32: Bounds-clamp counts at 50% retention on the additional layouts. Each row contains 12,000 fit–request pairs.

Policy / component	Base	Full	Selected	Complement
MLP / relative ridge	100	100	100	100
Set–E / estimate LEACE	100	100	100	100
Set–C / centroid LEACE	165	180	185	183

K Output-head Audit: Definitions, Controls and New Results

The new analyses are reviewer-motivated follow-ups to the retained evidence in Appendices A–J. The primary new-layout comparison evaluates frozen policies on layouts 14000–14047; the original 24-layout set remains a development cohort. No historical result is relabeled as a new training run or as independent external replication. All numeric tables in this appendix are generated from saved result files.

K.1 Output consistency and the native coordinate convention

Let $A : \mathbb{R}^d \to \mathbb{R}^2$ be the fixed linear head and $Q : \mathbb{R}^d \to \mathbb{R}^d$ a selected update map, not necessarily orthogonal. With the Euclidean metric in the network's native hidden coordinates, $R = A^+A$ projects onto the head row space and $N = I - R$ onto its kernel. The condition

$$AQN = 0 \quad \Longleftrightarrow \quad \exists T \in \mathbb{R}^{2\times 2} \text{ such that } AQ = TA \tag{6}$$

is necessary and sufficient for the selected raw output to be a function only of the complete raw output change. For sufficiency, $AQ\delta h = TA\delta h$. For necessity, if $A\delta h_1 = A\delta h_2$ always entails $AQ\delta h_1 = AQ\delta h_2$, then AQ annihilates $\ker A$, hence $AQ = AQR = AQA^+A$. Choose $T = AQA^+$. This criterion does not require a learned concept to be linearly decodable and does not establish semantic fidelity when it holds.

The null-mixing quantity is $AQN\delta h$, not $AN\delta h$, which is zero. Our row and null decomposition depends on the chosen hidden metric. We do not claim invariance to arbitrary non-orthogonal reparameterizations or interpret its terms as causal mediation shares. The scripts check both identities before aggregation. The largest mean linear closure residual across reported conditions is below 9×10^{-16}; exact per-request vectors were reported as retained or deterministically regenerated in the supplied study.

K.2 An analytic negative control

This example is a population-covariance calculation, not a radar simulation. Let $h = (u_x, u_y, z)$ have unit marginal variances, $\mathrm{Cov}(u_y, z) = \rho$ and all other cross-covariances zero. The action head reads only u_x, u_y:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad C = \mathrm{Cov}(h, z) = \begin{bmatrix} 0 \\ \rho \\ 1 \end{bmatrix}, \quad \delta h = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{7}$$

For $0 \le \rho < 1$, the covariance-orthogonal selected map is $Q_C = CC^\top/(C^\top C)$ and the LEACE selected map is $Q_L = Ce_3^\top$. Thus $A\delta h = 0$, but $AQ_C\delta h = (0, \rho/(1+\rho^2))^\top$ and $AQ_L\delta h = (0, \rho)^\top$. Each complement cancels its selected raw response. At baseline raw output $(1, 0)^\top$, separately normalizing these components produces nonzero finite movements and a nonadditive residual. A signed fraction relative to the zero complete raw response is undefined and is not reported.

K.3 Sequential maps and random oblique controls

For every primary Set–CNN fit, we fit a direction LEACE eraser P_d to the actual teacher loss labels on the original training requests. The residual-coordinate eraser is fitted to $(H - \mu)P_d^\top$ with the same coordinate labels. The selected increment is $Q_{z|d} = (I - P_{z|d})P_d$. Reversing the order instead yields $Q_{d|z} = (I - P_{d|z})P_z$. Each composition is evaluated on the same native δh, through the frozen head and finite decoder. Neither sequential increment must itself be an orthogonal projector. Small coordinate increments do not establish nonlinear forgetting or exclude coordinate computations that act through the erased direction.

Sixteen seeded random orthogonal matrices O_j are generated before new-layout evaluation. The control map $Q_j = O_jQ_LO_j^\top$ has the same rank, idempotence, singular values, spectral norm and Frobenius norm as Q_L. It does not generally satisfy the original coordinate covariance constraint, and its realized $\|Q_j\delta h_i\|$ need not match $\|Q_L\delta h_i\|$. The controls isolate orientation relative to the fixed network, not a permutation test of a semantic null. The same control index is pooled across fits

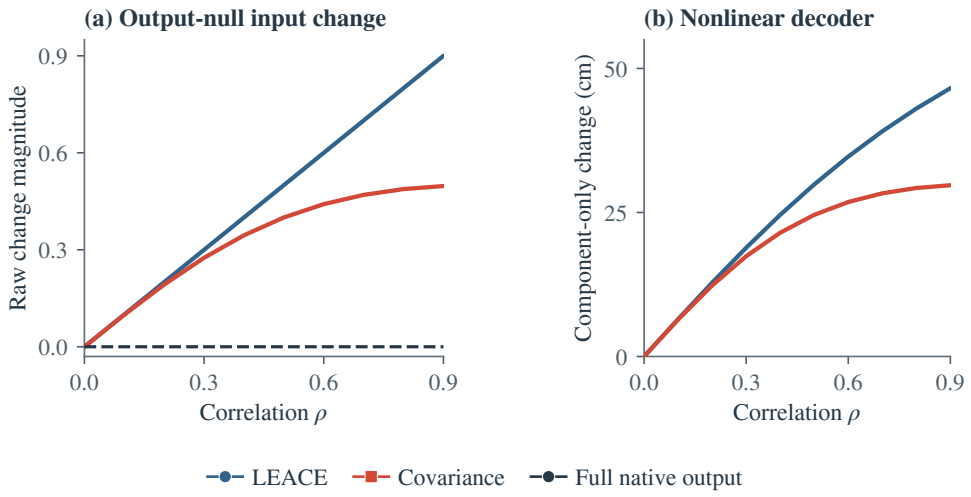


Figure 20: Analytic negative control. Changing only z is initially output-null, yet erasing its correlated direction creates a selected output. The complementary raw change cancels it exactly. Finite component-only movement follows the actual normalized 0.65 m decoder, without a room clamp in this analytic example.

for Figure 7; controls are not counted as sixteen independent trained models. Existing per-request norm-matched perturbations answer a different strength-control question.

The raw-output positive control uses LEACE with $z = Ah + b$ as its erasure label. On covariance support, this selected map satisfies $AQ = A$; its exact recovery is an algebraic implementation check. A covariance constraint alone does not imply the same output recovery for an arbitrary eraser, and the positive control does not establish a unique mechanism for an independently defined coordinate.

Table 33: New-layout action-head audit at 50% retention. Raw projections divide by complete raw-response energy: row plus null equals selected before rounding. Movement is measured from the same unedited endpoint after decoding each component separately; these magnitudes are not additive. Cells show estimates and joint fit/layout/noise 95% intervals.

Policy / map	Selected	Row	Null
Signed raw-response projections			
Set–E / Covariance	0.817 [0.759, 0.884]	0.609 [0.554, 0.677]	0.207 [0.161, 0.266]
Set–E / LEACE	0.940 [0.820, 0.999]	0.368 [−0.516, 1.280]	0.572 [−0.360, 1.460]
Set–C / Covariance	0.906 [0.885, 0.942]	0.549 [0.501, 0.609]	0.357 [0.303, 0.410]
Set–C / LEACE	0.895 [0.797, 0.960]	0.605 [−0.298, 1.639]	0.290 [−0.743, 1.193]
Finite component-only movement (cm)			
Set–E / Covariance	6.47 [4.93, 8.32]	4.75 [3.68, 6.05]	1.71 [1.21, 2.32]
Set–E / LEACE	6.56 [4.83, 8.71]	9.59 [6.71, 12.63]	9.82 [6.63, 13.57]
Set–E / After direction	0.16 [0.11, 0.22]	0.53 [0.32, 0.78]	0.62 [0.40, 0.88]
Set–C / Covariance	6.92 [5.43, 8.67]	4.11 [3.21, 5.19]	2.69 [2.04, 3.37]
Set–C / LEACE	6.27 [4.73, 8.16]	8.21 [4.47, 12.94]	9.36 [6.72, 12.39]
Set–C / After direction	0.06 [0.04, 0.09]	0.49 [0.23, 0.79]	0.50 [0.24, 0.81]

Table 34: Sequential and positive controls on new layouts at 50% retention. Cells show signed projections and joint 95% intervals. Sequential rows report increments under the named order, not marginal causal effects; raw-output erasure is an algebraic positive control.

Selected increment	Set–E Raw	Set–E Bounded	Set–C Raw	Set–C Bounded
Coordinate	0.940 $[0.820, 0.999]$	1.024 $[0.873, 1.117]$	0.895 $[0.797, 0.960]$	0.929 $[0.796, 1.011]$
Teacher direction	0.994 $[0.956, 1.018]$	1.029 $[1.006, 1.078]$	1.010 $[0.985, 1.036]$	1.013 $[0.992, 1.038]$
Coordinate after direction	-0.006 $[-0.012, -0.001]$	-0.006 $[-0.013, -0.001]$	-0.001 $[-0.005, 0.002]$	-0.001 $[-0.004, 0.001]$
Direction after coordinate	0.048 $[-0.007, 0.147]$	0.103 $[0.015, 0.159]$	0.115 $[0.055, 0.199]$	0.138 $[0.071, 0.200]$
Raw output	1.000 $[1.000, 1.000]$	1.000 $[1.000, 1.000]$	1.000 $[1.000, 1.000]$	1.000 $[1.000, 1.000]$

K.4 Retention levels and finite movement

The 25%, 50% and 75% interventions retain a fixed ceiling-rounded subset of each recorded scan. Reconstructed public state is used before computing hidden differences. For Set–CNN, estimate and centroid fields are not directly supplied as input coordinates. For the recurrent development control, only the current request changes, holding incoming Full-prefix state fixed.

Alongside raw selected, row and null vectors, the outputs store separate normalized directions and bounded endpoints. Two residuals have different meanings: $d_Q + d_{I-Q} - d$ closes the selected/complement split, while $d_{\mathrm{row}} + d_{\mathrm{null}} - d_Q$ closes the row/null split inside the selected response. Neither is silently discarded. Tables report component magnitudes without adding them, and the machine-readable results retain the raw cross term $2v_{\mathrm{row}}^{\top} v_{\mathrm{null}}$, component energies, hidden-update norms and boundary indicators. Positive, negative and greater-than-one signed projections are kept.

L New-layout Cohort and Nonlinear Action References

L.1 Cohorts and chronology

The original experiment uses training layouts 1000–1047 and validation layouts 2000–2011. The 600-request cohort from 3000–3023 has been reused in method development. The earlier frozen comparison used layouts 13000–13047, already reported in Appendix J. The present additional cohort uses 14000–14047, with noise seeds 500–504, six scans and five decisions per episode: all 240 episodes and all 1,200 requests are retained.

A local, timestamped plan fixed the new layout IDs, acquisition policy, requested contrasts and all retention rates before their generation. This is not external preregistration. Full Estimate teacher trajectories supply the recorded requests, so the cohort shares a policy-conditioned acquisition distribution with the original study. No new evidence-necessity generator was constructed. Policy weights, existing linear/nonlinear probes and geometry maps are frozen; the new audit does not perform policy checkpoint selection on these layouts.

Forty Set–CNN policies enter the pooled linear teacher comparison. The original and additional ten-seed cohorts are resampled independently along the fit axis, with shared layout/noise resamples. Twenty original-cohort Set–CNN policies enter the native geometry and historical nonlinear-probe confirmation. The MLP and GRU geometry extensions retain their original development requests rather than being silently counted as additional new-layout policies.

L.2 Gaussian-landmark reference specification

This reference was added after inspecting other new-layout diagnostics and remains exploratory. Each input is standardized with training means and standard deviations. We select 256 equally

Table 35: New-layout sensitivity across every retention rate. Cells show signed projections and joint 95% intervals. After dir. is the incremental own-coordinate update after teacher-direction erasure. Raw null is the output-null component of that map; bounded selected uses the actual endpoint decoder.

Policy / update	Projection	25% retained	50% retained	75% retained
Set–E / Cov.	Raw null	0.205 $[0.168, 0.252]$	0.207 $[0.161, 0.266]$	0.195 $[0.144, 0.254]$
	Bounded selected	0.923 $[0.852, 0.983]$	0.964 $[0.890, 1.026]$	0.959 $[0.827, 1.046]$
Set–E / LEACE	Raw null	0.562 $[-0.296, 1.437]$	0.572 $[-0.360, 1.460]$	0.682 $[-0.348, 1.590]$
	Bounded selected	0.981 $[0.887, 1.049]$	1.024 $[0.873, 1.117]$	1.160 $[0.821, 1.230]$
Set–E / After dir.	Raw null	-0.004 $[-0.054, 0.053]$	-0.005 $[-0.056, 0.056]$	-0.017 $[-0.072, 0.046]$
	Bounded selected	-0.010 $[-0.020, -0.002]$	-0.006 $[-0.013, -0.001]$	-0.005 $[-0.011, 0.003]$
Set–C / Cov.	Raw null	0.343 $[0.294, 0.393]$	0.357 $[0.303, 0.410]$	0.361 $[0.307, 0.417]$
	Bounded selected	0.927 $[0.872, 0.976]$	0.974 $[0.909, 1.025]$	0.992 $[0.838, 1.042]$
Set–C / LEACE	Raw null	0.278 $[-0.728, 1.119]$	0.290 $[-0.743, 1.193]$	0.523 $[-0.604, 1.582]$
	Bounded selected	0.900 $[0.813, 0.971]$	0.929 $[0.796, 1.011]$	1.007 $[0.767, 1.086]$
Set–C / After dir.	Raw null	0.023 $[-0.035, 0.079]$	0.022 $[-0.037, 0.080]$	0.025 $[-0.035, 0.084]$
	Bounded selected	-0.001 $[-0.004, 0.002]$	-0.001 $[-0.004, 0.001]$	-0.000 $[-0.004, 0.002]$

spaced training request indices as landmark rows. Gaussian features are

$$\phi_j(x) = \exp\bigl(-\|\tilde{x} - \ell_j\|^2/(ms)\bigr)\,, \tag{8}$$

where s is the median non-diagonal squared distance among training landmarks and $m \in \{0.25, 1, 4\}$. Feature means and scales also use only training requests. Ridge penalties are $\{0.01, 0.1, 1, 10, 100\}$; validation coordinate MSE selects the bandwidth and penalty separately for each label. The mean-centered two-coordinate output has a fitted intercept.

We compare H2; the four coordinates of teacher and student unit directions; and those directions concatenated with world pose and room bounds. The teacher direction is the actual pre-clamp loss label, regenerated from the public request where necessary. A direction inferred from a clamped endpoint is not substituted. The three decoded labels are public estimate, public centroid and evaluator target, normalized as in the original probes. Target labels never enter student training or action selection.

Twenty policy fits times three input sets times three labels yield 180 trained readouts. Landmark count and the selection grid match, but the induced functions and source information do not. This design provides achievable nonlinear references, not equal information content, a direct causal effect, or a formal probe-capacity equivalence theorem. The supplied study reports storing all predictions on development and new layouts, and the same paired resamples compare H2 to its action reference.

M Separating Eraser Support from Readout Support

The two sensitivity arms intentionally differ from the historical 24/24 half-layout split. In the eraser-varying arm, sixteen eraser layouts are sampled without replacement from 1000–1031 and the probe support is fixed at 1032–1047. In the probe-varying arm, the eraser is fixed at 1000–1015 while sixteen probe layouts are sampled from 1016–1047. Twenty complete index sets per arm are recorded before executing this support audit. Each fit uses 1,024 requests per side, and the two sides never overlap within an audit.

Table 36: Gaussian-landmark readouts and paired action-reference contrasts. All readouts use 256 features and the same validation grid. Actions contains teacher and student unit directions; + pose adds world pose and bounds. Cells report R^2 (or paired ΔR^2) with joint 95% intervals. The contrast is H2 minus the action-plus-pose reference; it is evaluated on new layouts only and does not identify natural coordinate use.

Label	Actions	+ pose	H2	Paired ΔR^2
Development / Estimate teacher				
Estimate	0.767 $[0.597, 0.875]$	0.942 $[0.925, 0.955]$	0.742 $[0.611, 0.845]$	–
Centroid	0.257 $[-0.077, 0.462]$	0.547 $[0.379, 0.664]$	0.738 $[0.644, 0.805]$	–
Target	0.072 $[-0.289, 0.168]$	0.068 $[-0.350, 0.245]$	0.159 $[-0.177, 0.279]$	–
Development / Evidence teacher				
Estimate	0.389 $[0.154, 0.560]$	0.654 $[0.542, 0.748]$	0.693 $[0.545, 0.809]$	–
Centroid	0.414 $[0.042, 0.642]$	0.844 $[0.779, 0.885]$	0.856 $[0.787, 0.902]$	–
Target	-0.033 $[-0.437, 0.091]$	-0.089 $[-0.583, 0.095]$	0.211 $[-0.071, 0.308]$	–
New layouts / Estimate teacher				
Estimate	0.884 $[0.783, 0.913]$	0.960 $[0.896, 0.976]$	0.900 $[0.748, 0.943]$	-0.059 $[-0.150, -0.028]$
Centroid	0.584 $[0.028, 0.641]$	0.725 $[0.308, 0.766]$	0.863 $[0.591, 0.903]$	0.138 $[0.086, 0.276]$
Target	0.084 $[-0.035, 0.138]$	0.064 $[-0.124, 0.192]$	0.079 $[-0.131, 0.198]$	0.015 $[-0.197, 0.210]$
New layouts / Evidence teacher				
Estimate	0.471 $[0.301, 0.569]$	0.646 $[0.519, 0.735]$	0.847 $[0.716, 0.898]$	0.200 $[0.067, 0.330]$
Centroid	0.721 $[0.207, 0.838]$	0.930 $[0.769, 0.962]$	0.948 $[0.769, 0.969]$	0.018 $[-0.025, 0.050]$
Target	-0.072 $[-0.273, 0.028]$	-0.053 $[-0.244, 0.055]$	0.046 $[-0.171, 0.180]$	0.098 $[-0.141, 0.328]$

When the eraser changes, keeping probe support fixed means keeping its source layouts and validation rule fixed. The probe coefficients are necessarily refitted in the transformed coordinate system. Consequently this design separates fitting-support changes, not all numerical properties of an eraser from those of a downstream readout. When only probe support changes, the same eraser matrix and held-out transformed activations are reused. Identity is paired within every support condition.

The released fit-level records include training covariance residual, held-out covariance norm, perturbation RMS, operator Frobenius norm, principal angles, validation penalty, a covariance-regularized Mahalanobis diagnostic, and squared-loss components. The orientation summary compares the selected-map row space to a fixed reference for the same trained policy; it is not an angle measured in a two-dimensional PCA display. Overlapping support resamples do not supply independent environments or confidence intervals for a population of tasks.

N Backend Calibration and Architecture-to-Code Map

The legacy backend is packaged separately from the current learning backend. The default calibration evaluates three scenario modes for seeds 40–49. Stress calibration evaluates the four fixed configurations in Table 1 under those same seeds. The supplied study reports retaining all seventy results and traces; the available package contains the plotted summaries but not those v16 trace files. A script compares all 28 displayed numeric stress cells with the supplied manuscript table; every cell agrees at its stated precision. The original plotting script's hard-coded fallback arrays and synthetic visual point spreading are not used to create the new calibration figure.

Table 37: Independent-support squared-loss diagnostics. Cells are median [range] of pooled fit means over twenty support resamples, in room-normalized coordinates. Within each fit, total loss equals squared mean residual plus centered residual energy; separate medians need not add. Centered residual energy is not irreducible noise.

Policy / varied support	Total MSE	Mean bias squared	Centered residual
Set–E / eraser	0.124 [0.112,0.136]	0.045 [0.032,0.054]	0.080 [0.069,0.088]
Set–E / probe	0.115 [0.099,0.130]	0.031 [0.008,0.047]	0.084 [0.076,0.093]
Set–C / eraser	0.066 [0.058,0.077]	0.032 [0.024,0.040]	0.035 [0.028,0.042]
Set–C / probe	0.064 [0.050,0.080]	0.030 [0.010,0.038]	0.035 [0.029,0.043]

The runtime lacked the optional spatial-index library. An exact linear AABB broad phase supplies candidate triangles to the unchanged ray–triangle intersection calculation. This affects acceleration, not the geometric test. Timing results from this replacement are not presented as backend performance measurements. Error standard deviations in the legacy table use the population convention of the original ten-seed summary.

Executed components. The policy/client slot contains hand-designed teachers and frozen Set–CNN, MLP and GRU policies; no language-model endpoint is evaluated. Tool validation enforces pose, bounds and finite actions. The LTrace/path branch supplies synthetic direct/specular returns, source/path annotations and clutter. Public updates reconstruct count maps and estimates after thinning without passing target truth to the policy. The evaluator scores all acquired scans, so policy blindness does not make scoring blind.

Architecture context. The tool-call text and miniature plots in the preserved artwork are illustrative, not training targets or experimental observations. Waveform/ADC panels describe optional backend context rather than an executed signal-level sensing cohort. Weather, peers and multi-radar modes appear in legacy calibration; the representation study uses static single-radar rooms.

Figures 1, 3 and 4 are included from the supplied original PDFs without altering their visual content. Figure 1 is set to 0.70 of the text width; Figures 3 and 4 span the text width. Their inherited dense labels are smaller than the main text. All regenerated experimental figures use a native 5.5-inch canvas and 9-point ordinary labels. These two asset classes have different typography constraints; the artifact reports them separately rather than claiming that the unchanged schematics meet the new chart-label size target.

O Execution Scope and Reproduction of This Revision

The supplied v16 manuscript reports an audit that recovers deterministic activations from selected weights, evaluates the action-head decomposition, regenerates seventy backend fixtures and adds 240 episodes with frozen-policy readouts. These are retained study results, not experiments rerun during this presentation revision. The available v11 source archive supports its historical analyses; the exact v16 plot primitives support the revised graphics. The missing v16 experiment scripts and source files are identified in the release documentation. Recovering deterministic caches is distinct from policy retraining.

The revision does not add a matched-quality policy cohort, a Transformer policy, a second sensing domain, a new evidence-necessary task, a crossed recipient-pose/donor-coordinate DAS test, or edited-policy closed-loop evaluation. These are remaining limitations, not experiments presumed successful from the reviewers' suggestions. The selected research direction is an audit of representation diagnostics, with a narrower mechanistic interpretation supported by its new negative controls.

The separate MATLAB file contains exact plotted primitives in manuscript order. Python generated the delivered vector figures from the same data. The embedded numerical payloads were checked against the supplied specifications; MATLAB runtime and pixel-identical rendering are not claimed.