CLEAR: CALIBRATED LINEAR ERASURE AUDITING OF REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

What does a performance drop after representation erasure reveal about a model's use of the targeted information? Erasure aims to remove target information from internal activations, but can also disrupt other task-relevant features. The resulting performance drop can therefore reflect both effects. Interpreting this drop requires checking how much target information remains, how sensitively residual tests can detect it, and how intervention choices affect behaviour. We introduce CLEAR, a framework for calibrated linear erasure auditing of representations. CLEAR combines behavioural change with the remaining linear target association, the sensitivity of the residual test, and the response to different intervention strengths and replacement values. Calibration makes the resolution of each residual verdict explicit, connecting the validity of an internal edit to the behavioural evidence it provides. This joint assessment exposes ambiguities that effect magnitude alone can hide. Synthetic controls show that error from fitting the eraser can produce behavioural effects even at zero target dependence. In the SSR planner, full ego state erasure increases 3 s average displacement error to 7.3 times its baseline value while failing residual checks at every audited layer. Error also grows beyond the intervention strength that minimizes target association and varies with replacement values. CLEAR makes explicit how residual information, detection sensitivity, and intervention design constrain the interpretation of behavioural changes induced by erasure.

1 INTRODUCTION

World models and end-to-end policies encode observations into representations used to produce predictions and actions (Hafner et al., 2023; Vafa et al., 2025). Understanding their computation requires identifying which encoded variables influence a particular output. In a driving planner, for example, accurate speed decoding establishes that an activation contains ego state information, while explaining a trajectory requires tracing how that information enters the planning computation. This distinction connects representation analysis to a concrete question about model behaviour: which internal variables account for the output produced on a given input?

Linear readouts measure how well a variable can be decoded from an activation (Ravichander et al., 2021; Belinkov, 2022). Representation erasure extends this analysis by removing directions that predict the target and measuring the change in behaviour (Elazar et al., 2021; Ravfogel et al., 2020). The appeal of this procedure is its direct link between an internal edit and an observable output. Its interpretation depends on what the edit changes. A fitted projection can leave target information on held-out data, displace features useful for other computations, and alter relationships among activation coordinates. The measured effect consequently combines target dependence with the consequences of editing the surrounding representation.

Prior work identifies completeness and selectivity as distinct requirements for interpreting interventions (Canby et al., 2025). Probe-guided removal can both retain target information and damage features that support the task (Kumar et al., 2022), motivating improved erasure operators (Belrose et al., 2023; Ravfogel et al., 2023; Holstege et al., 2025). These requirements also make the measurement procedure part of the inference. A residual test has a resolution set by target scale, activation geometry, and audit sample size. A passing verdict can arise from a small residual or from low sensitivity to the leakage that remains. Conversely, estimation error in the fitted eraser can produce residual failures even when a calibration injects no leakage. Interpreting either outcome requires measuring the test's response to known leakage under the same fitting and evaluation conditions.

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.

Behavioural magnitude introduces a related ambiguity. Increasing erasure strength changes both target association and the distance that activations move. Replacing the removed component with a constant or a sample from another scene changes the edited representation in different ways. These choices provide a means to examine the source of a performance drop: fixing the projection separates its magnitude from refitting, while varying replacement values tests sensitivity to the remaining feature relationships. Together with residual calibration, these comparisons connect the validity of an edit to the behaviour it produces.

We introduce CLEAR, a framework for calibrated linear erasure auditing of representations. Its central design is to evaluate an erasure effect jointly with the validity and resolution of the residual test. Given activations X, a target Z, and an output Y, we fit a separate eraser for each target on separate data and compare residual association on held-out data with a matched null. Injected leakage calibrates both the failure rate at zero injected leakage and the dose detected with 80% power. An oracle based on the exact projection provides a reference for the contribution of eraser estimation error. These quantities place residual verdicts on a measured scale, specific to each target and layer. Behavioural effects are then evaluated alongside intervention delivery and the reliability of auxiliary readouts. Dose sweeps with a fixed projection and comparisons across replacement values assess sensitivity to the intervention. Together, these measurements make the evidence connecting an internal edit to its behavioural effect.

We evaluate the framework through controlled tests and an end-to-end planner. A synthetic system isolates fitting error and target dependence: reducing the budget for fitting the eraser produces a behavioural effect even at zero dependence. LAW provides a complementary control with a verified computational path, in which speed is strongly decodable from a latent computed after the current trajectory. The SSR case study then connects calibration to a large planning error. Driving offers measured ego state targets and a continuous trajectory error endpoint, making it possible to compare residual association with behavioural changes across intervention strengths. Full ego state erasure increases 3 s ADE to $7.3\times$ its baseline value, while residual checks fail throughout the audited layers. Scaling one fixed projection and varying its replacement values show how the intervention itself contributes to this effect: planning error grows beyond the dose that minimizes target association, and resampling across scenes is more harmful than zeroing. Together, these experiments show how CLEAR characterizes both the residual target association and the intervention sensitivity that constrain behavioural attribution.

Our contributions are threefold:

- We introduce CLEAR, an auditing framework that combines held-out residual checks, calibrated detection sensitivity, and paired behavioural effects to interpret linear representation erasure.
- We calibrate matched residual tests with injected leakage and an exact projection oracle, quantifying failure floors for the fitted pipeline and detection limits specific to each target and layer.
- We evaluate CLEAR with synthetic controls, a LAW control based on a computational path, and an SSR case study, showing how eraser-fitting error, computational paths, and intervention choices shape the evidence for target use.

2 Related Work

Concept removal and its evaluation. Erasure operators include adversarial, kernelized, log-linear, and closed-form methods (Ravfogel et al., 2022a;b; 2023; Belrose et al., 2023). Task-preserving erasure (Holstege et al., 2025) and downstream deconfounding (Fan et al., 2025) extend their objectives. The removal failures studied by Kumar et al. (2022) and the intervention reliability framework of Canby et al. (2025) are the closest foundations for our analysis, alongside recent amnesic-probing evaluations (Dobrzeniecka et al., 2025). CLEAR augments erasure evaluation with a matched held-out null, failure floor $p(0)$ for the fitted pipeline, and detection limit β^*, which quantify the sensitivity of the residual test for a fitted operator.

Causal interventions. Causal probing formalizes readouts using latent variables (Jin & Rinard, 2024), while causal abstraction and distributed alignment search for representations of specified causal variables (Geiger et al., 2021; 2024). Activation patching results depend on metrics and intervention implementation (Zhang & Nanda, 2024). We characterize this sensitivity for a prespecified linear subspace through a fixed projection and comparisons across replacement values.

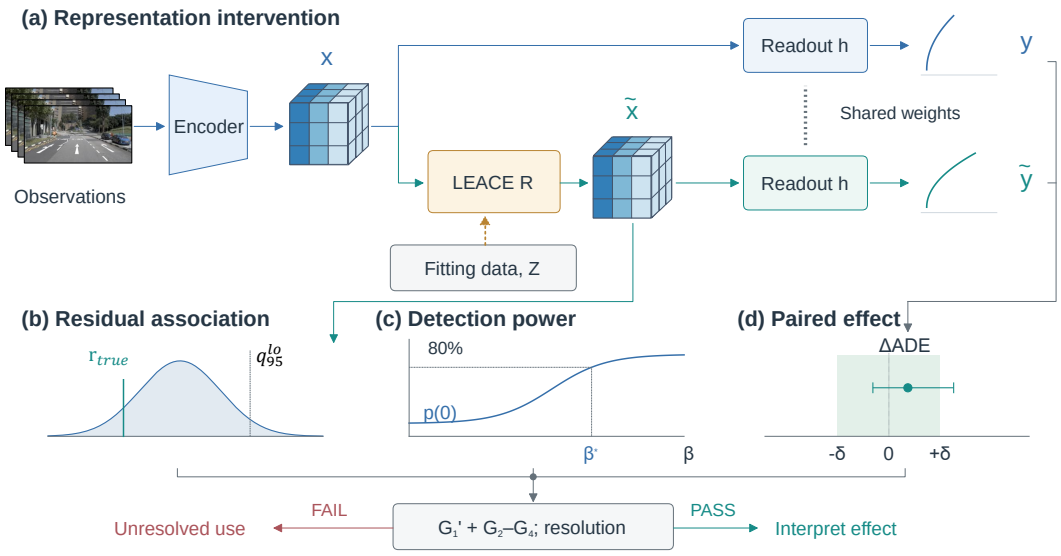


Figure 1: **The CLEAR auditing framework.** (a) Original and erased activations pass through the same frozen readout. LEACE is fitted on separate scenes. (b) Test residual linear association against a matched null. (c) Inject known leakage to calibrate power. (d) Measure the paired behavioural effect. Interpret the effect using residual and auxiliary checks together with calibrated resolution. The dashed amber arrow denotes fitting on separate scenes.

Driving representations. DriveWorld, LAW, SSR, and World4Drive learn representations for planning (Min et al., 2024; Li et al., 2024a; Li & Cui, 2025; Zheng et al., 2025). Ego-status studies identify shortcuts in the inputs to open-loop metrics (Li et al., 2024b; Weng et al., 2024), while NAVSIM and Bench2Drive separate non-reactive and closed-loop evaluation (Dauner et al., 2024; Jia et al., 2024). Our case study examines internal ego state edits and their effect on trajectory error in a frozen planner.

3 Method

CLEAR takes a frozen model, a target world state Z, a selected activation layer, and disjoint data for fitting the eraser and auditing. It fits and applies an erasure operator, measures residual association against a matched null, calibrates the test's resolution, and interprets the paired behavioural effect using these measurements. For each intervention arm, it reports residual association, the failure floor for the fitted pipeline, the detection limit, and the behavioural effect with uncertainty. Figure 1 shows this sequence. Auxiliary checks assess intervention delivery and readout reliability; attribution to selective removal additionally requires preservation of other information that supports the task after the edit.

3.1 Erasure operator

Let $x \in \mathbb{R}^d$ be an activation, $z \in \mathbb{R}^k$ its target, and h the remaining model computation. LEACE fits an affine eraser (Belrose et al., 2023),

$$\widetilde{x} = R(x) = x - P(x - \mu), \qquad y = h(x), \quad \widetilde{y} = h(\widetilde{x}), \tag{1}$$

where μ is the mean of the fitting set and P is the fitted, generally oblique projection onto directions that predict the target. This removes linear cross-covariance on the fitting data; the residual test evaluates generalisation to audit data. The input and model weights remain fixed, and R is never refitted on audit scenes. Separate semantic target blocks receive separate fits because their scales differ. Control interventions are matched to the measured target rank.

Model training, fitting the eraser, and auditing use data from pairwise disjoint sets of scenes. Both residual association and the paired output change are evaluated on audit scenes. This separates fitting success from intervention validity on unseen scenes.

Table 1: **Auxiliary checks and their pass criteria.** Readout checks use clean activations. Criteria apply to each listed layer; the corresponding joint check requires all applicable criteria to pass.

Check	Measurement and pass criterion
Delivery (G_2)	Add Gaussian noise at L2 with a standard deviation for each coordinate from the fitting set. The 95% interval for paired 3 s ADE change must exclude zero.
Target decodability (G_3)	At L1–L3, readouts for navigation commands and other agents selectivity must each have a positive 95% interval. At L0, ego state $R^2 > 0.5$ and the selectivity interval for random labels must contain zero with positive width.
Readout reliability (G_4)	At L0–L3, the 95% selectivity interval for labels chosen at random must contain zero and have width in $(0, 0.5]$.

3.2 Residual association statistic and criterion

For clean audit activations X, edited activations $\widetilde{X}$, and targets Z, define

$$r_{\text{true}} = \frac{\|\widehat{\text{Cov}}(\widetilde{X}, Z)\|_{\text{F}}}{\|\widehat{\text{Cov}}(X, Z)\|_{\text{F}}}, \qquad G_1' : \quad r_{\text{true}} \le q_{95}^{\text{lo}}. \tag{2}$$

The denominator normalizes by clean association. Because geometry and target rank affect this ratio, each intervention arm at each layer uses a reference computed from the same activations with 200 label reassignments across scenes. For scenes of unequal length, modular reassignment maps frame indices from the recipient scene to the donor scene, reusing or dropping donor frames as needed. We estimate the null's 95th percentile q_{95} and use the lower endpoint q_{95}^{lo} of a 95% bootstrap interval from 2000 resamples. An arm at a given layer is excluded as uninformative if $q_{95} \ge 1$, before applying the gate.

This criterion based on the relative null classifies an intervention as passing when residual linear association is at or below the matched threshold. Its empirical failure rate and detection sensitivity are estimated by the calibration below. Section A.2 evaluates sensitivity to the null's construction using a bijective alternative stratified by scene length.

3.3 Calibration of test resolution

To characterise the sensitivity of a passing result, we inject known leakage fractions β into matched semisynthetic activations and repeat the audit. Let

$$p(\beta) = \Pr(G_1' \text{ fails} \mid \beta), \qquad \beta^* = \min\{\beta \in [\min \mathcal{B}, \max \mathcal{B}] : \widehat{p}_{\text{lin}}(\beta) \ge 0.8\}, \tag{3}$$

where $\mathcal{B}$ is the registered dose grid and $\widehat{p}_{\text{lin}}$ linearly interpolates the empirical failure rates from repeated runs at adjacent grid points. The detection limit is the first crossing of 80% power. Leakage is controlled through erasure strength $\alpha = 1 - \beta$. For each target at each layer, we report both the detection limit β^* at 80% power and the failure rate at zero injected leakage $p(0)$, termed the *failure floor for the fitted pipeline*. The latter can be high because the fitted eraser leaves its own residual. We therefore calibrate the fitted pipeline and an oracle based on the exact projection separately. If power never reaches 80% on the grid, β^* is reported as a lower bound. Both scene and frame counts accompany calibration.

3.4 Conditional interpretation of behavioural effects

Three auxiliary checks assess intervention delivery G_2, target decodability from clean activations G_3, and readout reliability G_4, as defined in Table 1. Readout selectivity is the score difference between the target and its control with shuffled labels. Classification uses logistic regression and accuracy; regression uses ridge and R^2. Readouts use disjoint fitting and evaluation scenes. Intervals resample evaluation scenes with the fitted readout held fixed. A G_4 failure makes the corresponding G_3 interpretation unresolved. These checks characterize the measurement pipeline; preservation after erasure is a separate condition for attributing behavioural change to the target.

For driving, the behavioural endpoint is the paired change in 3 s average displacement error (ADE), $\Delta\text{ADE} = \text{ADE}(\widetilde{y}) - \text{ADE}(y)$, with uncertainty intervals over scenes. For arms that pass the residual check, a 90% interval contained in $[-\delta, \delta]$ meets the registered rule for behavioural equivalence; an interval entirely above δ meets the registered category for effects above the tolerance. Intervals

Table 2: **Confirmatory calibration by layer and target.** Detection limits are reported with fitted and oracle failure rates at zero injected leakage and detection rates for the fitted pipeline at the maximum dose $\beta = 0.30$. Bounds retain that maximum dose. All rates are percentages.

Layer	Target	β^*	Fitted $p(0)$	Oracle $p(0)$	$p(0.30)$
L0	acceleration	0.162	43.5	8.5	100.0
L0	rotation rate	0.141	36.7	8.1	100.0
L1	acceleration	0.217	32.9	9.3	100.0
L1	rotation rate	> 0.30	26.5	7.6	36.5
L2	acceleration	0.247	34.8	9.4	100.0
L2	rotation rate	> 0.30	28.7	9.8	51.5
L3	acceleration	0.257	32.1	8.6	100.0
L3	rotation rate	> 0.30	28.2	7.9	45.5

crossing the decision boundary remain inconclusive. These classifications concern effect size conditional on the residual check; attribution additionally depends on the auxiliary checks and calibrated power. The tolerance δ is derived from reproducibility across eraser refits as specified in Section A.8. All measured effects are reported.

Intervention sensitivity. To assess how the behavioural effect depends on the edit, we scale a frozen projection as $r_\alpha(x) = x - \alpha P(x - \mu)$ and compare the residual and behavioural responses across doses. Holding P fixed separates dose changes from eraser refitting. Complementary comparisons replace the removed component with the mean of the fitting set or a projection sampled from another scene. Divergent responses identify sensitivity to intervention magnitude or replacement values; they are interpreted alongside the residual checks and calibrated resolution.

4 Experiments

The experiments proceed in five stages. We first quantify residual test resolution and its dependence on the audit sample size. We then evaluate the protocol with synthetic controls under known target dependence. A separate LAW control tests whether decodability implies computational use. We next apply CLEAR to the SSR planner and finally vary intervention strength, replacement values, activation geometry, and eraser construction.

4.1 Calibration of the residual test

Calibration determines how much residual leakage the audit can detect and how often error from fitting the eraser alone causes a failure. On the exploratory split at L0, the exact-projection oracle fails in 9.4% of runs, compared with 28.7% for the fitted pipeline. The oracle therefore understates failures at zero leakage because it excludes the error introduced by estimating the eraser.

Table 2 reports calibration for 8 confirmatory combinations of target and layer. At L0, the fitted acceleration pipeline fails in 43.5% of runs with zero injected leakage, compared with 8.5% for the exact-projection oracle. Both synthetic calibration procedures pass the fidelity check within 1.5%. Controlled subsampling shows that more audit scenes narrow the null and expose more fitted residual, raising the pipeline floor, as quantified in Section A.3.

Resolution depends on the target and layer. The acceleration detection limit increases with depth despite similar floors. For rotation rate at layers L1 through L3, detection power remains below 80% at the maximum registered leakage dose, so the corresponding detection limits are reported as bounds.

Sampling dependence. A preregistered scaling law attributed null width to the number of held-out frames. A two-axis subsampling experiment instead identifies the number of held-out scenes as the main factor, as shown in Section A.3. A model fitted on seven of nine (S, F) settings predicts a held-out width ratio of 1.34, close to the observed 1.32; a frames-only model predicts 1.00. Both scene and frame counts are therefore reported with every floor. This result establishes the calibration consequence used below. Increasing the number of audit scenes makes the null narrower and makes fitted residual error more detectable.

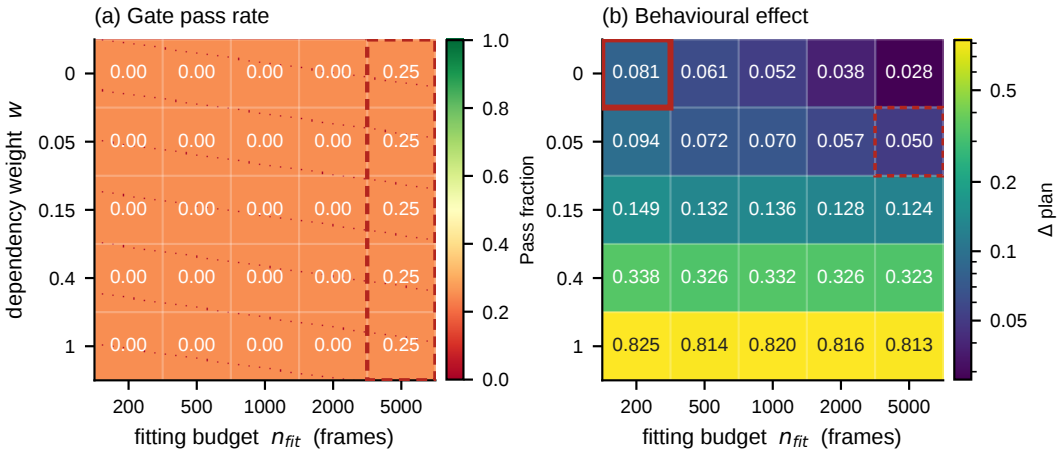


Figure 2: **Validity and effect respond to different factors.** The left panel shows how the gate responds to the fitting budget. The right panel shows the behavioural effect across dependency weights w. With $w = 0$, an eraser fitted on 200 frames produces a substantial effect. Cells average eight seeds. The dashed last column marks a censored 50% gate crossing because the pass rate reaches only 0.25 at 5000 fitting frames. In the effect panel, the solid box marks the zero-dependence condition at 200 fitting frames, with effect 0.0812. The dashed box marks the corresponding $w = 0.05$ condition at 5000 frames. The zero-dependence condition exceeds this comparator by $1.63\times$. Effect colours use a logarithmic scale.

4.2 Synthetic controls

The synthetic controls test whether the erasure effect tracks known dependence and whether intervention error can produce a behavioural effect.

Synthetic system. An ego-like variable enters the synthetic target with known weight w. The same erasure code and validity threshold are used throughout. Six preregistered variants vary dependence strength, fitting budget, collateral overlap, random-subspace displacement, and nonlinear coupling, with known dependence in every cell.

Validity follows fitting budget. In the fitting budget sweep, $w = 0.4$ is fixed while the amount of data used to fit the eraser varies. The gate pass rate rises from 0.00 at $n_{\text{fit}} = 200$ to 0.25 at 5000 frames, remaining below 50% throughout the grid. Residual failures also occur in the zero dependence control, showing that the validity decision is sensitive to error from fitting the eraser even when the target is unused. Section A.6 reports all six experiments and their outcomes against the preregistered predictions.

Behavioural effects depend on association and intervention error. Across the synthetic sweeps, the dependence sweep shows a monotonic increase in behavioural effect with w, with Spearman correlation 1.00. The collateral overlap sweep shows an increase at fixed w, with Spearman correlation 1.00. In the fitting budget sweep, reducing the fitting budget increases the effect from 0.3234 to 0.3383, a change of 1.6 times the registered tolerance. Thus, poorer fitting increases the effect even as residual validity deteriorates. The effect responds to both dependence and intervention error. Section A.6 gives the full results.

Effect at zero dependence. At $w = 0$, the measured effect is 0.0277 with the largest fitting budget and 0.0812 with 200 fitting frames, a $2.9\times$ increase. The latter falls within the range produced by a genuine small dependence with a well-fitted eraser, detailed in Section A.6. Effect magnitude alone therefore conflates target dependence with eraser error.

4.3 Decodability and computational contribution

LAW provides a control based on a verified computational path (Li et al., 2024a). We examine an early visual representation (W1), an intermediate waypoint representation (W2), and a world-model representation produced after waypoint prediction (W3) along this path. The trajectory is generated before W3 is returned, making W3 an auxiliary post-decision representation. It can therefore contain

Table 3: **Confirmatory ego state erasure in SSR.** Baseline 3 s ADE is 0.6294 m. Residual outcomes cover L0–L3. Effect intervals use scene-level resampling; their confidence levels are specified separately because the behavioural classification uses 90% intervals. The equivalence tolerance is δ =0.21007 m.

Erased target	ADE change	Interval in m	Level	Residual check
Full ego state	$+3.9602$ m	$[+3.5029, +4.4380]$	95%	Fail at all layers
Velocity	$+3.8593$ m	$[+3.3993, +4.3484]$	95%	Fail at all layers
Acceleration	$+0.2552$ m	$[+0.2144, +0.2973]$	90%	Pass at all layers
Rotation rate	$+0.0508$ m	$[+0.0371, +0.0656]$	90%	Pass at all layers

L1 comparison	Residual r_{true}	Threshold q_{95}^{lo}	Outcome
Full ego state	0.0826	0.0661	Fail
Rotation rate	0.3153	0.4761	Pass

information about the predicted world state without contributing to the current trajectory. This computational ordering provides a control for separating readout decodability from causal contribution to the output.

Longitudinal speed is strongly decodable from W3, with a median $R^2 = 0.9848$ across slots. W3 is formed from the predicted waypoints after trajectory generation, providing a direct example of a representation that is informative about the task while lying downstream of the current output.

The LAW readouts use 2002 frames over 50 scenes held-out samples from disjoint scenes. At the earlier visual tap W1, six slots predict longitudinal speed with R^2 between 0.9908 and 0.9928. These values exceed the best non-W1 slot, 0.5294, and also exceed the scene-identity ceiling of 0.6380. The same readouts predict heading angle and longitudinal acceleration weakly, with median R^2 values of -0.0159 and +0.0345, respectively. The strong W1 speed readout therefore captures within-scene visual speed variation, whereas W3 provides a post-decision readout of predicted world state. Section D provides further details on the computational path and reports the complete readout table.

4.4 CLEAR AUDIT OF SSR

CLEAR applies the same residual and behavioural criteria to the SSR planner. A fixed seed partitions 685 usable nuScenes scenes into three disjoint sets. The model training set contains 400 scenes and 16,083 frames, the eraser fitting set contains 150 scenes and 6,029 frames, and the audit set contains 135 scenes and 5,421 frames. Saved scene tokens verify pairwise disjointness. Section A.7 gives the scene assignments and sample counts. We audit ego state blocks at four representation layers, from L0 through L3. The three-column velocity block has rank one because two columns are constant, so rank-matched controls use this measured rank. Behaviour is evaluated on the frozen SSR checkpoint from one training seed using 3 s ADE.

The confirmatory audit evaluates full ego state, velocity, acceleration, and rotation-rate erasures at layers L0 through L3. Full ego state erasure, the registered primary intervention, produces the largest increase among the target erasures in Table 3 and fails the residual check at every audited layer. Velocity erasure shows the same pattern. Table 3 summarizes the confirmatory effects and their uncertainty.

The acceleration and rotation-rate interventions pass the residual check at every layer but yield different behavioural classifications. The rotation-rate interval lies within the equivalence tolerance. The acceleration interval lies just above it, making that classification sensitive to the chosen tolerance. The sensitivity analysis in Section A.8 gives the interval boundaries. These results separate behavioural impact from residual target association. The intervention can change planning behaviour, while the residual gate measures the remaining target association at the calibrated resolution.

The L1 comparison in Table 3 illustrates why a passing residual check requires calibration. Rotation has a larger residual than full ego state at L1, but also a wider null. Its detection rate at maximum injected leakage remains close to its failure floor at zero leakage, as reported in Table 2. The passing outcome is therefore accompanied by low test sensitivity at this target and layer.

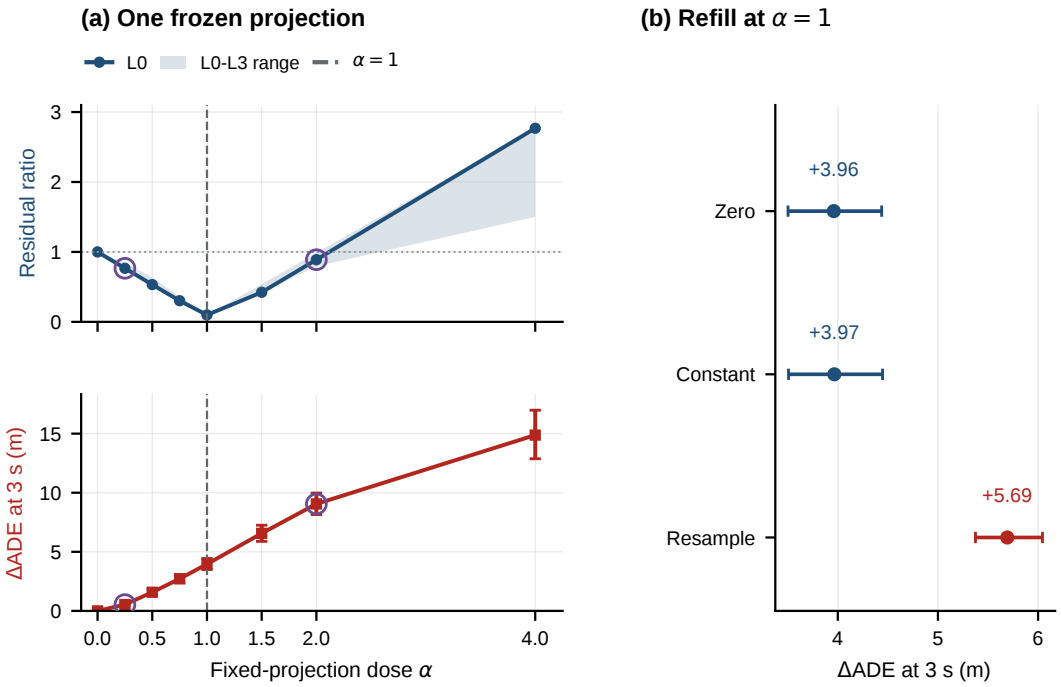


Figure 3: **Planning error increases beyond minimum target association.** (a) Residual association reaches its minimum at $\alpha = 1$, while ADE keeps increasing. The horizontal reference is the clean residual ratio $r = 1$. Ringed points form a descriptive comparison, outside the registered tests. (b) Resampling that preserves the marginal distribution produces more error than zeroing; constant replacement is comparable. The projection is fixed throughout, unlike the refitted dose sweep. Ringed doses $\alpha = 0.25$ and 2 have residuals 0.766 and 0.889 but ADE changes of $+0.54$ and $+9.05$ m. Error bars are 95% intervals over 135 scenes comprising 4611 valid frames. All three registered dose contrasts after erasure are positive and reject after Holm adjustment across the six-test dose-and-refill family.

The auxiliary checks identify a separate measurement issue. Intervention delivery and target decodability from clean activations pass, as do the residual-test controls and the rank-matched placebo check. However, the random-label readout control separates from its shuffled reference at L2, so L2 readout selectivity remains indeterminate. The corresponding intervals and scores are given in Table 4. The rank-matched random-subspace control fails the residual check at every layer.

We apply the same residual and behavioural criteria to each intervention at each layer, with the individual criteria evaluated separately. Full ego state remains the primary target. Section A.1 describes the intervention coverage and the outcome-dependent selection decision. The exploratory effects provide the same qualitative pattern as the confirmatory audit and are shown in Figure 8.

The refitted dose sweep in Section B.1 reduces residual association while increasing ADE. We next hold the projection fixed to isolate the effect of intervention strength and compare replacement values.

4.5 Intervention Sensitivity in SSR

Erasure of the full block increases 3 s ADE from 0.6294 m to 4.5895 m, reaching $7.3\times$ the baseline value. We test whether this effect tracks target association or the total representational change.

Dose sweep with a fixed projection. We apply the diagnostic that uses a fixed projection, defined in Section 3.4. Figure 3 shows that target association reaches its minimum at $\alpha = 1$, while planning error continues to increase at higher doses. All three registered contrasts for successive dose increases from $\alpha = 1$ onward are positive and reject after Holm adjustment across the six-test family. This divergence supports a contribution from intervention magnitude beyond the reduction in linear target association. The complete dose grid is given in Section B.1.

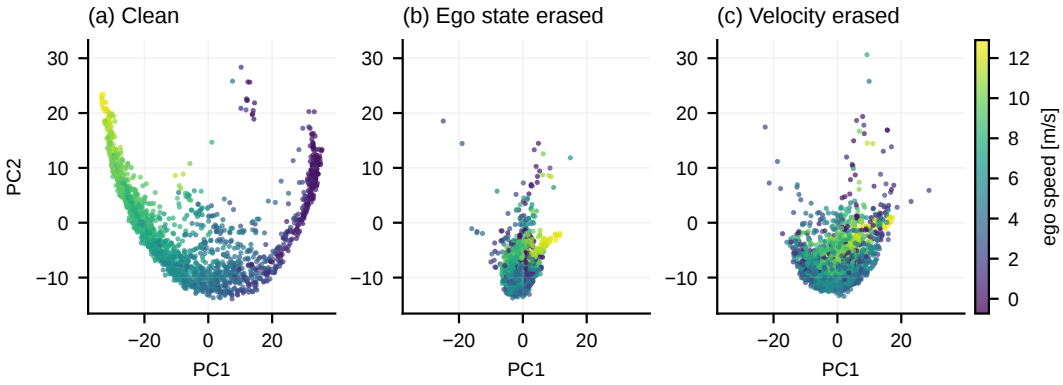


Figure 4: **Ego state erasure changes the geometry of dominant activation directions.** Clean, full ego state erased, and velocity erased L2 activations share the PCA basis fitted to the clean fitting set, centring mean, axis limits, and speed colour scale. PC1 and PC2 explain 63.7% and 10.1% of clean variance. Quantitative readout and displacement diagnostics are given in Figure 6.

Two refill interventions test sensitivity to replacement values. Constant replacement produces an effect comparable to erasure (paired difference, -0.0054 m; 95% CI, -0.0255 to 0.0151), whereas resampling across scenes increases error by 1.7322 m relative to erasure (95% CI, 1.5257 to 1.9426; $p_{\text{Holm}} = 0.0012$). This ordering establishes sensitivity to replacement values. Because resampling preserves the marginal distribution while changing its relationship to the current activation, the additional error may reflect disrupted feature relationships.

Activation geometry. On a shared principal basis, PC1 correlates with speed between -0.969 and -0.899 at every layer. Speed is therefore strongly associated with a leading axis of activation variance, as shown in Figure 4. Edits to speed-predictive directions may affect broad representational structure.

Eraser dependence. A preregistered comparison covers LEACE, iterative nullspace projection, and a projection based on the mean difference. At L1, the mean-difference operator passes the residual gate while LEACE fails. The mean-difference operator also removes most held-out activation variance, whereas the frozen INLP configuration is close to the identity. Rank matching therefore leaves intervention magnitude uncontrolled. These results show that the choice of eraser affects both residual validity and the amount of activation change.

Information retained after linear erasure. Across 20 combinations of four layers and five targets, an MLP recovers information after erasure with R^2 in 0.56 to 0.99. The linear readout is near chance except in one combination, with a maximum ridge R^2 of 0.1145 for full ego state erasure at L0, as shown in Figure 7. These readouts identify nonlinear information retained after the linear intervention.

Together, these analyses show that behavioural effect magnitude, residual validity, and intervention sensitivity capture distinct properties of an erasure.

5 Conclusion

CLEAR turns linear representation erasure into a calibrated audit that connects behavioural change with residual validity, detection resolution, and intervention sensitivity. Matched nulls and injected leakage establish the resolution of each residual verdict, while fixed-projection dose sweeps and replacement comparisons characterize how the intervention produces its behavioural effect. These experiments show how the measurements sharpen the interpretation of erasure. Synthetic controls reveal behavioural effects caused by eraser estimation even at zero target dependence. LAW provides a computational example in which information is strongly decodable from a representation downstream of the current output. In SSR, full ego state erasure increases ADE to $7.3\times$ its baseline value, with residual association at every audited layer and systematic variation across intervention strengths and replacement values. CLEAR brings these sources of evidence into a single empirical account of what an erasure changes and how that change relates to model behaviour.

AI USE STATEMENT

The authors used generative AI tools to polish author-written text for grammar, clarity, and readability and to assist in the interpretation and presentation of existing experimental results. These tools were not used to generate research ideas, methods, code, data, experiments, analyses, or reported numerical results. All AI-assisted interpretations and language revisions were reviewed and approved by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Data roles and statistical procedures are described in Sections A.7 and A.8; calibration results and synthetic controls appear in Sections A.5 and A.6. Sections B.1 and D document the additional intervention and readout analyses. Reported numerical values are populated from stored experimental results, and the manuscript build checks for unresolved values.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

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 (NeurIPS)*, 2023.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 857–878, 2025. doi: 10.18653/v1/2025.ijcnlp-long.47.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Alicja Dobrzeniecka, Antske Fokkens, and Pia Sommerauer. Improving causal interventions in amnesic probing with mean projection or LEACE. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12981–12993, 2025. doi: 10.18653/v1/2025.findings-acl.674.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics (TACL)*, 9:160–175, 2021.

Yu Fan, Yang Tian, Shauli Ravfogel, Mrinmaya Sachan, Elliott Ash, and Alexander Hoyle. The medium is not the message: Deconfounding document embeddings via linear concept erasure. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32124–32143, 2025. doi: 10.18653/v1/2025.emnlp-main.1634.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. In *International Conference on Machine Learning (ICML)*, 2023.

Floris Holstege, Shauli Ravfogel, and Bram Wouters. Preserving task-relevant information under linear concept removal. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 26880–26920, 2025. doi: 10.52202/085713-0907.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Charles Jin and Martin Rinard. Latent causal probing: A formal perspective on probing with causal models of data. *arXiv preprint arXiv:2407.13765*, 2024.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing classifiers are unreliable for concept removal and detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17994–18008, 2022.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018.

Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.18341.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model, 2024a.

Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M. Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. the BEV-Planner paper; arXiv:2312.03031.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, and Bin Dai. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15522–15533, 2024. doi: 10.1109/CVPR52733.2024.01470.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning (ICML)*, 2022a.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Kernelized concept erasure. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022b.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 3363–3377, 2021.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. *arXiv preprint arXiv:2507.06952*, 2025.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. PARA-Drive: Parallelized architecture for real-time autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15449–15458, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, XianPeng Lang, and Dongbin Zhao. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28632–28642, 2025. doi: 10.1109/ICCV51701.2025.02659.

A Audit and statistical details

A.1 SSR audit details

The confirmatory audit includes all prespecified interventions across the audited layers, including an L2-only full-state erasure and a Gaussian-noise delivery control. The yaw intervention at L3 is screened out by $q_{95} \geq 1$. Residual outcomes for the remaining conditions are reported with the confirmatory results.

Table 4: **Random-label readout control in the confirmatory SSR audit.** Selectivity is the difference between the random-label and shuffled-control scores. L2 fails because its 95% interval excludes zero; the intervals at L0, L1, and L3 contain zero.

L2 measurement	Value
Random-label score	0.3177
Shuffled-control score	0.3431
Selectivity, 95% interval	$[-0.0432, -0.0081]$

Exploratory calibration covers full ego state and velocity across the audited layers, and the confirmatory audit applies the individual criteria separately to each prespecified intervention condition. The preregistration specified full ego state erasure as the primary estimand and left multiplicity correction unspecified. Its erasure prerequisite failed at L0 in the exploratory split.

After that failure, velocity erasure was considered as a substitute. Because the choice depended on the observed prerequisite, it created a selection risk. Full ego state erasure remains the confirmatory primary arm; velocity erasure also fails at all four layers, as reported in Section 4.4. All prespecified conditions are reported, limiting outcome-dependent selection, although the individual criteria remain unadjusted for the familywise error rate.

A.2 Sensitivity to null construction

In the diagnostic audit, modular reassignment reuses or drops about $43/5389$ frames per draw because scene lengths differ. Figure 5 compares the original modular reassignment with a length-stratified bijective construction across the 20 gated slots in the exploratory audit. All verdicts agree, while threshold shifts vary in sign across targets and layers. Eight thresholds are stricter under the original construction; the passing margin for yaw erasure at L0 shrinks from 3.6% to 0.85% under the bijective construction.

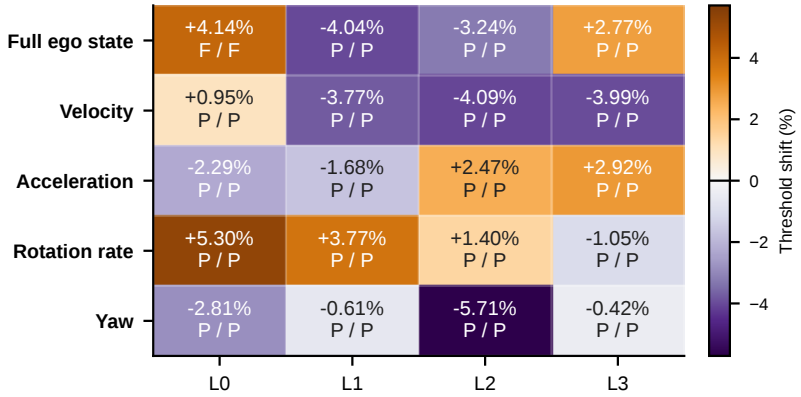


Figure 5: **Sensitivity to the null's construction.** Each cell shows the threshold shift and the original/bijective verdict pair for a gated slot in the exploratory audit. P and F denote passing and failing results. All 20 verdicts agree; eight thresholds are stricter under the original null. The yaw erasure at L0 passing margin shrinks from 3.6% to 0.85%. The sign of the threshold shift indicates which construction is stricter.

A.3 What sets the null's width

The failure floor depends on both fitted residuals and the width of the permutation null. The exploratory and confirmatory splits differ in checkpoint, scene pool, fitted eraser, activation geometry,

Table 5: **Null scaling under controlled subsampling at L0.** Floors are failure probabilities at zero injected leakage; arrows run from the full evaluation set to the smallest subset. Exponents describe decay with sample count. The last two rows report observed/predicted null ratios, ordered from the smallest to the largest sample size for frames. The registered tolerance is $\pm 10\%$; variation across draws ranges from 0.014 to 0.058.

Quantity	Acceleration	Rotation
Floor, varying frames	$0.450 \to 0.322$	$0.385 \to 0.274$
Floor, varying scenes	$0.450 \to 0.232$	$0.385 \to 0.213$
Frame exponent	0.150	0.163
Scene exponent	0.435	0.449
Frame ratios	$0.71, 0.80, 0.89, 0.95, 1.00$	$0.72, 0.81, 0.89, 0.95, 1.00$
Scene ratio range	0.936 to 1.000	0.952 to 1.001

Table 6: **Prediction on withheld sampling configurations.** Entries are predicted/observed null widths for two configurations held out when fitting the joint law. Values near one indicate accurate prediction.

Scenes	Frames	Acceleration	Rotation
135	2002	0.9938	0.9967
50	2010	0.9732	0.9811

and sample size. Their direct comparison mixes these factors. To isolate the evaluation set, a single fixed eraser is evaluated under subsampling that varies frame and scene counts separately, with five sizes and five independent draws per size. The eraser is fitted before the evaluation set is accessed, so replicate r uses the same eraser at every size.

With the eraser fixed, reducing the evaluation set lowers the failure floor and widens the null. The variation across repeated draws is smaller than the variation across sample sizes. Table 5 summarises both targets.

The scaling law depends on the sampling unit. When only the frame count varies, the observed null departs from the registered $n^{-1/2}$ law and falls outside its tolerance at the smallest sample size for both targets. When scene count varies, all five levels stay within tolerance. The total scaling exponent is compatible with the registered law, while its allocation between scenes and frames differs.

A joint law, validated out of sample. The null is built by permuting whole *scene* blocks, so its height is expected to depend mainly on the block count and weakly on the number of frames inside a block. Fitting both axes at once gives, at L0,

$$\mathrm{null}(S, F) \;\propto\; S^{-0.285}\, F^{-0.150} \quad \text{(acceleration)}, \qquad S^{-0.286}\, F^{-0.163} \quad \text{(rotation rate)}, \tag{4}$$

for S evaluation scenes and F evaluation frames. This reproduces all nine unique (S, F) points to within 0.92% and 1.19% respectively. The exponents sum to 0.435 and 0.449, both below the preregistered value of $1/2$.

To distinguish interpolation from prediction, a refit on seven points predicts two withheld configurations with similar frame counts but different scene counts. Table 6 gives the prediction-to-observation ratios for each target and configuration.

At the full evaluation size, the subsampling and independent calibration implementations agree to within 0.1%. Both targets show a weaker dependence on frame count than on scene count. These measurements concern L0, so the fitted exponents in Equation (4) are specific to that layer.

A.4 Representation geometry and nonlinear decodability



Figure 6: **Quantitative diagnostics of activation geometry.** (a) Speed readout using $k = 2$ or $k = 8$ leading PCs; the dashed line is the constant predictor with $R^2 = -0.0313$. (b) Displacement is $\|A_{\text{er}} - A\|_{\text{F}} / \|A - \bar{A}\|_{\text{F}}$; PC spreads are erased/clean ratios of standard deviations. The dashed reference is one.

Table 7: **Calibration on the residual scale on the exploratory split.** Observed residuals, thresholds from the matched null, and conditional transitions r^* are shown for each layer and target, with 95% intervals on r^*. These transitions describe realised residuals; power as a function of injected leakage is quantified separately by β^* and the fitted floor $p(0)$.

Layer	Target	r_{true}	q_{95}^{lo}	r^* [95% CI]	Gate
L0	Full ego state†	0.1188	0.1096	0.1495 [0.1478, 0.1509]	Fail
L0	Velocity	0.1593	0.1655	0.1801 [0.1782, 0.1819]	Pass
L1	Full ego state	0.0607	0.1090	0.1300 [0.1280, 0.1318]	Pass
L1	Velocity	0.0630	0.1192	0.1400 [0.1381, 0.1419]	Pass
L2	Full ego state†	0.0424	0.0987	0.1151 [0.1138, 0.1163]	Pass
L2	Velocity	0.0467	0.1026	0.1219 [0.1204, 0.1235]	Pass
L3	Full ego state†	0.0220	0.0444	0.0579 [0.0570, 0.0587]	Pass
L3	Velocity	0.0254	0.0575	0.0803 [0.0791, 0.0814]	Pass

† Experimental records created before explicit target annotation identify this target as full ego state.

A.5 Calibration by target and layer

Table 7 gives the exploratory residual-scale transitions defined in Section A.8. Rotation has the smallest clean cross-covariance norm, making a fixed leakage fraction a smaller absolute residual. This offers a possible explanation for its lower detection sensitivity.

The confirmatory calibration uses 5421 frames from 135 scenes; the exploratory calibration uses 2002 frames from 50 scenes. The controlled sampling comparison is specified in Section A.3.

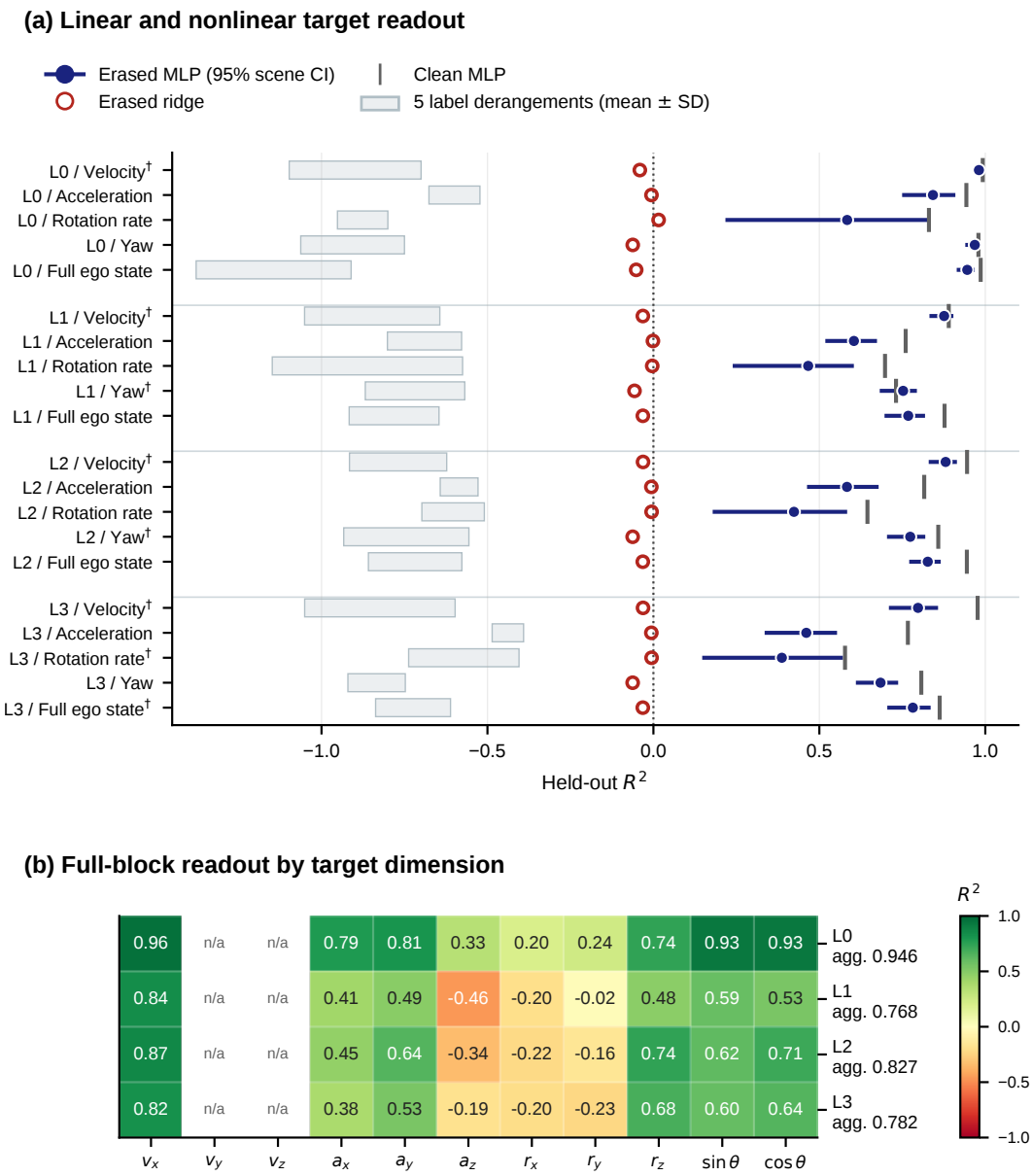


Figure 7: G'_1 **tests linear association.** Exploratory readouts on the same erased activations retain nonlinear predictability when ridge readouts are near zero. Filled points and intervals show MLP performance; open points show ridge, ticks show the clean MLP, and grey bands show the mean under shuffled labels ± one standard deviation. Daggers mark eight slots where variation across clean readout seeds exceeds the gap between erased and clean readouts. The heatmap preserves values for each dimension, including undefined constant targets. The separate confirmatory audit covers layers L0, L1, L2, L3 and retains one exception with residual linear information, with maximum ridge R^2 of 0.1145 for full ego state erasure at L0.

A.6 Details of the synthetic control

The synthetic experiments vary fitting budget and target dependence, including a zero-dependence condition, and separately test collateral overlap, rank-matched random-subspace displacement, and nonlinear coupling. The results below report the corresponding validity and effect measurements.

Fitting budget sweep. This experiment holds $w = 0.4$ and sweeps n_{fit} over the preregistered grid of $\{200, 500, 1000, 2000, 5000\}$ fitting frames. The preregistered falsifier requires a simultaneous collapse of the effect and the validity gate. The observed effect instead grows as the fitting budget

decreases, while gate passage also deteriorates. The narrow tolerance prediction fails, while the joint-collapse falsifier remains unmet. The pass rates in Section 4.2 remain below the 50% crossing criterion throughout the grid. We report n^*_{fit} as censored at the grid boundary; the bootstrap upper endpoint also reaches that boundary.

Zero dependence control. This control sets $w = 0$. The registered target is a rate of false refusals of at most 10% at every $n_{\text{fit}} \geq 2000$. The observed rates, 1.000 at 2000 frames and 0.750 at 5000 frames, exceed this target. A failed residual check withholds a use verdict. These rates quantify how often the protocol declines to interpret an intervention under known zero dependence.

Dependence and random subspace controls. In the dependence sweep, a genuine $w = 0.15$ dependency measured with an eraser fitted with sufficient data produces 0.1241, which exceeds the zero-dependence effect at the smallest fitting budget. Thus, fitting error and genuine dependence can produce effects of comparable magnitude. The random subspace control compares each target intervention with an erasure along a random subspace of matched rank. At $w = 0.4$ and the largest fitting budget, the ratio of target to random effects is $1.941\times$. The $w = 0$ effect remains below the random-subspace reference at every setting, supporting the use of that reference when interpreting effect magnitude.

Collateral overlap and nonlinear coupling. Increasing collateral overlap at fixed dependence yields a PASS rate of 0.125 throughout the sweep. The nonlinear coupling experiment yields velocity sensitivity $0.0229 \rightarrow 1.6772$ while the gate's PASS rate never exceeds 0.375. Both fail the registered prediction that the residual gate would pass throughout. The behavioural measurements increase with collateral overlap and nonlinear coupling, respectively.

Variation across dependence levels. Comparing the $w = 0$ and $w = 0.4$ rows budget by budget gives a maximum pass-rate difference of 0.000. Across all five dependence levels, behavioural effects span $29\times$ at the largest fitting budget in Figure 2.

A.7 Disjoint data roles

Pairwise disjointness of the scene sets in Section 4.4 is verified from scene-token lists. Each experimental record includes its data role, scene count, and frame count.

The 685 scenes are allocated unequally to retain sufficient data for fitting the eraser. Reducing the fitting set can increase both residual failure and behavioural effect. Reducing the audit set widens effect intervals, making behavioural equivalence harder to establish, but also widens the residual null and can reduce the residual-gate failure rate, as quantified in Section A.3.

A.8 Statistical interpretation of the audit

A.8.1 Residual criterion and behavioural equivalence

The behavioural interval rule in Section 3.4 follows equivalence testing with two one-sided tests (Lakens et al., 2018). The residual criterion in Section 3.2 uses an empirically calibrated threshold, with its error rate estimated from the calibration procedure. Below we detail target scaling, residual-scale calibration, and the construction of the behavioural tolerance.

Fitting an eraser for each block. The velocity block's scale is around 3.46 m/s while the acceleration and rotation rate blocks are two orders of magnitude smaller in their respective units. Consequently, an aggregate cross-covariance norm can obscure residuals in smaller-scale blocks. Separate block-specific fits and residual calibrations allow each semantic target to be evaluated on its own scale.

Residual-scale transition. In addition to the injected-leakage calibration in Section 3.3, we characterize the decision boundary on observed residuals. The residual-scale transition r^* describes the realised measurements. A logistic regression of the failure indicator on r_{true} defines r^* as the residual at which the fitted failure probability reaches 80%. The indicator and residual share a noise realization because the gate thresholds that same residual. Consequently, r^* describes the conditional

decision boundary for a particular generator, while β^* quantifies sensitivity to externally controlled leakage. At L0, r^* is 0.2314 for acceleration and 0.1777 for rotation rate. Equal-weight refits yield 0.2306 and 0.1785, respectively.

The exploratory calibration covers full ego state and velocity at four layers. Across these target–layer combinations, β^* spans 0.053 for full ego state at L3 to 0.155 for velocity at L0 and is 0.118 at the layer of the primary arm L0. The oracle operator's failure rate is 9% to 10% at every target–layer pair, while the fitted eraser's $\beta = 0$ floor runs from 18.4% to 28.7%. The fitted floor therefore exceeds the oracle failure rate across layers, with a gap that varies by target and layer.

Detection scale for clean readouts. The readout on clean activations uses a 95%-power detection dose. These doses lie in $[0.10, 0.20]$ and none reaches the preregistered negligible level 0.05. This calibration measures readout sensitivity, whereas β^* measures the sensitivity of the residual test after erasure.

The behavioural bound measures eraser reproducibility. The band on the planning effect is $\delta = \max(\delta_1, \delta_2)$ with $\delta_2 = 0.05\,\text{m}$ fixed in the preregistration and δ_1 set to twice the standard deviation across eraser refits. In the confirmatory run δ_1 dominates and the operative band is δ =0.21007 m. This is an operational tolerance based on eraser reproducibility, rather than a smallest effect of interest grounded in the task. A noisier eraser yields a wider band and can make behavioural equivalence easier to satisfy. The registered rule was retained for the confirmatory analysis, with sensitivity to the resulting tolerance examined below.

The acceleration interval's lower endpoint exceeds the registered tolerance by $4.3\,\text{mm}$. Tolerance sensitivity affects the classifications of acceleration and rotation-rate erasure. At any δ between 0.0656 and $0.2144\,\text{m}$, rotation-rate erasure is classified as equivalent and acceleration erasure as above tolerance. A δ above $0.2973\,\text{m}$ would make acceleration erasure equivalent as well, and one below $0.0656\,\text{m}$ would leave rotation-rate erasure uncertain. The other evaluated erasure conditions have effects of $+0.60$ to $+4.17\,\text{m}$, or $2.9\times$ to $19.9\times$ the operative band; the remaining condition is the Gaussian-noise positive control at L2. The sensitivity thresholds are derived from the observed intervals, with the preregistered rule for δ held fixed.

In the preregistered terminology, an effect above the tolerance contradicts behavioural equivalence.

B Intervention sensitivity and additional analyses

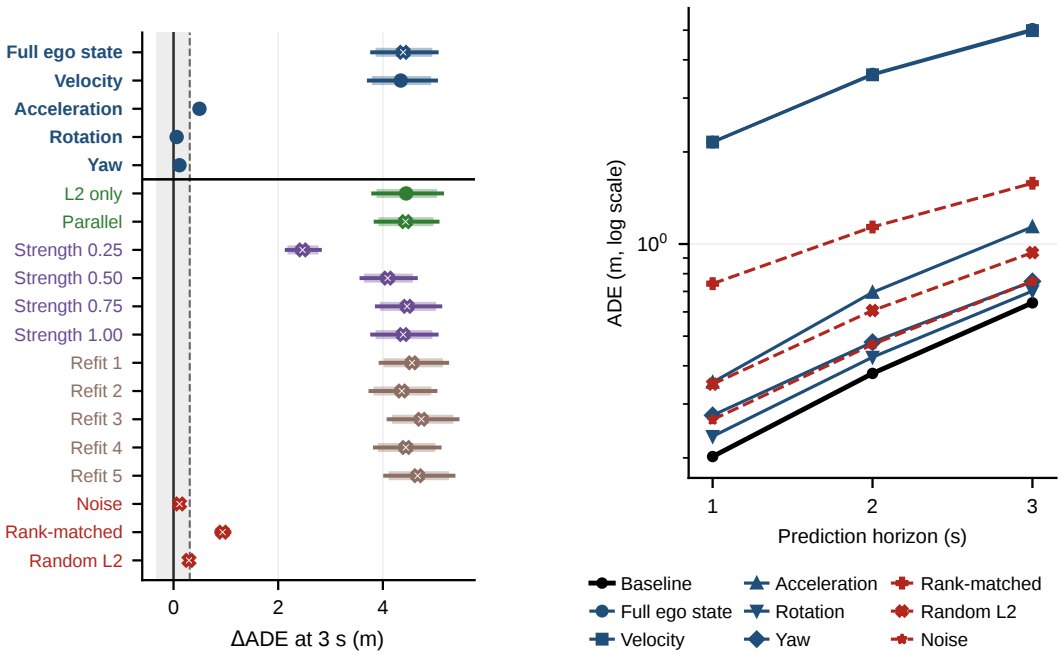
B.1 Response to intervention dose

The fixed-projection sweep in Section 4.5 uses $\alpha \in \{0, 0.25, 0.5, 0.75, 1, 1.5, 2, 4\}$ with one frozen projection. Its registered testing family contains three successive-dose contrasts above $\alpha = 1$, one test of the quadratic vertex against $\alpha = 1$, and two refill comparisons. Five of the six tests reject after Holm adjustment; the erasure-versus-constant-refill comparison is the exception. The refitted sweep applies $x - \alpha P_\alpha(x - \mu_\alpha)$, with a separate fit at each nominal dose. At L0, residual association decreases through $0.7658 \to 0.5325 \to 0.3022 \to 0.0989$ while the threshold stays within $[0.0703, 0.1165]$. ADE effects increase through $+2.3914 \to +3.6646 \to +3.9533 \to +3.9602$ and saturate by the third dose. Each point combines a change in nominal strength with a new fitted projection. The observed trend therefore describes the refitted intervention sequence. The fixed-projection sweep in Section 4.5 isolates scaling one projection.

The frequency of failure across refits depends on the reference distribution. A semisynthetic nested null gives $P(\geq 5 \text{ of } 6) = 0.270$; the corresponding probability on real evaluation activations is 0.753. Both differ substantially from 0.05, indicating that this refit statistic requires calibration to its evaluation distribution.

C Additional figures

These panels report exploratory behavioural effects, residual tests, null sensitivity, intervention dose responses, and behavioural controls.



(a) Effects per arm with intervals over scenes.

(b) Absolute ADE before and after intervention.

Figure 8: **Behavioural effects in the exploratory SSR audit.** The arm that erases the full block has the largest effect and the highest absolute error. Thin and thick intervals denote 95% and 90% bootstrap intervals over scenes; circles and crosses identify passing and failing residual checks; the grey band is $\pm\delta$ with $\delta = 0.309$ m. Results use 1702 valid frames from 50 scenes. Solid horizon curves show erasure interventions and dashed curves show controls. Confirmatory results are reported in Table 3.

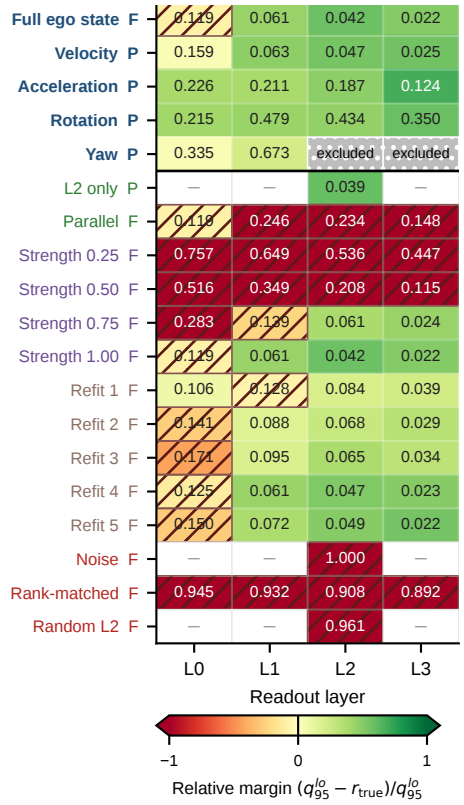
D LAW DIAGNOSTIC DETAILS

LAW's computational path distinguishes explicit ego inputs from visual speed correlates (Li et al., 2024a). At inference, the model constructs an ego-information vector containing historical trajectories, continuous kinematics, and a future command. The trajectory computation uses the discrete turn command but does not consume the continuous kinematic fields targeted in the SSR audit.

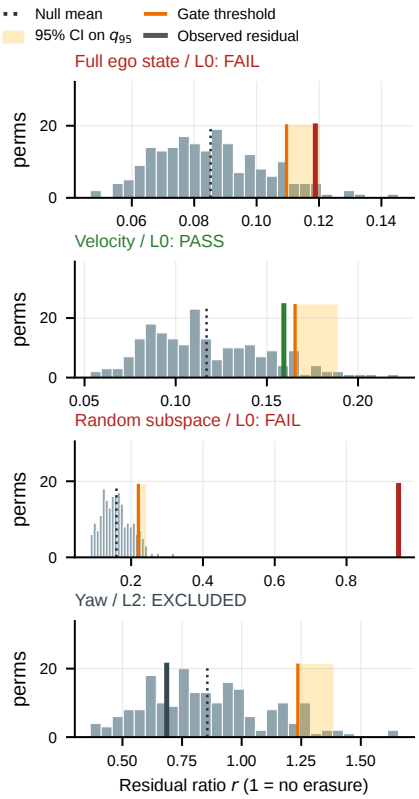
For the W3 path described in Section 4.2, predicted waypoints are broadcast across the world model tokens, and the resulting latent is returned as an auxiliary output. The LAW experiments evaluate frozen readouts directly, with the eraser held fixed.

Computational paths and readout checks. The LAW readouts use the same decision rule with three outcomes, the threshold for negligible effects, and the ridge-regularisation grid as the SSR decodability analysis. The taps and control tests were specified in a separate preregistration. Three taps of LAW's world model path were evaluated on 2002 frames over 50 scenes evaluation samples from disjoint scenes. The earliest tap, W1, contains a visual feature from a single frame. The computational path at this tap contains visual features rather than explicit ego quantities. Six slots at this tap decode longitudinal speed at R^2 between 0.9908 and 0.9928, against 0.5294 for the best slot outside them. W1 lies on the planning path and encodes speed through visual features, so its decodability has a different computational origin from W3. Those six slots correspond to one camera view's six tokens; the physical camera identity remains unverified.

Target and scene controls. First, the velocity target is three columns and the lateral and vertical columns are identically zero on every evaluation frame of this dataset, so they contribute nothing to the denominator of $1 - \mathrm{SSE}/\mathrm{SST}$ and a nonnegative amount to its numerator. Pooling preserves the reported R^2 interpretation because the zero-valued columns contribute no target variation. Second, a predictor that returns each evaluation scene's mean speed uses scene identity alone. Its R^2 is bounded by the variance attributable to differences between scenes, 0.6380. All six slots exceed

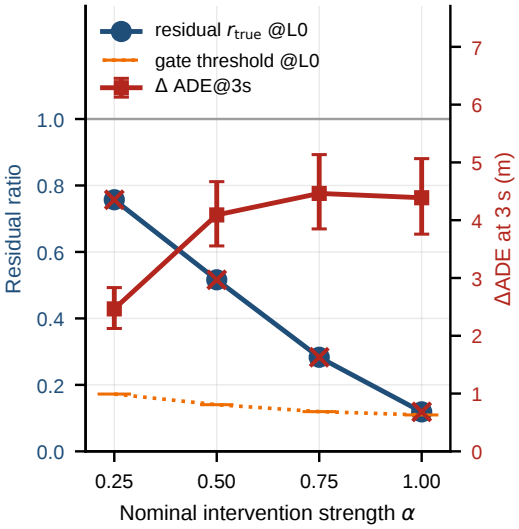


(a) Gate landscape across slots.

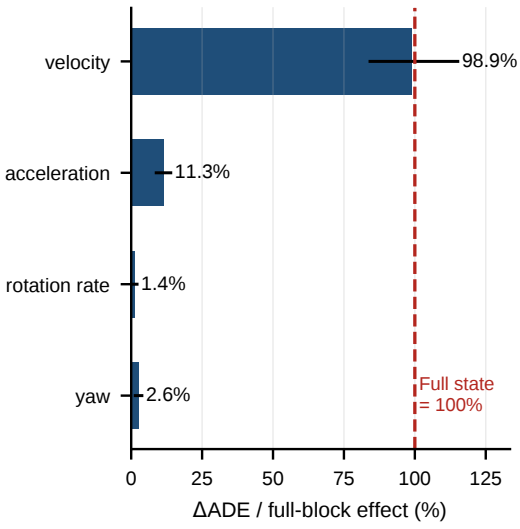


(b) Anatomy of the permutation null.

Figure 9: **Residual decisions and matched null distributions.** Exploratory residuals are shown alongside 200 scene-block label reassignments per slot, allowing the passing margin to be assessed for each arm. Hatching denotes failure; grey slots satisfy $q_{95} \geq 1$, and dashes mark slots without a valid gate measurement.

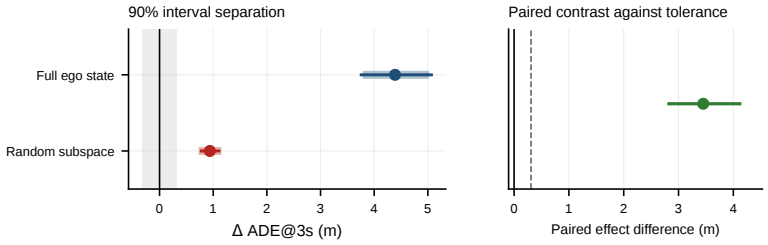


(a) Nominal intervention strength.

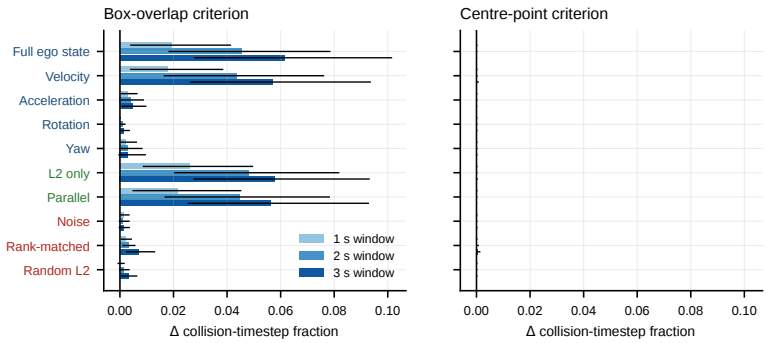


(b) Effects of individual blocks.

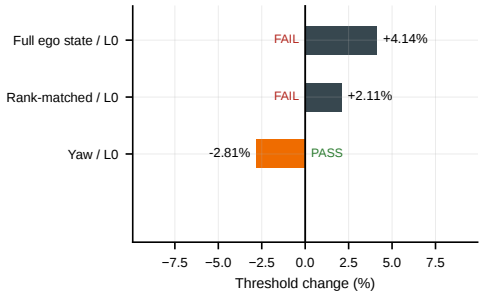
Figure 10: **Response to nominal intervention strength.** The x-axis gives the nominal intervention parameter. Residual linear association is measured separately by G_1'. Crosses denote residual failure at every dose at L0. The horizontal reference is $r = 1$ before erasure. Effects of individual blocks sum to 114.2% of the effect of erasing the full block and therefore do not form an additive decomposition.



(a) Placebo.



(b) Collision family.



(c) Permutation audit.

Figure 11: **Placebo, collision, and permutation controls.** Expected directions were specified before evaluation. (a) Thin/thick lines show 95%/90% intervals; the grey band is $\pm\delta$, and the dashed line marks δ. (b) Collision rates are fractions of timesteps, with both definitions on a shared scale; the largest absolute change under the centre point criterion is 0.00049. (c) Positive threshold changes indicate that the original null is stricter. All three verdicts agree; the passing yaw erasure at L0 margin falls from 3.6% to 0.9%.

this bound, indicating that their readouts capture speed variation within scenes. Third, on the same activations and split, the readout predicts heading angle at a median R^2 of -0.0159 and longitudinal acceleration at $+0.0345$. These target controls show that the strong readout is specific to speed among the quantities tested.

Scope of the LAW evidence. The scene-identity control excludes a predictor based solely on mean speeds between scenes. Occlusion and image-shuffling ablations were outside this analysis, so visual inference and training-set memorisation remain jointly represented. Readouts within the null band have detection floors of $[0.2, 0.3]$, above the negligible level 0.05, so these weaker representations remain indeterminate. The smallest detection floor across the LAW analysis is 0.15, which also remains above the negligible level. The two models use different temporal queues and unpaired evaluation sets, with 2002 frames over 50 scenes frames for LAW and 5421 frames over 135 scenes for SSR. Their detection limits are therefore calibrated and reported separately.

Effect of the instrument-validity check. A binary rule would label some LAW readouts as evidence of absence. The instrument-validity check instead classifies these cases as INSTRUMENT-INVALID and withholds a representational verdict.