WHEN DO CROSS-MODAL GAINS WARRANT A SEMANTIC ROLE? A ROLE-CONDITIONED AUDIT OF DERIVED PAIRING RULES

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal gains can change learning, but they do not by themselves show that the assigned relation explains the gain: a nuisance scaffold can reproduce the same endpoint pattern. The cross-modal alignment spuriousness (CMAS) audit makes this premise-failure gap explicit through three layers: phenomenon (whether a gain is observed), mechanism (which information and computational routes are active), and boundary (whether nuisance-relative utility clears declared margins). In a five-seed DOLOS case restricted to one corpus, carrier/backbone, split, and validation-only known-offset endpoint, the primary provenance–time contrasts are inconclusive at the declared margins, while phase is recoverable but phase-selected directions do not outperform rank-matched random subspaces. A second same-carrier role is likewise not authorized at this scale. The operational conclusion is bounded: cross-modal gains can change learning, but they do not by themselves authorize the assigned semantic role beyond declared alternatives.

1 INTRODUCTION

A cross-modal gain can be genuine yet fail to support its assigned relation: the same endpoint improvement can arise from that relation or from locality, identity, leakage, and routing. Which explanation does the gain authorize? Prior cross-modal studies establish learning effects (Arandjelović & Zisserman, 2017; Alwassel et al., 2020; Shi et al., 2022; Sarkar & Etemad, 2023), whereas semantic interpretation requires relation-specific utility beyond alternatives (Lapuschkin et al., 2019; Geirhos et al., 2020). We challenge the premise that gain identifies its causal source.

This premise failure is sharpest when a derived variable selects cross-modal positives. The rule creates *structural positives*: pairs attract by construction. Calling them *semantic positives* adds the falsifiable claim that the selected relation explains correspondence beyond declared nuisance structure. Our premise-failure question is: *when $X = f(M_1)$ selects which observations from M_1 and M_2 attract, what evidence authorizes that interpretation?* A positive endpoint alone does not identify the gain's causal source.

Existing diagnostics establish activity, recoverability, or route dependence (Hewitt & Liang, 2019; Ravichander et al., 2021; Rozanova et al., 2023); none alone shows utility beyond a declared alternative. For phase, a role-bearing direction should yield a different endpoint response from a rank-matched random intervention.

CMAS fixes the role claim, alternative, independent endpoint, decision scale, and route test before interpreting a gain (Figure 2). Three layers establish the phenomenon (gain), probe the mechanism (information and mask/weight paths), and delimit the boundary (authorization conditions). Later layers constrain what an earlier gain can mean.

Contributions and scope. *Paradigm.* We propose a cross-modal representation-learning paradigm: a gain is an authorization prompt, not a semantic conclusion; the relation earns support only after nuisance and route checks. Past cross-modal work often treats an endpoint gain as evidence for the assigned relation; this audit requires that gain to survive declared alternatives. *Instance.* CMAS fixes the role claim, alternatives, and discriminating evidence. *Interface.* Its role–control–endpoint–scale–route record joins interpretability and statistical evaluation. *Case.* A five-seed DOLOS audit

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.

instantiates it. It changes what a gain means: from a semantic conclusion to an authorization prompt. Sections 3–6 report the audit.

2 Related Work

What cross-modal supervision establishes. Co-occurrence, synchrony, cluster assignments, and discovered multimodal units drive useful representation learning (Arandjelović & Zisserman, 2017; Afouras et al., 2020; Chung & Zisserman, 2017; Alwassel et al., 2020; Shi et al., 2022). Relaxed-synchrony experiments show that temporal relations change representations (Sarkar & Etemad, 2023). CMAS tests whether the intended relation accounts for this usefulness beyond accompanying structure.

From task success to semantic authorization. Task scores leave their producing mechanism underdetermined (Lapuschkin et al., 2019; Geirhos et al., 2020). The dominant inference treats gain as evidence that the constructed relation supplied the useful signal, implicitly assuming a unique causal source. Yet a nuisance scaffold can reproduce the same success (Yüksekgönül et al., 2023; Yang et al., 2025; Arazo et al., 2020). CMAS adjudicates between these explanations instead of cataloguing another diagnostic.

What existing diagnostics leave open. Probes depend on capacity and controls (Hewitt & Liang, 2019), and recoverability need not imply task use (Ravichander et al., 2021). Interventional Probing compares removal or retention of probe-identified directions with rank-matched random subspaces (Ravfogel et al., 2020; Elazar et al., 2021; Rozanova et al., 2023). If phase-predictive directions receive privileged endpoint use, targeted interventions should differ from rank-matched random controls. We test this expectation on the DOLOS visual representation and correspondence endpoint.

The carrier is a worked audit case, not an external benchmark. The adapted comparison tests phase-predictive directions on the endpoint; CMAS tests the pairing role against nuisance-relative utility and applicable route criteria. These are distinct estimands, so the cited diagnostic traditions provide background reference rather than direct same-endpoint competitors.

Authorization gap. Derived-target methods establish usefulness, shortcut tests expose alternative cues, and representation interventions test selected directions. CMAS joins these observations into an identification contract: a role claim, nuisance scaffold, independent endpoint, meaningful-effect scale, and any applicable route jointly determine whether a pairing rule earns semantic support when intended and nuisance mechanisms predict the same gain.

3 Problem Formalization

To address the authorization gap, CMAS fixes the role claim, nuisance control, endpoint, decision scale, and applicable route before interpreting model behavior. The resulting estimands distinguish activity, recoverability, utility, and route dependence.

3.1 Derived Pairing Supervision

Let paired multimodal sequences be $\mathbf{M}_{1b} = \{m^1_{bj}\}_{j=1}^{T_1}$ and $\mathbf{M}_{2b} = \{m^2_{bi}\}_{i=1}^{T_2}$, and let $X_b = f(\mathbf{M}_{1b})$. A derived pairing rule maps this single-modality variable and any declared coordinates to an eligibility matrix and nonnegative weights,

$$S_X : (X_b, \mathbf{t}_b) \mapsto (G^X_b, W^X_b), \qquad G^X_b \in \{0,1\}^{T_2 \times T_1}, \quad W^X_b \in \mathbb{R}^{T_2 \times T_1}_{\geq 0}. \tag{1}$$

For an anchor with nonzero eligible mass, the rule-imposed pairing distribution is

$$\rho_X(j \mid b, i) = \frac{G^X_{bij} W^X_{bij}}{\sum_k G^X_{bik} W^X_{bik}}. \tag{2}$$

A pair with $G^X_{bij} = 1$ constitutes a *structural positive*: the construction selects it computationally. A *role claim* $C_X = (\mathcal{R}_X, \mathcal{N})$ interprets the X-specific variation in ρ_X as a stated cross-modal relation

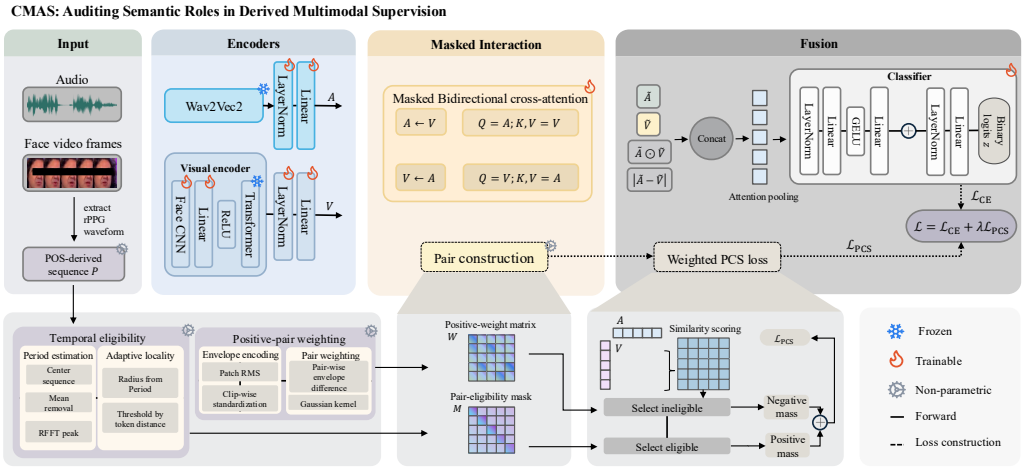


Figure 1: **The derived pairing rule can alter learning through two coupled computational paths, so an endpoint gain is non-identifying by itself.** The visual stream produces the POS-derived sequence P, which enters pair construction together with temporal and envelope rules. The resulting eligibility mask and positive-pair weights affect masked bidirectional interaction and the weighted PCS loss, while the fused representations feed the classifier. Frozen, trainable, and non-parametric components follow the legend in the diagram; solid and dashed arrows distinguish forward computation from loss construction. This architecture map specifies implementation paths only and does not by itself authorize the semantic role assigned to P.

$\mathcal{R}_X$ beyond nuisance information $\mathcal{N}$. We call the construction *derived pairing supervision*; semantic authorization is the evidence state sought for its role claim rather than a label supplied by Equation 1.

Training under rule S and seed ξ produces $\theta_S^{(\xi)} = \mathrm{Train}(\mathcal{D}_{\mathrm{tr}}; S, \xi)$. Semantic authorization describes an evidence state: relative to a valid nuisance rule, a role-relevant endpoint must support the implemented rule package, with compatible use evidence where applicable. In the case study, M_1 is video, M_2 is audio, $X = P$, and $\mathcal{N}$ contains physical-time locality. The claim concerns the complete phase/envelope pairing package beyond local time; it does not identify separate phase, mask, weight, speech–pulse, or physiological effects.

3.2 How Derived Supervision Acts

The audited mechanism instantiates $G^P = M^P$ and $W^P = w^P$. It is not an extra feature appended to the classifier. Instead, the rule acts through two coupled paths: the mask gates bidirectional cross-attention and therefore the carrier-task forward map, while the same mask and its weights define positives in an auxiliary contrastive loss. Given a normalized audio anchor a_{bi} and visual token v_{cj}, let $s_{bicj} = a_{bi}^\top v_{cj}$. Real-P defines same-clip eligibility and weights through phase ϕ, envelope E, and physical time t:

$$\begin{aligned} M^P_{bij} &= \mathbb{1}[|t_i - t_j| \le \delta_t]\,\mathbb{1}[|\operatorname{wrap}(\phi_{bi} - \phi_{bj})| \le \delta_\phi], \\ w^P_{bij} &= \exp\!\left[-(E_{bi} - E_{bj})^2/(2\sigma^2)\right]. \end{aligned} \tag{3}$$

The local-time rule retains only the first indicator, $M^{\mathrm{time}}_{bij} = \mathbb{1}[|t_i - t_j| \le \delta_t]$, and sets $w^{\mathrm{time}}_{bij} = 1$. For valid anchors with at least one eligible positive, define $A^S_{bi} = \sum_j M^S_{bij} w^S_{bij} e^{s_{bibj}/\tau}$ and $Z_{bi} = \sum_{c,j} e^{s_{bicj}/\tau}$. Under either rule $S \in \{P, \mathrm{time}\}$, the pairing loss is

$$\mathcal{L}^S_{\mathrm{PCS}} = -\frac{1}{BT}\sum_{b,i} \log \frac{A^S_{bi}}{Z_{bi}}, \qquad \mathcal{L}^S = \mathcal{L}_{\mathrm{CE}}(\theta; M^S) + \lambda \mathcal{L}^S_{\mathrm{PCS}}. \tag{4}$$

Figure 1 makes this implementation path explicit; it describes computation, not a semantic or causal conclusion, which is assessed by the independent endpoint and A1–A4 evidence roles.

This objective identifies the direct source of influence. Define the model distribution over all visual candidates and the similarity-tilted supervision-positive distribution as

$$\pi_{bicj} = \frac{e^{s_{bicj}/\tau}}{Z_{bi}}, \qquad q^S_{bicj} = \frac{\mathbb{1}[c = b] M^S_{bij} w^S_{bij} e^{s_{bibj}/\tau}}{A^S_{bi}}. \tag{5}$$

The exact similarity gradient of the averaged pairing loss, conditional on fixed contrast-token similarities, is

$$\frac{\partial \mathcal{L}^S_{\mathrm{PCS}}}{\partial s_{bicj}} = \frac{1}{BT\tau} \left(\pi_{bicj} - q^S_{bicj} \right). \tag{6}$$

The direct contrastive effect of changing rules is therefore

$$\left. \frac{\partial (\mathcal{L}^P_{\mathrm{PCS}} - \mathcal{L}^{\mathrm{time}}_{\mathrm{PCS}})}{\partial s_{bicj}} \right|_{\{s_{bicj}\} \text{ fixed}} = \frac{q^{\mathrm{time}}_{bicj} - q^P_{bicj}}{BT\tau}. \tag{7}$$

Thus P changes both which cross-modal interactions are visible and how the auxiliary loss reallocates attractive mass. Equation 6 clarifies why fit to the rule's own labels can be circular evidence for semantics: optimization directly moves the model distribution π toward the rule-induced target q^S. Equation 7 isolates this direct contrastive path; it is not a decomposition of the full parameter update because M^S also changes the forward computation. A1 empirically verifies that the weighted auxiliary gradient is active. Activity and imposed-label fit therefore document optimization influence, whereas role support requires utility on a P-independent correspondence endpoint relative to S_{time} and, where an explicit P residual exists, behaviorally relevant dependence on that residual. This distinction motivates A1–A4.

Boundary map from the two paths. The conceptual map separates coupled M1 attractive-mass and M2 forward-gating channels from four boundaries. W1 treats nuisance separability as a moderator: donor, derangement, and identity controls should weaken or reverse a gain caused by provenance or subject structure, while leakage remains outside their gates. W2 treats endpoint scale as a moderator: the finite grid tests whether the contrast clears, approaches, or falls below the declared anchors. W3 treats route compensation as a moderator: a specified deletion route should change the endpoint if the assigned residual is privileged, but natural A4 is unavailable here. W4 treats carrier/population scope as a transport boundary, not an observed moderator. The current controls therefore provide initial boundary tests but no complete strength/weakness/reversal curves; W1–W4 remain contracts rather than completed moderation results.

Mechanism competition after the boundary map. Time-only tests locality; donor/derangement test provenance and alignment; D1 tests subject-preserved identity on common support; split and endpoint gates address overlap and answer-key reuse. These controls narrow those explanations, while broader leakage remains open. Natural A4 is unavailable and route evidence is synthetic-only; mask, weight, density, geometry, smoothness, noise, compensation, and architecture-wide route explanations remain compatible, so transfer is out of scope.

3.3 CMAS Definition

For this pairing subtype, "failure" means failure of the declared semantic-authorization criterion at the stated endpoint and margin. An audit configuration $\mathcal{A} = (C_X, \mathcal{P}, S_{\mathrm{real}}, S_{\mathrm{ctl}}, V, \delta_{\mathrm{sem}}, \mathcal{G}, \delta_{\mathrm{del}})$ fixes the role claim, population, supervision rules, endpoint, margins, and applicable routes. A1 tests activity, A2 tests probe-bounded recoverability, A3 tests control-relative utility, and A4 tests a specified carrier route. Each margin is externally grounded or labeled author-chosen. Semantic authorization under $\mathcal{A}$ requires valid, sufficiently resolved evidence at the declared scale and compatible route evidence. A bounded diagnosis additionally requires A1 activity and an A3 upper bound below $+\delta_{\mathrm{sem}}$. When an explicit route applies, it also requires the A4 interval to lie within $[-\delta_{\mathrm{del}}, +\delta_{\mathrm{del}}]$. Boundary overlap is inconclusive, and unavailable optional routes are not applicable. These states describe C_X in the audited carrier and population.

Throughout, *derived pairing supervision* names the rule-defined construction, and *semantic authorization* names its evidence state relative to C_X; *role claim* denotes the interpretation under audit. The four roles guard against inert activity, probe-limited recovery, nuisance-explained utility, and dispensable routes.

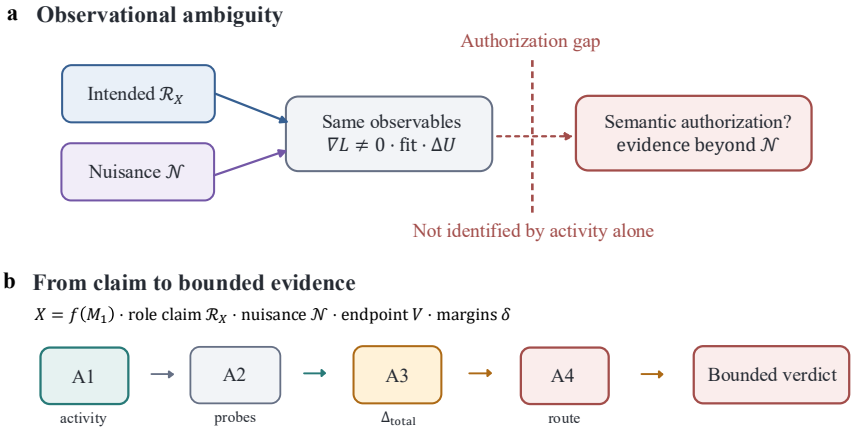


Figure 2: **Semantic authorization requires evidence beyond shared observables.** Panel (a) contrasts intended and nuisance explanations; panel (b) fixes $X = f(M_1)$, the role, nuisance, independent endpoint, and decision margins. A1–A4 are evidence roles rather than mandatory stages or significance votes. The natural DOLOS contract uses A3, leaves A4 N/A, and is inconclusive at the declared scale.

3.4 Estimands and Boundaries

Let $V(\theta; \mathcal{P})$ be a fixed endpoint on population $\mathcal{P}$, aggregated over the declared training seeds. For pairing supervision, the audited-control utility estimand is

$$\Delta_{\text{sem}} = \mathbb{E}_\xi \left[V(\theta^{(\xi)}_{\text{real}}; \mathcal{P}) - V(\theta^{(\xi)}_{\text{ctl}}; \mathcal{P}) \right]. \tag{8}$$

In our case, S_{ctl} preserves the physical-time window but does not match Real-P in pair density or weights. Thus, Δ_{sem} is baseline-conditional rather than a content-only effect. The corresponding mask contrast is package-level: changing the mask also changes positive density, pairing geometry, induced smoothness, and effective noise. The design therefore tests the implemented package, while leaving these channels unresolved as separate components.

For an explicit route, we compare two evaluations of the same trained model. The $H^{(+X)}_{\theta_\xi}$ arm retains the residual. The $H^{(-X)}_{\theta_\xi}$ arm deletes only that residual. Both arms share the weights and pairing mask. The deletion estimand therefore asks whether this specified route is needed after training, rather than whether X was useful anywhere in the training process. We define

$$\Delta_{\text{del}} = \mathbb{E}_\xi \left[V(H^{(-X)}_{\theta_\xi}; \mathcal{P}) - V(H^{(+X)}_{\theta_\xi}; \mathcal{P}) \right]. \tag{9}$$

Positive Δ_{del} means deletion improves the endpoint; negative values mean deletion harms it. Practical route dependence is assessed against $[-\delta_{\text{del}}, +\delta_{\text{del}}]$, not zero significance alone. The estimate therefore speaks only to the specified residual at the declared scale; it does not test compensation through other routes or effects acquired earlier in training.

For the component analysis, the intermediate R-unit keeps the Real-P mask but replaces its non-uniform weights by one. Time-only keeps only the physical-time mask. The three-arm comparison accounts for how the complete rule differs from the control, but it does not make mask and weight effects independent causal components. We write $R\text{-unit} = (M^P, \mathbf{1})$ and $\text{Time-only} = (M^T, \mathbf{1})$. We report the complete contrast and two conditional contrasts:

$$\begin{aligned} \Delta_{\text{total}} &= U(M^P, w^P) - U(M^T, \mathbf{1}), \\ \Delta_{\text{weight}} &= U(M^P, w^P) - U(M^P, \mathbf{1}), \\ \Delta_{\text{mask}} &= U(M^P, \mathbf{1}) - U(M^T, \mathbf{1}). \end{aligned} \tag{10}$$

The weight contrast changes weighting within the Real-P mask; the mask contrast adds phase-qualified selection at unit weight. Because R-unit is intermediate, $\Delta_{\text{total}} = \Delta_{\text{weight}} + \Delta_{\text{mask}}$ exactly.

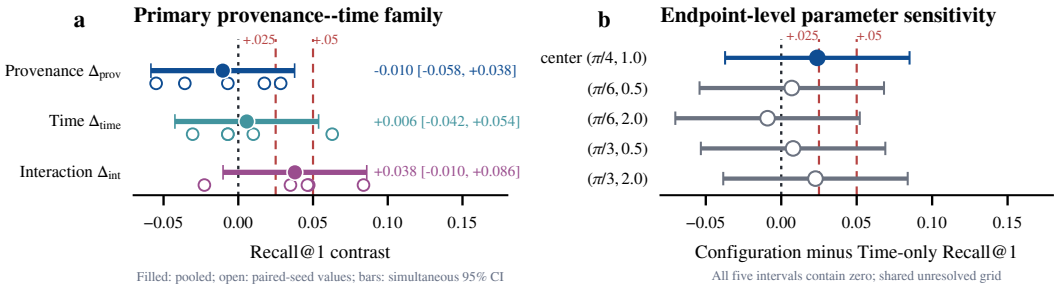


Figure 3: **Primary source–time contrasts are inconclusive at the predeclared boundaries.** Panel (a) uses Recall@1 contrast on the x-axis and contrast family on the y-axis; filled points are pooled five-seed estimates, open points are the five paired-seed values, and bars are familywise simultaneous 95% seed-first/source-group bootstrap CIs. Panel (b) uses configuration-minus-Time-only Recall@1 on the x-axis and endpoint-blind configuration on the y-axis; the center is filled and the four corners are open, each with its five-seed pooled interval. Both panels use five paired seeds, 191 validation clips, and six nonzero offsets (1,146 queries per cell); red lines mark $+0.025$ and $+0.05$. Exact sign-flip/Wilcoxon sensitivities appear in the tables; boundary crossing is not equivalence or absence.

Table 1: **Distinct evidence roles support a configuration-specific verdict.** The matrix separates generic interfaces from the five-seed DOLOS contract; the configured role, control, endpoint, and boundaries define the case verdict.

Gate	Evidence	Threshold	Case output	Contribution to verdict
A1	Update activity	nonzero update ratio/rotation	Generic; not used in current verdict	Establishes optimization influence.
A2	Probe recovery	named recovery gates	Generic; not used in current verdict	Establishes probe-bounded availability.
A3	Control-relative utility	CI vs. $0, +0.025, +0.05$	Primary CIs cross boundaries	Current resolution leaves the verdict unresolved.
A4	Route dependence	route intervention	N/A in this contract	Applies when the role names a route.
Sec.	Identity/parameter controls	common support; finite grid	Identity below zero; parameters unresolved	Controls narrow interpretation.

This is telescoping accounting; the mask contrast changes placement, density, geometry, and local structure jointly. Because neither conditional contrast fixes positive density, pairing geometry, induced smoothness, or effective noise, neither isolates an independent mask or weight channel.

4 Semantic Audit Framework

Equations 6–7 show how a rule changes interactions and attractive mass without deciding what the change means. Four evidence roles sit within the chain: A1 asks whether optimization is active, A2 whether role-relevant information is recoverable, A3 whether the complete package improves the independent endpoint beyond the nuisance rule, and A4 whether a specified route matters after training. They are not a four-way significance vote or a mediation decomposition. The audit then follows three layers: phenomenon (activity and availability), mechanism (package utility and routes), and boundary (population, seed unit, uncertainty, and margins). A3 is the natural package-level verdict; the synthetic lane exercises the otherwise unavailable A4 route as a positive control. Figure 1 maps the two paths, Figure 2 fixes the authorization contract, Figure 3 tests the primary boundary, and Figure 4 reports descriptive heterogeneity; Table 1 records each role.

We separate four roles because a single downstream gain cannot distinguish their failure modes: A1 rules out inert activity, A2 locates probe-bounded information but does not establish endpoint privilege, A3 tests nuisance-relative utility, and applicable A4 tests route necessity under compensation. They are evidence gates, not a mediation/moderation decomposition or four significance votes.

A1: From a rule on paper to a force on learning. A1 measures $r_\lambda = \|\lambda \nabla \mathcal{L}_{\mathrm{PCS}}\| / \|\nabla \mathcal{L}_{\mathrm{CE}}\|$ and the post-clipping rotation from the CE-only update. Finite, nonzero values show that the contrastive path changes optimization; the attention-mask path is assessed separately. A1 establishes the computational influence that a semantic interpretation must connect to the remaining role-specific evidence.

A2: Probe recoverability as convergent evidence. Recoverability depends on layer, probe family, and capacity (Hewitt & Liang, 2019; Ravichander et al., 2021). A single linear probe could confuse recoverability with probe capacity, so Ridge, a nine-token local TCN, and a leave-one-subject-out full-sequence TCN provide complementary sensitivity controls. Passing supplies localization evidence; failing leaves distributed or transient encodings unresolved. A2 establishes probe-bounded information availability, while A3 and applicable A4 determine how that information enters the authorization record.

A3: Whether the asserted rule improves over the audited nuisance rule. A3 is the central pairing test. Time-only retains physical-time eligibility while removing phase-qualified gating and envelope weighting. The contrast estimates what the complete P rule adds beyond local temporal structure. With Δ_{sem} and declared margin δ_{sem}, an interval above the margin supports an advantage at that scale, an upper bound below it withholds authorization at that scale, and overlap is inconclusive. Because Time-only also changes density, geometry, smoothness, and effective noise, the estimand is package-level rather than an isolated phase effect.

A4: Whether an available route becomes necessary. A4 tests whether a specified carrier route matters after training. If the P-specific route carries the asserted role, deleting its residual should lower the endpoint; if the apparent gain is carried by the local-time scaffold, deletion should be near-inert. We therefore create a residual containing $(\sin\phi, \cos\phi, E)$, verify its availability at the route entrance, and delete only that residual at inference while retaining weights and the pairing mask. With Δ_{del} defined as deletion minus injection, intervals beyond $\pm\delta_{\mathrm{del}}$ resolve route dependence and containment within the region supports route-specific practical indifference. The estimand concerns this residual route; compensation, other routes, and earlier-training effects remain outside it.

Joint verdict. The evidence is a progression, not a vote: A1 establishes activity, A2 recoverability, A3 nuisance-relative package utility, and applicable A4 route dependence. Support requires an A3 interval entirely above the margin plus compatible route evidence when specified; an upper bound below the margin, with A1 activity and applicable A4 indifference, yields bounded failure; overlap is inconclusive. A predeclared above-margin A3 interval with compatible route evidence would authorize the role at that scale; boundary overlap leaves the role undecided. These are configuration-specific decision rules, not a universal theorem.

The interfaces become testable only after fixing a role, nuisance control, independent endpoint, applicable route, and resolution rule. Here only A3 contrasts are natural-role evidence; A1, A2, and A4 are generic interfaces, with synthetic A3/A4 serving as calibration. The portable object is the authorization question and evidence structure; transfer requires a new carrier-specific contract.

5 Case Study and Setup

Why this design answers authorization. The independent known-offset endpoint, nuisance-control matrix, paired-seed bootstrap, and conditional route tests target authorization rather than downstream fit; a new carrier must redeclare each element.

Data and audited carrier. We study 955 DOLOS clips from five shared subjects under the existing parameter-efficient cross-modal carrier model (Guo et al., 2023), with a 764/191 train/validation split. The endpoint uses only the 191 validation clips; the full 955-clip corpus is retained for split-aware construction, donor maps, and leakage audits. The official test set is unread. Facial video is converted into a fixed POS-derived sequence P following the established construction (Wang et al., 2017). Audio and video become $T = 64$ physical-time-aligned tokens. The parameter center is $\delta_\phi = \pi/4$, $\sigma = 1.0$, and $\tau = 0.07$, with endpoint-blind corners $(\pi/6, 0.5)$, $(\pi/6, 2.0)$, $(\pi/3, 0.5)$, and $(\pi/3, 2.0)$. The endpoint is parameter-free known-offset direct Recall@1 over six nonzero

injected offsets. For an evaluator-injected offset, its negative is the correct recovery candidate and must rank first; the evaluator does not use the pairing mask as its answer key. This is a single-corpus, single-backbone case study: other datasets and backbones remain untested, and no external performance leaderboard is claimed.

The primary comparison includes Self-Aligned-P, Self-Deranged-P, D0-Donor-Aligned-P, D0-Donor-Deranged-P, D1-Subject-Preserved-Donor-P, and Time-only. D0 is a split-preserving donor permutation; D1 preserves donor subject identity on fixed common support and uses a prespecified fallback elsewhere; derangement preserves recipient-wise values while destroying alignment. All five paired seeds use the same FP32 model, data path, epoch-10 checkpoint rule, and endpoint evaluator.

Controls. Time-only retains local temporal eligibility with unit weights. Donor/derangement vary provenance and alignment; identity and parameter controls are secondary and endpoint-blind. Real-L and synthetic paired/decoy/deletion lanes are bounded controls, not additional tasks; no natural route-deletion or external-ranking claim is made, and Interventional Probing is diagnostic only.

Same-carrier adaptation of Interventional Probing. We adapt the protocol studied by Rozanova et al. (Rozanova et al., 2023) as a same-carrier diagnostic on the frozen DOLOS visual contrast-token representation. The adaptation retains its amnesic-removal, mnestic-retention, and rank-matched random-subspace logic. Position-shared no-intercept ridge probes are fitted on 764 training clips with four fixed iterations and evaluated at five fixed checkpoints on 191 validation clips and six nonzero offsets. The endpoint is DOLOS known-offset Recall@1 rather than the original NLI endpoint. Its arm-level means are listed in Table 3; the comparison is a diagnostic on this endpoint.

Metrics, uncertainty, and resolution. A3 uses five seeds and a 10,000-draw seed-first/source-group bootstrap with familywise 95% CIs. Exact sign-flip/Wilcoxon tests enumerate the five paired-seed effects (minimum two-sided resolution 0.0625) and answer a different question from the intervals. d_z is descriptive; anchors 0, $+0.025$, and $+0.05$ are task-scale references (Anvari & Lakens, 2021). The simultaneous CIs are primary: an interval above a margin supports that role, an upper bound below it withholds authorization at that scale, and overlap is inconclusive. Here all primary intervals contain 0 and $+0.025$; only provenance falls below $+0.05$. No prospective power analysis was conducted; Appendix A reports post-hoc sensitivity, not a design-stage MDE or CI replacement. These bounds do not support a no-effect claim.

6 Results: What the Audit Reveals

Only A3-style contrasts are natural-role evidence; A1, A2, and A4 remain generic interfaces. Results answer three linked questions: phenomenon—the primary A3 contrasts move but do not clear a declared boundary; mechanism—phase is recoverable, while matched-random interventions do not establish privileged endpoint use; boundary—identity/Real-L controls and the finite grid narrow a package-level, case-bounded interpretation. Natural A4 is N/A; these layers map conditions, not votes or a continuous moderator.

Primary endpoint evidence. The primary arms do not move together: donor, derangement, and subject-preserved controls span both sides of Time-only, so a positive arm cannot by itself establish relation-specific utility. Across five paired seeds, the full-191 arm means $\pm$ between-seed SD are Self-Aligned-P 0.36318 ± 0.02280, Self-Deranged-P 0.33857 ± 0.02701, D0-Donor-Aligned-P 0.35462 ± 0.04391, D0-Donor-Deranged-P 0.36789 ± 0.02365, D1-Subject-Preserved-Donor-P 0.31501 ± 0.01948, and Time-only 0.33927 ± 0.04598. These are descriptive aggregates; Figure 3a reports the paired contrasts with seeds as the unit. Paired effects mix directions; the provenance–time interaction has the largest positive point estimate, but every simultaneous interval contains at least two anchors. Task-scale contrasts are -0.0104, $+0.0057$, and $+0.0379$, with $d_z = -0.296$, $+0.162$, and $+0.983$; these standardize five-seed directions rather than semantic strength, and the interaction seed effects remain mixed. Figure 3 and Table 6 give exact values and seed tests. At the declared scale, the family is inconclusive: these intervals establish neither equivalence nor absence.

Mechanism and alternative diagnostics. The adaptation detects phase ($p_{\text{joint}} = 0.005$ from 199 synchronized trajectory permutations), but detectability does not establish endpoint use. Time-

only preserves local temporal eligibility, so Real-P's added gating may have limited room; density, geometry, and smoothness remain coupled explanations. Exploratory five-seed sign-flip tests give means/p values of $-0.0009/1.0000$ (removal) and $-0.1002/0.0625$ (retention): removal straddles zero, while retention is below matched random. Privileged linear-subspace use is unsupported here; nonlinear, distributed, or compensated use remains compatible.

On the fixed 138-clip common-support cohort, the subject-preserved donor interval is below zero (Table 6); the fallback-inclusive full-191 contrast is -0.0396 with exploratory 95% CI $[-0.0937, +0.0140]$, outside the simultaneous family.

Validation-only Real-L is inconclusive ($\Delta_{L,\text{total}} = -0.0201$, simultaneous 95% CI $[-0.0587, +0.0185]$); the 955-clip endpoint is excluded. The interval leaves the complete Real-L package unresolved at this scale, rather than identifying a missing component. Synthetic A3/A4 detects the constructed dependence (A3: $+0.6956$; 95% CI $[+0.6627, +0.7332]$; A4: -0.6890; 95% CI $[-0.7222, -0.6511]$), calibrating large effects rather than natural semantics or near-margin power.

Figure 3b and Table 5 give the finite grid: estimates range from -0.0091 to $+0.0239$, and every simultaneous interval contains all three anchors. This shared indeterminacy constrains the declared grid; it does not imply parameter irrelevance or a flat continuous response. The table reports five-seed mean $\pm$ sample SD, not the simultaneous intervals.

6.1 Endpoint-blind heterogeneity and failures

The 27 cases were fixed before endpoint inspection from subject-stratified mask, weight, and donor features; sign counts relative to Time-only are 11/10/6 for Self-Aligned-P and 9/9/9 for D0-Donor-Aligned-P (Figure 4), documenting descriptive heterogeneity rather than a failure taxonomy or population rate. The validation predictions also summarize the lowest-margin 19 (9.9%): Self-Aligned-P Recall@1 is 0.1667 versus 0.3632 overall, offset error 2.521 versus 1.076 tokens, and margin -0.0120 versus -0.0022; D0 changes Recall@1 by 0.0000 and D1 by $+0.0018$. This is observable concentration, not calibrated confidence or a causal taxonomy; temporal, identity-leakage, and phase-use labels are unavailable.

7 Discussion and Limitations

What the controls exclude and retain. Transfer to other corpora or backbones is untested. Time-only leaves locality compatible; donor, derangement, and identity controls narrow but do not isolate explanations. D1 is cohort-bounded and compatible with donor shift, lost nuisance variation, alignment loss, or package geometry; split/endpoint gates address overlap and answer-key reuse, while A4 is synthetic-only and the cases remain descriptive.

Phase-detectability paradox. The central reversal is a dissociation: privileged use predicts phase-predictive directions should outperform rank-matched random directions, yet phase is detectable ($p_{\text{joint}} = 0.005$), removal has no advantage, and targeted retention is lower. A2 establishes availability; A3 tests endpoint privilege. Phase-targeted interventions may therefore act on detectable information that the endpoint does not privilege. Compensation, nonlinear/distributed coding, and package-level A3 mixing remain compatible but unidentified. The boundary is failure to demonstrate privileged use in this carrier, endpoint, and five-seed contract.

8 Conclusion

CMAS changes the interpretation order: first record activity, then test mechanism, then ask whether package utility clears the declared boundary. Here source–time and Real-L are inconclusive, phase is recoverable, and matched-random privilege is not established. Treat gains as audit triggers: recoverability cannot substitute for endpoint privilege; semantic deployment requires a predeclared above-margin A3 contrast plus compatible route evidence.

Ethics and Reproducibility Statements

Ethics. No new participants; fairness and deployment validity remain unmeasured. **Reproducibility.** Appendix A and anonymous replay provide de-identified statistics.

AI Use Statement

Generative AI assisted conceptual framing, mathematical exposition, hypothesis and experiment-design feedback, interpretation of author-provided results, literature retrieval, translation, drafting and language editing, and figure/code editing. It did not generate data, clean or reformat datasets, conduct qualitative analysis, supply proofs, implement or train the audited model, or replace statistical validation. The authors verified citations, code, figures, numerical results, and claims and take responsibility for the final content.

References

Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision – ECCV 2020*, pp. 208–224, 2020. doi: 10.1007/978-3-030-58523-5_13.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9758–9770, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6f2268bd1d3d3ebaabb04d6b5d099425-Abstract.html.

Farid Anvari and Daniel Lakens. Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, 96:104159, 2021. doi: 10.1016/j.jesp.2021.104159.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017. doi: 10.1109/ICCV.2017.73.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks*, pp. 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207304.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, pp. 251–263, 2017. doi: 10.1007/978-3-319-54427-4_19.

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/.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex C. Kot. Audio-visual deception detection: DOLOS dataset and parameter-efficient crossmodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22135–22145, 2023. doi: 10.1109/ICCV51070.2023.02023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4.

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, 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*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Julia Rozanova, Marco Valentino, Lucas Cordeiro, and André Freitas. Interventional probing in high dimensions: An nli case study. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2489–2500, 2023. doi: 10.18653/v1/2023.findings-eacl.188. URL https://aclanthology.org/2023.findings-eacl.188/.

Pritam Sarkar and Ali Etemad. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 9723–9732, 2023. doi: 10.1609/aaai.v37i8.26162.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Z1Qlm11uOM.

Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. doi: 10.1109/TBME.2016.2609282.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? In *Advances in Neural Information Processing Systems*, volume 38, pp. 46535–46568, 2025. doi: 10.52202/085713-1388. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b6d18473eb525df8008868f1390cc8c-Abstract-Datasets_and_Benchmarks_Track.html.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01936.

A Audit Implementation Details

This appendix records implementation and provenance details for the current five-seed source–time contract. The 764/191 train/validation populations were disjoint; all endpoint evaluations used 191 validation clips and six nonzero offsets (1,146 predictions per cell); the official test set remained unread. The full 955-clip corpus was used only for split-aware construction, donor maps, and exposure audits.

The core labels used in the manuscript are fixed in Table 2; the same spellings apply to captions, table headers, and the replay bundle.

Current five-seed contract. The primary comparison uses Self-Aligned-P, Self-Deranged-P, D0-Donor-Aligned-P, D0-Donor-Deranged-P, D1-Subject-Preserved-Donor-P, and Time-only. The parameter-sensitivity family reuses the center $(\pi/4, 1.0)$ and evaluates four endpoint-blind corners $(\pi/6, 0.5)$, $(\pi/6, 2.0)$, $(\pi/3, 0.5)$, and $(\pi/3, 2.0)$. Each arm uses seeds $\{1111, 2222, 7777, 8888, 9999\}$, FP32 training, a fixed epoch-10 checkpoint, and the same validation evaluator.

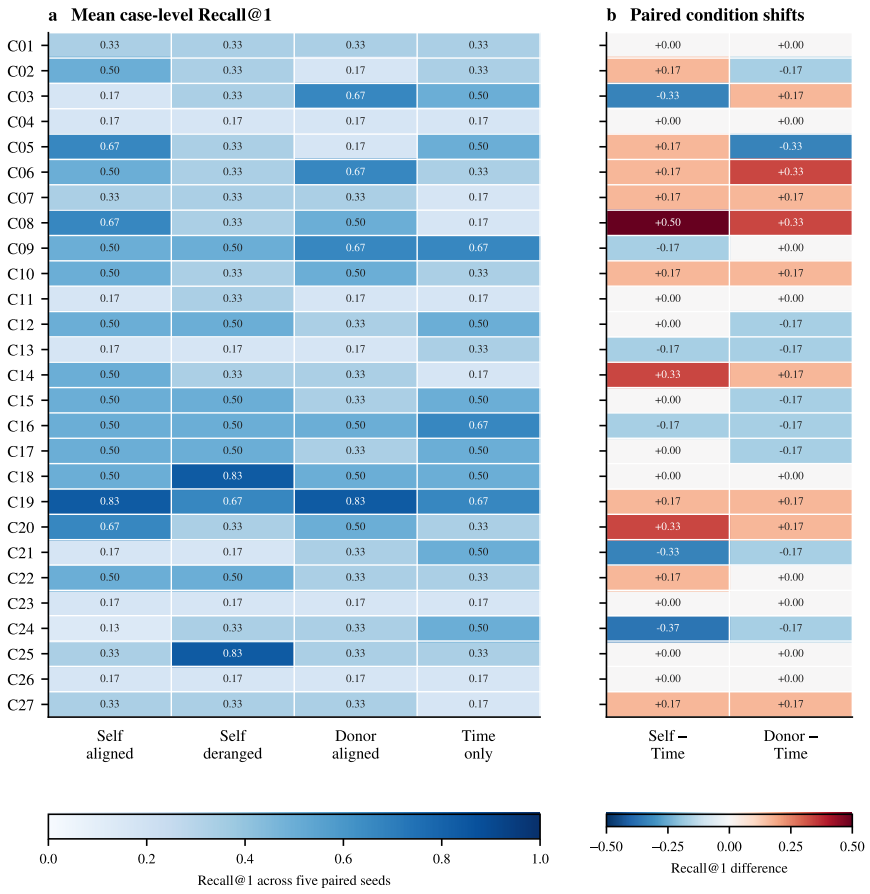


Figure 4: **Endpoint-blind cases mix helpful, tied, and harmful local responses.** The 27 cases were fixed from construction features before reading endpoint outcomes or labels. Panel (a) reports validation Recall@1 for four conditions (x-axis: Recall@1; y-axis: condition); each point is a five-seed mean for one case. Panel (b) reports Self-Aligned-P minus Time-only and D0-Donor-Aligned-P minus Time-only (x-axis: contrast; y-axis: case); sign determines helpful/tied/harmful with tie tolerance 10^{-12}. No interval is attached to a case point. This panel is descriptive, uses no inferential test or multiplicity correction, and does not define a mechanism class.

Table 2: **Core terminology used by the five-seed audit.** Labels are treated as protocol names, not interchangeable aliases.

Label	Meaning in this manuscript
Real-P	Natural phase/envelope pairing package
Time-only	Physical-time eligibility with unit weights
R-unit	Real-P mask with unit weights
AMNESIC-V	Remove phase-predictive visual directions
RAND-V	Remove rank-matched random visual directions
MNESTIC-V	Retain phase-predictive visual directions
MNESTIC-RAND-V	Retain rank-matched random visual directions

Construction and leakage gates. D0 is a split-preserving, recipient-wise bijection with no fixed points; D1 preserves donor subject identity on 138 validation common-support clips and uses a prespecified fallback for the remaining 53 clips. All controls were constructed without endpoint outcomes, labels, or official-test access.

A3. The endpoint population is 191 validation clips and six nonzero offsets per clip. Five fixed seeds ($\{1111, 2222, 7777, 8888, 9999\}$) feed the 10,000-draw seed-first/source-video-group bootstrap. The primary family is Δ_{prov}, Δ_{time}, and Δ_{int}; exact sign-flip tests use the five paired seed effects and have minimum nonzero two-sided resolution 0.0625. Exact Wilcoxon signed-rank sensitivity tests enumerate all sign assignments after midranking tied absolute differences; none of the five contrasts contains an exact zero. The boundaries 0, $+0.025$, and $+0.05$ are task-scale decision references, not equivalence or semantic thresholds. The resulting estimates and intervals are reported in Table 6; every primary interval contains at least two boundaries, so the current evidence is inconclusive at the declared scale. A post-hoc sensitivity to seed-level variability gives reference-model 80% thresholds of 0.0590, 0.0590, and 0.0648 for provenance, time, and interaction at two-sided $\alpha = 0.05$; it is not prospective power, a design-stage MDE, or a replacement for the simultaneous CIs.

Same-carrier Interventional Probing adaptation. This is an adaptation of the protocol studied by Rozanova et al. (Rozanova et al., 2023), not a reproduction of their NLI benchmark. For each of the five fixed checkpoints, position-shared no-intercept ridge probes were fitted on 48,896 training tokens with $\lambda = 1$ and four fixed iterations. Each iteration projected the representation into the fitted probe's null space; the stacked probe rows defined the mnestic subspace, and random controls matched rank and iteration count. The intervention acts on the visual contrast tokens and preserves the zero-padded temporal shifts used by the known-offset endpoint. Each arm was evaluated on 191 validation clips and six nonzero offsets (1,146 queries per arm and seed). The natural arm mean Recall@1 was 0.3412; AMNESIC-V and RAND-V were 0.3421 and 0.3429, while MNESTIC-V and MNESTIC-RAND-V were 0.1517 and 0.2518.

The phase detectability test used complete 64×2 phase trajectories as the permutation unit within each subject and gave $p_{\text{joint}} = 0.005$. We define $d_{\text{rand}} = R_1(\text{AMNESIC-V}) - R_1(\text{RAND-V})$ and $d_{\text{mnest}} = R_1(\text{MNESTIC-V}) - R_1(\text{MNESTIC-RAND-V})$. Their paired hierarchical 95% intervals were $[-0.0129, +0.0102]$ and $[-0.1513, -0.0517]$, respectively. These are method-native zero-referenced contrasts; they are not CMAS accuracy scores, NLI benchmark results, or evidence that one framework dominates the other.

Table 3: **The adapted Interventional Probing protocol supplies a complementary same-carrier diagnostic.** Means aggregate the five fixed checkpoints, each evaluated on 191 validation clips and six nonzero offsets. The comparison uses the DOLOS correspondence endpoint rather than the original NLI endpoint; it is not a performance leaderboard.

Arm	Representation operation	Mean Recall@1 $\pm$ between-seed SD
N	Unmodified visual tokens	0.3412 ± 0.0274
AMNESIC-V	Remove phase-predictive subspace	0.3421 ± 0.0280
RAND-V	Remove rank-matched random subspace	0.3429 ± 0.0252
MNESTIC-V	Retain phase-predictive subspace	0.1517 ± 0.0312
MNESTIC-RAND-V	Retain rank-matched random subspace	0.2518 ± 0.0293

Implementation precondition. The comparison fixes precision, initialization, sampler, epoch endpoint, checkpoint policy, donor maps, and validation evaluator before reading outcomes. This is a comparability condition, not a semantic test. A same-hardware forward-cost record for the frozen epoch-10 carrier checkpoint reports 70,234,627 total parameters (4,215,171 trainable), 364.235 GFLOPs per batch of eight (45.529 GFLOPs per clip), 96.154 ms per batch (12.019 ms per clip), and 83.20 clips/s on an NVIDIA GeForce RTX 3090 with FP32 and 30 warm-up/100 timed iterations. The record excludes P extraction, data loading, and training cost, and is not an external efficiency leaderboard.

Same-carrier Real-L analysis. The Real-L validation filter and three-contrast family were fixed before restricting 15 precomputed arm-by-seed prediction sets to the 191 validation IDs. Each Real-L, L-unit, and paired Time-only arm retained its fixed epoch-10 checkpoint and five seeds; every cell supplied 1,146 known-offset predictions, with zero train or official-test clips. The three contrasts form their own simultaneous family and are not pooled with source–time. Table 4 records this family and the separate synthetic calibration. $\Delta_{L,\text{weight}} = -0.0058$ (95% CI $[-0.0444, +0.0328]$)

and $\Delta_{L,\mathrm{mask}} = -0.0143$ (CI $[-0.0529, +0.0243]$); both remain unresolved, and their telescoping relation is accounting rather than a causal decomposition.

Synthetic functional-dependence calibration. The fixed paired-code/decoy-code comparison uses the same five seeds, 191 validation clips, and six offsets, but BF16 AMP and a deliberately constructed deterministic code. All paired-code kept arms reach Recall@1 $= 1.0$. The A3 contrast and paired visual-code deletion A4 contrast pass their predeclared $+0.05$ and -0.05 gates, respectively. This is a positive-control calibration for detecting a large constructed dependence, not natural semantic evidence, null specificity, near-margin power, or universal identifiability.

Table 4: **Real-L and synthetic controls answer different bounded questions.** Real-L intervals are simultaneous within its three-contrast family; synthetic intervals use the frozen seed-first/clip bootstrap. Neither lane is pooled with the source–time family.

Contrast	Estimate	CI low	CI high
Real-L total	-0.0201	-0.0587	$+0.0185$
Real-L weight	-0.0058	-0.0444	$+0.0328$
Real-L mask	-0.0143	-0.0529	$+0.0243$
Synthetic A3	$+0.6956$	$+0.6627$	$+0.7332$
Synthetic A4	-0.6890	-0.7222	-0.6511

Table 5: **Endpoint-level parameter sensitivity is inconclusive at the declared scale.** Estimates compare each endpoint-blind configuration with Time-only using the five paired seeds and the familywise simultaneous bootstrap. Exact two-sided sign-flip p values are included for completeness; the intervals, not p values alone, determine the boundary verdict. The finite five-point grid is the predeclared sensitivity object; no continuous interpolation is claimed.

Configuration	Estimate	SD	CI low	CI high	exact p
Center $(\pi/4, 1.0)$	$+0.0239$	0.0589	-0.0372	$+0.0850$	0.4375
$(\pi/6, 0.5)$	$+0.0070$	0.0392	-0.0541	$+0.0681$	0.6875
$(\pi/6, 2.0)$	-0.0091	0.0420	-0.0702	$+0.0520$	0.7500
$(\pi/3, 0.5)$	$+0.0079$	0.0535	-0.0533	$+0.0690$	0.6250
$(\pi/3, 2.0)$	$+0.0227$	0.0336	-0.0384	$+0.0838$	0.1875

Secondary evidence boundary. Identity and parameter controls are reported as case-bounded sensitivity analyses. They narrow which interpretations remain compatible with the observed endpoint but do not isolate a unique causal component, establish equivalence, or support cross-carrier generality.

B Anonymous Analysis Replay

With Python 3.10 or newer and the packages listed in `requirements.txt`, the anonymous bundle reconstructs the five-seed source–time point estimates, effect sizes, exact tests, and familywise simultaneous intervals from de-identified seed and source-group records. It separately replays the Interventional Probing arm means, targeted-versus-random hierarchical intervals, and synchronized permutation p values, verifies the released Real-L and synthetic seed summaries, and regenerates the primary and case-response figures. The bundle also releases the de-identified frozen-checkpoint manifest, parameter seed rows, endpoint-confidence stratum, and same-hardware forward-cost record used for the bounded analyses above. The optional `conditional_mde_sensitivity.py` script computes the post-hoc conditional sensitivity to seed-level variability reported above; it does not retrain a model. The released records omit identities, media, learned features, labels, checkpoints, and data-bearing training inputs; the bundle supports statistical replay rather than clean-room model retraining.

C Statistical Design and Sensitivity Analyses

The paired seed effects are reported in Table 7; the parameter-factorial decomposition is reported in Table 8. These tables expose seed-level direction mixing and finite-grid uncertainty without creating a new confirmatory family. For released seed-level contrasts, mean $\pm$ between-seed SD are provenance -0.0104 ± 0.0351, time $+0.0057 \pm 0.0351$, interaction $+0.0379 \pm 0.0385$, identity -0.0587 ± 0.0511, and Real-L total -0.0201 ± 0.0214. The frozen checkpoint manifest supplies five seed rows for each parameter configuration; the resulting sample SDs are reported in Table 5 and the de-identified rows are released as `source_time_parameter_seed_contrasts.csv`. The simultaneous intervals remain the inferential object.

Table 6: **Statistical units and margins bound each audit interpretation.** Point contrasts are paired five-seed means; intervals are familywise simultaneous 95% seed-first/source-group bootstrap CIs, not SD/SEM bars. Crossing zero is not interpreted as equivalence.

Block	Fixed population and seeds	Unit, interval, and margin	Permitted interpretation
A3 provenance	191 validation clips, six offsets, five seeds	Paired seed effects; familywise simultaneous CI $[-0.0585, +0.0377]$, $d_z = -0.296$, exact $p = 0.5625$.	No stable provenance effect at the declared boundaries.
A3 time	191 validation clips, six offsets, five seeds	Paired seed effects; CI $[-0.0424, +0.0537]$, $d_z = +0.162$, exact $p = 0.8125$.	Temporal alignment is inconclusive at the declared scale.
A3 interaction	Same paired matrix	CI $[-0.0102, +0.0859]$, $d_z = +0.983$, exact $p = 0.125$.	Positive point estimate does not identify interaction.
Identity	138 common-support validation clips, five seeds	CI $[-0.1162, -0.0012]$, $d_z = -1.149$, exact $p = 0.125$.	Secondary, case-bounded identity contrast.
Real-L	191 validation clips, six offsets, five seeds	Total CI $[-0.0587, +0.0185]$, $d_z = -0.939$, exact $p = 0.125$.	Unresolved same-carrier A3-only role.
Param.	Center plus four corners, five seeds	All five CIs contain all three anchors.	Shared indeterminacy.

Table 7: **The paired seed effects show sign mixing in the primary family.** Each entry is the within-seed contrast used by the corresponding estimand; seeds are paired across all arms and are not treated as independent clips.

Seed	Δ_{prov}	Δ_{time}	Δ_{int}	Δ_{local}	Δ_{identity}
1111	-0.0358	-0.0070	-0.0227	-0.0742	-0.0713
2222	-0.0550	$+0.0628$	$+0.0838$	$+0.0567$	-0.1208
7777	-0.0070	-0.0070	$+0.0471$	$+0.0428$	-0.0833
8888	$+0.0175$	-0.0305	$+0.0349$	$+0.0175$	$+0.0121$
9999	$+0.0284$	$+0.0100$	$+0.0462$	$+0.0768$	-0.0302

Table 8: **Factorial parameter contrasts remain unresolved.** The main effects and interaction are computed from the center and four endpoint-blind corners with shared paired seeds. Exact two-sided sign-flip p values are reported for completeness; the intervals contain the three declared boundaries.

Contrast	Estimate	SD	CI low	CI high	exact p
Δ_ϕ	$+0.0163$	0.0183	-0.0448	$+0.0774$	0.1875
Δ_σ	-0.0006	0.0325	-0.0617	$+0.0605$	1.0000
$\Delta_{\phi\times\sigma}$	$+0.0309$	0.0503	-0.0302	$+0.0920$	0.2500

Seed-level and case-level sensitivity. The paired seed effects are retained in Table 7, the released CSV, and the primary figure. Exact two-sided Wilcoxon sensitivity p values for provenance, time, interaction, Self–Time, and identity are 0.625, 1.000, 0.125, 0.4375, and 0.125, respectively; they supplement rather than replace the simultaneous intervals. The endpoint-blind panel fixes 27 validation cases before reading outcomes; it reports 11/10/6 positive/tied/negative Self–Time cases and 9/9/9 Donor–Time cases. These summaries are descriptive and do not create a second confirmatory family.

D NEAREST-WORK DECISION MATRIX

Table 9 summarizes how the audit evidence objects differ from the neighboring diagnostic traditions.

Table 9: **CMAS combines evidence objects that prior diagnostic traditions treat separately.** The Interventional Probing row is a same-carrier adaptation of the protocol studied by Rozanova et al.; it uses a different correspondence endpoint from their NLI benchmark and is a diagnostic comparison, not a performance leaderboard.

Tradition	Representative object	Existing verdict	CMAS distinction
Derived targets	Cross-modal target or representation (Alwassel et al., 2020; Shi et al., 2022)	Target fit or downstream transfer	Does not fix a nuisance baseline or test whether the asserted relation earns role-specific support.
Shortcut tests	Prediction under controlled inputs (Yüksekgönül et al., 2023; Yang et al., 2025)	Sensitivity or counterfactual error	Moves the decision upstream to the supervision rule.
Probing	Information in a representation (Hewitt & Liang, 2019; Ravichander et al., 2021)	Controlled recoverability	Uses recovery as bounded, convergent evidence only.
Interventional probing	Matched removal or retention (Ravfogel et al., 2020; Elazar et al., 2021; Rozanova et al., 2023)	Phase detectable; selected directions show no matched-random endpoint advantage.	Same-carrier adaptation with a different endpoint; not an NLI reproduction or CMAS-versus-method ranking.
CMAS record	Role claim under $\mathcal{A}$	Supported / bounded failure / inconclusive / N/A	Joint role, control, endpoint, resolution, and route decision.