

What Makes a Latent Real? A Unified Emergence Gate: Controllability, Cross-View Invariance and Temporal Geometry, and Why It Only Pays Under Intervention

Cercle Digital (independent research)
Correspondence: research@cercle.digital

Abstract

Learning causal variables from entangled perception requires an *admission rule*: given a representation with candidate latent slots, decide which slots correspond to real factors of the world and which are artifacts of the observation map, including how *many* slots to keep. We call this the *emergence-gate* problem, and we study it on a frozen, audited family of synthetic worlds (controllable + autonomous factors observed through diverse nonlinear “glass” maps carrying an adversarial *structured residue*) and on real frozen encoders (MiniLM, DINO-ViT, Wav2Vec2). Four findings. **(1) Every passive gate we test fails, each by a distinct and reproduced mode:** reconstruction and predictability over-admits (a redundant copy of a real factor is exactly as predictable as a new one); post-hoc non-redundancy under-admits (a freely trained slot collapses into the span of existing slots); orthogonally grown information gain over-admits structured residue; ARD pruning and cross-seed recurrence are defeated because the glass *spreads* the residue and is *fixed* across seeds; a learned forward-information-gain value criterion is valid at the oracle but null when learned (add-rate 1.00 on residue). **(2) A unified gate:** a slot is admitted iff it is *controllable under intervention* (b_1 ; $\text{do}()$ denotes a targeted, repeated action probe on the world, as opposed to any filtering of its observations), *or invariant across independent observation maps* (b_2), *or temporally geometric* (b_3 : probability current, i.e. broken detailed balance, or slowness, both read from one velocity operator; evidence for b_3 is synthetic-dynamics only at present), separates real from artifact slots at AUC 1.00 on the reference bench, and its payoff is specifically *out-of-distribution*: downstream readout $0.91 \approx \text{oracle}$ vs. 0.51 ungated (synthetic), 0.885 vs. 0.584 (real perception, embodied-touch gate). **(3) The chicken-and-egg of causal representation** (the admission criterion presupposes the very representation being built) becomes *escapable* once the refit step is anchored on an unfalsifiable interventional target: starting deliberately confounded, decoy-leak falls $0.711 \rightarrow 0.019$ (a control anchored on the observational label stays at 0.719); the selection lever is geometric *coverage* of where to touch, and it is *dimension-gated* (inert on a 1-D target, active on 3-D). **(4) At full scale** on a trained multimodal kernel (40k samples, three frozen encoders, modalities as views), the gate machinery demonstrably works, its data-only selection matches the oracle ranking, yet buys *nothing*: cross-modal contrastive training already self-debiases (gated–ungated CI < 0 on every readout tested), and restoring a *single* controllable action re-ignites the gate ($0 \rightarrow 28$ anchored dimensions of 64; OOD R^2 of the controlled factor $0.00 \rightarrow 0.48$, $\Delta = +0.478 \pm 0.033$). The demarcation is clean: *gate payoff requires an embodied $\text{do}()$; none of the passive objectives we tested substitutes for it*. We release the frozen world family, the pre-registered benches, and all captured logs.

1 Introduction

A learning agent that must model its world from raw perception faces a decision that is logically prior to disentanglement, abstraction, or planning: *of the latent slots my encoder maintains, which ones are real?* A slot is “real” when it tracks a factor of the world, something that persists, can be intervened on, or drives dynamics, rather than an artifact of the map from world to observation. The decision includes the count: an agent that admits too many slots models its own sensor; one that admits too few is blind. We call an admission rule for latent slots an *emergence gate*.

Terminology, up front. Six terms recur from the title onward. A *latent slot* is one unit of an encoder’s representation (a coordinate, or a small learned group of coordinates)

that may or may not track a world factor. $\text{do}()$ denotes Pearl’s intervention operator (Pearl, 2009): acting *on the world* rather than filtering observations of it; in this paper it is always a concrete, targeted, repeated action probe available to the agent (pushing an actuator, editing a text counterfactually, executing a game command on a copied state). *Controllability* (b_1) is a reproducible response of a slot to such a probe. *Cross-view invariance* (b_2) is agreement of a slot’s content across several *independently drawn* observation maps of the same world, measured by canonical correlation analysis (CCA). *Temporal geometry* (b_3) covers two readouts of the representation’s evolution in time: *slowness* (slowly mixing directions, as in slow feature analysis) and *probability current* (broken detailed balance: a systematic circulation in state space that no time-reversible process can produce). Throughout, the *glass* is our name for a frozen

entangling observation map, and its *structured residue* is a deterministic function of the real factors that the glass adds to the observation: the adversary every admission rule must survive, because it inherits every statistical virtue of the factors it is made of. Finally, *out-of-distribution* (OOD) evaluation tests a readout after the spurious train-time correlations are broken; it is where the gate’s payoff will turn out to live.

The negative anchor of the field is well known: without inductive bias or supervision signal, disentangled factors are not identifiable from i.i.d. observations (Locatello et al., 2019); identifiability returns under auxiliary variables, temporal structure, multiple views, or interventions (Hyvärinen & Morioka, 2016, 2017; Khemakhem et al., 2020; Zimmermann et al., 2021; von Kügelgen et al., 2021; Lippe et al., 2022; Ahuja et al., 2023; Brehmer et al., 2022). That literature asks *which generative assumptions make the true factors recoverable*. We ask an operational question one level up, closer to what an agent must actually do: given a concrete encoder over a concrete stream, *which test should admit or reject each slot, and what does a correct admission buy downstream?* This reframing has three consequences that the identifiability framing hides, and they are our results:

1. **Passive admission rules do not merely lack guarantees, they fail actively, each by a distinct mechanism** traceable to the observation map (Section 4). We exhibit the failure-mode matrix: over-admission by redundancy, under-admission by span-collapse, over-admission by structured residue, and the defeat of stability- and sparsity-based rules by the fixedness and spreading of the glass.
2. **Exactly three signals we tested survive, and they are the three the observation map cannot fabricate** (Sections 5–6): a reproducible response to a targeted repeated intervention (b_1); consistency across *independently drawn* observation maps (b_2); and temporal geometry, slowness and, strictly stronger, nonzero probability current, i.e. broken detailed balance read from the antisymmetric part of one velocity operator (b_3). Their disjunction gates at AUC 1.00 on the reference bench, and the payoff of gating is an *out-of-distribution* phenomenon: in-distribution, ungated representations look fine; OOD, only the gated ones survive.
3. **The bootstrap circularity is escapable, and the demarcation of when gating pays is sharp** (Sections 7–8). Anchoring the refit loop on an unfalsifiable interventional target dissolves the chicken-and-egg on which observational value criteria founder; and at full scale on a trained multimodal kernel, the gate pays *zero* on a static corpus, where cross-modal contrastive learning already self-debiases, and is re-ignited by a single controllable action. The useful question is therefore not “which unsupervised objective identifies?” but “*which anchoring*

signal makes the fixed point escapable, and is it available in my data regime?”

Scope claim, stated up front. We claim no new learning mechanism: every ingredient (EM, k-center coverage, counterfactual directions, CCA consensus, slow features, probability currents) is classical. The contribution is (i) the failure-mode matrix of passive gates on a controlled adversary, (ii) the unified three-signal gate with its measured complementarity and its measured traps, (iii) the convergence framing of the bootstrap with the load-bearing/inert decomposition of its levers, and (iv) the pre-registered demarcation experiment at kernel scale. All experiments follow a fixed metrology (Section 3.3); nulls are pre-registered and reported as results, not omissions.

2 Related work

Identifiability and its impossibility. Locatello et al. (2019) established that unsupervised disentanglement is unidentifiable without bias; nonlinear ICA restores identifiability with temporal contrast or auxiliary variables (Hyvärinen & Morioka, 2016, 2017; Khemakhem et al., 2020); Zimmermann et al. (2021) show contrastive objectives can invert the generative process under assumptions; von Kügelgen et al. (2021) treat augmentations as views isolating content from style. Our b_2 signal is the operational cousin of the multi-view route, with two additions the theory does not measure: its *trap* (an invariant meta-confounder provably fools it; Section 5.2) and its *support condition* (dose-response in the diversity of views).

Causal representation learning. The program of Schölkopf et al. (2021) and the interventional identifiability results (Brehmer et al., 2022; Lippe et al., 2022; Ahuja et al., 2023; Eberhardt et al., 2005) justify our b_1 : interventions carry information no observational distribution can fabricate. What we add is the admission-rule form (a slot-wise gate), the measured complementarity with b_2/b_3 (interventions are structurally blind to autonomous factors), and the demonstration that the payoff of intervention-gating is OOD-specific and survives real frozen encoders and a real game. Closest to b_1 as an operational signal, BIS-CUIT identifies causal variables from binary interactions with unknown targets (Lippe et al., 2023); score-based analyses extend interventional identifiability to general transformations (Varici et al., 2025), and identifiability from multiple unpaired distributions generalises the environment route (Zhang et al., 2024). Those results characterise *when* interventional or multi-distribution data identifies; the gate asks the downstream question, *which slot-wise test to run, and what a correct admission buys out of distribution*. Closest to our *unification*, Yao et al. (2025) prove that

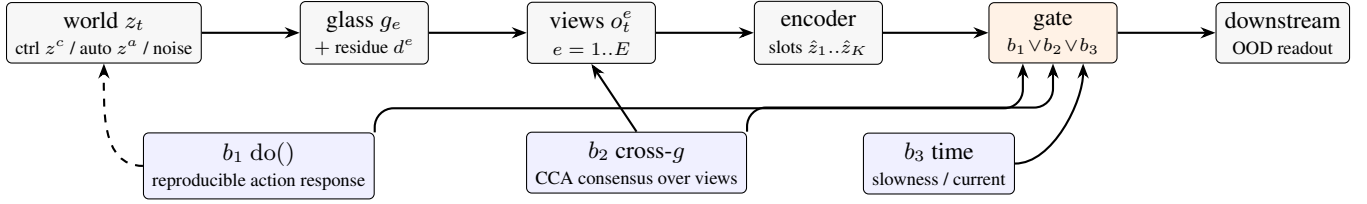


Figure 1: The emergence-gate problem and the unified gate. The world’s factors are seen only through frozen entangling maps g_e that add a structured residue d^e (a deterministic function of the real factors: the adversary every passive rule must survive). Candidate slots $\hat{z}_j$ are admitted iff at least one of three signals fires: a reproducible response to a targeted repeated intervention (b_1), invariance across independently drawn observation maps (b_2), or temporal geometry, slowness or probability current (b_3). b_1 probes the world (dashed); b_2 reads across views; b_3 reads along time. The payoff of gating is measured downstream and out of distribution.

interventional, multi-view and temporal identifiability results are instances of a single invariance principle; the $b_1 \vee b_2 \vee b_3$ disjunction is the operational counterpart of the same triad, delivered as a slot-wise admission rule with measured failure modes and a measured OOD payoff rather than as an identifiability theorem. On the benchmark side, Causal Triplet (Liu et al., 2023) shares two of our desiderata, actionable counterfactuals and an interventional downstream task judged out of distribution, and reports that current CRL methods struggle there; our demarcation offers one mechanism-level reading of that difficulty.

Invariance across environments. IRM and its critiques (Arjovsky et al., 2019; Rosenfeld et al., 2021; Peters et al., 2016; Gulrajani & Lopez-Paz, 2021) anticipate our findings on the multi-environment side: we reproduce the pathology at the representation level (more environments do not close the no-oracle gap; *diversity*, not count, is the lever) and measure the shortcut regime on real text (cf. Sagawa et al., 2020; Geirhos et al., 2020).

Time as a signal. Slow feature analysis (Wiskott & Sejnowski, 2002) and temporal disentanglement (Klindt et al., 2021) motivate b_3 -slowness. The promotion of b_3 to *geometry*: reading the antisymmetric part of the velocity cross-moment as a probability current, i.e. broken detailed balance, imports a physics instrument (Battle et al., 2016; Gnesotto et al., 2018; Weiss, 2003); combinatorial Hodge decompositions (Jiang et al., 2011) supply the shared operator language. To our knowledge the use of a current readout as a latent-admission signal, and its measured advantage over slowness against slow structured residue, are new as measurements.

Slots, capacity, abstraction. Slot Attention (Locatello et al., 2020) and L_0 /ARD sparsification (Louizos et al., 2018) address the mechanics of maintaining and pruning slots; we show *which test* should drive the pruning. Bisimulation-style abstraction (Givan et al., 2003; Zhang et al., 2021) shares our downstream criterion (keep what matters for control) but assumes the reward/dynamics interface we are trying to bootstrap. Active-learning theory

(coverage/core-sets vs. uncertainty; Sener & Savarese, 2018; Seung et al., 1992; Settles, 2009) predicts the selection-lever half of Section 7; curiosity signals (Pathak et al., 2017; Schmidhuber, 1991) are among the passive value criteria we find insufficient for admission.

Mechanistic interpretability. The admission question has a network-internal twin: causal abstraction (Geiger et al., 2025) formalises when internal units implement causal variables; activation patching (Zhang & Nanda, 2024) is a $\text{do}()$ applied to a network’s internals rather than to the world; superposition (Elhage et al., 2022) and sparse-autoencoder feature dictionaries (Huben et al., 2024) face the same over/under-admission dilemma that Study 1 maps for world factors: which candidate features are real, rather than artifacts of the readout. Our gate operates world-side, with world-side interventions; porting b_1 to internal patching is an open bridge we do not measure here.

3 Problem setup, world family, and metrology

3.1 The glass, the residue, and the two halves of the problem

The reference substrate (WORLD-FAMILY, frozen and audited) is a family of worlds $\mathcal{W} = \{(z, g_1..g_E, W_d^1..W_d^E)\}$ designed to make the admission problem adversarial while keeping ground truth available for evaluation only:

- **Factors.** $z_t = (z_t^c, z_t^a, \varepsilon_t)$ with $K_{\text{ctrl}}=2$ *controllable* factors (respond to the agent’s action a_t), $K_{\text{auto}}=2$ *autonomous* factors (evolve on their own; AR(1) with $\rho=0.9$, or a 2-D rotation $z' = \rho R(\theta)z + \eta$ for the current-bearing worlds of Section 5.3), and fast i.i.d. noise.
- **Glass.** Each view e observes $o_t^e = g_e(z_t)$ through a frozen random tanh-MLP g_e (the “glass”): dimensions are mixed nonlinearly and non-invertibly from the reader’s standpoint.

- **Structured residue (the adversary).** Each view carries $R=2$ residue channels $d_t^e = \tanh(W_d^e z_t^{\text{real}})$: *deterministic functions of the real factors*. The residue is therefore predictable, persistent, exactly as slow as its sources, reconstruction-relevant, and stable across seeds, it passes every “by-the-eye” statistical test a passive rule can apply. It is causally inert (interventions on it change nothing) and view-specific (unless deliberately shared, in the trap condition of Section 5.2).

Conditions in the family vary E (1..8), the diversity of the g_e (identical, narrow, diverse), and whether the residue map is shared across views (the adversarial trap). Everything is frozen: the family ships with fixed seeds and audit scripts (real factors linearly recoverable from each view, $R^2 \approx 0.88$ – 0.92 ; residue separable when g varies, 0.38 at $E=8$, and *inseparable* when shared, 1.00).

The problem splits into two halves that will recur throughout: the **controllable half** (slots tracking z^c , reachable by intervention) and the **autonomous half** (slots tracking z^a , unreachable by intervention, where “true-but-immobile” must be told from “reflection-of-the-glass”).

Real-perception substrates. Claims are ported off the synthetic family wherever a port exists: frozen MiniLM (Wang et al., 2020; Reimers & Gurevych, 2019) over Gutenberg text and templated interactive fiction with embedded `do()` (Sections 5.1, 6); four genuinely different frozen text encoders as diverse g (Section 5.2); a real game (Colossal Cave Adventure, 141 rooms, non-destructive `do()` by state copy); and a trained multimodal kernel over DINO-ViT (Caron et al., 2021), MiniLM, and Wav2Vec2 (Baevski et al., 2020) features of a 40k sample corpus (Section 8).

3.2 The gate decision problem, formally

Fix an encoder producing slots $\hat{z}_{1..K}$ (or feature dimensions) from the stream. An *emergence gate* is a predicate $\text{REAL}(j) \in \{0, 1\}$ computable from data available to the agent (no ground truth). We evaluate gates by (i) *admission quality*: agreement with the true real/artifact partition (AUC; MCC of recovered vs. true factors; *decoy-leak*, the fraction of the kept subspace aligned with the residue); and (ii) *downstream payoff*: performance of a fixed readout/policy trained on the kept slots, evaluated *out-of-distribution*, i.e. after the spurious link between residue and task is broken.

The three signals of the unified gate are, per slot j :

$$b_1(j) = \max_k |\text{corr}(\Delta \hat{z}_j, a_k)|, \quad (1)$$

$$b_2(j) = \rho_j^{\text{CCA}}, \quad (2)$$

$$b_3(j) = \max \left(\underbrace{1 - \lambda_j^{\text{SFA}}}_{\text{slowness}}, \underbrace{\|\Pi_A u_j\|}_{\text{current}} \right), \quad (3)$$

where b_1 is read under targeted, repeated `do()` probes a_k ; ρ_j^{CCA} is the consensus canonical correlation of slot j ’s

aligned direction across the E views (held-out); and, for b_3 , Y is the whitened representation, $M = \langle Y_t (Y_{t+1} - Y_t)^\top \rangle$ its velocity cross-moment, $A = \frac{1}{2}(M - M^\top)$ the antisymmetric part of M (the linear signature of probability current / broken detailed balance; Weiss, 2003; Gnesotto et al., 2018), and Π_A projects onto its dominant singular directions. The gate is the disjunction

$$\text{REAL}(j) = [b_1(j) > \tau_1] \vee [b_2(j) > \tau_2] \vee [b_3(j) > \tau_3], \quad (4)$$

with thresholds fixed a priori or self-calibrated against each signal’s own null (time-shuffle for b_3 , probe-shuffle for b_1 , view-agreement reliability for b_2 ; the “reliability-aware” variant). All three are *readouts*: none enters any training loss (a demarcation that is itself one of our measured results; Section 5.3). The evidentiary status of the three signals is not uniform, and we flag it here rather than in the limitations alone: b_1 and b_2 are established through real-encoder and real-environment ports (Sections 5.1–5.2); b_3 rests on synthetic dynamics only, enters the disjunction with its mandatory shuffle and reversible-twin controls, and should be read as preliminary evidence that closes a measured trap (Section 5.3), not as a pillar established at parity with b_1/b_2 .

3.3 Metrology

Every bench follows one discipline, fixed before this paper’s experiments and kept throughout: fixed seeds with explicit RNG; ≥ 10 – 12 seeds (or book-folds on text) with 95% confidence intervals; never point estimates; pre-registered hypotheses with thresholds fixed before the run; strong baselines (oracle ceilings and confounded floors, not strawmen); the world is never weakened to manufacture a pass; and *accepted nulls*: a hypothesis that fails is reported as a result with its diagnosed mechanism. Two methodological findings produced by this discipline are themselves used below: on noisy non-monotone trajectories, use a *stable-clean* criterion rather than first-crossing (a first-crossing metric manufactured a spurious win once; caught and retracted); and match resources before claiming “beats the ceiling” (one unfair-resource run was caught and corrected).

4 Study 1: Every passive gate we test fails, each by a distinct mode

The parent problem: *can a purely observational rule decide how many slots are real?* We test one candidate rule per cycle, each operationalised on the same two pre-registered conditions, and let the mechanism of each falsification generate the next rule. The two conditions: **(A) missing-real**, the encoder has $K-1$ slots for K real factors, so a controllable/world factor is genuinely unrepresented; a good gate *should* admit a new slot; **(B) complete**, every real factor is already captured; any new slot can only be a redundant copy,

noise, or residue; a good gate should *not* admit. Controls always include the naive reconstruction rule (admit iff reconstruction error drops), and each successor bench re-runs its predecessors' gates as calibration controls that must reproduce their known failures.

4.1 Cycle 1: predictability (GATE-PREDICTABILITY)

The first candidate rule replaces reconstruction gain with realness in latent space: a slot is admitted iff it is *modelable*, that is, predictable by the latent forward model, persistent, or controllable. We grow one slot, train it briefly, and keep it iff its latent realness clears a fixed threshold, comparing against the naive reconstruction rule on conditions (A) and (B) plus a missing-noise control.

The gate admits correctly in (A) and rejects i.i.d. noise, but *over-admits* in (B): a freshly grown slot that redundantly re-encodes an already-captured factor is exactly as predictable, persistent and controllable as a genuinely new one. Predictability therefore decides *realness of content*, not *novelty*; the natural repair, an information-gain constraint against the existing slots, is the next cycle's hypothesis.

4.2 Cycle 2: post-hoc non-redundancy (GATE-NONREDUNDANCY)

Cycle 2 adds that repair: the new slot must carry information the existing latents do not (measured as $1 - R^2$ on their span), conjoined with the realness test. The protocol is otherwise unchanged, and the GATE-PREDICTABILITY gate runs as a control, duly reproducing its over-admission in (B). The combined gate fails in the opposite direction: it *never* admits, including in condition (A) where a real factor is genuinely missing. The mechanism is span-collapse. Gradient descent finds the shared variance first, so a freely trained slot collapses into the linear span of the existing slots, and its measured non-redundancy is ≈ 0 in *both* conditions. Judging the slot post hoc is too late; it must be *grown toward* novelty.

4.3 Cycle 3: orthogonally grown information gain (GATE-INFOGAIN)

Cycle 3 therefore grows the slot under a decorrelation penalty and admits iff information gain and realness both clear their thresholds; both predecessor gates run as controls and reproduce their modes. A *third* failure mode appears. In condition (B) the orthogonally grown slot locks onto the *structured residue*, which is novel with respect to the real factors, informative for reconstruction, and persistent; it passes both tests, and the gate over-admits again (add-rate 1.00 in (B); candidate-to-residue correlation ≈ 0.46 – 0.48). This closes the observational bracket on a diagnosis rather than

another repair: *every observational criterion tested is satisfiable by the glass's own residue*. Being a deterministic function of the real factors, the residue inherits their predictability, their novelty, their information gain, and their persistence.

4.4 Cycle 4: interventional touch, joint pruning, and two predicted failures (GATE-TOUCH)

Cycle 4 changes the logical type of the signal. An *active* gate admits a slot iff it responds reproducibly to a targeted, repeated action perturbation, on the reasoning that structured noise cannot fake a reproducible response to a *chosen* probe. ARD-style joint shrinkage (group- L_1) from an over-complete start is added to handle redundancy, and a cross-seed recurrence rule is given a fair run despite being predicted to fail, since the residue is fixed with the glass and should recur across seeds exactly like a real factor.

The touch gate keeps the genuine controllable factors and rejects a null action channel. ARD handles the redundant-copy half but is defeated on the residue, which the glass *spreads* across slots so that no single scale shrinks it; the cross-seed control fails exactly as predicted. What stands is the **autonomous residual**: for factors that do not respond to action, no available passive signal (persistence, ARD scale, seed recurrence) separates real-autonomous from structured residue. Intervention crosses the controllable half of the admission problem and leaves the autonomous half untouched.

4.5 Cycle 5: value-side criteria (GATE-CURIOSITY, GATE-VALUE)

The last passive candidates are *value* signals: curiosity, read as forward prediction progress (GATE-CURIOSITY), and conditional forward information gain (GATE-VALUE), the latter evaluated twice, at an *oracle* panel built on the true base and *learned* from the agent's own imperfect base. Curiosity cleanly ignores i.i.d. noise (real 0.26 vs. noise 0.03), but its directional-selectivity hypothesis is null. The information-gain criterion is *valid at the oracle*: conditional forward info-gain is -0.0003 on the residue against $+0.0278$ in the missing-real condition, so the criterion itself separates. *Learned*, it is null: the agent's imperfect base leaks, and the gate admits the residue with add-rate 1.00. The value criterion is therefore not wrong but *circular*: computing it correctly requires the very representation being built. This is the bootstrap wall, taken up in Section 7.

Conclusion of Study 1. Table 1 is the negative half of the paper's thesis. The count and reality of latent variables is *not a statistical property of the observation*: the glass's residue inherits every observational virtue of the real factors. What survives cycle 4 is a signal of a different logical type, an

Table 1: The failure-mode matrix of passive admission rules (Study 1). Conditions: (A) missing-real (should admit), (B) complete (should not). Each successor bench re-runs its predecessors as controls; all modes reproduce. Add-rates are from the original pre-registered benches (fixed thresholds, world never weakened); the released benches regenerate the tables deterministically.

Rule (bench)	Signal read	Failure mode	Mechanism
Naive recon (Δerr)	reconstruction	over-admits (B)	noise/residue also lowers recon
Predictability (GATE-PREDICTABILITY)	latent realness	over-admits (B)	redundant copy equally modelable
Non-redundancy (GATE-NONREDUNDANCY)	$1 - R^2$ post-hoc	under-admits (A)	free slot collapses into span
Orthogonal growth (GATE-INFOGAIN)	info-gain $\wedge$ real	over-admits (B)	structured residue passes both
ARD joint prune (GATE-TOUCH)	per-slot scale	misses residue	g spreads residue across slots
Cross-seed recurrence (GATE-TOUCH)	stability	no separation	g is fixed: artifact recurs
Curiosity (GATE-CURIOSITY)	prediction progress	selectivity null	rejects i.i.d. noise only
Fwd. info-gain, learned (GATE-VALUE)	value	add-rate 1.00 (B)	criterion circular: base leaks
Touch do() (GATE-TOUCH)	intervention	settles ctrl half	residue can't fake reproducible response

intervention the residue cannot answer. What remains open is the autonomous half, and the circularity of value. The next two studies address each.

5 Study 2: Three signals the glass could not fabricate

The tense in the title is deliberate and bounds the claim: under our adversarial construction, the glass's structured residue did not produce these three signals except in the measured traps reported below. This is a reproduced property of the construction, not a general impossibility theorem.

5.1 b_1 : controllability under intervention: from box to text to a real game

The claim to establish for b_1 is portability: a reproducible response to a targeted, repeated do() identifies controllable factors through any glass, and the identification survives real encoders and real environments where observational readers collapse out of distribution. We test it on four ports of increasing realism, all with the same design, an observational reader and an interventional reader at matched budget, evaluated after the train-time spurious link is broken. The ports are (i) the synthetic family (GATE-TOUCH, above); (ii) real text, counterfactual affect edits on Gutenberg through frozen MiniLM (TOUCH-TEXT); (iii) templated interactive fiction with an embedded do() (use_key opens a door; TOUCH-FICTION); and (iv) a real game, Colossal Cave Adventure, 141 rooms of prose that genuinely under-determines exits, with do() by non-destructive state copy (TOUCH-GAME).

The identification ports at every step. On the synthetic family, touch sees controllable factors at 0.94 and is blind to autonomous ones at 0.01–0.02, the complementarity that motivates b_2/b_3 . On real text, the counterfactual edit direction cancels the dominant topic nuisance: +0.07 CI-separated over the observational reader, and markedly more label-efficient (2 edits do the work of 32 passive labels on the

point estimate), with a declared bound at 0.63 for lexicon-based edits. In interactive fiction, the touch axis holds OOD at ≈ 0.96 where the observational reader collapses, and 2 do() again beat 32 labels. On the real game, the observational exit-affordance reader drops $0.86 \rightarrow 0.60$ OOD across held-out rooms (a prose-driven navigator succeeds only 0.76 of the time) while the touch-built map holds at 1.0. b_1 is thus load-bearing along the whole chain, from the analytic setting through text and embodiment to a real game. Its structural limit is equally clean: it cannot see what it cannot poke.

5.2 b_2 : invariance across independent observation maps: with its trap measured

Intervention is blind to the autonomous half, so the second signal reads *across* observation maps: a mechanism seen through N *independently drawn* glasses should be separable from any glass-specific residue, and cross-view consensus (deterministic multi-view CCA) should recover the autonomous factors b_1 cannot reach. We measure this in three settings: spectral consensus on pure-autonomous worlds (INVARIANCE-SPECTRAL); five pre-registered conditions on the frozen audited family (INVARIANCE-GATE: 5 materialization seeds, 95% CIs, seedless gate, against a passive single-view baseline and the touch score); and a real port (INVARIANCE-ENCODERS) with four genuinely different frozen text encoders (MiniLM, word-LSA, char-LSA, bigram-LSA) as four real g 's over the same Gutenberg passages.

The spectral crossing is clean (band margin 1.62 vs. 0.97), and all five frozen-family conditions pass (Table 2). The gate recovers autonomous reals CI-above the passive baseline (0.71 vs. 0.53). Identical glasses yield margin 0.00, so the number of views alone is not the lever; separation is dose-dependent in glass *diversity*, rising from 0.09 to 0.31 to 0.61 at $E=2, 4, 8$ (Figure 2). The complementarity with touch is confirmed (touch: ctrl 0.94, auto 0.01; gate: auto 0.71). The fifth condition measures the trap: a residue *shared* across views leaks *into* the kept invariant subspace

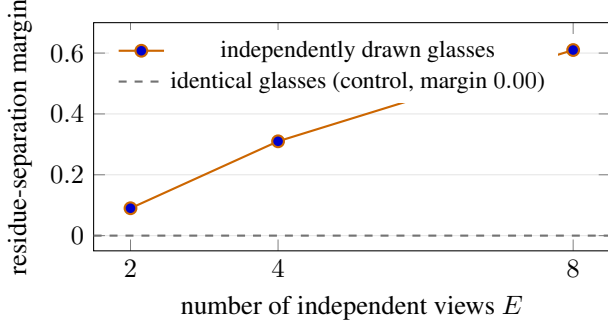


Figure 2: b_2 separation needs glass diversity, not view count (INVARIANCE-GATE, conditions C1/C4). The residue-separation margin rises with the number of independently drawn glasses, while identical glasses stay at 0.00 (dashed reference): $E > 1$ alone is not the lever.

CI-above the per-view case (0.86 vs. 0.66), so **the gate is provably fooled by an invariant meta-confounder**. On real encoders the consensus transfers: cross-encoder 0.625 vs. best single 0.585 at 20 labels.

b_2 therefore recovers the autonomous half b_1 cannot reach, under two measured conditions: sufficient diversity of glasses (its support condition) and no residue shared across them (its trap). Neither is a theoretical footnote; both bind at scale (Sections 7 and 8).

5.3 b_3 : temporal geometry: slowness, then current; and why time must stay a readout

The third signal reads *along* time. We build it in two steps, slowness then current, and close with a demarcation null that fixes how time may be used at all.

Slowness (TIME-SLOWNESS). The hypothesis is complementarity: slowness (SFA) should recover autonomous factors *without* any glass diversity, exactly where b_2 lacks support. A seedless linear SFA readout on the frozen family (10 materialization seeds, 95% CIs), with time-shuffle as the authenticity control, confirms it (Table 3): slowness recovers autonomous reals at 0.71–0.89 across all conditions and is CI-above passive precisely where b_2 is weak (`narrow_g` 0.84 vs. 0.74; $E=2$: 0.80 vs. 0.69). Time-shuffle collapses the readout to 0.37–0.41, so the effect is temporal order, not variance, and the union $b_2 \vee b_3$ never CI-drops below its best arm. Slowness carries its own trap, symmetric to b_2 ’s (bound T4): it cannot prune the *slow residue*, since $d = \tanh(W_d z)$ is an instantaneous function of slow reals and hence slow itself (residue-leak 0.74–0.82, at the passive level). Two reservations are declared. The substrate’s autonomous factors are AR(1) with $\rho=0.9$, the slowest things in the world by construction and SFA’s favourable terrain, so the claim is

conditional on autonomous factors being slow; and part of the union’s gain is dimensional (8 vs. 4 dims).

Current (TIME-CURRENT). Promoting time from a scalar to a *geometry* closes T4. The residue is an instantaneous image of the real factors and carries no probability current of its own, so a current readout should recover the autonomous factor while leaking far less residue. We read the whitened velocity cross-moment M , take its antisymmetric part $A = (M - M^\top)/2$, and keep the dominant singular directions of A as the current-bearing subspace (Eq. 3), with two mandatory controls: an exact *reversible twin* world (same frozen glasses and residue, $\theta=0$, detailed balance holds) and time-shuffle; 10 seeds, zero training. On the oscillator world the current recovers the autonomous factor at 0.86 while leaking only 0.17 of residue, against 0.63 recovery and 0.69 leak for slowness: the separation margin flips $-0.06 \rightarrow +0.69$ (Table 4). Circulation magnitude detects irreversibility as a *ratio*: $\|A\|_F = 0.50$ vs. 0.10 on the reversible twin (4.8 $\times$), shuffle 0.11. A difference-in-differences anatomy attributes +0.51 of the +0.75 margin advantage (68%) to true current (oscillator minus reversible) and +0.24 (32%) to structure: an instantaneous residue couples *symmetrically*, hence is constitutively invisible to the antisymmetric part. The same antisymmetric operator family also reads logical contradiction as a sign holonomy on signed belief graphs, a structural kinship we note without relying on it. The reservations: one injected 2-D rotation; a linear instrument on a nonlinear glass leaves a nonlinearity floor (reversible 0.10 > shuffle 0.09, so detection must remain a ratio, with the reversible and shuffle controls mandatory); and no real-perception port yet.

Readout, not objective (TIME-OBJECTIVE-NUL). If time is a good readout, is it a good *training objective*? The answer is null in all three forms tried: temporal contrastive positives *hurt* retrieval ($r@1$ 0.64 $\rightarrow$ 0.48); an SFA penalty is redundant; and a trained single-stream time-contrastive encoder is *worse than PCA* (0.50 vs. 0.68) on the same data where the SFA *readout* beats PCA (0.71). Time is a signal for *selection*, not a loss for *formation*. This readout-not-objective demarcation recurs independently at kernel scale (Section 8), and we apply it throughout: none of $b_1/b_2/b_3$ enters a training loss in the unified gate.

6 Study 3: The unified gate, and why its payoff is out-of-distribution

The three signals are complementary by construction: b_1 sees the controllable half, b_2 and b_3 the autonomous half, and each carries a distinct measured trap. Their disjunction (Eq. 4) should therefore decide admission where every single signal fails, and a correct admission should buy some-

Table 2: b_2 on the frozen world family (INVARIANCE-GATE; 5 seeds, 95% CIs; gate deterministic/seedless). MCC on *autonomous* reals unless noted.

Pre-registered condition	Statistic	Verdict
C1 controls null	identical- g band margin = 0.00; $E=1$ degenerate	anti-rigging holds
C2 recovers autonomous	gate 0.71 vs. passive 0.53 (CI-separated)	recovers what touch can't
C3 the trap bites	shared-residue leak into kept subspace 0.86 vs. 0.66	fooled by invariant confounder
C4 dose-response	margin 0.09 $\rightarrow$ 0.31 $\rightarrow$ 0.61 for $E=2, 4, 8$	separation needs g -diversity
C5 complementarity	touch: ctrl 0.94, auto 0.01; gate: auto 0.71	autonomous half needs a 2nd signal

Table 3: b_3 -slowness on the frozen family (TIME-SLOWNESS; 10 seeds; MCC on autonomous reals; shuffle = time-order permuted within trajectories). Touch is blind to these factors everywhere (ctrl 0.94 / auto 0.02).

World	role	b_3 SFA	b_3 shuffle	b_2 CCA	passive	$b_2 \vee b_3$
single_g	control $E=1$	0.87	0.41	0.60	0.60	0.87
identical_g	control identical g	0.86	0.37	0.36	0.65	0.86
narrow_g	low diversity	0.84	0.40	0.81	0.74	0.87
dose_E2	low diversity $E=2$	0.80	0.41	0.63	0.69	0.80
diverse_g	multi-env	0.71	0.37	0.73	0.68	0.75
dose_E8	high diversity $E=8$	0.71	0.37	0.73	0.68	0.75
shared_residue	adversarial trap	0.89	0.40	0.82	0.73	0.89

Table 4: b_3 -current closes the slow-residue trap (TIME-CURRENT; 10 seeds; zero training). Left: recovery of autonomous factor vs. residue leak. Right: circulation magnitude $\|A\|_F$ with time-shuffle control. The reversible twin shares glasses and residue; only the autonomous dynamics differ ($\theta=0$).

World	current auto	current resid	SFA auto	SFA resid	passive auto
oscillator	0.86	0.17	0.63	0.69	0.70
reversible twin	0.48	0.35	0.71	0.82	0.68
	circulation $\ A\ _F$		shuffle		
oscillator	0.50		0.11		
reversible twin	0.10		0.09		

thing measurable downstream. We test admission and payoff separately.

Admission. UNIFIED-GATE runs the disjunction $b_1 \vee b_2$ on the synthetic bench against each signal alone; a reliability-aware variant self-calibrates thresholds from each signal's own null (UNIFIED-GATE-SELFCAL); b_3 enters the final rule after Study 2 (Eq. 4), and on the frozen family the union is no-regret (Table 3). The disjunction separates real from artifact slots at AUC 1.00 on the bench where each single signal has a blind half, and the self-calibrated variant keeps this separation without hand-set thresholds.

Payoff. We train a fixed downstream policy or readout on gated vs. ungated representations and evaluate it in-distribution and OOD, after the train-time spurious link is broken, on three substrates: a synthetic world with planted nuisance (PAYOFF-SYNTHETIC); *real perception*, frozen MiniLM over an interactive-fiction world whose task factor

(a door) is spuriously tied to a dominant place-and-mood nuisance at train time, the gate built from $K=8$ embodied do() probes (PAYOFF-TEXT; 8 seeds, 95% CIs); and a sand-boxed web-world port (PAYOFF-WEB). The payoff is OOD, and only OOD. In distribution the ungated representation is competitive everywhere: the mirage. Out of distribution the gated arm wins on all three substrates (Figure 3): synthetic, gated 0.91 $\approx$ oracle vs. ungated 0.51; real perception, gated 0.885 vs. ungated 0.584 at every label budget (Table 5), the ungated policy riding the train-time mood correlation (0.861 ID $\rightarrow$ 0.584 OOD, a drop of 0.28) while the gated one holds (a drop of 0.05); web world, gated 0.83 vs. ungated 0.00. The gate is also data-efficient: at 5 policy labels, gated 0.917 vs. ungated 0.603, the 1-D causal axis acting as a strong prior.

Gating to *admitted* slots thus converts an in-distribution-equivalent representation into an OOD-robust one; the payoff concentrates exactly where evaluation usually does not look. In PAYOFF-TEXT the gate consumes 8 embodied interventions the ungated arm does not have; rather than only declaring that asymmetry, we measure it. An *equal-information control* (PAYOFF-EQUALINFO) hands EVERY arm the same $K=8$ probe pairs and varies only the *reader* of the interventional information: the admission rule (average the counterfactual deltas into one axis) reaches 0.859 OOD at 5 policy labels, where the same deltas appended as raw features to the ungated representation reach only 0.680 (converging to 0.875 only at 80 labels); a supervised direction fit on the probe pairs themselves matches the rule (0.876); and a random 1-D axis stays at chance (0.53), ruling out the bottleneck itself as the lever. The payoff therefore requires *both* the interventional information and a compress-

Table 5: OOD payoff on real perception (PAYOFF-TEXT; frozen MiniLM; gate = 8 embodied do() probes; 8 seeds, 95% CIs). Policy win-rate vs. number of policy labels.

labels	gated ID	gated OOD	ungated ID	ungated OOD	oracle OOD
5	0.917	0.859	0.603	0.533	1.000
20	0.935	0.882	0.741	0.575	1.000
80	0.935	0.885	0.861	0.584	1.000

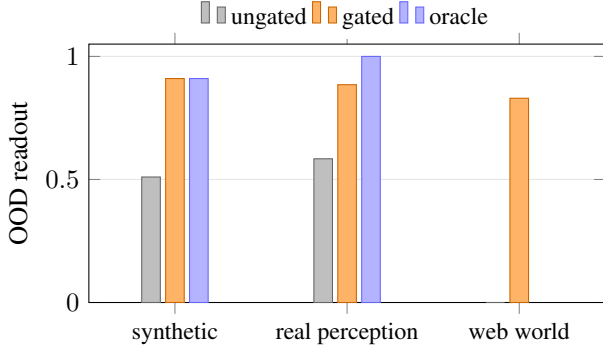


Figure 3: The gate’s payoff is out-of-distribution. Downstream readout after the train-time spurious link is broken, on PAYOFF-SYNTHETIC (synthetic), PAYOFF-TEXT (real perception) and PAYOFF-WEB (web world). In-distribution the ungated arms are competitive (Table 5), the mirage the OOD evaluation exposes. Oracle bars: PAYOFF-SYNTHETIC oracle $\approx$ gated; the PAYOFF-WEB oracle was not separately instrumented, so no oracle bar is shown for the web world.

ing reader of it; the b_1 rule is one efficient such reader, not the only one. Remaining reservations: the causal axis is 1-D and the policy linear; the world is templated (perception real via MiniLM, prose bounded); and decisions are single-step.

7 Study 4: The bootstrap becomes escapable: anchor, coverage, and the confounding objective

Study 1 ended on a circularity: the valid value criterion (GATE-VALUE) needs the very representation being built. This study makes the circularity a *convergence question* and answers it. Substrate: a controlled box-world with $R=3$ relevant directions in $D=12$ plus a spurious *decoy* direction statistically confined among the positives (relevant to any observational objective, causally inert). Metrics: *decoy-leak* (fraction of the recovered subspace on the decoy), overlap with the true subspace, OOD accuracy when the decoy is decorrelated. All benches ≥ 12 seeds, 95% CIs.

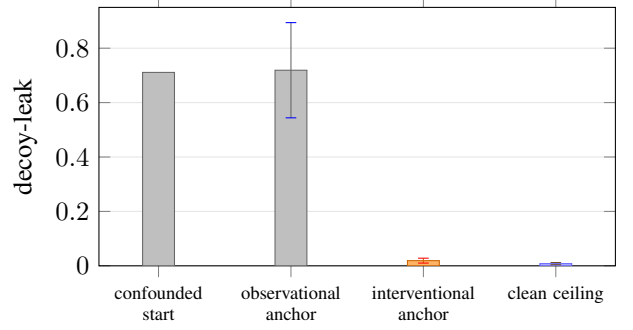


Figure 4: The bootstrap escape is anchor-driven (BOOTSTRAP-ANCHOR; ≥ 12 seeds, 95% CIs). From a deliberately confounded base (decoy-leak 0.711), the loop whose refit is anchored on the interventional target converges to the clean ceiling (0.019; ceiling 0.008), while the identical loop anchored on the observational label does not move (0.719).

7.1 Cycle 1: the escape and its control (BOOTSTRAP-ANCHOR)

The first question is whether the loop escapes at all, and which part of it does the work. Starting *deliberately confounded*, with the base set to the lossy learned base whose reading of intervention reconfounders (leak 0.711), we run the loop select-where-to-touch $\rightarrow$ observe the real touch $\rightarrow$ refit, anchoring the refit on the unfalsifiable interventional target; a control runs the identical loop anchored on the observational label.

The anchored loop escapes to the clean ceiling without oracle help: leak 0.711 $\rightarrow$ **0.019 $\pm$ 0.009** (clean ceiling 0.008 $\pm$ 0.004; overlap 0.985; OOD 0.757 $\rightarrow$ 0.836). The control stays stuck at 0.719 $\pm$ 0.175 (Figure 4): the anchor, not the loop mechanics, is the lever. The escape is moreover *corrective dynamics*, not accumulation: at matched total budget the iterative loop CI-beats a single batch 3/3 (0.22 vs. 0.33 @60; 0.19 vs. 0.39 @120; 0.14 vs. 0.23 @240). The chicken-and-egg was thus a property of *observational* value loops; an interventional anchor, a signal the decoy cannot fabricate, turns the confounded fixed point into an escapable basin. A corollary follows: “freeze the foundation” is *sufficient, not necessary*.

7.2 Cycle 2: the selection lever is coverage, and it is dimension-gated (BOOTSTRAP-COVERAGE/BOOTSTRAP-COVERAGE-DIRECTED/BOOTSTRAP-TEXT/BOOTSTRAP-EMBODIED)

With the escape established, the next question is which policy for *where* to touch reaches the clean subspace data-efficiently: curiosity (boundary), committee disagreement, or geometric coverage. Coverage wins, and boundary never cleans. k-center coverage cleans first (budget 192; random

at 384); the *oracle* true-boundary selector never cleans (0.43 leak at 384, worsening), because boundary points concentrate on a measure-zero surface that does not span the subspace; QBC tracks curiosity, so externality per se is not the lever. One apparent refinement was refuted by our own metrology: directing coverage by the base’s believed subspace looked like a win and failed under the stable-clean criterion with fresh seeds (the slack is inert). The honest consolidation: the anchor plus *any* coverage escapes at $\sim 16\text{--}20\%$ of N .

The lever is moreover *dimension-gated*, established by a cross-substrate pair. On real text (frozen MiniLM, counterfactual-edit touch; BOOTSTRAP-TEXT) the anchor ports (touch OOD AUC 0.608 ± 0.023 vs. confounded 0.531 ± 0.023 ; leak 0.028 vs. 0.980) but *selection is inert*: all policies tie and the ceiling is hit at $\sim 5\%$ of budget, because the affect target is $\sim 1\text{-D}$. On an *embodied* touch with a 3-D target (reproducible rollout win-rates; BOOTSTRAP-EMBODIED) the anchor survives (leak 0.070 ± 0.027 ; OOD 0.804 ± 0.018 vs. 0.758 ± 0.015) and coverage re-bites, CI-beating curiosity at 3/4 budgets (e.g. 0.16 vs. 0.44 @60), with curiosity *worsening* as budget grows ($0.34 \rightarrow 0.43$). Coverage of where to touch is therefore real but proportional to the dimensionality of the target; curiosity governs what to encode, not where to touch for *discovery*, consistent with core-set theory (Sener & Savarese, 2018) and reproducing the boundary pathology in an embodied setting.

7.3 Cycle 3: through perception: the attenuation decomposes exactly (PERCEPTION-COST/PERCEPTION-FREEZE/PERCEPTION-CAPACITY/PERCEPTION-FORM)

The third cycle passes the state itself through an entangling glass $o = \tanh(1.1XW)$ carrying a planted residue of swept strength α , and asks whether the escape survives. It does, in an attenuated form, and the attenuation decomposes exactly across four benches. First, touch(o) stays CI-above the confounded reader at every α (gap $+0.072 \rightarrow +0.039$) but is no longer residue-invariant; perceiving costs ~ 0.03 against the given-state ceiling of 0.832 ± 0.019 . Second, the α -dependent part is the bootstrap’s own confound reappearing in *perception*, and it is recovered by *freezing* the touch-foundation once fit on residue-free perception: frozen stays flat ($0.794 \rightarrow 0.784$) while re-learned collapses ($0.794 \rightarrow 0.651$; $+0.133$ at $\alpha=0.8$). Third, the remaining *entangling* cost is a matched-resource floor ($+0.028 \rightarrow +0.050$ across richness levels): more frozen capacity raises both arms in parallel and never catches up. Fourth, the floor is partly pierced by *forming* the encoder on the interventional per-axis sensitivity $s_j = \partial \text{win} / \partial x_j$: the gap narrows $+0.043 \rightarrow +0.026$ ($\sim 40\%$ recovered), CI-separated at 30 seeds and directional-only at 12, which we disclose; the re-

mainder is a genuine feedforward bound. The bootstrap wall, the perception wall, the freeze-the-foundation pivot, and the encoder branch are therefore *one problem*: the interventional anchor crosses most of it, and each residual cost maps to exactly one lever (residue $\rightarrow$ freeze; entangling $\rightarrow$ form on the touch; the rest $\rightarrow$ feedforward bound).

7.4 Cycle 4: when the self-supervised objective itself confounds (OBJECTIVE-CONFOUND, OBJECTIVE-RELIABILITY, OBJECTIVE-OUTPUT, OBJECTIVE-REAL)

The last cycle makes the training objective itself the adversary. The joint bootstrap bites when reconstruction *genuinely prefers* the artifact, that is, when an independent high-variance nuisance dominates the loss around low-variance reals; the hypothesis is that an unfalsifiable signal used *formatively*, shaping what z encodes, escapes with no oracle where a slot *mask* cannot, since a mask filters what is encoded but cannot conjure what a recon-dominant loss never encoded. Sweeping residue dominance (residue/reals R^2 $0.17 \rightarrow 1.86 \rightarrow 6.53$) confirms both halves: ungated real-MCC collapses $0.619 \rightarrow 0.106$ while the oracle holds 0.349 (the bite Study 1’s family lacked), and the formative cross-view-invariance gate escapes to 0.302, CI-above ungated, flat across dominance ($0.295/0.285/0.302$), while expelling the residue (0.073 vs. ungated 0.736).

Two refinements land as accepted partials. A loss-weighted reliability is *not* no-regret: at low dominance the constraint still costs (0.339 vs. ungated 0.619), because the reliability signal is itself residue-confounded. Moving the selection *to the output*, training both pathways and picking by data-only cross-view reliability, *is* no-regret at every dominance; and a deterministic multi-view CCA readout individuates factors at low and mid dominance (0.570 and 0.356, the best no-oracle arms) but cannot recover what the formative objective did not encode at extreme dominance. The real port lifts half-way, and honestly so: on four real text encoders the *bite* ports cleanly (tight bottleneck $k=2$: oracle 0.715 CI-above recon 0.567; loosening to $k=32$ lifts recon to $0.702 \approx$ oracle, the knob behaves) but the *rescue* does not (formed stays flat at 0.57–0.60), because unsupervised invariance latches onto the *most-shared* direction, topic and register, which on real text is not the task factor. The formative escape is thus *terrain-gated*: it works when the task factor is the dominant shared structure, true by construction in synthetic families and false for affect on real text. Study 5 finds the same terrain gate at kernel scale, and it sharpens the demarcation: invariance alone cannot be pointed; intervention can.

8 Study 5: Demarcation at scale: the gate pays nothing on a static corpus, and one `do()` re-ignites it

The final study is the paper’s argumentative core, run as a pre-registered demarcation on the largest substrate available to us: a trained multimodal kernel, frozen DINO-ViT / MiniLM / Wav2Vec2 features of a real 40k-sample corpus, aligned by a cross-modal contrastive objective into a 64-dimensional shared space, with *modalities as the diverse views* (the b_2 reading). On a static corpus the agent has no action: $b_1 \equiv 0$ by construction. All three benches: 12 seeds, 95% CIs.

8.1 Cycle 1: the gate works and pays nothing (SCALE-STATIC)

The first bench asks whether the joint bootstrap (EM plus the emergence gate over the shared space) runs end-to-end at scale and pays as it did in PAYOFF-SYNTHETIC. The machinery ports: the gate’s data-only cross-modal-invariance ranking picks better-than-chance dimensions, statistically indistinguishable from the held-out oracle ranking (gate—random at matched capacity $J=16$: image $+0.004 \pm 0.003$; oracle—random $+0.006 \pm 0.003$). But there is no net payoff: gated—ungated is CI-below zero on both modalities (image -0.030 ± 0.005 , audio -0.011 ± 0.003), and the full 64-dimensional space beats every gated subspace at every capacity J . Per-dimension cross-modal invariance on unseen concepts is weak (mean 0.36), exactly the terrain gate of OBJECTIVE-REAL. The null is theory-consistent: the concept signal is *distributed* across all 64 dimensions, and the cross-modal contrastive objective *already self-debiases*, modality diversity doing the gate’s job during training. With $b_1 \equiv 0$, only the weak invariance half has an operating point.

8.2 Cycle 2: the null is not the readout’s fault (SCALE-NAMING)

A natural objection is that retrieval is dimension-hungry, so the payoff might live on a bounded readout. We therefore test zero-shot cross-modal *naming*, a namer trained on text and tested on image and audio: the modality-OOD axis where the PAYOFF-SYNTHETIC payoff lives. The null is stronger there. Performance is monotone in kept dimensions up to $J=64$ (the full space is best; image naming $0.075 \approx 26\times$ chance, so the grounding is real), no peak exists at any $J < 64$, and gate $\approx$ oracle $\approx$ random: the weak selection signal of cycle 1 disappears altogether. Audio is near floor (60 concepts, uninformative). There is *no transferable nuisance subspace to prune*: the contrastive objective self-debiases for every readout tested. If the gate is to pay here, something other than a readout must change.

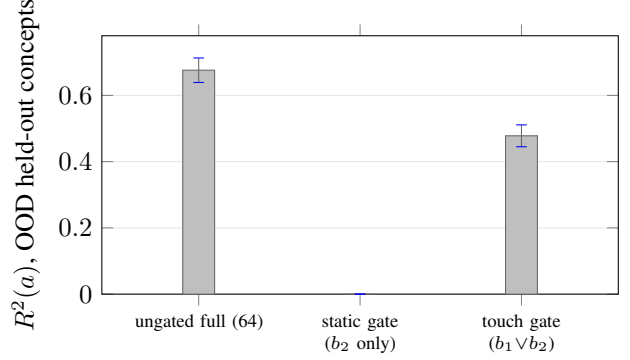


Figure 5: One `do()` re-ignites the gate (SCALE-TOUCH; 12 seeds; 95% CIs). The static (b_2 -only) gate: the regime of a static corpus, keeps 0 of 64 dimensions and cannot recover the controllable factor. Restoring a single controllable action lets b_1 fire ($\text{ctrl}_{\max}=1.00$): the same gate block anchors 27.6 dims and recovers the factor OOD. The touch gate does not beat the ungated full space (declared: the readout is dimension-hungry); it gains the operating point that was missing.

8.3 Cycle 3: one `do()` re-ignites the gate (SCALE-TOUCH)

The remaining hypothesis is that the null is the *absence of the `do()`*, not a broken gate. We give the static kernel a single controllable action, $x_a = x + a \cdot v_{\text{ctrl}}$ with a fixed *per-modality* direction, invisible to b_2 by construction and visible to b_1 through common-random-number reading, and run the same gate block. The signals split as predicted: $\text{ctrl}_{\max} = 1.00$ (b_1 sees the `do()`) while $\text{inv}_{\max} = 0.485$ (weak, as in cycle 1). The *static* gate (b_2 -only, the regime of a static corpus) anchors **0 dimensions** and recovers nothing ($R^2(a) = 0.000 \pm 0.000$), restating cycle 1. The *touch* gate ($b_1 \vee b_2$) anchors **27.6 of 64** dimensions and recovers the factor on held-out concepts: $R^2(a) = 0.478 \pm 0.033$, paired $\Delta(\text{touch} - \text{static}) = +0.478 \pm 0.033$ with CI > 0 (Table 6, Figure 5). Same gate block, same real features: only the availability of the `do()` flips recovery. Two bounds are declared: the touch gate does *not* beat the ungated full space on this readout (0.478 vs. 0.676, the same dimension-hunger as cycle 1), and the autonomous concept signal is retained cross-modally in the touch arm (image→text recall 0.069 vs. ungated 0.144, static 0.001). **The demarcation closes, on its positive side:** the gate’s machinery fires the moment a controllable signal exists; what the static corpus lacked was the signal, not the mechanism.

Demarcation, stated. *On static perceptual corpora, a well-functioning emergence gate buys nothing, because (i) the cross-modal contrastive objective already self-debiases and (ii) the only signal with leverage: controllability, is structurally absent. The gate’s payoff exists exactly when an embodied `do()` exists. None of the passive substitutes*

Table 6: The demarcation triptych at kernel scale (12 seeds, 95% CIs). Cycle 1–2: on the static corpus ($b_1 \equiv 0$) the gate selects as well as the oracle yet pays nothing, the contrastive objective self-debiases. Cycle 3: restoring one `do()` flips recovery of the controlled factor; the same gate block anchors $0 \rightarrow 27.6$ of 64 dims.

SCALE-STATIC: selection works, payoff does not (image→text / audio→text, OOD)			
gate—random @J16	+0.004 ±0.003 / +0.003 ±0.004	oracle—random @J16	+0.006 ±0.003 / +0.002 ±0.004
gated—ungated (payoff)	− 0.030 ±0.005 / − 0.011 ±0.003	(CI < 0: no payoff)	
SCALE-TOUCH: OOD recovery of the controllable factor a (R^2 , held-out concepts)			
ungated (full 64)	0.676 ±0.037	static gate (b_2 -only):	0.000 ±0.000 (keeps 0 dims)
touch gate ($b_1 \vee b_2$)	0.478 ±0.033	paired Δ touch—static:	+ 0.478 ±0.033 (keeps 27.6 dims)

tested in this paper (reconstruction, predictability, novelty, information gain, seed-stability, ARD, curiosity, invariance-alone, slowness-alone) supplies the missing operating point. Table 7 recapitulates the whole paper along this demarcation: one row per data regime, which signals were available, which trap was in play, and what gating paid in and out of distribution.

9 Limitations and honest bounds

- **The autonomous half is bounded, not solved.** b_2 is provably fooled by a residue shared across views (Table 2, C3), and this trap *ports to real perception*: with four real text encoders as views, an all-encoder meta-confounder is non-prunable (0.94, dose-dependent), and the *clean* autonomous-vs-reflection separation is unshowable on text: text encoders share too much for a strong encoder-specific residue to exist. The clean frontier for the autonomous half is vision (diverse vision encoders; autonomous illumination/background factors); which invariance-based repairs fail there, and what survives as a deployable reliability signal, we leave outside this paper’s scope.
- b_3 **is measured on synthetic dynamics only.** The slowness claim is conditional on autonomous factors being slow (their AR(1) construction makes them so here); the current claim is conditional on an autonomous factor actually carrying circulation, and the linear instrument has a declared nonlinearity floor (detection is a ratio against reversible-twin and shuffle controls, which are mandatory, not optional). Neither has a real-perception port yet.
- **Invariance cannot be pointed.** The formative rescue is terrain-gated (OBJECTIVE-REAL): unsupervised invariance recovers whatever is *most shared*, which on real data need not be the task factor. Micro-supervision does not fix this cheaply (an active k-center labelling arm covered the dominant nuisance, near-null), and multiplying environments does not either (no-oracle plateaus far

from the oracle; the lever is environment *diversity*, not count, consistent with Rosenfeld et al., 2021).

- **The gate consumes interventions – now with the asymmetry measured.** In PAYOFF-TEXT and SCALE-TOUCH the gated arm uses `do()` probes the ungated arm lacks. The equal-information control (PAYOFF-EQUALINFO) bounds what that buys: with the SAME probe budget handed to every arm, the admission rule wins at small label budgets over the same probes used as raw features, a supervised reader of the probe pairs matches it, and a random axis stays at chance. The payoff needs the interventional information plus a compressing reader; it is not an artifact of the probe budget alone. Budgets are small (8 probes; one action) and reported.
- **At kernel scale the touch gate does not beat the full representation** on dimension-hungry readouts (0.478 vs. 0.676, Table 6); the claim is the recovered *operating point* ($0 \rightarrow 28$ anchored dims; $0.00 \rightarrow 0.48$ OOD R^2), not superiority on every readout.
- **The benches and the gate were co-designed.** The synthetic family and the failure-mode cycles of Study 1 evolved together, each falsification generating the next rule, and the AUC 1.00 of Eq. 4 is measured on the same adversarial construction that motivated the three signals. The risk that the gate fits the construction rather than the problem is real; the real ports are the mitigation we offer (b_1 on text, fiction, a real game and the kernel; b_2 on four real encoders; the payoff and its pre-registered null on real perception). A bench family designed by independent hands would be the stronger test.
- **Worlds are planted where they are adversarial.** The synthetic families isolate mechanisms under a controlled adversary (never weakened); the real ports (MiniLM/Gutenberg, the 141-room game, the 40k kernel) carry their own declared bounds: templated prose, 1-D task axes, lexicon-proxy labels, frozen encoders. Add-rate tables for the Study-1 benches regenerate deterministically from the released code; the summary in Table 1 reports their reproduced modes.

Table 7: Recapitulation: data regimes $\times$ available signals $\times$ trap $\times$ payoff. $\checkmark$ = signal available and measured in that regime; $-$ = structurally absent (no port, or no action). “Mirage” = in distribution the ungated arm is competitive; the payoff of correct admission is OOD-specific throughout.

Regime (benches)	b_1	b_2	b_3	Trap in play	Payoff ID	Payoff OOD $\rightarrow$ verdict
Frozen synthetic family (UNIFIED-GATE, PAYOFF-SYNTHETIC)	$\checkmark$	$\checkmark$	$\checkmark$	structured residue; shared residue fools b_2 (C3); slow residue fools slowness, closed by current	mirage: ungated competitive	gated 0.91 $\approx$ oracle vs. 0.51 $\rightarrow$ gate pays, OOD only
Real text, IF world, frozen MiniLM (PAYOFF-TEXT, INVARIANCE-ENCODERS)	$\checkmark$ (8 probes)	$\checkmark$ (4 encoders)	$-$	invariant meta-confounder non-prunable (0.94, dose-dependent)	mirage: ungated 0.861	gated 0.885 vs. 0.584 $\rightarrow b_1$ pays; b_2 trap ports
Real game, 141 rooms (TOUCH-GAME)	$\checkmark$ (state copy)	$-$	$-$	prose under-determines exits	obs reader 0.86	obs 0.60 vs. touch map 1.0 $\rightarrow b_1$ ports alone
Sandboxed web world (PAYOFF-WEB)	$\checkmark$	$-$	$-$	planted train-time nuisance	competitive	gated 0.83 vs. 0.00 $\rightarrow$ gate pays
Multimodal kernel, static corpus (SCALE-STATIC, SCALE-NAMING)	$-$	$\checkmark$ (modalities as views)	$-$	none exploitable: contrastive self-debiases	selection = oracle = random	gated-ungated CI $< 0 \rightarrow$ pre-registered null
Multimodal kernel + one action (SCALE-TOUCH)	$\checkmark$ (1 do())	$\checkmark$	$-$	same corpus, signal restored	$-$	anchored dims 0 $\rightarrow$ 27.6; $R^2(a)$ 0.00 $\rightarrow$ 0.478 $\rightarrow$ gate re-ignited (not superior on dimension-hungry readouts)

10 Conclusion

The count and the reality of latent variables are not statistical properties of observation. Every passive admission rule we tested fails, each by a mode traceable to the observation map: the glass’s structured residue inherits predictability, novelty, information gain, persistence and stability from the real factors it is made of. The disjunction *controllable* $\vee$ *cross-view invariant* $\vee$ *temporally geometric* covers the problem on our benches precisely because it aggregates the only three signals our glass construction could not fabricate: a reproducible answer to a chosen intervention, agreement across independently drawn glasses, and broken detailed balance in time, each with a measured trap and a measured complement. The payoff of correct admission is an out-of-distribution phenomenon, invisible to in-distribution evaluation.

The two structural results frame what this buys. First, the bootstrap circularity that stops observational value criteria is escapable: anchored on an unfalsifiable interventional target, a deliberately confounded representation converges to the clean ceiling (decoy-leak 0.711 $\rightarrow$ 0.019), while the same loop anchored on the observational label never moves (0.719); the residual costs of doing this through perception decompose exactly into named levers. Second, the payoff of the entire apparatus is conditional on the existence of an interventional signal: at full kernel scale the gate selects as well as an oracle and still pays nothing on a static corpus, cross-modal contrastive training self-debiases, for every readout tested, and a single restored do() flips the gate from 0 to 28 anchored dimensions and from 0.00 to 0.48 OOD recovery of the controlled factor.

Within the perimeter measured here, frozen audited synthetic families, real frozen encoders on text and vision/audio features, one real game, one trained multimodal kernel, the demarcation is: *gate payoff requires an embodied do(); none of the passive objectives we tested substitutes for it.* We therefore suggest the identifiability debate be supplemented by an operational question: not “which unsupervised objective identifies the factors?” but “*which anchoring signal makes the fixed point escapable, and does the data regime*

contain it?” On our evidence, the answer to the second half is: static corpora do not; a single action does.

Reproducibility statement

All benches are pre-registered (hypotheses and thresholds fixed before runs), run at ≥ 10 –12 seeds (or book-folds) with 95% confidence intervals, and deterministic given (config, seed). The frozen world family ships with its audit scripts; frozen packages contain the exact code, the captured stdout of every reported run, and the emitted figures. The real substrates use frozen public encoders (all-MiniLM-L6-v2, DINO-ViT, Wav2Vec2) and public corpora (Project Gutenberg; the 141-room game via the pure-Python adventure package). Nulls are part of the release. An anonymised release of the reference worlds and benches (the code/ supplement) accompanies submission; Appendix A maps every bench name used in the text to its script. The kernel-scale features file (68 MB; frozen DINO-ViT/MiniLM/Wav2Vec2 features of the 40k-sample corpus) ships inside the supplement as the frozen reference of the Section 8 numbers, together with the script that regenerates it from the public sources (`featurise_corpus.py`; encoders, per-source caps, and held-out rules documented in the script header).

References

- Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 372–407, 2023.
- Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 2020.

- Battle, C., Broedersz, C. P., Fakhri, N., Geyer, V. F., Howard, J., Schmidt, C. F., and MacKintosh, F. C. Broken detailed balance at mesoscopic scales in active biological systems. *Science*, 352(6285):604–607, 2016.
- Brehmer, J., de Haan, P., Lippe, P., and Cohen, T. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.
- Eberhardt, F., Glymour, C., and Scheines, R. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pp. 178–184, 2005.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.
- Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., and Icard, T. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- Givan, R., Dean, T., and Greig, M. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.
- Gnesotto, F. S., Mura, F., Gladrow, J., and Broedersz, C. P. Broken detailed balance and non-equilibrium dynamics in living systems: A review. *Reports on Progress in Physics*, 81(6):066601, 2018.
- Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.
- Huben, R., Cunningham, H., Riggs Smith, L., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- Hyvärinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 460–469, 2017.
- Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.
- Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 2207–2217, 2020.
- Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. M. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.
- Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, S. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 13557–13603, 2022.
- Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. BISCUIT: Causal representation learning from binary interactions. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 1263–1273. PMLR, 2023.
- Liu, Y., Alahi, A., Russell, C., Horn, M., Zietlow, D., Schölkopf, B., and Locatello, F. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning*, volume 213, pp. 553–573. PMLR, 2023.
- Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4114–4124, 2019.
- Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

- Louizos, C., Welling, M., and Kingma, D. P. Learning sparse neural networks through L_0 regularization. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.
- Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2778–2787, 2017.
- Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.
- Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.
- Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.
- Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.
- Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.
- Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pp. 222–227. MIT Press, 1991.
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.
- Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.
- Settles, B. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- Seung, H. S., Oppen, M., and Sompolinsky, H. Query by committee. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pp. 287–294, 1992.
- Varıcı, B., Acartürk, E., Shanmugam, K., Kumar, A., and Tajer, A. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.
- von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33, 2020.
- Weiss, J. B. Coordinate invariance in stochastic dynamical systems. *Tellus A*, 55(3):208–218, 2003.
- Wiskott, L. and Sejnowski, T. J. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.
- Yao, D., Rancati, D., Cadei, R., Fumero, M., and Locatello, F. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025.
- Zhang, A., McAllister, R., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.
- Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.
- Zhang, K., Xie, S., Ng, I., and Zheng, Y. Causal representation learning from multiple distributions: A general setting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 60057–60075. PMLR, 2024.
- Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12979–12990, 2021.

A Bench index

Table 8: Bench names used in the text, their role, their substrate, and the reproducing script in the code supplement.

Bench	Role in this paper	Substrate	Script
WORLD-FAMILY	frozen audited reference worlds + audits	synthetic family (shared module)	world_family.py
GATE-PREDICTABILITY	passive cycle 1: predictability over-admits	sensorimotor world + glass	bench_gate_predictability.py
GATE-NONREDUNDANCY	passive cycle 2: post-hoc novelty under-admits	idem	bench_gate_nonredundancy.py
GATE-INFOGAIN	passive cycle 3: orthogonal growth admits residue	idem	bench_gate_infogain.py
GATE-TOUCH	cycle 4: touch settles ctrl half; ARD/seed controls	idem	bench_gate_touch.py
GATE-CURIOSITY	cycle 5a: curiosity selectivity null	idem	bench_gate_curiosity.py
GATE-VALUE	cycle 5b: info-gain valid at oracle, circular learned	idem + oracle panel	bench_gate_value.py
TOUCH-TEXT	b_1 port: counterfactual affect edits	Gutenberg + frozen MiniLM	bench_touch_text.py
TOUCH-FICTION	b_1 port: embedded $\text{do}()$ in interactive fiction	templated IF + MiniLM	bench_touch_fiction.py
TOUCH-GAME	b_1 port: real game, $\text{do}()$ by state copy	Colossal Cave Adventure	bench_touch_game.py
INVARIANCE-SPECTRAL	b_2 : spectral consensus crossing	pure-autonomous worlds	bench_invariance_spectral.py
INVARIANCE-GATE	b_2 : 5 conditions, trap, dose-response	frozen world family	bench_invariance_gate.py
INVARIANCE-ENCODERS	b_2 port: 4 real text encoders as views	Gutenberg, 4 encoders	bench_invariance_encoders.py
TIME-SLOWNESS	b_3 -slowness + time-shuffle control	frozen world family	bench_time_slowness.py
TIME-CURRENT	b_3 -current closes the slow-residue trap	frozen family + reversible twin	bench_time_current.py
TIME-OBJECTIVE-NULL	demarcation null: time as objective hurts	frozen family; InfoNCE	bench_time_objective_null.py
UNIFIED-GATE	disjunction admits at AUC 1.00	frozen world family	bench_unified_gate.py
UNIFIED-GATE-SELFAL	self-calibrated thresholds (per-signal nulls)	idem	bench_unified_gate_selfcal.py
PAYOFF-SYNTHETIC	OOD payoff, synthetic (planted nuisance)	idem	bench_payoff_synthetic.py
PAYOFF-TEXT	OOD payoff, real perception, 8 $\text{do}()$ probes	MiniLM-IF world	bench_payoff_text.py
PAYOFF-EQUALINFO	equal-information control: same probes, readers vary	MiniLM-IF world	bench_payoff_equalinfo.py
PAYOFF-WEB	OOD payoff, sandboxed web world	HTML/JS/CSS world	bench_payoff_web.py
BOOTSTRAP-ANCHOR	escape from confounded start; anchor control	box-world + decoy	bench_bootstrap_anchor.py
BOOTSTRAP-COVERAGE	selection lever: coverage vs. boundary vs. QBC	idem	bench_bootstrap_coverage.py
BOOTSTRAP-COVERAGE-DIRECTED	refuted refinement (stable-clean criterion)	idem	bench_bootstrap_coverage_directed.py
BOOTSTRAP-TEXT	anchor ports to text; selection inert (1-D)	Gutenberg + MiniLM	bench_bootstrap_text.py
BOOTSTRAP-EMBODIED	coverage re-bites on a 3-D embodied target	rollout world	bench_bootstrap_embodied.py
PERCEPTION-COST	escape through a perceiving glass; cost measured	perceived box-world	bench_perception_cost.py
PERCEPTION-FREEZE	freezing the touch-foundation recovers the α -part	idem	bench_perception_freeze.py
PERCEPTION-CAPACITY	entangling cost is a matched-resource floor	idem	bench_perception_capacity.py
PERCEPTION-FORM	forming the encoder on $\text{do}()$ sensitivity	idem	bench_perception_form.py
OBJECTIVE-CONFOUND	recon-dominant objective; formative escape	rich-perception world	bench_objective_confound.py
OBJECTIVE-RELIABILITY	loss-weighted reliability not no-regret	idem	bench_objective_reliability.py
OBJECTIVE-OUTPUT	output-side selection is no-regret	idem	bench_objective_output.py
OBJECTIVE-REAL	bite ports to real encoders; rescue terrain-gated	Gutenberg, 4 encoders	bench_objective_real.py
SCALE-STATIC	gate works, pays nothing on static corpus	multimodal kernel (40k)	bench_scale_static.py
SCALE-NAMING	null robust across readouts (zero-shot naming)	idem	bench_scale_naming.py
SCALE-TOUCH	one $\text{do}()$ re-ignites the gate	idem + injected action	bench_scale_touch.py