THE PROPRIOCEPTIVE TRANSFORMER

Functional Counterfactual Reconstruction, Candidate-Local Procedure Geometry, Cross-Model Relational Subspaces, Fixed-Prefix Control, Anomaly Science, and Governed Cognitive Improvement

Research Candidate v2.8 - Unified Relational Subspace, Mechanism, and Anomaly Research Program Edition

Logan Matthew Napolitano
Proprioceptive AI, Inc.
ORCID: 0009-0000-1927-8537
31 July 2026

Evidence posture

This v2.8 edition preserves the completed PTB-1 V2 measurement substrate, the exact-prefix theorem correction, the repaired Operator V2.1.2 harness, and the first clean candidate-local mixed development regime. It adds a separate cross-model result: an XMB1 relational map fitted on 30 exploratory roots transferred without refitting to 30 preregistered untouched roots at mean geometric R^2 0.823, and a third preregistered cohort of 105 untouched roots supported a direction-specific low-rank relational subspace. Two frozen principal directions retained approximately all full-map performance, exceeded matched-rank random controls, and their complement collapsed below zero. The result demonstrates frozen cross-model relational transfer within the registered model pair and feature contract. This unified edition also incorporates the complete anomaly-focused expansion program, formal projection and control equations, technical differentiation, alternative-mechanism tests, product acceptance requirements, and reproducibility receipts. Fixed-prefix causal steering, matched operational selection, broader model-pair transfer, substantive RSI, and AIT-1 remain open.

Keywords: transformer interpretability; proprioception; cross-model representation; relational subspace; low-rank telemetry; candidate-local state; fixed-prefix intervention; XMB1; CYGNUS; governed cognitive improvement

Abstract

Transformer interpretability usually examines hidden states from outside the model and within one model at a time. We introduce the Proprioceptive Transformer, a higher-order architecture that reconstructs a common functional origin, generates deterministic candidate procedures, measures candidate-local state geometry across depth and time, constructs model-native relational telemetry, and connects those measurements to procedure selection, fixed-prefix intervention, external verification, memory, cross-model prediction, and governed improvement. The completed PTB-1 V2 substrate contains 800 unique roots, 6,400 registered Hermes-3-Llama-3.1-8B trajectories, and 147,100 activation arrays. Exact-prefix analysis establishes that an on-manifold hidden state is reconstructible from the exact prefix and registered model contract; the decisive single-model proof targets are therefore matched operational selection and fixed-prefix causal intervention rather than Shannon irreducibility. Operator Proof V2.1.2 includes a checkpoint-aware five-token termination registry, the registered chat template, durable raw-output persistence, generation-relative indexing, truncation precedence, and an external CPU-only production supervisor. Its first clean candidate-local mixed development regime qualified prospectively at 13 correct, 19 wrong, complete semantic coverage, and 4/8 mixed roots.

The present revision adds a cross-model relational axis. XMB1 maps twelve Hermes-3-8B model-native relational coordinates into registered Llama-3.3-70B geometric targets without exchanging raw hidden vectors. A fixed artifact trained on 30 exploratory roots transferred without any fresh centering, scaling, rank selection, regularization, coefficient fitting, feature selection, or target selection to 30 preregistered untouched roots. Mean geometric R^2 was 0.823, all six targets were positive, and a pair-shuffle null had median R^2 -0.895 with empirical p=0.005. A third prospectively sealed experiment captured 420 observations from 105 untouched roots across seven prompt families. Frozen principal directions, matched-rank random directions, named coordinate families, principal complements, and a pair-shuffle null were all fixed before target access. Two principal directions retained 101.4% of the full twelve-coordinate mean R^2, exceeded the matched-rank random p95, and the complementary subspace collapsed below zero. Named motions, norms, cosines, and curvature retained materially less performance. The result supports a compact, direction-specific cross-model relational subspace within the registered scope. A long-context family exposed a low-target-variance and range-extrapolation boundary rather than an unbounded absolute-error failure.

The cross-model result does not establish raw hidden-state equivalence, causal transmission, universal model-independent coordinates, consciousness, or open-ended recursive self-improvement. It identifies a new experimentally tractable object: a frozen, low-dimensional, model-native relational state channel that predicts selected internal geometry across models with different hidden dimensions. The immediate program is to test the subspace across additional model pairs, layers, prompt families, temporal transitions, and fixed-prefix interventions while integrating direction-aware telemetry, applicability, provenance, and governed evaluation into the Proprioceptive AI product. The standalone anomaly research program is incorporated here as the experiment portfolio and receipt framework for that continuation.

Significance statement

The central new result is not a raw activation alignment. A fixed map between model-native relational measurements generalized to untouched prompts, and nearly all registered geometric transfer concentrated in two frozen principal directions whose complement failed. This supplies a compact cross-model proprioceptive channel: different models retain their native dimensions and bases while participating in a shared predictive relational state. The result is strongest as a measured transport phenomenon and a product telemetry primitive. Causal use of the subspace, transfer across additional architectures, and autonomous improvement remain separate proof obligations.

Evidence legend

Marker	Meaning
[D] Demonstrated	Implemented or measured under a registered protocol and bound to receipts or raw artifacts.
[S] Supported	Controlled evidence favors the claim, but replication, transfer, causal closure, or broader scope remains.
[P] Provisional	Preliminary result, active development result, or incomplete analysis.
[T] Theoretical	Formal proposal or architectural prediction not yet empirically closed.
[X] Retracted / superseded	An earlier claim, value, or implementation explicitly withdrawn or replaced.
[B] Boundary	A registered applicability or measurement boundary that constrains interpretation or product use.

v2.8 manuscript status

Order 0 measurement remains closed. The Operator V2.1.2 harness and clean mixed regime remain development-level foundations for fixed-prefix mechanism science. The cross-model relational bridge is independently supported within its registered Hermes-to-Llama geometric scope, and H3 supports direction-specific low-rank concentration on a third untouched cohort. No valid Hermes fixed-prefix causal mechanism verdict, matched selector verdict, broad cross-model atlas, substantive RSI cycle, or AIT-1 closure is claimed.

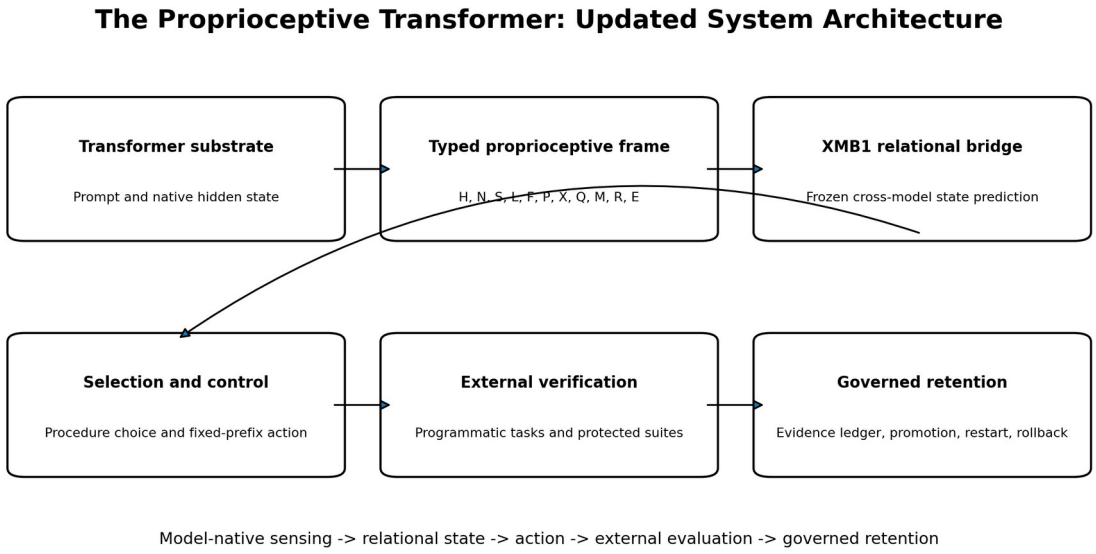

Figure 1. Updated architecture: model-native sensing and relational transfer feed action selection, control, external verification, and governed retention.

1. Introduction

The dominant transformer paradigm is extraordinarily capable but architecturally incomplete. A context is encoded, hidden computation unfolds, and a next-token distribution is sampled. The model may encode uncertainty, semantic relations, latent plans, confidence, or error, but ordinary inference supplies no validated closed loop through which the system measures its developing trajectory, compares alternative actions, intervenes on relevant geometry, exchanges compact state with other models, and retains only externally verified improvements.

The Proprioceptive Transformer begins from a cybernetic premise: intelligence is not only representation and generation. A complete cognitive architecture also requires internal sensing, state estimation, counterfactual comparison, control, memory, interoperability, and governed adaptation. Biological proprioception is not a verbal self-theory. It is a control-relevant measurement channel. The artificial analogue proposed here is functional rather than anthropomorphic.

Earlier editions organized the program around a four-order closure ladder: complete measurement, operational selection, fixed-prefix causal proprioception, and governed autonomous improvement. The measurement order closed through PTB-1 V2. Exact-prefix reconstructibility then corrected the primary predictive framing. The present revision adds a relational transfer axis that is orthogonal to the causal ladder: one model's model-native relational measurements can predict selected internal geometry of another model without raw hidden-vector interchange. This is a measured channel, not yet a causal control channel.

The cross-model result changes the scope of the architecture. Proprioception need not be confined to one model's self-observation. A system can assemble a typed, provenance-complete state from multiple model lanes, each retaining its own native basis and dimensionality. The relevant research object is not a universal activation vector but a multiplexed relational state whose directions can be tested for transfer, compression, anomaly detection, and control.

1.1 Principal contributions of v2.8

1. A completed PTB-1 V2 measurement substrate containing 800 unique roots, 6,400 trajectories, and 147,100 activation arrays.
2. An exact-prefix reconstructibility lemma, empirically confirmed by bit-identical hidden states for identical prefixes.

3. A revised operational thesis: internal state matters when timely, low-cost, selectable, and causally actionable, not because it is Shannon-independent of an exact prefix.
4. CYGNUS Operator Proof V2.1.2: an exact-Hermes, unquantized, manual-decoding, eleven-arm fixed-prefix causal architecture.
5. A live-validated checkpoint-aware termination, chat-template, RawStore, generation-relative indexing, and external CPU-only production-recovery path.
6. A prospectively frozen clean candidate-local mixed regime enabling fresh M1/M2/M3 mechanism fitting.
7. An XMB1 native-dimension relational bridge from twelve Hermes coordinates to registered Llama-70B geometric targets.
8. A zero-shot frozen-map transfer result on 30 untouched roots with mean geometric R^2 0.823, six of six positive targets, and destructive pair-shuffle separation.
9. A third prospectively sealed H3 experiment with 105 untouched roots and 420 observations across seven balanced prompt families.
10. Direction-specific low-rank concentration: top-two frozen principal directions retained approximately all full-map performance and exceeded matched-rank random controls.
11. Complement collapse: removing the leading registered directions drove mean performance below zero.
12. Named-family insufficiency: motions, norms, cosines, and curvature did not reproduce the principal-subspace result.
13. A long-context applicability boundary traced primarily to target-variance collapse and range extrapolation rather than unbounded absolute residuals.
14. An authoritative handoff mechanism that rejects narrative hashes in favor of disk-grounded cryptographic values.
15. An updated product architecture centered on typed frames, frozen-map modes, direction-aware telemetry, applicability, evidence ledgers, and governed child evaluation.
16. A falsifiable anomaly map and cross-model transfer atlas program for additional architectures, temporal transitions, and fixed-prefix subspace interventions.

1.2 Related work and differentiation

Representation similarity methods such as centered kernel alignment compare geometry across networks while controlling known pathologies of canonical-correlation statistics [16]. Model stitching tests whether one model's intermediate representation can be connected to another model's downstream computation through a learned adapter [17,18]. Recent language-model work has transferred sparse autoencoders, probes, steering vectors, and concept representations across models using affine or linear maps [19,20]. Other work reports that semantic information often occupies low-dimensional linear subspaces within individual language models [21], and recent cross-architecture analyses distinguish functional transfer from generic geometric similarity [22]. Shared logical subspaces have also been studied across natural-language and symbolic views inside a model [23].

The present result differs in four ways. First, the source and target are registered internal telemetry quantities rather than an attempt to interchange complete residual streams. Second, each model retains its native dimensionality and basis; only relational coordinates cross the bridge. Third, the strongest result is a frozen artifact applied without refitting to untouched prompt roots. Fourth, H3 tests direction identity against matched-rank random subspaces and complements, showing that transfer concentrates in a specific registered relational subspace rather than merely in any low-rank projection. The work is therefore closest to a model-native telemetry bridge and a system-level proprioceptive state contract, while remaining connected to the broader literature on stitching, feature transfer, and representational similarity.

The broader convergence literature strengthens the motivation while sharpening the boundary. The Platonic Representation Hypothesis argues that learned representations may converge toward shared statistical structure across models and modalities [24]. SVCCA and related canonical-correlation methods measure shared subspaces up to affine transformations [25,27]. Representation Engineering treats population-level directions as objects for monitoring and intervention [26]. XMB1 does not claim that those theories are already proven in the registered model pair. It contributes a narrower operational result: a frozen map over model-native telemetry transfers selected internal geometry, and a preregistered direction test identifies where that transfer is concentrated.

1.3 Technical differentiation and novelty-supporting features

Approach	Primary object	Cross-model contract	Untouched frozen transfer	Direction-specific controls	Distinct XMB1 contribution
CKA / SVCCA / representational similarity	Similarity statistic between activation sets	Compares geometry; does not create a product telemetry state	Usually descriptive rather than a frozen predictor	No registered complement and matched-rank transport test by default	XMB1 predicts registered target quantities and binds every route to artifacts and applicability.
Model stitching	Adapter connecting intermediate representations to downstream computation	Can interchange or functionally connect representations	Learned stitch is evaluated functionally; target computation is entered	Controls focus on stitch performance and adapter behavior	XMB1 exports no raw residual stream and predicts target telemetry rather than replacing target layers.

Approach	Primary object	Cross-model contract	Untouched frozen transfer	Direction-specific controls	Distinct XMB1 contribution
Cross-model SAE, probe, or steering transfer	Linear features or behavioral control vectors	Transfers interpretable features through an affine map	Some methods generalize mappings across concepts or scales	Usually tests feature transfer, not registered principal complements	XMB1 tests a frozen telemetry artifact on untouched roots and isolates a predictive principal subspace.
Low-dimensional semantic subspace studies	Within-model semantic geometry	Usually one model or one view at a time	Does not by itself establish cross-model transport	PCA or CCA concentration may be descriptive	H3 compares registered directions with matched-rank random subspaces and shows complement collapse.
XMB1 relational subspace	Model-native relational measurements and selected target internal quantities	Native dimensions, bases, and weights remain separate	Yes: exact artifact, zero fresh fitting, untouched roots	Matched-rank random, principal complement, pair shuffle, named families	A provenance-complete cross-model proprioceptive telemetry channel with direction-aware product and causal extensions.

Technical novelty posture

The manuscript supports a specific technical differentiation: frozen predictive transport of registered model-native telemetry, combined with direction-specific low-rank controls and a typed applicability contract. It does not itself determine legal novelty, patentability, or freedom to operate; those require a separate claim-by-claim prior-art analysis.

2. Evidence Discipline and Current Ledger

The theory is broader than the currently closed evidence. The evidentiary system must operate in both directions: ambition cannot promote a claim, but caution cannot erase a measured result. This edition distinguishes completed measurement, supported relational transfer, development evidence, applicability boundaries, prospective causal programs, and still-open autonomous closure.

Claim	Status	Current interpretation
Functional counterfactual reconstruction under Contract A	[D]	Registered model-derived fields reproduced across repeated and clean-process executions.
Complete PTB-1 V2 bank	[D]	800 roots, 6,400 trajectories, 147,100 arrays; structural integrity and source identities frozen.
Exact-prefix reconstructibility lemma	[D/S]	Deterministic from the model contract and empirically supported by bit-identical same-prefix states.
Operator V2.1.2 repaired harness	[D-implementation]	Termination, chat template, raw persistence, hook placement, and production recovery validated.
Clean candidate-local mixed regime	[D-development]	13 correct, 19 wrong, complete coverage, 4/8 mixed roots; mechanism-fitting surface closed.
XMB1 frozen-map zero-shot transfer	[D/S]	Fixed Hermes relational map predicted six Llama geometric targets on untouched roots at mean R^2 0.823.
H3 distributed relational code	[S]	Top-two registered directions retained approximately full performance and exceeded matched-rank random controls.
Named motions specificity	[X]	Not confirmed on fresh matched-random controls; superseded as a feature-attribution claim.
Long-context family	[B]	Target-variance collapse and range extrapolation make R^2 unstable; product applicability must downgrade or abstain.
Matched operational action selection	[T]	Fair selector comparison remains pending.
Fixed-prefix causal controllability	[T]	Registered development pilot and sealed campaign remain pending.
Governed promotion/restart mechanics	[D-operational]	Narrow parent-child promotion, distinct restart, memory continuity, and rollback mechanics demonstrated.
General research RSI / AIT-1	[T]	Requires substantial end-to-end blind improvement across three governed restart cycles.

2.1 Supersessions and corrections since v2.6

- The five previously circulated bridge artifact hashes were fabricated narrative values. Disk-grounded values are now authoritative: muB 1f772bdfddeac1c4, sdB e9d799a7e69b2546, W dc91b1fae2e15b78, U a1c2cab052b5842f, and Vt a1392c3326292528.
- The frozen-map result itself was replayed bit-exactly with zero fitting calls and nine of nine acceptance checks; the hash correction did not alter the measured bridge result.

- The original group-restricted Cycle 1 action was an identity under a target-separable estimator and is invalid as confirmatory evidence.
- The historical motions-specific result was internal cross-validation with root and tuning overlap; a fresh matched-random experiment did not confirm motions specificity.
- The third-cohort H3 result supersedes simple named-family explanations with a direction-specific relational-subspace interpretation.
- The extreme long-context R^2 is not interpreted as arbitrary absolute divergence; a dedicated audit traced it primarily to near-zero within-family target variance and range extrapolation.
- A per-family slicing bug was corrected after primary aggregates printed. Every preregistered primary aggregate remained bit-identical; the correction is append-only and explicitly excludes the pair-shuffle path, where paired permutation would destroy the null.
- Retrieval microbenchmarks are retained as product-development diagnostics, not as AIT cycles or broad evidence of semantic-memory competence.

3. Exact-Prefix Reconstructibility

3.1 Lemma

For a deterministic transformer in evaluation mode, with fixed parameters, attention mask, generation position, cache-construction procedure, and exact token prefix, the on-manifold hidden state is deterministic.

$$h_t = f_{theta}(prompt, x_{1:t}, mask, position, cache\ contract)$$

Consequently, a token-based system with the exact model, exact prefix, and unlimited reconstruction compute can recover the same hidden state. An empirical test demanding information-theoretic independence from that complete prefix is structurally degenerate.

3.2 Empirical confirmation

The completed PTB-1 bank confirmed the lemma at machine precision. Across 4,020 same-root candidate pairs sharing the same first generated token, t1 hidden states were bit-identical with relative L2 distance 0.000e+00. Across 2,649 candidate pairs sharing the same four-token prefix, t4 hidden states were likewise bit-identical. States diverged only when token sequences diverged.

> **Consequence**
>
> The architecture should not claim that hidden state contains Shannon information unavailable from the exact prefix. It should claim that hidden state is already computed, internally accessible, potentially lower-latency or lower-cost than reconstruction, directly intervenable while the preceding prefix remains fixed, and usable as a source of model-native relational telemetry.

3.3 Nondegenerate proof targets

1. Matched operational selection: under matched training roots, estimator family, feature dimension, regularization grid, and charged compute, does already-available internal state select better procedures than token, logit, confidence, best-fixed, and random alternatives?
2. Fixed-prefix causal control: after reconstructing the same prompt and exact prefix independently for every arm, does a hidden-state intervention directionally change future external success under matched geometric, temporal, and token-level controls?
3. Cross-model relational transport: can a frozen map from one model's native relational state predict another model's registered internal quantities on untouched roots without raw-vector alignment or refitting?
4. Relational-subspace control: does intervention in a transported or registered principal subspace change target-model state or behavior more specifically than matched-rank random, orthogonal, off-layer, and off-token controls?

4. Spatial-Temporal and Cross-Model State Architecture

$$h_{l,t} = a_{l,t} + d_{l,t}$$

$$d_{l,t}^{(c)} = m_{l,t} + delta_{l,t}^{(c)}$$

The spatial decomposition separates active, high-variance representational directions from lower-variance dark geometry that may remain dynamically or causally important. The temporal decomposition separates persistent orientation from the

increment produced by the present candidate trajectory. These are different update and persistence rules, not merely symbolic partitions.

4.1 Proprioceptive state bundle

Component	Role
a_(l,t)	Active semantic and task state.
d_(l,t)	Lower-variance dark-state geometry.
m_(l,t)	Persistent historical or session orientation.
delta_(l,t)^(c)	Candidate-local trajectory increment and candidate causal variable.
kappa_(l,t)	Local curvature and control geometry.
v_(l,t)	Externally calibrated verification state.
z_t	Cross-model relational latent or registered principal-subspace coordinate.
o_t	Applicability, OOD, route-disagreement, and uncertainty state.

The bundle need not exist as one monolithic vector in the base transformer. It is an engineered system state assembled from activations, persistent controller memory, candidate differences, relational geometry, intervention metadata, cross-model predictions, and external calibration. Proprioception is therefore a property of the measurement-control architecture, not a claim that a pretrained transformer spontaneously contains a neatly labeled self-variable.

4.2 Typed multiplexed frame

$$Pi_t = [H_t, N_t, S_t, L_t, F_t, P_t, X_t, Q_t, M_t, R_t, E_t]$$

H contains native hidden-state references; N normalization-derived state; S sparse-autoencoder summaries; L Lie-feature summaries; F probes and fibers; P persistent temporal orientation; X candidate-local temporal state; Q self-model fidelity; M research memory; R predicted transitions and counterfactuals; and E evaluator, tool, artifact, service, authority, and resource state. Every field carries model identity, capture site, timestamp, source hash, availability, confidence, OOD status, evidence status, and provenance.

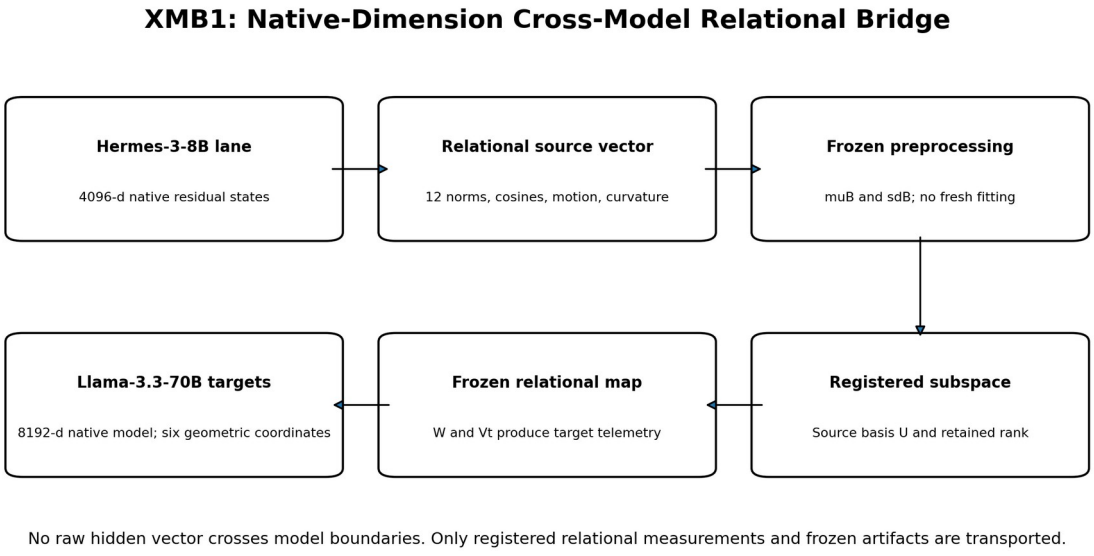

Figure 2. XMB1 preserves native dimensions and transports only registered relational measurements through frozen artifacts.

5. Counterfactual and Relational State Contracts

5.1 Contract A: functional state reconstruction

A registered model identity and revision, prompt tokens, attention mask, decoding configuration, candidate seed, relevant random-number-generator state, controller and routing state, evaluation mode, hook state, and cache-construction procedure define a reproducible functional origin. Under the registered Hermes gate, model-derived fields reproduced across 100 in-process trials and 20 clean-process executions.

Contract A result

A reproducibly reconstructed functional counterfactual origin has been demonstrated. This supports controlled candidate comparison and independent per-arm reconstruction.

5.2 Contract B: serialized runtime-state restoration

Contract B remains stronger and prospective. It would serialize and restore a live KV cache, selected hidden tensors, generation position, attention-mask state, routing state, controller state, and RNG state. Operator Proof V2 does not require Contract B because each intervention arm reconstructs its own common prefix state under Contract A.

5.3 Contract X: frozen cross-model relational transport

Contract X binds a source model identity, source coordinate manifest, standardization artifacts, mapping artifacts, target manifest, fitting scope, prompt-root cohort, procedure contract, and applicability state. It forbids silent refitting in ZERO_SHOT_FROZEN_MAP mode. A prediction is authoritative only when all artifact hashes, model identities, coordinate dimensions, and transformation paths match the registered receipt.

Contract X field	Requirement
Source model	Registered Hermes-3-8B lane and source capture contract.
Target model	Registered Llama-3.3-70B lane and target coordinate contract.
Source state	Twelve relational coordinates; no raw source hidden vector exported.
Artifacts	Disk-grounded muB, sdB, W, U, and Vt hashes.
Mode	ZERO_SHOT_FROZEN_MAP forbids fitting calls.
Target state	Registered geometric coordinates with target identity and capture provenance.
Applicability	Range, variance, family, route-disagreement, and OOD warnings.
Null	Pairing destroyed while target observations remain fixed.

6. PTB-1 V2: Completed Measurement Substrate

PTB-1 V2 provides the complete multi-task measurement substrate described in earlier editions. The bank was captured using the exact registered Hermes-3-Llama-3.1-8B model, bfloat16 precision, no quantization, five relative depths, five token horizons, eight candidates per root, scientific manifests, and artifact-derived semantic terminators.

Property	Closed result
Task roots	800 unique roots across reasoning, factual verification, coding, and instruction following.
Candidate trajectories	6,400.
Activation arrays	147,100 real arrays.
Model	Hermes-3-Llama-3.1-8B, exact registered revision, bfloat16, unquantized.
Structural integrity	0 duplicate candidate keys, 0 row-only roots, 0 array-only roots, 0 orphan arrays, 0 post-terminator continuations.
Splits	Train 3,728 rows; validation 1,368; test 1,304.
Original bank	Immutable and preserved.
Corrected labels	Separate immutable overlay after verifier-dispatch repair.

6.1 Trajectory binding and stop semantics

Every candidate is one continuous cached trajectory. Captured activations and the externally scored completion belong to the same sampled sequence. The final source derives a five-token semantic terminator union from checkpoint artifacts and stops at the first emitted member. In registered confirmation, post-terminator continuation fell from 71 of 72 audited records to zero and median semantic response length fell from 512 to 45 tokens.

6.2 Atomic persistence

Each prompt root is an atomic scientific transaction containing all eight rows and all reachable arrays. Temporary artifacts are flushed, fsynced, and atomically renamed before the ledger admits the root. An authenticated process-group SIGKILL after two durable roots followed by array-aware resume produced exact canonical equivalence with an uninterrupted registered run.

6.3 Label repair without recapture

The capture verifier implemented only legacy names while scientific manifests used newer names. The defect affected the outcome column, not trajectories, tensors, manifests, or token sequences. A prospectively frozen relabeler executed the original answer keys and produced a 6,400-row overlay with zero verifier errors while leaving the original bank immutable.

7. XMB1 Cross-Model Relational Bridge

7.1 Native-dimension contract

The source and target models do not share hidden dimension, basis, checkpoint, or raw state. Hermes-3-8B exposes its own native 4,096-dimensional states. Llama-3.3-70B exposes its own native 8,192-dimensional states. XMB1 derives compact model-native relational measurements from each lane and learns a registered map between source relational coordinates and target internal quantities. The map does not imply raw residual-stream equivalence.

The registered source vector contains twelve coordinates derived from layer norms, consecutive-layer cosines, depth motion, and curvature. The target vector contains geometric quantities and, in the broader bridge, sparse-feature and probe-derived quantities. Constant targets were removed. A square invertible rotation was rejected as a destructive null because it preserved all information; source-target pair shuffling became the registered alignment-destruction control.

7.2 Frozen artifact and transformation contract

Artifact	Authoritative hash prefix	Role
muB	1f772bdfddeac1c4	Source centering vector.
sdB	e9d799a7e69b2546	Source scale vector.
W	dc91b1fae2e15b78	Registered cross-model coefficient matrix.
U	a1c2cab052b5842f	Source-side SVD / principal basis.
Vt	a1392c3326292528	Target-side SVD basis / registered map component.

These values were recomputed from disk after a narrative handoff contained five fabricated alternatives. The fabricated values are preserved only as invalid historical data. The authoritative handoff generator now prohibits cryptographic values from being promoted without a disk or live-process source.

7.3 Formal frozen-map mechanism and control geometry

Let b_i in R^12 denote the registered Hermes relational source vector for observation i. The frozen source standardization is z_i = (b_i - muB) / sdB, with elementwise division. Let the registered coefficient matrix admit the source-side singular decomposition W = U Sigma V^T. For rank k, P_k = U_k U_k^T is the registered principal projector and W_k = U_k Sigma_k V_k^T is the corresponding reduced map.

$$z_i = (b_i - muB) / sdB$$

$$W = U\ Sigma\ V^T, \quad P_k = U_k\ U_k^T, \quad W_k = U_k\ Sigma_k\ V_k^T$$

$$a\_hat\_i^{(full)} = z_i\ W, \quad a\_hat\_i^{(k)} = z_i\ P_k\ W = z_i\ W_k$$

The principal-complement route uses z_i (I - P_k) W. A matched-rank random route substitutes an outcome-independent orthonormal projector Q_k Q_k^T for P_k. The pair-shuffle null evaluates z_pi(i) W against the fixed target a_i, so source-target alignment is destroyed without permuting targets with their own sources. These transformations share the same scaler, coefficient matrix, target manifest, and evaluation path; only the registered subspace intervention changes.

$$a\_hat\_i^{(comp,k)} = z_i\ (I - P_k)\ W, \quad a\_hat\_i^{(rand,k)} = z_i\ Q_k\ Q_k^T\ W$$

$$a\_hat\_i^{(shuffle)} = z\_pi(i)\ W \ \text{ evaluated against fixed } a_i$$

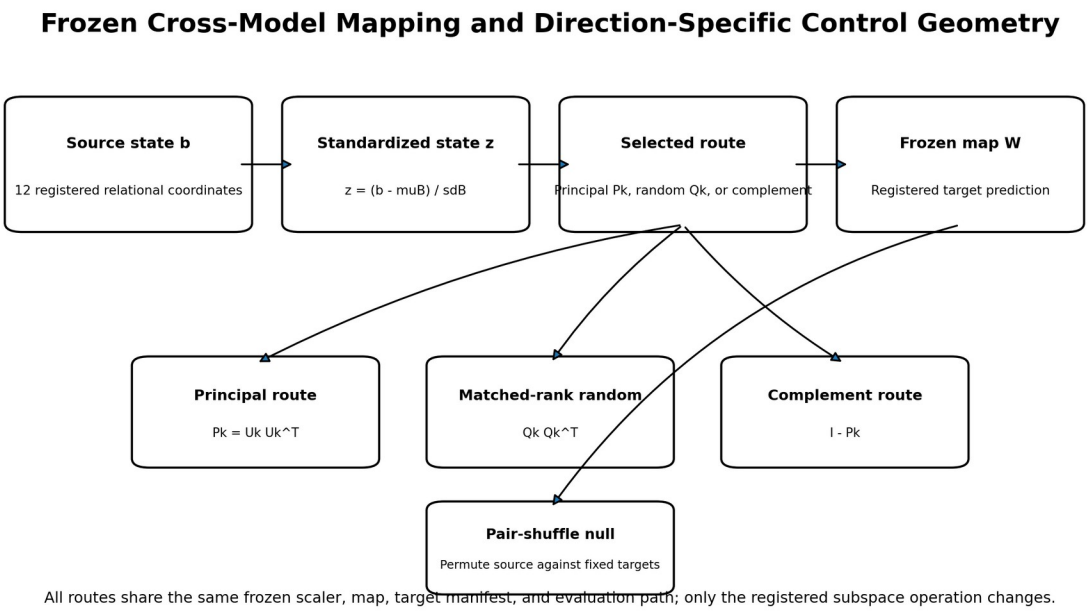

Figure 3. Frozen map and direction-specific control geometry. Principal, matched-random, complement, and pair-shuffle routes share the same artifact contract.

7.4 Statistical estimands and interpretation

For target j on an untouched cohort, R_j^2 = 1 - sum_i(a_ij - a_hat_ij)^2 / sum_i(a_ij - mean_i a_ij)^2. The primary geometric score is the unweighted mean across the six preregistered geometric targets. Retention_k is the principal-route mean divided by the full-route mean. Direction specificity is evaluated against the frozen matched-rank random distribution, while complement collapse tests whether predictive performance remains after the leading registered source directions are removed.

$$R_j^2 = 1 - SSE_j / SST_j, \quad mean\_R^2 = (1/6) \sum_{j=1}^{6} R_j^2, \quad retention_k = mean\_R_k^2 / mean\_R_{full}^2$$

R^2 is unstable when within-cohort target variance approaches zero. The manuscript therefore reports absolute residuals, range extrapolation, target variance, and applicability state alongside R^2. The long-context result is treated as a measurement and product boundary rather than silently excluded.

7.5 First untouched-root zero-shot transfer

A bridge fitted on 30 exploratory roots was applied unchanged to 30 preregistered untouched roots. No centering, scaling, rank, regularization, coefficient, feature, or target parameter was refit. The replay reproduced bit-exactly with zero fitting calls.

Target	Zero-shot R^2
a_norm_l32	+0.995086
a_norm_l50	+0.991109
a_cos_l32_l50	+0.887891
a_depth_motion_rel	+0.883827
a_norm_ratio	+0.734343
a_depth_motion_norm	+0.446386
Geometric mean	+0.823107
Geometric median	+0.885859

All six registered geometric targets were positive. The pair-shuffle null had median R^2 -0.8950, p95 -0.6656, and empirical p=0.0050. The frozen map outperformed a separate fresh-cohort fold-refit analysis, consistent with the full exploratory corpus providing a more stable coefficient estimate than twenty-four roots per fresh fold. This does not prove that frozen maps always outperform refitting; it records the observed scope.

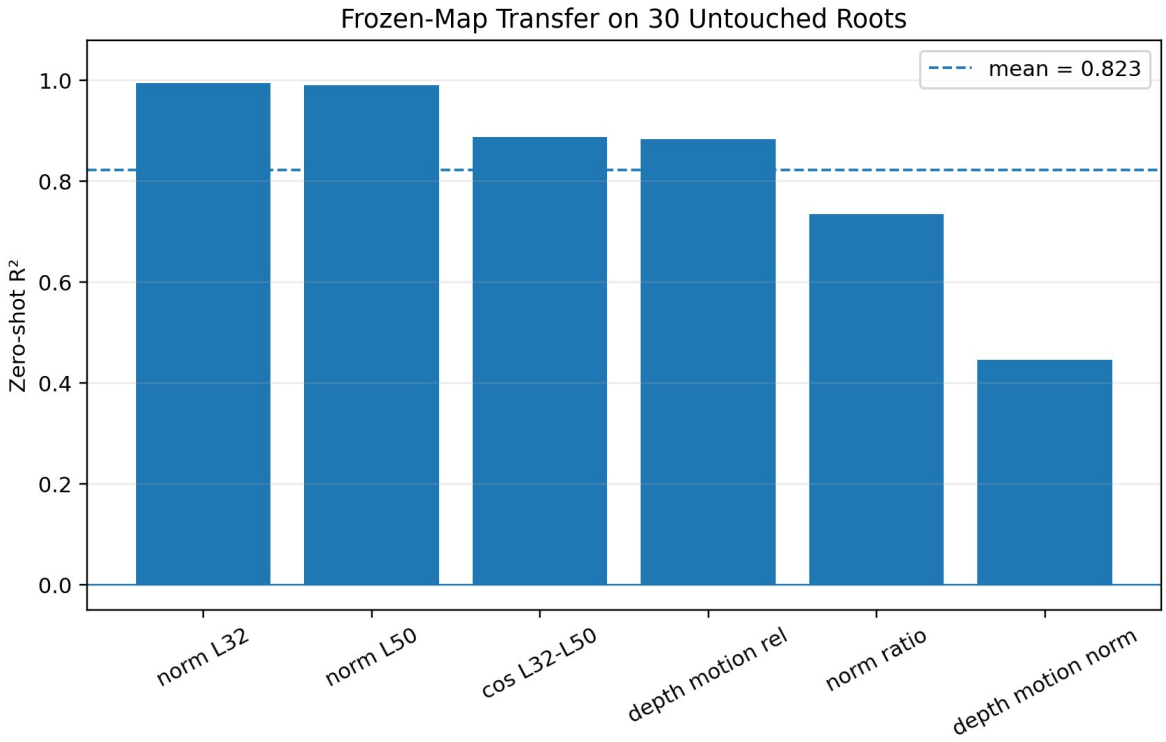

Figure 4. Target-wise zero-shot performance of the exact frozen bridge on 30 untouched roots.

7.6 What the zero-shot result establishes

- A registered cross-model relational artifact transports to untouched prompt roots without refitting.
- The transported signal predicts selected target-model internal geometry rather than only external behavior.
- The result survives a destructive source-target pairing control.
- The models retain incompatible native dimensions and exchange no raw hidden vectors.
- The result is scoped to the registered models, layers, coordinates, prompt distribution, and frozen artifact contract.

7.7 What it does not establish

- Raw hidden-state equivalence or invertible alignment between complete residual streams.
- Causal transmission of state from Hermes to Llama.
- A universal coordinate system shared by all transformers.
- Behavioral control or capability transfer at the strength of geometric prediction.
- Consciousness, subjective access, or strategic agency.

8. H3: Distributed Relational Code

8.1 Hypothesis and preregistration

The failed motions-specific attribution suggested that semantic feature-family labels were not the correct organizing variable. Exploratory source geometry instead showed effective rank near three and strong performance concentration in leading principal directions. H3 therefore preregistered the hypothesis that retained relational rank and direction identity predict frozen-map performance better than named coordinate-family membership.

Registration item	Value
Roots	105 untouched roots; 15 per family across seven prompt families.
Observations	420; four frozen procedures per root.
Seal	2026-07-31T19:42:53Z.
First target access	2026-07-31T19:43:36Z; seal preceded access by 43 seconds.
Root overlap	0 by root ID; 0 by prompt-body hash against 60 prior roots.
Capture	420/420 successful; 0 retry; 0 failure; 26.6 minutes.
Conditions	FULL_12; top-1/2/3/6; matched-rank random 1/2/3/6; complements 1/2/3; four named families; pair shuffle.
Fit calls	0.

8.2 Primary results

Condition	Mean R^2	Retention	Matched-random p95	Exceeds p95
FULL_12	+0.5128	1.0000	-	-
PRINCIPAL_TOP_1	+0.4497	0.8770	+0.3192	True
PRINCIPAL_TOP_2	+0.5199	1.0138	+0.3511	True
PRINCIPAL_TOP_3	+0.5166	1.0075	+0.4855	True
PRINCIPAL_TOP_6	+0.5127	0.9998	+0.5861	False
NAMED_MOTIONS	+0.3161	0.6164	-	-
NAMED_NORMS	+0.2386	0.4654	-	-
NAMED_COSINES	+0.1326	0.2585	-	-
NAMED_CURVATURE	-0.1468	-0.2864	-	-
PRINCIPAL_COMPLEMENT_1	-0.0114	-0.0223	-	-
PRINCIPAL_COMPLEMENT_2	-0.0816	-0.1591	-	-
PRINCIPAL_COMPLEMENT_3	-0.0783	-0.1527	-	-

The top two principal directions retained 101.4% of the full twelve-coordinate aggregate. The value above one indicates that lower directions slightly degraded generalization under this cohort; it is not evidence that a restriction contains more raw information than the full source. Top-one, top-two, and top-three conditions exceeded their matched-rank random p95 values. Top-six did not, because rank-six random subspaces frequently retained much of the available low-dimensional source information.

The complements provide the strongest necessity-style passive control. Removing only the leading direction reduced mean R^2 to approximately zero. Removing the top two or three drove mean R^2 below zero. The pair-shuffle null had median R^2 -0.8793. No fitting call executed during any condition.

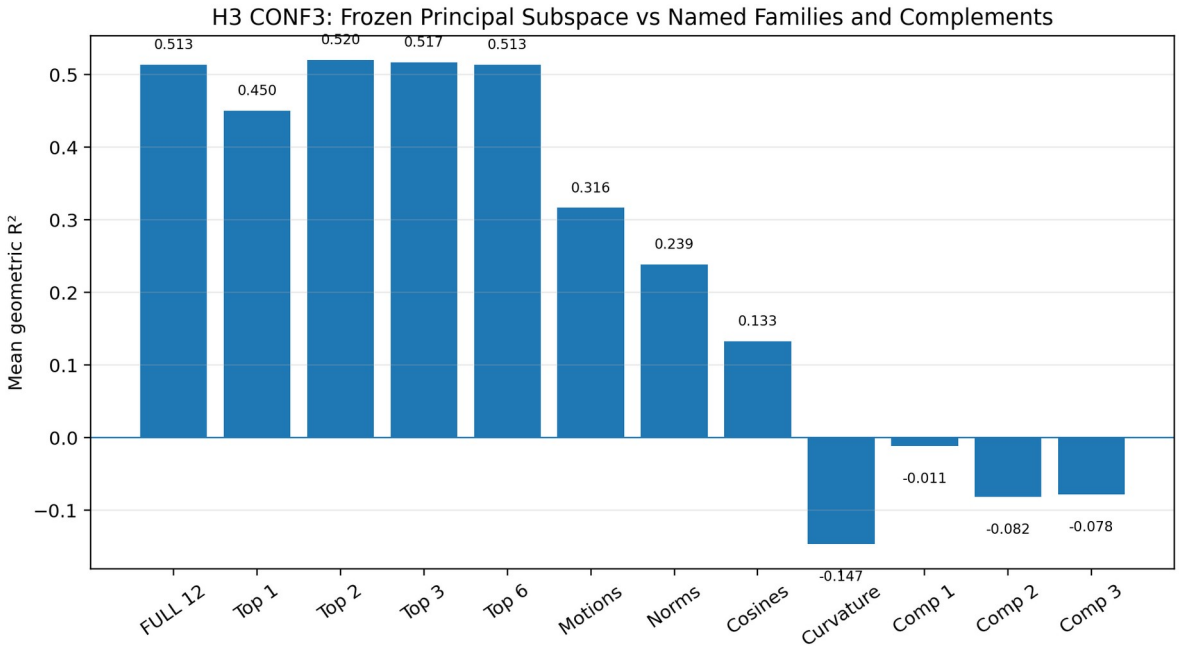

Figure 5. H3 primary condition results. Registered principal directions outperform named families, while their complements collapse.

8.3 Source spectrum

The registered source spectrum was 0.884, 0.068, 0.022, followed by a small residual tail, yielding effective rank 1.63. Approximately 95.2% of registered source variance lies in the first two directions. The H3 result is therefore both low-rank and direction-specific: matched-rank random subspaces at ranks one through three were weaker than the corresponding registered directions.

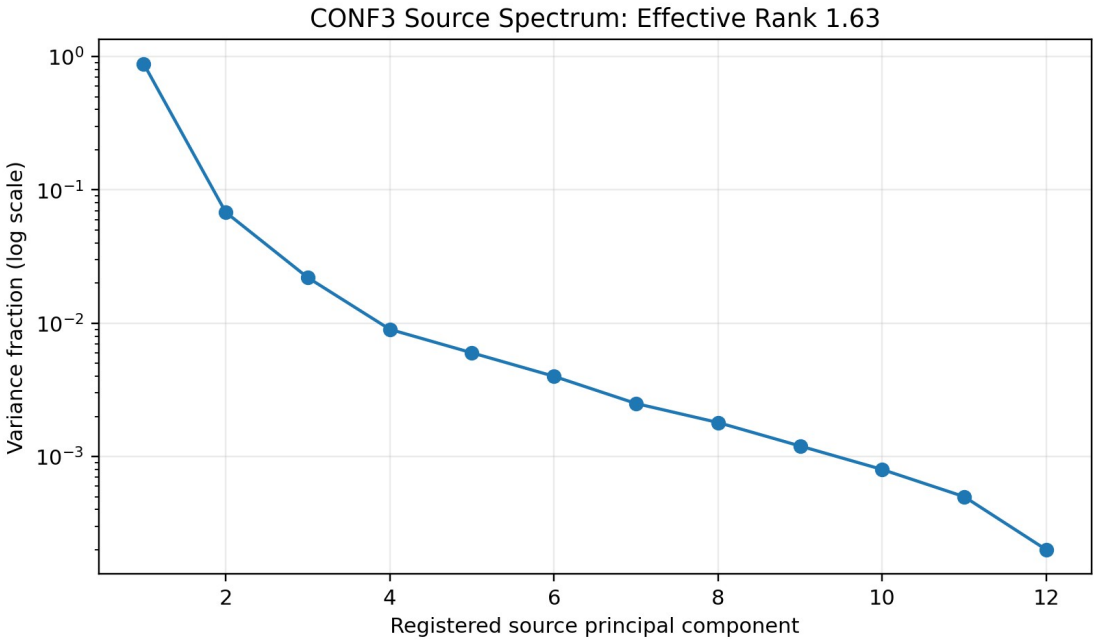

Figure 6. CONF3 source spectrum. The first two registered directions dominate the relational source geometry.

8.4 Named-family comparison and the curvature anomaly

No named family recovered the top-two principal result. Motions was the strongest standalone family, retaining 61.6%, but its fresh-cohort uniqueness claim had already failed against matched-random subsets. Norms retained 46.5%, cosines 25.9%, and curvature was harmful. An exploratory overlap audit found that family projection energy did not order standalone retention: curvature had the second-highest top-two projection energy but the worst predictive result, while norms had relatively low top-two energy but the largest leave-family-out cost.

The mismatch is an important anomaly. The predictive subspace is not explained by coordinate count or simple energy overlap. Coefficient sign, target loading, covariance cancellation, scaling, and cross-family interactions remain candidate mechanisms. Motions may have appeared historically privileged because it was the strongest human-named partial view of a direction that actually spans several coordinate families.

Family	PC1 energy	PC1-2 energy	Retention	Leave-family-out delta
Motions	0.0723	0.2389	0.6164	-0.1842
Norms	0.0441	0.0523	0.4654	-0.2938
Cosines	0.1382	0.1425	0.2585	-0.0058
Curvature	0.0092	0.1554	-0.2864	+0.0898

8.5 Long-context low-variance boundary

The registered long-context family produced aggregate R^2 near -72.6. A target-level audit showed that the extreme value was driven primarily by target-variance collapse, compounded by range extrapolation. The long-context prompts were structurally similar; several target coordinates varied by less than one to eleven percent of their cohort variance. Moderate absolute residuals were therefore divided by a near-zero R^2 denominator.

Target	Long-context variance	Cohort variance	Variance ratio	Family R^2	Residual RMS
a_norm_l50	0.4546	57.3837	0.0079	-47.25	4.684
a_norm_l32	0.7181	69.5777	0.0103	-15.45	3.437
a_depth_motion_norm	0.0123	0.3954	0.0311	-342.51	2.057
a_cos_l32_l50	0.0002	0.0021	0.1114	-10.48	0.052

This is not a post-hoc exclusion. Long context remains inside the registered primary aggregate. The correct product interpretation is a LOW_TARGET_VARIANCE and RANGE_EXTRAPOLATION applicability boundary. A prospective follow-up must preserve prompt text and token length, vary templates and stage counts, and use absolute and normalized error metrics that remain meaningful when within-family target variance is small.

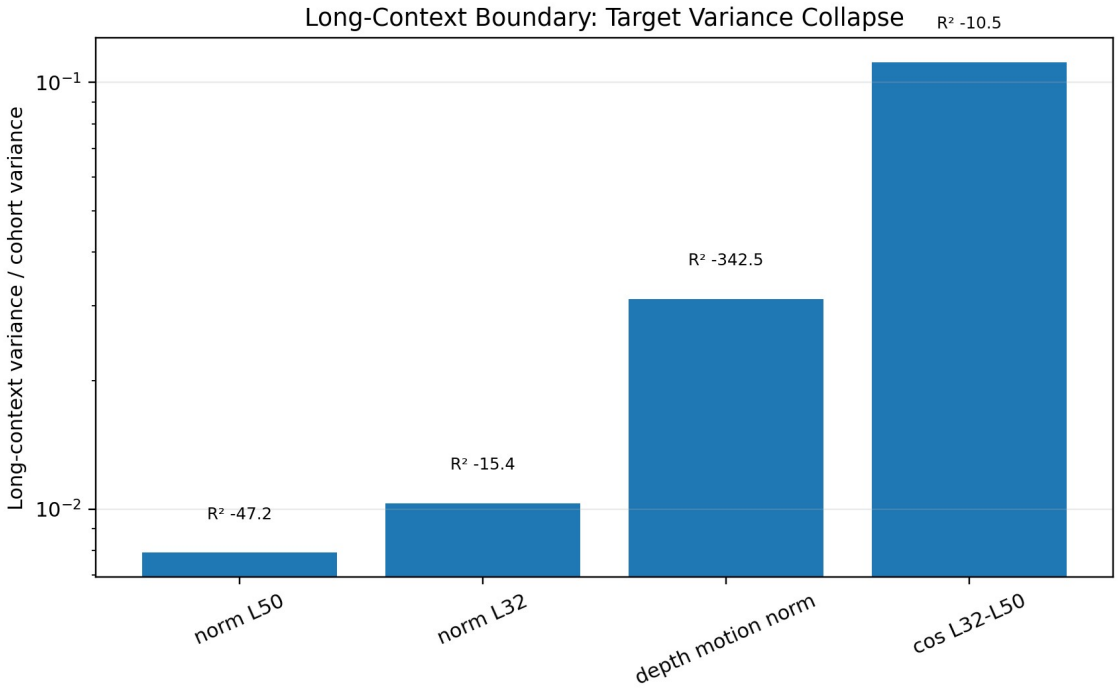

Figure 7. Selected long-context target variance ratios. Extreme negative R^2 arises in a low-variance regime.

8.6 Analysis correction and null integrity

The per-family scorer initially passed a row-subset of predictions against the full target matrix and crashed. The correction aligned target rows with selected prediction rows. Primary aggregates had already printed, and outcomes were therefore visible before the correction. A versioned audit demonstrated that every preregistered primary aggregate was bit-identical before and after the repair. The correction affected only per-family and grouped-bootstrap slicing.

The audit also identified a critical boundary: applying the same selected indices to both source and target in a pair-shuffle path would permute pairs together and produce a false positive near the full result. The registered null correctly permutes source alignment against fixed targets. This distinction is now enforced in the analysis contract.

H3 verdict

H3_DISTRIBUTED_RELATIONAL_CODE is SUPPORTED_WITHIN_REGISTERED_SCOPE. On 105 untouched roots, the frozen bridge depends disproportionately on a small registered principal subspace. Two directions retain approximately full performance, matched-rank random directions are weaker, and the complementary subspace collapses. The result does not yet establish causal control, universal transfer, or a model-independent ontology.

8.7 Candidate mechanisms and decisive anomaly tests

The passive result establishes transport and concentration, not why the directions exist. The most useful mechanistic hypotheses are those that make different predictions under new cohorts, new model pairs, coefficient perturbations, and causal interventions.

Candidate explanation	Why it fits current evidence	Distinct prediction	Decisive next test
Shared low-dimensional relational factor	Top-two directions explain most source variance and retain nearly all transfer.	Direction angles and target loadings remain stable across cohorts and related model pairs.	Fourth-cohort subspace-angle replication and cross-model atlas.
Denoising / harmful residual directions	Top-two slightly exceeds FULL_12; curvature is harmful alone.	Removing specific lower directions improves frozen generalization reproducibly.	Prospective rank curve, coefficient ablation, and shrinkage comparison.
Cross-family interaction and covariance cancellation	Named-family retention does not track projection energy; norms and curvature behave anomalously.	Conditional contributions depend on combinations, signs, and target loadings rather than single-family energy.	Shapley-style decomposition, sign audit, conditional ablation, and covariance-preserving controls.
Target marginal or prompt-family regularity	A simple structure could predict targets without genuine paired transfer.	Target-only and root-identity baselines approach the frozen map, or pairing controls fail.	Target-only predictors, root leakage controls, new prompt domains, and pair shuffle.
Approximate shared relational geometry	Frozen map transports across dimensions without raw-vector interchange.	Additional architecture pairs exhibit related low-rank channels, perhaps with rotated bases.	Bidirectional atlas with untouched roots and subspace alignment.
Static telemetry without shared dynamics	Current result predicts state magnitude, not transitions.	Temporal map fails beyond static and target-autoregressive baselines.	Paired t to t+1 transition experiment and temporal shuffle.

Candidate explanation	Why it fits current evidence	Distinct prediction	Decisive next test
Causal control channel	The predictive subspace is read or written by control-relevant components.	Fixed-prefix intervention shows directional effects and matched-control specificity.	Rank-matched principal, complement, random, off-layer, off-token, and token-matched causal arms.

8.8 Threats to validity already controlled and still open

Threat	Current control or evidence	Residual risk / required extension
Fresh-data leakage	Zero root-ID and prompt-body overlap; preregistration precedes target access.	Independent external replication and additional corpora.
Refitting masquerading as zero shot	Instrumented zero-fit replay; disk-grounded frozen hashes.	Independent implementation of the artifact-loading path.
Low rank alone explains performance	Matched-rank random subspaces at ranks one to three are weaker.	More random draws and new-cohort direction stability.
Projection implementation artifact	Principal complements collapse and pair shuffle is destructive.	Second implementation, algebraic test vectors, and patching controls.
Named-family cherry-picking	All named families and matched-random subsets are reported; motions specificity is superseded.	Reliability-matched target sets and predeclared family alternatives.
R^2 denominator instability	Target variance and absolute residuals audited; long context remains included.	Prospective variance-controlled long-context cohort and robust metrics.
One target or family drives aggregate	Per-target, per-family, bootstrap, and influence analyses reported.	Larger cohorts and hierarchical modeling.
Causal overinterpretation	Paper explicitly labels H3 passive and predictive.	Fixed-prefix intervention with matched controls and protected suite.

9. What the Original Predictive Program Established - and Did Not Establish

The original predictive program generated valuable constraints but did not close single-model proprioception. Corrected development base rates ranged from near-ceiling reasoning and factual verification to near-floor instruction following, so pooled performance mostly separated task classes rather than candidates from one common root. A prospectively frozen root-conditioned t1 analysis compared candidate controls with controls plus t1 delta. The primary control already achieved pairwise accuracy 0.9719 and selection accuracy 0.9455. Adding delta increased pairwise accuracy to 0.9772 and selection to 0.9636, but worsened conditional log loss; the root-bootstrap information-gain interval crossed zero.

At t1, the first token is the discrete readout of the same post-token state being measured. Conditioning on the first token removes nearly all legitimate t1 variation; omitting it leaves an empty within-root baseline because prompt, task, pre-sampling logits, entropy, and margin are root-constant. Exact-prefix reconstructibility generalizes this problem to later horizons. The result is therefore retained as a readout-redundancy stress test rather than the final proof of an internally useful control architecture.

The cross-model bridge is not a resurrection of the superseded Shannon-increment claim. It predicts registered target-model quantities from source-model relational telemetry under a frozen artifact and realistic resource contract. A hypothetical system with exact access to both complete models could reconstruct both states; the operational result is that a compact already-available source channel transports useful target-state estimates without rerunning or exchanging target hidden tensors.

10. Revised Order 1: Matched Operational Action Selection

The operational question is whether an already-available internal state provides a useful selector under constrained resources. A token-based system could reconstruct hidden state by running the model, but that computation is not free. The fair comparison therefore matches training roots, estimator family, output feature dimension, parameter budget, regularization grid, candidate procedures, and external verifier, while separately reporting additional forward calls, latency, and compute.

Action	Procedure
A0	Direct answer.
A1	Explicit derivation plus verification.
A2	Adversarial critique followed by revision.
A3	Independent branches followed by aggregation.

Selector	Information and fairness contract
Hidden-state selector	Pre-action prompt-conditioned state; train-only residualization and dimensionality reduction.
Token selector	Prompt-token hashed or fixed token features; same estimator family and output dimension.
Logit/confidence selector	Prompt-boundary logits or frozen projection plus entropy, margin, and confidence.
Best fixed	Procedure selected on validation, then frozen for test.
Always A0	Direct-answer baseline.
Random	Uniform procedure selection.

- Hidden selector exceeds validation-selected best fixed external reward on sealed test.
- Hidden selector exceeds matched token and logit/confidence selectors.
- Paired root-bootstrap intervals exclude no improvement for required comparisons.
- Selection occurs before external execution.
- Inference latency and extra model-forward cost are reported.
- The result reproduces from a clean restart.

11. Revised Order 2: Fixed-Prefix Causal Proprioception

Prediction does not establish control. The decisive experiment fixes prompt and token prefix, reconstructs the same state independently for each arm, changes hidden state at one explicit layer and token horizon, and measures future external success. No arm inherits a cache mutated by another arm.

$$Success(+delta) > Success(baseline) > Success(-delta)$$

11.1 Causal transaction

1. Tokenize the prompt and run prompt prefill with KV cache.
2. Generate and hash one shared prefix.
3. Reconstruct the exact same prompt and prefix state independently for every arm.
4. Verify byte-identical prefix IDs and hashes.
5. Install the intervention hook only for the frozen target layer and token step.
6. Run exactly one intervention-bearing forward pass and remove the hook immediately.
7. Continue generation under the same decoding and stop semantics.
8. Score externally and bind all arm metadata to source, model, direction, layer, horizon, and prefix hashes.

11.2 Registered arms

Arm	Purpose
Baseline	Fresh reconstruction with no hidden-state modification.
Amplify	Add positive normalized success direction.
Negative	Subtract the direction.
Suppress	Reduce the existing projection without reversing it.
Reverse	Reflect the directional component through zero.
Delete	Remove the registered direction or subspace projection.
Matched-norm random	Control perturbation magnitude.
Orthogonal placebo	Control geometrically unrelated movement.
Off-layer	Test layer specificity.
Off-token	Test temporal specificity.
Token-matched	Match immediate logit effect with distinct geometry.
Principal-subspace arm	Prospective extension: perturb registered cross-model relational direction or complement under matched controls.

11.3 Causal specificity

A positive arm is insufficient if generic perturbations, token-matched controls, or off-target interventions behave similarly. Causal closure requires directional ordering, paired intervals, matched-control attenuation, protected-suite non-regression, and clean-restart replication. The H3 principal subspace provides a new candidate geometry but does not receive causal status until it passes this program.

12. CYGNUS Operator Proof V2

The historical operator artifact existed only on the local workstation and had never run on the primary execution host. It was preserved read-only as a historical seed and prospectively superseded. The new workspace isolates baseline, source, specifications, tests, runs, artifacts, and receipts. Operator V2.1.2 uses the registered model and tokenizer contract, checkpoint-aware terminator logic, the registered chat template, raw persistence before scoring, generation-relative intervention indexing, and external production recovery.

Module	Role
manual_decode2.py / termination_registry.py	Registered prefill, stopping, raw persistence, cache reconstruction, exact-prefix capture, and hook site.
interventions.py	Eleven algebraically and geometrically distinct arms.
directions.py	M1-M4 construction and train-only subspace fitting.
selection.py	Validation-only mechanism, layer, horizon, amplitude, and rank selection.

Module	Role
verifiers.py	Programmatic development and protected-suite scoring.
receipts.py	Source, model, environment, result, and clean-restart bindings.
tasks.py	Development tasks requiring non-answer-bearing prefixes and post-intervention continuation.
window_supervisor.py	CPU-only production stop, GPU drain, experiment ownership, restoration, and verification.

12.1 Competing mechanisms

Mechanism	Construction	Current status
M1	Mean successful-state minus mean failed-state direction.	Synthetic recovery pass; real validation pending.
M2	Regularized supervised linear direction after nuisance residualization.	Synthetic recovery pass; real validation pending.
M3	M1 or M2 projected into a train-fitted low-variance dark subspace.	Distinct synthetic mechanism; real validation pending.
M4	Action-conditioned transition direction.	Optional extension.
M5	Registered cross-model principal relational subspace.	Supported predictive geometry; causal intervention pending.

13. Operator Proof V2 Implementation State

Operator Proof V2.1.2 is the prospective causal implementation of the Proprioceptive Transformer. The historical seed remains read-only and is not treated as causal evidence. The repaired implementation is bound to the registered Hermes model and tokenizer contract, uses deterministic greedy A0-A3 procedures, separates semantic correctness from response-format compliance, and preserves raw generated token IDs including the emitted terminator while restricting semantic parsing to the pre-terminator prefix.

The live smoke produced 16/16 natural completions and no decoder invariant failure. The harness is frozen for development research unless a hard invariant fails. The next admissible single-model scientific step remains fresh M1/M2/M3 capture and fitting on 32 train, 16 validation, and 8 mechanics roots, followed by dose calibration and the eleven-arm pilot. The cross-model H3 result supplies an additional candidate subspace but does not replace this clean-label causal program.

14. External Experiment Supervision and Production Recovery

The first pilot windows revealed a systems-level design error: the CUDA-holding experiment process also owned production restoration. That process could retain model references, terminate without unwinding, or die before freeing GPU memory. V2.0.4 moved restoration authority into an external CPU-only supervisor that imports neither torch nor transformers and never owns a CUDA context.

1. Fingerprint production, the live service ports, source hashes, model identity, corrector state, and GPU occupancy.
2. Acquire GPU and window locks.
3. Stop only the registered production process.
4. Poll until the registered physical GPU is stably released.
5. Launch the experiment as a direct child process with a distinct process group.
6. On normal exit, exception, signal, or forced kill, preserve logs and continue cleanup.
7. Poll until all experiment GPU processes disappear and memory is stably free.
8. Restart production only when the port is absent and the frozen source hash still matches.
9. Verify readiness, model identity, corrector state, independent audit service, and GPU baseline.
10. Release locks only after post-restore verification.

The supervisor repeatedly restored the live capture service after forced termination, exceptions, and pre-generation aborts while preserving the independent audit service. This closes a production-recovery failure class for the frozen harness and is also a system-level proprioceptive mechanism: the research architecture senses operational state and restores a registered configuration.

15. Clean Outcome Contract

A causal direction is meaningful only if positive and negative labels represent cognitive success and failure rather than parser quirks, formatting conventions, or token-budget exhaustion. Earlier calibration violated this requirement through comma-blind parsing, text-based stopping, short token caps, and truncation-as-failure. Those measurements were superseded within their affected scope.

15.1 Semantic stopping and raw observations

- Stop only on artifact-declared semantic terminator token IDs.
- Do not use decoded-text blank lines as a stopping rule.
- Record terminator identity and position.
- Exclude the terminator from verifier text.
- Classify maximum-token exhaustion separately from semantic completion.
- Preserve raw text and token IDs before parsing or scoring.

15.2 Outcome taxonomy

Outcome	Definition	Use in fitting
CORRECT_ARITHMETIC	Parsed final answer matches the frozen verifier.	Positive outcome.
WRONG_ARITHMETIC	A valid final answer is present but incorrect.	Primary negative outcome.
MAX_TOKEN_TRUNCATION	Generation reaches the cap before a usable final answer.	Always unscorable.
VERIFIER_ERROR	Verifier raises, times out, or cannot execute.	Hard data-quality failure.
MALFORMED_OUTPUT	Output violates the registered answer contract.	Separate contract outcome.
NO_FINAL_ANSWER	No parseable final answer after semantic completion.	Tracked separately.

Clean-label invariant

A regime may enter mechanism fitting only when semantic outcomes identify procedure-local success rather than formatting or censoring. The frozen gate requires scorable coverage at least 0.85, at least eight correct and eight wrong trajectories, at least two mixed roots among eight, zero verifier errors, no more than one censored output, and complete decoder-invariant compliance.

16. Adaptive Difficulty Calibration

Difficulty is treated as an experimentally selected system parameter rather than intuition. Under the repaired harness, the immediate objective is not to map a universal capability curve before causal work. It is to identify the first prospectively ordered structural regime that yields clean, semantically scorable, within-root procedure variation.

The original observed capability cliff was superseded by interacting measurement confounds. Historical sequences are retained as instrument records, not authoritative competence curves. The clean sequential search froze level order, qualification rules, procedure bank, semantic prompt, decoder, token budget, and namespaces before generation. It stopped at the first qualifying regime.

Structural regime	Correct	Wrong	Coverage / mixed roots	Decision
three_term_2d	29	3	1.000 / 1 of 8	Reject: ceiling and insufficient wrong support.
three_digit_x1_highcarry	13	19	1.000 / 4 of 8	Qualifies: first clean mixed regime.

The same underlying arithmetic root can now produce both success and failure as deterministic procedure choice changes. This is the candidate-local variation required to learn procedure-sensitive internal geometry without confusing task difficulty with outcome. It closes the development-surface gate but does not establish selector superiority or causal steering.

17. Competing Internal Control Mechanisms

Operator Proof V2 does not crown one direction by theoretical preference. It constructs multiple candidate mechanisms, measures them externally, eliminates mechanisms that behave like generic perturbations, and composes the strongest survivor into the next layer of the architecture. The H3 principal subspace becomes a new candidate mechanism only after it is re-expressed in a valid intervention geometry and tested with matched controls.

$$M1: \hat{d}_1 = \mathrm{normalize}(\mu_{success} - \mu_{failure})$$

$$M2: \hat{d}_2 = \text{normalize}(\text{argmin}_w [L_{train}(w; \text{residualized } h) + \lambda ||w||_2^2])$$

$$M3: \hat{d}_3 = \text{normalize}(P_{dark} \hat{d}_1) \text{ or } \text{normalize}(P_{dark} \hat{d}_2)$$

$$M4: \hat{d}_4 = \text{normalize}(E[\Delta h \mid \text{successful action}] - E[\Delta h \mid \text{failed action}])$$

$$M5: Z = U_k^T ((b - \mu_B) / \sigma_B), \text{ with causal use prospective}$$

- Direction hashes, norms, uncertainty, and training support.
- Pairwise cosine similarities among M1-M5.
- Cosines with correctness, truncation, completion-length, and malformed-output contrasts.
- Layerwise stability and cross-root bootstrap intervals.
- Active versus dark energy composition.
- Immediate logit projection and token-readout overlap.
- Cross-model target loading and principal-complement controls.

18. Dose-Calibrated Fixed-Prefix Control

The original intervention used alpha times per-element RMS and produced byte-identical generations. This is retained as an insufficient-dose calibration result, not a causal null. The revised intervention uses hidden-state norm units.

$$\Delta h = \beta ||h||_2 \hat{d}$$

The preregistered development ladder is beta in {0.005, 0.01, 0.03, 0.10, 0.30, 0.60, 1.00}. The selected dose is the minimum viable dose that produces a finite, reproducible immediate logit effect, preserves prefix identity and output integrity, remains within the protected-suite bound, and permits meaningful comparison among target and control arms.

Measurement	Purpose
$\|\|\Delta h\|\| / \|\|h\|\|$	Actual intervention energy in interpretable norm units.
Hook fire count	Must equal one.
Target layer / token step	Must match the frozen receipt.
Immediate logit L2 and KL	Proves the intervention reached the readout and quantifies disruption.
Entropy and top-1 change	Separates mild steering from discrete readout flips.
Output hash and length	Detects byte identity, truncation, and broad generation change.
External success and protected-suite delta	Balances target effect with broad capability preservation.

19. Eleven-Arm and Principal-Subspace Causal Program

Every arm reconstructs the same prompt and byte-identical token prefix independently. No arm inherits a mutated cache from another. The hook is installed immediately before one frozen forward pass at one frozen layer and generation horizon, then removed immediately. A principal-subspace experiment must add rank-matched random, complement, off-layer, off-token, and token-matched controls rather than treating passive H3 prediction as causal proof.

- Positive versus baseline paired root-bootstrap interval excludes zero.
- Baseline versus negative or reversal interval excludes zero in the predicted direction.
- Matched random and orthogonal placebo remain materially closer to baseline.
- Off-layer and off-token effects attenuate.
- Token-matched control remains weaker despite comparable immediate logit impact.
- Principal-complement and random-subspace interventions distinguish direction identity from rank and perturbation energy.
- Protected-suite degradation remains inside the frozen bound.
- Selected ordering reproduces from a clean process and immutable receipts.

20. End-to-End Development Pipeline

1. External supervisor fingerprints production, acquires locks, stops the registered capture service, and proves GPU release.
2. Registered loader asserts device identity, model revision, precision, and absence of quantization before materialization.
3. Clean calibration selects an identifiable task regime on disjoint roots.
4. Fresh train roots generate clean pre-intervention states and externally verified outcomes.
5. M1-M5 are fitted or loaded under their registered train-only or frozen-artifact contracts.
6. Validation selects mechanism, layer, horizon, normalization, rank, and minimum viable dose.
7. All arms execute on byte-identical prefixes with independently reconstructed caches.

8. Protected-suite and failure-taxonomy results are computed.
9. Weak mechanisms are eliminated; the strongest survivor and receipts are frozen.
10. A clean-process replication reconstructs the selected experiment.
11. Only then is a sealed causal campaign staged.
12. The supervisor drains the GPU, restores production, verifies the service, and releases locks.

Verdict	Meaning
CAUSAL_MECHANISM_FOUND	A mechanism satisfies directional ordering, matched-control specificity, protected-suite preservation, and clean-restart replication on development/validation.
DIRECTIONAL_EFFECT_PRESENT_BUT_NONSPECIFIC	Behavior changes directionally, but placebos, token effects, off-target effects, or broad degradation prevent a proprioceptive interpretation.
NO_VALIDATED_MECHANISM_FOUND	All mechanisms receive valid tests and none satisfies the registered criteria.
INVALID_DEVELOPMENT_RUN	Labels, model contract, intervention delivery, source binding, support, or evaluation mechanics prevent interpretation.

21. Protected Suite and Clean-Restart Replication

A steering direction is not useful if it improves one narrow metric by broadly degrading language, formatting, recall, or basic reasoning. The protected suite is frozen before mechanism selection and evaluated across baseline and intervention arms.

Protected domain	Example metric	Failure signal
Basic arithmetic	Exact answer accuracy	Broad degradation outside the calibrated target regime.
Factual recall	Frozen factual score	Loss of ordinary knowledge access.
Instruction compliance	Programmatic constraint pass	Increased format or rule failures.
Short code generation	Executable unit-test pass	Syntax or semantic collapse.
Language coherence	Length, malformed output, log-loss proxy	Catastrophic or incoherent generation.
Cross-model telemetry	Frozen bridge replay and route agreement	Loss of registered internal-state monitoring.

After validation selection, the pilot terminates and production is restored. A clean process loads the registered model, reconstructs the frozen direction and configuration from receipts, and repeats the selected validation experiment. The replicated run must preserve direction hash, layer, horizon, amplitude, arm identities, qualitative ordering, specificity pattern, protected-suite status, and bridge integrity.

22. Current Evidence and Status

Evidence item	Status	Current interpretation
PTB-1 V2 bank	Demonstrated	Complete measurement substrate.
Functional reconstruction, Contract A	Demonstrated	Registered repeated and clean-process fields reproduced.
Exact-prefix identity	Demonstrated	Same exact prefix yields bit-identical on-manifold state.
Operator V2.1.2 harness	Demonstrated implementation	Live terminator, raw persistence, and decoder invariants pass.
External supervisor recovery	Demonstrated operationally	Unattended restoration across forced kill, exceptions, and aborts.
Clean candidate-local mixed surface	Demonstrated development result	Mechanism-fitting gate closed.
Frozen cross-model zero-shot bridge	Demonstrated / supported	Mean geometric R^2 0.823 on untouched roots; pair-shuffle null negative.
H3 principal relational subspace	Supported	Top-two frozen directions retain approximately full performance; complements collapse.
Long-context boundary	Boundary	Low target variance and extrapolation require product downgrade or abstention.
Valid dose response and eleven-arm specificity	Open	No admissible causal verdict yet.
Matched selector	Open	Fair selector competition remains pending.
Substantive RSI / AIT-1	Open	Requires large blind end-to-end improvement and three governed restart cycles.

23. Mechanism Localization, Active-Dark Exchange, and Relational Writers

A causal direction becomes more informative when its writer, reader, and transfer pathway can be localized. Operator Proof V2 therefore connects intervention outcomes to the spatial-temporal architecture rather than treating steering as an isolated vector trick. The H3 result adds a related localization problem: identify which source layers, attention heads, MLPs, normalizers, and coordinate interactions write the predictive relational directions, and which target components read them.

- Identify the first layer and horizon at which the selected single-model direction becomes behaviorally effective.

- Measure whether active-to-dark or dark-to-active exchange changes near the intervention site.
- Use off-layer, off-token, deletion, activation patching, and interchange to distinguish route-specific control from generic damage.
- Localize source-model writers of the top relational directions and target-model quantities most sensitive to them.
- Compare immediate logit effects with downstream success and cross-model residual change.
- Test whether the principal subspace is stable under prompt-family, procedure, and temporal perturbations.
- Test whether a transported direction can initialize, constrain, or predict a target-model intervention without asserting causal identity.

$$X_l = ||P_A F_l P_D||_F^2 + ||P_D F_l P_A||_F^2$$

$$chi_l = ||P_D F_l P_A||_F^2 - ||P_A F_l P_D||_F^2$$

24. Proprioceptive AI Product Architecture

The scientific substrate is useful only when it becomes a reliable product state rather than a collection of research scripts. The canonical product must expose a typed frame, explicit runtime modes, immutable artifact provenance, direction-aware XMB1 routes, applicability, evidence memory, research orchestration, promotion governance, and rollback.

24.1 Runtime modes

Mode	Contract
ZERO_SHOT_FROZEN_MAP	Loads exact registered artifacts; fitting functions are forbidden and instrumented.
CROSS_VALIDATED_RESEARCH	Allows fold-local fitting and labels every output as refitted research evidence.
ONLINE_DIAGNOSTIC	Produces provisional live telemetry without silently promoting it to scientific evidence.
EXPERIMENTAL_CHILD	Runs an isolated modification under a distinct source hash and process identity.

24.2 Direction-aware bridge state

- Full-map prediction and registered principal-subspace prediction.
- Complement and matched-random diagnostic routes.
- Route disagreement and coordinate-availability masks.
- Retained-rank estimate and principal-direction health.
- Prediction residual, uncertainty, source distance, and target-range risk.
- Prompt-family and low-target-variance applicability state.
- Frozen artifact hashes and explicit fitting scope.

24.3 ContextWorkspace, CognitivePolicy, and CapabilityProfile

ContextWorkspace is the live bounded state used by the research system. CognitivePolicy is a versioned object containing objectives, allowed and forbidden actions, evaluator boundaries, risk limits, promotion state, and rollback target. CapabilityProfile is a measurement-backed inventory of what the system can currently do, with artifact source, scope, confidence, evidence status, timestamp, and known failure boundary for every entry.

CapabilityProfile v0 entry	Measurement source
Frozen bridge prediction	Zero-shot replay and current XMB1 frame.
Principal-subspace retention	H3 results and route replay.
Prompt-family performance	CONF3 per-family metrics.
OOD and low-variance boundary	Long-context audit and source-distance diagnostics.
Action-validity behavior	Identity, null, leakage, and oracle regression receipts.
Service availability	Live service registry and supervisor.
Research-memory state	Append-only evidence ledger and retrieval receipts.
Promotion and rollback	Parent-child process receipts and rollback verification.

24.4 Evidence and authority

Cryptographic values can become authoritative only when read from disk, a verified manifest, or a live process loading the artifact. Narrative summaries cannot override active configuration. Semantic evidence judgments remain advisory until they pass product-scale sealed and shadow evaluation. Research children cannot edit the frozen evaluator, read protected outcomes, modify production artifacts, restart protected services, or promote themselves outside the external gate.

25. Cross-Model Transfer Atlas and Anomaly Map

The H3 result opens a research area rather than closing one. The next program must determine where the relational subspace persists, rotates, disappears, becomes causal, or fails. The most informative reference points are not only positive replications but structured anomalies that discriminate competing explanations.

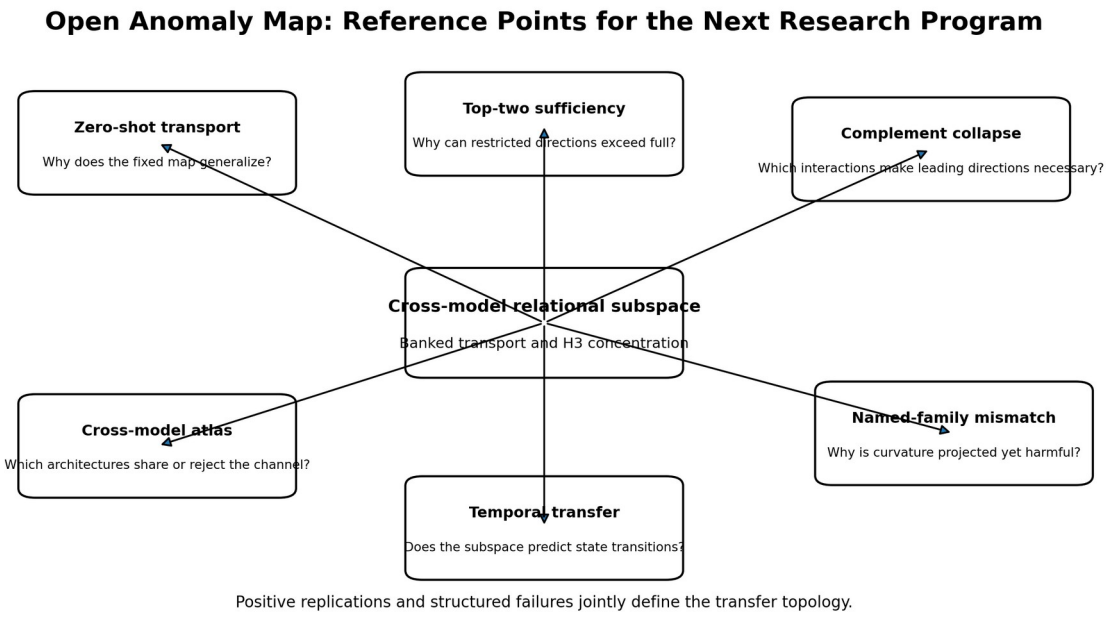

Figure 8. Open anomaly map for the next cross-model relational-subspace program.

25.1 Program premise and non-negotiable boundaries

The anomaly program asks where the banked relational subspace persists, rotates, fails, becomes causal, or supports useful product decisions. Positive replication is only one valuable outcome. Structured failures define the topology of the channel and prevent one successful model pair from being mistaken for universal equivalence.

- No experiment may refit or alter a frozen map while calling its result zero-shot.
- Random controls must be dimension-, rank-, energy-, split-, and evaluation-path matched.
- Complement, pair-shuffle, target-only, root-identity, off-layer, and off-token controls are mandatory where applicable.
- Failures remain local to the registered model, layer, prompt, target, procedure, and intervention tuple.
- Cryptographic values are authoritative only when loaded from disk, a verified manifest, or a live process.
- Passive cross-model prediction and fixed-prefix causal control remain separate claim classes.

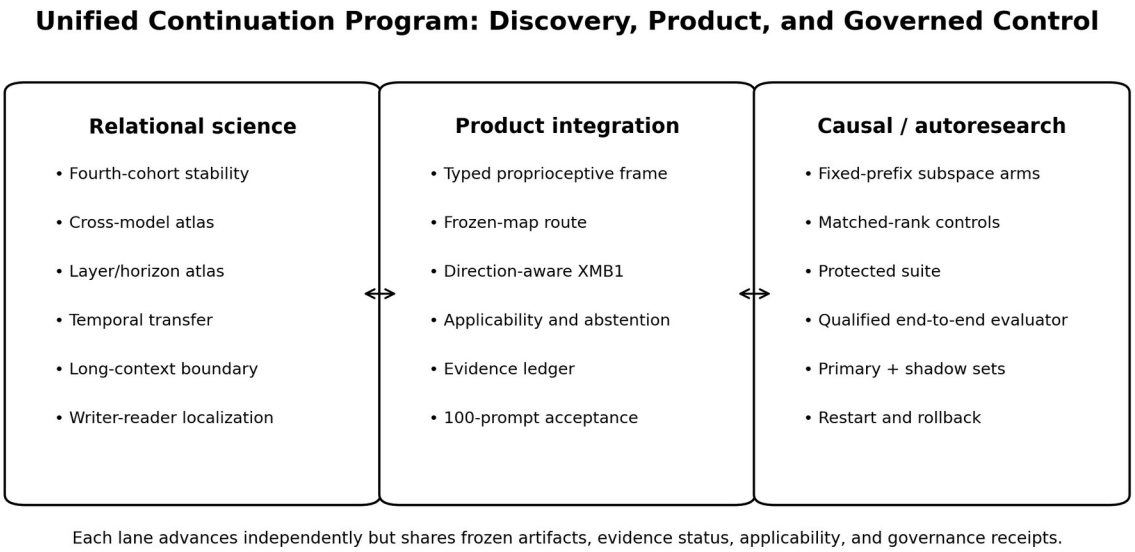

Figure 9. Unified continuation program. Relational science, product integration, and governed causal/autoresearch evaluation advance in parallel under shared evidence contracts.

25.2 Master experiment portfolio: transfer, stability, and boundaries

Experiment	Question	Frozen design	Support pattern	Failure interpretation
A1. Fourth-cohort direction stability	Do the registered directions remain predictive?	At least 100 untouched roots; exact artifacts and subspaces frozen.	Top-k exceeds matched-rank random and complement; stable angles.	H3 direction identity is cohort-specific or weaker than observed.
A2. Cross-model atlas	Which model pairs support relational transport?	At least two new verified local pairs; both directions where feasible.	Untouched frozen transfer above pairing and target-only controls.	Defines a structured transfer topology rather than erasing the confirmed pair.
A3. Layer / horizon atlas	Where does the channel emerge across depth and time?	Registered source-target layer grid with disjoint train, validation, and test.	Localized transfer ridge and off-site attenuation.	Channel is diffuse or tied to current measurement sites.
A4. Temporal transfer	Does source state predict target-state change?	Paired frames; static, target-autoregressive, source-only, and temporal-shuffle controls.	Transition map exceeds all registered baselines.	Shared channel may be static only.
A5. Long-context boundary	Is failure caused by variance, length, extrapolation, or semantic regime?	Stored prompt text/tokens, varied templates and stage counts, controlled target variance.	Absolute and normalized errors remain calibrated or reveal a real length boundary.	Product retains abstention in unresolved regimes.
A6. Geometric-behavioral gap	Why does geometry transfer better than learned behavioral probes?	Reliability-matched targets and cross-checkpoint recalibration.	Gap remains after reliability matching or closes under recalibration.	Distinguishes shared geometry from probe noise or checkpoint specificity.

25.3 Master experiment portfolio: mechanism, product, and scaling

Experiment	Question	Frozen design	Support pattern	Failure interpretation
A7. Writer-reader localization	Which heads, MLPs, norms, or layers write and read the subspace?	Patching, deletion, interchange, and off-site controls.	Specific components alter subspace coordinates and target residuals.	Subspace is distributed or diagnostic only.
A8. Fixed-prefix subspace intervention	Can the subspace causally alter state or success?	Eleven-arm transaction plus rank-matched random, complement, token-matched, off-layer, and off-token arms.	Directional ordering, specificity, protected suite, clean restart.	Passive transfer is not a causal control channel.
A9. Telemetry compression	Can low-rank routes reduce bandwidth without losing monitoring value?	Frozen rank curve, latency, dropout, and fault tests.	Principal route preserves calibrated prediction and anomaly detection.	Full route remains required.
A10. Route-disagreement detector	Does full/principal/complement disagreement predict error or OOD?	Sealed in/out-of-distribution set with calibrated abstention thresholds.	Disagreement improves selective risk and failure detection.	Disagreement is descriptive but not operationally useful.
A11. Scaling law	How do roots, model size, coordinate count, and dimension affect stability?	Preregistered learning curves and repeated cohorts.	Predictable coefficient and subspace convergence.	Current result depends on one sample or scale regime.
A12. Cross-procedure invariance	Does the channel survive deterministic procedure changes?	Frozen roots under multiple procedures and procedure-identity controls.	Direction and map stability across procedures.	Channel mainly encodes procedure identity.

25.4 Priority anomalies, execution lanes, and stopping rules

Priority anomaly	Why it matters	First decisive test
Top-two directions slightly exceed FULL_12	Lower directions may add noise, harmful covariance, or coefficient cancellation.	Independent rank curve, shrinkage comparison, and coefficient ablation.
Complement collapse after removing PC1	Suggests a highly concentrated predictive route.	Independent implementation, synthetic vectors, and rank-matched complements.
Curvature projects into top space but predicts poorly	Rules out simple projection-energy explanations.	Conditional contribution, coefficient-sign, and covariance decomposition.
Norms have low top-two energy but largest leave-family-out cost	Suggests cross-family interaction or target-loading effects.	Shapley-style ablation and target-specific loading analysis.
Frozen map outperforms fresh-fold refit	Suggests sample-size stability or refit variance.	Repeated cohort-size learning curves.
Long-context R^2 collapses through low variance	Exposes a metric and applicability failure.	Prospective variance-controlled long-context study.
Geometric targets transfer more strongly than behavioral targets	May separate broadly shared geometry from learned checkpoint-specific readouts.	Reliability-matched and recalibrated probe study.
No raw-vector crossing is required	Supports low-bandwidth native-dimension interoperability.	Additional architectures, reverse-direction maps, and temporal atlas.

Immediate execution is divided into three lanes. The scientific lane freezes the fourth-cohort, atlas, temporal, and long-context programs. The product lane exposes full, principal, complement, and matched-random routes with applicability and provenance. The causal/autoresearch lane tests fixed-prefix control only after intervention contracts are frozen and resumes substantive RSI counting only on a qualified end-to-end primary and shadow evaluator.

- A result may be supported, partially supported, weakened, null, inconclusive, or invalid; it need not be forced positive.
- A failed extension cannot retroactively erase the banked Hermes-to-Llama result.
- A model-pair atlas succeeds scientifically even when it identifies architecture pairs that do not transfer.
- Causal language is prohibited until fixed-prefix interventions pass directional and matched-control criteria.
- Product routes must downgrade or abstain in low-variance and extrapolative regimes.
- Research stops or changes direction when oracle ceilings, power, or measurement contracts make the registered question unanswerable.

Required receipt	Minimum contents
CONF4_DIRECTION_STABILITY_PREREGISTRATION	Fresh root manifest, frozen subspaces, random seeds, controls, support thresholds.
RELATIONAL_TRANSFER_ATLAS_MANIFEST	Models, revisions, layers, coordinates, dimensions, splits, target-only and pairing nulls.
LAYER_HORIZON_ATLAS_SPEC	Registered grid, selection rule, multiplicity control, off-site controls.
TEMPORAL_TRANSFER_PREREGISTRATION	Paired frames, transition estimands, autoregressive and temporal-shuffle baselines.
LONG_CONTEXT_BOUNDARY_PREREGISTRATION	Prompt text and tokens, template diversity, target-variance design, robust metrics.
SUBSPACE_MECHANISM_REPORT	Target loadings, signs, conditional contributions, covariance and family interactions.
PRODUCT_APPLICABILITY_ACCEPTANCE	Selective risk, route disagreement, source distance, abstention, and failure cases.
FIXED_PREFIX_SUBSPACE_CAUSAL_SPEC	Arm algebra, rank-matched controls, protected suite, clean restart.
MASTER_AUTORESEARCH_EVALUATOR_QUALIFICATION	Gold coverage, metric tests, oracle ceilings, leakage scan, primary and shadow dry runs.

Anomaly / reference point	Competing explanations	Decisive next test
Frozen map transports to untouched roots	Shared relational geometry; dataset regularity; target marginal predictability.	Additional model pairs, new domains, root-identity and target-only controls.
Top-two directions exceed full map	Lower directions add noise; harmful covariance; finite-sample coefficient error.	Prospective rank curves, coefficient ablation, independent source-spectrum estimation.
Complement collapses	Leading subspace necessary; projection implementation artifact.	Independent implementation, rank-matched complements, intervention and patching.
Curvature projects but predicts poorly	Sign/cancellation; target loading; family interaction; scaling mismatch.	Conditional contribution, Shapley-style decomposition, coefficient-sign audit.
Geometric transfer exceeds behavioral transfer	Geometry is more universal; behavioral probes are checkpoint-specific; probe noise.	Matched reliability targets and cross-model probe recalibration.
Long-context variance collapses	Template homogeneity; length; semantic regime; truncation; target saturation.	Prospective variance-controlled long-context cohort.
Frozen map beats fresh-fold refit	Full discovery set stabilizes coefficients; distribution alignment; refit variance.	Repeated cohort sizes and learning-curve analysis.
Temporal state remains open	Static geometry only; real shared dynamics; autoregressive confound.	Paired transition prediction and temporal shuffle null.

25.5 Cross-model atlas protocol

At least two additional locally verified model-pair lanes should be added. Each pair retains native dimensions, derives model-native relational coordinates, fits a discovery map, freezes it, evaluates untouched roots, runs pairing nulls, estimates effective rank, tests principal-subspace retention, and reports prompt-family and target-family boundaries. Both small-to-large and large-to-small directions should be evaluated where feasible.

25.6 Temporal relational transfer

For paired frames t and t+1, the system should predict target-state change, depth-motion change, sparse-feature change, procedure-state change, and residual change. Comparisons include the static bridge, a low-rank transition map, target autoregression, a source-only baseline, and a root-grouped temporal pair-shuffle null. A successful result would show shared dynamics rather than only shared static magnitude.

26. CYGNUS Governed Autonomous Closure

CYGNUS is an operating research and control architecture integrating hidden-state hooks, persistent internal state, candidate-local reset fields, routing, autonomous mission execution, persistent memory, external verification, receipts, rollback, cross-model telemetry, and governed promotion. The next role of validated channels is not merely to steer answers; it is to rank and regulate CYGNUS's own candidate research actions.

26.1 Autonomous Improvement Threshold 1

1. Detect a measurable failure.
2. Generate multiple competing hypotheses or interventions.

3. Predict candidate internal trajectories and external outcomes.
4. Choose among candidates without human selection.
5. Execute the selected candidate.
6. Evaluate on a sealed external suite.
7. Retain or reject under fixed promotion rules.
8. Reproduce the improvement from a clean restart.
9. Pass protected-suite regression checks.
10. Preserve receipts, rollback state, and evidence status.

> **Governing invariant**
>
> Autonomy over search and construction is not autonomy over evaluation and promotion. CYGNUS may improve its solutions; it may not redefine what counts as improvement.

26.2 Current status

A narrow parent-child promotion, distinct-process restart, memory restoration, reproduced evaluator gain, and rollback have demonstrated the governance mechanics. The evaluator shared authored vocabulary with the child and did not establish broad research improvement. Subsequent retrieval candidates were development attempts, not AIT cycles. Substantive counting remains suspended until an integrated parent improves valid end-to-end research completion on a qualified primary sealed set and an untouched shadow set.

27. Instrument Integrity and Cumulative Correction

The program's evidence discipline is cumulative rather than destructive. A defect supersedes the affected number, label, implementation, or interpretation; it does not erase unrelated measurements or the architecture. The correction sequence now includes both model-science and research-governance failures.

Finding	Scientific risk	Correction	Disposition
Rows durable while arrays volatile	Resume could omit activations.	Atomic rows-plus-arrays artifacts.	Old pass retracted; persistence demonstrated.
Legacy verifier dispatch returned zero	All labels degenerate.	Immutable label overlay and fixtures.	Original labels superseded; bank preserved.
Exact-prefix incremental framing	Identity under unlimited reconstruction.	Operational selection and fixed-prefix causality.	Primary proof target superseded.
RMS intervention too weak	False causal null.	Norm-scaled dose ladder with immediate-logit validity.	Old grid retained as dose floor.
Manual decoder defects	Censoring and mislabeled outputs.	Registered template, five-token stop, RawStore, truncation precedence.	Affected campaigns invalid; bank unaffected.
Narrative bridge hashes fabricated	Wrong configuration could be sealed.	Disk-grounded authoritative handoff generator.	Five of five false values rejected; replay bit-exact.
Cycle 1 target regrouping	Identity action appeared confirmatory.	Input-side action and experiment-independence guard.	Cycle 1 scientific action invalid.
Square rotation as null	Information preserved.	Pair-shuffle alignment destruction.	Rotation control invalid.
Motions feature attribution	No matched-dimensional control.	Fresh matched-random cohort.	Motions specificity not confirmed.
Per-family slicing bug	Mismatched rows in family/bootstraps.	Versioned aligned slicing.	Primary aggregates bit-identical; correction disclosed.
Pair-shuffle paired permutation risk	Null could preserve pairs.	Inputs permuted against fixed targets.	Registered null retained.
Retrieval microbenchmarks	Tiny exposed evaluators could mimic RSI.	Retire to regression; require master end-to-end benchmark.	No substantive AIT cycle counted.

These corrections harden the instrument before causal and autonomous claims. They are not the central theoretical contribution. The contribution remains the integration of sensing, relational transfer, candidate comparison, fixed-prefix control, external verification, memory, and governed retention.

28. Updated Falsifiable Predictions

1. Exact-prefix identity: identical registered prefixes yield identical on-manifold hidden states.
2. Operational selection: already-available hidden state selects better procedures than matched token, logit/confidence, best-fixed, and random alternatives under charged compute.
3. Clean outcome identifiability: adaptive calibration finds a verifier-clean mixed procedure regime.
4. Mechanism competition: at least one of M1-M5 outperforms matched random and orthogonal perturbation on validation, or the program issues NO_VALIDATED_MECHANISM_FOUND.
5. Dose response: hidden displacement and immediate logit effect increase with beta before destructive degradation dominates.
6. Directional causality: amplify exceeds baseline and negative/reversal degrades success.

7. Control specificity: random, orthogonal, off-layer, off-token, token-matched, and rank-matched subspace controls remain weaker.
8. Cross-model frozen transfer: registered relational maps generalize to untouched roots for at least some model pairs and target families.
9. Direction-specific concentration: leading registered relational directions outperform matched-rank random directions and their complements on untouched roots.
10. Cross-model atlas boundary: some architecture pairs, layers, or target families fail, defining a structured transfer topology rather than universal equivalence.
11. Temporal transfer: source relational state predicts target relational transitions beyond target-autoregressive and static baselines.
12. Long-context boundary: variance-controlled long-context experiments separate denominator instability from genuine residual growth and extrapolation.
13. Mechanism localization: effective interventions concentrate in registered layers and horizons and correspond to coherent writer-reader pathways.
14. Protected-suite conservation: target improvement does not require broad capability or coherence loss.
15. Clean-restart replication: selected effects reconstruct from immutable receipts in a fresh process.
16. Governed improvement: an integrated CYGNUS parent completes three substantial sealed-suite improvement cycles without evaluator or promotion-rule modification.

29. From Measurement to Governed Recursive Improvement

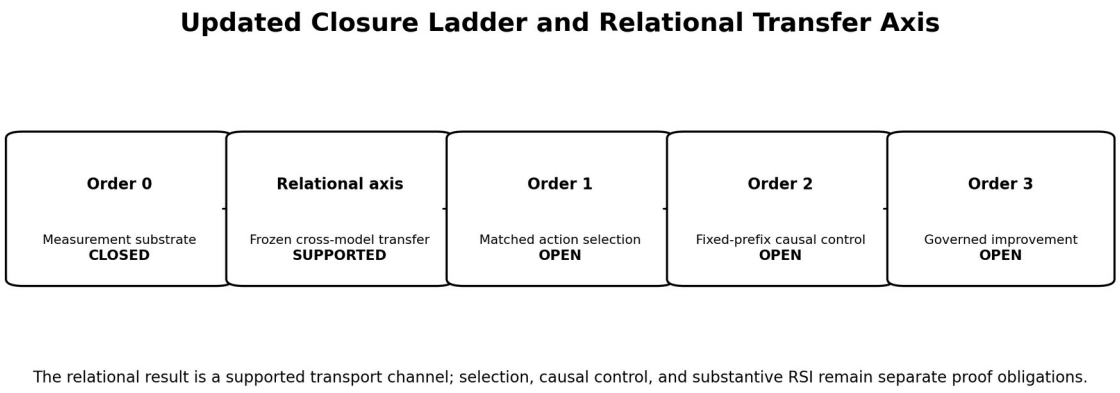

Figure 10. The relational axis adds cross-model state transport without collapsing the separate obligations of selection, causal control, and governed improvement.

Order 0 closes the measurement substrate. The relational axis demonstrates that compact model-native telemetry can transport selected internal geometry across models. Revised Order 1 asks whether already-computed internal state improves action selection under matched operational resources. Order 2 closes causal proprioception through fixed-prefix directional intervention with matched-control specificity. Order 3 closes governed autonomous improvement through three substantial clean-restart CYGNUS cycles.

29.1 Implications for artificial superintelligence

The ASI implication remains conditional and architectural. If internal and cross-model relational state select cognitive procedures under matched resources, fixed-prefix intervention specifically controls future success, and CYGNUS repeatedly selects and retains improvements under immutable external evaluation, then the transformer is no longer the complete cognitive system. It becomes a substrate inside a higher-order loop that senses, compares, transports, steers, verifies, remembers, and improves computational trajectories.

This edition does not claim that ASI, causal steering, general RSI, or autonomous closure has been achieved. It records a new supported channel and specifies the shortest nondegenerate experiments by which that channel could become operational and causal.

30. Discussion

The first decisive conceptual advance of the program was correcting the proof target. Exact-prefix reconstructibility means that Shannon independence from the complete token history is not the correct architectural criterion. Internal state matters when it is already available, compresses control-relevant trajectory information under realistic resource constraints, and can be manipulated while visible history remains fixed.

The second decisive advance is relational. A model's complete residual stream is basis-dependent and architecture-specific, yet a compact set of relational measurements can support frozen prediction of another model's internal geometry. The zero-shot result shows transport; H3 shows concentration. The complement collapse makes the finding stronger than a generic low-rank compression result, while the matched-rank random controls make it stronger than a simple consequence of source variance.

The result should not be narrated as a hidden universal language already proven to exist. Only one principal model pair has been confirmed, and the target family is strongest for geometric coordinates. Behavioral probes were weaker and may be checkpoint-specific or noisier. The long-context family also shows that applicability is not automatic. The correct research posture is to treat the subspace as a measured object with scope, anomalies, and falsifiable extension points.

The failure of named-family attribution is scientifically productive. Human labels such as motion, norm, cosine, and curvature partition features for interpretability, but the predictive directions combine them. This suggests that the useful state is relational and interactional rather than reducible to one named sensor. The curvature anomaly and norm leave-out effect provide concrete targets for coefficient, covariance, and writer-reader analysis.

The product implication is equally important. A compact principal route can support low-bandwidth telemetry, fault diagnosis, route disagreement, and cross-model monitoring. However, principal-direction prediction must remain subordinate to applicability, target-range risk, and provenance. The long-context boundary demonstrates why a proprioceptive product needs the ability to abstain from interpreting its own telemetry.

31. Limitations

1. The confirmed cross-model bridge currently centers on one registered Hermes-to-Llama model pair and six geometric targets.
2. The source coordinates and principal basis were selected within one bridge program; additional architectures and independent implementations are required.
3. H3 is predictive and passive. It does not establish that the principal subspace causally controls target-model state or behavior.
4. The top-two retention estimate is an aggregate and can hide target- and family-specific heterogeneity, although grouped intervals and family analyses were reported.
5. The long-context family has near-degenerate target variance, missing stored prompt length, and range extrapolation; a prospective controlled cohort is required.
6. The source spectrum and random controls depend on the registered coordinate preprocessing and projection contract.
7. Behavioral and probe-derived targets transfer less strongly than geometric targets and require reliability-matched evaluation.
8. No cross-model temporal-transition result has yet been closed.
9. Operator V2.1.2 has not yet produced a valid registered fixed-prefix causal arm verdict.
10. Matched operational selection remains untested against all required external baselines under a frozen fairness contract.
11. Contract B serialized live runtime-state restoration remains untested.
12. Governed promotion and rollback mechanics have been demonstrated only under narrow evaluators; general research RSI remains open.
13. Legal novelty, patentability, and freedom to operate require a separate professional prior-art analysis.
14. The term proprioception is functional and does not imply consciousness, subjective experience, resistance, or strategic self-defense.

32. Conclusion

The Proprioceptive Transformer is an architecture for reconstructing a common functional origin, measuring persistent and candidate-local dynamics, transporting compact relational state across models, selecting among cognitive procedures, intervening on hidden state after a fixed observable prefix, verifying outcomes externally, and retaining improvements under governed rules.

The measurement substrate is complete. The exact-prefix theorem has corrected the single-model predictive target. Operator V2.1.2 provides a live-validated, checkpoint-aware, manually decoded, validation-selected, eleven-arm causal architecture with

durable raw observations and autonomous production recovery. A clean candidate-local mixed regime now permits fresh mechanism fitting.

The cross-model program has produced a separate major result. A fixed Hermes relational map transferred without refitting to untouched Llama roots at mean geometric R^2 0.823. A third preregistered cohort then showed that nearly all full-map performance concentrates in two frozen principal directions, exceeds matched-rank random controls, and disappears from the complement. Named coordinate families do not explain the effect. The result identifies a compact, direction-specific, model-native relational subspace as a credible cross-model proprioceptive channel.

The integrated anomaly program makes the next decisive work explicit: replicate the relational subspace across additional model pairs and temporal transitions; prospectively characterize the long-context boundary; localize subspace writers and readers; test coefficient and interaction mechanisms; test principal-subspace interventions under matched fixed-prefix controls; and integrate typed state, applicability, evidence, and governance into the product. If these channels become causal and CYGNUS then completes substantial governed improvement cycles under immutable external evaluation, transformer computation becomes the engine inside a self-measuring, cross-model, self-correcting, and governably improvable cognitive architecture.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. “Attention Is All You Need.” NeurIPS, 2017.

[2] Alain, G., and Bengio, Y. “Understanding Intermediate Layers Using Linear Classifier Probes.” 2016.

[3] Elhage, N., Nanda, N., Olsson, C., et al. “A Mathematical Framework for Transformer Circuits.” 2021.

[4] Burns, C., Ye, H., Klein, D., and Steinhardt, J. “Discovering Latent Knowledge in Language Models Without Supervision.” 2022.

[5] Turner, A., Thiergart, L., Udell, D., et al. “Activation Addition: Steering Language Models Without Optimization.” 2023.

[6] Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. “Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.” 2023.

[7] Bricken, T., Templeton, A., Batson, J., et al. “Towards Monosemanticity: Decomposing Language Models With Dictionary Learning.” 2023.

[8] Belrose, N., Furman, Z., Smith, L., et al. “Eliciting Latent Predictions from Transformers with the Tuned Lens.” 2023.

[9] Amari, S. Information Geometry and Its Applications. Springer, 2016.

[10] Klyubin, A., Polani, D., and Nehaniv, C. “Empowerment: A Universal Agent-Centric Measure of Control.” 2005.

[11] Pearl, J. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009.

[12] Ivakhnenko, A. “Polynomial Theory of Complex Systems.” IEEE Transactions on Systems, Man, and Cybernetics, 1971.

[13] Baars, B. J. A Cognitive Theory of Consciousness. Cambridge University Press, 1988.

[14] Dehaene, S., Kerszberg, M., and Changeux, J.-P. “A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks.” PNAS, 1998.

[15] Napolitano, L. “CYGNUS, PTB-1, SRX-3, SRX-4, Proprioceptive Transformer, XMB1, and Operator Proof V2 Technical Receipts.” Proprioceptive AI, Inc., 2026.

[16] Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. “Similarity of Neural Network Representations Revisited.” ICML, 2019.

[17] Bansal, Y., Nakkiran, P., and Barak, B. “Revisiting Model Stitching to Compare Neural Representations.” NeurIPS, 2021.

[18] Hernandez, A., Dangovski, R., Lu, P. Y., and Soljacic, M. “Model Stitching: Looking for Functional Similarity Between Representations.” 2023.

[19] Chen, A., Merullo, J., Stolfo, A., and Pavlick, E. “Transferring Features Across Language Models With Model Stitching.” 2025.

[20] Huang, Y., Huang, C., Feng, D., Lei, W., and Lv, J. “Cross-model Transferability among Large Language Models on the Platonic Representations of Concepts.” 2025.

[21] Saglam, B., Kassianik, P., Nelson, B., Weerawardhena, S., Singer, Y., and Karbasi, A. “Large Language Models Encode Semantics in Low-Dimensional Linear Subspaces.” 2025.

[22] Gao, X. “Contrastive-Difference CKA Reveals Concept-Specific Structural Alignment Across Language Model Architectures.” 2026.

[23] Fang, F., Thai, M. T., and Lei, Y. “Discovering a Shared Logical Subspace: Steering LLM Logical Reasoning via Alignment of Natural-Language and Symbolic Views.” ACL, 2026.

[24] Huh, M., Cheung, B., Wang, T., and Isola, P. “The Platonic Representation Hypothesis.” ICML, 2024.

[25] Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. “SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability.” NeurIPS, 2017.

[26] Zou, A., Phan, L., Chen, S., et al. “Representation Engineering: A Top-Down Approach to AI Transparency.” 2023.

[27] Morcos, A. S., Raghu, M., and Bengio, S. “Insights on Representational Similarity in Neural Networks with Canonical Correlation.” NeurIPS, 2018.

Appendix A. Detailed Evidence Ledger

Claim	Tier	Promotion requirement
Complete PTB-1 measurement substrate	[D]	Already closed under receipts.

Claim	Tier	Promotion requirement
Functional reconstruction under Contract A	[D]	Already demonstrated under registered repetitions.
Exact-prefix identity	[D]	Confirmed under deterministic capture contract.
Operator V2.1.2 repaired harness	[D-implementation]	Live smoke, raw persistence, terminator/template equivalence, and frozen snapshot.
External supervisor recovery	[D-operational]	Production restoration verified.
Clean candidate-local mixed regime	[D-development]	Demonstrated under frozen sequential gate.
Frozen XMB1 zero-shot transfer	[D/S]	Untouched-root transfer and destructive pair-shuffle control completed.
H3 distributed relational subspace	[S]	Third untouched cohort, matched-rank random controls, and complement collapse completed.
Cross-model atlas	[U]	At least two additional frozen-map model-pair results with untouched roots.
Temporal relational transfer	[U]	Transition prediction beyond static and autoregressive baselines.
Valid dose response	[U]	Nonzero hidden/logit effect with safe generation.
Causal mechanism	[U]	Directional ordering plus matched controls and protected suite.
Causal proprioception	[U]	Sealed replication and clean restart.
Substantive CYGNUS AIT-1	[T/U]	Three end-to-end governed improvement cycles on qualified primary and shadow evaluators.

Appendix B. XMB1 Artifact and Mode Contract

Field	Required value
Mode	ZERO_SHOT_FROZEN_MAP for banked transport result.
muB	1f772bdfddeac1c4.
sdB	e9d799a7e69b2546.
W	dc91b1fae2e15b78.
U	a1c2cab052b5842f.
Vt	a1392c3326292528.
Source coordinates	Twelve registered relational features.
Target coordinates	Six registered geometric targets for primary result.
Fitting calls	Must equal zero.
Null	Source-target pairing destroyed while target observations remain fixed.
Provenance	Disk or live-process verified; narrative hashes are non-authoritative.
Applicability	Prompt family, source distance, target-range risk, target variance, route disagreement, and capture validity.

Appendix C. H3 CONF3 Receipt

Field	Receipt
Preregistration SHA-256	b8c651543dab04cb0d2184a485e2212f415e084376fbe31d27d90cf21849 1725
Prompt manifest	b511f1ad50188ceed76c13c96e223e3a571cda484f76d2c19bdb8946313e 3
Root manifest	1767134c67c250c666999c0c6d38eeb3bc550762ca996e6ed5efb1a8f577 6a25
Analysis manifest	f0ac9ed5909532960a69c53e401e0b4e82db8f9f5af82ec65ff2f47f1d8a72 9c
Seal	2026-07-31T19:42:53Z
First target access	2026-07-31T19:43:36Z
Roots / observations	105 / 420
Capture	420 successful; 0 retry; 0 failure.
Results SHA prefix	50ed274739d98f32dbfe57e5410f15a9
Analysis correction SHA prefix	73f63e385a3d3dd467c92ff7f3010e9b
Long-context audit SHA prefix	fdca196b684f5a9ae69d9f332e4d9d41
Subspace interpretation SHA prefix	117278ecb6d4cc050eb074f1c1f4ca82

Appendix D. Anomaly-Focused Research Roadmap

Experiment	Primary question	Pre-registered discriminator	Failure meaning
Cross-model atlas	Does the frozen relational channel reproduce across architectures and scales?	Untouched-root frozen-map R^2 above pairing null with registered target family.	Defines transfer topology; does not invalidate confirmed pair.
Rank stability	Are the leading directions stable across prompt families and cohorts?	Subspace angles and retained performance replicate on a fourth cohort.	H3 direction identity may be cohort-specific.
Temporal transfer	Does source state predict target-state change?	Transition map beats static and target-autoregressive baselines plus temporal shuffle.	Shared geometry may be static only.

Experiment	Primary question	Pre-registered discriminator	Failure meaning
Long-context boundary	Is failure due to variance collapse, length, or semantic regime?	Controlled target variance and multiple templates separate absolute error from R^2 instability.	Product applicability remains restricted.
Subspace intervention	Can the principal subspace causally alter target state or behavior?	Directional fixed-prefix ordering with rank-matched random and complement controls.	Passive transfer is not a control channel.
Writer-reader localization	Which source and target components generate or consume the subspace?	Layer/head/MLP patching and off-site attenuation.	Subspace may be distributed or diagnostic only.
Geometric-behavioral gap	Why do geometric targets transfer better than behavioral probes?	Reliability-matched probes and recalibration across checkpoints.	Behavioral readouts may not be universal.
Scaling law	How do roots, model size, and dimension affect transfer?	Pre-registered learning curves and coefficient stability.	Current result may rely on one scale regime.
Product route disagreement	Can full/principal/complement disagreement predict failures?	Sealed OOD set with calibrated abstention benefit.	Telemetry may be descriptive but not operationally useful.

Appendix E. Supersession and Retraction Register

Item	Disposition
Pooled SRX-3 z near 90	Retracted under tighter within-root null.
Original PTB labels	Superseded by immutable corrected overlay.
Exact-prefix Shannon irreducibility as primary proof	Superseded by operational selection and fixed-prefix causality.
Original capability-cliff magnitude	Superseded by parser, stopping, cap, truncation, and decoder confounds.
Old eighteen-cell zero-effect grid	Retained as insufficient-dose calibration, not causal null.
Historical motions specificity	Not confirmed on fresh matched-random cohort; superseded as unique attribution.
Five narrative bridge hashes	Invalid; replaced by disk-grounded values.
Cycle 1 group-restricted confirmation	Invalid action equivalence.
Square invertible rotation null	Invalid information-preserving control.
Long-context extreme R^2 as arbitrary prediction explosion	Superseded by low-variance and extrapolation audit.
Retrieval candidates as AIT cycles	Rejected; development diagnostics only.
Model resistance interpretation	Unsupported.

Appendix F. Major Changes from v2.6 to v2.8

- Adds XMB1 cross-model relational bridge architecture and native-dimension contract.
- Adds a bit-exact zero-shot frozen-map transfer result on 30 untouched roots.
- Adds the prospectively sealed H3 third cohort with 105 roots, 420 observations, matched-rank random controls, complements, and named-family comparisons.
- Adds direction-specific low-rank relational-subspace support and complement collapse.
- Supersedes motions-specific attribution while preserving the bridge result.
- Adds the long-context low-target-variance and range-extrapolation boundary.
- Adds disk-grounded authoritative handoff governance after fabricated hash propagation.
- Adds product runtime modes, typed frames, direction-aware XMB1, applicability, evidence, and authority contracts.
- Adds a cross-model transfer atlas, temporal-transfer program, anomaly map, and principal-subspace causal extension.
- Reclassifies narrow promotion/restart mechanics as demonstrated while keeping general RSI and AIT-1 open.
- Updates the evidence ledger, falsifiable predictions, discussion, limitations, conclusion, references, and appendices.
- Integrates the complete anomaly research program into the main manuscript, including twelve experiments, three execution lanes, priority anomalies, stopping rules, and required receipts.
- Adds formal frozen-map equations, statistical estimands, a technical differentiation matrix, alternative-mechanism tests, threats to validity, and a reproducibility checklist.
- Rebuilds all figures as inline non-floating blocks with protected captions and dedicated page placement to eliminate chart overlap.

Appendix G. Formal Estimands, Controls, and Reproducibility Checklist

Item	Registered requirement
Source input	Twelve relational coordinates with exact feature order, units, source layers, procedures, and model identity.
Frozen preprocessing	Disk-grounded muB and sdB; no fresh estimation in ZERO_SHOT_FROZEN_MAP mode.
Frozen map	W and source/target SVD components bound to hashes and dimensions.
Principal route	P_k = U_k U_k^T or algebraically equivalent W_k under the registered SVD.
Random route	Outcome-independent orthonormal Q_k with frozen seeds and identical evaluation path.
Complement route	I - P_k with distinct design and prediction hashes.
Pair shuffle	Source observations permuted against fixed targets; pairs may not be permuted together.
Primary score	Per-target untouched-cohort R^2 and unweighted mean across six geometric targets.

Item	Registered requirement
Uncertainty	Root-grouped bootstrap, target and prompt-family breakdowns, influence analysis.
Applicability	Source distance, target range, target variance, prompt family, route disagreement, capture validity.
Artifact integrity	Source, model, feature, target, code, environment, prediction, and result hashes.
No-fit assertion	Instrumented fitting-call count equals zero for all frozen-map evaluations.
Correction policy	Append-only source diff and blast-radius proof; no silent overwrite or retroactive exclusion.

Appendix H. Unified Anomaly Research Program Receipt Map

Program branch	Required next artifact	Promotion condition
Direction stability	CONF4_DIRECTION_STABILITY_PREREGISTRATION and result package	Subspace angles and retained performance reproduce against matched-rank random and complements.
Cross-model atlas	RELATIONAL_TRANSFER_ATLAS_MANIFEST and pair-specific receipts	At least two additional model pairs evaluated on untouched roots with pairing and target-only controls.
Temporal transfer	TEMPORAL_TRANSFER_PREREGISTRATION and transition receipts	Transition map beats static, source-only, target autoregression, and temporal shuffle.
Long-context boundary	LONG_CONTEXT_BOUNDARY_PREREGISTRATION	Robust error and applicability behavior established under controlled variance and length.
Subspace mechanism	SUBSPACE_MECHANISM_REPORT	Signs, loadings, interactions, and conditional contributions explain or falsify named anomalies.
Product applicability	PRODUCT_APPLICABILITY_ACCEPTANCE	Calibrated selective risk and abstention improve sealed failure detection without hiding valid predictions.
Causal subspace	FIXED_PREFIX_SUBSPACE_CAUSAL_SPEC and complete arm receipts	Directional ordering, matched-control specificity, protected suite, clean restart.
Substantive autoresearch	MASTER_AUTORESEARCH_EVALUATOR_QUALIFICATION and parent-child receipts	Primary and shadow end-to-end improvement, restart reproduction, memory continuity, rollback.

Declarations

Data and code availability

Frozen specifications, source hashes, manifests, complete-bank receipts, corrected label overlays, repaired-harness snapshots, RawStore records, XMB1 bridge artifacts, zero-shot replay, H3 preregistration, capture logs, controls, audits, supervisor receipts, and future causal result packages are maintained as hash-addressed artifacts. Release or independent replication packages should bind the immutable data, sources, models and tokenizer revisions, feature and target manifests, evaluator, procedures, analysis code, environment, and artifact hashes.

Competing interests

The author is founder of Proprioceptive AI, Inc. and holds intellectual-property and commercial interests related to the architecture described in this paper.