JOINT HOLONOMY MATCHING: COMPLETE SUPERVISION OF FINITE REPRESENTATION CONNECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

What information must a student preserve to reproduce a teacher's representation connection up to local coordinate changes? We introduce Joint Holonomy Matching (JHM), which aligns a fundamental loop family under one shared root gauge. Using the standard spanning-tree classification of graph connections, we establish that zero JHM is equivalent to gauge equivalence of the complete observed connections. Among supervision schemes retaining a subset of a fixed fundamental generator family, the full family is deletion-minimal for a uniform completeness guarantee: omitting any generator can leave inequivalent connections indistinguishable. Independent per-loop alignment is also incomplete in general, as matching individual conjugacy classes can miss incompatible joint compositions. We prove that every finite orthogonal connection on a simple graph admits an exact realization by local orthonormal frames and their Procrustes transports. A non-Abelian $SO(3)$ realization exhibits $L_{\text{separate}} = 0$, $L_{\text{joint}} = 5 - 2\sqrt{2}$, and a group-commutator discrepancy of $\sqrt{6}$. A quantitative bound controls errors on new compositions of observed edges. In a protocol-frozen three-seed neural pilot, Joint reduced composition error relative to Separate by 1.16 percent on the common evaluable subset of the designated test set, with the same direction across all seeds. In this implementation, Separate required 3.35 times as much loss-and-alignment time and 1.36 times as much full-update time as Joint. The structural effect was not independently established on validation, and unequal effective geometric regularization prevents attribution to the shared-gauge constraint alone. These results establish a complete connection-level supervision target, identify compositional information lost by independent alignment, and provide preliminary neural evidence for its optimization.

1 INTRODUCTION

A representation-matching objective determines which discrepancies between a teacher and a student are treated as meaningful. Logit losses supervise outputs, feature losses compare coordinates after a chosen map, relational losses preserve selected pairwise structure, and topological losses preserve selected invariants. These objectives can be useful without determining how local representation changes compose along paths. The present work isolates that compositional object and asks what information is sufficient to supervise it completely on a finite set of observations.

Each matched observation is assigned a low-dimensional orthonormal frame. Neighboring frames induce orthogonal Procrustes transports, ordered products of those transports define path maps, and closed products define holonomies. Once all loops are based at one root, their transports act on the same coordinate space. If each loop is allowed to fit that root space independently, every loop can agree up to conjugacy while mixed compositions remain incompatible. JHM prevents this loss of information by aligning a complete fundamental loop family under one shared root gauge.

The central question is therefore not whether holonomy, gauge transformations, or spanning-tree coordinates exist. Those objects are classical. The question is which gauge-invariant transport observations form a complete supervision target for a finite observed teacher-student connection. We show that a fixed fundamental loop family, compared under one common conjugation, is sufficient and deletion-minimal within that family. Separate per-loop alignment is incomplete in general. The separation is not merely an abstract matrix example: every finite $O(k)$ connection on a simple

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

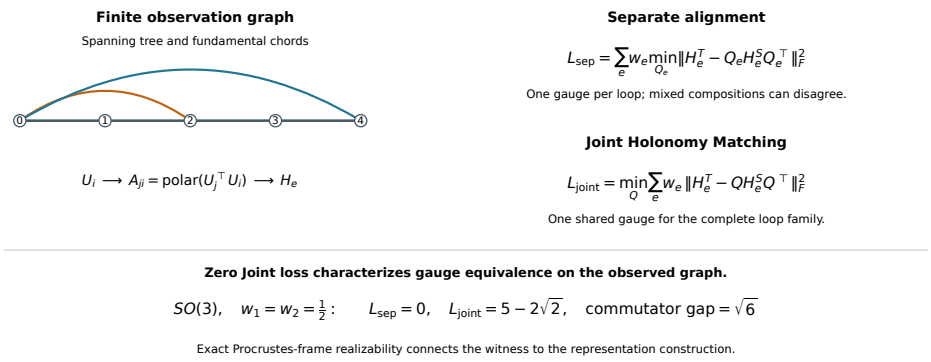

Figure 1: JHM uses one shared root gauge for a complete fundamental loop family. Separate matching can fit each loop perfectly while losing composition compatibility. The displayed $SO(3)$ witness uses $w_1 = w_2 = 1/2$.

graph can be realized exactly by the same local-frame Procrustes construction used by the method. Generator residuals also control errors on new compositions of observed edges.

The neural experiment is intentionally secondary to the structural result. It tests whether the objective can be optimized through a trainable bridge while the decoder remains frozen. The designated TEST contrast favors Joint over Separate on the common valid subset, with the same sign in all three seeds. A second held-out set does not reproduce that ordering, and unequal effective geometric regularization prevents a causal attribution of the observed contrast to the shared-gauge constraint alone. We therefore treat the empirical result as evidence that JHM is trainable in this setting, not as evidence of a general downstream advantage.

The scope of the claims is deliberately finite. Completeness concerns the observed connection and does not imply equality of full hidden representations, predictions, or statistical generalization. Deletion minimality concerns subsets of one fixed fundamental family and is not an information lower bound for arbitrary invariant encodings.

2 Related work and positioning

Structural knowledge distillation already extends beyond logits. Relational Knowledge Distillation matches distances and angles (Park et al., 2019), Geometric Knowledge Distillation transfers graph structure (Yang et al., 2022), and TopKD studies persistence-based topology (Kim et al., 2024). Geometry-aware representational alignment also uses Procrustes and Gram structure (Bhattarai et al., 2025). JHM does not claim novelty for geometry, topology, or orthogonal alignment. Its supervised object is a family of local transports whose products must remain mutually compatible.

Representation-alignment work studies equivalence classes and common coordinate systems (Han, 2026; Achara et al., 2026; Gupta et al., 2026). JHM likewise uses an orthogonal alignment, but the shared variable acts by simultaneous conjugation on all based loop operators. This makes their multiplication structure part of the supervised target rather than an incidental consequence of separate matches.

Local subspaces and orthogonal transports are classical in geometric data analysis (Singer and Wu, 2012). Tree gauges and chord holonomies are standard coordinates in group synchronization and cycle-consistency methods (Lerman and Shi, 2022). JHM does not estimate an unknown connection from noisy group ratios; it compares two prescribed finite connections induced by matched observations. Representation holonomy has also been studied as a diagnostic of a single trained network using loops in input space and rotation-only Procrustes transports (Sevetlidis and Pavlidis, 2026). The present contribution is the teacher-student supervision target, the proof that one shared alignment is complete on the observed graph, a realizable strict separation from independent loop alignment, and a neural pilot that exposes both the promise and the limitations of the objective.

3 CONNECTIONS AND JOINT MATCHING

Let $G = (V, E)$ be a finite connected simple undirected graph. For $X \in \{T, S\}$, choose frames $U_i^X \in \mathbb{R}^{d_X \times k}$ with $(U_i^X)^\top U_i^X = I_k$. For an oriented edge $i \to j$, assume the overlap is nonsingular and define

$$A_{ji}^X = \mathrm{polar}\big((U_j^X)^\top U_i^X\big), \qquad A_{ij}^X = (A_{ji}^X)^\top. \tag{1}$$

Under a local frame change $U_i^X \mapsto U_i^X R_i^X$, the edge map transforms as

$$A_{ji}^X \mapsto (R_j^X)^\top A_{ji}^X R_i^X. \tag{2}$$

Choose a rooted spanning tree $\mathcal{T}$ and let P_i^X denote its root-to-vertex transport. A chord $e = (i \to j)$ defines the based holonomy

$$H_e^X = (P_j^X)^\top A_{ji}^X P_i^X. \tag{3}$$

There are $r = |E| - |V| + 1$ chord generators; let $\mathcal{C}$ denote this chord set.

Proposition 1 (standard tree-gauge characterization)**.** *Teacher and student connections satisfy $A_{ji}^T = C_j A_{ji}^S C_i^\top$ on every edge for some $C_i \in O(k)$ if and only if one $Q \in O(k)$ satisfies $H_e^T = Q H_e^S Q^\top$ for every chord.*

Proof. Edge agreement implies $P_i^T = C_i P_i^S C_o^\top$, which gives the chord condition with $Q = C_o$. Conversely, define

$$C_i(Q) = P_i^T Q (P_i^S)^\top. \tag{4}$$

This matches every tree edge identically. Substituting the common chord identity then matches every chord as well. □

For positive weights w_e, define

$$F(Q) = \sum_{e \in \mathcal{C}} w_e \left\| H_e^T - Q H_e^S Q^\top \right\|_{\mathrm{F}}^2, \qquad L_{\mathrm{joint}} = \min_{Q \in O(k)} F(Q). \tag{5}$$

Compactness gives a minimum, and Proposition 1 identifies the zero set of JHM with full graph-connection gauge equivalence. The weaker separate objective is

$$L_{\mathrm{sep}} = \sum_{e \in \mathcal{C}} w_e \min_{Q_e \in O(k)} \left\| H_e^T - Q_e H_e^S Q_e^\top \right\|_{\mathrm{F}}^2 \le L_{\mathrm{joint}}. \tag{6}$$

Separate matching gives each loop its own root gauge and therefore discards cross-loop compatibility.

Corollary 1.1 (deletion minimality for a fixed fundamental family)**.** *Fix the graph, the spanning tree, and its r fundamental generators. The full family compared under one common conjugation is complete. No proper subset is uniformly complete over all $O(k)$ connections on the graph.*

Proof. Sufficiency follows from Proposition 1. For a proper subset, choose an omitted chord e_* and some $R \neq I_k$ in $O(k)$. In tree coordinates, set every teacher edge to the identity and every student edge to the identity except $A_{e_*}^S = R$. Every retained generator agrees, but the omitted generator is I_k for the teacher and R for the student. Conjugation cannot map a nonidentity matrix to the identity. □

This statement is deletion minimality within a fixed fundamental family, not an information lower bound over all possible invariant encodings. Separate introduces r orthogonal alignment matrices, whereas Joint introduces one. Their nominal continuous parameter counts are $rk(k-1)/2$ and $k(k-1)/2$, although stabilizers can reduce the identifiable dimensions.

4 FINITE PROCRUSTES REALIZABILITY

An abstract separation matters for the method only if it can arise from the Procrustes construction in Equation 1.

Proposition 2 (finite Procrustes realization)**.** *Let G be a finite simple graph with n vertices and maximum degree $\Delta \geq 1$. Assign arbitrary $A_{ji} \in O(k)$ to its oriented edges, with $A_{ij} = A_{ji}^{\top}$. For every $0 < \varepsilon < \Delta^{-1}$ there exist frames $U_i \in \mathbb{R}^{nk \times k}$ such that*

$$U_i^{\top} U_i = I_k, \qquad U_j^{\top} U_i = \varepsilon A_{ji} \tag{7}$$

for each edge. Hence every prescribed edge is recovered exactly by the Procrustes polar factor.

Proof. Form the symmetric block Gram matrix K with $K_{ii} = I_k$, edge blocks $K_{ji} = \varepsilon A_{ji}$, and zero nonedge blocks. For $x = (x_i) \neq 0$, orthogonality and $2ab \leq a^2 + b^2$ give

$$x^{\top} K x \geq \sum_i \big(1 - \varepsilon \deg(i)\big) \|x_i\|^2 \geq (1 - \varepsilon\Delta)\|x\|^2 > 0. \tag{8}$$

Factor $K = B^{\top} B$ and let U_i be the ith block of k columns of B. Then the stated inner products follow, and $\operatorname{polar}(\varepsilon A_{ji}) = A_{ji}$. □

The centered local cloud $[U_i, -U_i]$ has the same selected rank-k subspace, so the construction also yields explicit local feature clouds up to frame gauge. The ambient dimension nk is an existence bound. It is not claimed to be dimension-optimal, and it does not imply that an arbitrary prescribed neural architecture realizes every connection.

5 STRICT NON-ABELIAN SEPARATION

Take two fundamental loops with $w_1 = w_2 = 1/2$. Let R_x and R_y be quarter turns about the x and y axes, and set

$$(H_1^T, H_2^T) = (R_x, R_y), \qquad (H_1^S, H_2^S) = (R_x, R_x). \tag{9}$$

Individually, R_x and R_y are conjugate in $SO(3)$, so $L_{\mathrm{sep}} = 0$. Under one common conjugation, the two student matrices remain equal and become quarter turns about a single unit axis u. Rodrigues' formula gives

$$F(u) = 6 - 2(u_x + u_y) - (u_x^2 + u_y^2) \geq 5 - 2\sqrt{2}, \tag{10}$$

with equality at $u = 2^{-1/2}(1, 1, 0)$. Therefore

$$L_{\mathrm{sep}} = 0, \qquad L_{\mathrm{joint}} = 5 - 2\sqrt{2}. \tag{11}$$

For proper rotations, conjugacy in $O(3)$ and $SO(3)$ is the same. If a, b are unit axes and $\alpha, \beta \in [0, \pi]$ are canonical rotation angles, an individual residual depends only on those angles:

$$\min_Q \|R(a, \alpha) - QR(b, \beta)Q^{\top}\|_{\mathrm{F}}^2 = 8 \sin^2\left(\frac{\alpha - \beta}{2}\right). \tag{12}$$

Separate supervision therefore identifies individual conjugacy classes without constraining the relative axis geometry shared across loops.

Using the group commutator $C(H_1, H_2) = H_1 H_2 H_1^{-1} H_2^{-1}$, the student commutator is the identity and

$$\left\|C(H_1^T, H_2^T) - QC(H_1^S, H_2^S)Q^{\top}\right\|_{\mathrm{F}} = \sqrt{6} \tag{13}$$

for every Q. Proposition 2 realizes this witness with five vertices, four tree edges, two chords, $k = 3$, and $\varepsilon = 0.2 < 1/3$. One historical implementation check used the matrix Lie bracket rather than the group commutator. Both yield teacher–student Frobenius discrepancies of $\sqrt{6}$ for this quarter-turn witness, but they are not the same object in general.

6 EDGE INTERPRETATION AND COMPOSITION CONTROL

For $C_i(Q)$ from Equation 4, every tree-edge residual is zero and each chord-edge residual equals its loop residual:

$$\|A^T_{ji} - C_j(Q)A^S_{ji}C_i(Q)^\top\|_{\mathrm{F}} = \|H^T_e - QH^S_eQ^\top\|_{\mathrm{F}}. \tag{14}$$

Thus JHM can be read as chord-edge alignment subject to exact tree-edge alignment. For any path γ from a to b, let $n_e(\gamma)$ count traversals of chord e and let $s_e(Q) = \|H^T_e - QH^S_eQ^\top\|_{\mathrm{F}}$. A telescoping argument gives

$$\|A^T_\gamma - C_b(Q)A^S_\gamma C_a(Q)^\top\|_{\mathrm{F}} \le \sum_e n_e(\gamma)s_e(Q) \le \sqrt{\sum_e \frac{n_e(\gamma)^2}{w_e}}\,\sqrt{F(Q)}. \tag{15}$$

The same argument bounds any word in chord holonomies and their inverses. This controls compositions of observed edges; it does not establish statistical generalization to unseen inputs. Exact compatible edge agreement already determines every path.

JHM is also weaker than exact global orthogonal feature alignment when the local frames are estimated compatibly, because an exact shared orthogonal feature map induces gauge-equivalent local connections. The converse does not hold: connection agreement does not determine feature scale or model outputs. The strict comparison established here is Joint against Separate loop matching.

7 NEURAL PILOT ON OBSERVABLE CONNECTIONS

7.1 SETUP AND EXECUTED OBJECTIVE

A scene is a paired teacher-student token sequence of up to 80 positions generated for an aerial image. The teacher is a large language model for which the top 128 token probabilities were stored at each position. The student is a frozen decoder whose per-position hidden states were cached once. The only trainable component is a bridge with 25 tensors and 5.29 million parameters, mapping cached states to log probabilities over a fixed chart of 832 leaves. A group is a pre-existing cluster of scenes from one source image pair, and groups never cross TRAIN, VAL, and TEST. LEVIR scenes follow the LEVIR-CD source construction (Chen and Shi, 2020). Four additional aerial sources are denoted DET, UAV, SOT, and SUES because their public identities were not independently verified for this submission.

The fixed token-to-leaf map aggregates teacher support into 832 leaves, floors values at 10^{-12}, renormalizes to unit mass, and applies the square-root transform, which places points on the unit Hellinger sphere. The teacher side uses the saved top-128 observable without imputing its unknown tail, while the student side uses the bridge's full 832-leaf output. Saved top-128 support retains at least 0.998 teacher mass on every kept position, so the declared 0.95 admission rule removes no paired sequence. The shared base objective instead reads the saved top-64 target, whose median per-scene minimum retained mass is about 0.89.

Of 26,756 paired sequences that passed pairing checks, 1,408 were selected deterministically by whole groups into 1,024 TRAIN, 128 VAL, and 256 TEST scenes. All arms start from the same pretrained Base bridge checkpoint within a seed. The TRAIN scenes were already part of that checkpoint's training set, and the VAL scenes were already part of its validation set. TEST was absent from both stages. Consequently, VAL and TEST are held out with respect to JHM fine tuning, but only TEST is untouched by the initial checkpoint.

For a valid center q, a centered seven-position window yields the leading rank-three left singular subspace $U_q \in \mathbb{R}^{832\times 3}$. Consecutive centers form a chain tree rooted at the first window, while skip-one edges form chords. A sequence of length L therefore produces $L-6$ vertices and $r = L-8$ chords, with equal weights r^{-1}. Transports are polar factors of $U_j^\top U_i$ without reflection correction. A geometric scene is invalid when the relative rank-boundary gap $(\sigma_3 - \sigma_4)\sigma_1^{-1}$ or the minimum overlap singular value is below 10^{-5}. Invalid contributions were omitted during training and produce an explicit evaluation failure. About 0.88 percent of scene steps were invalid.

For bridge parameters ψ, the executed arm objective is

$$L_{\mathrm{arm}}(\psi) = L_0(\psi) + \lambda\ \mathrm{mean}_{\mathrm{scene}}\, L^{\mathrm{geo}}_{\mathrm{arm}}(\psi), \qquad L^{\mathrm{geo}}_{\mathrm{arm}} \in \{0, L_{\mathrm{sep}}, L_{\mathrm{joint}}\}. \tag{16}$$

Every geometric squared Frobenius residual uses the scale $(4k)^{-1}$ with $k = 3$. There is no additional feature loss. The common L_0 is the pretrained bridge's four-term chart-alignment objective with leaf, class, sheet, and vocabulary weights $1, 1, 1, 0.1$, and it is identical in all arms. The coefficient $\lambda = 0.21287788$ was fixed before the arm matrix on ten TRAIN batches of eight scenes by

$$\lambda \ \mathrm{RMS}\{\|\nabla L_{\mathrm{sep}}\|, \|\nabla L_{\mathrm{joint}}\|\} = 0.1 \ \mathrm{mean} \, \|\nabla L_0\|, \tag{17}$$

using the pooled RMS across both geometric arms. Numerically, this was $0.1 \times 4.397 \times (2.066)^{-1}$. The realized calibration gradient shares were 0.054 for Separate and 0.131 for Joint, which is an important confound discussed below.

The inner Q problem used a modified Newton method on $SO(3)$ with retraction, Armijo search, 14 fixed starts, float64 arithmetic, stopping gradient norm 10^{-12}, and at most 2,000 iterations. For odd k, $SO(k)$ covers all $O(k)$ conjugations because Q and $-Q$ induce the same conjugation. The minimizing Q^* is detached and bridge parameters receive the envelope gradient. Finite-difference relative errors were 1.39×10^{-6} for Separate and 3.85×10^{-7} for Joint. The frozen protocol prose named an earlier Levenberg-Marquardt solver with a 200-iteration cap; the executed, hashed code used Newton. This is a protocol-prose mismatch, not a post-result change to the executable campaign. All nine runs used seeds 17, 29, and 43, paired initialization and minibatches within seed, 2,000 updates, batch size eight scenes, and learning rate 5×10^{-5}. Each run continued a Base bridge checkpoint previously trained for 3,500 updates at batch 128, with optimizer state reset identically across arms.

7.2 Evaluation endpoint

For each held-out scene, chords are split by parity of a fixed SHA-256 hash of scene identity and chord index into calibration and evaluation halves, with at least three chords in each half. One common Q is fitted to the calibration half and then frozen. Student parameters are never updated on TEST, but the evaluation is not zero-shot alignment because Q is calibrated against teacher holonomies inside each held-out scene.

On the evaluation half, a fixed hash rule composes 32 words, with eight words at each length 2, 4, 8, and 16, from held-out holonomies and their transposes. Each word uses at least two distinct generators and excludes adjacent inverse pairs and repeats. Its error is

$$\mathrm{err}(w) = \frac{\|H_w^T - Q H_w^S Q^\top\|_{\mathrm{F}}^2}{4k} \in [0, 1]. \tag{18}$$

The scene value averages words within each length and weights the four lengths equally. Scenes are averaged within groups, groups are averaged within a set, and the three seeds are averaged last. This endpoint is a composed-word analogue of the shared-gauge objective. It tests the hypothesized Joint versus Separate ordering, rather than an independent downstream criterion.

7.3 Structural results

The planned full TEST primary is not fully computable. Eleven of 256 scenes, 4.30 percent, fail on at least one arm under the frozen degeneracy rule. The common valid portion contains 245 scenes in 168 groups, and under the preregistered protocol its contrasts are secondary to the planned full-set primary. Refusals are arm-dependent: one scene fails in all nine checkpoint cells, while the other ten fail in one cell each. Joint accounts for 8 of the 19 failing cells, Base for 6, and Separate for 5. Five sensitivity treatments, including arm-specific valid sets, three imputations, and scene-level resampling, preserve the sign and interval exclusion of zero for Joint minus Separate.

On the common valid TEST portion, mean composition error is 0.3636 for Base, 0.3643 for Separate, and 0.3601 for Joint. The Joint minus Separate contrast is

$$-0.004223, \qquad 95\% \text{ interval } [-0.007473, -0.001112], \tag{19}$$

which is 1.16 percent of the approximately 0.364 error level. Seed-wise differences are -0.003913, -0.007020, and -0.001737. Their signs agree, but their range, 0.00528, is larger than the aggregate estimate. Per-seed intervals exclude zero only for seed 29. The aggregate bootstrap interval is conditional on these three trained models and does not estimate uncertainty over newly trained models.

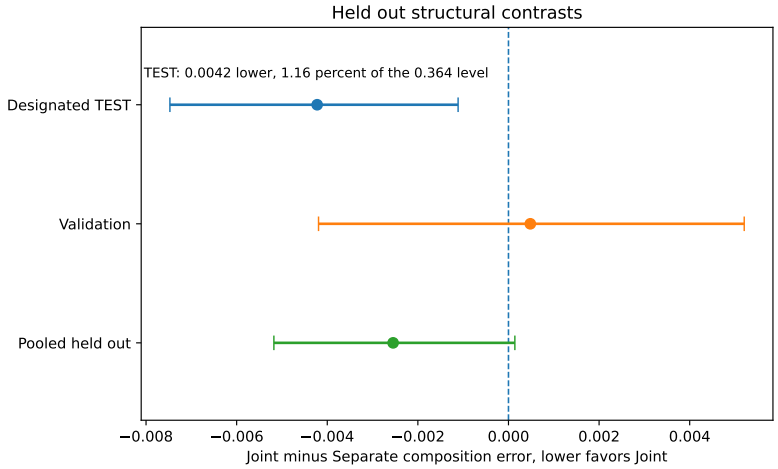

Figure 2: Joint minus Separate composition error on the common valid portions of the held-out sets. TEST favors Joint, while VAL and the pooled held-out interval include zero. Intervals are paired group-bootstrap intervals conditional on the three training seeds.

On VAL, using 118 valid scenes in 93 groups, Joint minus Separate is $+0.000482$ with interval $[-0.004193, +0.005204]$. Pooling the two nonoverlapping held-out sets over 261 groups gives -0.002546 with interval $[-0.005178, +0.000141]$. The TEST minus VAL contrast is -0.004705 with interval $[-0.010382, +0.000819]$. These held-out results are compatible with a small negative effect and with no effect. The designated TEST ordering is observed, but a neural advantage across held-out data is not established.

The calibrated common Q leaves median calibration-chord residual 0.104 and a maximum of about 0.199, so the one-percent arm difference sits on top of substantial teacher-student mismatch. Separate minus Base is $+0.000748$ with interval $[-0.002243, +0.003726]$, which is unresolved. Joint minus Base is -0.003475 with interval $[-0.007003, -0.000087]$, but that exclusion is fragile: best-value imputation gives $[-0.006887, +0.000093]$, the LEVIR-only interval covers zero, each seed interval covers zero, and VAL gives $+0.000318$ with interval $[-0.004292, +0.005453]$. We therefore treat Joint minus Base as unresolved.

TEST is close to a single-source evaluation. Of 168 groups, 162 are single LEVIR scenes, four are SUES groups, and DET and UAV contribute one group each. The Joint minus Separate contrast is -0.00419 on LEVIR, whose interval excludes zero, and -0.00649 on SUES, whose interval includes zero. The corresponding values are -0.00538 on UAV and $+0.00133$ on DET, but no source-specific inference is possible from one group. TEST contains no SOT scene and VAL contains no UAV scene. We make no transfer claim across source families. The four word-length intervals also cover zero and overlap, so the data do not resolve dependence of the contrast on composition length. The held-out chord-residual contrast under frozen Q is unresolved as well. Equation 15 controls composition error from generator residuals, but it does not explain the observed neural arm contrast.

7.4 Regularization load, cost, and optimization regime

A single λ did not equalize geometric gradient load. At calibration, the effective shares of $\|\nabla L_0\|$ are 0.131 for Joint and 0.054 for Separate. During training their median shares are approximately 0.102 and 0.039. The Joint minus Separate contrast is therefore confounded with regularizer magnitude and cannot be causally attributed to the shared-gauge constraint by this campaign. A matched-load replication would be needed for that attribution and has not been run.

Median full-update time is 0.4632 seconds for Base, 0.6454 for Joint, and 0.8786 for Separate. Joint is 1.39 times Base, Separate is 1.90 times Base, and Separate requires 1.36 times the Joint full-update time. The structural loss and alignment block takes 0.0988 seconds for Joint and 0.3312 seconds for Separate, so Separate requires 3.35 times the Joint block time in this implementation. The structural

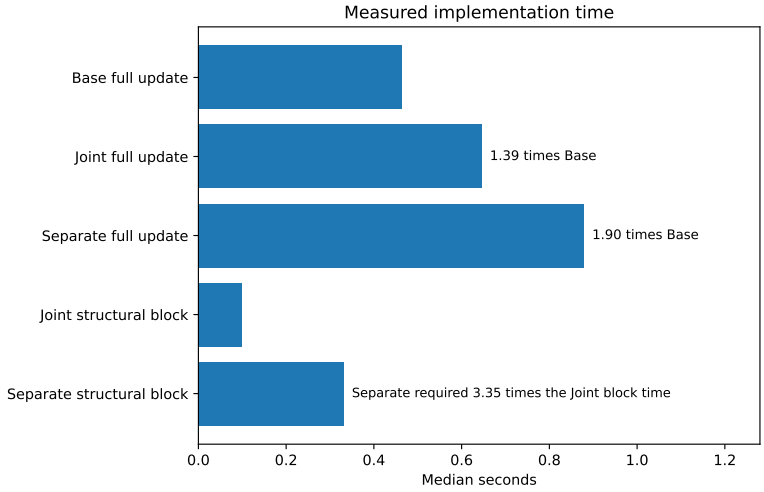

Figure 3: Measured implementation time. Joint adds overhead relative to Base but is less expensive than the implemented Separate control. These timings are implementation-specific and are not a complexity theorem.

block runs on CPU in float64 alongside the GPU bridge pass. Joint solves one alignment problem per scene, while Separate solves one per chord, about 72 chords per scene. Data and update counts are matched; FLOPs and wall time are not compute-matched.

The nine-run matrix took 3.51 hours on one RTX 4060. The broader 3.69-hour campaign window also includes preflight, coefficient calibration, and final evaluation. Training had not reached a plateau at 2,000 updates: over the final 20 percent of updates, TRAIN base loss still fell by 10 to 13 percent of its total decrease in every arm and seed. Nothing is claimed about arm ordering at convergence. A post-campaign measurement of the training-time base objective shows held-out L_0 rising by about $+0.053$, roughly 13 percent of its initial level, for every arm including Base, while TRAIN falls by about -0.14. The inherited continuation regime therefore overfits the 1,024 TRAIN scenes. All arms used the same continuation regime, so the reported arm contrasts are conditional on that regime. This deterioration also makes the current recipe unsuitable as a deployment improvement. No downstream generation score or human semantic evaluation was measured.

8 Discussion

JHM isolates a precise supervision problem: how to preserve the connection induced by a finite collection of local representation frames up to changes of local coordinates. A complete fundamental loop family under one shared root gauge determines that observed connection up to gauge. Deleting one generator can destroy the guarantee, and independent loop alignment can match every individual conjugacy class while missing joint composition. The finite Procrustes realization shows that this distinction is compatible with the local-frame construction used by the method, rather than existing only for unconstrained abstract matrices. The path bound then connects generator residuals to new compositions on the same observed graph.

The neural pilot provides a narrower result. On the common valid portion of designated TEST, Joint has lower composition error than Separate in all three seeds, but validation does not reproduce the ordering and the pooled held-out interval includes zero. The endpoint itself is structural, the effective geometric loads are unequal, and no downstream or external criterion was measured. These facts prevent a general claim about neural utility or deployment. They do not weaken the exact structural statements, whose zero set, deletion minimality, realizability, strict non-Abelian witness, and composition bound are independent of the size of the neural contrast.

A focused next experiment would compare Joint and Separate after TRAIN-only calibration of their geometric gradient loads, while measuring how those loads evolve during training. Broader

source coverage and an independent downstream criterion would address distinct questions of generalization and practical utility. JHM already provides a complete supervision target for finite observed connections up to gauge, distinguishes compositional incompatibilities missed by separate alignment, and admits optimization through neural bridge parameters.

REPRODUCIBILITY STATEMENT

The main paper states the mathematical assumptions and proofs and records the executed neural objective, observable, graph construction, endpoint, failure rule, and principal numerical contrasts. The anonymous supplement contains the witness generator, selected training and evaluation code, the frozen protocol fields needed to interpret the campaign, sanitized summary tables, solver checks, and a manifest of cryptographic hashes. It excludes private paths, names, nonpublic README material, and model weights. The surviving sanitized artifacts identify the starting bridge state by seed and training stage, namely the Base checkpoint after 3,500 updates at batch 128, but they do not contain a recoverable immutable teacher-model identifier, student-model revision identifier, or original checkpoint filename and hash. We therefore do not invent those identifiers. Numerical summary tables can be checked without rerunning training; complete neural re-execution additionally requires the cached tensors and checkpoint assets described in the supplement index.

AI USE STATEMENT

Generative AI tools were used as research-assistance tools during the development of this work. They assisted with literature discovery and summarization, refinement of the conceptual and methodological framing, formulation and checking of mathematical claims and proof arguments, code drafting and debugging, experiment specification, numerical cross-checking, interpretation of stored experimental outputs, figure preparation, and drafting and editing portions of the manuscript. This assistance included work on the finite Procrustes realizability argument, the non-Abelian separation witness, the composition-control analysis, and the neural evaluation methodology. The neural pilot uses aerial-image data together with saved teacher-generated token probabilities as supervision. These model-generated targets are distinguished from the source images and from the AI assistance used to develop and write this work.

All mathematical statements and derivations were subsequently checked by the authors against explicit algebraic or numerical calculations where applicable. AI-assisted code and numerical analyses were inspected and executed by the authors, and reported numerical results were checked against stored experimental outputs. Literature references and factual claims were reviewed by the authors before inclusion. The authors take responsibility for the methodology, proofs, experiments, interpretations, citations, and all other content of the submission.

REFERENCES

Park, W., Kim, D., Lu, Y., and Cho, M. Relational Knowledge Distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Yang, C., Wu, Q., and Yan, J. Geometric Knowledge Distillation: Topology Compression for Graph Neural Networks. In *Advances in Neural Information Processing Systems*, 35, 2022.

Kim, J., You, J., Lee, D., Kim, H. Y., and Jung, J. H. Do Topological Characteristics Help in Knowledge Distillation? In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Bhattarai, P., Amjad, M., Zhylko, D., and Alhanai, T. Knowledge Distillation through Geometry Aware Representational Alignment. arXiv:2509.25253, 2025.

Han, S. I. Teacher Supervision over Representation Equivalence Classes. arXiv:2607.03572, 2026.

Achara, A., Gaintseva, T., Mahaut, M., Chakraborty, P., Johansson, V. S., Barsbey, M., Rodolà, E., and Crisostomi, D. Multi Way Representation Alignment. arXiv:2602.06205, 2026.

Gupta, S., Kansal, S., Jegelka, S., Isola, P., and Garg, V. Canonicalizing Multimodal Contrastive Representation Learning. arXiv:2602.17584, 2026.

Sevetlidis, V. and Pavlidis, G. Gauge invariant Representation Holonomy. In *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026.

Singer, A. and Wu, H. T. Vector Diffusion Maps and the Connection Laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012.

Lerman, G. and Shi, Y. Robust Group Synchronization via Cycle Edge Message Passing. *Foundations of Computational Mathematics*, 22:1665–1741, 2022.

Schönemann, P. H. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31:1–10, 1966.

Le Jan, Y. Holonomies and Gauge Fields. In *Random Walks and Physical Fields*, pages 139–167. Springer, 2024.

Chen, H. and Shi, Z. A Spatial Temporal Attention Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12(10):1662, 2020.

A Additional theory scope and rank-two witness

For the rank-two witness, take $\theta = \pi/3$, $w_1 = w_2 = 1/2$, teacher pair (R_θ, R_θ), and student pair $(R_\theta, R_{-\theta})$. A reflection matches the second loop, so Separate is zero. One residual remains under a common $O(2)$ element and gives $L_{\mathrm{joint}} = \frac{1}{2}\|R_\theta - R_{-\theta}\|_{\mathrm{F}}^2 = 3$. The teacher product is $R_{2\theta}$ and the student product is the identity, with gap $\sqrt{6}$. On the neural $(4k)^{-1}$ scale, the rank-two value 3 becomes 0.375 and the rank-three value $5 - 2\sqrt{2}$ becomes 0.1809644.

The finite realization excludes nonidentity self-loop transports and arbitrary parallel-edge assignments. The nk ambient dimension is an existence bound. The construction does not show that every prescribed neural architecture, metric-neighborhood rule, or low-dimensional manifold realizes arbitrary connection data.

B Constructive witness and numerical checks

The anonymous supplement contains a standalone NumPy witness program that builds the block Gram matrix, factors it, reconstructs local frames, recomputes every Procrustes edge, forms chord holonomies, and verifies the exact $O(2)$ and $SO(3)$ witnesses. In the five-vertex construction, four edges form a chain tree and chords connect vertices 0 to 2 and 0 to 4. The maximum degree is 3 and $\varepsilon = 0.2 < 1/3$, so Proposition 2 applies directly. The recorded $SO(3)$ loss is $5 - 2\sqrt{2}$ to numerical precision and recovered edge errors are at machine precision. The supplement also contains independent algebra checks for gauge invariance and the path bound.

The main text uses the group commutator in the strict separation result. One historical implementation check used the matrix Lie bracket. The two expressions are not interchangeable, even though both yield teacher–student Frobenius discrepancies of $\sqrt{6}$ for the quarter-turn witness. All claims in the paper refer to the group commutator.

C Neural detail tables

The corresponding Joint minus Separate seed intervals are $[-0.009664, +0.001760]$, $[-0.012989, -0.001018]$, and $[-0.007648, +0.004040]$. The group bootstrap used 4,000 paired resamples with one shared resampling draw across arms.

Of the 168 common valid TEST groups, 162 are single LEVIR scenes, four are SUES groups, and DET and UAV contribute one group each. The clustering correction therefore acts on six groups only. Independence of distinct LEVIR pairs is inherited from the source group map rather than established here. The source-specific Joint minus Separate values are -0.004194 for LEVIR with interval

Table 1: TEST composition error on the common valid portion.

Seed	Base	Separate	Joint
17	0.359714	0.360378	0.356466
29	0.364136	0.367901	0.360881
43	0.366933	0.364749	0.363012

$[-0.007566, -0.000833]$, -0.006494 for SUES with interval $[-0.015141, +0.002154]$, -0.00538 for UAV, and $+0.00133$ for DET. The one-group sources do not support source-specific inference.

The four composition-length contrasts for lengths 2, 4, 8, and 16 are -0.002420, -0.004845, -0.005991, and -0.003637. All four intervals cover zero and overlap. The data therefore do not resolve a word-length interaction. Measured after the campaign, the common base loss rises from about 0.413 at TEST initialization to about 0.466 after every arm and from about 0.398 to about 0.449 on VAL, while TRAIN drops by about 0.14. These are secondary diagnostics measured after the protocol-frozen campaign and are not substitutes for the structural endpoint.

D SOLVER AND PROTOCOL NOTES

The simultaneous-conjugation solver is a modified Newton method on $SO(3)$ with 14 deterministic starts, retraction, Armijo line search, float64 arithmetic, gradient-norm tolerance 10^{-12}, and at most 2,000 iterations. The prose in the frozen protocol retained an older Levenberg-Marquardt description with a 200-iteration cap. The executed source, hashes, and all reported results correspond to the Newton solver. The discrepancy was documented before manuscript preparation and changes no stored result.

The starting bridge checkpoint differs by seed but is shared across Base, Separate, and Joint within that seed. Optimizer state is reset identically. The initial checkpoint had already used the current TRAIN scenes for training and the current VAL scenes for validation, whereas current TEST scenes were absent from both. TEST is therefore the only split untouched by the initial bridge checkpoint. Initial base losses are approximately 0.349 on TRAIN, 0.398 on VAL, and 0.413 on TEST.