PATH-LOCKED NEURAL COMPOSITION: EXACT EVALUATION-PATH INVARIANCE WITH GEOMETRIC ALGEBRA

Anonymous authors
Paper under double-blind review

ABSTRACT

We combine learned motion calibration with exact geometric algebra (GA) composition. Paired motor products in projective GA form a mergeable 16-coordinate summary, followed by a learned gate. On development-selected, family-held-out ETH3D odometry, this gate attains 0.440 m translation error with 577 head parameters, compared with 0.915 m for identity motion and 0.526 m for feature-matched direct regression, reductions of 51.9% and 16.4%. Its throughput is $2.17\times$ that of a contextual gated recurrent unit (GRU) variant at batch 256 on one CPU thread, excluding odometry extraction. Feature-matched controls support the translation benefit of input-conditioned motor blending. We characterize evaluation-path invariance by associativity on reachable states and evaluate it on four public recorded-data benchmarks. Fixed-input tests identify *path-locking*: MovieLens-1M accuracy falls from 0.4319 to 0.3483 when the same bilinear-residual model uses a right rather than left fold; TUM RGB-D motion error increases from 6.83 to 36.53 cm. Exact GA composition has a maximum observed path discrepancy of 2.28×10^{-15} m on TUM. Intermediate-state regularization reduces empirical path sensitivity without certifying invariance. Associative motor composition supports accurate learned calibration and consistent chunk merging, while predictive improvements remain dataset- and horizon-dependent.

1 INTRODUCTION

Projective geometric algebra (PGA) represents 3D rigid motions as motors whose geometric product implements physical motion composition (Hestenes & Sobczyk, 1984; Gunn, 2019). This product is bilinear, generally noncommutative and associative. It preserves motion order while allowing serial, balanced and chunked reductions to represent the same transform. We use this structure to separate learned calibration from the composition law: a pointwise network corrects motion estimates, two motor products summarize the raw and calibrated sequences, and a learned gate produces the final motion. The complete paired summary is mergeable without a recurrent encoder.

On family-held-out ETH3D odometry (Schöps et al., 2019), this construction reduces translation error relative to identity motion and feature-matched direct and residual regression, with a compact readout and lower measured computation cost than the contextual GRU variant. These results concern learned correction of recorded sensor estimates. TUM RGB-D (Sturm et al., 2012) separately tests the composition law using tracked relative motions, and KITTI (Geiger et al., 2012) evaluates the architectures on outdoor stereo estimates. The MovieLens-1M matrix model (Harper & Konstan, 2015) is an associative-algebra control, not evidence specific to GA.

The choice of *evaluation path*, or reduction tree, affects parallel depth, chunk boundaries and cached recomputation. An unconstrained learned binary operator need not agree across these paths, even when its trained-fold accuracy generalizes to longer sequences. We call prediction degradation under reparenthesization *path-locking*. Our diagnostic fixes the checkpoint, ordered observations and readout, changing only the tree. We compare exact products, mixed-path training and primitive/intermediate-state associator penalties. The results distinguish empirical consistency from

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.

associativity on all reachable states. Uncertainty estimates use users, trajectories or families as resampling units, rather than individual overlapping windows.

1. **Learned calibration with a mergeable PGA motor state** on recorded ETH3D: a 577-parameter gate reduces translation error without a recurrent encoder. Feature-matched controls, frozen-checkpoint horizon tests and computation measurements characterize its accuracy and efficiency.
2. A **fixed-input evaluation-path diagnostic**, with five-seed MovieLens and TUM results, based on classical associativity on reachable states.
3. **Controlled comparisons of structural and learned consistency**: exact products, mixed paths and primitive/intermediate-state penalties, including query-depth sensitivity and full-summary cache verification.

2 Related work

Associative computation. Parallel prefix (Ladner & Fischer, 1980; Hillis & Steele, 1986; Blelloch, 1990), distributed aggregation (Dean & Ghemawat, 2004; Akidau et al., 2015), linear attention/recurrences (Martin & Cundy, 2018; Katharopoulos et al., 2020) and state-space models (Gu et al., 2022; Smith et al., 2023; Gu & Dao, 2024) use associative operations. Our contribution concerns their use in learned motor calibration and evaluation-path diagnostics, not associative reduction itself. Prefix-Scannable Models (Yau et al., 2026) also permit faithful online and parallel execution of non-associative aggregators by preserving a prescribed tree. We require arbitrary regrouping of ordered summaries, as when changing motion-chunk boundaries or updating cached intervals. Preserving a prescribed tree is an alternative execution model, distinct from arbitrary regrouping.

Learning algebraic laws. Learnable commutative monoids regularize associativity in graph aggregation (Ong & Veličković, 2022); associativity-biased log-depth recurrences extend this to ordered sequences (Pert et al., 2026). The latter's intermediate-state cosine penalty motivates our intermediate-state controls; the primitive-only penalty does not reproduce that method. A concurrent CatLPN manuscript imposes associativity, identity and closure losses on latent programs (Anonymous, 2026). We add a fixed-input behavioral test and compare exact structure with path exposure and primitive/intermediate penalties on recorded histories.

Geometric algebra representations. Projective GA represents rigid motions as motors, including dual-quaternion coordinates (Hestenes & Sobczyk, 1984; Gunn, 2019). GA Transformers and Clifford neural layers exploit multivectors for equivariant learning (Brehmer et al., 2023; Brandstetter et al., 2023; Ruhe et al., 2023; Filimoshina & Shirokov, 2025); PGA representations also appear in robot policies (Sun et al., 2025), while dual-quaternion networks regress poses on $SE(3)$ (Li et al., 2022). We study agreement across evaluation trees. Equivariance alone does not imply this property; the fixed geometric product supplies it directly.

DPC-Net learns $SE(3)$ corrections to classical odometry (Peretroukhin & Kelly, 2018); dual-quaternion blending is also established (Kavan et al., 2008). Our construction combines pointwise calibration, an exactly mergeable paired-motor state and a learned readout, testing utility separately from invariance. Motion-only controls do not replicate image-conditioned DPC-Net.

Learned relational composition. Relational models use additive (Bordes et al., 2013), complex/rotational (Trouillon et al., 2016; Sun et al., 2019) and bilinear composition (Nickel et al., 2011; Socher et al., 2013; Yang et al., 2015). Path-query and recurrent systems learn repeated relational updates (Guu et al., 2015; Neelakantan et al., 2015; Hochreiter & Schmidhuber, 1997; Cho et al., 2014). Unlike length-generalization tests, our intervention fixes the inputs. It complements behavioral tests of capabilities not resolved by aggregate accuracy (Ribeiro et al., 2020).

3 Evaluation-path consistency

3.1 Ordered trees and reachable closure

Let $E \subseteq \mathbb{A} \subseteq \mathbb{R}^d$ be the set of encoded events and let $\mu : \mathbb{A} \times \mathbb{A} \to \mathbb{A}$ be a binary composer. A full ordered binary tree T over $x_{1:L} \in E^L$, $L \geq 1$, recursively applies μ without permuting its leaves;

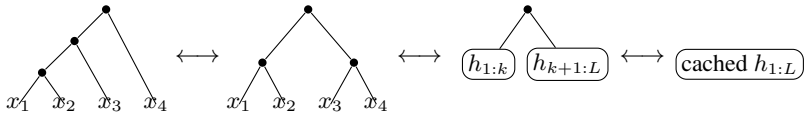

Figure 1: Evaluation paths preserve leaves and order but change parentheses. Associativity on reachable states makes them equivalent. The diagnostic compares these paths while holding the checkpoint and recorded history fixed.

write its output as $h_T(x_{1:L})$. There are $C_{L-1} = \frac{1}{L}\binom{2(L-1)}{L-1}$ such trees. The left fold is the path used for ordinary recurrent training.

The reachable closure $\mathcal{C}(E)$ is the smallest set containing E and closed under μ; closure here is algebraic, not topological. Define the associator

$$\mathcal{A}(a, b, c) = \mu(\mu(a, b), c) - \mu(a, \mu(b, c)). \tag{1}$$

Theorem 1 (Closure-level path invariance)**.** *For a domain $\mathcal{C}$ closed under μ, the following are equivalent: (i) $\mathcal{A}(a, b, c) = 0$ for every $a, b, c \in \mathcal{C}$; and (ii) every two ordered full binary trees have equal outputs for every nonempty finite sequence in $\mathcal{C}$.*

The reverse direction follows immediately from the two length-three trees. For the forward direction, any full binary tree can be transformed into the left comb by associativity rotations; each rotation preserves its value. The closure condition is required because an internal node is generally not distributed like a primitive event. For $\mathcal{C} = \mathcal{C}(E)$, invariance of all trees over E also implies this law: graft event trees realizing any $a, b, c \in \mathcal{C}(E)$ into the two length-three trees (Appendix A). This is full-state equality; predictions from a non-injective readout need not reveal non-associativity.

Proposition 1 (Operational consequences)**.** *Under Theorem 1, (a) a balanced tree equals a serial fold while using $\lceil \log_2 L \rceil$ algebraic rounds given sufficient parallel resources; (b) any ordered partition into contiguous chunks can be reduced independently and merged without changing the result; and (c) replacing one summary in a balanced cache of N chunks requires $O(\log N)$ ancestor updates, excluding recomputation within that chunk. Appending a summary has the same bound with balanced insertion and cached products maintained during rebalancing.*

These bounds concern associative reduction. Hardware, communication, numerical precision and surrounding neural modules determine wall-clock performance. In the contextual ETH3D model, the GRU still processes the full sequence: only its raw and calibrated product branches admit this cache/merge guarantee. Holding that context computation fixed preserves the final prediction under product-tree changes, but separately gated chunk predictions have no established merge law.

Representation boundary. Under the standard even-subalgebra identification, a unit PGA motor $q + \epsilon d$ is a unit dual quaternion and double-covers an element of $SE(3)$; its geometric product maps to rigid-motion composition (Dorst et al., 2007; Gunn, 2019). Thus an exact dual-quaternion or homogeneous-transform implementation supplied with identical corrected motions is representation-equivalent. The experiments do not establish an accuracy advantage over equivalent coordinates. We test an exact geometric composition operator with learned calibration and readout, rather than learned global aggregation. PGA provides the motor representation, product and geometric action used in this architecture.

3.2 Path-locking and diagnostics

For a trained readout g, path-locking is degradation in predictive performance of $g(h_T(x))$ when changing from the trained path T_0 to another tree T on the same x. We report the paired accuracy change and representation dispersion

$$D_T(L) = \mathbb{E}_x \frac{\|h_T(x) - h_{T_0}(x)\|_2}{\|h_{T_0}(x)\|_2 + 10^{-12}}. \tag{2}$$

Both statistics use unchanged task data. The paired performance change measures predictive sensitivity, while dispersion measures the change in representation.

A local associator penalty $\mathbb{E}_{a,b,c\sim E}\|\mathcal{A}(a,b,c)\|^2$ need not enforce Theorem 1: after one reduction, the operands lie in $\mu(E,E)$, and later triples lie deeper in $\mathcal{C}(E)$. Zero error on E^3 therefore does not establish the identity outside its sampled support. Mixed-path training samples intermediate states from several reduction trees, but finite paths and lengths cannot certify a global law. An associative parameterization sets the associator to zero by construction.

4 Experimental methods

4.1 MovieLens dataset, task, and held-out split

MovieLens-1M contains 1,000,209 anonymous ratings by 6,040 users on 3,706 observed movies (Harper & Konstan, 2015). We use the official archive and verify MD5 `c4d9eecfca2ab87c1945afe126590906`. Ratings are sorted within user by the published timestamp, with source-file order as the deterministic tie break. A SHA-256 hash of user ID assigns 4,857 users to training, 596 to development, and 587 to confirmation. No user crosses a partition.

For an endpoint t, the input is L preceding observed pairs $(\text{movie}, \text{rating})$ followed by the queried movie at t with its rating masked. The target is the recorded 1–5 star rating. We train at $L = 8$ and evaluate $L \in \{4, 8, 16\}$. The confirmation split contains 96,340, 93,992, and 89,296 eligible endpoints respectively. No negative examples, artificial transformations, simulated observations, or data augmentation are used.

The protocol was fixed before substantive training. Development established an acceptance criterion: a learned arm had to exceed the strongest train-only baseline, degrade materially under at least two alternative tree families with user-clustered intervals excluding zero, and the exact arm had to remain invariant within 10^{-5}. Only after all five development seeds passed were confirmation labels evaluated. Confirmation did not select arms, seeds, lengths, or metrics.

4.2 Composition models and controls

All models map movie and rating IDs to a width-nine event state, reduce the $L + 1$ states, apply a final layer normalization, and use the same five-class linear head. Full-tree tests include the masked query as the last leaf, whose depth changes from one in the left fold to L in the right fold. A separate frozen-checkpoint test fixes the query at the final merge and changes only the history subtree (Appendix G).

Exact matrix. The event state is reshaped into a 3×3 matrix, $X = I+0.05\tanh(z)$, and $\mu(X,Y) = XY$. Matrix multiplication is generally noncommutative but associative. This arm learns operands and readout with a fixed composition law: 33,494 total parameters, none in composition.

Bilinear-residual. For a learned tensor $W \in \mathbb{R}^{9\times9\times9}$,

$$\mu(a,b) = \mathrm{LN}\left(\tfrac{1}{2}(a+b) + \tfrac{1}{3}\sum_{i,j=1}^{9} a_i W_{ij:} b_j\right). \tag{3}$$

It has 747 composition parameters and 34,241 total parameters. The full rule, including residual addition and layer normalization, is not a bilinear map.

Nonlinear MLP. A two-layer GELU MLP maps $[a;b] \in \mathbb{R}^{18}$ through hidden width 25 to $\mathbb{R}^9$, followed by layer normalization. It has 727 composition parameters, within 2.7% of the bilinear rule, and 34,221 total.

Interventions. The *mixed-path* arm uses the bilinear model but uniformly trains on left, right, balanced, or sampled order-preserving trees. The *local-penalty* arm trains on the left fold and adds $\frac{0.1}{9}\|\mathcal{A}(a,b,c)\|^2$, averaged over the batch, at one sampled consecutive-triple position shared by the batch of encoded sequences. This primitive-only control is supplemented by all-primitive and intermediate-state penalties (Appendix G). No intervention changes an event or label.

All arms use AdamW for 3,000 steps, batch size 256, learning rate 2×10^{-3}, weight decay 10^{-4}, and gradient clipping at one. Five seeds change initialization and sampling over training endpoints. Baselines are a train-only global rating mode, per-movie mode with global fallback, and per-movie

empirical rating distribution. Evaluation uses the left fold, right fold, balanced tree, and 16 independently seeded order-preserving random-tree schedules. Confidence intervals resample whole users with their complete sets of endpoints (2,000 draws).

4.3 TUM RIGID MOTIONS: THE GEOMETRIC-ALGEBRA INSTANTIATION

The TUM RGB-D benchmark provides recorded camera trajectories measured by an external high-accuracy motion-capture system (Sturm et al., 2012). We use all 47 official ground-truth trajectory files and verify every source checksum. A deterministic name hash assigns 25 complete trajectories to training, 9 to development, and 13 to confirmation. We select the first recorded pose at or after each 0.1-second grid point, without interpolation, then form consecutive relative motions. The target is the relative transform between the two recorded window endpoints. We train at $L = 8$ (nominal 0.8 s) and evaluate $L \in \{4, 8, 16, 32\}$; no image, pose, perturbation, or label is generated.

A relative rigid motion is encoded as a unit dual quaternion $M = q + \epsilon d$, $d = \frac{1}{2}tq$, with t a pure translation quaternion, equivalently a motor in the even subalgebra of 3D projective GA (Gunn, 2019). The exact composer is its geometric product,

$$(q_1 + \epsilon d_1)(q_2 + \epsilon d_2) = q_1 q_2 + \epsilon(q_1 d_2 + d_1 q_2), \tag{4}$$

with no normalization inside the reduction. This is the physically correct semidirect composition and is associative by construction. An invariant but physically incomplete control multiplies rotations but adds translations without the rotational action. Learned arms begin at this control and add a zero-initialized bilinear or parameter-matched MLP residual in eight motor coordinates, followed by projection to a unit motor. Mixed-path training and the local associator penalty match the MovieLens interventions. We optimize five seeds for 5,000 AdamW updates using scaled squared translation error plus a sign-invariant quaternion loss (Appendix E). Intervals resample whole trajectories 10,000 times.

4.4 ETH3D: LEARNED CALIBRATION OVER EXACT MOTOR PRODUCTS

TUM derives leaves from tracked poses; ETH3D tests imperfect sensor estimates. We use all 56 real motion-capture training sequences in its SLAM benchmark (Schöps et al., 2019), excluding five Structure-from-Motion sequences and all synthetic data. Open3D 0.19 hybrid RGB-D odometry (Zhou et al., 2018) estimates consecutive motions from unchanged RGB and active-stereo depth; motion capture supplies endpoint targets. Failed pairs break runs, and no observation or pose is imputed, perturbed, or generated.

Let $M_{\mathrm{raw}} = \prod_i M_i$ and $M_{\mathrm{cal}} = \prod_i c_\theta(M_i)$, using the PGA product in recorded order. The main model reads only the paired state $S = (M_{\mathrm{raw}}, M_{\mathrm{cal}})$ through a 577-parameter $16 \to 32 \to 1$ GELU/sigmoid gate, producing $\alpha \in [0, 1]$. Its prediction is the unit-motor projection of $(1 - \alpha)M_{\mathrm{raw}} + \alpha M_{\mathrm{cal}}$. The state merges as $(A, B) \star (C, D) = (AC, BD)$: independent chunks retain both motors, and the readout is applied only after merging. The complete 16-coordinate summary is associative; the projected prediction is not itself the object to merge. Projection requires a nonzero real quaternion. The original blend does not align endpoint signs; Appendix L tests sign alignment separately without changing its reported results.

We hold out 16 name-defined capture families indivisibly in five balanced folds. Within each fold and seed, a 280-parameter calibrator is trained at $L = 8$ and frozen; heads are trained at primary $L = 32$. The original contextual comparator uses a bidirectional GRU of raw leaves, pooling its outputs with both endpoints to predict α (10,689 head parameters). A two-stage budget-matched 10,846-parameter direct GRU receives calibrated leaves and the same batches, but neither raw leaves nor exact endpoint features. Thus this original comparison matches training, not information interfaces. Families are weighted equally over five seeds. This protocol followed earlier ETH3D development and is a grouped robustness test, not untouched-dataset confirmation; Appendix F retains the selection history.

Attribution controls. Post-hoc controls give direct and residual heads the *same* raw-leaf GRU and both exact endpoint features as the contextual gate; a constant gate tests input dependence. All outputs are projected to unit motors. These controls distinguish scalar blending from direct or residual coordinate prediction, not valid motors from invalid poses. The GRU-free model uses the same frozen calibrators and fitting budget (Appendix H).

5 RESULTS

5.1 LEARNED CALIBRATION WITH A MERGEABLE GA STATE

The mergeable GA gate reaches 0.440 m family-equal translation error at the frozen nominal 3.2-second endpoint (Table 1). Identity motion reaches 0.915 m on the same 11,219 recorded endpoints: a 51.9% reduction, improving 15/16 families, with conditional difference interval $[-.608, -.315]$ m. Against feature-matched direct regression (0.526 m), the reduction is 16.4%, also improving 15/16 families, with interval $[-.120, -.054]$ m. These comparisons establish utility beyond correcting large raw-odometry errors (4.156 m), on development-selected ETH3D data.

Table 1: ETH3D family-held-out results at $L = 32$, averaged over five seeds and 16 named families. Lower is better. The lower panel is a post-hoc attribution comparison with a separate shared initialization stream; identity is a parameter-free control.

Method	Translation (m)	Rotation (deg)
Raw exact motor product	4.156	36.03
Identity motion	0.915	45.07
GA leaf calibration	0.530	41.12
Calibrated direct GRU (original)	0.503	26.53
Contextual GA (original)	0.440	28.09
Constant motor gate	0.618	40.09
Feature-matched direct head	0.526	24.81
Feature-matched residual head	0.543	**24.65**
Contextual GA (matched features)	**0.436**	27.88
Mergeable GA gate (no GRU)	0.440	28.60

The improvement is not uniform: identity has lower translation error on camera-shake (.174 versus .639 m for mergeable GA). Direct and residual heads have lower mean rotation error. Conditional intervals do not account for overlapping training sets; fold-block sign-flip sensitivity for mergeable versus feature-direct gives $p = .0625$, not a population-level significance claim. The original contextual-versus-direct comparison is reported separately in Appendix F.

With matched raw-leaf context and exact endpoints, contextual GA reaches .436 m versus .526/.543 m for direct/residual heads, improving 14/16 families against each; the constant gate reaches .618 m. This supports input-conditioned motor blending within these budgets, not associativity alone. The scalar gate couples rotation and translation; these comparisons do not establish that coupling as the cause of the accuracy trade-off (Appendix H).

The mergeable gate uses 577 head parameters (857 including calibration), versus 10,689 for contextual GA. Its translation difference from the feature-matched contextual model is .0037 m, with interval $[-.0318, .0389]$ m; no non-inferiority margin was prespecified. Its paired-motor state also supports chunk merging and cache updates. At batch 256 on one CPU thread, throughput is 82.5k versus 38.0k recorded windows/s, a $2.17\times$ ratio. This profile includes calibration, both product branches and the head, but excludes odometry extraction (Appendix H). At batch one, mergeable/contextual latencies are 2.158/2.360 ms; the throughput ratio is specific to batched aggregation.

Frozen-model scope. Without refitting any of the 25 checkpoints, the mergeable model has the lowest mean translation and rotation among eight arms at $L = 4, 8$. At $L = 64$ (median 6.41 s), it reaches .602 m versus identity's 1.220 m on 15 available families. Restricting every horizon to the same 3,435 recorded starts retains the translation advantage through $L = 64$; sign-aligned readout changes primary error only from .4399 to .4396 m. Appendices I and L report all outcomes.

5.2 EVALUATION-PATH SENSITIVITY ON MOVIELENS

Table 2 reports the untouched confirmation result at the training length. The per-movie mode reaches 0.3992 accuracy. The bilinear and MLP left folds reach 0.4319 and 0.4325, respectively, exceeding this baseline. Changing only the reduction tree reduces bilinear accuracy by 8.37 percentage points on the right fold and 6.08 on the balanced tree; MLP accuracy decreases by 14.70 and 9.66 points.

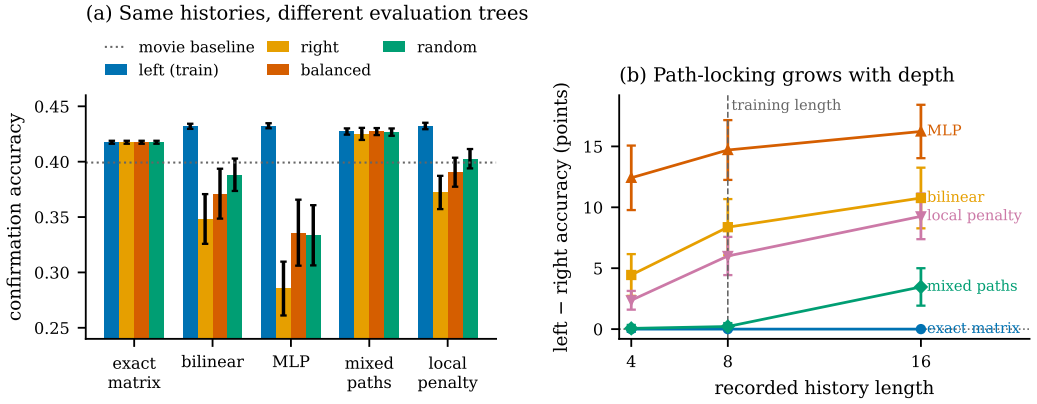

Figure 2: Five-seed confirmation on recorded MovieLens-1M histories. Error bars are standard deviations over training seeds. (a) Left-trained bilinear and MLP composers exceed the per-movie baseline on the trained path but lose accuracy under reparenthesization. Exact composition is invariant; mixed-path training reduces the trained-length accuracy difference between paths. (b) The left-to-right drop grows with history length for unconstrained and locally penalized models. The dashed line marks the training history length.

The 16 random trees also reduce accuracy for both models. Path sensitivity therefore occurs in both learned parameterizations on unchanged observations.

Table 2: Confirmation accuracy at history length 8, mean ± standard deviation over five training seeds. Parentheses are paired changes from the same arm's left fold. Random averages 16 order-preserving tree schedules.

Composer	Left (trained)	Right	Balanced	Random
Exact matrix	0.4175±.0014	0.4175 (+.0000)	0.4175 (+.0000)	0.4175 (+.0000)
Bilinear-residual	0.4319±.0023	0.3483 (−.0837)	0.3711 (−.0608)	0.3882 (−.0438)
Matched MLP	0.4325±.0023	0.2854 (−.1470)	0.3359 (−.0966)	0.3335 (−.0989)
Mixed paths	0.4272±.0028	0.4251 (−.0021)	0.4272 (−.0000)	0.4267 (−.0005)
Local penalty	0.4322±.0029	0.3722 (−.0600)	0.3905 (−.0418)	0.4027 (−.0296)
Train-only movie mode	0.3992	–	–	–

Bilinear right-fold accuracy decreases by 5.27 to 10.32 points across seeds, and balanced accuracy by 3.28 to 8.41. Every seed's user-clustered 95% interval excludes zero for both comparisons. Macro-F1, MAE and NLL also deteriorate (Appendix C).

Query-depth sensitivity. The full right fold also changes the depth at which the queried movie is processed. A post-hoc test of all 25 frozen checkpoints keeps the query at the final merge and reparenthesizes only the history. Bilinear and MLP accuracy still falls from .4319/.4325 to .4032/.3637: losses of 2.87/6.88 points, smaller than the full-tree losses but not eliminated by fixing query placement. The full-tree losses therefore combine history regrouping with changed query depth. Both history-right and history-balanced intervals exclude zero in every bilinear, MLP, and original-penalty seed (Appendix G).

5.3 Accuracy and Numerical Invariance of Exact Composition

The exact matrix arm changes no predicted rating across any tested tree. Its maximum absolute logit difference is 2.86×10^{-6} in float32 and 6.22×10^{-15} in float64 over all confirmation examples, lengths, seeds, and paths, below the prospectively fixed 10^{-5} tolerance. This is finite-precision agreement, not a claim of bitwise equality.

At $L = 8$, the matrix model reaches 0.4175 accuracy, above the 0.3992 per-movie baseline and 1.44 percentage points below the bilinear left fold. Associativity ensures agreement across paths for the learned function, independently of its task accuracy; it does not imply an optimal task parameterization.

5.4 Mixed-path training and associator regularization

Mixed-path training reaches 0.4272 accuracy and reduces the right and balanced changes to -0.0021 and approximately zero at $L = 8$, with a 0.47-point reduction from the bilinear left-fold score. The local penalty retains 0.4322 left-fold accuracy but leaves 6.00- and 4.18-point decreases. Training on multiple evaluation trees is more effective in this comparison than regularizing one primitive triple per batch.

The representations also change across paths. At $L = 8$, mean right/balanced dispersion is 0.648/0.544 for the bilinear arm, 1.019/0.709 for the MLP, 0.391/0.331 after mixed-path training, 0.522/0.426 with the local penalty, and below 2×10^{-7} for exact composition in float32.

Empirical consistency depends on sequence length. At $L = 16$, bilinear and MLP left-fold accuracies are 0.4332 and 0.4323, but their right-fold losses grow to 10.76 and 16.22 points (Figure 2b). Mixed-path training reaches 0.4295 on the left and loses 3.46 points on the right; balanced and random losses are smaller at 0.75 and 0.64 points. Exact composition remains unchanged. Length generalization along the trained path therefore does not imply evaluation-path generalization.

Intermediate-state regularization. A post-hoc development-selected weight sweep compares all-primitive and multilevel intermediate-state penalties at the same training budget. Three-seed development selection reduces the $L = 8$ left/right gap to 1.57 points with multilevel MSE (.4310/.4153), versus 3.33 for all-primitive MSE (.4316/.3983). Multilevel cosine reaches .4296/.4164, but at $L = 16$ falls from .4248 to .3697. Intermediate-state penalties reduce the observed path discrepancy but do not establish a global invariance guarantee. Appendix G reports all three penalty families, support diagnostics, and selection costs.

5.5 Exact composition and learned residuals on TUM

The TUM confirmation set contains 6,102 length-32 windows from 13 untouched trajectories. The invariant decoupled control has 14.28 cm median endpoint error. The learned residual reduces this error on its trained left fold: its five seed medians range from 5.44 to 8.28 cm, with an across-seed mean of 6.83 cm. Under the right fold, the same models and motions reach 36.53 cm. Every seed's 95% trajectory-clustered interval for right minus left mean error lies strictly above zero; the lower bounds range from 14.35 to 36.50 cm. Mean rotation error simultaneously rises from 8.91° to 81.49°. Thus path-locking is not specific to discrete recommendation, classification, or a learned readout.

Mixed-path training reduces the difference between paths (Figure 3), with left/right median translation errors of 5.22/4.58 cm. The local penalty reduces left-fold error to 6.86 cm but has 33.83 cm right-fold error, with all five clustered difference intervals above zero. The GA motor product reconstructs the endpoint independently of tree, with worst observed path discrepancy 2.28×10^{-15} m across every trajectory, length, and tree. This agreement follows from associativity of the geometric product used to represent physical motion composition.

6 Limitations

Outdoor generalization on KITTI.

A separately specified KITTI-trained replication uses unchanged LIBVISO2 stereo odometry (Geiger et al., 2011): sequences 00–06 train, 07–08 develop, and 09–10 form the study holdout. Five seeds and all eight arms are retained. On 245 overlapping 100 m segments, mergeable GA reaches 1.168 m translation error versus raw odometry's 1.062 m (10.0% worse), while mean rotation error falls from .01473 to .01365 degrees/m (7.4% better). Translation error is lower in only 2/5 seed-specific, two-drive means. Numerical tree and equivalent-matrix checks pass, but the prespecified predictive improvement criterion is not met. These are independently predicted segment-relative

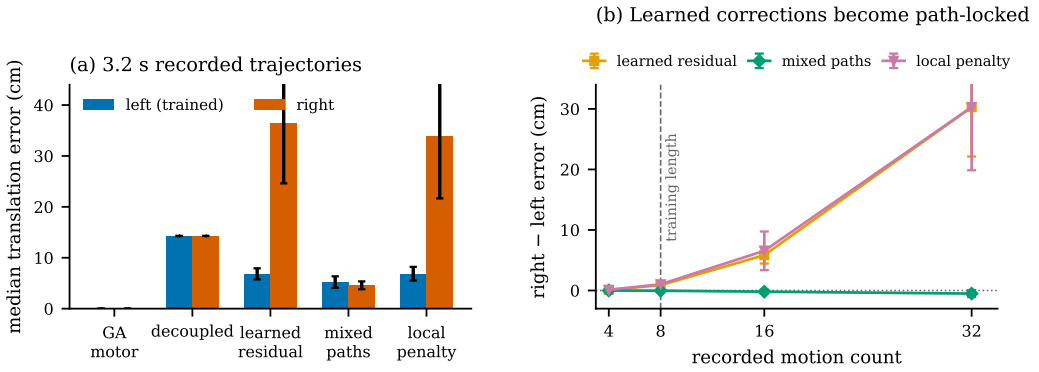

Figure 3: Confirmation on public TUM motion-capture trajectories. Error bars are standard deviations over five training seeds. (a) At 3.2 s, the learned residual improves the invariant but physically incomplete decoupled control on its trained fold but has higher error under reparenthesization. The GA motor product is exact in real arithmetic. Mixed-path training reduces the observed path discrepancy more than the local penalty. (b) Trajectory-equal right-minus-left mean error grows beyond the 0.8-second training horizon. No recorded motion or target changes between paths.

errors, not official KITTI trajectory scores. Appendix K reports all arms, both drives and the secondary horizons.

Task and comparator scope. MovieLens is diagnostic, not recommender state of the art: a simple history/movie-frequency mixture reaches .4297, above the exact arm (Appendix G). TUM verifies geometry, not learned perception; ETH3D tests a fixed odometry front end, not end-to-end SLAM. Learned operators could discover associativity; trained-path accuracy alone does not establish it.

Selection and confirmation boundary. MovieLens uses held-out confirmation. TUM's development-selected long-horizon hypothesis was tested once with fixed checkpoints on untouched trajectories. ETH3D's family protocol preceded fold-specific fitting but followed dataset inspection. Regularization, query-last, identity, feature-matching and horizon controls are post-hoc. Local hashes establish file integrity, not independently timestamped preregistration or inspection history.

Dataset and horizon scope. The studies cover recorded recommendation histories and rigid motions. At $L = 128$ on ETH3D (12.79 s, 13 families), the mergeable model's translation difference from identity is $-.132$ m, with exploratory conditional interval $[-.582, .383]$ m; rotation error is higher. These additional horizons reuse existing data and do not establish independent generalization. Oxford did not satisfy its development criteria and has unresolved frame, target-quality and continuity issues. Its apparent improvements are excluded; confirmation remains unopened (Appendix J). No simulated or generated observations are used.

Numerical precision and computational scope. Associativity guarantees equality in exact arithmetic, not bitwise floating-point agreement. The contextual GRU remains sequential; the GRU-free paired state admits logarithmic algebraic reduction depth. The measured throughput improvement reflects removal of recurrent computation, not parallel-tree acceleration or end-to-end visual SLAM throughput. Reparenthesization preserves operand order; no result permits reordering recorded motions.

7 Conclusion

Exact PGA motor products combine learned calibration with a compact associative summary. On family-held-out ETH3D, the mergeable gate more than halves translation error relative to identity motion and reduces measured computation relative to the contextual variant. MovieLens and TUM tests demonstrate prediction sensitivity to reparenthesization in learned operators. Predictive improvements are specific to datasets, horizons and comparators; the paired-state construction provides consistent tree reduction, ordered chunk merging and cached recomputation.

REPRODUCIBILITY STATEMENT

Code and retained experimental records are available at https://github.com/conf-anonymous/path-locked-composition-artifact. The anonymous artifact contains checksum-verified loaders, deterministic user/trajectory splits, model and intervention implementations, frozen checkpoints, development and confirmation records, clustered bootstraps, exactness audits, and scripts that fail if the frozen gates are not met. MovieLens, TUM, ETH3D and KITTI source data must be obtained from their official providers under the applicable terms; original dataset files are not redistributed. The artifact retains all KITTI arms and horizons, including its failed utility criterion, and separately records the excluded Oxford attempts and reference investigation. Record/hash audits are distinguished from sensor extraction and fresh model replay, which require the original data and stated environments.

AI USE STATEMENT

The authors established the research objectives and conducted the development of the methods and investigation of hypotheses, retaining control over core production, scientific judgments and methodological decisions. Generative AI tools assisted code generation and testing, data processing, experiment automation, hypothesis refinement and experimental-control design, analysis of run records, proofreading, and artifact preparation. Agentic workflows supported concurrent experimentation with the aim of reducing implementation and evaluation time and cost; this study does not quantify those workflow savings. AI-generated outputs were treated as untrusted and incorporated only after author review, verification and any necessary corrections or adaptations. Empirical observations were drawn from public recorded datasets, and reported experimental results were obtained by executing the research implementations. No synthetic observations or targets were generated or used. The artifact includes implementations, tests, frozen checkpoints, retained experimental records and automated consistency checks. Reported numerical results were checked against retained run records using automated audits and author review. The authors bear sole responsibility for the research design and conduct, the correctness of the implementations and mathematical arguments, the analysis and interpretation of results, and all claims, conclusions and decisions in the manuscript.

REFERENCES

Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J. Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, and Sam Whittle. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8(12): 1792–1803, 2015.

Anonymous. On why form shapes reason: Structuring latent program networks with category-theoretic constraints. ICLR 2026 OpenReview manuscript, 2026. URL `https://openreview.net/forum?id=BpQIvcNvnp`.

Guy E. Blelloch. Prefix sums and their applications. Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University, 1990.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K. Gupta. Clifford neural layers for PDE modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Johann Brehmer, Pim de Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *Symposium on Operating Systems Design and Implementation (OSDI)*, 2004.

Leo Dorst, Daniel Fontijne, and Stephen Mann. *Geometric Algebra for Computer Science: An Object-Oriented Approach to Geometry*. Morgan Kaufmann, 2007.

Ekaterina Filimoshina and Dmitry Shirokov. GLGENN: A novel parameter-light equivariant neural networks architecture based on clifford geometric algebras. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 17153–17188. PMLR, 2025.

Andreas Geiger, Julius Ziegler, and Christoph Stiller. StereoScan: Dense 3D reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium (IV)*, 2011. URL https://www.cvlibs.net/software/libviso/.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. URL https://www.cvlibs.net/datasets/kitti/.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

Charles Gunn. Projective geometric algebra: A new framework for doing euclidean geometry. *arXiv preprint arXiv:1901.05873*, 2019.

Kelvin Guu, John Miller, and Percy Liang. Traversing knowledge graphs in vector space. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.

David Hestenes and Garret Sobczyk. *Clifford Algebra to Geometric Calculus: A Unified Language for Mathematics and Physics*. D. Reidel, 1984. doi: 10.1007/978-94-009-6292-7.

W. Daniel Hillis and Guy L. Steele. Data parallel algorithms. *Communications of the ACM*, 29(12): 1170–1183, 1986.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020.

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Transactions on Graphics*, 27(4), 2008. URL https://users.cs.utah.edu/~ladislav/kavan08geometric/kavan08geometric.html.

Richard E. Ladner and Michael J. Fischer. Parallel prefix computation. *Journal of the ACM*, 27(4): 831–838, 1980.

Wenjie Li, Wasif Naeem, Jia Liu, Dequan Zheng, Wei Hao, and Lijun Chen. PDQ-Net: Deep probabilistic dual quaternion network for absolute pose regression on $SE(3)$. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180, pp. 1118–1127. PMLR, 2022.

Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. doi: 10.1177/0278364916679498.

Will Maddern, Geoffrey Pascoe, Matthew Gadd, Dan Barnes, Brian Yeomans, and Paul Newman. Real-time kinematic ground truth for the oxford robotcar dataset. *arXiv preprint arXiv:2002.10152*, 2020.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. In *International Conference on Learning Representations (ICLR)*, 2018.

Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. Compositional vector space models for knowledge base completion. In *Association for Computational Linguistics (ACL)*, 2015.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning (ICML)*, 2011.

Euan Ong and Petar Veličković. Learnable commutative monoids for graph neural networks. In *Proceedings of the First Learning on Graphs Conference*, volume 198, pp. 43:1–43:22. PMLR, 2022.

Oxford Robotics Institute. Robotcar dataset sdk. `https://github.com/ori-mrg/robotcar-dataset-sdk`, 2022. Reference audit used commit 16ce3329223ca418fe5106277b91aea8d9b672b2.

Valentin Peretroukhin and Jonathan Kelly. DPC-Net: Deep pose correction for visual localization. *IEEE Robotics and Automation Letters*, 3(3):2424–2431, 2018. doi: 10.1109/LRA.2017.2778765.

Charles Pert, Dalal Alrajeh, and Alessandra Russo. Length generalization with log-depth recurrent units. *arXiv preprint arXiv:2605.26035*, 2026.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Association for Computational Linguistics (ACL)*, 2020.

RobotCar Dataset SDK issue tracker. Issue 40: The extrinsic of rtk is not consistent with the picture in ijrr paper. `https://github.com/ori-mrg/robotcar-dataset-sdk/issues/40`, 2021. User-reported frame discrepancy; not an official calibration correction.

David Ruhe, Jayesh K. Gupta, Steven de Keninck, Max Welling, and Johannes Brandstetter. Geometric clifford algebra networks. In *International Conference on Machine Learning (ICML)*, 2023.

Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 134–144, 2019. doi: 10.1109/CVPR.2019.00022.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

Xiatao Sun, Yuxuan Wang, Shuo Yang, Yinxing Chen, and Daniel Rakita. Hybrid diffusion policies with projective geometric algebra for efficient robot manipulation learning. *arXiv preprint arXiv:2507.05695*, 2025.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations (ICLR)*, 2019.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning (ICML)*, 2016.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations (ICLR)*, 2015.

Morris Yau, Sharut Gupta, Valerie Engelmayer, Kazuki Irie, Jacob Andreas, and Stefanie Jegelka. Sequential–parallel duality in prefix-scannable models. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2506.10918`.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018.

A Proofs and formal details

A.1 Proof of Theorem 1

If all trees agree, apply the claim to the two trees over (a, b, c) to obtain $\mu(\mu(a, b), c) = \mu(a, \mu(b, c))$ for arbitrary $a, b, c \in \mathcal{C}$. Conversely, assume associativity on $\mathcal{C}$. A tree rotation replaces a subtree $\mu(\mu(a, b), c)$ by $\mu(a, \mu(b, c))$, where closure ensures that the three subtree values lie in $\mathcal{C}$. The rotation preserves the output. Every ordered full binary tree can be transformed into the left comb through a finite sequence of such rotations, so any two trees have the same output. For the event-only version, each element of the algebraically generated closure $\mathcal{C}(E)$ is realized by a finite event tree. Given arbitrary a, b, c in that closure, graft their realizing trees into $((a, b), c)$ and $(a, (b, c))$. The resulting event trees have identical ordered leaves, so event-level tree invariance equates their outputs and proves associativity on $\mathcal{C}(E)$. Prediction-level equality under a non-injective g does not imply this result: a readout can discard changing state components or saturate. Agreement of predictions is therefore not a substitute for a state-level associativity guarantee. The separate dispersion diagnostic measures internal changes.

A.2 Proof of Proposition 1

A balanced full tree is one of the trees covered by Theorem 1; its longest root-to-leaf path is $\lceil \log_2 L \rceil$ when balanced as evenly as possible. For chunks, first reduce each contiguous subsequence and then treat its result as one subtree; closure and Theorem 1 equate the merged result with the full left fold. A balanced cache over N chunk summaries has height $O(\log N)$. Replacing one cached summary changes only its ancestors. If the replacement chunk contains K leaves, recomputing its summary first takes $O(K)$ work; the ancestor bound does not include that cost. A height-balanced insertion scheme, such as an AVL tree with subtree products, also supports append with $O(\log N)$ worst-case product updates, including rebalancing. These bounds assume constant-size summaries and fixed leaf encoding parameters. They apply to one cached endpoint, not materializing all L prefix outputs, which requires $\Omega(L)$ output work. The recorded cache experiment checks fixed-tree replacement, not dynamic insertion or GRU recomputation.

A.3 Support of primitive associator regularization

Theorem 1 requires associativity on every reachable internal state. A loss supported only on E^3 constrains the operation on primitive triples, but a four-leaf tree already evaluates triples involving $\mu(E, E)$. Unless the penalty distribution covers the reachable closure, or the parameterization enforces a global identity, zero sampled loss does not imply path invariance. This is a support statement rather than a claim that local regularization is necessarily ineffective.

B Complete experimental protocol

B.1 Data provenance and eligibility

The official `ml-1m.zip` archive has MD5 `c4d9eecfca2ab87c1945afe126590906` and SHA-256 `a6898adb50b9ca05aa231689da44c217cb524e7ebd39d264c56e2832f2c54e20`. All files pass archive integrity checking. We parse only `ratings.dat`; demographic attributes are unused. Each row supplies user ID, movie ID, 1–5 rating, and timestamp. The archive's 1,000,209 rows are preserved. There are 790,115/770,687/731,831 eligible training endpoints, 89,594/87,210/82,442 development endpoints, and 96,340/93,992/89,296 confirmation endpoints at $L = 4/8/16$.

User assignment is `int(SHA256("ml1m:" + user)[:8],16) mod 10`: buckets 0–7 train, 8 development, and 9 confirmation. This rule is deterministic and independent of ratings. The archive is public recorded data. Random initialization, minibatch sampling over observed endpoints, bootstrap resampling, and sampled binary evaluation schedules create no observations or labels.

B.2 Optimization and parameter accounting

Movie and rating embeddings are summed. The rating vocabulary has six entries: mask plus the five observed values. The exact arm applies $I + 0.05\tanh(z)$ elementwise after reshaping. Learned arms apply event layer normalization. All arms share final layer normalization and a linear five-class head. Cross-entropy is the sole task loss. The local-penalty arm adds the mean squared coordinate difference between the two reductions of one consecutive encoded triple at a batch-shared random position, including the masked query when it lies in the triple.

The exact, bilinear/mixed/penalty, and MLP models contain 33,494, 34,241, and 34,221 parameters. Their composition rules contain 0, 747, and 727 trainable parameters. The exact arm is not exactly parameter-matched and uses a different event parameterization; these architectural differences preclude attributing its task-accuracy gap solely to associativity. Its count includes an unused 18-parameter event-normalization module retained by the shared constructor.

C Additional confirmation metrics

Table 3: Confirmation metrics at $L = 8$, means over five seeds. F1 is macro-F1; MAE is in rating stars. Lower is better for MAE and NLL.

Arm	Path	Accuracy	F1	MAE	NLL
Exact	left/right/tree/random	.4175	.3485	.7635	1.3028
Bilinear	left	.4319	.3603	.7376	1.2749
	right	.3483	.2977	.9496	1.4725
	balanced	.3711	.3296	.9058	1.4141
	random	.3882	.3402	.8581	1.3713
MLP	left	.4325	.3578	.7323	1.2742
	right	.2854	.2167	1.1783	1.7178
	balanced	.3359	.2742	.9976	1.4917
	random	.3335	.2684	1.0106	1.5155
Mixed	left	.4272	.3546	.7440	1.2816
	right	.4251	.3462	.7515	1.2871
	balanced	.4272	.3514	.7462	1.2826
	random	.4267	.3505	.7454	1.2833
Penalty	left	.4322	.3605	.7371	1.2741
	right	.3722	.3207	.8928	1.3975
	balanced	.3905	.3454	.8603	1.3641
	random	.4027	.3505	.8236	1.3338

The train-only global-mode accuracy is 0.3441. The train-only per-movie mode is 0.3992, with macro-F1 0.3101 and MAE 0.7813. The train-only per-movie empirical distribution has NLL 1.3319. Learned left-fold models improve all four metrics.

D PER-SEED CONFIRMATION INTERVALS

Table 4: Bilinear paired accuracy changes at $L = 8$ with 95% user-clustered bootstrap intervals. Every confirmation interval excludes zero.

Seed	Right minus left	Balanced minus left
0	$-.0969$ $[-.1048, -.0891]$	$-.0744$ $[-.0814, -.0673]$
1	$-.0653$ $[-.0697, -.0607]$	$-.0328$ $[-.0369, -.0287]$
2	$-.1003$ $[-.1084, -.0921]$	$-.0718$ $[-.0787, -.0651]$
3	$-.1032$ $[-.1118, -.0951]$	$-.0841$ $[-.0921, -.0760]$
4	$-.0527$ $[-.0566, -.0486]$	$-.0411$ $[-.0455, -.0367]$

Development preceded confirmation and showed the same ordering. At $L = 8$, mean left accuracies were 0.4225 (exact), 0.4400 (bilinear), 0.4419 (MLP), 0.4361 (mixed), and 0.4403 (penalty), against a 0.4117 per-movie baseline. Bilinear right/balanced changes were $-0.0872/-0.0639$; MLP $-0.1494/-0.0999$; mixed $-0.0025/+0.0000$; and penalty $-0.0640/-0.0450$. Every per-seed bilinear interval excluded zero for both primary paths before confirmation was opened.

E TUM PROVENANCE, PROTOCOL HISTORY, AND INTERVALS

We downloaded the 47 ground-truth text files listed by the official TUM RGB-D benchmark and recorded their SHA-256 checksums. Each source row contains a timestamp, translation, and quaternion. Eight repeated timestamp strings in `freiburg2_xyz` were discovered during loader validation, before any training or outcome inspection; the frozen amendment retains the first row in immutable file order. Quaternions receive only a sign-continuity convention, which represents the same rotation. No pose is averaged, interpolated, or altered. At $L = 8$, the split has 13,682/6,255/6,414 windows in train/development/confirmation; at $L = 32$, confirmation has 6,102 windows.

The TUM evaluation followed three documented protocols. Experiment 017 prospectively tested unrestricted eight-coordinate bilinear and MLP laws; its development gate failed because their trained-fold error was worse than the invariant decoupled control. Confirmation remained unopened. Experiment 018 froze a residual-over-control parameterization, but its original development gate also failed: it demanded improvement from every bilinear seed and four of five MLP seeds already at 0.8 s, plus a balanced-tree condition. Its development result motivated a 3.2-second evaluation hypothesis. Experiment 019 froze that endpoint and gate, reused the unchanged Experiment 018 checkpoints without retraining, and then evaluated the 13 confirmation trajectories once. We therefore call this result *development-selected and confirmation-frozen*, not preregistered before development. All five conjunctive conditions passed.

Training objective. For decoded prediction $(\hat{q}, \hat{t})$ and recorded endpoint (q, t), with unit quaternions, the task loss is

$$\mathcal{L}_{\mathrm{TUM}} = \mathbb{E}_{\mathrm{batch}} \left[\left\| (\hat{t} - t)/s \right\|_2^2 + 2 \left(1 - \min\{1, \langle \hat{q}, q \rangle^2\} \right) \right]. \tag{5}$$

The fixed scale s is the median training-set $L = 8$ endpoint displacement, with a .001 m floor. Squaring the quaternion dot product removes the sign ambiguity; capping it at one prevents negative numerical rotation loss. The local-penalty arm adds 0.1 times the batch/coordinate mean squared associator of a sampled consecutive primitive triple. All residual arms use 5,000 AdamW updates, batch 256, learning rate .002, weight decay 10^{-5} and gradient clipping at one. The task loss differs from the robust smooth-L_1 head loss used on ETH3D.

Table 5: TUM confirmation at $L = 32$. Errors and trajectory-clustered 95% intervals are in cm. Δ is trajectory-equal mean right-minus-left error; the path columns are endpoint-error medians.

Arm/seed	Left median	Right median	Mean Δ [95% interval]
Bilinear 0	8.28	38.55	28.95 [22.85, 34.87]
Bilinear 1	7.44	53.68	41.65 [36.50, 47.38]
Bilinear 2	6.22	39.90	34.57 [29.48, 40.04]
Bilinear 3	5.44	26.97	25.67 [18.65, 32.52]
Bilinear 4	6.79	23.58	20.56 [14.35, 26.82]
Mixed-path mean	5.22	4.58	-0.51
Local-penalty mean	6.86	33.83	30.30
Decoupled control	14.28	14.28	0.00
GA motor	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$

Across seeds, mixed-path right-minus-left intervals all include zero. Every local-penalty interval indicates higher right-fold error, with lower bounds between 13.84 and 39.64 cm. The exact motor's maximum decoded path discrepancy over all lengths and evaluated trees is 2.28×10^{-15} m.

F ETH3D PROVENANCE, SELECTION HISTORY, AND FAMILY RESULTS

The official ETH3D site identifies 56 real training sequences recorded inside a motion-capture system and five additional Structure-from-Motion sequences (Schöps et al., 2019); we use exactly the former. The downloader verifies 112 official modality archives. Open3D odometry yields 17,974 valid adjacent pairs and one recorded front-end failure. At $L = 32$, 11,219 windows from 53 sequences are evaluable. `plant_1`, `plant_3`, and `plant_4` contain no qualifying contiguous 32-leaf run and contribute no evaluation window; the plant family remains represented by six sequences.

The ETH3D evaluation followed four documented protocols. Experiment 021 prospectively froze a sequence-hash split and six-arm gate. Its development conjunction failed: unconditional GA leaf calibration reduced aggregate error but increased error on accurate sequences, and confirmation remained unopened. Experiment 022 then logged a new contextual-gating hypothesis. After five-seed development, its checkpoints and criteria were fixed and hashed; its one-time sequence-held-out confirmation passed. Audit then identified related named families across that split. Experiment 023 therefore froze the stricter family partition before grouped-fold training and retrained both stages within each fold. Experiment 024 froze and ran the two-stage budget-matched direct control. Each protocol retains its original acceptance decision. The grouped result is a robustness evaluation selected after earlier ETH3D inspection, not untouched-dataset confirmation.

Family membership is determined from official name prefixes, retaining all lighting and motion variants together. A deterministic size-balanced assignment holds out: (0) mannequin/kidnap/repetitive; (1) plant/planar; (2) sofa/desk/ motion; (3) Einstein/ceiling/vicon-light; and (4) cables/camera-shake/table/ large-loop/reflective. Table 6 reports the resulting family-equal primary metric. Each entry first averages windows within sequence, then sequences within family, then five seeds for learned methods.

Table 6: ETH3D $L = 32$ translation error (m) for wholly held-out capture families. Calibrated direct matches stages and budgets, not input features.

Family	Raw	Context	Direct	Family	Raw	Context	Direct
Cables	10.260	.932	.950	Camera shake	23.671	.657	.617
Ceiling	3.119	.704	.880	Desk	.227	.206	.293
Einstein	3.021	.391	.463	Kidnap	1.221	.380	.416
Large loop	3.524	.538	.531	Mannequin	16.419	.797	.625
Motion	.912	.724	.700	Planar	.041	.050	.235
Plant	1.157	.451	.435	Reflective	2.083	.260	.322
Repetitive	.059	.066	.167	Sofa	.555	.597	.701
Table	.092	.067	.185	Vicon light	.129	.219	.528

The four families not improved over raw by contextual GA are planar, repetitive, sofa, and vicon-light; their absolute increases are .009, .007, .042, and .090 m. The prespecified family bootstrap retains all 16 families. For contextual GA minus raw, its interval is $[-7.200, -0.980]$ m; for contextual GA minus the budget-matched calibrated direct GRU it is $[-0.116, -0.013]$ m. These are translation results. The matched direct GRU has

lower rotation error than contextual GA (26.53° versus 28.09°), while both improve raw (36.03°). Contextual GA reduces mean translation error from the direct GRU's .503 m to .440 m, improving 11/16 families; it improves 12/16 families over raw.

Frozen-output robustness and failure analysis. These diagnostics were specified after the model freeze and do not replace the prespecified interval. Contextual-minus-direct translation has family median $-.0669$ m. All 16 leave-one-family-out means remain negative; the least favorable is $-.0468$ m after removing vicon-light. Removing the three largest raw-error families (camera-shake, mannequin, and cables) leaves a $-.0926$ m mean against the matched direct control and a $-.884$ m mean against raw. Enumerating all 2^{16} family-level sign assignments gives one-sided $p = .0174$ and two-sided $p = .0348$. Raw error correlates strongly with improvement over raw (Spearman $\rho = .935$), but negatively with advantage over the matched direct model ($\rho = -.691$): both learned methods reduce the largest odometry errors, and direct has lower translation error on camera-shake and mannequin. Contextual GA has lower rotation error in 10/16 families, but camera-shake contributes a 65.7° increase and the family-equal mean is 1.565° higher. The aggregate improvement relative to direct is therefore translation-specific.

Conditional uncertainty and fold dependence. The intervals resample 16 named families while retaining all five fitted seeds. They condition on the learned models and do not include retraining uncertainty; fold training sets overlap, and families in one fold share a fitted model. Name-based grouping does not independently verify distinct physical environments. For contextual minus direct, the five fold-mean translation differences are $+.0114, -.0848, -.0555, -.1858, -.0303$ m. Flipping entire fold blocks, retaining family-equal weighting, yields two-sided $p = .125$ over 32 assignments. This is sensitivity to the assumed unit of exchangeability, not a corrected population p-value; neither family nor fold sign enumeration proves its own exchangeability assumptions. The reported point estimates and per-family errors therefore do not establish population-level cross-dataset superiority.

Identity-motion utility control. The fixed prediction is the identity motor: zero translation and rotation, evaluated against the same recorded reference endpoints without fitting. It has .915057 m family-equal translation error and 45.072° rotation error, compared with .439936 m and 28.094° for contextual GA. Translation improves in 15/16 families and in all five fold aggregates; conditional family-bootstrap difference is $-.475121$ m, interval $[-.615718, -.309058]$ m. Camera-shake is the exception (.174166 m identity versus .656838 m contextual). Only 5.95% of windows have recorded displacement at most .15 m. This control establishes improvement relative to a stationary prediction on predominantly nonstationary recordings. It does not isolate the contribution of gating or compare against equivalent rigid-motion coordinates.

Exact computation and projection domain. The contextual gate is

$$\alpha = \sigma(W_2\ \mathrm{GELU}(W_1[\overline{\mathrm{BiGRU}(M_{1:L})}; M_{\mathrm{raw}}; M_{\mathrm{cal}}] + b_1) + b_2).$$

The GRU has hidden size 32 in each direction and the gate has hidden width 32. The final gate weights start at zero, bias -2. The leaf calibrator is an $8 \to 16 \to 8$ GELU residual network with zero-initialized output, projected to a motor. Both stages use 5,000 AdamW updates, batch 256, weight decay 10^{-4}, clipping one, and learning rates .002/.001. Calibration samples $L = 8$ training windows; the primary $L = 32$ stage samples sequences equally then their valid windows. With translation scale s fixed to the median training $L = 8$ reference displacement (floor .001 m), the primary loss is the sum of coordinatewise smooth-L_1 terms for $(\hat{t} - t)/s$, plus $2(1 - \langle \hat{q}, q \rangle^2)$. The calibration stage uses squared coordinate translation error instead of smooth-L_1, with the same rotation term; quaternion dot squares are capped at one for numerical safety. Every compared head shares the same fold-specific training scale and rotation weight, including the original direct control and feature-matched heads. The comparison does not use different specified task-loss weights for different heads. The same objective can nevertheless yield different translation/rotation trade-offs across parameterizations; we have not performed a loss-weight sensitivity study and do not attribute the trade-off to a single cause.

For coordinates (r, d), projection uses $n = \|r\|$, $q = r/n$, and $d' = d/n - \langle q, d/n \rangle q$. This yields a unit motor only for $n > 0$. The implementation clamps n below 10^{-8} for arithmetic safety; it is not a proof of unit norm at a zero real quaternion. Reference quaternions use sign-continuous representatives, but the original raw/calibrated endpoint blend does not align antipodal signs. Thus the blended prediction is representation-dependent and can be singular in principle. We make no global nondegeneracy claim. Appendix L gives a separate sign-aligned readout audit; the original model and all its results remain unchanged. The product-tree guarantee holds with the same full-sequence context and a well-defined deterministic final projection, not by merging independently gated chunk predictions. Recorded $L = 32$ windows span 3.1334–3.2440 seconds (median 3.2071); 3.2 seconds throughout denotes the nominal sampling horizon.

Post-freeze systems profile. Fold-0/seed-0 models were deterministically retrained and had to reproduce stored family metrics before timing (maximum discrepancy 4.28×10^{-9}). On macOS ARM with PyTorch

2.12, CPU, one thread, and recorded length-32 ETH3D operands, median latency at batches 1/32/256 was 2.456/3.069/6.719 ms for contextual GA and .241/.729/3.734 ms for the calibrated direct GRU. Balanced exact motor reduction took 1.127/1.173/1.454 ms. Across the original 25 runs, median primary training time was 99.05 s for contextual GA and 79.32 s for the direct control, plus the shared 18.14 s calibrator. These implementation-level measurements show bounded absolute cost but no speed advantage for the original contextual model over that direct control; the GRU-free variant is profiled separately in Appendix H.

G POST-HOC MOVIELENS CONTROLS

These analyses followed inspection of the original confirmation outcomes. The original checkpoints and results are unchanged; these analyses do not constitute a new independent confirmation. Every example and label is an unchanged official recorded rating. Query-last evaluation includes all 93,992 length-eight endpoints and all five seeds per original arm. Only the history subtree changes; its result is combined with the masked query at one final merge.

Table 7: Post-hoc query-last sensitivity, five-seed mean accuracy at $L = 8$.

Original arm	Left	History-right	History-balanced
Exact matrix	.4175	.4175	.4175
Bilinear-residual	.4319	.4032	.4198
MLP	.4325	.3637	.3970
Mixed paths	.4272	.4239	.4272
One-triple penalty	.4322	.4096	.4228

Intermediate-state penalties and development-only selection. Experiment 027 fixes the original bilinear-residual model and 3,000-update budget. Primitive MSE averages coordinatewise associators over *all* consecutive primitive triples. Multilevel MSE includes every consecutive triple at each adjacent-pair reduction level, averages within levels and then across levels, and retains odd last states. A third variant uses squared cosine distance, inspired by the intermediate-state penalty of Pert et al. (2026); it is not a reproduction of that paper's architecture. The final common weight grid is $\{.1, 1, 10, 100, 1000, 10000\}$. We select each variant by the mean of five path accuracies (left/right/balanced/history-right/history-balanced) across *three development seeds*, retaining collapsed fits. Selected weights are 10 (primitive MSE), 1 (multilevel MSE), and 1000 (multilevel cosine); no held-out score selects a weight or checkpoint. Each is evaluated at five seeds.

An initial seed-zero grid $\{.1, 1, 10\}$ selected its upper boundary; expansion revealed seed-sensitive collapse at larger primitive-MSE weights. We then fixed the common grid and three-seed selection above before the remaining fits. Earlier selections/results are retained as historical, not alternative confirmation claims. Across all stages, 66 distinct fits consumed 1195.6 summed per-run training wall seconds, excluding evaluation, with one CPU thread per worker and two concurrent workers. This post-hoc development search is not exhaustive optimization.

Table 8: Primitive and intermediate-state penalties: means over five seeds. A_E is primitive coordinate MSE; A_{multi} averages all sampled reduction levels.

Penalty	L	Left	Right	Balanced	A_E	A_{multi}
All-primitive MSE	8	.4316	.3983	.4166	.000339	.013777
Multilevel MSE	8	.4310	.4153	.4243	.001820	.002328
Multilevel cosine	8	.4296	.4164	.4233	.002718	.003093
All-primitive MSE	16	.4341	.3658	.4003	.000314	.016256
Multilevel MSE	16	.4304	.3672	.3967	.001722	.003842
Multilevel cosine	16	.4248	.3697	.3906	.002644	.005176

At $L = 8$, the selected primitive penalty achieves a smaller primitive associator than the selected multilevel penalties, yet a larger multilevel discrepancy. In the retained common-weight-10 five-seed comparison, primitive and multilevel MSE have similar primitive errors (.000339/.000374), but multilevel errors .013777/.000387, differing by about $36\times$. This supports the relevance of intermediate-state coverage without isolating it from every optimization effect. The trained-length improvement does not extend uniformly to longer compositions: even small average sampled associators need not control accumulated errors or downstream readout sensitivity. All per-seed path accuracies and user-clustered intervals are retained.

History-aware prediction baseline. We also mix a train-only Laplace-smoothed per-movie rating distribution with the empirical distribution of the eight preceding recorded ratings. Development selects history weight .25 from $\{0, .25, .5, .75, 1\}$; confirmation accuracy is .4297, versus .3947 for history mode alone. This simple baseline nearly matches the learned left folds and exceeds the exact matrix arm. Accordingly MovieLens supports a controlled path-locking diagnosis, not superior recommendation quality or a uniquely beneficial learned use of history.

H ETH3D FEATURE ATTRIBUTION AND MERGEABLE-STATE ABLATION

Experiments 028–029 were specified after adversarial review, using the same already-inspected ETH3D recordings and fixed five-fold, five-seed design. Every leaf calibrator is reproduced and its metrics checked against the original held-out values before the new primary fit. No outer-fold score chooses an arm, seed, training duration or hyperparameter. All arms and outcomes are retained. These are grouped robustness ablations, not independent confirmation tests.

Feature-matched heads and output constraints. All three GRU-based heads encode the raw leaves with a width-32 bidirectional GRU and concatenate its pooled outputs with both exact endpoints. A common 80-to-32 hidden layer precedes either one mixing coefficient (contextual) or eight motor coordinates (direct/residual). GRU and first-head-layer initialization are identical within a seed. Direct regression starts at the identity motor, avoiding a random near-zero real quaternion; residual regression starts at zero over the same initial raw/calibrated blend as the contextual gate. The constant control learns a single sigmoid coefficient without a GRU. All heads use the same 5,000 updates, sequence-balanced batches, robust loss, optimizer and seed streams, with fixed calibration. Thus information and training are matched for the GRU heads, while output constraints, parameter counts and initialization priors remain explicitly different. The direct/residual heads already benefit from exact GA endpoint features; they are not non-associative binary composers.

Mergeable paired-motor state. The GRU-free head is a $16 \to 32 \to 1$ GELU/sigmoid network with 577 parameters (857 including the shared 280-parameter calibrator). Its only inputs are the two exact motor products. The paired state S has componentwise geometric-product merge, an identity pair and closure; Theorem 1 therefore applies to the *complete summary*. The projected final motor is a readout of S, not the object to merge. Pointwise leaf calibration, parallel paired products and a fixed-size head permit logarithmic algebraic depth; a streaming update needs two motor products plus the readout. Changing all calibrator parameters still requires updating all affected leaves, not one logarithmic cache edit. This ablation uses the same fixed training budget and all 25 original fold/seed calibrators, with no held-out tuning.

Its family-equal error is .439929 m/28.600°, compared with .436224 m/27.882° for the feature-matched contextual model. The translation difference interval $[-.0318, .0389]$ m includes meaningful advantages in either direction; we did not prespecify a non-inferiority margin. Its translation error is lower than the feature-matched direct/residual heads in 15/16 families each, with conditional translation intervals $[-.1201, -.0544]$/$[-.1382, -.0678]$ m, while both heads retain lower mean rotation error. All per-seed/fold/family errors and conditional intervals are included in the artifact, without an independent-population claim. Original protocol intervals retain their frozen bootstrap draws; the new control/identity intervals use 10,000 family resamples with seed 280906. For the feature-matched contextual head, translation difference intervals against direct and residual heads are $[-.142, -.041]$ and $[-.159, -.056]$ m, with lower error in 14/16 families against each.

Across all 25 feature-matched contextual fits, float64 left/right/balanced/chunked product trees differ by at most 6.53×10^{-14} m and 4.34×10^{-15} rad; the mergeable gate reaches 1.29×10^{-13} m and 1.06×10^{-14} rad. Minimum un-clamped real-quaternion norms on recorded predictions are .3895 for contextual and .4935 for mergeable; across the four attribution heads the minimum is .0590. No singularity is observed on these recordings; this does not establish global nonsingularity. All 25 calibrators reproduce their original metrics exactly. Experiments 028 and 029 respectively sum to 4542.0 and 150.2 per-run training wall seconds (the former includes calibration); runs use two workers, one CPU thread each.

Recorded-data cache verification. On all 2,028 fold-zero held-out length-32 windows, cached endpoint features and the original online contextual computation give identical float32 coordinates with the same parameters. A separate balanced motor-product-cache audit replaces predicted leaves by their learned calibration of those same recordings, without changing sensor observations or reference labels. Updating all 32 positions uses 160 product calls per branch after 31 initialization calls, versus 992 calls for 32 complete recomputations. Every update equals full reduction in coordinates in this audit; maximum decoded translation discrepancy is zero and angular discrepancy 1.57×10^{-16} rad. These counts audit one product branch; the paired state maintains two such branches. They are numerical checks, not measured wall-clock speedups or a claim that the original GRU context can be cached this way.

Same-input implementation profile. After training completed, fold-zero/seed-zero online predictors were measured with 20 warmups and 100 timed calls per batch size on recorded odometry leaves, using macOS ARM64, PyTorch 2.12 and one CPU thread. Both include leaf calibration, both product branches and the head, but exclude RGB-D odometry extraction. Contextual versus mergeable median latency is 2.360/2.158 ms at batch one, 3.027/2.276 ms at batch 32, and 6.728/3.103 ms at batch 256. At batch 256 this is 38,050/82,491 windows per second, a $2.17\times$ throughput ratio. Removing the GRU reduces head parameters from 10,689 to 577 (94.6%); including the calibrator the counts are 10,969 and 857. These are measured implementation tradeoffs on one machine, not a parallel-algorithm speedup or a deployment guarantee. Cached and online predictions match exactly for both fitted models.

I FROZEN-CHECKPOINT HORIZON AND SCOPE AUDIT

Experiment 030 evaluates every fixed Experiment 028/029 fold/seed checkpoint at $L \in \{4, 8, 16, 32, 64, 128\}$, with no new fit, selection or loss adjustment. Heads were trained at $L = 32$ and leaf calibrators at $L = 8$. We reuse the exact valid contiguous runs of at least 32 increments admitted by Experiment 023; longer windows do not bridge odometry failures. Both inputs and relative endpoint targets come from the unchanged recorded data. All eight arms use identical windows within a horizon. The 150 fold/seed/horizon records reproduce all original $L = 32$ learned-head sequence means exactly in both metrics.

Table 9 reports equal-family means after equal-sequence weighting within each family and averaging five seeds. Median actual durations are .405, .811, 1.585, 3.207, 6.414 and 12.792 seconds. The first four horizons retain 53 sequences/16 families; $L = 64$ retains 44 sequences/15 families (no vicon-light), and $L = 128$ retains 23 sequences/13 families (also no kidnap or table). Thus comparisons between arms at a fixed horizon are paired, but cross-horizon means do not isolate length from changes in recording coverage.

Table 9: Frozen-model ETH3D scope audit. Lower is better. Eight arms and all six horizons are retained, including unfavorable outcomes. These are post-hoc robustness results, not new independent confirmation.

Length	4	8	16	32	64	128
Recorded windows	15,699	15,059	13,779	11,219	7,278	3,435
Translation error (m)						
Identity	.159	.299	.539	.915	1.220	1.388
Raw motor	.804	1.409	2.323	4.156	7.481	23.630
Calibrated motor	.096	.170	.300	.530	.794	1.379
Constant gate	.119	.207	.355	.618	1.016	2.158
Contextual gate	.194	.225	.290	**.436**	**.597**	**1.183**
Feature-matched direct	.218	.284	.369	.526	1.118	2.011
Feature-matched residual	.526	.496	.461	.543	.759	1.308
Mergeable GA gate	**.092**	**.153**	**.259**	.440	.602	1.257
Rotation error (degrees)						
Identity	10.07	16.59	28.05	45.07	62.32	63.62
Raw motor	13.25	19.07	25.77	36.03	42.52	67.82
Calibrated motor	8.20	13.91	24.26	41.12	62.68	96.47
Constant gate	8.10	13.67	23.71	40.09	61.05	94.97
Contextual gate	8.28	10.91	**16.89**	27.88	43.22	77.84
Feature-matched direct	12.22	15.46	19.50	24.81	53.88	98.49
Feature-matched residual	20.91	21.26	20.80	**24.65**	**37.42**	**60.18**
Mergeable GA gate	**6.36**	**10.24**	17.17	28.60	40.20	71.10

At $L = 4, 8$, the mergeable model has the lowest mean translation and rotation errors among the evaluated methods. This does not establish statistical superiority over each comparator. For example, translation differences from calibrated-only are $-.0045$ and $-.0172$ m, with conditional 95% intervals $[-.0151, .0063]$ and $[-.0345, .0005]$ m. At $L = 8$, the rotation difference from contextual is $-.664°$, interval $[-1.597, .184]°$.

The $L = 64$ translation difference from identity is $-.617$ m, interval $[-.863, -.332]$ m, improving 13/15 families; differences from direct and residual heads are $-.516$/$-.157$ m, intervals $[-.718, -.331]$/$[-.216, -.092]$ m. At $L = 128$, only 8/13 families improve translation over identity and its difference interval includes zero; mean rotation is $71.10°$ versus identity's $63.62°$. Exact merging does not guarantee predictive accuracy at this horizon. At the trained $L = 32$, contextual has slightly lower mean errors in both metrics, while the mergeable model uses fewer head parameters and supports associative merging of its complete summary.

All intervals here use 10,000 paired resamples of available families, seed 300906, conditional on fitted checkpoints, named families and overlapping folds. They are exploratory, without multiplicity adjustment across arms and horizons. Original reported $L = 32$ intervals retain their original draws. Minimum observed pre-projection real norm for mergeable across the new audit is .1752; this is not a global nonsingularity guarantee. The audit verifies utility across horizons, not a new exhaustive numerical tree test.

J EXCLUDED OXFORD DEVELOPMENT ATTEMPT AND REFERENCE AUDIT

Oxford RobotCar (Maddern et al., 2017; 2020) was planned as an outdoor check using official visual-odometry (VO) increments and independently released RTK poses. The original study and a later architectural extension were locally frozen before outcome inspection. Source-schema and timestamp-eligibility corrections were documented before training; original code, manifests and seals are retained. Of 58 acquired VO/RTK pairs, one had non-overlapping timestamps, leaving 34 training, 14 development and nine unopened confirmation traversals. No missing traversal was substituted, and no pose was interpolated or perturbed.

All 50 prescribed fits completed. Although individual numerical and utility checks passed, the original conjunctive development gate failed its learned-arm competence, relative path-gap magnitude and mixed-intervention conditions. The shared confirmation gate therefore remained closed. These records cannot establish confirmation of either Oxford hypothesis and are not included in the paper's supporting accuracy results.

Coordinate-frame validation. After the development outcomes were known, we audited the frame and measurement pipeline against the official SDK's homogeneous-matrix implementation (Oxford Robotics Institute, 2022). We independently accumulated every recorded VO transform in file order, selected the same recorded timestamps without calling the SDK's interpolator, and reconstructed all four horizons on all 14 development traversals. Window counts and raw-error aggregates reproduced the motor pipeline; the maximum aggregate discrepancy was 2.1×10^{-10} m. Nevertheless, the median angle between VO and target displacement on moving windows was approximately 90° in every traversal. The SDK issue tracker contains an earlier report of a corresponding RTK-axis discrepancy (RobotCar Dataset SDK issue tracker, 2021). That report is corroborating context, not an authoritative calibration fix.

A fixed quarter-turn frame diagnostic, with no learned parameters, reduced the apparent raw error. It changes the interpretation of coordinates, not a source recording, and is retained only as a post-outcome diagnostic. It is not substituted for the frozen targets or reported as a corrected benchmark. The large learned-versus-raw gains are consequently confounded by frame interpretation and cannot be attributed to odometry correction. Agreement with SDK code alone does not establish physical frame validity.

Subsequent process investigation. All 48 training/development VO archives matched their recorded official MD5s; VO and RTK CSVs matched their archive members byte-for-byte. Extraction corruption did not explain the discrepancy. The source RTK data nevertheless contain a 1,647 m displacement in 0.101599 s, and a diagnostic > 50 m/s rule flags 1,262 training and 18 development edges. Separately, the original loader had not required consecutive VO edge timestamps to join: 652 of 36,120 development $L = 32$ windows cross a disconnected edge. The speed threshold does not validate the remaining targets, and arithmetic consistency does not establish continuity across missing motion estimates.

Two post-outcome sensitivity campaigns completed 50 additional fits, keeping the original input coefficients and retaining both all-window and screened evaluations. Their different outcomes remain conditional on an unvalidated frame interpretation, not corrected benchmark evidence. An input-lift prototype was stopped when a quaternion-sign mismatch was detected; its invalid status and the subsequent coefficient-identity tests are preserved. The original experimental outcome is retained separately.

A four-date INS/RTK comparison of 90,959 recorded timestamp pairs, including 85,088 good-status pairs, reproduces the approximately 90° yaw offset and roll sign reversal. A fixed yaw/roll-sign probe still leaves median attitude disagreements of 1.868–1.884°. These are correlated rows from four traversals; INS and RTK process the same physical GNSS/IMU observations, so cross-export agreement is not independent physical ground truth. The full VO-to-reference frame contract remains unresolved.

Exclusion from accuracy evidence. Both Oxford studies remain excluded from the positive evidence base. All development records, failed gate decisions, source hashes and audit diagnostics are retained separately in the artifact. The nine confirmation traversals have not been opened. A future Oxford study would require an independently justified frame contract, a documented corrected protocol and renewed development evaluation before any confirmation decision. The matrix audit is not an equivalent-coordinate performance comparison, and neither its diagnostic quarter-turn nor the invalidated gains support GA superiority claims.

K KITTI OUTDOOR ARCHITECTURAL REPLICATION AND ITS LIMITS

Scope and acquisition. KITTI odometry provides 11 publicly labeled driving sequences and 11 official test sequences (Geiger et al., 2012). We split the labeled recordings into 00–06 training, 07–08 development and 09–10 study holdout. Official test sequences 11–21 are not used for training, tuning or evaluation, and no evaluation-server submission is involved. The stereo images, calibration, timestamps and poses are original recordings. No interpolation, augmentation, injected noise or generated motion is used. Archive integrity and original source identities are retained.

LIBVISO2 (Geiger et al., 2011) runs unchanged with its default matcher, 200 RANSAC iterations and upstream random seed, in a fresh process per drive. Its SSE implementation is compiled for x86_64 and run through Rosetta; this is not a native ARM front-end benchmark. The adapter receives the rectified stereo images and calibration, without reference poses. The previous-to-current point transform is inverted once, following the upstream convention, to obtain composable camera motions. Quaternion signs are fixed once at the leaves; there is no intermediate sign canonicalization or normalization inside products. Failures split continuous components rather than inserting identity motion or bridging a gap. All primary holdout segments have complete observed coverage.

Locally prospective protocol and fitting. The scientific specification was fixed after a disclosed 401-frame sequence-00 front-end integration test, before full extraction outcomes or GA fitting. It is a local prospective record, not independently timestamped public preregistration. The implementation and training-only scale were sealed before any parameter update. Study-holdout release required complete fits, valid sources, front-end validation and numerical checks, *not* a positive development gain.

This is a KITTI-trained replication of the calibration/readout architectures, not zero-shot ETH3D transfer or another learned-binary-composer path-locking experiment. Each of five seeds fits a 280-parameter pointwise calibrator and five heads: constant, contextual, feature-direct, feature-residual and the 577-parameter mergeable gate. The calibrator is then frozen for every head; the six modules share logged minibatch schedules within each seed. All 30 fits complete 5,000 AdamW steps with batch size 64, weight decay 10^{-4} and gradient clipping at one. Learning rates are .002 for calibration and .001 for heads. Recorded training lengths $32, 64, 128, 256$ are sampled uniformly, then recordings and valid starts are sampled uniformly. Translation is scaled by the training-only median length-128 endpoint displacement, 108.0190 m; all six modules use coordinatewise smooth-L_1 translation loss on $(\hat{t} - t)/s$ plus $2(1 - \langle \hat{q}, q \rangle^2)$, with $s = 108.0190$ m and the quaternion dot square clamped at one. The raw-motion/endpoint information interfaces match Appendix H. No seed, checkpoint or correction threshold is selected using holdout performance. The raw and identity arms require no fitting.

Evaluation unit. The primary endpoint is the first recorded position strictly beyond 100 m of reference traveled distance from starts spaced every ten frames. We retain the original devkit's float32 cumulative-distance convention for endpoint selection. Reference distance selects the scoring window; it is not an input feature. The error transform is $\hat{T}^{-1}T$, scored by its translation norm and rotation angle. At nominal 100 m, translation in meters is numerically equal to percentage error; rotation below is degrees per nominal meter. Longer distances use their own denominators. The neural readout predicts each segment independently and need not define one globally coherent trajectory. These are *KITTI-distance segment-relative errors*, not official full-trajectory leaderboard scores or a claim to state-of-the-art visual odometry.

There are 147 primary windows on drive 09 and 98 on drive 10. We average windows within each recording, then seeds and recordings equally. The 245 overlapping windows and five fits do not create 1,225 independent trials; only two held-out drives are observed. No population-level confidence interval or significance claim is made from this split.

Table 10: KITTI study holdout: all eight prespecified arms, five seeds, 100 m segments. Per-drive columns are mean translation errors; the joint mean weights drives equally, not by their window counts. Lower is better.

Arm	Drive 09 (m)	Drive 10 (m)	Mean (m)	Rotation (deg/m)
Identity	94.435	96.444	95.440	.37893
Raw motor product	1.237	.887	1.062	.01473
GA leaf calibration	2.676	3.531	3.103	.03519
Constant gate	1.169	1.090	1.130	.01419
Contextual GA	1.252	1.100	1.176	.01352
Feature-direct	6.758	7.814	7.286	.05699
Feature-residual	1.348	1.317	1.333	.00828
Mergeable GA	1.206	1.130	1.168	.01365

Primary outcome and seed variation. The mergeable model's 100 m translation error is 10.02% higher than raw, despite 7.37% lower mean rotation error. Its per-seed, two-drive translation means are 1.3293, 1.4005, 1.0784, 1.0450 and .9879 m; raw is 1.0619 m for each seed. Drive 09 improves in the seed mean, while drive 10 worsens. The model has lower joint mean translation error than the feature-direct and feature-residual heads, but higher error than the raw estimator and constant gate. The locally fixed support criterion required at least 5% improvement in the joint translation mean against raw, identity, direct and residual controls, an improving direction on both drives and at least four seeds, and rotation no more than 10% worse than raw. This criterion is not met. Development translation error is also higher than raw (1.5528 versus 1.5379 m); training error is lower (1.6447 versus 1.7166 m). The primary task and metric are unchanged.

The pointwise calibrator alone raises the primary translation error from 1.062 to 3.103 m and rotation error from .01473 to .03519 degrees/m. This localizes an observed failure within the learned correction pipeline, not the exact composition law. The calibrator receives motion coordinates, not images, tracking residuals or sensor-confidence features. The replication therefore tests motion-only calibration, not image-conditioned pose correction or a complete visual SLAM system. It does not establish that missing visual context causes the failure or that such context is necessary; those hypotheses would require controlled comparisons.

Structural checks. All 55 recording/seed reports pass their technical checks. Across 5,375 audited window/seed cases at fixed lengths $32, 64, 128, 256, 512$, float64 raw, calibrated and mergeable predictions are checked under left, right, balanced, 16 random trees and a three-chunk reduction. Maximum decoded path differences are 1.75×10^{-11} m and 3.97×10^{-14} rad; the maximum difference between equivalent homogeneous-matrix and calibrated-motor coordinates is 3.64×10^{-12}. These audits use the first 20 eligible starts per recording/seed/length, or all starts if fewer, not an exhaustive tree enumeration. The minimum observed pre-projection real norm over all scored windows is .02825, not a global nonsingularity guarantee. Equivalent coordinates reproduce the same corrected geometry; this is not evidence that GA is more accurate than an equivalent matrix implementation.

Secondary outcomes and interpretation. Table 11 gives the complete distance sweep against raw. Translation is slightly better at 500 and 600 m, but worse at the other distances, and the rotation advantage reverses at 700–800 m. Fixed-length $32, 64, 128, 256, 512$ translation means are .3314, .6608, 1.4830, 4.0806 and 13.6205 m for mergeable, versus raw's .3142, .6174, 1.3934, 3.9826 and 13.8366 m. All secondary arms, drives and seeds are retained in the artifact. These descriptive outcomes do not replace the prespecified primary result.

Table 11: KITTI study holdout distance sweep. Translation is in meters and rotation in degrees, *not* normalized by distance in this table. Each cell weights the two drives and five seeds equally. Overlapping distance windows have different endpoints and coverage; these are descriptive sensitivity results.

Distance (m)	Raw trans. (m)	Mergeable trans. (m)	Raw rot. (deg)	Mergeable rot. (deg)
100	1.062	1.168	1.473	1.365
200	2.718	2.824	2.550	2.301
300	5.344	5.429	3.508	3.211
400	9.037	9.100	4.355	4.024
500	13.550	13.404	5.332	5.068
600	19.167	19.141	6.140	6.092
700	25.387	26.498	6.559	6.984
800	30.804	32.614	6.662	7.391

Exact PGA composition provides an associative merge operation on outdoor motion estimates. The KITTI result limits the predictive claim: learned correction does not consistently reduce the error of this stereo odometry front end. The ETH3D improvement and KITTI outcome establish dataset-dependent accuracy for the same class of geometric architectures.

L FROZEN READOUT AND MATCHED-START SENSITIVITY

Experiment 040 specifies two post-hoc sensitivity analyses after inspection of the earlier results. It freezes the 25 Experiment 028/029 checkpoints, folds, recorded inputs and targets; it does not fit or select a model. Its 150 fold/seed/horizon records reproduce every original Experiment 030 sequence-level metric exactly for all eight arms. Source and checkpoint hashes are verified before and after execution. All outcomes, including the two readout variants below, are retained.

L.1 ANTIPODAL SIGNS AND FINAL-PROJECTION CONDITIONING

Motors M and $-M$ encode the same rigid motion, but their linear blends need not agree. To isolate this issue, compute the original gate coefficient α from the unchanged input coordinates and frozen weights, then replace only the final blend by

$$B_+ = (1-\alpha)M_{\mathrm{raw}} + \alpha\eta M_{\mathrm{cal}}, \qquad \eta = \begin{cases} -1, & \langle q_{\mathrm{raw}}, q_{\mathrm{cal}}\rangle < 0, \\ +1, & \text{otherwise.} \end{cases} \tag{6}$$

The sign multiplies all eight calibrated motor coordinates, following the standard antipodal-alignment rule for dual-quaternion blending (Kavan et al., 2008). No intermediate product is canonicalized. We apply the existing final projection to B_+, separately for the contextual and mergeable gates. Since both gates still consume the original coordinate representatives, this modification does not establish invariance to arbitrary sign changes of their inputs.

For unit real parts, writing $\rho = |\langle q_{\mathrm{raw}}, q_{\mathrm{cal}}\rangle|$ gives

$$\|\operatorname{real}(B_+)\|^2 = (1-\alpha)^2 + \alpha^2 + 2\alpha(1-\alpha)\rho \;\geq\; \tfrac{1}{2}. \tag{7}$$

Thus the aligned final blend has real norm at least $1/\sqrt{2}$ for $\alpha \in [0,1]$. For nonunit, nonzero real parts, the corresponding lower bound is $\min(\|q_{\mathrm{raw}}\|, \|q_{\mathrm{cal}}\|)/\sqrt{2}$, obtained by dropping the nonnegative cross term and bounding both squared norms by their minimum. This is a bound on the final normalization denominator, not a guarantee of task accuracy, leaf-calibrator nonsingularity, or a globally continuous readout at zero endpoint dot product.

Table 12: Full-coverage frozen readout sensitivity. Values are family-equal five-seed means; contextual denotes the feature-matched fit. Each pair reports original/aligned prediction. The sign rate counts window/seed cases, not independent observations.

	Opposite	Mergeable GA		Contextual GA	
L	signs (%)	Translation (m)	Rotation (deg)	Translation (m)	Rotation (deg)
4	1.01	.0919/.0919	6.36/6.36	.1936/.1902	8.28/8.12
8	1.98	.1530/.1530	10.24/10.24	.2247/.2209	10.91/10.79
16	3.23	.2585/.2586	17.17/17.17	.2900/.2897	16.89/16.84
32	5.18	.4399/.4396	28.60/28.61	.4362/.4352	27.88/27.85
64	9.55	.6023/.5935	40.20/40.06	.5971/.5954	43.22/43.22
128	17.09	1.2568/1.2377	71.10/71.37	1.1835/1.1811	77.84/77.78

At $L = 32$, 2,905 of 56,095 window/seed cases have negative endpoint real-part dot products. Sign alignment changes mergeable translation error by $-.000284$ m and rotation error by $+.0060^\circ$ (Table 12). The primary translation benefit is therefore insensitive to this readout change; it is not an accuracy gain attributable to sign alignment. Across all six horizons, the aligned minimum real norm is .7129 for mergeable and .7086 for contextual, versus .1752 and .1776 before alignment. Every observed aligned blend satisfies the finite-input bound; no output reaches the 10^{-8} arithmetic clamp. The original predictor is retained, rather than selecting a readout from these held-out outcomes.

A float64 check uses the first 20 eligible starts per sequence, seed and horizon, or all starts if fewer: 27,215 window/seed/horizon cases. Left, right, balanced and three-chunk paths change no alignment branch on this sample. The maximum aligned prediction discrepancy is 9.65×10^{-14} m and 2.91×10^{-14} rad across both heads. This is a bounded numerical audit, not exhaustive tree enumeration or a global floating-point guarantee.

L.2 COMMON SEQUENCES AND IDENTICAL RECORDED STARTS

To separate horizon from recording coverage, we evaluate two additional populations without choosing either by performance. The common-sequence cohort retains the 23 sequences in 13 families that support $L = 128$, using all eligible starts at each shorter horizon. The matched-start cohort further fixes all 3,435 triples of (sequence, admitted contiguous run, start) that support $L = 128$. Shorter windows are prefixes of those same recorded runs; their targets are the corresponding recorded reference endpoints. No failure is bridged and no observation is generated. This conditions on 128-step run availability, not the full original recording population.

Table 13: Frozen original mergeable model on fixed recording populations. Family-equal means over five seeds; lower is better. Both panels retain the same 23 sequences/13 families at every horizon. The lower panel also fixes within-sequence starts. Full original coverage remains in Table 9.

		Translation (m)		Rotation (deg)	
L	Windows	Identity	Mergeable	Identity	Mergeable
Common sequences; all eligible starts at each horizon					
4	10,425	.145	.090	10.04	6.87
8	10,105	.271	.149	15.83	10.89
16	9,465	.480	.246	26.12	18.49
32	8,185	.782	.410	40.85	30.93
64	6,060	1.072	.672	55.84	43.77
128	3,435	1.388	1.257	63.62	71.10
Identical recorded starts at every horizon					
4	3,435	.139	.074	9.36	6.56
8	3,435	.265	.124	14.72	10.28
16	3,435	.480	.222	25.35	17.84
32	3,435	.819	.358	42.54	29.64
64	3,435	1.146	.666	61.34	43.40
128	3,435	1.388	1.257	63.62	71.10

Translation remains lower than identity through $L = 64$ on matched starts, with conditional difference intervals excluding zero at each of these five horizons. At $L = 32$, the difference is $-.461$ m, interval $[-.649, -.244]$ m; at $L = 64$, it is $-.480$ m, interval $[-.800, -.064]$ m. Restricting only sequence membership gives the same direction through $L = 64$, whose interval is $[-.667, -.051]$ m. At $L = 128$, the interval is $[-.571, .385]$ m and rotation error exceeds identity. Thus changing recording coverage is not the sole explanation for the loss of predictive advantage at the longest horizon. The scalar readout couples rotational and translational correction, but these analyses do not identify that coupling as the cause.

All eight original arms and both readout sensitivities are retained for every population and horizon. We average windows within sequence, sequences within family, then families and five seeds equally. Intervals use 10,000 paired family resamples with seed 160926, conditional on the fitted models, without multiplicity adjustment. They neither remove training-fold dependence nor provide independent confirmation. Previously reported intervals retain their original resampling draws.