PHASE-TTC
From Motion to Time

Causal Event–RGB Fusion for Signed Time-to-Contact Estimation

Álvaro Schwiedop Souto

Student, Universidade de Santiago de Compostela

alvaro@schwiedop.com

schwiedop.com/PHASE-TTC | LinkedIn | Repository

Abstract. PHASE-TTC estimates signed time-to-contact by combining event and RGB measurements over causal history. Each modality estimates motion separately; a compact head corrects the RGB estimate using twenty phase, timing, and support features. Grouped out-of-fold predictions train the later stages. With up to 650 ms of history, the submitted TRAIN40 system scores 35.80 Overall MiD over 6,762 GarlTTC Test tokens, a 20.49% reduction relative to Garl-TTC in the 28 October 2026 record. The broader evidence is mixed: on 468 labeled, development-exposed EvTTC queries, fusion trades a lower empirical p95 against worse mean and extreme errors, with paired sequence-bootstrap intervals that include zero. This expanded article retains those findings and adds the subsequent research. A compact quality adapter combines frozen neural estimates and causal box expansion in signed inverse TTC; on a corrected local evaluation of 434 historically exposed queries, MAE decreases from 0.8250 to 0.3359 s. An unresolved official-evaluator discrepancy prevents interpreting that local gain as a new official result. A separate canonical-cadence intervention improves one Slider recording but yields a road interval containing zero and fails downstream promotion controls. New predictive-learning, observability, sensor, and detection diagnostics preserve negative results rather than establish a universal JEPA benefit. A paced replay retains numerical parity but reaches 29.14 ms at p95, exceeding the 20 ms deadline. Together, the original experiments and subsequent controls characterize what causal event–RGB fusion can measure, where adaptation helps, and which claims still require independent validation.

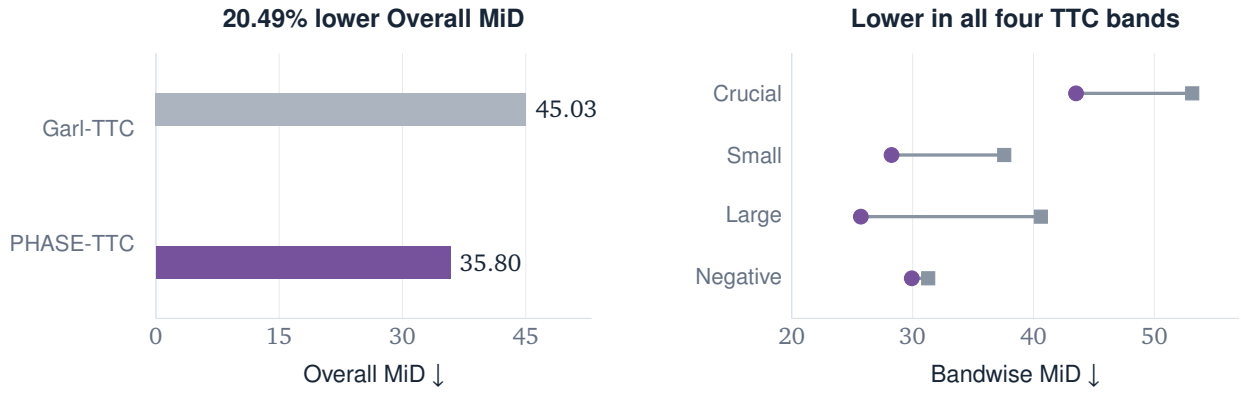


Figure 1. GarlTTC Test leaderboard, 28 October 2026. Overall MiD appears on the left and the four TTC bands on the right. Circles denote PHASE-TTC and squares Garl-TTC. Both entries report 0% scorer-defined failures across 6,762 tokens. [B01,B02]

Contents

1 Introduction 3

2 The final PHASE-TTC system 4

3 Training and evaluation 6

4 Official GarlTTC benchmark result 9

5 Frozen transfer and modality comparisons 11

6 Competition inputs: historical coverage and disagreement 13

7 Quality-conditioned adaptation of the frozen system 14

8 Canonical cadence as a controlled intervention 17

9 Design analysis and controlled ablations 19

10 Runtime and causal service 21

11 Discussion and limitations 22

12 Conclusion 24

A Technical foundations and historical method families 27

B Evaluation details and causal checks 32

C Paired transfer uncertainty and numerical cases 34

D Competition adaptation and complete prediction traces 36

E The initial representation-learning programme 41

F Temporal information and phase refinement 45

G Transfer, output stability, and runtime 48

H From event representations to multimodal measurement 50

I Representation and physical-measurement follow-ups 54

J Detection, geometric consistency, and the localization boundary 56

K Experimental chronology 57

L Reproducibility 59

M Evidence index 61

N Expanded evidence: 29 October–2 November 2026 65

Evaluation overview

GarlTTC Test: scored benchmark performance (6,762 tokens).
Labeled EvTTC transfer: errors and constituent controls (468 eligible queries).
EvTTC competition: preserved frozen-model coverage/disagreement and corrected local adaptation scores (434 queries).
Canonical-cadence development: 2,608 queries; historical exposure and rejected promotion controls.
Paced replay: complete-output response time (250 updates).

1. Introduction

An approaching object's image expands before contact. Recovering time-to-contact (TTC) from that expansion requires preserving small changes in scale through cropping, feature extraction, and temporal aggregation. Event cameras make this a particularly demanding measurement problem: they record asynchronous brightness changes, and the number of events depends on texture, illumination, contrast, and sensor thresholds as well as motion. Features that predict event activity accurately may therefore carry little information about approach speed.

PHASE-TTC combines geometric measurements from events and RGB images with their recent history. Each modality first produces its own motion estimate. The final fusion head receives named phase, timing, and support features and learns a correction to the RGB estimate. This lets the two branches share interpretable measurements without requiring their hidden coordinates to align. We train the later stages on out-of-fold predictions generated from separate sequence groups. [M22,M23]

The resulting TRAIN40 system ranks first on the public GarlTTC Test leaderboard as of 28 October 2026. It scores 35.80 Overall MiD, compared with 45.03 for Garl-TTC, with a 0% scorer-defined failure rate over 6,762 tokens [4]. This 20.49% reduction is measured in the benchmark's weighted phase error. To understand the result beyond the leaderboard, we examine the contribution of each modality, the effects of temporal context, and the time required to deliver a complete prediction. [B01,B02]

We also apply frozen checkpoints to 434 queries from ten EvTTC competition sequences. PHASE-TTC returns a prediction for every query; reproduced Garl full, event-only, and visual-only models return 424, 434, and 426 predictions. At the 28 October snapshot, numeric TTC targets had not been verified for these exact identities, so the original comparison measured coverage and disagreement (Section 6). A subsequent target-source audit recovers all 434 targets and enables a separate local comparison of the original PHASE-TTC output with a quality-adapted variant. The historical disagreement analysis remains intact; the new local error analysis neither converts disagreement into accuracy nor replaces the earlier 468-query labeled transfer evaluation. [T01,N01]

The extension through 2 November 2026 adds three compact quality routers over frozen neural and causal geometric estimates. On the corrected 434-query local evaluator, the delivered adaptation reduces MAE from 0.824976 to 0.335899 s. These recordings had influenced earlier development, and Slider-750 still disagrees with the organizer-provided score. We therefore describe the result as exposed local adaptation, not independent benchmark validation. A separate intervention corrects the distinction between actual observation cadence and the benchmark's canonical phase interval. Its modest development gains coexist with failed downstream promotion controls. [N01--N03]

1.1 Contributions

Our contributions are the phase-based estimator and its training procedure, an evaluation that includes both modalities separately, and a study of prediction delivery under paced arrivals. The design experiments explain which changes helped and which failed. A post-hoc analysis compares fusion and RGB errors on the labeled transfer queries, with uncertainty estimated by resampling complete sequences (Appendix C).

The expanded article develops two further system contributions: a quality-conditioned adapter that fuses signed inverse TTC while preserving frozen upstream models (Section 7), and a controlled cadence intervention that rebuilds temporal inputs before testing the complete downstream chain (Section 8). The accompanying predictive-learning, observability, sensor, and geometry controls distinguish a lower intermediate loss from a useful TTC estimate. We retain negative results and rejected candidates alongside positive measurements, and keep detection diagnostics in a separate appendix. The contribution lies in the implemented system and controlled evidence; the physical TTC identities and quality-weighted fusion principle are not claimed as new laws or unique ideas. [N01--N06]

1.2 Relation to prior work

Event cameras provide high temporal resolution through asynchronous brightness changes, but downstream algorithms must still construct task-relevant spatial and temporal representations [3]. DSEC demonstrates synchronized event and frame acquisition in demanding driving conditions [5].

I-JEPA learns by predicting target representations from context [1]. VICReg regularizes representation variance and covariance [2]. These approaches motivate the early predictive-learning experiments, whose losses and regularizers are reported with each configuration.

Predictive joint embeddings can favor slowly varying distractors over a changing state of interest [11]. We test this possibility by changing nuisance dynamics independently of the TTC signal. The resulting controls help distinguish predictable appearance from useful motion information.

Prior event-based TTC methods estimate motion through geometric fitting and spatio-temporal registration [6]. EvTTC supplies synchronized observations and a dedicated TTC benchmark [12]. The eAP study introduces geometry-aware representations for event-enhanced perception and the Garl models used in our comparisons [7]. We retain the evaluation populations defined by each experiment when comparing these methods. [T02]

2. The final PHASE-TTC system

2.1 Deployed model composition

PHASE-TTC deploys a causal chain of nine fitted components derived from the V16 PHASE-ONLY family: four modality-specific producers, two PAIR heads, two context refiners, and one fusion head. RGB and events each use an A5 baseline producer and a coarse-to-fine C2F producer. The PAIR readouts extract phase features from their A5 representations, after which the context refiners incorporate temporal history. The fusion head anchors its correction on the refined RGB estimate. Figure 2 traces this path; the five auxiliary rotations serve training only. [M22,M23]

Final fitting uses 88,744 TRAIN40 queries from forty sequences. The former P, H, and V groups all contribute to this refit, so their earlier scores characterize development rather than independent validation of the deployed chain. Each producer receives 49,932 updates, equivalent to eighteen epochs at effective batch size 32. The event producers carry forward the earlier TRAIN40 weights; RGB training resumes to the same endpoint. [L43]

2.2 Object coordinates and causal history

Each query includes supplied object coordinates. A common spatial frame covers both endpoint regions and preserves apparent scale change, which independent tight crops can erase. The evaluated system therefore assumes external localization. [M03]

Both branches use observations available by query time t, with up to 650 ms of causal history. Runtime begins at its declared measurement boundary and ends when the host receives the TTC output at $t + L$. Figure 3 separates the sensing interval from response time; its observation marks are schematic. [M22,L43,L44]

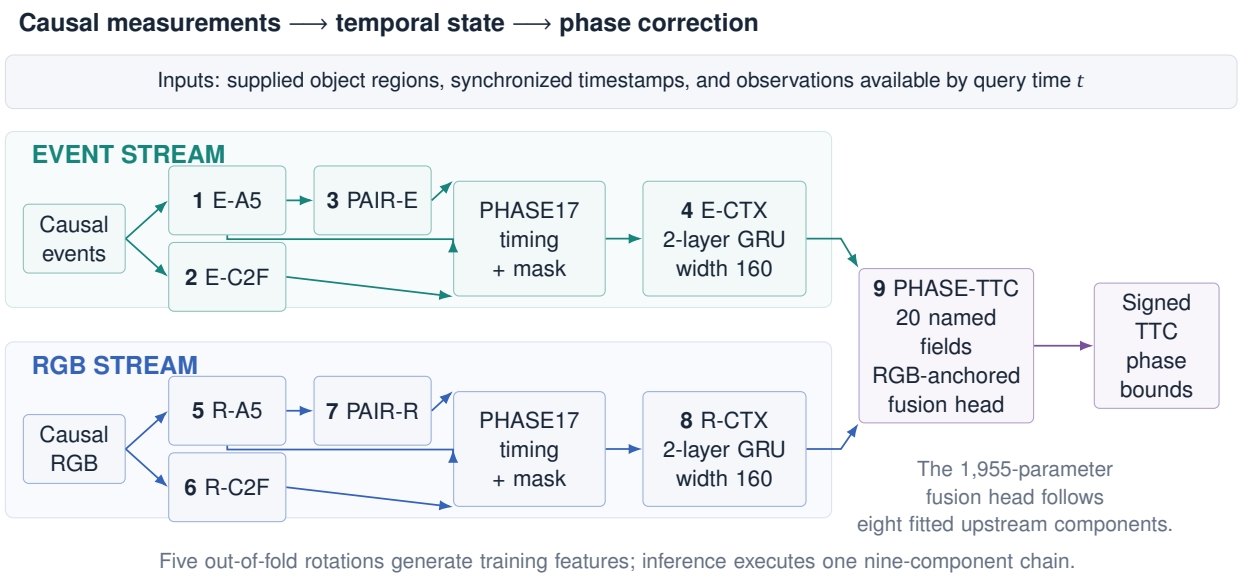


Figure 2. Causal nine-component inference graph. Supplied object regions feed parallel RGB and event producers, PAIR heads, and temporal refiners. Twenty named fields enter the final RGB-anchored fusion head. [M03, M22, M23]

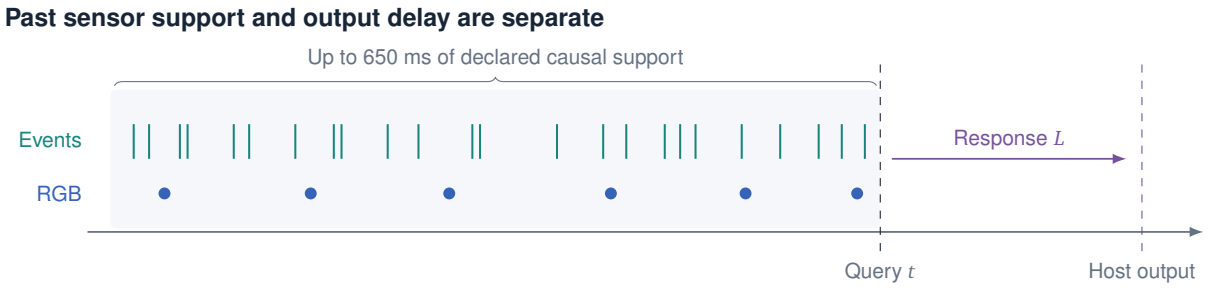


Figure 3. Causal support and response-time boundaries. RGB and event observations precede query time t. The latency clock spans the declared runtime boundary to the complete host TTC output at $t + L$.

2.3 Compact fusion interface

Each context refiner is represented for final fusion by a constructed 170-coordinate block. Coordinates 0–159 are its learned hidden state. Coordinates 160–169 are an explicit interface containing, in order, one raw location, one emitted phase, three expert phases, four current timing values, and one valid-history count divided by eight. Concatenating the RGB block followed by the event block gives the 340-coordinate container

$$\mathbf{s} = \left[\mathbf{s}_R^{170}; \mathbf{s}_E^{170}\right] \in \mathbb{R}^{340}. \tag{1}$$

Fusion reads $\mathbf{s}_{160:170}$ and $\mathbf{s}_{330:340}$, yielding twenty named phase, timing, and support inputs while the 160-dimensional hidden states remain within their modality branches. Within each ten-coordinate block, the first five phase coordinates are divided by 0.03; timing and support keep their native scaling. The shared interface aligns quantities with consistent meanings across the separately fitted refiners. [M22]

The selected vector passes through

$$20 \to 32\ [\text{LayerNorm}, \text{SiLU}] \to 32\ [\text{LayerNorm}, \text{SiLU}] \to 3. \tag{2}$$

Zero initialization of the last affine layer makes the initial fusion output equal to the RGB anchor. If c_0, c_1, c_2 are the three head outputs, ℓ_R is the RGB raw-location anchor, and $\mathbf{w}_R$ is the separately supplied

pair of RGB width logits, the deployed computation is

$$\ell_F = \ell_R + 0.03c_0, \tag{3}$$

$$\boldsymbol{\sigma}_F = 0.03\ \mathrm{softplus}(\mathbf{w}_R + [c_1, c_2])\,, \tag{4}$$

$$\widehat{\mathrm{TTC}} = \mathrm{phase_to_ttc}(\ell_F), \tag{5}$$

$$q_{10}^{\phi} = \mathrm{clip}(\ell_F - \sigma_{F,0},\ \phi_{\min}, \phi_{\max}), \tag{6}$$

$$q_{90}^{\phi} = \mathrm{clip}(\ell_F + \sigma_{F,1},\ \phi_{\min}, \phi_{\max}). \tag{7}$$

The native bounds are $\phi_{\min} = -\log 2 + 10^{-6}$ and $\phi_{\max} = 8$. For the point prediction, the inverse map clamps ℓ_F to these bounds, computes $q = -\mathrm{expm1}(-\phi)/0.1$, and emits $\mathrm{sgn}_+(q)/\max(|q|, 1/60)$ seconds, where $\mathrm{sgn}_+(0) = 1$. Thus the emitted magnitude is at most 60 s and zero phase maps to +60 s. The reported point phase is recomputed from this finite TTC; the learned-width phase-interval endpoints are clamped separately. Saturation is a numerical convention, not a measured 60-second collision time or a calibration result. [M22, R02]

The compact fusion head contains 1,955 trainable parameters: 672 + 64 + 1,056 + 64 + 99 across three affine layers and two LayerNorms. Figure 4 summarizes the slicing and anchored output map. [M22]

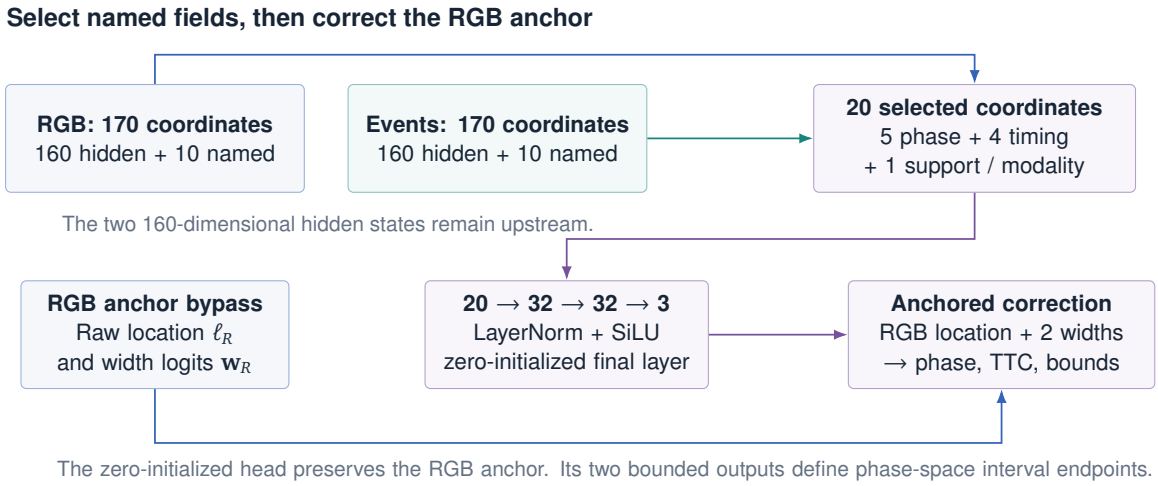


Figure 4. Named-feature fusion interface. Twenty phase, timing, and support coordinates are selected from the 340-value container. The RGB anchor and width logits follow a separate bypass into the three-output correction head. [M22]

2.4 Submission coverage and benchmark status

The official submission contains finite predictions without fallback for 6,762 query tokens across twelve sequences: 5,805 physical queries and 957 aliases. Aliases identify repeated physical queries and do not represent additional inference work. [L44]

Section 4 reports the official test result for this V16-derived PHASE-ONLY graph. The later V18 controls evaluate a different configuration. [B01, B02, L44]

3. Training and evaluation

3.1 Grouped out-of-fold training and final refitting

Five rotations of eight-sequence blocks generate leakage-controlled inputs for the upper stages. In each rotation, twenty-four groups fit the four producers and two PAIR heads, eight different groups fit the context refiners and normalizers, and the remaining eight groups generate out-of-fold features. A query group never contributes to supervised fitting of the models that generate its upper-stage features. [M23]

We pool PHASE17 RGB and event features across the out-of-fold rotations to train the context heads and normalizers. Those models then generate the PHASE20 features, expressed through the 340-coordinate container, for fusion-head training. Rotations pool named fields with shared semantics, while learned hidden coordinates stay within their originating models. Targets join the data only after out-of-fold feature generation. [M23]

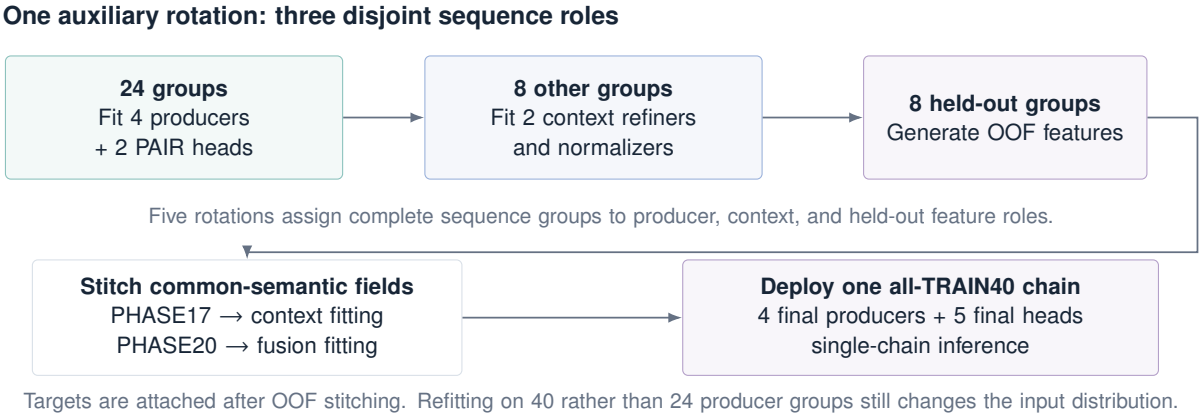


Figure 5. Five-rotation fitting scheme. Each rotation separates producer fitting, context fitting, and held-out feature generation. Final refitting produces one all-TRAIN40 inference chain. [M23]

Inference executes one all-TRAIN40 chain comprising the four producers, two PAIR heads, two context refiners, and fusion head. Auxiliary producers see twenty-four groups per rotation, whereas deployed producers see all forty. This stage-wise exposure difference can shift the upper-stage feature distribution after refitting. Earlier development use also remains relevant to architecture-selection claims.

Training comprised 690,934 auxiliary endpoint updates and 58,660 final-head updates across forty-five fits. Final PAIR endpoints used 11,264 updates; each context model and the final head used 12,044. Schedulers and warmup scaled with the fitting population. Update counts describe optimization exposure rather than wall-clock GPU effort. [L45]

Producer training combines relational DINO teacher signals, foreground geometry supervision, and task-specific adaptation. The deployed chain no longer evaluates the generic teacher, but its supervision remains part of the training recipe.

3.2 Metrics and target coordinates

TTC error in seconds, inverse-TTC error, benchmark phase error, and detection of urgent risk measure different aspects of behaviour. We use the signed benchmark phase

$$\phi(T) = -\log\left(1 - \frac{0.1\,\mathrm{s}}{T}\right), \qquad T < 0 \text{ or } T > 0.1\,\mathrm{s}. \tag{8}$$

Negative TTC denotes recession. The interval $[0, 0.1]$ s is outside this transform's domain. The official benchmark scorer defines motion-in-depth (MiD) error as

$$\mathrm{MiD}(\widehat{T}, T) = 10^4 \left|\log\left(1 - \frac{0.1}{T}\right) - \log\left(1 - \frac{0.1}{\widehat{T}}\right)\right| = 10^4 |\phi(\widehat{T}) - \phi(T)|. \tag{9}$$

MiD measures phase displacement. At large positive T, the first-order approximation $\phi(T) \simeq 0.1/T$ compresses differences in seconds, while sensitivity rises near the positive domain boundary. We therefore report MAE in seconds alongside MiD. [M10, M15]

The scorer groups target TTC into four right-closed intervals: crucial $c : (0, 3]$ s, small $s : (3, 6]$ s, large $l : (6, 10]$ s, and negative $n : (-10, 0]$ s. The grouping includes zero in the negative band, although the phase transform is undefined at zero; metric validity is checked separately. Targets outside these bands do not contribute to the band-weighted score.

The scorer computes the band-weighted score

$$\text{overall_MiD} = 0.5\,\text{MiD}_c + 0.3\,\text{MiD}_s + 0.1\,\text{MiD}_l + 0.1\,\text{MiD}_n, \tag{10}$$

and a sample-weighted mean MiD. Missing bands remain missing, without weight redistribution. The complete band-weighted score requires targets in all four bands. The evaluation fixes the scorer by SHA-256 and tests a perfect four-band example, a hand-computed crucial-band error, and an absent negative band. [M15]

Metric reductions retain failures explicitly. The wrapper records invalid predictions by band and leaves its strict mean unavailable when any row is invalid. The official failure rate differs from the false-negative rate for urgent collision warnings. We therefore report prediction coverage, invalid-MiD count, and risk recall with their respective metrics. A finite saturated TTC may still carry an unsupported-measurement flag.

3.3 Official benchmark and exploratory evaluations

Table 1. Evaluation populations and the quantities each supports. Historical46 and Extended-new share three sequence groups but no query identities. The 434-query inference snapshot and corrected local comparison share identities but answer different questions.

Population	Size	Interpretation
Official GarlTTC Test	6,762 scored tokens	Private targets; band-weighted MiD and scorer-defined failures. [B01, B02]
Published Garl EvTTC	3 sequences	Published RTE; target/query correspondence to our run unverified. [T02]
EvTTC competition snapshot	434 queries; 10 sequences	Original 28 October predictions, coverage and disagreement; targets not verified at that snapshot. [T01]
Corrected local EvTTC	Same 434 queries	Original versus adapted PHASE-TTC error; historically exposed, local scoring, Slider-750 official equivalence unresolved. [N01]
EvTTC Extended-new	468 targets; 32 sequences	Labeled, development-exposed transfer with annotated ROI. [L01]
EvTTC Historical46	46 targets; 3 sequences	Labeled, development-exposed transfer with annotated ROI. [L01]
Cadence development	2,608 queries; 25 recordings	Recording-macro paired comparison; 24 road recordings and Slider-500; historically exposed. [N02]
Native eAP replay	5,805 physical queries; 6,762 template tokens	Local output sensitivity and coverage; private Test12 targets not scored for the candidate. [N03]

The official GarlTTC Test submission covers twelve sequences and uses TRAIN40 event–RGB inputs with up to 650 ms of causal history. Section 4 reports the public ranking as of 28 October 2026 over 6,762 scored tokens, including aliases of repeated physical queries [4]. The evaluator scores submitted TTC values without standardizing temporal support, localization, supervision, or runtime across systems. Separate development experiments probe mechanisms within their tested configurations. [B01, B02, L44]

3.4 Corrected local targets and later development cohorts

The 434-query competition inference snapshot originally supported only coverage and model disagreement. The subsequent target-source audit corrected the public assignment for two road recordings and recovered eleven previously absent targets, giving 434 supported targets. The local error comparison reports query-weighted MAE and mean relative error. Both original and adapted columns apply the organizer-described interpolation of negative predictions between valid outputs for scoring only; their archived prediction files retain the signed values unchanged. The original and adapted files contain two

and one negative values, respectively. Nine of ten sequence relative-error rows reproduce the organizer-provided original result after rounding, while Slider-750 remains discrepant. The corrected local table therefore does not constitute a new official score. [N01]

The cadence experiment has a separate population: 2,531 queries from twenty-four road recordings and 77 from Slider-500. Its primary summaries average recordings equally within each domain, preserve signed predictions without clipping, and use relative denominator $\max(|T|, 10^{-6})$. Availability is reported independently of paired error, including the frozen F0 route's one missing Slider output. The protocol requires preserved coverage, no worsening of MAE or relative error in either domain, and an improvement in at least one metric. A failed fixed 149-query screen remains a failure even when the larger-population aggregate improves. [N02,N03]

The final quality-router fit also uses 2,608 development queries and excludes the eight competition road recordings and Slider-750/1000 from optimizer updates. Earlier experiments had already inspected those recordings, so this exclusion does not establish independent generalization. The later paired refit and canonical-cadence runs are separate model variants. F0 passes its numerical development criteria, but the router/refit gates fail and native accuracy conservation remains unverified. Neither replaces the delivered quality adapter. [N01--N03]

3.5 Independence and paired comparisons

Complete sequences form the units of separation and resampling. A sample-weighted sequence bootstrap carries all rows from each sampled group into the replicate and recomputes the ratio of summed losses to summed counts; sequence-macro summaries weight groups equally. Both reductions inherit any development exposure. Appendix B gives the corresponding lineage, causality, and latency checks. [M15--M20]

Three head seeds quantify optimization variability conditional on their parents; they do not independently replicate the upstream representation. The additional transfer bootstrap is post hoc: it resamples saved paired predictions without changing a model, split, or historical interval (Appendix C).

For canonical cadence, 10,000 paired road-recording bootstrap replicates with seed 7 estimate descriptive intervals over the twenty-four development recordings. The intervals retain the cohort's historical exposure. Slider-500 is one recording reused from the screen and receives no between-recording confidence interval. A separate eight-fold, three-seed, two-arm router comparison comprises 48 fits; those fits do not create 48 independent datasets or erase upstream exposure. [N02,N03]

4. Official GarlTTC benchmark result

4.1 Public ranking and bandwise accuracy

As of 28 October 2026, PHASE-TTC ranks first among the three displayed submissions on the official GarlTTC Test leaderboard. Submission 1016512, listed as *PHASE-TTC (TRAIN40, 650 ms causal history)*, records Overall MiD 35.80 over 6,762 scored tokens with a 0% scorer-defined failure rate. The next displayed score is 45.03 for the Garl-TTC reference. [4] [B01]

The scorer returns Overall MiD 35.8015 at full precision, consistent with the rounded public entry. Relative to Garl-TTC's published 45.03, PHASE-TTC reduces the ranking metric by 20.49%. All four TTC bands improve (Table 2); the large positive-TTC band has the largest relative reduction and the negative band the smallest. Percentage reductions combine the unrounded PHASE-TTC values with Garl-TTC's published two-decimal scores. [B01,B02]

Table 2. Official GarlTTC Test scores on 28 October 2026 (lower is better). PHASE-TTC scores are shown to four decimals and Garl-TTC at published precision. Relative reductions use unrounded PHASE-TTC scores. Both entries have 6,762 scored tokens and 0% scorer-defined failures. [B01, B02]

Metric	PHASE-TTC	Garl-TTC	Decrease
Overall	35.8015	45.03	20.49%
Crucial	43.5190	53.13	18.09%
Small	28.2568	37.58	24.81%
Large	25.7179	40.61	36.67%
Negative	29.9314	31.29	4.34%

4.2 Where the weighted difference comes from

For band b, the additive contribution is $w_b(\mathrm{MiD}_b^{\mathrm{Garl}} - \mathrm{MiD}_b^{\mathrm{PHASE}})$. The crucial and small bands contribute 82.4% of the Overall difference under weights 0.5 and 0.3. The large band's 0.1 weight limits its contribution despite its larger relative reduction; the negative band contributes least.

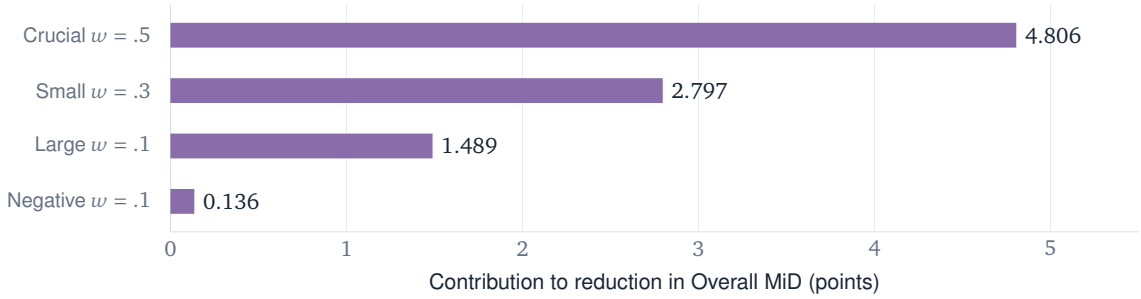


Figure 6. Bandwise contributions to the Overall MiD difference. Crucial and small TTC account for about 82% under the benchmark's 0.5/0.3/0.1/0.1 weights. Table 29 gives the rounding reconciliation. [B01, B02]

4.3 Metric and evaluation population

The four TTC bands cover $(0, 3]$, $(3, 6]$, $(6, 10]$, and $(-10, 0]$ seconds. MiD measures error in the motion-in-depth coordinate defined in Section 3. The ranking uses the weighted Overall MiD; the scorer also reports `mean_MiD`, an unweighted average over finite per-query errors. [9]

The 6,762 scored tokens represent 5,805 distinct physical queries and 957 aliases in the submitted package. Aliases reuse predictions at the same sensor query. The official failure criterion marks nonfinite TTC predictions or predictions with $|\widehat{T}| < 0.1$ s. Band scores average finite MiD values. The available aggregate outputs support bandwise analysis; studying individual errors requires per-query scores. [L44, B02]

4.4 Scope of the comparison

The submission declares RGB and event inputs, TRAIN40 fitting, and up to 650 ms of causal history. Because the evaluator scores submitted TTC values, the leaderboard compares complete systems with their own temporal support, localization, supervision, and runtime choices. Section 9 examines modality and temporal dependencies within the tested development configurations.

The official result measures the submitted system on private test targets. The exploratory EvTTC transfer compares frozen routes on development-exposed, positive-TTC sequences with annotated object regions. Section 10 evaluates execution time separately, including deadline misses.

5. Frozen transfer and modality comparisons

LABELED TRANSFER / 468 eligible queries / 32 development-exposed sequences / annotated ROI

On the labeled transfer queries, PHASE-TTC had lower error than the earlier Garl adapter, but the RGB-only model remained stronger on average. The full error distribution reveals a mixed comparison with RGB: fusion had a lower empirical 95th percentile and more of the largest errors.

5.1 Evaluation data and scoring

This evaluation uses labeled EvTTC sequences and the earlier frozen Garl adapter. Both differ from the 434-query competition experiment in Section 6.

We froze the final candidate before evaluating transfer on EvTTC, with no further fitting or checkpoint selection. We registered sixteen uniformly spaced query times in each of thirty-two sequences with compatible sensors and object annotations, and retained forty-six queries from the earlier evaluation. Removing four duplicates left 554 physical queries. Predictions were fixed before using TTC labels for scoring. The selected query times differ from those in the earlier 946- and 3,640-query H8 evaluations, including on shared sequences.

Each query received the nearest public TTC label within 50 ms, with the earlier timestamp breaking ties. Forty queries had no label within that tolerance. We scored Historical46 and Extended-new separately, using 46 and 468 labeled queries. Their query identities are disjoint after duplicate removal, although all three Historical46 sequences also appear among the thirty-two Extended-new sequences. All retained TTC labels are positive.

Relative time error was defined as

$$\mathrm{RTE} = 100\,\frac{|\widehat{T} - T|}{|T|}, \tag{11}$$

without clipping the signed nonzero target. Pooled summaries give each query equal weight; sequence-macro summaries give each physical sequence equal weight. We resample whole sequences for confidence intervals to preserve correlation between nearby query times. EvTTC had already informed development, so the evaluation is exploratory. Inputs also use annotated boxes, and temporal support differs: PHASE-TTC can use up to 650 ms, while Garl uses two frames and event segments separated by 0.1 s. Both use float32 tensors, but backend settings were not matched.

Table 3. Frozen TRAIN40 transfer on `extended_new`: 468 GT-valid queries, 32 sequences, positive TTC only, annotated ROI. Pooled metrics; lower is better. Source [L01].

Model	MAE (s)	RTE (%)	Median AE (s)	P95 AE (s)
Event context	1.451	65.70	0.857	4.397
Garl reproduced	3.258	140.07	0.674	5.252
Phase correction	1.023	41.23	0.464	3.243
RGB context	1.015	38.56	0.408	3.799

5.2 Average error and the RGB comparison

PHASE-TTC's advantage over the earlier Garl full adapter extended across most scenes: it had lower sequence-mean RTE on twenty-six of thirty-two Extended-new sequences. Its pooled RTE was 41.23% and MAE 1.023 s (Table 3). Giving each sequence equal weight produced an RTE difference of −94.83 percentage points, with a 95% bootstrap interval of $[-210.50, -14.49]$. The wide interval reflects substantial variation across scenes. [L01]

RGB alone remained the stronger model on average, with RTE of 38.56% and MAE of 1.015 s on the same queries. It also had lower pooled errors on Historical46. The RGB branch is therefore a more demanding reference for assessing the value of event fusion than the earlier Garl adapter.

The ordering changes farther into the saved error distribution. RGB also has the lower median error, while fusion's empirical 95th percentile is 3.24 s against RGB's 3.80 s (Figure 7). Fusion has lower sequence-mean RTE on fourteen of thirty-two sequences (Figure 19). [L01, L46]

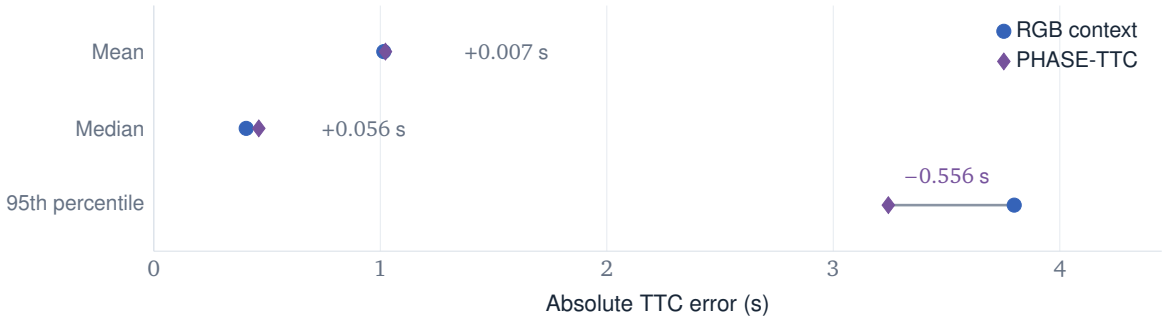


Figure 7. Absolute-error summaries for fusion and RGB. Connected points compare the mean, median, and 95th percentile on the same 468 Extended-new queries. RGB has lower mean and median error; PHASE-TTC has a lower 95th percentile. Table 4 gives intervals for the paired differences.

5.3 Error exceedance and paired uncertainty

The saved predictions contain fewer moderately large fusion errors but more extreme ones. PHASE-TTC has fewer absolute errors above 3 s (29 versus 37) and 5 s (9 versus 13), yet more above 10 s (5 versus 2). Its maximum error reaches 32.5 s, nearly twice RGB's 16.6 s. We computed these threshold counts post hoc, retaining every eligible prediction in Figure 8. [R01, L46]

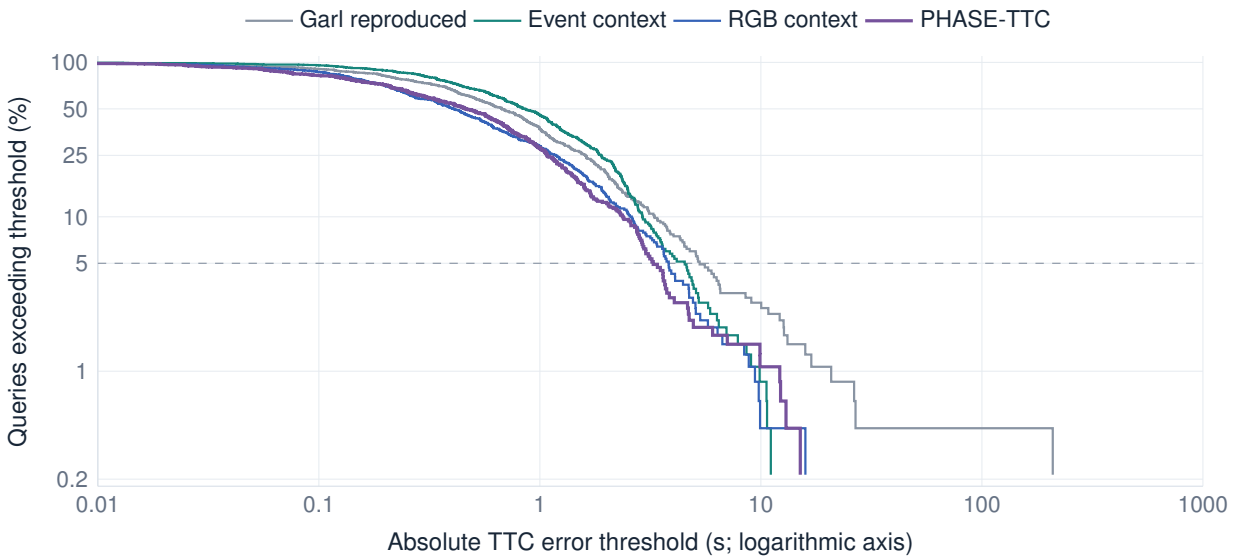


Figure 8. Fraction of queries exceeding each absolute-error threshold. Curves use all 468 labeled queries, with equal query weights and logarithmic axes. The dashed line marks 5%. The plotted range includes the earlier Garl adapter's outlier above 600 s. [R01, L46]

Resampling complete sequences gives a PHASE-minus-RGB p95 difference of −0.556 s, with a 95% interval of $[-1.126, +0.241]$ s. The mean and median difference intervals also include zero. These exploratory intervals use 10,000 paired replicates, a fixed seed, and linear quantiles; they describe the saved models on sequences already used during development. Appendix C provides the procedure, every sequence result, and two numerical cases selected post hoc. [R01]

Table 4. Paired differences in absolute error across 468 labeled queries from 32 sequences. Percentile intervals use 10,000 paired sequence resamples. [R01;derivedfromL46]

Metric	RGB (s)	PHASE-TTC (s)	Difference [95% interval] (s)
Mean absolute error	1.015	1.023	+0.007 [−0.110, +0.129]
Median absolute error	0.408	0.464	+0.056 [−0.038, +0.145]
95th percentile	3.799	3.243	−0.556 [−1.126, +0.241]

6. Competition inputs: historical coverage and disagreement

28 OCTOBER INFERENCE SNAPSHOT / 434 query identities / 10 sequences / coverage and prediction disagreement

We applied three released Garl-TTC variants to the same ten-sequence EvTTC inputs as the frozen PHASE-TTC reference. The 434 queries span eight road and two Slider sequences, with sequence, row, frame index, and timestamp matched across models. The comparison contains 1,302 Garl outputs and 434 PHASE-TTC outputs. [T01]

At the time of this frozen inference comparison, no verified numeric TTC targets mapped to the 434 query identities. The analysis therefore measured coverage and paired prediction disagreement:

$$d_i^{(m)} = \widehat{T}_{i,m}^{\mathrm{Garl}} - \widehat{T}_i^{\mathrm{PHASE}}, \qquad a_i^{(m)} = |d_i^{(m)}|. \tag{12}$$

The differences in Equation 12 compare model outputs rather than prediction error. This population differs from both the labeled transfer in Section 5 and Garl's published three-sequence RTE evaluation. Appendix D gives the adapter, timing measurements, and ten sequence traces.

The corrected target audit reported on 1 November supersedes the earlier absence of local labels: it recovers all 434 targets after resolving the public source assignment for two road sequences. Section 7 uses those targets to compare original and adapted PHASE-TTC under one documented local evaluator. The results below preserve the original Garl-versus-PHASE outputs and their coverage rather than retroactively presenting them as a reproduced error ranking. The published three-sequence result remains a separate protocol, and the Slider-750 discrepancy prevents claiming official equivalence for the new local scores. [T01,T02,N01]

6.1 Frozen checkpoints and native input support

The released full, event-only, and visual-only state dictionaries load strictly and run in float32 with frozen weights. RGB-dependent routes use two real causal frames, while event-only uses two observed causal intervals. Startup queries without the required observations remain unavailable. PHASE-TTC uses up to 650 ms of context, so the models share query identities under different temporal support. [T01]

Supplied boxes and calibration condition each route. Rotation-only projection omits translation because object depth is unavailable. For Slider, the adapter derives rotation from the supplied camera poses because the data provide no named event-to-RGB extrinsic. We treat the resulting runs as frozen-checkpoint adaptations.

6.2 Coverage, sign, and disagreement tails

PHASE-TTC and event-only Garl each emit 434 finite values; full Garl emits 424 and visual-only 426. The first query in each road sequence lacks two causal RGB frames, creating eight gaps in both RGB-dependent variants. Full Garl has two further Slider-750 gaps because the earlier event endpoint precedes recorded coverage. All invoked models return finite values; nulls correspond to unavailable inputs.

On shared finite rows, median absolute disagreement from PHASE-TTC is 0.26 s for full Garl, 0.68 s for event-only, and 0.24 s for visual-only. Their p95 disagreements are 1.10, 5.62, and 1.52 s. Full Garl

Table 5. Predictions on 434 EvTTC competition queries. The table reports finite-output coverage, signed TTC range, and negative counts. Median and P95 summarize absolute differences from PHASE-TTC where both models produce finite outputs. [T01]

Model	Finite	Range (s)	< 0	Median (s)	P95 (s)
PHASE-TTC	434/434	[-42.225, 51.869]	2	–	–
Garl full	424/434	[0.907, 52.705]	0	0.261	1.098
Garl event-only	434/434	[-42.211, 406.095]	10	0.681	5.623
Garl visual-only	426/434	[-23.832, 38.767]	5	0.236	1.524

	PHASE-TTC	Garl full	Garl event-only	Garl visual-only
CCRs1-low	50/50	49/50	50/50	49/50
CCRs1-medium	43/43	42/43	43/43	42/43
CCRs1-high	30/30	29/30	30/30	29/30
CCRs2-low	46/46	45/46	46/46	45/46
CCRs2-medium	55/55	54/55	55/55	54/55
CCRs2-high	59/59	58/59	59/59	58/59
CCRm-low	42/42	41/42	42/42	41/42
CCRm-medium	40/40	39/40	40/40	39/40
Slider-750	51/51	49/51	51/51	51/51
Slider-1000	18/18	18/18	18/18	18/18

Counts show finite predictions per sequence; tinted cells mark queries with insufficient causal input.

Figure 9. Finite prediction coverage by model and sequence. Tinted cells mark the eight missing road RGB pairs and two additional early Slider-750 event gaps in full Garl. [T01]

emits no negative TTC, compared with two negative PHASE-TTC outputs, ten event-only outputs, and five visual-only outputs. [T01]

Slider-750 contains the largest excursions: event-only reaches 406.10 s, with p95 absolute disagreement of 241.68 s. The native height-ratio conversion becomes ill-conditioned as the two predicted endpoint heights approach equality. Figures 20 and 21 preserve the sign and magnitude of these outputs across all ten sequences. [T01]

Timing remains descriptive because the variants ran once in fixed order with shared cache and warm-state effects, and PHASE-TTC used a different adapter and longer history. [T01]

7. Quality-conditioned adaptation of the frozen system

LOCAL ADAPTATION / 2,608 fitting queries / 434 corrected evaluation queries / historical exposure retained

The original TRAIN40 graph supplies several estimates of the same signed quantity: its final fused TTC, the RGB context estimate, and the event context estimate. Their errors need not agree, and their reliability changes with temporal support and apparent motion. The subsequent PHASE-TTC-Adapt experiment leaves all nine upstream components frozen and learns a small router over those three outputs and a causal box-expansion estimate. This is adaptation of the fusion rule. It is not a new end-to-end fit of the visual encoders, and it does not incorporate the later canonical-cadence intervention in Section 8. [N01]

7.1 Causal box geometry and its quality descriptors

For each query, the geometric branch admits supplied boxes with timestamps in the preceding 650 ms, rejects nonfinite or nonpositive extents, and requires strictly increasing observation times. It fits three apparent sizes: width, height, and square root of area. Let s_i denote one such size, s_n its latest value, and u_i time relative to the query in seconds. The branch constructs $\rho_i = s_n/s_i$ and estimates a line using median pair slopes,

$$b = \operatorname{median}_{i<j} \frac{\rho_j - \rho_i}{u_j - u_i}, \qquad a = \operatorname{median}_i(\rho_i - b u_i), \tag{13}$$

$$q_s = -b/a \quad (a > 0). \tag{14}$$

The median of the three q_s values supplies the geometric inverse TTC. This is an apparent-scale model with a query-time intercept; it does not recover metric depth or compensate for arbitrary rotation, occlusion, or annotation jitter. At least two boxes spanning 10 ms are needed for a fit. Near-zero expansion, nonfinite descriptors, or an endpoint older than 150 ms makes the geometric estimate unavailable. A two-point zero residual is not interpreted as zero uncertainty. [N01]

Twenty descriptors expose the conditions of that fit: the three axis/area rates, their median and short-history counterpart, rate disagreement, residual and residual-to-motion ratio, temporal span, normalized box count, admitted-box fraction, endpoint age, quantization support, short/long disagreement, aspect ratio and its change, center motion, and slope scatter. Absolute coordinates and absolute clock time are absent. The evaluation supplies object regions, so this branch does not establish that an autonomous detector can provide equally useful observations.

7.2 Router architecture and exact output policy

Each quality router receives 33 values. Thirteen describe the three neural magnitudes, their availability, pairwise sign agreements, geometric availability, and neural/geometric sign agreement. The remaining twenty are the geometric quality descriptors, with the first five rate coordinates replaced by their magnitudes. This design lets the weights respond to agreement and support without supplying a sequence identifier or a target-derived error. The fitted normalizer and componentwise clipping to $[-8, 8]$ precede a $33 \to 18 \to 4$ network with a tanh hidden layer. Its two affine maps contain 688 trainable parameters. The final affine weights start at zero with bias $(2, 0, 0, 0)$, favoring F0 when other experts are available without forcing an exact initial F0 output. [N01]

Let m_j indicate an available expert and let $g_j(\mathbf{x})$ be its gate logit. The router forms normalized masked weights

$$w_j = \frac{m_j \exp g_j(\mathbf{x})}{\sum_k m_k \exp g_k(\mathbf{x})}, \qquad \widehat{q} = \sum_{j:m_j=1} \frac{w_j}{\widetilde{T}_j}, \tag{15}$$

where $\widetilde{T}_j$ preserves the expert sign and limits its magnitude to $[0.1, 60]$ s. The implementation excludes zero-valued experts from inverse-coordinate fusion. It applies a rate-magnitude floor of $1/60\ \mathrm{s}^{-1}$ before inversion, yielding a finite signed output with magnitude at most 60 s. Exact cancellation, $\widehat{q} = 0$, returns the original fused prediction F0 rather than choosing an approach direction.

The delivered ensemble contains routers trained with seeds 7, 13, and 21. Each router performs Equation 15 and emits seconds; the ensemble then averages those three TTC predictions in seconds. It does not average their inverse rates. The ensemble returns F0 exactly when no valid box history is supplied, when endpoint age is nonfinite, or when the latest box is older than 150 ms. A valid geometric estimate is a separate requirement: a fresh but too-short box history can leave geometry unavailable while allowing the router to combine available neural estimates. These distinctions preserve the implemented fallback semantics. [N01]

Table 6. Corrected local EvTTC comparison on historically exposed data. MAE is in seconds; RAE is a percentage. Both columns apply the organizer-described interpolation of negative predictions for scoring only. These are query-weighted local scores, not new official results. Slider-750 remains unresolved against the organizer's score.

Population	n	MAE orig.	MAE adapt.	RAE orig.	RAE adapt.
All queries	434	0.8250	0.3359	49.894	17.250
Road	365	0.3054	0.1852	11.917	7.469
Excluding Slider-750	383	0.3296	0.1933	18.730	9.417
Slider-750	51	4.5448	1.4071	283.931	76.071

7.3 Fitting population and exposure

The final fit uses 2,608 queries: 2,531 from twenty-four road recordings and 77 from Slider-500. Each seed runs 300 AdamW updates with learning rate 0.01 and weight decay 0.01. The eight competition road recordings and Slider-750/1000 do not enter this final optimizer fit. They had, however, influenced preceding development and startup-recipe inspection. We therefore treat the subsequent comparison as historically exposed local evaluation. Excluding recordings from a final refit does not erase earlier model-selection exposure. [N01]

This delivered recipe must also remain distinct from earlier all-data fits with 2,677 queries and from the separate all-32-sequence F1 refit. Those experiments have different data membership and upstream fitting. Three router seeds quantify optimization variability conditional on the shared frozen producers; they do not independently replicate sensor acquisition or representation learning.

7.4 Corrected evaluation of the delivered adapter

The later target audit resolves the source assignment of CCRs1-low and CCRs1-medium using the public download mapping. It recovers eleven previously missing targets without fitting new clock offsets, making all 434 query identities scoreable locally. Earlier 423-row summaries are superseded for this comparison. The original October submission and the adapted submission retain their original bytes, including two and one negative predictions, respectively. For scoring only, both columns use the organizer-described interpolation of negative predictions between valid outputs. This transformation differs from the runtime's signed output policy. [N01]

Over all 434 queries, MAE decreases from 0.824976 to 0.335899 s and mean relative error from 49.894242% to 17.249803%. The respective relative reductions are 59.28% and 65.43%. Both errors decrease in each of the ten sequences under this evaluator. The 365 road queries improve from 0.305401 to 0.185200 s MAE, so the aggregate gain is not solely a consequence of Slider-750. These reductions are query-weighted; they should not be combined with the recording-macro cadence summaries below.

Slider-750 remains a protocol discrepancy. The original model's locally reconstructed relative error is 283.931058%, whereas the organizer-provided result is 132%. The other nine sequence rows agree after rounding. The adapter's 76.071442% on Slider-750 is therefore a local value and cannot be compared directly with the organizer's 132%. No verified new official score for the adapted submission is available in this evidence snapshot. The gain supports further controlled evaluation; it does not establish independent transfer superiority or preservation of the original GarlTTC Test score.

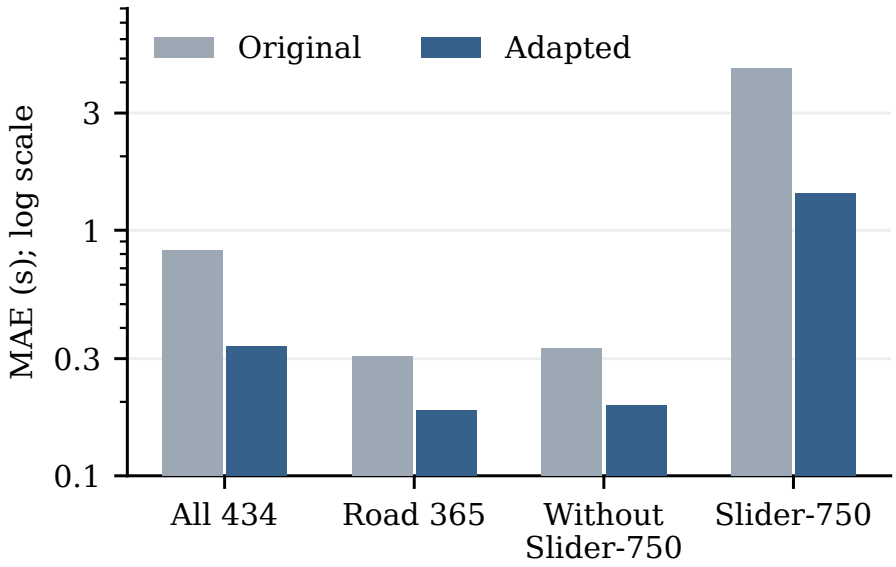


Figure 10. Local effect of quality adaptation. The corrected evaluator compares the original and adapted outputs on the same queries. A logarithmic vertical axis separates the road and Slider error scales. This figure reports exposed development evidence, with organizer-style negative-output interpolation applied to both variants. [N01]

8. Canonical cadence as a controlled intervention

FROZEN-GRAPH INTERVENTION / 149-query screen and 2,608-query development population / candidate not promoted

The metric phase uses a fixed interval, whereas the sensor supplies observations at measured intervals. These two time quantities must remain separate. The canonical-PAIR experiment asks whether normalizing the learned readout to a reference interval improves its consistency, then tests whether any local improvement survives the downstream contexts and fusion. The intervention introduces no optimizer updates into the original chain. [N02]

8.1 Recomputing the readout before rescaling

The intervention first rebuilds the 133-coordinate PAIR input with its three time-derived fields recomputed for $d_0 = 0.1$ s: duration, log-duration, and inverse duration. The frozen PAIR network processes this input and emits raw phase ϕ_c. Given the actual interval Δ, the adapter then computes

$$q_c = -\operatorname{expm1}(-\phi_c)/d_0, \tag{16}$$

$$q_\Delta = q_c d_0/\Delta, \tag{17}$$

$$\phi_\Delta = -\log(1 - d_0 q_\Delta). \tag{18}$$

Only after this transformation does it apply the original emission policy. The implementation checks finiteness and $1 - d_0 q_\Delta > 0$; unsupported transformations remain explicit failures. When Δ equals the FP32 representation of 0.1 s, a direct identity path preserves the original bytes. This is not a division of an already emitted TTC, nor a rescaling of the original phase with its time-conditioned features unchanged. [N02]

Weights, normalizers, observation cutoffs, and producer batches remain fixed. The run reconstructs the complete PHASE17 feature vector from corrected PAIR features, including disagreements, and reruns temporal contexts and the final fusion head. Predictions and missing-output status are sealed before a separate process joins targets. Each new query first reproduces a paired original execution on the same platform; an old cache is not substituted as the control.

A small time-warp diagnostic motivated this population test. It altered prepared observations and their clocks, rather than reselecting fixed-duration windows from a time-scaled raw stream. Under speed factor a, the ideal comparison is $aT_{\text{warped}}/T_{\text{original}} = 1$. Canonicalization improved all four tested F0 contrasts, yet

Table 7. Frozen F0 with canonical PAIR timing on the exposed development population. Metrics average recordings equally within each domain; availability is explicit. Road ΔMAE has a descriptive 95% recording-bootstrap interval [−0.0005047, 0.0000931] s. Slider-500 is one previously used recording.

Domain	Paired/all	MAE orig.	MAE canon.	RAE orig.	RAE canon.
Road	2531/2531	0.810620	0.810490	30.6004	30.5910
Slider-500	76/77	1.178156	1.120809	83.8859	82.4420

the road event-context ratio at $a = 0.5$ moved from 0.673790 to 1.891432. The downstream regression prevents an interpretation in which every intermediate correction must benefit the complete estimator. [N02]

8.2 Fixed screen and modality localization

The initial screen contains 72 road queries and 77 Slider-500 queries. A later four-arm localization control holds this population fixed and corrects events only, RGB only, both, or neither. In that paired CPU control, road macro MAE is 0.748852855 s for the original and event-only arms, and 0.748672302 s for RGB-only and both. The observed road change therefore comes from the RGB intervention in this control. On the 76 available Slider outputs, the corresponding values are 1.178155983, 1.163044442, 1.141746554, and 1.120808948 s. Both modalities contribute there. This localization experiment diagnoses a fixed candidate; it does not select a new recipe after inspecting the four results. [N02, N05]

8.3 Complete development population

The full experiment reuses the sealed screen and adds 2,459 road queries, yielding 2,531 road queries across twenty-four recordings and 77 Slider-500 queries. The original F0 and its canonical counterpart both return all road predictions and 76 Slider predictions. The previous Slider absence remains absent. Neither route introduces a new missing prediction or a sign change. [N02]

Road recording-macro MAE changes from 0.810620018 to 0.810489612 s. A paired bootstrap over complete road recordings, using 10,000 resamples and seed 7, gives a mean difference of −0.000130406 s and a descriptive 95% interval of [−0.000504700, 0.000093150] s. The interval includes zero. Slider-500 improves from 1.178155898 to 1.120808950 s, a 4.87% relative reduction, but consists of one acquisition already inspected in the screen. The extra road queries do not provide an independent Slider replication.

The F0 intervention passes the fixed numerical development criteria: preserve coverage, avoid worsening either error metric in either domain, and improve at least one metric. Passing that screen is narrower than establishing a statistically independent benefit or permission to replace the delivered system. The data remain historically exposed, and the measured road effect is small.

8.4 Why the downstream candidate was rejected

The frozen adapted router improves aggregates on the complete population but fails the earlier road screen. That failed criterion remains binding. A subsequent paired refit evaluates eight folds, three seeds, and original/canonical feature arms, totaling 48 fits. The canonical ensemble improves full-population road MAE from 0.539608 to 0.537801 s and Slider MAE from 0.364987 to 0.356515 s. On the fixed screen, however, road MAE worsens from 0.484752 to 0.511821 s and relative error from 16.720860% to 17.431738%. No seed passes both evaluation views. The refit is rejected; no final refit or export replaces PHASE-TTC-Adapt. These are out-of-fold ensemble scores, not the scores of the delivered all-fitting-data router. [N03]

8.5 Native replay: output sensitivity versus accuracy

The same-platform eAP comparison covers 5,805 physical queries and their 6,762 submission-template tokens, including aliases. Both arms preserve availability and sign, and the event context remains identical. For F0, median absolute output change is 1.4305×10^{-5} s, p95 is 7.6771×10^{-5} s, and the maximum is 0.006496 s. These are differences between model predictions, not errors relative to targets. The private Test12 labels were not used, so the experiment establishes no accuracy or ranking conservation. [N03]

Historical parity is a separate issue. The recaptured original is exactly repeatable locally, but every one of the 5,805 queries fails some comparison with historical intermediate states, and 504 fail the historical TTC tolerance. Local repeatability cannot remove those failures or assign their cause. The paired original/canonical design isolates the new intervention within the present platform while retaining the older reproducibility discrepancy.

A separate 120-query warm-weight raw-path benchmark measures an original p95 of 963.353 ms and a canonical p95 of 1,008.124 ms. It uses supplied boxes and potentially warm operating-system caches; it excludes detection, physical capture, and sensor transport. These measurements cover a different execution path from the five-second paced service in Section 10. Their percentiles cannot be subtracted from precomputed-input percentiles to recover preprocessing cost. The result provides no sustained 50 Hz evidence and adds a runtime cost to the proposed intervention. [N07]

9. Design analysis and controlled ablations

Three findings shaped PHASE-TTC: preserving scale change mattered more than representation stability alone; correcting an expert's phase could improve on selecting an expert; and changing the loss often traded one TTC range against another. The studies below used different data splits and pretrained models. Appendices E, F, and H give the full configurations and unsuccessful variants.

9.1 Preserve the motion variable before predicting it

Using a common object region preserves apparent expansion across observations. Independently resizing each endpoint can remove that signal before the network sees it (Figure 16). An explicit object query also distinguishes objects that share one full-frame input but have different TTC targets. The estimator consequently requires externally supplied localization. [M03]

The synthetic experiments showed why embedding stability alone was insufficient. With slowly changing nuisance features, a temporal-residual objective recovered more dynamic information than a variance regularizer. Changing those nuisance features independently every frame reversed the ordering. The geometry experiments exposed a related difficulty: stable foreground maps and accurate absolute heights did not ensure accurate changes in scale. Some models also received RGB-derived SAM masks or DINOv3 relations during training, even when their inference inputs contained only events. [E02, E12, E25, E26]

Removing paths from a trained V8 model identified which inputs its prediction used. Analytic scale alone scored 373.18 MiD, and adding the residual without transport gave 373.60. Restoring transport reduced the score to 163.23; restoring history reduced it further to 155.82. The weights remained fixed throughout these tests. Separate JEPA controls did not establish a benefit from predictive pretraining. [E21, E23, E24]

9.2 Correct phase rather than only selecting an expert

A hard router is limited to the predictions its experts already make. SIMPLEX-T can move beyond them by learning a correction around their median phase from masked temporal context. The auxiliary cost head still describes the original experts, so it does not directly estimate the corrected prediction's risk. The phase quantiles likewise require calibration before they can be interpreted as collision probabilities. [M_TEMPORAL_MODEL, M_PHASE_CONTRACT]

Extending history from H8 to H16 lowered MiD for each of three head seeds on OLD_DEV. The paired interval across all seeds excludes zero. However, one seed had already informed exploration; the two new seeds alone give a difference of −2.42 MiD with interval [−5.24, 0.04]. All heads shared the same upstream producers. Figure 11 shows both summaries; H8 remained the selected candidate. [M_SIMPLEX]

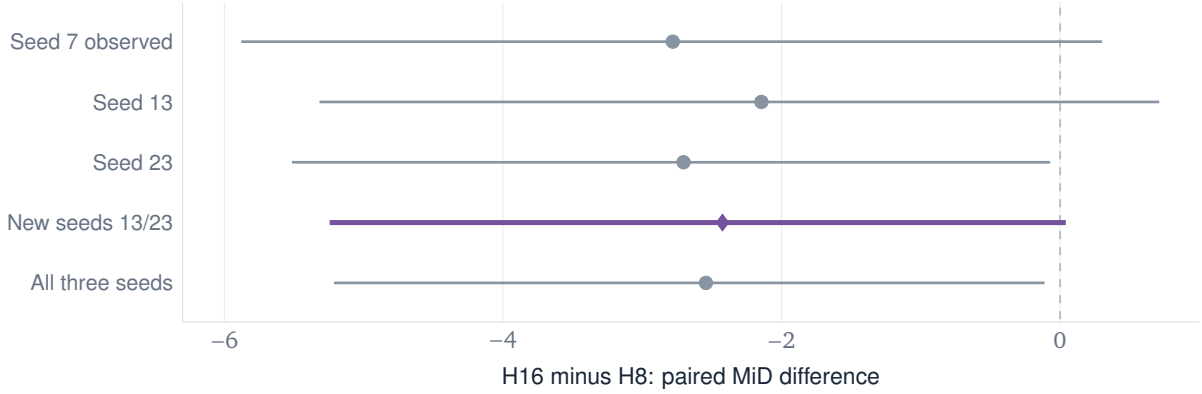


Figure 11. H16 minus H8 on nine OLD_DEV sequences. Points show paired MiD differences and lines show the recorded hierarchical 95% intervals. The highlighted estimate uses only the two new head seeds; its interval includes zero. [M_SIMPLEX]

Keeping all experts produced the lowest score in the removal study: FULL-C0 reached 121.40 MiD, compared with 127.79 for A5-only, 126.70 for C2F-only, and 124.67 for A5 plus PAIR. [M_SIMPLEX]

In the temporal-aggregation comparison, set aggregation with age gave 122.44 MiD; removing explicit time gave 123.69. Because the frozen experts can already encode temporal information, the remaining differences cannot be assigned to recurrence alone. [M_SIMPLEX]

Transport probes improved on destructive controls, but stronger calibrated or pooled alternatives narrowed that advantage. [M_STAGE61,M_STAGE62]

9.3 Adaptation trades error between TTC ranges

Changes to the readout often improved imminent-contact estimates at the expense of receding motion. V14 M1–L0 reduced crucial MiD from 156.83 to 134.47 but increased negative MiD from 61.38 to 84.21. V17's geometry bypass lowered the aggregate score on the development-used V split from 106.545 to 102.762 while worsening MAE; its head seeds shared encoders. Across V14–V17, loss reweighting, residual range, sampling, numerical precision, clipping, geometry bypasses, and height readouts failed to meet all acceptance criteria. Invalid height-derived predictions were retained in scoring. The next experiments therefore allowed the measurement model itself to adapt. [L02,L50--L52]

V18 full adaptation reduced aggregate MiD on H from 112.29 to 78.92 and crucial MiD from 157.27 to 95.85, while negative MiD rose from 65.52 to 87.09. The model passed the overall sign criterion but remained inaccurate in the magnitude of receding TTC. V19 increased the training weight on negative targets and recovered some negative-range accuracy and MAE, at the cost of aggregate and crucial-range error. [L10,L24,L25,L30]

9.4 Events help in the matched V18 comparison

Supplying real events improved V18 C1 relative to its equal-capacity ZERO control. TRUE-event fusion achieved aggregate MiD 87.29 ± 0.87, compared with 92.47 ± 0.16; dispersion is the sample standard deviation of the scalar scores across seeds. The paired comparison over complete groups also favored TRUE. Each seed repeated C1 adaptation and head fitting from the same A5 initialization and RGB context model. This experiment uses V18 C1, whereas the final TRAIN40 system derives from V16. [L03,L31--L33]

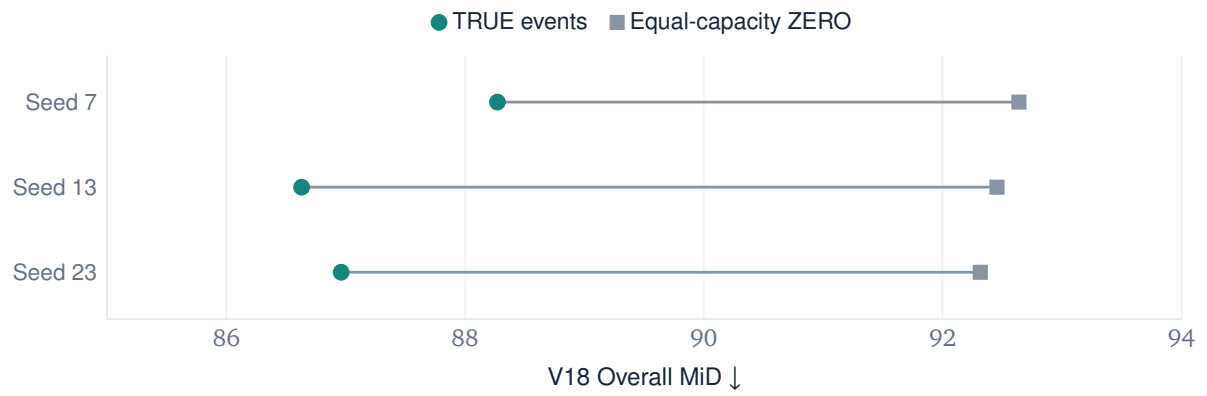


Figure 12. Real versus zeroed events in V18 C1. Each line pairs TRUE and ZERO scores for one adaptation/head seed. The models share the A5 initialization and RGB context model, and TRUE uses the observed event features. [L03]

The V18 crucial-band gain remained below its registered threshold. Its comparison with the preserved V16 model gave an interval of approximately $[-3.69, 0.14]$, and no V18 candidate met all acceptance criteria. Testing ZERO events and equal temporal support in the final TRAIN40 recipe would be needed to extend the attribution result to the submitted system.

10. Runtime and causal service

The persistent service completed every update in its five-second paced replay and preserved numerical parity, but it missed the 20 ms response target. The 95th percentile was 29.14 ms. The distinction between model execution and delivery of a complete prediction explains why a faster inner kernel is not enough to meet the deadline.

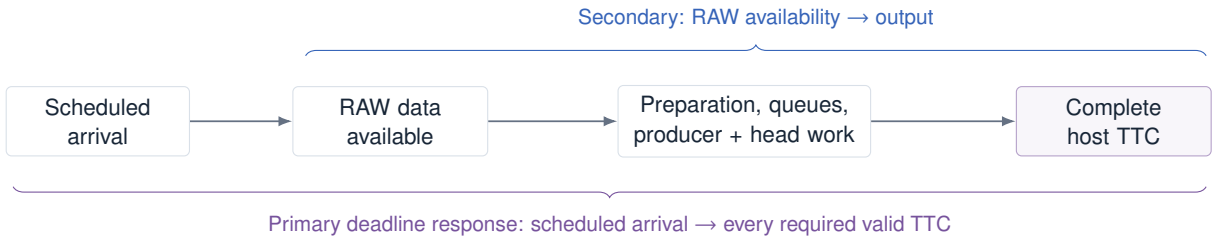


Figure 13. Response clocks for one query. Deadline response starts at scheduled arrival; the secondary clock starts after RAW acquisition, copy, and payload sealing. Both stop when all required TTC outputs reach the host.

10.1 Latency measurement boundaries

Earlier timing experiments measured different parts of the computation. V18 used warm raw-sensor queries with its original producer batching. The final test12 diagnostics timed a prepared eager path, and the EvTTC transfer measured adapter wall time, including preparation and cache behavior. Hardware and inputs also varied. We report these measurements separately and use the scheduled-arrival clock for the deadline test below. The saved timing record contains aggregate statistics, so the analysis uses quantiles and deadline counts rather than reconstructing a per-arrival distribution.

The service reduced repeated work by preparing RGB within the object region, extending event prefixes incrementally, and reusing reverse-transport summaries exactly. It also combined top-k normalization with native query graphs and parallel branches, overlapping event execution with RGB preparation. RGB features were reused; event encodings were recomputed. A separate diagnostic verified all 462 recorded calls to the prepared transport component. [L41]

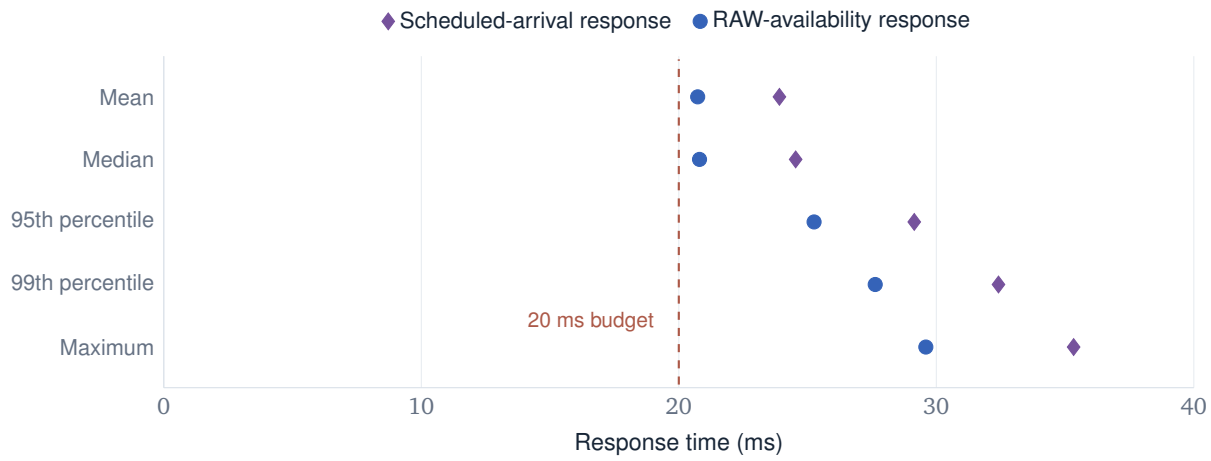


Figure 14. Response-time summaries for 250 completed updates. Scheduled-arrival latency has a p95 of 29.14 ms. Of the 250 responses, 201 (80.4%) meet or exceed 20 ms. The second clock begins after RAW acquisition, copy, and payload sealing. [L40]

Numerical checks rejected several faster alternatives. Compiled transport arithmetic changed the final TTC on sensitive examples despite agreeing on most others. Processing smaller event packets preserved the tested state buffers, representations, and final predictions, but that experiment did not time the computation or run a 200 Hz source through the paced service. It established a viable numerical path for a later timing experiment. [L42]

10.2 Paced-service latency at 50 Hz

We ran the replay on an NVIDIA GeForce RTX 5070 Ti Laptop GPU using Python 3.11.15 and PyTorch 2.11.0+cu128, built with CUDA 12.8. Computation used FP32 with TF32 disabled, two PyTorch CPU threads, eight RGB preparation workers, and a queue capacity of two. Model loading, synthetic shape priming, and initialization from preceding history finished before measured arrivals. Table 30 gives these setup intervals and the parity tolerance. [L40,R02]

The five-second replay processed 250 single-object updates from preloaded RAW data accessible only to the CPU feeder. All updates completed with zero drops and zero source errors, and independent post-run checks confirmed TTC parity. Nevertheless, 200 scheduled arrivals found earlier work unfinished. Of the 250 responses, 201 met or exceeded 20 ms. Scheduled-arrival latency reached 29.14 ms at p95, 32.41 ms at p99, and 35.33 ms at maximum. Starting the clock after RAW acquisition, copy, and payload sealing gave a lower p95 of 25.25 ms. [L40]

The measurement excludes disk I/O and covers one object over five seconds. It therefore leaves physical-camera input, sixty-second operation, and larger object counts for further testing. Failed replays remain in separate reports, including dropped updates and rapid error acknowledgements. We count latency only for successful TTC outputs, so a fast acknowledgement cannot improve the prediction-latency summary.

11. Discussion and limitations

11.1 What the benchmark result establishes

The archived 28 October ranking shows that the submitted PHASE-TTC system performs well on GarlTTC Test. Its fusion head operates on phase, timing, and support features whose meanings remain comparable across the two modalities. This avoids requiring independently trained hidden coordinates to align. The ranking alone cannot determine how much of the gain comes from that interface, the longer history, or

the upstream models; those questions require matched comparisons. This score concerns the original TRAIN40 graph, not the later quality adapter or canonical-cadence candidate. [M22,M23,N01--N03]

The out-of-fold training procedure addresses a narrower problem: later stages should not train on features produced by a model fitted to the same sequence group. At deployment, however, they receive features from producers trained on all 40 groups rather than the 24 used in each auxiliary fit. That shift remains, as does the exposure involved in earlier architecture selection. The JEPA experiments provide a separate result: they did not establish a benefit from predictive pretraining, even when embeddings retained variance and effective rank. Subsequent supervised or teacher-assisted gains do not resolve that question.

11.2 Where fusion helps, and where it fails

The labeled transfer experiment gives a mixed answer about adding events. Fusion has a lower empirical p95 absolute error than RGB alone, but a higher mean and more errors above ten seconds. All three paired intervals for mean, median, and p95 differences include zero. [L01,L02,L30,R01]

The range experiments reinforce this point. Improving imminent-contact predictions sometimes worsened estimates for receding objects, because phase error, error in seconds, and warning behavior emphasize different parts of the problem. The labeled transfer includes only positive TTC and cannot test the full signed domain or a complete warning policy.

The original 434-query competition experiment answers a different question. It establishes prediction coverage and disagreement between models under the label status recorded on 28 October. Its Garl visual-only checkpoint also differs from PHASE-TTC's RGB context model. Matching scenario names is insufficient to combine these results with the labeled transfer; the recordings and query identities must match as well. The later corrected target audit enables a separate original-versus-adapted PHASE-TTC error comparison on all 434 queries. It does not change the meaning of the historical disagreement plots or reproduce the published three-sequence Garl protocol. [T01,T02,N01]

11.3 Adaptation, cadence, and intermediate objectives

The quality adapter produces the strongest recent local improvement, reducing query-weighted MAE from 0.824976 to 0.335899 s under the corrected evaluator. The road-only reduction, from 0.305401 to 0.185200 s over 365 queries, shows that the aggregate gain is not confined to Slider-750. However, earlier startup and model-selection experiments had inspected the competition recordings. Excluding them from the final 2,608-query fit cannot restore an untouched test set. The adapted predictions also have no verified new official score in this evidence record. [N01]

Canonical PAIR timing tests a narrower hypothesis. It preserves availability and improves Slider-500, while the descriptive road MAE interval includes zero. The frozen router fails its fixed road screen, and a subsequent 48-fit comparison also fails the combined development criteria. The larger-population aggregate does not revoke that failure. This distinction matters: a correction consistent with units or ideal geometry still has to improve the complete learned estimator under the prespecified controls. [N02,N03]

The physical and representation experiments reinforce that distinction. Lower future-state loss does not establish a more useful log-height-change representation; algebraic observability does not guarantee a reliable normal-flow estimate; and ideal cuboid inversion does not guarantee better depth from predicted geometry. These negative results delimit the tested mechanisms without proving that predictive learning or physical auxiliaries can never help. Their methodological details remain part of the technical supplement. [N04--N06]

11.4 Generalization and deployment

Most development comparisons reuse sequences or upstream models. Repeating the head with different seeds measures variation conditional on those models, and nearby windows remain correlated observations of the same scene. Independent retraining and new sequences would test how much of

the observed performance survives beyond these choices. Deployment also requires obtaining object regions causally; the current evaluations supply them, and rotation-only projection between sensors approximates alignment without depth.

The compact head has 1,955 parameters, but the full system includes all upstream components. A complete parameter and peak-memory inventory remains to be reported. The paced-run records specify the GPU, software, execution settings, and initialization excluded from timing, while omitting the CPU model and GPU driver version. Learned phase widths and propagated variances likewise need empirical calibration before they can support probability statements.

The runtime study identifies an operational bottleneck: all predictions remain numerically consistent, yet their p95 response exceeds 20 ms. Input preparation and scheduling must improve along with kernel speed. A longer run with physical-camera input and multiple objects would test whether those improvements hold under sustained load. [L40--L42]

Native cadence replay introduces a further reproducibility boundary. The recaptured original is exactly repeatable on its current platform, yet historical parity fails for intermediate states in 5,805 queries and for TTC in 504. Small original-versus-canonical output changes on those queries measure local sensitivity, not conservation of Test12 accuracy or restoration of historical numerical identity. Private Test12 targets are required to score the new candidates. [N03]

12. Conclusion

PHASE-TTC combines geometric measurements and recent history from events and RGB images, then corrects the RGB estimate through a compact phase-based head. It ranks first in the archived public GarlTTC Test leaderboard of 28 October 2026, reducing Overall MiD by 20.49% relative to Garl-TTC. The original labeled transfer still shows a lower p95 alongside worse extreme errors, and the paced service still exceeds its deadline.

The expanded evidence adds a useful but historically exposed quality adaptation on corrected local EvTTC targets. It also shows why temporal normalization, predictive objectives, and physical consistency require complete-system controls: several plausible interventions failed those controls and were not promoted. Independent evaluation of the delivered adapter, resolution of Slider-750's scorer discrepancy, better rare-error behavior, and sustained response deadlines remain central requirements for stronger generalization and deployment claims.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv:2301.08243*, 2023. URL https://arxiv.org/abs/2301.08243.

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. URL https://arxiv.org/abs/2105.04906.

[3] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. URL https://arxiv.org/abs/1904.08405.

[4] GarlTTC Benchmark. GarlTTC Test public leaderboard. https://www.codabench.org/competitions/17289/, 2026. Accessed 28 October 2026. PHASE-TTC submission 1016512.

[5] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. URL https://arxiv.org/abs/2103.06011.

[6] Jinghang Li, Bangyan Liao, Xiuyuan Lu, Peidong Liu, Shaojie Shen, and Yi Zhou. Event-aided time-to-collision estimation for autonomous driving. In *ECCV*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/07043.pdf.

[7] Jinghang Li, Shichao Li, Qing Lian, Peiliang Li, Xiaozhi Chen, and Yi Zhou. Toward deep representation learning for event-enhanced visual autonomous perception: the eap dataset. *arXiv:2603.16303*, 2026. URL https://arxiv.org/abs/2603.16303.

[8] Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv:2603.14482*, 2026. URL https://arxiv.org/abs/2603.14482.

[9] NAIL-HNU. Garl-TTC official benchmark evaluator. https://github.com/NAIL-HNU/Garl-TTC/blob/main/garl_ttc_benchmark/codabench/scoring_program/score.py, 2026. Accessed 28 October 2026.

[10] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. Dinov3. *arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

[11] Vlad Sobal et al. Joint embedding predictive architectures focus on slow features. *arXiv:2211.10831*, 2022. URL https://arxiv.org/abs/2211.10831.

[12] Kaizhen Sun, Jinghang Li, Kuan Dai, Bangyan Liao, Wei Xiong, and Yi Zhou. Evttc: An event camera dataset for time-to-collision estimation. *IEEE Robotics and Automation Letters*, 2025. URL https://arxiv.org/abs/2412.05053.

Methods, controls, and complete evidence

Detailed methods, controlled interventions, sequence-level comparisons, and source mappings accompany the main results. Separate appendices document historical experiments, official evaluation, transfer analysis, and frozen competition inference, followed by the corrected local evaluation and subsequent controls.

Original evidence snapshot: 28 October 2026. Expanded research coverage: 2 November 2026.

The build uses saved evidence; separate later model fits retain their own protocols and exposure history.

A. Technical foundations and historical method families

PHASE-TTC grew through four model families: predictive embedding, geometry-constrained endpoint estimation, phase-domain temporal refinement, and RGB-anchored event fusion. Each family tests a different mechanism with its own objective and inference path. [M01--M14]

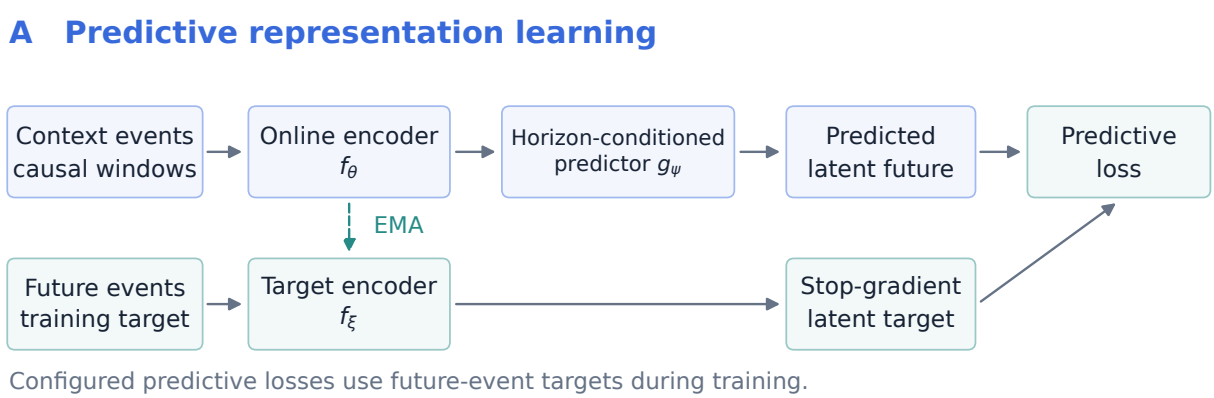


Figure 15. Selected model families. (A) Predictive latent learning uses a training-only target branch. (B) Causal scale and transport map endpoint geometry to signed motion. (C) Phase-domain fusion combines event and RGB evidence. Panel A contains the JEPA objective; Sections A.5 and 2 detail the later temporal and fusion paths.

A.1 Events, temporal windows, and object coordinates

An event stream is a sequence $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$, with integer microsecond timestamps and normalized polarity $p_i \in \{-1, +1\}$. `EventBatch` stores aligned NumPy arrays for these fields. Each context window is anchored at its end time. Indexing rejects nonpositive durations, mismatched reference times, and future windows that begin before context ends. Causality tests cover crops, features, checkpoints, and calibration inputs. [M01]

The dense voxel representation distributes an event between adjacent temporal bins. For K bins, duration Δt, and normalized time $u_i = (K-1)(t_i - t_0)/\Delta t$, the unnormalized representation is

$$V_{p,k}(x, y) = \sum_i \mathbf{1}[x_i = x, y_i = y, p_i = p] \max(0, 1 - |u_i - k|). \tag{19}$$

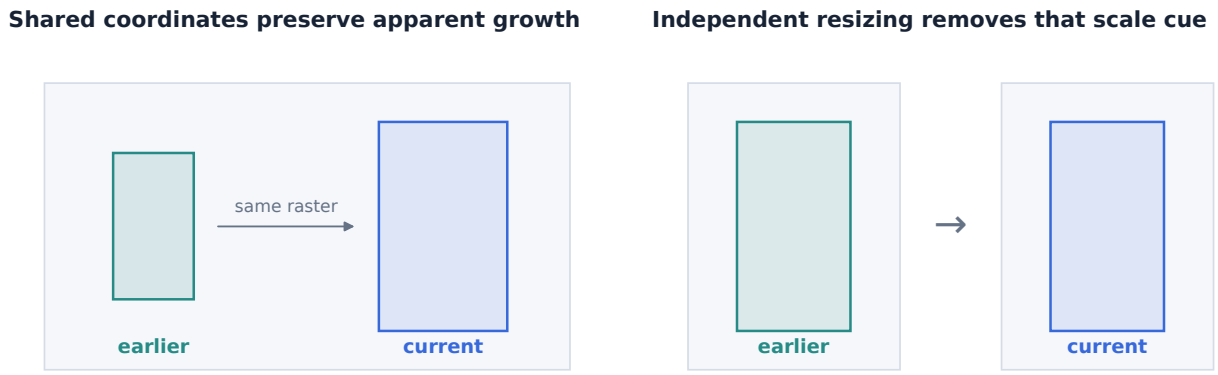


Figure 16. Shared coordinates preserve apparent expansion. A common crop maps both endpoints at one spatial scale, whereas independent tight crops normalize away part of the size change. The diagram is schematic.



Figure 17. Event and RGB appearance for one object region. FCWD1:0007 illustrates the cross-modal alignment problem. The orange region is the original visualization overlay.

Separate polarity yields $2K$ channels. Two flattened weighted histograms accumulate the adjacent-bin contributions without constructing a buffer with twice the event count. Optional robust scaling divides occupied voxels by the 95th percentile of their absolute magnitude; empty voxels remain zero. Median centering is omitted because equal occupied counts would collapse to zero. Event count, time surface, and sparse tokens remain separate representation alternatives. [M02]

Object-centred experiments use a common square region that covers both endpoint boxes, expands by a fixed margin, and maps both event windows into one raster coordinate system. Tight independent crops can erase apparent scale change. The resulting tensor is *oracle-box-conditioned*: the event backbone assumes external localization even though its forward signature has no box argument. Causal-scale tests confirm that only the designated endpoint boxes affect the crop. [M03]

The eAP adapter reconstructs local targets from public three-dimensional tracks, whereas the benchmark uses private smoothed labels. Results therefore keep reconstructed eAP targets, EvTTC labels, and official GarlTTC labels in separate evaluation populations. [M01]

A.2 Predictive embedding learning

The generic JEPA implementation encodes a context representation with an online encoder f_θ and predicts a future latent conditioned on a horizon identifier:

$$z_t = f_\theta(V_t), \qquad \widehat{z}_{t+h} = g_\psi(z_t, e_h), \qquad z^\star_{t+h} = \mathrm{sg}\big(f_\xi(V_{t+h})\big). \tag{20}$$

The target starts from the online weights and is updated by an exponential moving average,

$$\xi \leftarrow m\xi + (1-m)\theta. \tag{21}$$

For multihorizon samples, a validity mask excludes unavailable future pairs from the alignment objective. The trainer supports normalized Smooth-L1 alignment, dense token prediction, spatial or tubelet masking, and variance or `visreg` regularization. Each experiment records its active loss configuration. [M04]

The V8 attribution experiment copies the `CausalScaleTTC` endpoint encoder and predicts three views of an unlabeled third endpoint from two earlier endpoints: dense features, a global geometry token, and foreground logits. Its objective is

$$\mathcal{L}_{\mathrm{V8}} = \frac{1}{3}\left(\mathcal{L}_{\cos}(\widehat{D}_2, D_2^\star) + \mathcal{L}_{\cos}(\widehat{g}_2, g_2^\star) + \mathcal{L}_{\mathrm{SL1}}(\widehat{F}_2, F_2^\star)\right). \tag{22}$$

The scalar foreground-logit view uses Smooth-L1 because one-dimensional cosine similarity is almost everywhere constant. A frozen target encoder processes the future endpoint in evaluation mode under `no_grad`. TTC, boxes, masks, category, metadata, geometry losses, contrastive losses, VICReg, and router outputs remain outside this objective. [M05]

V8 schedules EMA momentum by cosine interpolation between configured endpoints. It measures dispersion, effective rank, covariance trace, intersample similarity, and norm for the dense, global, and foreground views. A persistent-collapse rule aborts training when these diagnostics remain below their thresholds. Downstream probes then measure TTC utility.

V8 includes strict encoder transfer, random-frozen and scratch controls, partial-finetuning allowlists, and a shuffled-future control. The shuffled control uses a deterministic cross-track derangement that preserves the future marginal while holding the batch schedule, model, optimizer, update count, masking, augmentation, and EMA schedule fixed. Track identities construct the derangement and are excluded from model inputs. The declared gates require improvement over these controls before attributing a gain to predictive temporal structure. [M05]

A.3 Spatial structure and temporal memory

The high-resolution tubelet path preserves tokens as $[B, T, P, D]$, with distinct temporal and spatial patch axes. Images are padded on the right and bottom to a ceiling-based patch grid; invalid temporal positions and spatial padding carry masks. Windowed attention mixes spatial positions within a frame before a temporal mixer receives each patch trajectory. This separation prevents flattened raster order from being interpreted as time. [M06]

A block-causal Transformer permits attention among all spatial patches at the same instant and all earlier instants, while masking later frames. If token i belongs to time t_i and token j to time t_j, its attention mask is $A_{ij} = \mathbf{1}[t_j > t_i]$. A triangular token mask would impose an arbitrary left-to-right spatial ordering within each frame, so the implementation supplies the temporal block mask explicitly.

The recurrent delta-memory alternative maintains a separate matrix state per spatial patch and attention head. Writing the channelwise retention as α_t, the implemented update is

$$\widetilde{S}_t = \mathrm{diag}(\alpha_t) S_{t-1}, \tag{23}$$

$$e_t = v_t - k_t^\top \widetilde{S}_t, \tag{24}$$

$$S_t = \widetilde{S}_t + \beta_t k_t e_t^\top, \qquad o_t = q_t^\top S_t. \tag{25}$$

Invalid patch steps retain the previous state and emit zero. State accumulation uses FP32, and the output returns to the input value dtype. Chunked execution passes explicit state between chunks and returns a detached final state, avoiding state shared between sequences. The causal recurrence uses no custom CUDA kernel. Tests establish equivalence between chunked and unchunked computation; runtime experiments determine deployment speed. [M06]

Before allocation, the implementation estimates quadratic attention-score memory, $BH(TP)^2 b$ bytes for H heads and b bytes per value, and rejects layouts above the configured safety limit. Runtime evaluation reports observed memory, preprocessing time, and throughput.

A.4 Causal scale, local transport, and TTC anchors

The geometry-constrained model learns a foreground probability field and differentiable visible extent for each endpoint. It forms geometry tokens from the encoder token, log-height, centroid, foreground fraction, and sensor support. The analytic log-scale difference is supplemented by a bounded residual,

$$r = \log h_2 - \log h_1 + \delta(a,b), \qquad \delta(a,b) = r_{\max} \tanh\left(\frac{u(a,b) - u(b,a)}{2}\right). \tag{26}$$

With deterministic evaluation and reversed transport inputs, the residual changes sign under endpoint reversal. The primary output remains geometry-bound; the auxiliary direct inverse-TTC head does not feed into it. [M07]

The timestamp anchor matters. Under constant relative speed and apparent size inversely proportional to distance, the *current-endpoint* inverse TTC is

$$q_2 = \frac{e^r - 1}{\Delta t}, \qquad T_2 = \frac{\Delta t}{e^r - 1}. \tag{27}$$

The learned-height-ratio reference implements

$$T_1 = \frac{\Delta t}{1 - h_1/h_2} = \frac{\Delta t}{1 - e^{-r}} = T_2 + \Delta t, \tag{28}$$

which is anchored at the earlier endpoint under the same assumptions. Each route uses the convention determined by its label timestamp and evaluator. The benchmark phase defined below is a metric coordinate distinct from the current-endpoint log-scale difference. [M07, M10]

The transport extension derives event-native motion evidence from dense features D_1, D_2. It builds cosine correlations in a local neighbourhood $\mathcal{N}(p)$ and obtains expected displacement by

$$P(\delta \mid p) = \frac{\exp\big(\langle \bar{D}_1(p), \bar{D}_2(p+\delta)\rangle/\tau\big)}{\sum_{\delta' \in \mathcal{N}(p)} \exp\big(\langle \bar{D}_1(p), \bar{D}_2(p+\delta')\rangle/\tau\big)}, \quad d(p) = \sum_{\delta \in \mathcal{N}(p)} P(\delta \mid p)\,\delta. \tag{29}$$

Out-of-bounds candidates are masked before softmax. The resulting summaries include translation, anisotropic and isotropic divergence, flow magnitude, confidence margin, entropy, and forward–reverse cycle error, with additional foreground-weighted variants. Coarse-to-fine models mix measurements from two scales through a learned gate. Shifted views replace a $[B, C, K, H, W]$ unfolded tensor to reduce intermediate memory. [M08]

Training-time RGB distillation introduces privileged visual supervision. The A4 DINOv3 teacher [10] supplies local cosine-similarity maps at six fixed feature-grid offsets. An L1 loss transfers these spatial relationships without requiring equal channel dimensions; slicing avoids circular wraparound. The student consumes events alone at inference, making this route privileged distillation rather than label-free self-supervision. [M09]

A.5 Phase-domain routing and temporal refinement

The later benchmark coordinate is

$$\phi(T) = -\log\left(1 - \frac{d}{T}\right), \qquad d = 0.1\ \text{s}, \quad T < 0 \text{ or } T > d. \tag{30}$$

Its inverse uses $q = -\operatorname{expm1}(-\phi)/d$. The implementation clamps phase to its configured support and saturates TTC magnitude at 60 seconds; zero phase maps to +60 seconds. Saturation defines finite numerical output, while signed negative values represent recession. [M10]

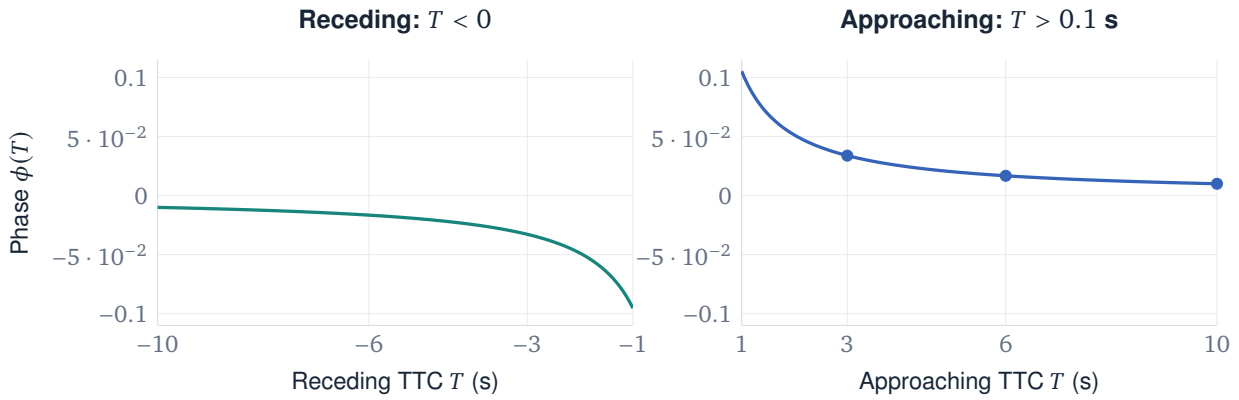


Figure 18. Signed phase geometry. Analytical curves show $|T| \in [1, 10]$ s on common vertical scales, with approaching-band markers at 3, 6, and 10 s. The zoom excludes $[0, 0.1]$ s and the adjacent singularities.

The Simplex risk router selects among frozen expert TTCs by predicting relative expected absolute-phase costs. For three sorted expert phases, convexity and the 1-Lipschitz property of $\mathbb{E}|a - Y|$ imply ordered secant slopes $-1 \le s_0 \le s_1 \le 1$. Softmax increments parameterize these constraints; integrating the slopes across expert-phase gaps yields relative costs. Hard selection returns one original expert prediction under a fixed tie order. The learned quantities are relative selection costs rather than calibrated collision probabilities. [M11]

The temporal refiner predicts a correction from a bounded feature history and four timing channels using a masked two-layer GRU; a causal Transformer provides the matched control. Histories require a contiguous valid suffix, and missing prefixes leave recurrent state unchanged. Residual mode anchors the phase at the median of the *current* experts,

$$\widehat{\phi} = \mathrm{median}(\phi_1, \phi_2, \phi_3) + 0.03\, a_\theta(H). \tag{31}$$

Separate heads predict phase-quantile widths and relative costs for the original experts; the latter do not estimate risk for the corrected output. Training combines phase absolute error, pinball losses, and relative-cost regression. Globally defined query mass supplies the weights without minibatch renormalization. Sparse-history variants retain selected slots from a longer history and recompute timing gaps from integer anchors, preserving cold starts and the original query population. [M12]

A.6 Multimodal complement and uncertainty boundaries

The later joint-measurement complement anchors on RGB phase and learns a correction from RGB features, event features, and the RGB phase itself:

$$\widehat{\phi} = \phi_{\mathrm{RGB}} + s\, g_\theta([z_{\mathrm{RGB}}, z_{\mathrm{event}}, \phi_{\mathrm{RGB}}]). \tag{32}$$

The final layer starts at zero, so the initial prediction reproduces the RGB anchor up to native phase conversion. A matched ZERO control removes the complete event-feature block, including admitted event-clock features, while retaining network capacity. When RGB is absent, the model copies the event TTC without applying correction or averaging. The paired TRUE–ZERO comparison isolates the contribution of event information to the learned RGB correction. [M13]

Uncertainty has multiple meanings across these families. The causal-scale route predicts log-variance in log-ratio space and propagates it through the local Jacobian. For $q = \mathrm{expm1}(r)/\Delta t$,

$$\mathrm{Var}(q) \approx \frac{e^{2r}}{\Delta t^2}\,\mathrm{Var}(r), \qquad \mathrm{Var}(T) \approx \frac{\Delta t^2 e^{2r}}{(e^r - 1)^4}\,\mathrm{Var}(r). \tag{33}$$

The singularity near zero expansion requires separate finite-output and known indicators. The implementation marks unsupported or near-zero log-ratio predictions unknown and derives model-based threshold probabilities from an inverse-TTC Gaussian approximation. Empirical calibration requires measurements

of coverage, calibration error, and corruption response. The temporal refiner's phase quantiles carry a different meaning from collision-class probabilities. [M07, M12]

A.7 Optimization and execution semantics

Component losses keep their true counts over each effective batch, and tail top-k selection uses the complete effective batch instead of separate microbatches. Activation recomputation preserves random-number-generator state, keeping gradient accumulation and recomputation equivalent to the stated objective. The joint-measurement loss includes teacher relations, unlike the label-free JEPA loss. [M14]

Deployment optimizations include bounded event history, chunked memory, persistent input state, reusable feature caches, and graph execution. Because each changes state lifetime, numerical-parity tests accompany every admitted path. Runtime measurements state the associated preprocessing and inference boundary.

B. Evaluation details and causal checks

This appendix describes data separation, causal checks, and runtime measurement in detail.

B.1 The unit of independence

Windows from one trajectory share appearance, geometry, object identity, and adjacent event history. We therefore partition complete sequence groups and resample at the sequence level. Small disjoint splits remain development evidence when the project has reused their sequences for diagnosis. [M16]

Recovery experiments keep every learned parent fold-local. Outer-development rows stay outside encoder fitting, teacher-derived caches, feature normalizers, and expert parents that feed the downstream model. A frozen parent still carries its prior exposure, so the analysis follows sample membership through preprocessing, checkpoints, cached features, normalization, and model selection.

A deterministic event representation can be reused when preprocessing and sample identity stay fixed. A checkpoint fitted on a larger population changes the exposure of every downstream result. File identity verifies which representation or model was used; training membership determines the strength of the resulting claim. [M17]

B.2 Paired inference at the sequence level

Comparisons pair the same eligible rows and sequences. For per-row metric differences d_{si} in sequence s, the sample-weighted estimate is

$$\widehat{\Delta} = \frac{\sum_s \sum_{i=1}^{n_s} d_{si}}{\sum_s n_s}. \tag{34}$$

A paired sequence bootstrap samples sequence identities with replacement and carries every row from each selected sequence into the replicate. Sample-weighted resampling recomputes the ratio of summed differences to summed counts; the sequence-macro calculation averages sequence means. The scorer-based comparison uses the sample-weighted form with a fixed bootstrap seed, complete sequence clusters, and no removal of invalid MiD rows. [M15]

Intervals based on few sequences have limited precision and inherit development exposure and experiment multiplicity. Random seeds measure optimization variation for a fixed split, whereas out-of-fold partitions vary the held-out groups. Results report seeds, folds, and sequences separately.

B.3 Causality tests

Prefix-invariance tests determine whether later inputs alter an earlier output. The causal-scale suite checks foreground logits, geometry tokens, pair tokens, analytic and residual log-ratios, and TTC. A legacy symmetric smoother provides a positive control by violating prefix invariance; the causal-left mode must remain invariant under the same perturbation. [M18]

Network masks alone cannot establish causality. Tests also perturb crop coordinates, normalization inputs, teacher frames, and labels. Self-supervised samples must remain unchanged when labels change, and crop tests restrict geometry to the designated endpoint boxes. [M03,M18]

Streaming tests cover state reset and update semantics. The bounded object-ROI estimator rejects timestamp rollback until reset, validates aligned packets and coordinates, and discards events outside its history. It consumes ROI-local event coordinates and supplied object observations. Paced replay then adds acquisition, queuing, temporal state, multiple objects, and host-visible output. [M19]

B.4 Matched controls and the scope of attribution

Architectural ablations match rows, seeds, training budgets, parent exposure, selection rules, and model capacity. A zeroed-input model removes modality information while retaining capacity. A shuffled-future control disrupts temporal correspondence while preserving the future marginal and optimization schedule. A frozen random encoder measures what the downstream head can recover without learned features.

A hard router selects among fixed expert predictions; a residual phase model can correct them. Teacher distillation measures the effect of privileged RGB supervision on an event backbone. Failed and vetoed arms remain in the analysis because they constrain mechanism-level explanations.

B.5 Latency, memory, and sustained service

A tensor-forward benchmark starts after tensor preparation. The paced runner instead includes a RAW producer, absolute arrivals, a bounded queue, and host-visible output. Completion occurs when the consumer returns every required prediction. Fixed arrivals expose backlog, while the results record drops, acquisition errors, incomplete histories, and invalid outputs. [M20]

The primary deadline-to-output response is

$$L_i = t_i^{\text{host output}} - t_i^{\text{scheduled arrival}}. \tag{35}$$

A secondary clock starts after RAW acquisition. Deadline evaluation reports tail quantiles, maximum response, coverage, backlog, and drops because mean response can hide missed arrivals. Warmup uses data available before the first admitted anchor. The 50-Hz milestone requires sustained cadence, numerical parity, and complete finite outputs. [M20]

The available RAW sequences are shorter than the sustained-service milestone. The replay therefore characterizes short-run latency for its hardware, population, numerical tolerance, and duration.

Implementation checks and their scope are reported in Appendix L. [M21] The official leaderboard score evaluates prediction accuracy; it provides no measurement of sensor-to-output latency or sustained service.

C. Paired transfer uncertainty and numerical cases

We analyze the saved query-level predictions in `data/transfer_queries.csv` (L46). Source key R01 maps the derived results to `revision/derived_transfer_analysis.json`; `scripts/prepare_revision.py` regenerates them.

C.1 Exact paired sequence bootstrap

The analysis uses the 468 eligible Extended-new predictions and their 32 physical-sequence groups. Each of 10,000 replicates samples 32 group identities with replacement and includes every query from each sampled group. Repeated groups contribute repeated queries. PHASE-TTC and RGB share the sampled indices. For each model, we compute pooled mean, median, and 95th-percentile absolute error, then form the PHASE-minus-RGB difference. NumPy's `linear` quantile convention and `default_rng(20261028)` define the 2.5th and 97.5th percentile endpoints.

The point differences are +0.007 s for the mean, +0.056 s for the median, and -0.556 s for p95. All three intervals contain zero (Table 4). The exploratory intervals condition on the fitted models and development-exposed sequences; they do not incorporate representation-training variability or adaptive model selection.

Differences use full-precision predictions. The mean difference rounds to +0.007 s; subtracting the separately rounded means 1.023 and 1.015 instead gives +0.008 s.

C.2 Numerical cases

The two queries below maximize and minimize the PHASE-minus-RGB absolute-error difference within the saved predictions. These post hoc examples show the range of model disagreement; their causes would require inspection of the corresponding sensor observations.

Table 8. Queries with the largest positive and negative PHASE-minus-RGB absolute-error differences in the labeled transfer set. [R01, L46]

Query	Target (s)	PHASE (s)	RGB (s)	Δ AE (s)
`CPNAO-high:13700430`	2.996	6.604	-12.903	-12.290
`CCRs-1-medium-100-overlap-100:5950352`	6.765	-25.740	-9.880	+15.861

C.3 Complete sequence comparisons



Figure 19. Sequence-level relative TTC error differences. Both panels show the 32 labeled Extended-new sequences in a common order; parentheses give query counts and negative values favor PHASE-TTC. The symmetric-log axes are linear between −5 and +5 percentage points. Display labels omit repeated overlap suffixes; Table 9 gives the full identities. [R01,L46]

Table 9. Query counts and mean RTE (%) for all 32 Extended-new sequences. Differences are in percentage points. [R01, L46]

Sequence	n	PHASE	RGB	Garl	Δ_{RGB}
CPNAO-low	16	46.95	41.09	1283.26	+5.86
CPLA-medium	15	45.43	35.53	1265.98	+9.90
CPNAO-medium	15	41.92	20.94	134.36	+20.98
CPNA-high	15	66.50	66.14	150.71	+0.36
CPNAO-high	14	55.86	75.18	137.89	-19.31
CPNA-medium	13	59.61	43.68	141.20	+15.93
CPLA-high	15	62.29	54.92	143.22	+7.37
CCRm-medium-0-overlap-0	13	14.61	14.57	57.78	+0.04
CCRm-low-0-overlap-0	16	16.40	11.85	49.69	+4.56
CPNA-low	12	35.90	44.34	62.54	-8.44
CCRs-1-high-100-overlap-100	15	19.23	22.75	44.19	-3.52
CCRs-1-high-50-overlap-50	15	18.25	17.55	41.09	+0.71
CCRm-low-50-overlap-50	15	17.70	11.67	30.65	+6.02
CCRs-3-low-100-overlap-100	13	8.35	15.08	18.67	-6.73
CCRm-medium-50-overlap-50	15	20.93	18.41	30.93	+2.53
CCRm-medium-100-overlap-100	14	15.25	13.79	25.04	+1.47
CCRs-1-low-50-overlap-50	16	4.98	8.93	13.15	-3.95
CCRm-low-100-overlap-100	13	20.69	15.96	28.85	+4.73
CCRs-1-medium-0-overlap-0	14	6.72	6.17	13.14	+0.55
CCRs-1-high-0-overlap-0	15	17.32	21.49	23.61	-4.17
CCRs-1-low-0-overlap-0	13	9.81	9.73	15.58	+0.07
CCRs-2-high-100-overlap-100	11	13.75	12.35	18.65	+1.40
CCRs-2-medium-100-overlap-100	15	6.84	7.39	11.36	-0.55
CCRs-2-low-100-overlap-100	16	15.67	13.62	19.44	+2.05
CPLA-low	15	25.19	26.09	28.38	-0.90
CCRs-1-medium-50-overlap-50	15	5.73	11.32	8.66	-5.60
CCRs-3-medium-100-overlap-100	16	15.32	18.54	9.91	-3.21
CCRs-side-high	16	21.54	25.26	15.77	-3.72
CCRs-1-low-100-overlap-100	15	395.63	273.53	378.78	+122.10
CCRs-side-low	16	41.76	63.28	24.48	-21.52
CCRs-side-medium	16	46.06	65.11	27.92	-19.05
CCRs-1-medium-100-overlap-100	15	111.28	131.87	83.19	-20.60

D. Competition adaptation and complete prediction traces

This appendix details input construction and timing for the original 434-query competition inference snapshot of 28 October. Linear and symmetric-log views show the full signed prediction traces. The corrected target audit and quality-adapted PHASE-TTC comparison added on 1 November appear in Section 7; they do not alter the frozen predictions shown here. [T01, N01]

D.1 Frozen models and causal input construction

We use the public Hugging Face full, event-only, and visual-only checkpoints with their released configurations and source revision. Strict state-dictionary loading and float32 inference keep each model fixed. The release identifies the checkpoint bytes but does not document complete training membership, so the evaluation represents frozen-checkpoint adaptation rather than a claim of clean zero-shot transfer.

Garl-TTC predicts the ratio of object heights at two endpoints and converts it to TTC as

$$\widehat{T} = \frac{\Delta t_{\text{measured}}}{1 - \widehat{h}_0/\widehat{h}_1}. \tag{36}$$

The released configuration sets the native model field dT = 0.1 s. For Equation 36, the adapter substitutes the measured separation of each endpoint pair while leaving the released representation unchanged. The checkpoint metadata do not establish the cadence used during training.

The full and visual routes select the latest real RGB frame at or before a query, then the preceding frame closest to 100 ms earlier. Full Garl builds one causal event window per RGB endpoint with support capped at $\min(\Delta t_{\text{measured}}, 100\,\text{ms})$. Event-only Garl uses two consecutive 100 ms windows after startup and splits the observed causal prefix into two intervals for earlier queries. Each endpoint uses the released 20-plane, polarity-agnostic time volume. Queries without sufficient history remain gaps.

PHASE-TTC uses its declared 650 ms causal history. The shared query population therefore compares the frozen systems under their respective temporal support.

Supplied boxes use zero-based, half-open event-plane coordinates. The adapter selects the latest box available at each endpoint; if the endpoint precedes all query-causal boxes, it uses the earliest such box and records availability at query time. Road sequences pair right Blackfly RGB with right Prophesee events. Slider uses its supplied streams and a rotation derived from camera poses because its calibration lacks a named event-to-RGB extrinsic. Calibrated intrinsics, distortion, and rotation-only ray projection map the event box into RGB; unavailable object depth precludes translational correction. A common square edge spans the endpoint boxes, with endpoint-specific event centers and the current center applied to both RGB frames.

D.2 Published three-sequence comparison

Table VI of the eAP paper reports Garl-TTC on three EvTTC sequences, with an unweighted mean RTE of 10.60% [7]. That study uses a separate labeled population; shared scenario names do not establish shared recordings or query identities. [T02]

Table 10. Garl-TTC's published zero-shot relative TTC error on three EvTTC sequences, as reported in the eAP paper, Table VI. [T02]

Published sequence	RTE (%)
CCRs2-medium	8.31
CCRs2-high	10.56
CCRm-medium	12.93
Unweighted average	**10.60**

Table 11. Garl–PHASE prediction comparisons for all ten sequences. Each entry gives shared-finite coverage and median absolute prediction difference in seconds. [T01]

Sequence	Full n/N; med.	Event n/N; med.	Visual n/N; med.
CCRs1-low	49/50; 0.465	50/50; 0.776	49/50; 0.213
CCRs1-medium	42/43; 0.308	43/43; 0.863	42/43; 0.200
CCRs1-high	29/30; 0.207	30/30; 0.519	29/30; 0.111
CCRs2-low	45/46; 0.221	46/46; 0.339	45/46; 0.243
CCRs2-medium	54/55; 0.178	55/55; 1.045	54/55; 0.254
CCRs2-high	58/59; 0.188	59/59; 0.682	58/59; 0.223
CCRm-low	41/42; 0.270	42/42; 0.703	41/42; 0.283
CCRm-medium	39/40; 0.151	40/40; 0.839	39/40; 0.350
Slider-750	49/51; 0.408	51/51; 0.470	51/51; 0.258
Slider-1000	18/18; 0.288	18/18; 0.119	18/18; 0.067

D.3 Runtime boundary and interpretation

Table 12. Offline processing time in seconds per query. Model time covers the synchronized forward pass and host output; full time includes tensor preparation. The Garl variants ran in full/event/visual order; PHASE-TTC used its own adapter and history length. [T01]

Model	Model med.	Model P95	Full med.	Full P95
PHASE-TTC	0.753	7.820	2.687	9.073
Garl full	0.039	0.045	0.542	1.009
Garl event-only	0.007	0.011	0.350	0.806
Garl visual-only	0.007	0.011	0.158	0.221

Timing separates tensor preparation from the synchronized model-to-host call and excludes loading, indexing, calibration, source hashing, and result serialization. The variants run once in full, event-only, visual-only order, so operating-system and HDF5 caches and GPU warm state can carry between runs. Table 12 reports these observations alongside PHASE-TTC, whose adapter and temporal support differ.

Independent verification accounts for each planned record and replays the float32 height-to-TTC conversion. Full and visual Garl remain partial because of input gaps; event-only contains all 434 rows. Across the ten sequences, the four routes exhibit distinct signed behavior and disagreement tails. [T01]

D.4 Target audit and evaluation boundary

The later audit recovers all 434 local TTC targets after correcting the public source mapping for CCRs1-low and CCRs1-medium. It preserves the recorded clock offsets rather than fitting timing shifts to match errors. Earlier 423-target summaries are superseded. The corrected original-versus-adapted comparison applies the same documented handling of negative predictions to both columns and reports complete population counts. The historical signed traces below retain their original values and gaps. [N01]

For Slider-750, the corrected local evaluator gives 283.931058% relative error for the original PHASE-TTC predictions, compared with 132% in the organizer-provided result. The other nine rows agree after rounding. The adapted local value of 76.071442% therefore cannot be compared directly against the organizer's 132% as if both used an equivalent scoring contract. The unresolved target/evaluator correspondence and the cohort's historical development exposure remain explicit limitations. This appendix supplies no new error ranking for the three Garl variants. [N01]

D.5 All ten sequences, with missing inputs preserved

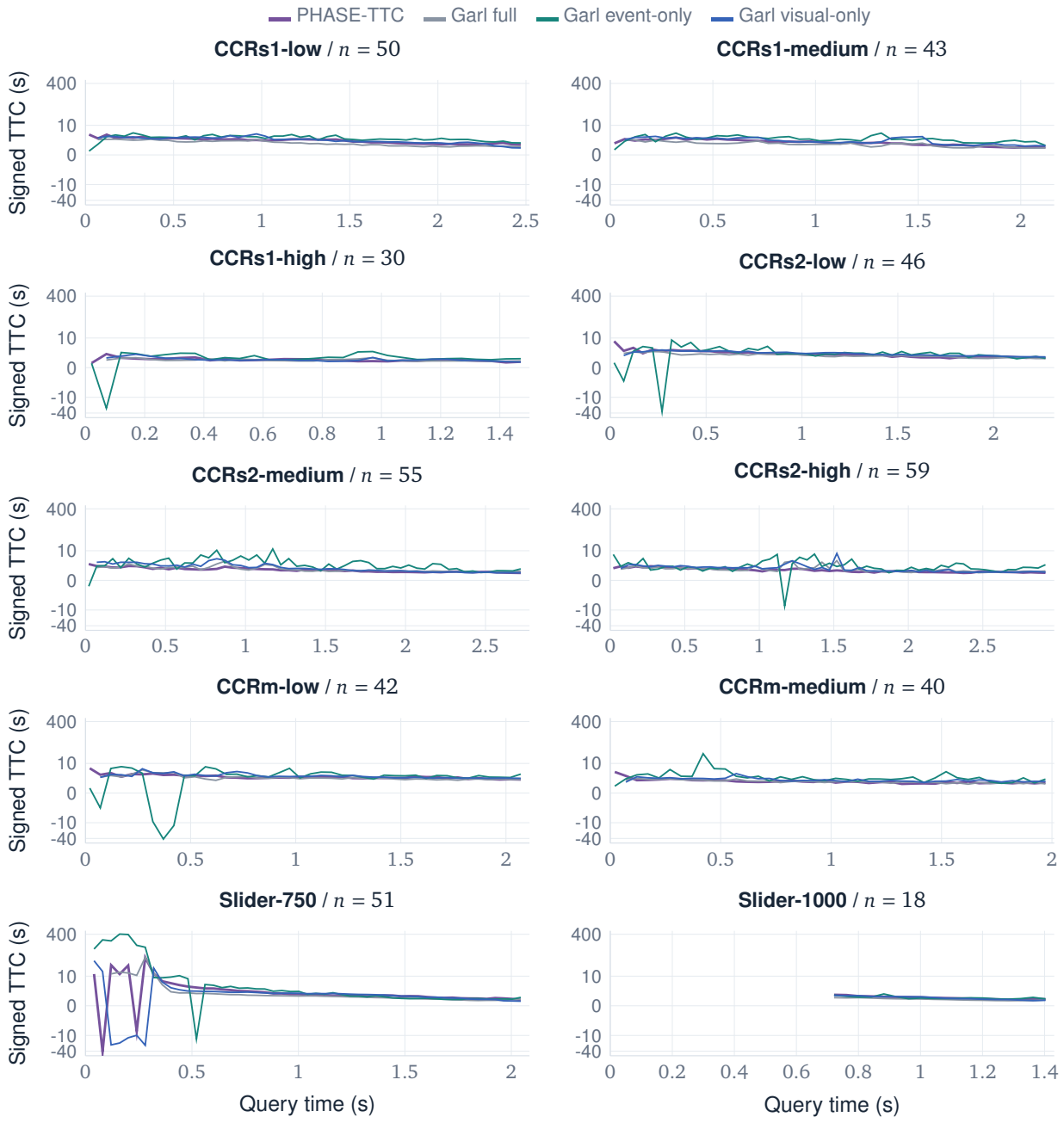


Figure 20. Signed predictions on a symmetric-log scale. PHASE-TTC is purple, Garl full gray, event-only teal, and visual-only blue. All panels are linear inside $[-2, 2]$ s and logarithmic outside; gaps mark unavailable inputs. [T01]

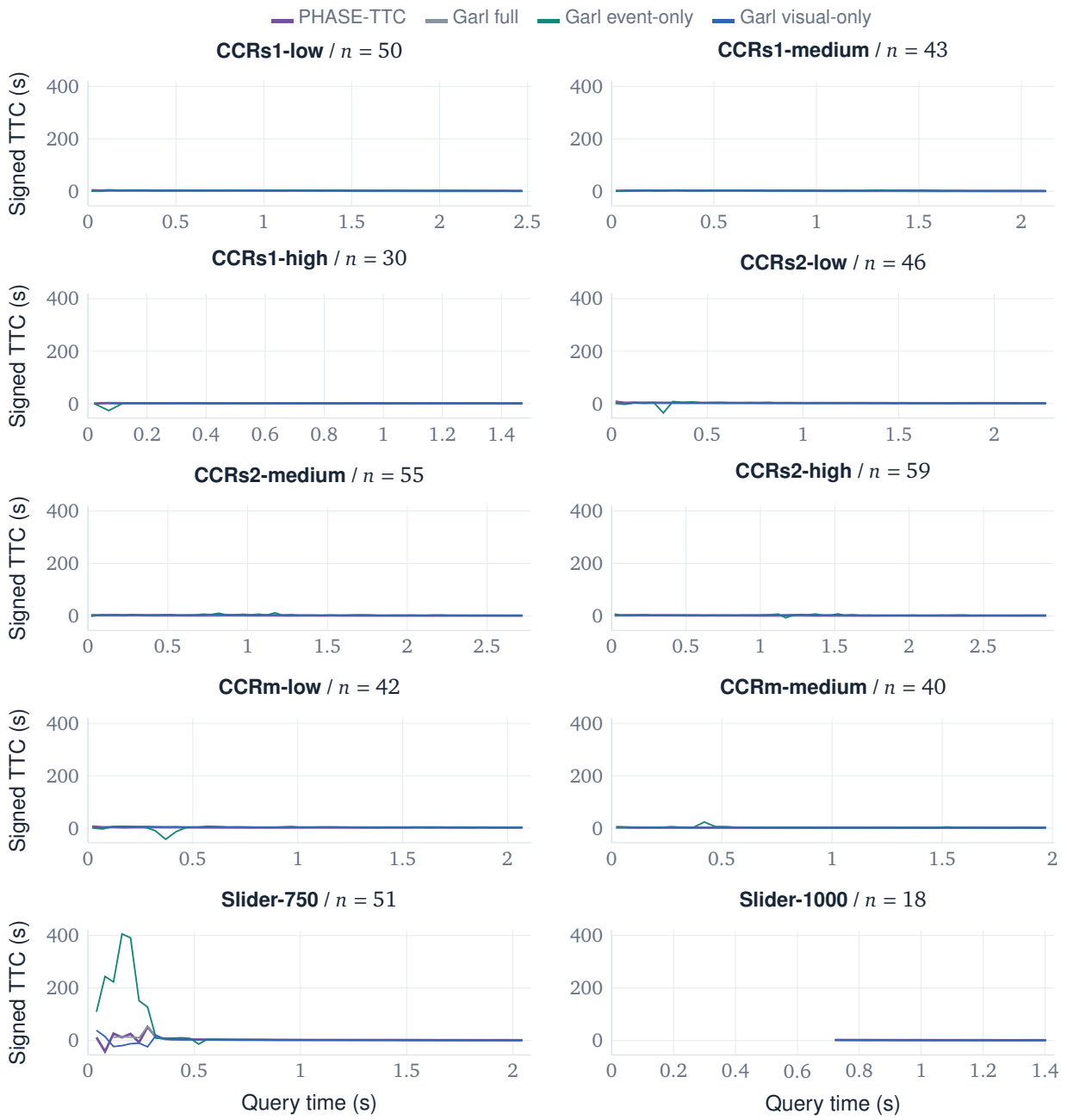


Figure 21. Signed predictions on a common linear scale. The same outputs as Figure 20 show the full excursion range, which compresses the central values. [T01]

E. The initial representation-learning programme

PHASE-TTC began as E-JEPA-TTC, a study of predictive event representations for time-to-contact estimation. The programme covered pooled and spatial embeddings, apparent expansion, and transport-conditioned residual dynamics. Some models improved TTC while losing the geometric behaviour that motivated them, so the experiments assessed accuracy, geometric consistency, and representation content separately.

The historical names distinguish otherwise similar version numbers. *Causal Scale v5–v8* refers to synthetic operator tests, whereas *Scientific Recovery V5–V8* refers to later real-data studies. The EvTTC and eAP A0/A1 experiments also used different data and metrics.

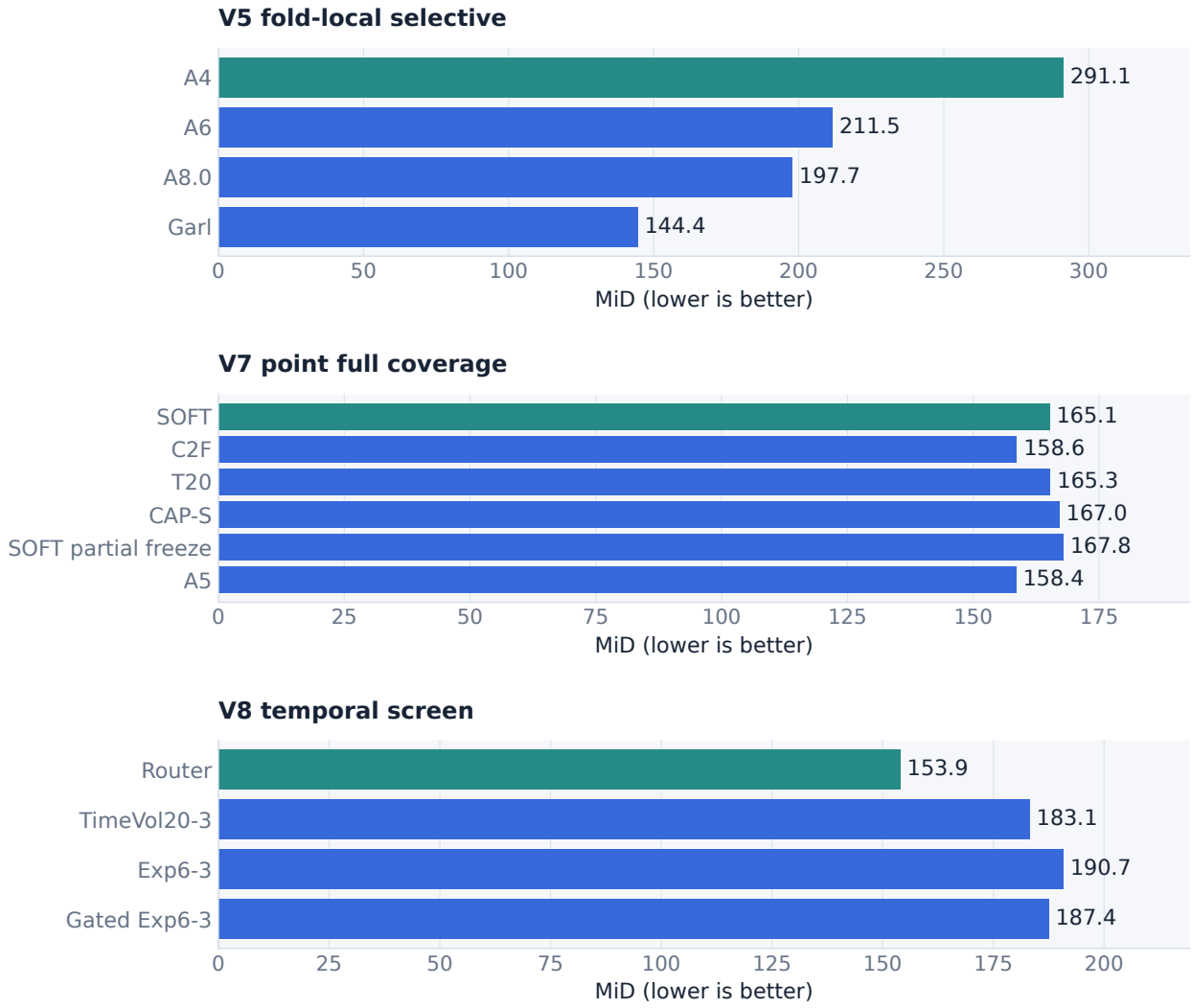


Figure 22. Early recovery results by experiment family; lower MiD is better.

E.1 Global and dense representations on EvTTC

Global A0 remained preferable to Dense A1 in the matched EvTTC comparison. Five grouped folds and seeds 7, 13, and 21 produced fifteen paired runs per model (Table 13). [E01]

Dense A1 reduced mean absolute error, but increased both relative error and the registered composite selection score. It also used 2.16 times A0's measured latency and 1.58 times its training time. We retained the smaller global model. [E01]

Architecture	MAE (s)	Relative error (%)	Selection score
Global A0	1.0109	30.25	0.58452
Dense A1	1.0067	30.55	0.59312

Table 13. Global A0 and Dense A1 means over five EvTTC folds and three seeds.

Attention residuals, temporal KDA/Object-KDA, object-conditioned crops, and FlowMimic-style alignment all failed their selection gates. Because those screens used different data or criteria, they do not extend the matched A0/A1 ranking.

E.2 Representation variance and dynamic information

The temporal-residual objective captured the desired dynamics under a slow synthetic nuisance process. Its mean dynamic R^2 was about 0.72, versus 0.15 for variance regularization. The latter maintained high embedding variance while preserving nuisance information, showing that variance alone was a poor proxy for useful temporal content. [E02]

Changing the nuisance on every frame reversed the result: the residual objective's dynamic R^2 became negative, while the variance control reached about 0.74. Later probes kept both level and change features and varied the nuisance process. Neither effective rank nor per-dimension standard deviation demonstrated TTC-relevant content on its own. [E02, E12]

E.3 Object identity and crop geometry

Object-level eAP/GarlTTC requires an explicit query because several annotated objects can share one event frame while having different TTC values. Object-LHR resolved the ambiguity by cropping the selected object at both endpoints before estimating temporal scale.

Independent resizing erased much of the apparent scale change required by the geometric head. Object-Expansion v2 stabilized the output by predicting sign and log inverse-TTC magnitude; Object-Signed-Expansion v3 combined them in one expansion coordinate and added ordered temporal features plus box motion. Perturbation tests showed strong dependence on box motion and little response to events, so this accuracy could not be credited to event representation learning.

Object Event v4 placed three causal endpoints in one common crop, preserving apparent growth and separating the neural event path from the motion ablation. The object query and its coordinate system became part of the model specification. Earlier variants used different conditions and remain separate experiments.

E.4 Physical tests of the scale operator

The frozen v4.31 matcher passed representation-stability checks but failed the zoom, reversal, and translation requirements. We also found uncommitted source changes and a missing evidence stage, so the run was ineligible for selection. Its foreground, height, pseudoflow, and correspondence diagnostics motivated a new scale estimator.

For two apparent heights h_{t-1} and h_t separated by Δt, the Causal Scale operator defines

$$r_t = \log h_t - \log h_{t-1}, \qquad \widehat{\tau_t^{-1}} = \frac{\exp(r_t) - 1}{\Delta t}. \tag{37}$$

Under idealised constant-speed approach, this gives inverse TTC at the newer endpoint. Changes in shape, viewpoint, occlusion, or relative motion can violate the geometric assumptions. For $\hat{\tau}_t = \Delta t/(\exp(r_t) - 1)$, the sensitivity is

$$\frac{\partial \hat{\tau}_t}{\partial r_t} = -\frac{\Delta t \exp(r_t)}{(\exp(r_t) - 1)^2}. \tag{38}$$

The derivative grows without bound as r_t approaches zero. Output clipping can contain the resulting TTC value, but it cannot improve the scale measurement.

The learned foreground failed to reproduce the ideal operator on held-out synthetic data. Causal Scale v5 reached correlation 0.92 with translation leakage 0.027, missing both requirements. V7 reduced leakage to 0.0034 and symmetric TTC error to 0.25, but correlation stayed at 0.92 and the validation result did not transfer to the new synthetic group. [E13, E14]

Causal Scale v8 selected on both mean and worst-group performance across independent synthetic groups. Fixed symmetric smoothing supported sparse endpoints, and a final CVaR loss emphasized large ratio errors. Correlation remained below the gate in every group. We nevertheless continued to a real-data screen, treating it as exploratory. [E15]

E.5 Foreground supervision in the first eAP screens

The first eAP screens used a common object crop, three event endpoints, and oracle boxes; they did not evaluate detection. A0 learned a weak rectangular foreground target. A1 replaced it with height, width, and centre supervision. A1-FR added a full-resolution decoder, A1-DF used deeper features, and A1-DF-R supervised the pair ratio directly. A3 kept the A1 architecture and trained with RGB-derived SAM masks. All screens reused public validation data.

Arm	MiD ↓	Failure (%)
A1 geometry	346.829	9.961
A3 SAM distilled	353.635	10.889
Garl trained matched	203.634	0.000

Table 14. Early eAP public-validation results for A0, A1, A3, and matched Garl.

A3 did not improve on A1, and both scored worse in aggregate MiD than matched Garl. The diagnostics showed that estimating absolute height did not guarantee recovery of its temporal derivative. Garl achieved lower aggregate error but predicted positive TTC for receding cases, leaving signed motion unresolved. [E25, E26]

A4 distilled local spatial relations from a frozen DINOv3 RGB teacher into an event student; A4D also learned changes in those relations across endpoints. Both used RGB during training and events at inference. A5 and A6 then varied TTC specialization and transport while monitoring the inherited geometry.

E.6 Scientific Recovery V5: fold-specific producer training

The first grouped-development run inherited an A4 parent trained on all nine development sequences, including each held-out outer group. We excluded it from promotion and retrained A4 inside every outer fold. The corrected study initialized A6 and A8.0 from the matching fold-local parent and trained Garl from scratch. [E03]

The comparison matched samples, targets, training budget, metric, and oracle-ROI access. The E-JEPA-family branch used three endpoints and twelve channels; Garl used two endpoints and forty channels. Checkpoint selection still used the outer-development split, and TTC buckets shaped the initial 8,192-token sample. Thus, the correction removed parent exposure without turning the study into a blind test or a target-free sample. [E03]

A8.0, which freed transport while retaining the frozen parent geometry, improved MiD over A6 by 13.8 points; the paired 95% interval was $[-17.7, -10.1]$. It still exceeded the preregistered MiD threshold of 175 and remained behind Garl, so V5 produced no candidate. [E03, E05]

Sequence-scoped track identifiers and preserved token/target IDs made the fold-local correction auditable through the full training lineage.

Model	Parameters	MiD ↓	Failure (%)
A4 parent	355,118	291.088	9.448
A6 transport	498,130	211.509	7.849
A8.0 dual transport	627,827	197.691	7.019
V6.1 radius 2	627,827	194.122	6.689
A5 causal	424,274	155.472	4.761
Garl local matched	24,674,178	144.353	0.000

Table 15. Scientific Recovery V5–V6 results on 8,192 tokens from nine sequences, three folds, and seed 7.

E.7 Scientific Recovery V6–V7: accuracy and geometric retention

Increasing the transport radius from one to two did not yield a clear V6 gain: V6.1 reduced mean MiD relative to A8.0, but the paired interval $[-8.2, 1.0]$ crossed zero. A5 achieved lower TTC error than the geometry-frozen variants while weakening their geometric dynamics, which kept accuracy and mechanism retention as separate criteria. [E04, E06]

V7 tested soft geometric distillation, coarse-to-fine transport, extra temporal bins, greater capacity, and partial freezing. Under full-coverage point evaluation, A5 scored 158.4 MiD and served as the reference. Promotion required both lower TTC error and retained geometric behaviour. [E07, E08, E09, E10, E11]

Arm	Point MiD	Delta vs A5	Paired 95% interval
A5 reference	158.449	–	–
SOFT	165.116	+6.668	[+3.084, +10.547]
C2F	158.573	+0.125	[-3.025, +3.432]
T20	165.260	+6.812	[+2.290, +11.514]
CAP-S	167.025	+8.576	[+3.735, +13.630]
SOFT partial freeze	167.826	+9.378	[+5.359, +13.272]

Table 16. V7 full-coverage MiD relative to A5, using seed 7 and three outer folds.

Every V7 arm failed at least one mechanism, geometry, or confirmation gate, and none advanced to confirmation seeds. More temporal bins, capacity, and partial freezing did not close the comparator gap. The JEPA objective was inactive in these runs, so their results provide no evidence about JEPA pretraining. [E07, E08, E09, E10, E11]

A post-hoc A5/C2F analysis found complementary errors. We separated a deployable feature-based router from a target-aware oracle and used the result only to design V8's prospective nested-routing test.

E.8 Scientific Recovery V8: nested routing and JEPA attribution

Only the nested A5/C2F router passed V8's candidate screen; all three temporal-representation interventions failed it. [E16, E17, E18, E19]

Arm	MiD	Finite (%)	Stored gate
router	153.877	100.00	True
timevol20_3	183.092	95.25	False
exp6_3	190.711	94.10	False
gated_exp6_3	187.400	94.80	False

Table 17. V8 seed-7 MiD and finite coverage for the router and temporal-representation arms.

The seed-7 router improved on A5 by 4.6 MiD points. Its scores were 153.9, 153.3, and 156.0 for seeds 7, 13, and 23. All three runs reused the same development sequences and estimate optimization stability, not independent external replication. Matched Garl seed replications were unavailable. [E16, E20]

The stored V8 summaries support point estimates, finite coverage, and gate outcomes. They do not specify enough aggregation detail to reconstruct valid uncertainty intervals, so none are reported.

A post-training path replay located the trained model's dependence. The analytic height path had high error, and adding a residual without transport changed little. Transport plus the residual produced the main improvement; history added a smaller gain. Because the replay disabled paths without retraining, it measures use inside this trained graph rather than the effect of retraining each ablation. [E21]

Replay intervention	MiD ↓
analytic only	373.176
analytic residual	373.596
analytic residual transport	163.231
analytic residual transport history	155.824
analytic transport	373.176

Table 18. V8 post-training path replay showing MiD for the analytic, residual, transport, and history paths.

The H2 diagnosis was shortcut/residual dominance: the analytic height path did not explain the trained model's advantage. We next tested a measurement interface with less dependence on explicit height estimation. The diagnosis did not rule out useful geometry or transport features. [E22]

V8 tested JEPA attribution through D0–D4. The meta-router had no single transferable encoder, so A5 supplied the downstream reference. D2 froze the JEPA encoder, D3 fine-tuned part of it, and scratch, random-frozen, and shuffled-future arms controlled for downstream capacity and temporal correspondence. [E24]

Arm	1%	5%	10%	25%	100%
D2	379.95	379.88	379.94	379.79	379.99
D3	379.52	379.38	379.39	379.32	379.67

Table 19. V8 JEPA attribution across nominal label fractions for D0–D4.

D2 and D3 produced finite predictions with no recorded failures, but neither passed the causal-benefit gate against scratch, random-frozen, and shuffled-future controls. V8 provided no positive JEPA attribution, and the next experiments focused on the measurement interface and feature usefulness. [E23]

F. Temporal information and phase refinement

V9–V12 asked whether causal temporal information improved phase estimates. The controls removed transport fields, disrupted correspondence or order, reduced the expert set, and replaced recurrence with unordered aggregation. Prespecified comparisons used sequence-level uncertainty and kept separately trained producer families distinct.

F.1 E-Clock: incremental information in transport features

E-Clock represented signed TTC in benchmark phase. With $\Delta = 0.1$ seconds,

$$\phi(T) = -\log(1 - \Delta/T), \qquad d(\hat{T}, T) = 10^4 |\phi(\hat{T}) - \phi(T)|. \tag{39}$$

The phase exists only on the admissible signed-TTC domain, and d measures phase error rather than seconds. Historical A5 replay, raw inverse-TTC evaluation, and bounded deployment output follow different numerical conventions and are reported under their original definitions.

Transport helped X0's direct-phase model, but did not close the gap to A5. With architecture, capacity, matcher, and input structure fixed, DYN-U scored 225.0 MiD versus 270.6 for the zeroed BASE-U control. Canonical A5 and PAIR-U scored 158.4 and 158.1. X0.5 next conditioned the transport fields on A5. [M_X0]

Table 20. Pair routing (Stage 61) and local-field experiments (Stage 62).

Arm	MiD	Queries	Finite
Stage 61			
R1	152.855	8,192	100%
R2	150.876	8,192	100%
R2_SHUFFLE	185.429	8,192	100%
RouterR	153.877	8,192	100%
Stage 62			
RouterR	153.877	8,192	100%
X2_A5_REPLAY	162.205	8,192	100%
X2_GLOBALPOOL	155.351	8,192	100%
X2_LOCALFIELD	155.463	8,192	100%
X2_SHUFFLEFIELD	155.827	8,192	100%
X2_TIMESWAP	155.585	8,192	100%

Sequence-macro MiD, evaluated on nine sequences and three folds per stage. Lower is better. [M_STAGE61,M_STAGE62]

X0.5 found no confirmed incremental value from nine frozen transport fields. DYN9 scored 155.2 MiD, against 158.7 for calibrated A5 and 175.0 for a target-blind within-sequence shuffle. The difference from calibration was −3.5 with hierarchical 95% interval $[-11.4, 3.5]$; the interval against raw A5 also crossed zero. Leave-one-sequence-out tuning kept normalization and ridge fitting outside the outer meta-test. DYN9 failed the utility gates, so the conditional X1 adapter was not trained. [M_X05,M_X05_GATE]

F.2 Pair routing and local motion fields

R2 gave Stage61's lowest point estimate but failed the utility gate. Across 8,192 queries from nine sequences, MiD was 150.9 for R2, 152.9 for R1, 153.9 for RouterR, and 185.4 for shuffled R2. The paired hierarchical intervals comparing R2 with R1 and RouterR crossed zero. [M_STAGE61]

Stage62 could not isolate a local-field benefit. LOCALFIELD scored 155.5 MiD, close to GLOBALPOOL at 155.4, SHUFFLEFIELD at 155.8, and TIMESWAP at 155.6. Although all improved on this producer family's 162.2 A5 replay, LOCALFIELD failed the mechanism gates against its controls. [M_STAGE62]

The canonical V7 A5 and the A5 retrained inside the nested router use different checkpoints and training exclusions. Their scores remain separate despite overlap in architecture and queries.

F.3 Temporal features, regret estimation, and geometry support

Stage63 repaired the experimental base by standardizing temporal bins and restoring one FP32 producer state. Equivalent fixed-grid pooling also made CUDA backward deterministic on the prescribed shapes. [M_STAGE63_REPAIR]

RAW-L1 failed Stage64's preregistered gates at 162.2 MiD. COUNT-L1 scored 157.2, temporal permutation 164.3, and STATE-L1 150.7; lower control scores did not replace the registered primary arm. Stage65's RISK17 router reached 146.0, but its recovered record retained `INTEGRITY_BLOCKED`, so it was not accepted. [M_STAGE64,M_STAGE65,M_STAGE69_DECISION]

SIMPLEX17 improved on frozen RISK17 by 2.0 MiD points, but both uncertainty intervals crossed zero. Its score was 144.0, with hierarchical 95% interval $[-5.5, 1.2]$ and sequence-only interval $[-4.7, 0.3]$ for the difference. SIMPLEX8GATE, UNCONSTRAINED17, and PAIRPERM scored 143.9, 144.5, and 149.2. The four arms completed 12 fits and 18,000 CPU float64 updates across three folds, with endpoints frozen before outer evaluation. SIMPLEX17 remained primary and failed the utility gate. [M_STAGE67_DIAGNOSTICS,M_STAGE67_DECISION]

Stage68 stopped before fitting because the normal-flow estimator supplied too little valid geometry. Only 368 of 8,192 tokens had any valid window; 583 of 16,384 windows supported the full-affine fit,

Table 21. H16 and H8 temporal heads on OLD_DEV.

Head seed(s)	H8 MiD	H16 MiD	Δ MiD	Hierarchical 95% CI
7 (observed)	121.646	118.866	-2.781	[-5.879, +0.301]
13	121.841	119.697	-2.144	[-5.317, +0.712]
23	121.865	119.162	-2.703	[-5.514, -0.071]
New seeds 13/23	121.853	119.429	-2.424	[-5.244, +0.041]
All three seeds	121.784	119.241	-2.543	[-5.213, -0.111]

Nine sequences, 8,192 queries, and three folds. Delta is H16 minus H8; multi-seed rows average paired losses. Seed 7 had been evaluated previously. All heads share fixed expert producers. [M_SIMPLEX]

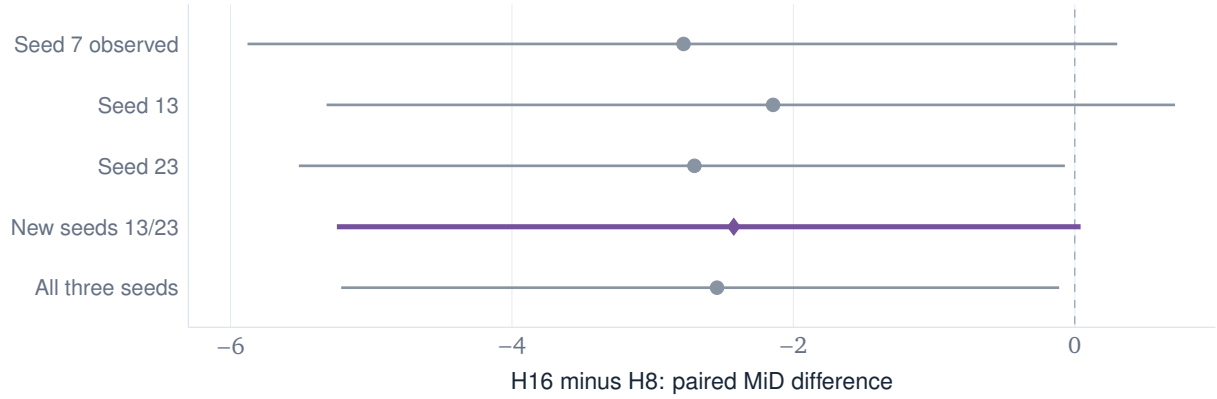


Figure 23. Paired H16-minus-H8 MiD differences for three head seeds on OLD_DEV.

and none supported the reduced-isotropic fit. Fold support rates of 4.9%, 5.6%, and 3.0% all missed the fixed 20% threshold. We skipped both the ridge and Stage69 neural fits. The failed support gate applies to this estimator, not to event motion information in general. [M_GEOMETRY]

F.4 SIMPLEX-T: residual phase refinement

Expert selection constrains the output to the available expert phases. SIMPLEX-T instead anchors a residual predictor at their label-free median,

$$\hat{\phi}_t = \Pi\big(\text{median}_k\, \phi_{t,k} + 0.03\, g(h_t)\big)\,, \tag{40}$$

where h_t summarizes masked history and Π enforces the numerical output bounds. Unlike expert selection, this residual can leave the experts' convex hull. A separate cost head estimates relative losses of the original experts; it does not estimate the corrected prediction's risk. Both the recurrent and bounded-window Transformer versions mask missing history and start with zero-initialized output heads.

The primary recurrent model uses two 160-unit GRU cells. It maps expert and timing features to phase, auxiliary phase quantiles, and relative expert costs. The quantiles are not calibrated collision probabilities. History is retrospective within the current query ROI; the experiments did not verify persistent object tracking. [M_TEMPORAL_MODEL, M_PHASE_CONTRACT]

The temporal campaign completed 24 endpoints and 60,000 saved updates on nine sequences. It tested H16, reduced expert sets, and set aggregation while retaining H8 as the registered candidate. Seeds 7, 13, and 23 retrained only the heads; all shared the same expert producers. [M_SIMPLEX]

H16 reduced MiD in every head seed. Across all three, mean paired loss was 119.2 for H16 and 121.8 for H8, a difference of −2.5 with hierarchical 95% interval $[-5.2, -0.1]$. Seed 7 had already informed exploration. Seeds 13 and 23 alone gave −2.4 with interval $[-5.2, 0.04]$, which crosses zero. The analysis averages losses across shared-producer heads, rather than ensemble predictions or full-system replications. [M_SIMPLEX]

The full expert/context model outperformed every registered reduction. FULL-C0 scored 121.4 MiD, versus 127.8 for A5-only, 126.7 for C2F-only, and 124.7 for A5 plus PAIR. Unordered aggregation with age reached 122.4; removing explicit time gave 123.7. Each reduction failed the accuracy screen or its small-excess-error bound. The experts themselves can encode temporal information, so this comparison does not isolate recurrence as the source of the gain. [M_SIMPLEX]

The analytic EWMA/transport control lost to H8 on every sequence, scoring 155.4 versus 121.6 MiD; the hierarchical interval for the 33.8-point difference was $[17.1, 51.0]$. WIDE averaged 119.0 across three head seeds, but its intervals against both H8 and H16 crossed zero. It was not promoted. [M_EWMA, M_WIDE]

F.5 Capacity and teacher-feature coverage

Stage70–76 and the V11 capacity extension planned direct, pairwise-predictive, trajectory, and temporal-permutation comparisons at approximately 2M, 4M, and 8M parameters. The 4M trajectory model was primary. The proposed 30,000-query set covered 27 sequences; extra timestamps increased density, not the number of independent acquisitions.

The V11 audit found teacher features for 4,775 historical tokens and none for the other 25,225. Matching token IDs also left ROI construction, RGB preprocessing, and teacher provenance unresolved. [M_CAPACITY_PROTOCOL, M_CAPACITY_AUDIT]

Stage70 ended with both completed and unfinished endpoints. The partial record supplies no capacity ranking and no evidence that unlabeled JEPA pretraining improved SIMPLEX-T. [M_STAGE70_LEDGER]

G. Transfer, output stability, and runtime

G.1 Development, training, and transfer populations

The transfer studies answer different questions from OLD_DEV and TRAIN40. OLD_DEV had already guided development, and TRAIN40 includes fitted examples. EvTTC and Forward Collision Warning Data (FCWD) instead evaluate frozen models with dataset-specific adapters. FCWD contains three driving sequences with paired events, frames, and LiDAR-derived TTC [6]. The expanded EvTTC set adds timestamps to the same 32 sequences rather than a new holdout.

The earlier H8/TRAIN40 route predates the final refit in Section 2. A5 and C2F each received 49,932 updates, the pair head 6,840, and each of three H8 heads 2,500. Its 88,744 training predictions diagnose fit but cannot estimate unseen-scene performance. Inference used events, while training also used an RGB teacher and geometric supervision. [M_SCIENCE_AUDIT]

The initial EvTTC evaluation had 946 valid targets among 1,024 queries from 32 sequences; the expansion had 3,640 among 3,947. FCWD contributed 597 eligible targets among 630 queries from three previously exposed sequences. We left rows without ground truth unscored, and all reported methods returned finite predictions on the eligible rows. Full RGB/event Garl was omitted from FCWD because the RGB-to-event mapping had not been verified. [M_V12_EVTTC_946, M_V12_EVTTC_3640, M_V12_FCWD_597]

G.2 Input and output conditions

H8 used retrospective event history and supplied oracle ROI metadata. Garl event-only used its native pair of event windows, while Garl full added RGB. The systems differed in context and input construction. EvTTC projected RGB to events with calibrated rotation but no depth-dependent parallax correction. These transfer adapters do not reproduce the published benchmark protocol.

We corrected RGB synchronization before running full Garl. Selected images had to lie within 1 ms of each event endpoint and about 100 ms apart; the largest admitted offset was 453 μs. The runner strictly loaded the released checkpoint and upstream normalization. Because we designed the RGB extension after inspecting event-only results, this comparison remains exploratory even though the weights stayed frozen. [M_TRANSFER_REPORT]

Table 22. Frozen H8 and Garl models evaluated on EvTTC and FCWD.

Method	MAE (s)	Median AE (s)	RMSE (s)	RTE (%)
EvTTC: 946 eligible / 1,024 queries; 32 sequences				
H8 / head 7	1.321	0.766	3.027	62.35
H8 / head 13	1.353	0.731	3.206	62.92
H8 / head 23	1.409	0.738	3.392	68.87
Garl event	8.048	0.997	159.247	293.47
Garl RGB + event	1.868	0.692	7.875	76.77
EvTTC expanded: 3,640 / 3,947; same 32 sequences				
H8 / head 7	1.361	0.722	3.475	53.70
H8 / head 13	1.395	0.699	3.618	54.52
H8 / head 23	1.456	0.702	4.003	59.15
Garl event	4.560	0.958	83.896	161.12
Garl RGB + event	94.510	0.661	5561.622	5180.18
FCWD: 597 / 630; three sequences				
H8 / head 7	1.778	0.675	3.588	30.20
H8 / head 13	1.597	0.580	3.258	27.41
H8 / head 23	1.569	0.598	3.109	27.48
Garl event	2.146	0.602	5.072	26.89

Scores use queries with valid TTC targets; prediction coverage is 100% on these queries. H8 bounds TTC at ±60 s, while Garl retains its native signed outputs. The FCWD comparison uses event-only inputs. [M_V12_EVTTC_946, M_V12_EVTTC_3640, M_V12_FCWD_597]

H8 bounds TTC at ±60 seconds. Garl keeps its native scale-ratio division, which emits very large finite values when the estimated scales nearly match. We retained those predictions in the full-population metrics.

G.3 Transfer results and metric-dependent rankings

H8 had the lowest mean error on the first 946 EvTTC targets: its three heads reached 1.32–1.41 seconds MAE, versus 8.05 for Garl event-only and 1.87 for Garl full. Garl full had the best median, 0.69 seconds versus 0.73–0.77 for H8. [M_V12_EVTTC_946]

The 3,640-target expansion showed the same split. H8 achieved 1.36–1.46 seconds MAE, Garl event-only 4.56, and Garl full 94.51. Yet Garl full retained the best median at 0.66 seconds, ahead of H8's 0.70–0.72. Rare near-singular scale ratios drove Garl full's mean, so native-output and matched-clipping results remain separate. [M_V12_EVTTC_3640]

On FCWD, H8 led in MAE at 1.57–1.78 seconds versus 2.15 for Garl event-only. Garl led in mean relative TTC error at 26.9%, compared with 27.4–30.2% for H8. These rankings weight the TTC distribution differently, and only three sequences were available for clustered uncertainty. [M_V12_FCWD_597]

G.4 Head and full-route runtime

Prepared-input CPU profiling measured three rotated-order blocks of 500 batch-one calls per head. Median blockwise p95 was 9.28 ms for H1, 19.47 ms for H8, and 31.45 ms for H16. The timer covered canonical TTC emission, but not event reading, ROI construction, expert inference, detection, or tracking. [M_HEAD_COST]

The shared-GPU route was much slower than the isolated heads. Across 1,728 supplied-ROI requests from 64 TRAIN queries, H8 p95 was 9.35 and 8.48 seconds in two warm blocks; H16 reached 13.20 and 15.30 seconds. The application reopened HDF5 inputs between cold measurements, while the operating system could retain cached pages. [M_ROUTE_COST]

Table 23. Head-only and full-pipeline runtime.

Model	MiD	Head p95 (ms)	Route 1 p95 (s)	Route 2 p95 (s)
H1_SEED7	132.832	9.28	2.944	2.902
H8_SEED7	121.646	19.47	9.349	8.479
H16_SEED7	118.866	31.45	13.201	15.304
FULL_C0	121.405	21.07	8.477	9.337
A5_ONLY_C0	127.785	21.42	7.928	7.471
C2F_ONLY_C0	126.705	19.28	8.185	8.256
A5_PAIR_C0	124.669	19.83	7.086	7.502
SET_AGE_C0	122.443	10.00	7.628	8.303
SET_NOTIME_C0	123.687	9.91	7.727	7.967

Head: median of three blockwise p95 values, CPU FP32, prepared inputs, 500 measurements per block. Route: p95 on 64 TRAIN queries with supplied ROI, shared GPU, and resident producers. Accuracy uses OLD_DEV seed 7, nine sequences and three folds. [M_HEAD_COST, M_ROUTE_COST]

ROI/voxel reading and preparation consumed 85.5% of H8's measured time and 90.4% of H16's in the second warm block. We focused V12 on preparation, bounded caching, and transfer scheduling rather than head speed alone. [M_ROUTE_REPORT]

CUDA graphs reduced recorded A5 launches from 5,877 to 1,270 while preserving the tested outputs, losses, RNG state, and gradients. A later window measured 41.78 versus 32.89 updates/min, but host load and data wait also changed, so the run did not isolate graph speedup. Adaptive caching reached 26.31 updates/min against a 36.75 reference and failed its throughput test. [M_ENGINEERING]

H. From event representations to multimodal measurement

The later experiments assembled spatial measurement, temporal transport, signed-TTC readout, and multimodal correction into one causal estimator. PHASE-TTC combines task adaptation and teacher supervision with event/RGB producers, context models, and a compact fusion head. No matched factorial separated all of these components, so the final result does not establish a JEPA contribution.

H.1 V13: a parallel RGB route and a common temporal interface

V13 added an RGB route with the event branch's A5, C2F, PAIR, and temporal-context decomposition. The common interface made it possible to test event information after an image-based TTC estimate. Each modality still used its own checkpoints, normalization, training data, and temporal support.

Three disjoint acquisition-group roles separated fitting stages. P supplied 53,890 queries from 24 sequences for producers and PAIR. H supplied 18,422 queries from eight sequences for downstream normalizers and heads, while V supplied 16,432 queries from eight sequences for internal evaluation. Earlier work had already used V, so its results remain development evidence.

PHASE17 connected both modalities to their context models through named phase, timing, and support fields. Their meanings remain stable across folds, whereas independently trained encoders can rotate their latent coordinates. The explicit interface enabled grouped out-of-fold training without a separate latent-alignment model.

H.2 V14: changing the emphasis of the objective

V14 could not improve short-TTC error without harming other ranges. Its six-cell factorial combined three sample-mass schemes with two losses, holding the frozen event producers, PAIR parent, and seed-7 draws fixed. Each 160-unit, two-layer GRU head received 2,500 updates over eight masked history slots and seventeen features; timestamps represented gaps without dense imputation.

Table 24. V14 factorial study on exposed V: apparent crucial-range gains can coexist with non-crucial regressions. Source [L02].

Model	MiD	Crucial	Negative	MAE (s)
RT-M0-L0-S7	106.21	156.83	61.38	2.51
RT-M0-L1-S7	101.94	149.59	61.01	2.03
RT-M1-L0-S7	102.23	134.47	84.21	2.98
RT-M1-L1-S7	104.04	136.19	80.04	2.34
RT-M2-L0-S7	104.32	134.87	87.62	3.37
RT-M2-L1-S7	104.83	134.59	79.93	2.38

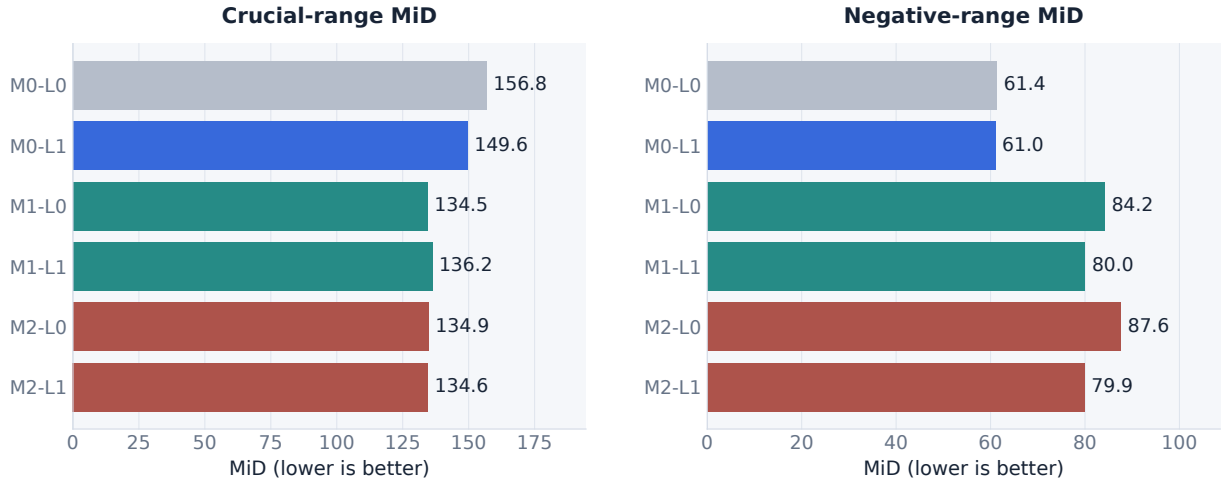


Figure 24. V14 MiD changes across crucial, small, large, and negative TTC bands.

M0 retained the native mass vector. M1 preserved mass per acquisition group while allocating TTC bands in the ratio 0.5/0.3/0.1/0.1. M2 also balanced crucial and long bands, bounding each row between 0.25 and four times its M1 weight. Uniform sampling with replacement kept the number of episodes unchanged. L0 retained the phase, quantile, and auxiliary objectives; L1 added

$$\mathcal{L}_1 = \mathcal{L}_0 + 0.25\,\mathbb{E}_{i\sim\text{batch}}\left[Nm_i\,\text{SmoothL1}\left(\frac{\widehat{T}_i - T_i}{\max(|T_i|, 0.5\,\text{s})}, 0\right)\right], \tag{41}$$

where m_i is the registered sample mass. Both losses kept the native TTC mapping, clipping, and flat-gradient regions; neither used a straight-through estimator.

Every V14 candidate failed the full gate. M1–L0 reduced crucial-band MiD from 156.8 to 134.5, while large-band error rose from 52.2 to 76.4 and negative-band error from 61.4 to 84.2. M0–L1 achieved the best aggregate and MAE in the matrix, but missed the required crucial-band improvement and other safeguards. All outputs were finite; the failures came from measured regressions. [L02]

Only 76 V queries fell in the urgent interval $(0.1, 1]$ seconds. Across the factorial, 81.6–100% received late warnings; negative TTC predictions did not count as positive warnings. Such miss rates preclude a safety claim regardless of false-alert rate.

V14 closed as `NO_CANDIDATE`, so the conditional replication and transfer stages did not run. The factorial exposed a tradeoff between urgent-range accuracy and the other signed-TTC ranges without producing an admissible improvement.

H.3 V15 and V17: testing readout and geometry explanations

Neither V15 intervention passed all gates. Sampling proportional to sample mass, corrected by w_i/p_i, preserved V14's cross-band tradeoff. Expanding the A5 residual range through an equal-slope hyperbolic

tangent left crucial-band performance almost unchanged. The modified producer received regenerated PHASE17 fields and H normalizers.

V17 tested frozen geometry from several angles. A bypass exposed geometry and transport tokens to the temporal head; other arms examined learned NLL precision, grouped ridge probes, directional differences, height calibration and supervision, vertical position, gradient clipping, and nonlinear spatial-height readout. Each had a dedicated control. Directional differences remove static levels while also reducing dimensionality, so they do not isolate appearance memorization.

The geometry bypass carried a useful but unstable downstream signal. On V, its three-seed mean improved crucial MiD from 157.1 to 148.4 and aggregate MiD from 106.5 to 102.8, while MAE worsened. The effect was stronger on H than V, and the three heads shared encoders rather than repeating representation training. [L50]

The first height-regression arm emitted predictions outside the admissible MiD domain. We kept those rows and left its aggregate missing instead of clipping them away. Temporal supervision restored coverage but not competitive TTC accuracy. Preserving vertical positions improved training height fit while worsening H TTC, and the nonlinear readout also failed. The tested frozen-feature calibration was insufficient, though the bypass result still showed usable geometric signal.

Removing gradient clipping changed little. Learned precision improved crucial-band error in every head seed, but missed the required magnitude and safeguards. The recorded gradient norms describe one loss component rather than the full-dataset objective. None of V17's nine research lines produced a candidate. [L51,L52]

H.4 V18: adaptation of the measurement stage

V18 moved adaptation into the measurement stage. Its primary comparison covered direct and correspondence-based pair readouts, trained and frozen encoders, and a native-system control. C1 tested geometric readout, D1 predictor width, B1 auxiliary correspondence supervision, and A1 adaptation depth. Some arms also updated correspondence, so equal TTC-update counts did not mean equal computation.

C1 improved bandwise geometry over an unconstrained pair-correction readout but failed as a complete TTC estimator. D1 found no benefit from greater predictor width. B1 beat a TTC-only control at the same number of TTC updates, yet its negative-band error blocked promotion over A5. A1 then compared full adaptation with partial retention of the correspondence-trained representation.

A1-FULL delivered V18's largest TTC gain and its clearest tradeoff. On H, aggregate MiD fell from 112.3 for A5-P2 to 78.9, and crucial MiD from 157.3 to 95.8; negative MiD rose from 65.5 to 87.1. It passed the global sign-error safeguard and beat last-block adaptation overall, but did not preserve accurate negative magnitude or demonstrate retention of the correspondence representation. [L10,L24,L25]

C1 also tested event information after the RGB context estimate. TRUE fusion and an equal-capacity ZERO-event control used seeds 7, 13, and 23. Each seed repeated C1 adaptation and head fitting, while all shared the historical A5 initialization and RGB context parent. The comparison isolates event input within this configuration, not variability from complete pretraining.

TRUE fusion improved on ZERO within V18. Across seeds, aggregate MiD was 87.3 ± 0.9 for TRUE and 92.5 ± 0.2 for ZERO; TRUE's crucial MiD was 126.6 ± 0.9. A paired whole-group bootstrap favored TRUE in aggregate loss. The crucial-band gain missed the preregistered 10% threshold, and the interval against the preserved V16 PHASE-ONLY family crossed zero at about $[-3.7, 0.1]$. Thus, events complemented RGB in this model without establishing superiority over V16. [L03,L31,L32,L33]

No V18 arm satisfied the complete promotion gate, and the campaign closed as `NO_CANDIDATE`. A later portability replay verified archived metrics and exports without adding training or performance evidence.

Table 25. V18 primary and adaptive experiments on development subset H (18,422 rows). Sources [L10--L14,L20--L25].

Model	MiD	Crucial	Negative	MAE (s)
A5-P2-FROZEN	112.29	157.27	65.52	5.64
PAIR-DIRECT-S7	103.87	117.98	148.17	5.00
PAIR-CORR-S7	103.51	115.56	156.00	5.50
PAIR-CORR-FROZEN-S7	106.72	123.61	146.99	5.40
NATIVE-P2-FT-S7	104.05	145.12	61.26	4.65
PAIR-CORR-GEOMETRY-S7	90.95	114.35	106.23	6.95
PAIR-CORR-WIDE-S7	115.12	124.48	206.42	6.79
B1-TTC-CONTROL-STAGE2-S7	85.98	106.71	96.63	6.11
B1-CORRESPONDENCE-STAGE2-S7	81.44	100.22	84.12	5.48
A1-LAST-BLOCK-S7	90.64	111.03	97.96	7.11
A1-FULL-S7	78.92	95.85	87.09	5.45

Table 26. V18 C1 equal-capacity TRUE/ZERO event controls on exposed V (16,432 rows), using adaptation/head seeds 7, 13, and 23. Source [L03].

Head / seed	Overall MiD	Crucial MiD	MAE (s)
FUSION-TRUE-S7	88.270	127.661	1.817
FUSION-ZERO-S7	92.638	136.989	1.842
FUSION-TRUE-S13	86.630	125.976	1.807
FUSION-ZERO-S13	92.455	136.500	1.837
FUSION-TRUE-S23	86.961	126.069	1.818
FUSION-ZERO-S23	92.315	136.314	1.836

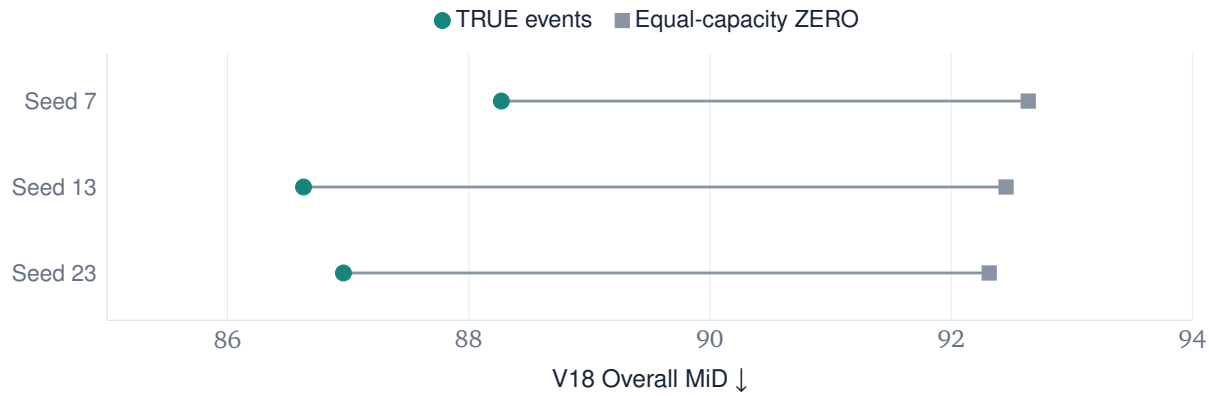


Figure 25. V18 measurement adaptations and paired TRUE/ZERO event-fusion controls.

H.5 V19: repairing negative motion changes the tradeoff

V19 tested loss weighting as the source of A1's negative-band regression. L0 and LW shared the correspondence-trained B1 parent, full adaptation, seed 7, uniform sampling, and 12,000 updates. L0 assigned crucial/small/large/negative masses of 50/30/10/10%. LW doubled negative mass to 20% and set the positive bands to 44.44/26.67/8.89%, preserving their relative proportions. The encoder did not change.

LW repaired part of the negative-motion error but worsened the primary aggregate. Relative to L0, it reduced negative MiD by 18.7%, MAE by 15.3%, and sign error by 1.55 percentage points, while aggregate and crucial MiD increased. The same pattern appeared on P training data, ruling out an H-only generalization failure. Because the replication trigger required lower aggregate error, LW did not advance. [L30]

Table 27. V19 decision recovery on exposed H. LW trades a small overall MiD increase against improved MAE and negative-range behavior. Source [L30].

Model	MiD	Crucial	Negative	MAE (s)
A5	112.29	157.27	65.52	5.64
L0	78.92	95.85	87.09	5.45
LW	79.87	100.42	70.78	4.61

Later two-expert selectors also failed the system gate. They required both fitted experts and the measured feature path, rather than forming a jointly trained sparse mixture. Grouped validation limited fitting leakage, while reused H data and fixed experts constrained the conclusion to these selectors.

I. Representation and physical-measurement follow-ups

The new experiments extend the historical controls with narrower questions: whether future-state prediction preserves useful expansion increments, whether a scalar motion quantity is observable under the fitted model, and whether a sensor or clock residual predicts a better TTC estimate. Each experiment retains its own population and reference model. An improvement in an auxiliary metric does not substitute for a TTC comparison. [N04, N05]

I.1 A closed predictive-learning pilot

The event student learns against spatial features from a frozen, distilled V-JEPA 2.1 ViT-B teacher [8]. For each of seeds 7, 13, and 21, the pilot runs static distillation, one-step prediction, and a cross-sequence shuffled-future control for 1,000 updates with batch size 16. TTC labels do not enter these student losses. The final checkpoint is fixed before evaluation. After training, a ridge probe with regularization parameter one reads temporal differences of the frozen encoder, with the predictor removed. It fits on 512 P samples from twenty-four sequences and evaluates 256 H samples from eight different sequences. Box-based localization remains privileged information. [N04]

The primary target is annotated log-height change. Its error is dimensionless: it is neither a time-normalized rate nor TTC error in seconds. The criterion requires the one-step arm to beat both static distillation and shuffled futures, with the upper endpoint of each sequence-bootstrap difference interval below zero. The bootstrap resamples eight complete sequences while keeping the three seeds together, so its interval remains conditional on those seeds and on historically exposed H sequences.

The candidate's probe MAE is 0.03564004, compared with 0.03763239 for static distillation and 0.03506189 for shuffled futures. Its candidate-minus-static difference is −0.00199235, with interval $[-0.00496210, 0.00081898]$; candidate-minus-shuffled is 0.00057816, with interval $[-0.00217005, 0.00292590]$. Neither superiority criterion passes. The random-initialization control gives 0.03719641. This failed continuation criterion is not an equivalence test or evidence that predictive learning can never help.

The auxiliary latent metrics clarify the limitation. Future teacher-state MSE improves relative to both controls, but teacher-increment MSE worsens. The corresponding candidate-minus-static changes are −0.01719163 for state and +0.03848016 for increment; candidate-minus-shuffled changes are −0.05317388 and +0.00447422. Predicting a future state can compensate for present-state errors or model variation irrelevant to expansion, but the experiment does not identify either explanation causally. Its results establish that lower state-prediction error was insufficient for the prescribed physical probe. The pilot closed without rollout, scaling, or integration into the released estimator. [N04]

Table 28. Frozen-encoder probe after the fixed-teacher pilot, averaged over seeds 7, 13, and 21. The target is annotated log-height change (dimensionless), not TTC in seconds. A ridge probe fits 512 samples from 24 sequences and evaluates 256 samples from eight other, historically exposed sequences. Neither prespecified superiority test passes.

Training arm	Probe MAE
Static distillation	0.03763239
One-step prediction	0.03564004
Shuffled future	0.03506189
Random initialization	0.03719641

I.2 Target-specific observability and model error

For a linearized observation model $\mathbf{y} = H\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and target $q = \mathbf{c}^\top\boldsymbol{\beta}$, noiseless identifiability requires $\mathbf{c}$ to lie in the row space of H. In the affine image-motion parameterization, isotropic expansion uses $\mathbf{c} = (1/2, 0, 0, 1/2, 0, 0)^\top$. The target can be estimable without every affine coefficient being identifiable. Conversely, a pseudoinverse always returning a number does not certify that the target is supported. This algebraic condition is classical and is used here as a diagnostic. [N05]

On the 149-query screen, the diagnostic marks q estimable in 63 queries. Among those estimates, 26 have an incorrect sign, ten lie outside the canonical phase domain, and eight have zero residual degrees of freedom. A small fitted residual in the last group cannot demonstrate reliability. The target itself also need not equal local isotropic image expansion when rotation, parallax, or changing visible surfaces violate the measurement model.

The gain-prediction experiment compares physical descriptors against simpler support descriptors with leave-one-acquisition-out fitting. Physical-feature MSE is 0.919267453, versus 0.492043720 for support features; their difference is 0.427223732 with interval [0.080550690, 1.010417232]. This is MSE for predicting correction gain, not TTC MAE. The result rejects the proposed descriptor gate. A historical robust readout returns only eight usable estimates, four with the wrong sign; after preserving F0 fallback, its MAE is 0.779543461 s versus 0.766024827 s for the matched R0 baseline. The comparison records unavailable measurements and the original missing output; the error of the eight supported measurements is not presented as complete-population performance. [N05]

I.3 Sensor ambiguity and the missing reference

The sensor-identifiability bank constructs 24 pairs of matching event streams, with pixel, polarity, and timestamp agreement to a tolerance of one picosecond. Twelve pairs have different inverse TTC. Under the bank's log-affine texture model, changing motion and inversely scaling the texture gradient can preserve the increments that generate events; the distinct initial references are absent from the event stream. Providing a known reference helps distinguish the restricted conditions. This construction diagnoses an ambiguity in the tested model, not a general impossibility theorem for event-based estimation. [N05]

The real-data proxy uses 72 road queries from twenty-four acquisitions. Slider is excluded from the photometric fit because its alignment remains provisional. On common support, observed RGB change gives sensor-residual MSE 96.657290 versus 145.648563 for the physical proxy. Its improvement over an activity control is less decisive: the corresponding descriptive difference interval includes zero. Sensor-residual reductions cannot be interpreted as TTC improvements, and the inventory does not establish calibrated contrast thresholds, gain, initial reference, or refractory response.

The direct rate readout tests that implication. The affine photometric model permits translation, deformation, and global illumination variation and returns 72 valid rate estimates. With the original fallback retained, its macro TTC MAE is 9.168523738 s, compared with 0.766024827 s for R0. The spatially permuted control is worse still, at 45.629594312 s. Beating that control is insufficient when the matched estimator deteriorates so strongly. The experiment opens no producer or JEPA training and supplies no calibrated confidence signal for fusion. [N05]

I.4 Surfaces and dynamic clocks

For positive depth Z, signed inverse TTC is $q = -\dot{Z}/Z$. Differentiation gives

$$\dot{q} = q^2 - \ddot{Z}/Z. \tag{42}$$

If axial relative velocity is constant in this coordinate, its finite-time evolution is

$$q(t+\delta) = \frac{q(t)}{1-\delta q(t)}, \qquad T(t+\delta) = T(t) - \delta. \tag{43}$$

The transport cannot cross the contact singularity. Acceleration, camera rotation, or a change in the visible surface can invalidate the assumptions even when an object has constant velocity in world coordinates. These relations follow from kinematics; the experiment tests their utility in the estimator rather than claiming a new physical law. [N05]

A three-frame surface route supplies 133 of 149 measurements and falls back on the other sixteen. Its system MAE is 3.103982972 s versus R0's 0.766024827 s, with a paired difference interval [0.934231309, 4.198244944] s. The crop audit finds no changing crop within the 146 auditable histories; 53 contain a constant anisotropic resize. This rules out one proposed changing-zoom explanation for that cohort, not errors from occlusion, backgrounds, or parallax. [N05]

The constant-axial-velocity transport law above is exact only under its model assumptions. Reapplying its composition identity inside an exact implementation adds no new measurement. The follow-up compares fixed persistence, smoothing, transport, and observer rules against their matched event context after canonical PAIR. Context macro MAE is 1.941890262 s; EWMA gives 1.952822072 s, the observer 1.981365164 s, and persistence/transport 3.104231759 s. These comparisons do not establish a benefit. Their baseline is the event context, which must remain distinct from the multimodal R0 baseline used above. [N05]

Together, these experiments distinguish implementation correctness, model identifiability, and useful downstream prediction. No rejected gate becomes a successful learned physical model because its equations are interpretable. Further training would require a new, bounded hypothesis supported by these failures and evaluated with appropriate controls; it is not a result already achieved by this revision.

J. Detection, geometric consistency, and the localization boundary

The TTC evaluations in this paper condition on supplied object regions. A separate eAP study develops a portable RGB MonoDGP-R50 variant, a proposal ranker, and an interface from tracked detections to PHASE-TTC. We report this work to examine the localization dependency. Its detection scores, geometric diagnostics, and integration checks do not establish improved end-to-end TTC accuracy. The detector is also distinct from the event representation and predictive-learning experiments. [N06]

J.1 Fixed proposals and an aligned evaluator

On 23,958 development-exposed frames, the aligned eight-class evaluator gives mAP 0.05701726 for the retained Dfull detector and 0.06204229 after a TRAIN-fitted shared logistic ranker. Both evaluations use the same 1,197,900 proposals and 190,469 raw targets. Ranking changes the scores while preserving proposal geometry, classes, and velocities. Bicycle and other-vehicle AP remain zero; the aggregate gain does not imply uniform class improvement. The study uses one detector seed and a selection population reused during development. [N06]

Evaluation reconciliation matters here. Earlier AP101 reports used a restricted target population and a precision envelope. The aligned calculation uses greedy class matching in ego-XY coordinates, strict distance thresholds of 0.5, 1, 2, and 4 m, recall interpolation without the envelope, and the declared minimum-recall and minimum-precision rules. The two calculations are different estimands. We retain their historical records rather than comparing their numeric magnitudes as model improvements. The

ranked submission obtains official AP 0.05129508 on 35,940 test frames; no paired baseline submission exists, so the local ranking gain cannot be transferred to the official test. [N06]

J.2 Ideal cuboid inversion and predicted geometry

The geometric audit asks whether a height-only depth prior has a bias even with ideal inputs. Let $\boldsymbol{\delta}_i$ be a rotated cuboid-corner offset, $\mathbf{u}$ the projected center, and K the camera intrinsic matrix. Define

$$\mathbf{a}_i = K_{1:2,:}\boldsymbol{\delta}_i - \mathbf{u}\delta_{iz}. \tag{44}$$

For a positive center-to-edge distance t_e, cuboid containment gives

$$z_e = \max_i \left(\frac{s_e(\mathbf{a}_i)}{t_e} - \delta_{iz} \right), \tag{45}$$

where s_e selects the signed horizontal or vertical numerator for that edge. The implementation takes a median over non-clipped edges; a clipped image boundary is not treated as an equality. This is a pinhole consistency calculation, not a new physical law. [N06]

With target-derived dimensions, yaw, and projected edges, the admitted set contains 25,424 objects from 4,096 TRAIN frames. On the same admitted objects, the height prior has depth MAE 4.923050 m, whereas cuboid-edge inversion reaches 2.1085×10^{-6} m. The near-zero error verifies consistency with the ideal target generator. It does not measure sensor calibration or learned depth accuracy. [N06]

A second experiment substitutes predicted geometry on a *different* 4,096-frame TRAIN panel. It preserves the proposal pool, class, score, size, and yaw, but replaces the learned depth output and removes its learned residual. Baseline mAP 0.21064779 falls to 0.16478485 with cuboid-edge depth and to 0.08308035 with the height-only prior. Invalid geometric cases use the declared fallback. Thus the better ideal inversion does not improve this direct substitution. No model is promoted, and the experiment does not rule out retraining a compatible residual or estimator. [N06]

J.3 Negative screens and integration scope

The later range, class-budget, query-ranking, KITTI-transfer, and MonoCoP campaigns do not justify replacing the retained detector. Short, single-seed screens and differing upstream training exposure restrict their interpretation; they are not evidence that every related architecture or transfer recipe must fail. The unpromoted candidates remain in the experimental record. [N06]

An eight-frame integration replay produces five available signed TTC outputs and one unconfirmed track in the last frame. This demonstrates execution of the detector–tracker–TTC interface. Frozen TRAIN40 producers had already seen these sequences, and the replay supplies no independent end-to-end TTC accuracy estimate. The admission protocol also distinguishes target time from availability: the target is anchored at RGB exposure start, while inference can consume observations through exposure end. The resulting estimate has a measured availability lag and must not be described as causal at the earlier target instant. These boundaries remain necessary when replacing supplied boxes with predictions. [N06]

K. Experimental chronology

Development began in June 2026 with the E-JEPA-TTC hypothesis: predictive latent learning might improve event-based TTC estimation. The research then moved through geometry and cross-time transport to temporal context and RGB–event fusion. Several evaluation corrections changed which results could support a claim.

K.1 June–July: predictive learning and evaluation corrections

The first experiments established indexed EvTTC access, voxel representations, a TinyCNN, and geometric/event-rate baselines. JEPA pretraining with latent-variance regularization led to multihorizon prediction, frozen-encoder probes, and low-label comparisons. Token transformers, tubelets, deep token supervision, navigation conditioning, and rollout probes tested other forms of temporal information.

Centered derivatives gave the geometric baseline access to future observations. We replaced them with past-only derivatives and reported this annotation-assisted baseline on its supported queries, separate from models with different inputs or coverage.

On 13 July, we reclassified the inspected CPLA-high results as reused-test diagnostics (`26480260`). Their three fine-tuning seeds shared one SSL initialization, so the spread reflects downstream optimization conditional on that parent. We removed the runs from final-test claims and added checkpoint, cache-version, nested-subset, and matched-scratch checks. No repartition of the inspected sequences could recreate a blind holdout.

An eAP smoke-validation defect wrote success markers without running the matrix (`1f9e2f99`). We replaced the faulty path with the matrix invocation and added a regression test; the old markers were discarded as evidence.

K.2 August: geometry, transport and fold-local parents

August introduced physics-constrained FlowMimic, dense LevelDynamics, object-centric representations, and matched Garl comparisons. The A-series varied foreground geometry and decoder capacity, then added SAM supervision and DINOv3 relational distillation. A4 used an RGB teacher; A5 learned local cross-time transport on disjoint training rows. These stages no longer tested the original label-free JEPA objective alone.

On 13 August, we found that several grouped runs inherited an A4 parent trained on their held-out outer groups (`de3349fd`). Those summaries remain parent-exposed diagnostics. The corrected study trained A4 inside each outer fold, keeping the held-out group out of both representation learning and downstream fitting.

V6 and V7 produced no promoted model. A post-hoc A5/C2F routing result from V7 became a prospective V8 hypothesis. V8 then tested temporal mechanisms and nested routing after correcting ROI quantization, timestamp bins, and bounded HDF5 reads. Unsupported predictions stayed in the evaluation as named failures.

K.3 September–October: context and multimodal measurement

X0/X1 and Stages 61–76 tested phase dynamics, raw timing, routing, risk, and context under fixed optimization budgets. Nested producers excluded held-out groups from training, and incomplete runs were not promoted to completed evaluations.

SIMPLEX-T completed 24 head fits and 60,000 updates. Its H8/H16 comparisons retrained heads while sharing upstream producers. Efficient-context work then addressed TRAIN40 inputs and execution cost; the RGB port added native sensor clocks and transfer studies. The record also retains failed geometry, clipping, and nonlinear-readout controls from the range and representation studies.

The multimodal V16 phase-correction family supplied the final-refit recipe. V18 tested joint measurement, correspondence, geometric readout, predictor capacity, and encoder adaptation. Some arms improved individual bands, but none passed every promotion criterion; V18 closed as `NO_CANDIDATE`. V19 then tested negative-loss weighting and fixed expert selection.

K.4 October: final refit and causal runtime

The final TRAIN40 refit used group-separated auxiliary producers for out-of-fold inputs and full-TRAIN40 producers for deployment. It followed the V16 recipe but trained a new model lineage. The submission preserved query-token identity and complete prediction coverage, and its official score placed PHASE-TTC first on the public GarlTTC Test leaderboard on 28 October 2026 (Section 4).

Runtime experiments covered persistent histories, exact normalization, graph execution, crop reuse, and multiobject replay. Timing reports separated startup from measured arrivals and counted valid TTC outputs separately from failed acknowledgements. The 5 ms ingestion experiment kept 20 ms output queries and established numerical parity for that setting; it did not measure throughput.

K.5 29 October–2 November: adaptation and controlled follow-ups

The subsequent work recovered target correspondence for all 434 EvTTC queries, fitted a compact quality adapter over frozen TRAIN40 components, and tested canonical elapsed-time handling. The delivered adapter and the later rejected canonical refits remain distinct. Physical-measurement and fixed-teacher predictive-learning experiments produced negative controls; a separate eAP detection branch investigated localization, evaluation alignment and geometric priors. Sections 7 and 8, together with Appendices I and J, describe these additions with their own evidence and exposure histories.

L. Reproducibility

The release contains the manuscript source, analysis scripts, derived-result ledgers, and selected query-level predictions needed to regenerate the reported tables and figures. Repeating model training or sensor inference also requires the datasets, checkpoints, and experiment configurations listed in the evidence registry.

L.1 Manuscript and result regeneration

Install Python, NumPy, Matplotlib, and a LaTeX distribution. From the manuscript source root, run:

```
python scripts/build_expanded.py
```

The script verifies the distributed files, checks the preserved historical figures and tables, regenerates the new tables, and runs pdfLaTeX, BibTeX, and the required cross-reference passes. Once the dependencies are installed, the build uses bundled ledgers and requires neither network access nor raw sensor data.

Each ledger records the evaluation population, source key, and measured value. The verification scripts match source identities and recompute metrics from the distributed query-level predictions. Hardware-specific inference parity requires the corresponding checkpoints, inputs, and runtime environment.

L.2 Numerical verification

The verification suite checks 13 data-file hashes, eight transfer summaries, twelve official-scorer fields, and weighted Overall MiD. The competition verifier recomputes coverage, signed ranges, disagreements, and timing summaries from all 1,736 model–query rows. The dated leaderboard snapshot supplies the official ranking.

Table 29. Additive decomposition of the official Overall MiD difference. Each band contributes $w_b(\mathrm{MiD}_b^{\mathrm{Garl}} - \mathrm{MiD}_b^{\mathrm{PHASE}})$. The 0.0010-point reconciliation residual comes from Garl-TTC's published two-decimal band scores and appears separately below. [B01, B02]

Band	Weight	MiD difference	Weighted contribution
Crucial	0.5	9.6110	4.8055
Small	0.3	9.3232	2.7970
Large	0.1	14.8921	1.4892
Negative	0.1	1.3586	0.1359
Band subtotal	1.0	–	9.227548419
Published-score rounding residual	–	–	0.0010
Official Overall difference	–	–	9.228548419

The transfer analysis recomputes mean, median, p95, and relative error for four Extended-new methods from 468 eligible rows. It then constructs error-exceedance curves and paired complete-sequence bootstrap intervals. The fixed seed, quantile convention, bootstrap differences, and source-file digest reside under `revision/`. This post-hoc analysis uses fixed predictions and does not repeat model training or inference.

L.3 Paced-run configuration

The L40 receipts identify the final model and component digests used for the PHASE-TTC competition inference, while recording a separate replay input and timing protocol. Synthetic startup priming used no raw observations or targets and made no optimizer updates. The update clock includes unexpected shape preparation. The receipts omit the CPU model and GPU-driver version. [L40, R02]

Table 30. Paced-replay hardware, configuration, and initialization timings. Startup stages were timed separately and excluded from the arrival clock; unexpected shapes were processed within the measured update. [L40, R02]

Item	Recorded value
GPU	NVIDIA GeForce RTX 5070 Ti Laptop GPU
Software	Python 3.11.15; PyTorch 2.11.0+cu128; CUDA build 12.8
Precision	FP32; TF32 disabled for matrix multiplication and cuDNN
CPU / preparation	2 PyTorch CPU threads; 8 RGB workers; queue capacity 2
Parity tolerance	Absolute and relative tolerance both 10^{-5}; all 250 outputs checked
Outside the measured arrival clock	
RAW preload	2.769 s
Model and service initialization	0.516 s
Synthetic shape priming	12.095 s; 7 layouts; synthetic tensors only, no RAW/future sensor data
History initialization	0.132 s for 600 ms of preceding sensor history
Source warmup	0.051 s; 13 already-available RGB frames
Unrecorded hardware fields	CPU model and GPU driver version

L.4 Code verification

Selected tests cover event representations, target updates, prefix causality, label rejection, source tracking, fusion fallbacks, and paced-runtime accounting. The environment used Python 3.11.15, PyTorch 2.11.0+cu128, and pytest 9.1.1. Across two batches, 178 tests passed and four skipped, with no failures. Four additional tests ran against the SHA-pinned official evaluator and passed with no skips or failures. They cover perfect predictions across four bands, crucial-band weighting, absent-band handling, and invalid-output retention. [M21, M24]

For the competition adapter and verifier, 41 focused tests passed; the CPU-oriented suite skipped one conditional CUDA case. A separate CUDA test confirmed strict loading and native-model forward parity for all three public Garl checkpoints. Post-run validation covered 1,302 Garl terminal records, 434 PHASE predictions, input identities, causal timing, signed FP32 height-ratio conversion, and complete or partial result packages. [T01]

L.5 Evidence integrity and source revision

SHA-256 checks identify 95 original result snapshots. Appendix M lists their digest prefixes and numerical projections. Entries without a projection preserve source identity but cannot support metric recomputation from the distributed data alone. For the official benchmark, the scorer JSON agrees with the dated leaderboard after display rounding. Because the scorer output contains no submitted-prediction hash, it does not link the reported score to a specific prediction file at the byte level.

The scientific source snapshot is `0219a84f0fc3b07c171295a8a65176a7e0f631c1` in Kripta-Studios/phase-ttc.

The competition adapter used source revision `1d8de7d95b2ff8ccba884dbd40141c48745a1e95`. The public competition projection contains the numerical predictions, measurement boundaries, checkpoint identities, and source digests required to regenerate coverage, disagreement, and timing summaries without distributing raw sensor files.

`ASSETS.md` documents the supplied visual material and its attribution. The visual-source manifest records the sensor illustration's checksum and transformation; original dataset and material rights remain with their owners.

L.6 Expanded edition

The new result tables regenerate from the copied JSON ledgers under `updates/`. Their manifest verifies source hashes before rendering. The expanded build retains all original source sections and figures, updates dated conclusions where later evidence supersedes them, and appends the new evidence registry (Appendix N). Rebuilding the article does not rerun the model fits reported in the later experiments.

M. Evidence index

Source keys map each reported measurement to its result artifact. Numerical ledgers drive table and figure regeneration, while `data/public_evidence_registry.json` records full digests and field selectors. The index separates original-artifact hashes from distributed-file hashes. *Digest only* entries preserve source identity without distributing the fields required for metric recomputation. B02 denotes the official scorer JSON.

M.1 Result-source identities

Source key	Distributed ledger	Original SHA-256
E01	`data/early_results.json`	`55b5494b40f97f1b`
E02	`data/early_results.json`	`27ee1c9a21cca0ec`
E03	`data/early_results.json`	`aa46f6d199bcf5a4`
E04	`data/early_results.json`	`02c215aa4678f307`
E05	*Digest only*	`6fe8c8761d0a14fb`
E06	*Digest only*	`d38021db5b3b5b62`
E07	`data/early_results.json`	`d53b6ce406fa7b80`
E08	`data/early_results.json`	`d69f8b9082c883da`
E09	`data/early_results.json`	`9bf197c3d64635a3`

Continued on next page

Source key	Distributed ledger	Original SHA-256
E10	data/early_results.json	d47face1e36d1d2c
E11	data/early_results.json	6db419fd7c7fe536
E12	data/early_results.json	f33dc9abbad8fd08
E13	*Digest only*	ac26c1b2ad87c6c9
E14	*Digest only*	f947b32f05e655a1
E15	*Digest only*	7c60bc0fa9f1c5f0
E16	data/early_results.json	be5fdbc0b2e9773d
E17	data/early_results.json	d3fb5a5feb848ac7
E18	data/early_results.json	a3c4765f5278a3f4
E19	data/early_results.json	f1471599d502f2de
E20	data/early_results.json	3e02fd81751e3f86
E21	data/early_results.json	5c7b1eb24c07bd08
E22	*Digest only*	2e2eef6698ef28a4
E23	data/early_results.json	bc540cc21ddb317f
E24	*Digest only*	dab563a02450abe6
E25	data/early_results.json	06619299d1ceb785
E26	data/early_results.json	06dc45cf3ea7a3e0
M_SIMPLEX	data/middle_results.json	93f83e71f86dfcad
M_HEAD_COST	data/middle_results.json	8d96bbc9bd66d8fb
M_ROUTE_COST	data/middle_results.json	d1d834ab36d14fbd
M_WIDE	*Digest only*	c15e32a8dea0eece
M_EWMA	*Digest only*	47cca03ccda6f054
M_STAGE61	data/middle_results.json	20f1990bb70e5de7
M_STAGE62	data/middle_results.json	c4a0e1ae20d79094
M_GEOMETRY	*Digest only*	fa917902b5696e4b
M_STAGE67_DECISION	*Digest only*	ab1d3095955477f4
M_STAGE69_DECISION	*Digest only*	dfd1956775492584
M_ENGINEERING	*Digest only*	e756ba5c7a433efd
M_SCIENCE_AUDIT	*Digest only*	06073f4182de1a41
M_STAGE63_REPAIR	*Digest only*	f96b7fe795637cc5
M_CAPACITY_PROTOCOL	*Digest only*	1be3a4c9f9b64a08
M_CAPACITY_AUDIT	*Digest only*	07ef97a56b4159fa
M_STAGE70_LEDGER	*Digest only*	62597e162571def0
M_TRANSFER_REPORT	*Digest only*	d92c9074516354bd
M_ROUTE_REPORT	*Digest only*	1f29a459c91f6b68
M_TEMPORAL_MODEL	*Digest only*	4c26ad5ecf9e8354
M_PHASE_CONTRACT	*Digest only*	7ee447190ef7b218
M_STAGE64	*Digest only*	68409fb88499899f
M_X0	*Digest only*	61d8a0e40b6d87ad
M_X05	*Digest only*	c4aa55dd6a944d01
M_X05_GATE	*Digest only*	61eae7d38fb6ea5a
M_STAGE65	*Digest only*	4f88f676eecc7e90
M_STAGE67_DIAGNOSTICS	*Digest only*	885ac3604160cf27
M_S67_SIMPLEX17	*Digest only*	969cf5db29b2a251
M_S67_UNCONSTRAINED17	*Digest only*	1419522be60fc952
M_S67_SIMPLEX8GATE	*Digest only*	e094512164179869
M_S67_SIMPLEX17_PAIRPERM	*Digest only*	451821bb98c1d0c2
M_S65_RISK17	*Digest only*	b964e12180efbcbc
M_S65_RISK8	*Digest only*	75d98c9e0be9ddc3
M_V12_EVTTC_946	data/middle_results.json	03c57b4503ca95f4
M_V12_EVTTC_946_PAIRED	*Digest only*	5fee507ffe8a4aca
M_V12_EVTTC_946_ROWS	*Digest only*	d284e52a0c25a897
M_V12_EVTTC_3640	data/middle_results.json	792a5042dfa8be8b
M_V12_EVTTC_3640_PAIRED	*Digest only*	2619cc3d11be7320
M_V12_EVTTC_3640_ROWS	*Digest only*	20f58c1200b3834d
M_V12_FCWD_597	data/middle_results.json	a1c0c4a741b08a36
M_V12_FCWD_597_PAIRED	*Digest only*	2581bd042f29a466
M_V12_FCWD_597_ROWS	*Digest only*	68f4cd5610f50c8b
L01	data/late_results.json	9136babd4a07411e
L02	data/late_results.json	844e0771bab79717
L03	data/late_results.json	4af996900a99c585

Continued on next page

Source key	Distributed ledger	Original SHA-256
L10	data/late_results.json	f5bf12b2873a2ee4
L11	data/late_results.json	22e82a376b18c9f0
L12	data/late_results.json	fafd9cefbeddcd49
L13	data/late_results.json	86ce201ca25bcc12
L14	data/late_results.json	0588956e96dab44f
L20	data/late_results.json	6d15d1acfad38592
L21	data/late_results.json	2b33ba213186f3c5
L22	data/late_results.json	b1ad71b126b20e87
L23	data/late_results.json	0698e433731da0d3
L24	data/late_results.json	5f3789bd7efebce3
L25	data/late_results.json	774c751e4aaa91d7
L30	data/late_results.json	1073ab21e307ae26
L31	data/late_results.json	54ebca7a19d5a9eb
L32	data/late_results.json	d4fbc7494769a5d1
L33	data/late_results.json	025a058d846a7ceb
L40	data/late_results.json	c27d5d980f2bb1c1
L41	*Digest only*	d8df9ca6444c6dc0
L42	*Digest only*	06ab49e521db7255
L43	*Digest only*	aed0bdd40c58d827
L44	*Digest only*	e447c5147e4e1c25
L45	*Digest only*	4f643a15cbf3d963
L46	data/transfer_queries.csv	1fbff40f000c879b
L50	*Digest only*	3069072fb5180272
L51	*Digest only*	ad4a2a0660bd2da9
L52	*Digest only*	46e5516c40511837
B01	data/official_benchmark.json	0c7114c5acfded5e
B02	data/official_scoring_scores.json	e06e99ba1e5d6c4e
T01	data/evttc_competition_summary.json	96c849f29383c35f
T02	data/evttc_competition_summary.json	ceb5c6b5f3ef1b93

M.2 Implementation and verification references

Implementation keys map source revisions, method descriptions, and selected verification results to their distributed manifests.

Source key	Distributed manifest	Ledger selector
M01	research/architecture_sources.json	id=M01
M02	research/architecture_sources.json	id=M02
M03	research/architecture_sources.json	id=M03
M04	research/architecture_sources.json	id=M04
M05	research/architecture_sources.json	id=M05
M06	research/architecture_sources.json	id=M06
M07	research/architecture_sources.json	id=M07
M08	research/architecture_sources.json	id=M08
M09	research/architecture_sources.json	id=M09
M10	research/architecture_sources.json	id=M10
M11	research/architecture_sources.json	id=M11
M12	research/architecture_sources.json	id=M12
M13	research/architecture_sources.json	id=M13
M14	research/architecture_sources.json	id=M14
M15	research/architecture_sources.json	id=M15
M16	research/architecture_sources.json	id=M16
M17	research/architecture_sources.json	id=M17
M18	research/architecture_sources.json	id=M18
M19	research/architecture_sources.json	id=M19
M20	research/architecture_sources.json	id=M20
M21	research/architecture_sources.json	id=M21
M22	research/architecture_sources.json	id=M22
M23	research/architecture_sources.json	id=M23

Continued on next page

Source key	Distributed manifest	Ledger selector
M24	research/architecture_sources.json	id=M24

M.3 Distributed data integrity

Digest prefixes identify the numerical files and manifests used for regeneration; the build verifies each full hash before analysis.

File ID	Distributed file	SHA-256 prefix
D01	data/early_results.json	bbc3e4b30c44e6af
D02	data/early_tables.json	e7c769877f14ecf2
D03	data/git_summary.json	40b8eac8566ed3de
D04	data/late_results.json	e7093f1b75a4e96b
D05	data/middle_results.json	236eb5dab7db1932
D06	data/official_benchmark.json	f7bc1a8bd786f814
D07	data/official_scorer_verification.json	98ccc01371304d3a
D08	data/official_scoring_scores.json	e06e99ba1e5d6c4e
D09	data/transfer_queries.csv	2fd4c33716a6c48e
D10	data/visual_source_manifest.json	7cd53750ff473479
D11	research/architecture_sources.json	a544f7aab9d12e34
D12	data/evttc_competition_summary.json	9ad728d4168c0b41
D13	data/evttc_competition_predictions.csv	5e58c4155d90cf5e

M.4 Post-hoc paired transfer analysis

R01 maps revision/derived_transfer_analysis.json and the complete paired-bootstrap differences to scripts/prepare_revision.py. The analysis uses the L46 transfer-query predictions and retains their source digest. Appendix C gives the sequence-resampling procedure and numerical cases.

M.5 Deployment metadata and output map

R02 maps revision/deployment_contract.json to the native phase-output transformation and L40 paced-run configuration. It records original source digests, the executed code revision, model identities, hardware, precision, and initialization boundaries. R02 contains metadata from the recorded run rather than a new performance measurement. The build verifies its full digest before regenerating Table 30.

N. Expanded evidence: 29 October–2 November 2026

The preceding index preserves the original article's source identifiers and historical measurements. New identifiers N01–N07 refer to the later work; they do not overwrite E01–E08 or any original experiment key. The source package includes the new ledgers under `updates/evidence/`, their full SHA-256 digests, and the review's branch and document-coverage manifests.

Source	Scope	Distributed source and interpretation
N01	Final quality adapter; corrected 434-query comparison	`updates/evidence/adapt.json`; corrected local evaluator with historical exposure and unresolved Slider-750 equivalence. Model description in the copied 1 November delivery report.
N02	Canonical PAIR, complete development population	`updates/evidence/canonical.json`; recording-macro errors, explicit availability, and paired recording-bootstrap intervals.
N03	Rejected paired refit; native output sensitivity	`updates/evidence/refit_result.json` and `updates/evidence/native_result.json`; the refit fails selection and the native comparison measures output changes without Test12 targets.
N04	Fixed-teacher predictive-learning pilot	`updates/evidence/jepa.json`; physical-increment probe and static/shuffled controls, with neither superiority criterion passed.
N05	Physical measurement diagnostics	`updates/review/council_physics_claims.json` and the copied `updates/evidence/support/` records; source paths, exact field selectors, and original hashes accompany the interpretation.
N06	eAP detection follow-up	`updates/evidence/detection/source_manifest.json`; six branch-pinned JSON ledgers, including distinct ideal and predicted geometry panels.
N07	Runtime and local execution boundaries	`updates/evidence/runtime/summary.json`; the original successful-output paced population. The later raw-path comparison has its own copied receipts and a different workload and timing boundary.

N.1 Review coverage and scientific selection

The review fetched remote refs without modifying experiment branches and inventoried 241 commits across 72 refs in the lookback beginning 28 October. The 411 directly touched Markdown paths comprise 406 semantic or source-verified documents and five generated historical inventories checked structurally. The latter include older records outside the requested period. The manifest distinguishes these categories and preserves divergent branch versions.

The council reviewed transfer and fusion, runtime, eAP detection, physical measurement, and typography in separate lanes. A subsequent scientific review checked the new manuscript's equations, metric units, evaluation populations, causality, and unsupported-claim boundaries. These reviews guide inclusion; they are not independent experimental replications. No new training or official benchmark submission was performed to create this expanded article.