SPAR: MONOCULAR SPARSE 3D POINT TRACKING WITH EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous 3D point tracking is fundamental to dynamic scene understanding. Monocular event streams provide sparse and intermittent image-plane observations, making it challenging to maintain query identity and recover metric motion without direct longitudinal measurements. We propose SPAR, a sparse 3D point tracker that uses only monocular events after image-based initialization. Its query-conditioned interval–geometric matching combines observed-event support with sampled pre-trigger response history through Hit–Survival and refines correspondences with reliability-weighted local motion. An interval-evidence marginal objective further incorporates this candidate-conditioned history into correspondence learning. For 3D inference, we introduce a Query-conditioned Anchor-Relative Trajectory head (QART) that separates one-time metric anchoring from event-driven relative motion and carries correspondence confidence, support reliability, and local geometry into temporal 3D inference. We construct EventTraj3D from BlinkVision, comprising 13.630 billion asynchronous events and 6,494 annotated query trajectories. Experiments show lower 2D and 3D errors on EventTraj3D and the lowest RMSE among compared methods on E-3DTrack, while initialization perturbations show reduced sensitivity to depth-prior errors.

1 INTRODUCTION

Continuous 3D point tracking maintains correspondence to the same physical point while recovering its spatial motion. Event cameras record brightness changes with high temporal resolution and high dynamic range, providing fine-grained observations of fast motion (Gallego et al., 2022). A monocular configuration eliminates the need for a stereo baseline and cross-view matching but removes direct disparity-based depth constraints. Tracking must therefore recover both image correspondence and motion along the viewing direction from a single view.

Existing event point trackers estimate 2D trajectories through local matching and temporal association (Messikommer et al., 2023; Hamann et al., 2025; Han et al., 2025). Video-based 3D trackers use depth-supported representations or joint geometry and camera-motion estimation (Xiao et al., 2024; Zhang et al., 2025; Xiao et al., 2025), while existing event-based 3D trackers rely on stereo geometry or per-time-step depth inputs (Li et al., 2024a; Shen et al., 2026). Our method instead tracks sparse metric 3D points from monocular events after image-based initialization, without stereo or subsequent depth inputs. A pretrained depth prior (Hu et al., 2024) is refined during initialization. Subsequent event-driven states jointly represent motion and compensation for remaining anchor bias.

We identify a candidate-evaluation gap in event point tracking. DeepEvT (Messikommer et al., 2023), ETAP (Hamann et al., 2025), and MATE (Han et al., 2025) strengthen correspondence through feature correlation, motion guidance, and temporal aggregation, but do not explicitly compare the pre-trigger brightness histories implied by competing correspondences. A wrong neighboring edge can therefore receive strong feature and event support while implying an incompatible pre-trigger history. Observed-event similarity alone can leave this ambiguity unresolved. In monocular 3D tracking, an identity error also corrupts the motion evidence used for spatial estimation.

To address this gap, we associate inter-trigger history with each candidate motion path. Under an ideal model with accurate reset, no missing events, and symmetric thresholds, a pixel's brightness change remains between the thresholds until the next trigger (Gallego et al., 2022). Plausible endpoint

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

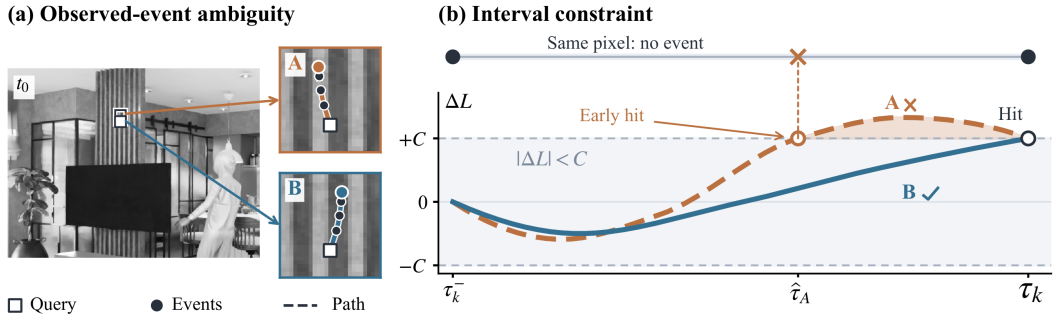

Figure 1: **Candidate ambiguity and pre-trigger evidence.** (a) Similar event support can favor neighboring edges A and B. (b) Both reach the endpoint response, but A predicts an unobserved earlier threshold crossing under the ideal event model. Our method compares sampled pre-trigger response ordering to distinguish such histories.

support need not imply a plausible preceding history (Figure 1(b)). Our key step is to make this history a query-conditioned correspondence criterion, evaluated separately for competing paths.

Hit–Survival implements this query-conditioned criterion through a soft, normalized response-ordering score. Validity checks, pixel-balanced averaging, and reliability-gated fusion limit unreliable interval evidence. A marginal objective incorporates the same evidence into correspondence learning. Local geometry provides a complementary motion constraint, but spatial proximity does not guarantee shared motion near object boundaries. We therefore weight support contributions by motion compatibility rather than treating all neighboring points as equally reliable.

Reliable correspondence still leaves longitudinal motion ambiguous. Local scale changes provide relative-depth cues only under assumptions on motion and geometry (Yang and Ramanan, 2020; Wan et al., 2025). We therefore introduce a Query-conditioned Anchor-Relative Trajectory head (QART) that retains matching evidence, support reliability, and local geometry when estimating initial-depth-normalized, anchor-relative 3D displacement. The refined initial depth provides the metric reference for subsequent event-driven spatial updates.

In summary, our contributions are as follows:

- We introduce Hit–Survival to evaluate query correspondence through pre-trigger history. To our knowledge, this is the first use of candidate-conditioned pre-trigger interval evidence for correspondence scoring and learning in monocular sparse 3D point tracking.
- We couple interval–geometric matching with the trajectory head to estimate initial-depth-normalized, anchor-relative trajectories from correspondence and local-motion evidence. This enables sparse metric 3D tracking from monocular events after image-based initialization, without stereo or per-time-step depth inputs.
- We construct EventTraj3D from BlinkVision, providing aligned 2D–3D query trajectories for monocular event-based tracking across diverse scenes, motions, depths, and event densities.
- Experiments demonstrate that our method achieves the lowest 2D and 3D tracking errors on EventTraj3D and the lowest RMSE on E-3DTrack among the compared methods.

2 Related Work

2.1 Point Tracking in Videos and Event Streams

Video point trackers use appearance and temporal context to maintain correspondence. PIPs (Harley et al., 2022) aggregates temporal correlations, TAPIR (Doersch et al., 2023) combines per-frame matching with temporal refinement, and CoTracker (Karaev et al., 2024) models multiple points jointly. TAP-Vid (Doersch et al., 2022) provides a common benchmark for video point tracking. Early event feature trackers use asynchronous multi-hypothesis updates or photometric matching

(Alzugaray and Chli, 2019; 2020; Gehrig et al., 2020). Related optical-flow methods use correlation features or continuous motion curves to estimate dense image motion (Gehrig et al., 2021; 2024).

DeepEvT (Messikommer et al., 2023) learns reference-image/event associations through cross-track attention and local patch matching. ETAP (Hamann et al., 2025) uses global context and feature alignment, while MATE (Han et al., 2025) combines kinematic and historical features for temporal consistency. BlinkTrack (Shen et al., 2025) fuses event and image estimates through differentiable Kalman filtering, and TAPFormer (Liu et al., 2026) models inter-frame evolution with reliability-weighted frame–event fusion. These methods strengthen correspondence through representation, motion modeling, and temporal fusion, but do not explicitly incorporate the pre-trigger response consistency implied by candidate motion into their matching criteria.

2.2 Monocular Depth and 3D Point Tracking

Monocular depth models learn relations between appearance and geometry. Depth Anything V2 (Yang et al., 2024) uses teacher–student training, while Metric3D v2 (Hu et al., 2024) learns metric geometric priors. For events, Zhu et al. (2019) train separate networks for optical-flow estimation and joint depth–egomotion estimation without ground-truth motion supervision. E2Depth uses a depth-supervised recurrent encoder–decoder (Hidalgo-Carrió et al., 2020). Depth AnyEvent transfers image-depth knowledge through cross-modal distillation, with a recurrent variant further aggregating event history (Bartolomei et al., 2025).

SpatialTracker (Xiao et al., 2024) and TAPIP3D (Zhang et al., 2025) use depth-supported 3D representations, whereas SpatialTrackerV2 (Xiao et al., 2025) jointly estimates scene geometry, camera motion, and point trajectories from monocular video. TAPVid-3D (Koppula et al., 2024) provides a benchmark for 3D point tracking. Among event-based 3D trackers, E-3DTrack (Li et al., 2024a) uses calibrated stereo geometry, while E-TraMamba (Shen et al., 2026) supports stereo events or monocular events with per-time-step depth inputs.

Another line of work estimates longitudinal motion from changes in projective scale. Optical expansion relates image scale to relative depth (Yang and Ramanan, 2020). EMoTive (Wan et al., 2025) derives depth ratios under non-rotating rigid motion with constant world-frame velocity. Han et al. (2026) combine directional expansion modulation with depth-aware contrast maximization to estimate dense optical flow and motion in depth, represented as depth ratios, for scene-flow recovery.

Persistent 3D point tracking associates longitudinal-motion cues with the same query across time. Our method combines query correspondence and local geometry to estimate metric trajectories from monocular events after image-based initialization.

2.3 Event-Generation Models for Motion Estimation

Event-generation models relate brightness changes to trigger timing for motion estimation. Contrast maximization scores motion hypotheses by the concentration of compensated events (Gallego et al., 2018), but its alignment objective does not itself test whether a candidate predicts an unobserved earlier trigger. Contraction can also cause event-collapse degeneracies (Shiba et al., 2022). EKLT (Gehrig et al., 2020) predicts event responses from reference-image gradients, with its patch-wise constant-flow approximation becoming restrictive across motion boundaries.

Bardow et al. (2016) use event and no-event data terms for joint optical-flow and intensity estimation. We instead use candidate-induced pre-trigger histories to evaluate query correspondence, integrating interval evidence into matching scores and correspondence learning.

3 Method

Our framework comprises three stages (Figure 2). Initialization refines the depth prior and establishes query templates. Hit–Survival scores candidate paths through event support and pre-trigger history, followed by reliability-weighted geometric reranking. The trajectory head aggregates correspondence and motion evidence to update metric 3D trajectories relative to the refined anchor.

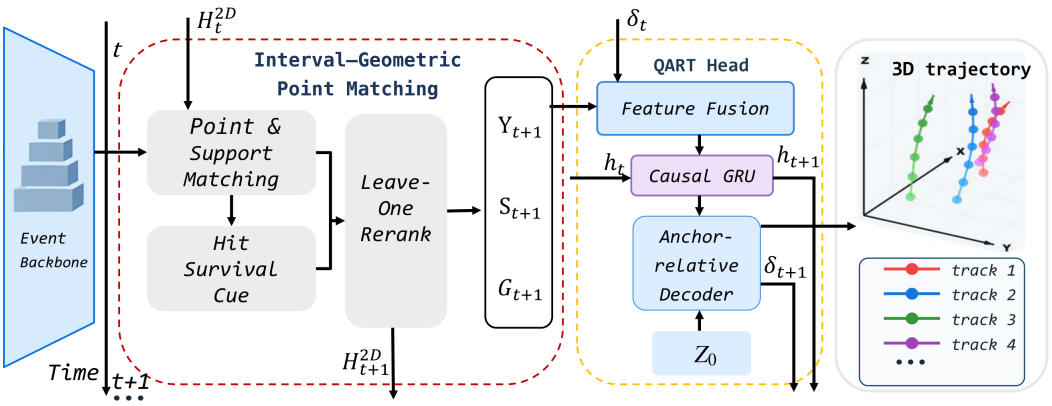

Figure 2: **Method overview.** Query and support candidates are scored by Hit–Survival and reranked by local geometry. The trajectory head integrates this evidence over time and decodes anchor-relative 3D positions using the refined initial depth. Only monocular events are added after initialization.

3.1 Inputs and Event Encoding

Given query pixels $\mathbf{q}_i$ $(i = 1, \ldots, N)$, our method predicts image positions $\widehat{\mathbf{u}}_{i,t}$ and metric positions $\widehat{\mathbf{X}}^0_{i,t}$ in the initial camera frame, for $t = 0, \ldots, T-1$. Inputs are monocular events, reference image I_0, and intrinsics K. Metric3D v2 supplies depth prior d_i^P (Hu et al., 2024). Each query has N_s local supports (Figure 3). The index $r = (i, \ell)$ denotes the query ($\ell = 0$) or a support ($\ell = 1, \ldots, N_s$).

Local supports. Harris corners are selected in a window around each query, with minimum spacing from the query and previously selected supports. Horizontal, vertical, and diagonal offsets complete the set when needed (Appendix A.5). Spatially spread supports provide a two-dimensional motion configuration for the local fit. Each support is first matched independently. Its displacement and confidence then contribute to leave-one-out prediction and geometric reweighting of query candidates.

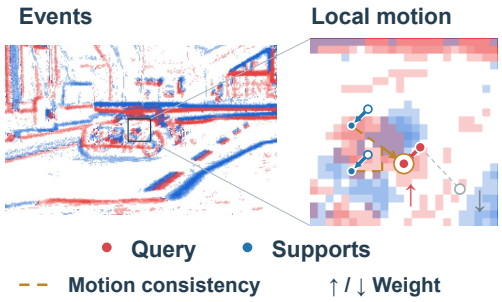

Figure 3: **Local motion consistency.** Support motion favors the highlighted query candidate over the gray alternative.

Event encoding. Time-binned event voxels yield feature maps for correspondence, while reference image patches provide query and support templates. Per-pixel event sequences retain predecessor times and trigger order. This preserves the temporal detail needed to evaluate candidate-induced responses within each interval.

Initialization. A jointly trained initializer uses the reference patch and first four event observations to predict a log-depth residual υ_i (Appendix A.4). This residual adapts the image-derived depth prior to the initial event observations. Positive and negative residuals increase and decrease the prior before clipping to configured bounds $0 < Z_{\min} < Z_{\max}$:

$$Z_{i,0} = \exp\big[\mathrm{clip}\big(\log d_i^P + \upsilon_i, \log Z_{\min}, \log Z_{\max}\big)\big]\,. \tag{1}$$

The anchor $\mathbf{a}_i = Z_{i,0} K^{-1} \widetilde{\mathbf{q}}_i$, refined depth, and templates are then fixed. Here $\widetilde{\mathbf{q}}_i$ denotes homogeneous coordinates. Subsequent events drive updates of the current 3D positions (Section 3.3).

3.2 Interval–Geometric Point Matching

Each update evaluates time-resolved candidate paths, combines Hit–Survival evidence with reliability-weighted local motion, and reads out the refined 2D location.

Candidate matching. A shared matcher combines current event features, initial templates, and temporal features to propose endpoints $\mathbf{c}_{rj,t}$ with scores $z_{rj,t}$. Each endpoint connects to the previous estimate $\widehat{\mathbf{u}}_{r,t-1}$ through the within-window path

$$\boldsymbol{\gamma}_{rj}(\tau) = (1 - \mu(\tau))\widehat{\mathbf{u}}_{r,t-1} + \mu(\tau)\mathbf{c}_{rj,t}, \tag{2}$$

where $\mu(\tau)$ is normalized time within the current update window (Appendix A.4). Candidates share the same retained trigger intervals but induce different paths through the reference patch. Evaluating these paths compares alternative explanations of a common observation, tying interval evidence to the correspondence being tested.

Hit: observed-event support. Hit evaluates the path center against eight surrounding samples in each of B temporal bins. Activity-weighted center–surround concentration is aggregated and scaled by temporal coverage to form $s^H_{rj,t}$. Valid sampling and at least two active bins establish spatially concentrated, temporally supported evidence (Appendix A.1).

Survival: pre-trigger history. For each same-pixel interval k, a candidate maps the event pixel into its reference patch at the predecessor, M interior times, and the trigger. Polarity-aligned log-intensity changes from the predecessor give interior responses a_{krjm} and endpoint b_{krj}. This sampled response-ordering test assigns a soft cost to pre-trigger excess and endpoint polarity disagreement:

$$\mathcal{U}_{krj} = \frac{\lambda_{\text{pre}}}{M} \sum_{m=1}^{M} \psi\left(\frac{a_{krjm} - b_{krj}}{\theta_{\text{resp}} d_{krjm}} \right) + \lambda_{\text{dir}} \psi\left(\frac{-b_{krj}}{\theta_{\text{resp}} d^{\text{end}}_{krj}} \right), \tag{3}$$

where $\psi(x) = \log(1 + \exp x)$, $\theta_{\text{resp}} > 0$ controls softness, and $\lambda_{\text{pre}}, \lambda_{\text{dir}} \geq 0$ weight the two terms. The detached denominators $|a_{krjm}| + |b_{krj}|$ and $|b_{krj}|$ are lower-bounded by $\epsilon_{\text{num}} > 0$. The magnitude bound $1/\theta_{\text{resp}}$ on each normalized argument limits individual response influence. Using each endpoint as its own response reference avoids fitting an absolute contrast threshold. This is a sampled, one-sided ordering constraint rather than a continuous test of the full no-event condition.

Robust aggregation and fusion. Only finite, ordered intervals with valid predecessors and in-patch samples contribute. Costs are averaged within each pixel and then across pixels, balancing high-rate and low-rate support. Negating, centering, and clipping yields $s^S_{rj,t}$. Log differences and response normalization remove constant gain and common response scale away from numerical floors and saturation. Cross-pixel consensus, Hit–Survival agreement, margins, and coverage set $g_{r,t} \in [0, 1]$. The joint evidence $e_{rj,t} = s^H_{rj,t} + g_{r,t} s^S_{rj,t}$ refines the distribution:

$$\bar{e}_{rj,t} = \text{clip}(\text{sg}(e_{rj,t}), -e_{\max}, e_{\max}), \qquad p_{rj,t} = \text{softmax}_j(z_{rj,t} + \bar{e}_{rj,t}). \tag{4}$$

Here sg stops gradients. Queries and supports receive bounded corrections, with candidate log-odds changing by at most $2e_{\max}$. At zero reliability, feature-plus-Hit matching remains (Appendix A.1).

Local geometry. A leave-one-out affine fit excludes point r and iteratively downweights supports with large residuals. Its similarity component predicts $\mathbf{u}^G_{r,t}$. Fit quality controls the strength of geometric reranking (Appendix A.2):

$$\widetilde{p}_{rj,t} \propto p_{rj,t} \exp\left[-\frac{\lambda_G w^G_{r,t}}{2\sigma^2_G} \left\| \mathbf{c}_{rj,t} - \mathbf{u}^G_{r,t} \right\|^2_2 \right], \tag{5}$$

where $w^G_{r,t}$ combines support spread, confidence, and similarity-fit residuals, reducing geometric influence as neighborhood motion becomes less consistent. Parameters λ_G and σ_G set weight and distance scale. Setting $w^G_{r,t} = 0$ preserves the pre-geometric distribution. Excluding point r prevents its matched endpoint from determining the geometric prediction used to rerank its own candidates. Independent matching supplies the observations, while residual reweighting controls their influence. Thus, local support is conditional on motion compatibility rather than spatial proximity alone.

2D readout. The query estimate adds a learned residual to the expectation under normalized $\widetilde{p}_{rj,t}$:

$$\widehat{\mathbf{u}}_{i,t} = \Pi_\Omega \left(\sum_j \widetilde{p}_{(i,0)j,t} \mathbf{c}_{(i,0)j,t} + \boldsymbol{\rho}^{2D}_{i,t} \right). \tag{6}$$

Here Π_Ω clips to the image domain. The query-specific residual $\boldsymbol{\rho}^{2D}_{i,t}$ refines the probability-weighted location. Confidence and local-motion features pass to the trajectory head.

3.3 ANCHOR-RELATIVE 3D TRAJECTORY INFERENCE

Our trajectory head (QART) estimates initial-depth-normalized displacement from correspondence features, support motion, and local geometry. Similar probability-weighted 2D locations can arise from distributions with different concentration and support agreement. A coordinate-only readout discards this distinction. Retaining the candidate and support features lets the spatial update use both inferred motion and its reliability, rather than treating every 2D estimate as equally informative.

Feature fusion and temporal aggregation. For query i, $\mathbf{y}_{i,t}$ contains matching features, $\mathbf{S}_{i,t}$ aggregates reliable supports, and $\mathbf{G}_{i,t}$ describes local geometry. A reference descriptor $\boldsymbol{\xi}_{i,t}^{\mathrm{ref}}$ encodes the initial ray, scale, time interval, and previous $\boldsymbol{\delta}_{i,t-1}$. Their encodings form observation $\mathbf{f}_{i,t}$. A causal GRU updates the 3D hidden state $\mathbf{h}_{i,t}$, separate from 2D memory $H_{i,t}^{2D}$ (Appendix A.3).

Anchor-relative decoding. The decoder predicts a normalized displacement and restores metric units using the fixed initial reference:

$$\boldsymbol{\delta}_{i,t} = \beta \tanh(\mathbf{W}_d \mathbf{h}_{i,t} + \mathbf{b}_d), \qquad \widehat{\mathbf{X}}_{i,t}^{0} = \mathbf{a}_i + Z_{i,0}\boldsymbol{\delta}_{i,t}. \tag{7}$$

$\mathbf{W}_d$ and $\mathbf{b}_d$ are decoder weights and bias, and $\beta > 0$ bounds each displacement component. Each update predicts the current offset from $\mathbf{a}_i$ instead of accumulating successive 3D increments. Initial-depth normalization expresses motion relative to the query's metric scale, while the fixed anchor keeps all updates in a common reference. Local geometry and fit reliability guide this update.

The current depth evolves as $\widehat{Z}_{i,t}^{0} = Z_{i,0}(1 + \delta_{i,t}^{z})$, where $\delta_{i,t}^{z}$ is the third displacement component in Equation (7). To characterize this representation, let $\mathbf{X}_{i,t}^{0,\mathrm{gt}}$ denote the ground-truth position at a scored time. The corresponding offset decomposes as

$$\boldsymbol{\delta}_{i,t}^{\mathrm{req}} = \frac{\mathbf{X}_{i,t}^{0,\mathrm{gt}} - \mathbf{a}_i}{Z_{i,0}} = \frac{\mathbf{X}_{i,t}^{0,\mathrm{gt}} - \mathbf{X}_{i,0}^{0,\mathrm{gt}}}{Z_{i,0}} + \frac{\mathbf{X}_{i,0}^{0,\mathrm{gt}} - \mathbf{a}_i}{Z_{i,0}}. \tag{8}$$

The two terms represent motion and compensation for remaining anchor bias. Within $\|\boldsymbol{\delta}_{i,t}^{\mathrm{req}}\|_\infty < \beta$, one event-conditioned offset represents both components. The trajectory head estimates this state from correspondence confidence, reliable local motion, and temporal history. Appendix F derives the error relationship. Section 4.4 reports the paired test-set perturbation study.

3.4 JOINT LEARNING OBJECTIVE

Training combines Charbonnier supervision of 2D and 3D positions with evidence-weighted correspondence learning and robust geometric consistency (Appendix B.1).

Hit–Survival marginal objective. On common valid domain $\mathcal{J}_{r,t}$, define detached weights $\omega_{rj,t} = \mathrm{softmax}_{j\in\mathcal{J}_{r,t}}(\mathrm{sg}(e_{rj,t})/\theta_{\mathrm{HS}})$, with temperature $\theta_{\mathrm{HS}} > 0$ and zero weights outside the domain. Let $\mathcal{A}_t$ contain queries and supports with $|\mathcal{J}_{r,t}| \geq 2$ and $g_{r,t} > 0$. We minimize

$$\mathcal{L}_{\mathrm{HS}} = -\frac{1}{N_p(T-1)} \sum_{t=1}^{T-1} \sum_{r\in\mathcal{A}_t} \frac{\log\left[\sum_{j\in\mathcal{J}_{r,t}} \max(p_{rj,t}, \epsilon_p)\, \omega_{rj,t}\right]}{\log|\mathcal{J}_{r,t}|}. \tag{9}$$

Here $N_p = N(N_s + 1)$ counts queries and supports, $\epsilon_p > 0$ stabilizes probabilities, and $p_{rj,t}$ is the pre-geometric distribution. The loss favors evidence-supported candidates. Evidence weights are held fixed while gradients update the matcher. Inactive point–time pairs contribute zero. Different candidate distributions can yield the same expected location, so position supervision need not distinguish their underlying evidence. The interval objective explicitly favors distributions supported by candidate-specific histories before their features enter spatial inference.

Local geometric consistency. Image-normalized robust penalties supervise valid local fits (Appendix B.3). The joint loss is $\mathcal{L} = \mathcal{L}_{2D} + \mathcal{L}_{3D} + \alpha_{\mathrm{HS}}\mathcal{L}_{\mathrm{HS}} + \alpha_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}}$, with nonnegative weights.

4 EXPERIMENTS

We evaluate tracking accuracy on two datasets and examine the proposed mechanisms through component ablations, fixed-candidate history controls, and paired depth-prior perturbations.

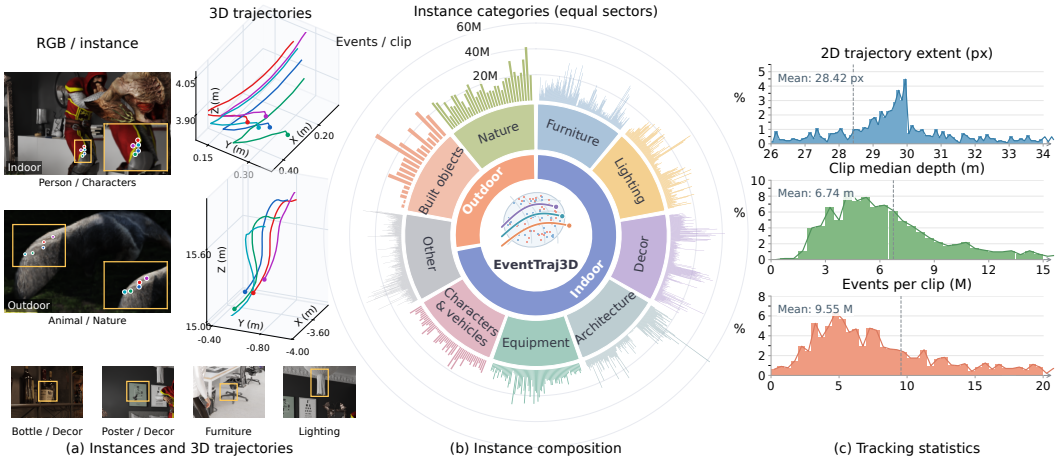

Figure 4: **EventTraj3D overview.** (a) Image instances and corresponding metric 3D trajectories. (b) Equal-sized instance-category sectors with radial bars showing events per clip, not category frequency. (c) Distributions of 2D trajectory extent, clip median depth, and event count.

4.1 Dataset and Evaluation Setup

EventTraj3D. We construct EventTraj3D from BlinkVision (Li et al., 2024b), with 13.630 billion events and 6,494 aligned 2D–3D query trajectories (Figure 4). Its 1,427 clips are split into 999 training, 214 validation, and 214 test clips. A comparison with existing point-tracking datasets is provided in Appendix C.1, Table C2.

E-3DTrack. We also evaluate on E-3DTrack (Li et al., 2024a) using monocular events, with all methods retrained on this dataset.

Experimental setup. All methods use the same initialization model. We train our model for 100 epochs using AdamW (Loshchilov and Hutter, 2019). Evaluation and implementation details are provided in Appendix C and Appendix B.4.

4.2 Quantitative Comparison

Our method achieves the lowest 2D EPE and FDE on EventTraj3D (Table 1). The gains cover both average localization and final tracked positions, consistent with more reliable temporal correspondence. Section 4.4 examines the contribution of interval history through component ablations and controls with fixed proposals and matcher weights.

Table 1: **EventTraj3D 2D errors (px).** Best in **bold**, second best underlined.

Method	EPE (px) ↓	FDE (px) ↓
DeepEvT	4.604	8.908
ETAP	<u>4.189</u>	<u>8.276</u>
TAPFormer	6.034	12.395
SPAR (Ours)	**3.213**	**6.345**

Our method also achieves the lowest 3D EPE, ADE, and FDE (Table 2). Under shared initialization, the gains cover both absolute position and anchor-relative displacement. This supports using correspondence and local-motion evidence for temporal 3D inference rather than relying only on depth estimates along tracked pixels. The model uses 44.80 M parameters, the second-lowest count, and 1.10 GiB peak GPU memory. Thus, the accuracy improvement does not require greater model capacity than most compared pipelines.

4.3 Evaluation on E-3DTrack

On E-3DTrack, our method achieves the lowest RMSE and best FA/TFR at 0.15 m and 0.2 m, while ranking second at 0.1 m (Table 3). These results support improved overall spatial accuracy and tracking at moderate error tolerances on a second dataset. The stricter 0.1 m comparison indicates that lower aggregate error does not imply uniformly better fine-scale localization.

Table 2: **EventTraj3D 3D errors and resources.** Errors in mm. ADE/FDE measure anchor-relative displacement. DAE-R: DepthAnyEvent-R. Best in **bold**, second best underlined.

Method	EPE↓	ADE↓	FDE↓	Params (M)↓	Peak GPU (GiB)↓
DeepEvT + E2Depth	657.98	615.06	1149.00	45.60	1.02
ETAP + E2Depth	656.06	613.91	1134.31	56.19	2.99
TAPFormer + E2Depth	660.20	617.91	1187.82	**40.16**	2.04
DeepEvT + DAE-R	511.53	480.11	985.98	61.01	**0.87**
ETAP + DAE-R	509.04	478.54	981.17	71.59	2.84
TAPFormer + DAE-R	514.79	485.11	988.71	55.57	1.90
SPAR (Ours)	**400.69**	**376.56**	**821.74**	44.80	1.10

Table 3: **Monocular 3D tracking on E-3DTrack.** FA/TFR are percentages at thresholds in meters. RMSE is in meters. DAE-R denotes DepthAnyEvent-R. Best in **bold**, second best underlined.

Method	$FA_{0.1}$↑	$FA_{0.15}$↑	$FA_{0.2}$↑	$TFR_{0.1}$↑	$TFR_{0.15}$↑	$TFR_{0.2}$↑	RMSE (m)↓
DeepEvT + E2Depth	3.36	7.87	13.64	6.76	13.76	22.19	0.3580
ETAP + E2Depth	7.34	15.12	23.79	22.02	33.63	44.12	0.2566
TAPFormer + E2Depth	7.12	14.46	22.93	20.22	31.17	41.32	0.2958
DeepEvT + DAE-R	4.65	9.43	16.01	8.06	15.72	24.51	0.3491
ETAP + DAE-R	**10.39**	18.08	27.40	**23.36**	35.72	46.14	0.2539
TAPFormer + DAE-R	7.64	16.56	25.29	21.32	33.45	43.16	0.2918
SPAR (Ours)	9.73	**22.74**	**39.00**	22.55	**41.73**	**57.49**	**0.2134**

4.4 Ablation and Robustness Analysis

Removing Hit–Survival produces the largest increase in 2D error and also degrades 3D tracking (Table 4). This links interval-based matching to query localization and subsequent spatial estimation. Removing local geometry increases 3D errors more despite a smaller 2D change. The different 2D and 3D rankings show that image-plane error alone is insufficient to assess spatial tracking quality. Together, these ablations support complementary roles for interval history in correspondence and neighborhood motion in 3D estimation.

Table 4: **Ablation.** FDE is anchor-relative. Best in **bold**, second best underlined.

Variant	2D EPE (px) ↓	3D EPE (mm) ↓	3D FDE (mm) ↓
w/o Hit–Survival	5.735	434.19	928.77
w/o local geometry	4.390	462.96	973.79
Direct XYZ	3.886	1167.57	1644.48
SPAR (full)	**3.213**	**400.69**	**821.74**

All variants share initialization, splits, and training. Direct XYZ retains encoders, GRU, and input features, with comparable loss normalization, validity masks, and gradient scaling. Output units and bounds are adapted to direct coordinates. The direct head has much higher 3D EPE and FDE even though its 2D EPE remains closer to the full model. This comparison supports the complete normalized, anchor-relative output parameterization, which improves spatial estimation without a commensurate change in image-plane error.

Initial-depth sensitivity. With fixed weights and three paired seeds, our method incurs less added 3D EPE than a direct 3D head under 5–30% multiplicative prior perturbations on EventTraj3D. Noise levels are relative standard deviations, and added error is measured from each model's unperturbed baseline (Appendix F.2). The smaller paired increases support reduced propagation of initialization error through the combined initialization-and-tracking pipeline. Error still grows with perturbation strength, so the result does not remove dependence on the metric anchor.

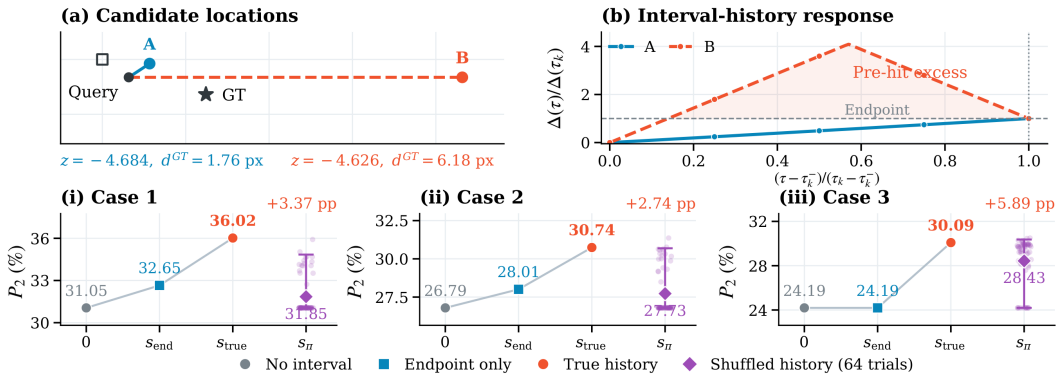

Figure 5: **Candidate-history analysis.** (a–b) Candidate locations and endpoint-normalized responses. (i–iii) Probability within 2 px of ground truth under fixed-candidate history controls.

Hit–Survival analysis. Candidate B scores slightly above A despite a larger localization error (6.18 *vs* 1.76 px). Only B's response exceeds its endpoint reference (Figure 5(a–b)). This example exposes the candidate-evaluation gap: similar endpoint support can conceal different pre-trigger histories, and response ordering supplies a criterion favoring the more accurate correspondence.

With weights, candidates, and base scores fixed, true history increases probability mass within 2 px of ground truth over endpoint-only evidence in all three cases (Figure 5(i–iii)). Posterior-weighted error decreases in all three cases, although the highest-probability candidate changes only in (iii). Thus, history improves confidence allocation as well as correspondence selection, without different proposals or matcher retraining. Both the 2D readout and the matching features passed to the trajectory head use the candidate distribution. Confidence changes can therefore affect the tracking update even when the highest-scoring candidate remains unchanged. True history also improves on the mean of 64 accepted shuffles per case (Appendix D). These shuffles preserve spatial locations, polarities, event counts, and interval-duration statistics. The comparison therefore links the gain to predecessor–trigger association rather than additional event support.

4.5 Qualitative Results

In the selected examples, our trajectories follow the ground-truth direction, curvature, and spatial extent more closely than ETAP + DAE-R (Figure 6). The left case retains the post-turn segment, the middle follows the curved motion, and the right reduces path deviation despite residual endpoint offsets. The closer paths support improved trajectory geometry beyond endpoint agreement. Additional examples on both datasets are provided in Appendix G.

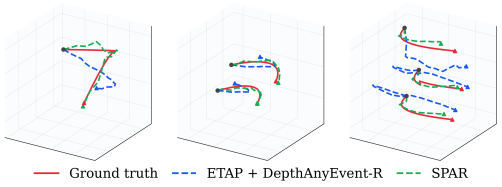

Figure 6: **3D trajectories.** Ground truth (red), ETAP + DAE-R (blue), and our method (green).

5 Conclusion

We presented a framework for sparse metric 3D point tracking from monocular events after image-based initialization. Hit–Survival evaluates the sampled pre-trigger histories implied by competing correspondences, while reliability-weighted local geometry supplies complementary motion evidence. The anchor-relative trajectory head preserves this evidence for temporal 3D inference rather than reducing each update to an image coordinate. Experiments on EventTraj3D and E-3DTrack demonstrate improved tracking, and paired perturbations show reduced initialization sensitivity. Together, these results support coupling candidate-level observation modeling with scale-normalized spatial prediction. Future work will investigate boundary-aware support selection and longer-horizon interval reasoning to better handle occlusion, deformation, and dense point tracking.

REFERENCES

Ignacio Alzugaray and Margarita Chli. Asynchronous Multi-Hypothesis Tracking of Features with Event Cameras. In *International Conference on 3D Vision (3DV)*, pp. 269–278, 2019.

Ignacio Alzugaray and Margarita Chli. HASTE: multi-Hypothesis Asynchronous Speeded-up Tracking of Events. In *British Machine Vision Conference (BMVC)*, 2020.

Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous Optical Flow and Intensity Estimation From an Event Camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 884–892, 2016.

Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth AnyEvent: A Cross-Modal Distillation Paradigm for Event-Based Monocular Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19669–19678, 2025.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A Benchmark for Tracking Any Point in a Video. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking Any Point with Per-Frame Initialization and Temporal Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10061–10072, 2023.

Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A Unifying Contrast Maximization Framework for Event Cameras, With Applications to Motion, Depth, and Optical Flow Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3867–3876, 2018.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Joerg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (1): 154–180, 2022.

Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. EKLT: Asynchronous Photometric Feature Tracking Using Events and Frames. *International Journal of Computer Vision*, 128 (3): 601–618, 2020.

Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras. In *International Conference on 3D Vision (3DV)*, 2021.

Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense Continuous-Time Optical Flow From Event Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (7): 4736–4746, 2024.

Friedhelm Hamann, Daniel Gehrig, Filbert Febryanto, Kostas Daniilidis, and Guillermo Gallego. ETAP: Event-based Tracking of Any Point. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27186–27196, 2025.

Han Han, Wei Zhai, Yang Cao, Bin Li, and Zheng-jun Zha. MATE: Motion-Augmented Temporal Consistency for Event-based Point Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2025.

Han Han, Wei Zhai, Tiesong Zhao, Bin Li, Yang Cao, and Zheng-jun Zha. Unsupervised 3d Motion Estimation Using Event Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8067–8076, 2026.

Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. In *European Conference on Computer Vision*, pp. 59–75, 2022.

Chris Harris and Mike Stephens. A Combined Corner and Edge Detector. In *Alvey Vision Conference*, pp. 147–151, 1988.

Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning Monocular Dense Depth from Events. In *International Conference on 3D Vision*, 2020.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (12): 10579–10596, 2024.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It Is Better to Track Together. In *European Conference on Computer Vision*, pp. 18–35, 2024.

Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, João Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. TAPVid-3D: A Benchmark for Tracking Any Point in 3D. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19855–19865, 2023.

Siqi Li, Zhikuan Zhou, Zhou Xue, Yipeng Li, Shaoyi Du, and Yue Gao. 3D Feature Tracking via Event Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18974–18983, 2024a.

Yijin Li, Yichen Shen, Zhaoyang Huang, Shuo Chen, Weikang Bian, Xiaoyu Shi, Fu-Yun Wang, Keqiang Sun, Hujun Bao, Zhaopeng Cui, Guofeng Zhang, and Hongsheng Li. BlinkVision: A Benchmark for Optical Flow, Scene Flow and Point Tracking Estimation Using RGB Frames and Events. In *European Conference on Computer Vision*, pp. 19–36, 2024b.

Jiaxiong Liu, Zhen Tan, Jinpu Zhang, Yi Zhou, Hui Shen, Xieyuanli Chen, and Dewen Hu. TAP-Former: Robust Arbitrary Point Tracking via Transient Asynchronous Fusion of Frames and Events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37053–37062, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Nico Messikommer, Carter Fang, Mathias Gehrig, and Davide Scaramuzza. Data-Driven Feature Tracking for Event Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5642–5651, 2023.

Juwei Shen, Yujie Wu, and Changwen Chen. E-TraMamba: A New Paradigm for Efficient Long-Term 3D Feature Tracking with Event Cameras. *arXiv preprint arXiv:2607.02866*, 2026.

Yichen Shen, Yijin Li, Shuo Chen, Guanglin Li, Zhaoyang Huang, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. BlinkTrack: Feature Tracking over 80 FPS via Events and Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9308, 2025.

Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Event Collapse in Contrast Maximization Frameworks. *Sensors*, 22 (14): 5190, 2022.

Zengyu Wan, Wei Zhai, Yang Cao, and Zhengjun Zha. EMoTive: Event-guided Trajectory Modeling for 3D Motion Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9342–9351, 2025.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. SpatialTracker: Tracking Any 2D Pixels in 3D Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20406–20417, 2024.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: Advancing 3D Point Tracking with Explicit Camera Motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6726–6737, 2025.

Gengshan Yang and Deva Ramanan. Upgrading Optical Flow to 3D Scene Flow Through Optical Expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1334–1343, 2020.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Bowei Zhang, Lei Ke, Adam W. Harley, and Katerina Fragkiadaki. TAPIP3D: Tracking Any Point in Persistent 3D Geometry. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 989–997, 2019.

A Supplementary Method Details

A.1 Candidate Distribution and Interval Reliability

This subsection suppresses the update index t for within-update quantities unless it is needed explicitly.

Event-generation reference. At fixed pixel $\mathbf{x}_k$, define actual log-brightness change $\Delta L(\mathbf{x}_k, \tau) = L(\mathbf{x}_k, \tau) - L(\mathbf{x}_k, \tau_k^-)$. With accurate reset, no missing events, and symmetric threshold magnitude $C > 0$, the ideal condition is

$$\Delta L(\mathbf{x}_k, \tau_k) = \pi_k C, \qquad |\Delta L(\mathbf{x}_k, \tau)| < C \quad (\tau_k^- < \tau < \tau_k). \tag{A1}$$

Motivated by Equation (A1), we use a candidate-conditioned, one-sided response-ordering score at finite samples of Equation (A5). Validity filtering, normalization, and reliability-gated fusion control the contribution of each interval.

Hit normalization. For each temporal bin b, we sample the candidate path center and the eight nonzero offsets in $\{-2, 0, 2\} \times \{-2, 0, 2\}$. Let E_{rjb}^{ctr} denote center activity and E_{rjb}^{ring} the mean activity over the eight neighboring offsets. Their center–surround concentration is

$$\kappa_{rjb} = \frac{E_{rjb}^{\mathrm{ctr}} - E_{rjb}^{\mathrm{ring}}}{E_{rjb}^{\mathrm{ctr}} + E_{rjb}^{\mathrm{ring}} + \epsilon_{\mathrm{num}}}. \tag{A2}$$

Let A_{rjb} be total activity over all nine locations, $\eta_{rjb} = A_{rjb} / \max_{b'} A_{rjb'}$, and $A_{rj}^{\mathrm{tot}} = \sum_b A_{rjb}$. Let n_{rj}^{act} count active bins among B. A candidate is in $\mathcal{J}_r^H$ if all sample locations are valid and at least two bins are active. Its uncentered score is

$$\widetilde{s}_{rj}^H = \frac{\sum_b \eta_{rjb}\kappa_{rjb}}{\sum_b \eta_{rjb}} \cdot \frac{A_{rj}^{\mathrm{tot}}}{\max_{j' \in \mathcal{J}_r^H} A_{rj'}^{\mathrm{tot}}} \cdot \frac{n_{rj}^{\mathrm{act}}}{B}. \tag{A3}$$

Subtracting the valid-candidate mean and bounding the result gives

$$s_{rj}^H = \mathrm{clip}\big(e_{\max}\left[\widetilde{s}_{rj}^H - \langle \widetilde{s}_{rj'}^H \rangle_{j' \in \mathcal{J}_r^H}\right], -e_{\max}, e_{\max}\big). \tag{A4}$$

All normalization denominators are lower-bounded by $\epsilon_{\mathrm{num}} > 0$. Unsupported scores are zero, and the final s_{rj}^H is detached. Angle brackets denote averaging over the indicated candidate set.

Reference response and interval sampling. At pixel $\mathbf{x}_k$, interval k has predecessor τ_k^-, trigger τ_k, and polarity $\pi_k \in \{-1, +1\}$. Sample $\tau_{k,m} = \tau_k^- + m(\tau_k - \tau_k^-)/(M+1)$ for $m = 1, \ldots, M$. RGB channels are averaged, bilinearly sampled at the candidate-mapped location, clamped below by $1/1024$, and logged. Denote this sample-then-log operation by $I_r^{\log}$ and the patch center by $\mathbf{o}_0$:

$$\Delta_{krj}(\tau) = \pi_k\left[I_r^{\log}\big(\mathbf{x}_k - \boldsymbol{\gamma}_{rj}(\tau) + \mathbf{o}_0\big) - I_r^{\log}\big(\mathbf{x}_k - \boldsymbol{\gamma}_{rj}(\tau_k^-) + \mathbf{o}_0\big)\right]. \tag{A5}$$

Set $a_{krjm} = \Delta_{krj}(\tau_{k,m})$ and $b_{krj} = \Delta_{krj}(\tau_k)$. Equation (3) uses $d_{krjm} = \max\{\mathrm{sg}(|a_{krjm}| + |b_{krj}|), \epsilon_{\mathrm{num}}\}$ and $d_{krj}^{\mathrm{end}} = \max\{\mathrm{sg}\,|b_{krj}|, \epsilon_{\mathrm{num}}\}$. These define the normalized response scale.

Validity and pixel-balanced aggregation. Retain finite same-pixel intervals with $\tau_k > \tau_k^-$. Let $\mathcal{X}_r$ contain pixels with nonempty retained interval sets $\mathcal{K}_r(\mathbf{x})$, and let $\mathcal{J}_r^S$ contain candidates whose samples are all finite and in-patch over these same intervals. The pixel-balanced aggregate is

$$\overline{\mathcal{U}}_{rj} = \frac{1}{|\mathcal{X}_r|} \sum_{\mathbf{x} \in \mathcal{X}_r} \frac{1}{|\mathcal{K}_r(\mathbf{x})|} \sum_{k \in \mathcal{K}_r(\mathbf{x})} \mathcal{U}_{krj}. \tag{A6}$$

Negating, candidate-centering, and clipping yields

$$s_{rj}^S = \mathrm{clip}\big(-s_{\max}\left[\overline{\mathcal{U}}_{rj} - \langle \overline{\mathcal{U}}_{rj'} \rangle_{j' \in \mathcal{J}_r^S}\right], -s_{\max}, s_{\max}\big). \tag{A7}$$

Here $s_{\max}$ bounds Survival. Reference samples, paths, and scores are detached. Empty pixel or candidate sets give zero Survival reliability.

Interval reliability. The common domain is $\mathcal{J}_r = \mathcal{J}_r^H \cap \mathcal{J}_r^S$. Indicator χ_r is one only if $|\mathcal{J}_r| \geq 2$, the Hit optimum over $\mathcal{J}_r^H$ belongs to $\mathcal{J}_r$, and it agrees with the Survival optimum over $\mathcal{J}_r$. Otherwise it is zero. The gate is

$$g_r = \chi_r \left(w_r^{\mathrm{corr}} w_r^{\mathrm{pix}} w_r^{\mathrm{margin}} w_r^{\mathrm{cand}} w_r^{\mathrm{sample}} w_r^{\mathrm{count}} \right)^{1/6}. \tag{A8}$$

The factors are respectively nonnegative score correlation, cross-pixel agreement, the smaller Hit/Survival margin reliability, candidate coverage, interval-sample coverage, and $w_r^{\mathrm{count}} = \max(|\mathcal{X}_r| - 1, 0)/\max(|\mathcal{X}_r|, 1)$. Each lies in $[0, 1]$, and g_r is detached. Agreement between the Hit and Survival optima enables reliable interval evidence to adjust the base ranking.

Candidate distribution. Base scores $z_{rj,t}$ include feature matching, candidate-group priors, and cardinality normalization. The same $e_{\max}$ bounds Hit and final joint evidence. Both queries and supports receive correction. The query retains independent 2D memory. The learning weights are formed from unclipped $e_{rj,t}$.

Evidence weighting. Center–surround contrast measures spatial concentration, and temporal coverage weights support across the update window. Pixel-balanced aggregation weights valid pixels equally. Cross-pixel consistency and coverage regulate the evidence entering the bounded fusion.

Response normalization. Let $J_r(\mathbf{v})$ be the bilinearly sampled scalar reference intensity before clamping. A constant gain $a_{\mathrm{gain}} > 0$ obeys

$$\log\big(a_{\mathrm{gain}} J_r(\mathbf{v}_1)\big) - \log\big(a_{\mathrm{gain}} J_r(\mathbf{v}_0)\big) = \log J_r(\mathbf{v}_1) - \log J_r(\mathbf{v}_0). \tag{A9}$$

The cancellation holds away from intensity floors and saturation. Positive common scaling of a_{krjm}, b_{krj} also cancels in Equation (3) when the denominator floor is inactive. Reliability gating further attenuates photometrically inconsistent interval evidence before the final bounded fusion.

Bounded influence of accepted evidence. The denominators satisfy $d_{krjm} \geq |a_{krjm}| + |b_{krj}|$ and $d_{krj}^{\mathrm{end}} \geq |b_{krj}|$ in the forward pass. Consequently, the arguments of both softplus terms lie in $[-1/\theta_{\mathrm{resp}}, 1/\theta_{\mathrm{resp}}]$ and

$$0 \leq \mathcal{U}_{krj} \leq (\lambda_{\mathrm{pre}} + \lambda_{\mathrm{dir}}) \psi(1/\theta_{\mathrm{resp}}). \tag{A10}$$

The joint clip additionally bounds its effect on correspondence. For valid candidates j, j' with finite base scores, Equation (4) gives

$$\left| \log \frac{p_{rj,t}}{p_{rj',t}} - (z_{rj,t} - z_{rj',t}) \right| = |\bar{e}_{rj,t} - \bar{e}_{rj',t}| \leq 2e_{\max}. \tag{A11}$$

Thus the implemented score provides a bounded adjustment to candidate ranking even when individual responses have large amplitudes.

A.2 Local Geometric Prediction and 2D Estimation

Excluding point r, a confidence-weighted affine map fits the other points' historical positions to independently matched endpoints. Iteratively reweighted least squares updates each weight from base confidence $c_{r'}^{\mathrm{match}}$ and affine residual $\varepsilon_{r'}^{\mathrm{aff},(n)}$:

$$\vartheta_{r'}^{(n+1)} = \frac{c_{r'}^{\mathrm{match}}}{\sqrt{1 + (\varepsilon_{r'}^{\mathrm{aff},(n)}/d_{\mathrm{rob}})^2}}, \qquad r' \neq r. \tag{A12}$$

Here $d_{\mathrm{rob}} > 0$ sets residual scale. Invalid points have zero weight. From a valid positive-determinant affine matrix $\mathbf{A}_{-r,t}$, extract proper rotation $\mathbf{R}_{-r,t}$ and isotropic scale $\zeta_{-r,t} = \sqrt{\det \mathbf{A}_{-r,t}}$. Translation $\mathbf{b}_{-r,t}^G$ aligns the weighted centroids under this similarity component:

$$\mathbf{u}_{r,t}^G = \zeta_{-r,t} \mathbf{R}_{-r,t} \widehat{\mathbf{u}}_{r,t-1} + \mathbf{b}_{-r,t}^G. \tag{A13}$$

Reliability is computed from the residual of the extracted similarity component. Writing the remaining bounded support, confidence, conditioning, and robust-weight factors as $c_{r,t}^G \in [0, 1]$, it has the form

$$w_{r,t}^G = \iota_{r,t}^{\mathrm{fit}} c_{r,t}^G \exp\!\left(-\varepsilon_{-r,t}^{\mathrm{sim}}/\tau_{\mathrm{fit}}\right). \tag{A14}$$

Here $\varepsilon^{\mathrm{sim}}_{-r,t}$ is the weighted RMS similarity residual, $\tau_{\mathrm{fit}} > 0$ its scale, and $\iota^{\mathrm{fit}}_{r,t}$ requires finite, nondegenerate geometry with enough effective support. A larger similarity residual under local deformation reduces reliability. At zero reliability, Equation (5) preserves independent matching. Each point retains its own candidate distribution and learned residual. The 2D residual uses query features, posterior-weighted features, and entropy. Refined features and confidence update $\mathbf{y}_{i,t}$ and the 2D memory carried into the next observation window.

A.3 Local Observation Representation

$\mathbf{y}_{i,t}$ contains query features, refined matching features, and confidence. Support records provide matching features, relative positions, 2D displacements, and Hit–Survival reliability. Shared encodings, weighted by confidence times distribution concentration, form $\mathbf{S}_{i,t}$. Descriptor $\mathbf{G}_{i,t}$ contains scale, rotation, translation, residuals, spread, and reliability. The reference descriptor $\boldsymbol{\xi}^{\mathrm{ref}}_{i,t}$ adds the initial ray, metric scale, time interval, and previous relative state.

Encoders E_y, E_G, E_ξ and fusion map F form the temporal observation:

$$\mathbf{f}_{i,t} = F\big([E_y(\mathbf{y}_{i,t}), \mathbf{S}_{i,t}, E_G(\mathbf{G}_{i,t}), E_\xi(\boldsymbol{\xi}^{\mathrm{ref}}_{i,t})]\big)\,. \tag{A15}$$

The causal update is

$$\mathbf{h}_{i,t} = \mathrm{GRU}(\mathbf{f}_{i,t}, \mathbf{h}_{i,t-1}). \tag{A16}$$

Figure 2 suppresses query indices: Y_t, S_t, G_t collect $\mathbf{y}_{i,t}, \mathbf{S}_{i,t}, \mathbf{G}_{i,t}$, while $h_t, H^{2D}_t, \boldsymbol{\delta}_t$ denote 3D state, 2D memory, and normalized displacement.

A.4 Initial Anchor and Candidate Paths

The positive prior is $d^P_i = \mathcal{D}(I_0, K)(\mathbf{q}_i)$. Initial patch features and four event observations predict υ_i in Equation (1), combining shared log-scale and zero-mean point corrections. The prior is detached. Setting $\upsilon_i = 0$ recovers it before clipping. After initialization, $Z_{i,0}$, $\mathbf{a}_i = Z_{i,0}K^{-1}\widetilde{\mathbf{q}}_i$, and templates stay fixed while the trajectory head updates the relative state (Appendix F). For $\tau^{\mathrm{obs}}_{t-1} < \tau^{\mathrm{obs}}_t$, the path interpolation is $\mu(\tau) = (\tau - \tau^{\mathrm{obs}}_{t-1})/(\tau^{\mathrm{obs}}_t - \tau^{\mathrm{obs}}_{t-1})$.

A.5 Local Support Selection

We detect Harris corners (Harris and Stephens, 1988) within a local image window centered on each query, excluding candidates too close to the query or already selected supports. If necessary, predefined horizontal, vertical, and diagonal offsets provide additional points within the image boundaries. Feasible point sets are selected for spatial spread and non-collinearity.

B Training Objectives and Implementation Details

B.1 Position Supervision and Normalization

For vector residual $\mathbf{v}$, the Charbonnier penalty is $\rho_\epsilon(\mathbf{v}) = \sqrt{\|\mathbf{v}\|_2^2 + \epsilon^2} - \epsilon$. Valid set $\mathcal{V}_2$ contains 2D supervision after the initial query time. Its loss is

$$\mathcal{L}_{2D} = \frac{1}{|\mathcal{V}_2|} \sum_{(i,t)\in\mathcal{V}_2} \rho_{\epsilon_2}\left(\frac{\widehat{\mathbf{u}}_{i,t} - \mathbf{u}^{\mathrm{gt}}_{i,t}}{D_{\mathrm{img}}}\right). \tag{B1}$$

The 3D supervision set $\mathcal{V}_3$ starts after the four-observation initialization. Its depth-normalized loss is

$$\mathcal{L}_{3D} = \frac{1}{|\mathcal{V}_3|} \sum_{(i,t)\in\mathcal{V}_3} \rho_{\epsilon_3}\left(\frac{\mathcal{G}_{\eta_\nabla}(\widehat{\mathbf{X}}^0_{i,t}) - \mathbf{X}^{0,\mathrm{gt}}_{i,t}}{d^P_i}\right). \tag{B2}$$

Here $D_{\mathrm{img}} = \sqrt{H^2_{\mathrm{img}} + W^2_{\mathrm{img}}}$, $\epsilon_2 = 1/D_{\mathrm{img}}$, $\epsilon_3 = 0.01$, and superscript gt denotes ground truth. The depth prior d^P_i is fixed and positive. Operator $\mathcal{G}_{\eta_\nabla}$ preserves the forward value and scales its backward gradient by $\eta_\nabla = 0.1$.

B.2 Interval-Evidence Marginalization

Equation (9) uses detached, unclipped joint evidence on the common domain. Queries and supports contribute only with at least two common valid candidates and positive reliability. Terms are divided by log domain size and averaged over all slots and updates, including inactive slots as zeros.

B.3 Geometric Consistency and Joint Optimization

For local fit residual $\varepsilon_{i,t}^{G}$ and validity indicator $\iota_{i,t}^{G}$, the geometric regularizer is

$$\mathcal{L}_{\mathrm{geo}} = \frac{1}{N(T-1)} \sum_{t=1}^{T-1} \sum_{i=1}^{N} \iota_{i,t}^{G} \sqrt{\left(\varepsilon_{i,t}^{G} / D_{\mathrm{img}}\right)^2 + \epsilon_g^2}. \tag{B3}$$

Here D_{img} is defined in Appendix B.1 and ϵ_g smooths the penalty. The joint objective is $\mathcal{L} = \mathcal{L}_{2D} + \mathcal{L}_{3D} + \alpha_{\mathrm{HS}}\mathcal{L}_{\mathrm{HS}} + \alpha_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}}$.

B.4 Implementation Details

We train our model for 100 epochs using AdamW (Loshchilov and Hutter, 2019), with fixed learning rates of 2×10^{-6} for the 2D tracking and initialization branches and 10^{-3} for the 3D trajectory and geometry branches. Eight RTX 4090 GPUs each process one clip (global batch size 8), using BF16 mixed precision and per-parameter-group gradient-norm clipping at 1. Overlapping truncated backpropagation uses 6-observation windows with stride 3.

Each query has $N_s = 8$ supports. We use $M = 5$ interval-path samples and $B = 5$ event bins. The 2D and 3D position losses have unit weights, with $\alpha_{\mathrm{HS}} = 0.5$ and $\alpha_{\mathrm{geo}} = 0.25$.

C Detailed Evaluation Protocol

C.1 Point Trajectories and Dataset Splits

EventTraj3D derives from BlinkVision's event and geometric resources (Li et al., 2024b), with statistics and a leakage-free split in Table C1. The clips span indoor and outdoor scenes. Instance-associated queries provide aligned 2D/3D tracks from native particle annotations. We retain event-observable segments with valid geometric labels, expressed in the initial left-camera frame.

Table C1: EventTraj3D statistics and train/validation/test splits.

Statistic	**Count or definition**
Asynchronous events	13.630 billion
Annotated point trajectories	6,494
Total clips	1,427
Training clips	999
Validation / test clips	214 / 214
Challenging clips (union)	173 / 1,427 (12.12%)
3D storage frame	Left camera at the initial tracking time

In Figure 4, category sectors have equal sizes and radial bars show event counts per clip.

Challenging cases. The union comprises 173 of all 1,427 clips (12.12%): repetitive appearance ambiguity, intermittent event support, or non-rigid deformation, including articulated motion. Clips satisfying multiple conditions are counted once in this full-dataset union.

Table C2: **Comparison of point-tracking datasets.** A checkmark indicates directly provided data or annotations, while "–" denotes a field not provided in the compared tracking format. Motion descriptions refer to target motion rather than camera motion. Dataset sources are Doersch et al. (2022), Zheng et al. (2023), Koppula et al. (2024), Hamann et al. (2025), and Li et al. (2024a).

Dataset	Events	Metric 3D tracks	Scene coverage	Target motion
TAP-Vid	–	–	Indoor / outdoor	Rigid + non-rigid (incl. articulated)
PointOdyssey	–	✓	Indoor / outdoor	Rigid + non-rigid (incl. articulated)
TAPVid-3D	–	✓	Indoor / outdoor	Rigid objects + articulated humans
EventKubric	✓	–	Synthetic scenes	Rigid objects
E-3DTrack	✓	✓	Indoor capture	Rigid-object motion
EventTraj3D (Ours)	✓	✓	**Indoor / outdoor**	**Rigid + non-rigid (incl. articulated)**

EventTraj3D combines event observations with metric 3D point trajectories across indoor and outdoor scenes, while covering rigid, articulated, and non-rigid target motion. Unlike EventKubric, it provides metric 3D trajectories, while its target-motion coverage also differs from the rigid-object annotation model used by E-3DTrack. Table C2 characterizes dataset coverage rather than ranking dataset difficulty.

C.2 Geometric Reference and Evaluation Metrics

Each clip uses its initial camera frame as a fixed reference. Homogeneous position $\widetilde{\mathbf{X}}^0_{i,t}$ projects into the current image as

$$\mathbf{u}_{i,t} = \mathrm{proj}\Big(K\mathbf{T}_{t\leftarrow 0}\widetilde{\mathbf{X}}^0_{i,t}\Big). \tag{C1}$$

$\mathbf{T}_{t\leftarrow 0}$ is the 3×4 initial-to-current camera transform supplied by the dataset, and $\mathrm{proj}([x,y,z]^\top) = [x/z, y/z]^\top$. Superscript c below denotes positions after this transform. Query trajectories receive position supervision. Auxiliary supports establish local correspondence from their reference patches.

Over valid point–time set $\mathcal{V}_2$, 2D endpoint error is

$$\mathrm{EPE}_{2D} = \frac{1}{|\mathcal{V}_2|}\sum_{(i,t)\in\mathcal{V}_2} \|\widehat{\mathbf{u}}_{i,t} - \mathbf{u}^{\mathrm{gt}}_{i,t}\|_2. \tag{C2}$$

Absolute 3D error uses the corresponding current-camera positions over $\mathcal{V}_3$:

$$\mathrm{EPE}_{3D} = \frac{1}{|\mathcal{V}_3|}\sum_{(i,t)\in\mathcal{V}_3} \|\widehat{\mathbf{X}}^c_{i,t} - \mathbf{X}^{c,\mathrm{gt}}_{i,t}\|_2. \tag{C3}$$

Both metrics average Euclidean distances over valid point–time pairs. A common rigid transform preserves 3D EPE. Evaluation uses original-resolution pixels and millimeters. Dataset poses align the supervision and evaluation coordinates. Subsequent tracking inputs are monocular events.

3D ADE and FDE measure anchor-relative displacement. The predicted anchor $\mathbf{a}_i$ and ground-truth initial point $\mathbf{X}^{0,\mathrm{gt}}_{i,0}$ are separately transformed into the current camera frame and subtracted from their respective trajectories. ADE first averages displacement errors over valid times within a trajectory, then equally across trajectories. FDE uses each trajectory's last valid scored time. Two-dimensional FDE also uses the last valid time. The 2D set excludes the initial query. Scoring in 3D starts at observation five after the four-observation initialization.

C.3 Comparison Methods

DeepEvT, ETAP, and TAPFormer supply 2D tracks. Pairing each with E2Depth or DepthAnyEvent-R gives six 3D pipelines. Results use all 214 EventTraj3D test clips (Tables 1–2) and the E-3DTrack evaluation (Table 3). All methods share the initialization model and protocol on both datasets.

E-3DTrack provides stereo events and 250 Hz ground-truth 3D feature trajectories (Li et al., 2024a). For the comparison in Table 3, all methods use monocular events and are retrained on E-3DTrack. For TAPFormer, the post-initialization image branch receives images formed by accumulating events over the corresponding time windows, while the event branch receives the corresponding event representation.

The ablations in Table 4 share initialization, dataset splits, and training protocol. Direct XYZ retains the encoders, GRU, and input features, with comparable position-loss normalization, validity masking, and gradient scaling. Output units and bounds are adapted to direct coordinates. The comparison evaluates the complete output parameterization.

The six comparison pipelines combine 2D tracks from DeepEvT, ETAP, or TAPFormer (Messikommer et al., 2023; Hamann et al., 2025; Liu et al., 2026) with depth from the recurrent E2Depth or DepthAnyEvent-R models (Hidalgo-Carrió et al., 2020; Bartolomei et al., 2025). Thus they predict depth along tracked pixels, whereas our trajectory head consumes correspondence and local-motion features directly. Parameter counts in Table 2 exclude the initialization module.

D Hit–Survival Analysis Details

D.1 Cases and Measurements

Three selected history controls fix weights, candidates, and base scores while varying interval evidence (Figure 5(i–iii) and Table D1). The pair in Figure 5(a–b) is a separate example.

P_2 is the probability mass of valid candidates within 2 px of ground truth. E_{post} is the posterior-weighted expected localization error over all candidates. E_{mode} is the localization error of the highest-probability candidate. These statistics describe the local candidate distribution.

Table D1: Candidate support and error in Figure 5. Shuffled rows report means over 64 permutations. Within each case, best values are **bold** and second-best values are underlined. Ties share a rank.

Case	Setting	P_2 (%)↑	E_{post} (px)↓	E_{mode} (px)↓
(i)	No interval	31.055	8.943	**14.264**
(i)	Endpoint only	32.652	8.843	**14.264**
(i)	**True history**	**36.022**	**8.457**	**14.264**
(i)	Shuffled mean	31.849	8.870	**14.264**
(ii)	No interval	26.794	13.192	**9.254**
(ii)	Endpoint only	28.007	13.111	**9.254**
(ii)	**True history**	**30.743**	**12.796**	**9.254**
(ii)	Shuffled mean	27.727	13.093	**9.254**
(iii)	No interval	24.193	12.784	14.519
(iii)	Endpoint only	24.193	12.784	14.519
(iii)	**True history**	**30.087**	**12.404**	**1.950**
(iii)	Shuffled mean	28.435	12.511	4.306

D.2 Candidate Responses and Association Controls

Candidates A and B have base scores of -4.684 and -4.626 and localization errors of 1.76 and 6.18 px, respectively (Figure 5(a–b)). The plotted response is $\Delta_{krj}(\tau)/\Delta_{krj}(\tau_k)$, with candidate indices suppressed in the figure. Each curve is normalized by its own endpoint response. Both end at one, but only B exceeds one within the interval. Thus, endpoint agreement can mask incompatible preceding evolution, motivating a candidate-specific pre-trigger history criterion.

In the separate control case (iii), the mode error falls from 14.519 to 1.950 px (Table D1), showing that interval history can change the selected correspondence. In cases (i–ii), the mode is unchanged but posterior-weighted error decreases. The benefit therefore includes confidence redistribution toward spatially better candidates, rather than requiring an argmax change at every update.

The four settings are no interval, endpoint only, true history, and shuffled history. No interval retains base matching and the Hit term. Endpoint only sets the pre-trigger term in the interval score to zero. Shuffling re-associates predecessor and trigger times within polarity groups while preserving spatial locations, polarities, event counts, and the multiset of interval durations. Permutations retain the original valid-candidate mask, and each uses recomputed reliability conditions.

Each case contains 64 accepted permutations. Dots show all outcomes, diamonds their means, and whiskers the empirical 2.5th–97.5th percentiles. True history has higher P_2 and lower posterior-weighted error than the shuffled mean in all three cases. Because shuffling preserves event counts, spatial locations, polarities, and interval-duration statistics, this comparison links the improvement to temporal association rather than simply more event support. True-history P_2 in case (i) exceeds the plotted percentile range, while case (iii) remains within it. The controls therefore support a benefit in these selected cases, not uniform separation from every shuffled realization.

Together, the fixed-candidate controls in Table D1 connect pre-trigger history to both correspondence confidence and, when the ranking changes, localization. This provides a mechanism-level explanation for the tracking degradation observed when interval evidence is removed.

E Ablation Visualization

Direct XYZ produces the largest increases in absolute and final-displacement errors (Figure E1). Removing Hit–Survival increases 3D EPE and FDE. Removing local geometry causes larger increases in both 3D metrics despite a smaller 2D degradation. Thus, a change in query localization alone does not fully describe the information needed for 3D inference. The distinct error patterns support interval evidence for correspondence, local motion for spatial inference, and the complete normalized, anchor-relative parameterization for trajectory prediction.

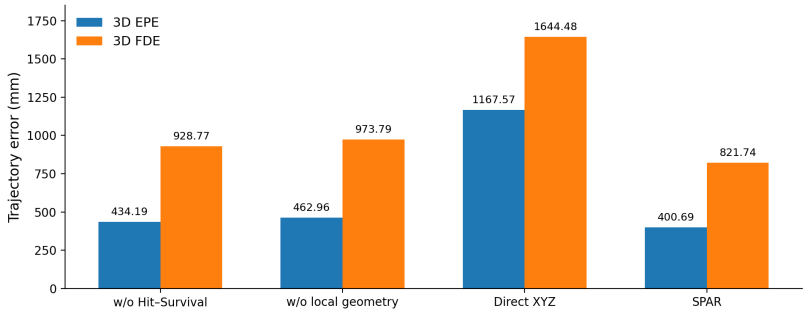

Figure E1: **Component ablation.** Values reproduce Table 4. EPE measures absolute 3D position error, and FDE measures final anchor-relative displacement error. Lower is better.

F Spatial State Parameterization and Initialization Robustness

F.1 Spatial State Analysis

Projection and parameterization. A pixel defines the viewing ray $\mathcal{R}(\mathbf{u}) = \{zK^{-1}\tilde{\mathbf{u}} \mid z > 0\}$, and positive rescaling preserves its projection: $\mathrm{proj}(K(\nu\mathbf{X})) = \mathrm{proj}(K\mathbf{X})$ for $\nu > 0$. Equation (7) equivalently gives $\boldsymbol{\delta}_{i,t} = (\widehat{\mathbf{X}}^0_{i,t} - \mathbf{a}_i)/Z_{i,0}$. The initial depth sets the metric unit of the relative state.

Longitudinal ambiguity and error propagation. For a current-camera point $\mathbf{X}^c = (X^c, Y^c, Z^c)^\top$ with $Z^c > 0$ and zero-skew pinhole intrinsics, the projection Jacobian is

$$\mathbf{J}_{\mathrm{proj}}(\mathbf{X}^c) = \begin{bmatrix} f_x/Z^c & 0 & -f_x X^c/(Z^c)^2 \\ 0 & f_y/Z^c & -f_y Y^c/(Z^c)^2 \end{bmatrix}, \qquad \mathbf{J}_{\mathrm{proj}}(\mathbf{X}^c)\mathbf{X}^c = \mathbf{0}. \tag{F1}$$

The viewing-ray null direction differs from a fixed camera axis for off-axis points. For $\mathbf{X}^c = Z^c K^{-1}\widetilde{\mathbf{u}}$, the first-order lifting error is

$$\mathrm{d}\mathbf{X}^c = K^{-1}\widetilde{\mathbf{u}}\,\mathrm{d}Z^c + Z^c K^{-1}[\mathrm{d}u, \mathrm{d}v, 0]^\top. \tag{F2}$$

Depth error survives perfect image localization, while pixel errors have larger metric effects at greater depth. Our trajectory head combines correspondence confidence, local motion, and temporal history to estimate longitudinal evolution. The 3D metrics measure full spatial error, and Appendix F.2 evaluates sensitivity to depth-prior perturbations.

Remaining anchor error and representational correction. Substitution of the motion-plus-anchor-bias identity in Equation (8) into Equation (7) gives

$$\widehat{\mathbf{X}}^0_{i,t} - \mathbf{X}^{0,\mathrm{gt}}_{i,t} = Z_{i,0}\big(\boldsymbol{\delta}_{i,t} - \boldsymbol{\delta}^{\mathrm{req}}_{i,t}\big). \tag{F3}$$

An offset can represent bias compensation together with motion. Equation (B2) still penalizes absolute bias because both positions share the same positive normalization.

For the depth component in the initial camera frame, $\delta^{z,\mathrm{req}}_{i,t} = \frac{Z^{0,\mathrm{gt}}_{i,t}}{Z_{i,0}} - 1, \qquad \widehat{Z}^0_{i,t} - Z^{0,\mathrm{gt}}_{i,t} = Z_{i,0}\big(\delta^z_{i,t} - \delta^{z,\mathrm{req}}_{i,t}\big)$. Here $Z^{0,\mathrm{gt}}_{i,t}$ is the third position component. The decoder can represent this compensation when $\|\boldsymbol{\delta}^{\mathrm{req}}_{i,t}\|_\infty < \beta$. Its estimation is guided by correspondence and geometric evidence. The initial anchor remains the reference for all subsequent states.

Absolute-position and anchor-relative displacement errors must also be distinguished. At each scored time,

$$(\widehat{\mathbf{X}}^0_{i,t} - \mathbf{a}_i) - (\mathbf{X}^{0,\mathrm{gt}}_{i,t} - \mathbf{X}^{0,\mathrm{gt}}_{i,0}) = (\widehat{\mathbf{X}}^0_{i,t} - \mathbf{X}^{0,\mathrm{gt}}_{i,t}) - (\mathbf{a}_i - \mathbf{X}^{0,\mathrm{gt}}_{i,0}). \tag{F4}$$

This identity relates absolute-position error, displacement error, and anchor bias under a common rigid transform. Appendix F.2 evaluates the resulting initialization sensitivity.

F.2 Robustness to Initial Depth Errors

Perturbation setting. With fixed model weights, we perturb each query's Metric3D v2 prior once at initialization, preserving its original estimation error. The perturbed prior is

$$d^{P\prime}_i = d^P_i \exp\left(\sigma_{\log}\varepsilon^{\mathrm{init}}_i - \frac{\sigma^2_{\log}}{2}\right), \qquad \varepsilon^{\mathrm{init}}_i \sim \mathcal{N}(0,1), \quad \sigma_{\log} = \sqrt{\log(1+\sigma^2_{\mathrm{init}})}. \tag{F5}$$

We use $\sigma_{\mathrm{init}} \in \{0.05, 0.10, 0.20, 0.30\}$ and baseline zero. The multiplier has mean one and standard deviation σ_{init}, which specifies the relative noise strength. The perturbation is sampled once at initialization and retained throughout the sequence.

Paired evaluation. We compare our method with a pretrained direct 3D prediction head on Event-Traj3D test sequences. The two models receive the same perturbations for seeds 20260925, 20260926, and 20260927. For each model and seed, we subtract its own unperturbed absolute 3D EPE, $\Delta\mathrm{EPE}(\sigma_{\mathrm{init}}) = \mathrm{EPE}(\sigma_{\mathrm{init}}) - \mathrm{EPE}(0)$. Table F1 reports the mean of these increases across the three seeds. The relative reduction is $\mathrm{Reduction} = 100\left(1 - \frac{\overline{\Delta\mathrm{EPE}}_{\mathrm{SPAR}}}{\overline{\Delta\mathrm{EPE}}_{\mathrm{direct}}}\right)$. The reported percentage quantifies the reduction in initialization-induced added error.

Table F1: Sensitivity to initial depth perturbations. Added absolute 3D EPE is measured from each model's own unperturbed baseline and averaged over three paired seeds. At each noise level, best and second-best added errors are **bold** and underlined.

Noise level	SPAR ΔEPE (mm)↓	Direct head ΔEPE (mm)↓	Reduction (%)↑
5%	**131.56**	184.93	28.86
10%	**272.72**	377.77	27.81
20%	**578.14**	796.54	27.42
30%	**904.26**	1207.96	25.14

Evidence of robustness. Across the tested noise levels, our method incurs less mean added error, with a smaller increase for every paired seed at every level. This supports reduced propagation of depth-prior perturbations through the combined initialization-and-tracking pipeline, consistent with a relative state that can represent motion and residual anchor compensation. Added error still grows with perturbation strength, so the metric anchor remains consequential. This establishes reduced initialization sensitivity for the combined pipeline, not independence from the depth prior.

G Additional Qualitative Results

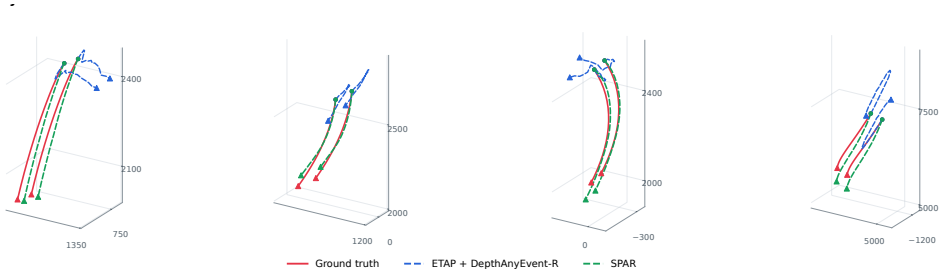

Figure G1: **Additional qualitative results on EventTraj3D.** Ground truth is shown in red, ETAP + DepthAnyEvent-R in blue, and our method in green.

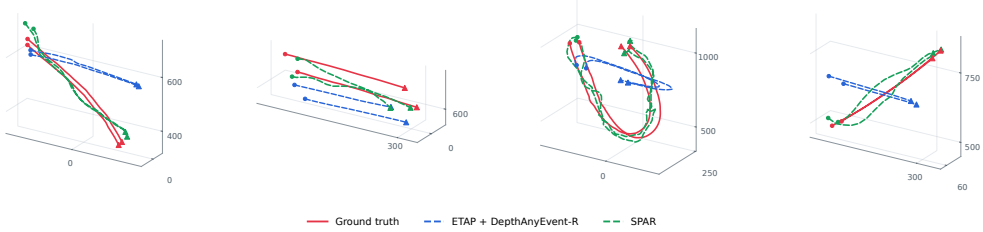

Figure G2: **Additional qualitative results on E-3DTrack.** Ground truth is shown in red, ETAP + DepthAnyEvent-R in blue, and our method in green.

The additional examples show that our predictions more closely follow the ground-truth trajectory shape and spatial extent across both datasets.

H AI Use Statement

AI tools were employed to polish the manuscript text, fix grammatical issues and enhance readability. Nevertheless, all research ideas, technical approaches, experimental schemes, data analysis and

conclusions of this work originate solely from the authors' independent research efforts. The authors have thoroughly checked and revised the whole manuscript and take full responsibility for the validity, integrity and authenticity of this work.