TEMPOTRACK: TEMPORAL DISTRIBUTION MATCHING FOR OPEN-VOCABULARY MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary multi-object tracking requires identity association for categories without task-specific identity supervision. Existing tracking methods retain appearance histories, yet prototypes or pooled similarities compress this history and can obscure how past observations support the current query. We propose **TempoTrack**, which represents each query–identity pair through the empirical distribution of its historical matching responses. Recent and long-term distributional evidence is combined by a candidate-specific temporal gate and candidate-relative scorer, while a moment-generating formulation provides a compact second-order parameterization without explicit covariance construction. TempoTrack introduces a lightweight, frame-causal association layer on top of frozen visual frontends and is trained using only Base identity supervision. On TAO, TempoTrack reaches 43.3 Novel AssocA on validation and 33.6 on test. These results support query-conditioned temporal distributions as a lightweight and effective representation of historical identity evidence for open-vocabulary association.

1 INTRODUCTION

Open-vocabulary multi-object tracking (OVMOT) extends video perception beyond the categories with task-specific supervision. In addition to localizing and recognizing objects, a tracker must determine whether observations at different times correspond to the same physical instance (Li et al., 2023; 2025a). Category recognition and identity association therefore require different evidence: two observations can share a category without sharing an identity. Pose changes, occlusion, viewpoint variation, and scale changes make this distinction particularly important when identity supervision is unavailable for the evaluated categories (Qin et al., 2024; Aydemir et al., 2025). We study this setting by learning association from Base identities while keeping the detector and visual representation frozen. Here, unseen categories are defined by the task-specific supervision protocol, separately from the knowledge acquired during visual pretraining.

Existing OVMOT methods provide increasingly rich evidence for association through open-vocabulary appearance matching (Li et al., 2023), transferable instance representations (Li et al., 2024b), temporal representation learning (Qian et al., 2025a), and multi-cue or trajectory-level reasoning (Qian et al., 2025b; Li et al., 2025c). Our question concerns how the available history supports a particular identity decision. A prototype provides a compact appearance reference, while a pooled similarity summarizes agreement with historical observations. Neither summary alone distinguishes every pattern of support: two identities can have similar average responses but different dispersion or disagreement between recent and longer histories. TempoTrack studies this distinction by organizing query–history responses into temporal matching distributions and learning how their statistics should determine the final association score. This focus complements improvements to visual representations and candidate retrieval.

We first distinguish candidate retrieval from correct identity selection. We analyze three representative OVMOT methods and observe a consistent pattern: the correct historical identity is often retained in the candidate pool, yet an incorrect identity receives a higher score. Figure 1 illustrates this ambiguity. This leads to our central question: how can historical support be organized to construct a

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.

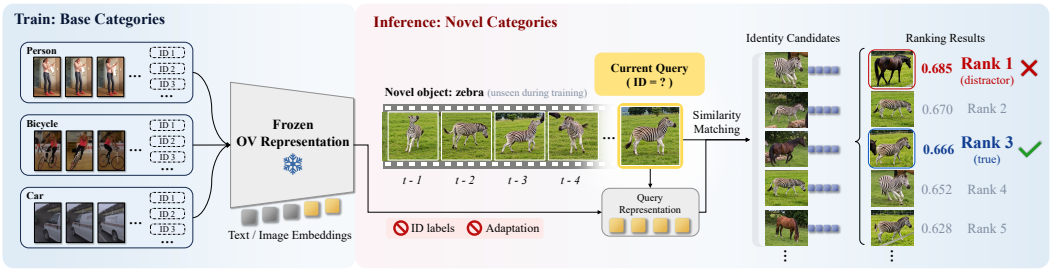


Figure 1: **Identity ranking ambiguity in open-vocabulary multi-object tracking.** A frozen open-vocabulary frontend retrieves multiple historical identity candidates for a Novel-category query, but the correct zebra identity is ranked below a visually similar distractor.

more discriminative identity score for the current query? We address this question through query-conditioned temporal matching distributions.

We propose **TempoTrack** to construct identity scores from query-conditioned temporal matching distributions. For each candidate, the current observation is compared with retained historical features before the resulting responses are summarized. Their mean and dispersion characterize historical support over recent and long-term ranges. A candidate-specific gate combines these ranges, and a structured score with a residual branch incorporates competition among identities. The residual learns candidate-conditioned interactions around the structured distributional score. The model learns with Base identity supervision and updates its predicted history after each frame-level assignment. To summarize, our main contributions are listed as below:

- We formulate query-conditioned temporal matching distributions as identity evidence and connect their empirical response moments to projected historical feature statistics through a tractable moment-generating representation.
- We develop an adaptive distributional association mechanism that combines complementary temporal ranges, explicit support–dispersion scoring, and candidate-relative interactions within a lightweight causal association layer.
- We evaluate TempoTrack on multiple public tracking benchmarks, achieving consistent gains in association performance and demonstrating its broad applicability across different tracking settings.

2 Diagnosing Association in Open-Vocabulary Tracking

2.1 Problem Setting

At frame t, a frozen frontend produces current observations $X_t = \{\mathbf{x}_t^n\}_{n=1}^{N_t}$ and maintains identities $\mathcal{T}_{t-1}$ established in previous frames. For identity i, its feature history is $H_i^{t-1} = \{\mathbf{h}_i^1, \ldots, \mathbf{h}_i^{L_i}\}$. Association determines which existing identity matches each observation, or passes the observation to new-track handling. In our analysis we examine this decision using the baseline association scores. Ground truth is used only after prediction to label evaluation events. It does not affect candidate generation or scoring because both rely solely on the predicted history.

2.2 Candidate Availability and Identity Ranking

An association error can arise because the correct identity is absent from the candidate set or because it is present but an incorrect identity receives a higher score. We distinguish these outcomes using the original frontend predictions, independently of TempoTrack scoring. Figure 2 separates correct rank-1 selection, correct identities retained at ranks 2–64, and identities outside the diagnostic candidate budget for each frontend and category group. Each current observation is matched to same-category ground truth at IoU ≥ 0.5, and predictions strictly before the current frame establish its historical identity correspondence. Ground truth labels the evaluation events after prediction and is not an input to candidate generation or scoring. The diagnostic population consists of matched observations with an identifiable correct prior history, including events in which that identity

is missing from the retained candidate set. Recall@K is the fraction of this population whose correct identity occurs among the first K candidates; absent identities count as retrieval failures.

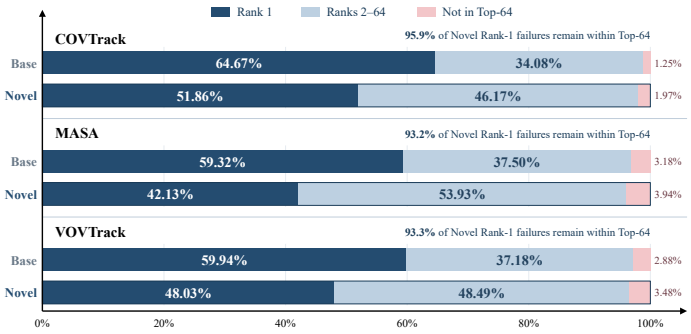


Figure 2: **Candidate-rank decomposition across three frozen OVMOT frontends.** For association events with an identifiable prior identity, each bar partitions the correct identity into Rank 1, Ranks 2–64, or outside the Top-64.

We analyze COVTrack, MASA, and VOVTrack, which adopt different open-vocabulary association designs and provide public implementations. The three frontends exhibit a consistent separation between candidate coverage and rank-1 selection. Novel Recall@1 ranges from 42.13% to 51.86%, whereas Recall@64 reaches 96.06%–98.03%; the Base–Novel gap decreases from 11.91–17.19 percentage points at rank 1 to at most 0.76 points at rank 64. These statistics describe the original frontend rankings on the diagnosed event population. TempoTrack uses a separate inference prefilter based on the last actual observation, retaining at most 64 identities for context and 8 for assignment. Therefore, the diagnostic Recall@64 describes broad historical candidate availability, whereas actual assignment eligibility is determined by that inference prefilter. The two quantities concern different candidate orders. Complete diagnostic curves and population counts are provided in Appendix D.

2.3 Temporal Evidence across Appearance Changes

We examine how same-identity visual evidence changes with temporal gap using frozen features. Similarity decreases for both Base and Novel categories. Across the evaluated bins, the Novel–Base differences range from -0.077 to $+0.017$ on validation and from -0.059 to $+0.032$ on test, without a progressively negative trend. The video-cluster intervals in Appendix D characterize uncertainty in these comparisons. This pattern motivates an instance-conditioned treatment of history: the relevance of retained observations depends on the current appearance and temporal context, rather than a fixed preference determined by the Base/Novel category group. The analysis concerns representation behavior in the sampled identities, separately from the end-to-end association comparison.

2.4 Design Implication

The design question is how to distinguish candidate histories with similar average support but different response patterns. For dot-product matching, averaging the responses is algebraically equivalent to matching an unnormalized mean feature: $L^{-1}\sum_k \mathbf{x}^\top \mathbf{h}_k = \mathbf{x}^\top (L^{-1}\sum_k \mathbf{h}_k)$. TempoTrack therefore builds its evidence from response dispersion, complementary temporal ranges, and candidate-relative interactions in addition to the mean. These components specify testable sources of improvement over prototype and pooled-similarity scoring.

3 TempoTrack: Temporal Distribution Matching

3.1 Overview and Temporal Identity History

TempoTrack constructs an identity score from the current query and the predicted history available before the frame-level assignment (Figure 3). Each identity maintains a bounded, temporally ordered bank of individually normalized historical features with cosine-based redundancy suppression. The recent history H_i^r comprises the most recent portion of the retained observations, whereas the long-term history H_i^l contains the complete retained bank, so $H_i^r \subseteq H_i^l$. These ranges capture recent appearance agreement and historical support across a broader set of observed states. The corresponding prototypes, $\mathbf{p}_{i,t-1}^r$ and $\mathbf{p}_{i,t-1}^l$, use distinct fixed update rates to provide complementary appearance references at different temporal scales. These prototype updates are distinct from the

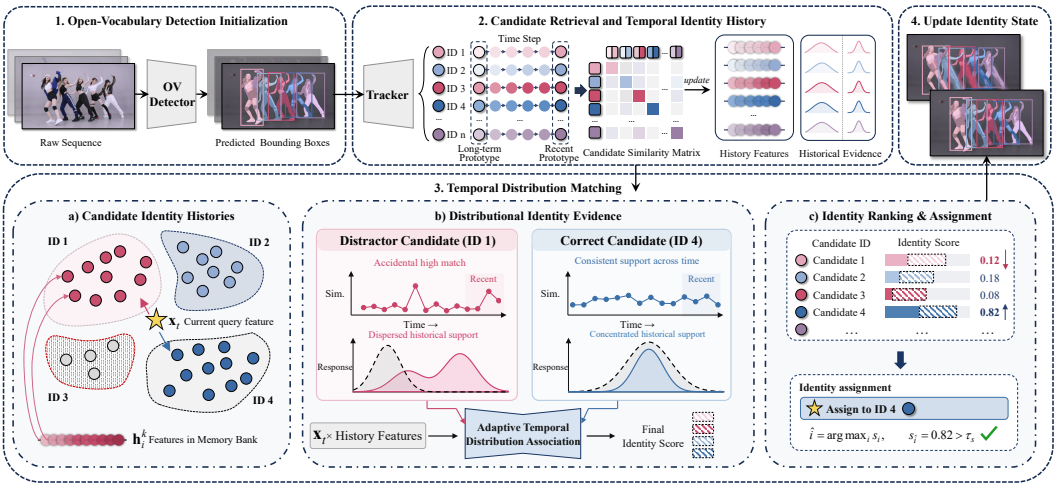


Figure 3: **Overview of TempoTrack.** A frozen open-vocabulary detector provides current observations, while each identity maintains recent and long-term history. TempoTrack constructs query-conditioned matching distributions from query–history responses, converts them into identity scores for ranking and assignment, and updates the identity history only after the association decision.

learned temporal gate that combines the distributional evidence. Candidate descriptors are computed from the pre-assignment state, and accepted observations update that state after the frame-level decisions are committed. The history capacity, temporal-range configuration, redundancy suppression, and prototype update rules are specified in Appendix A.

3.2 Query-Conditioned Temporal Matching Distributions

For a candidate identity i and temporal range $\rho \in \{r, l\}$, let $H_i^\rho = \{\mathbf{h}_i^k : k \in \mathcal{I}_i^\rho\}$ denote the corresponding retained history, where $\mathcal{I}_i^\rho$ is its index set and $L_i^\rho = |\mathcal{I}_i^\rho|$. Given the current query feature $\mathbf{x}_t$, we compare it with every retained historical feature in H_i^ρ. Since both query and history features are normalized, each pair produces a cosine matching response $d_i^k = \mathbf{x}_t^\top \mathbf{h}_i^k \in [-1, 1]$. Instead of collapsing these responses into a single prototype similarity or pooled score, we retain the complete set of query-to-history responses, which characterizes how the current query is supported by the different historical appearances of candidate i. We therefore define the *temporal matching distribution* as the empirical probability measure

$$\mathcal{D}_i^\rho(\cdot \mid \mathbf{x}_t) = \frac{1}{L_i^\rho} \sum_{k \in \mathcal{I}_i^\rho} \delta_{d_i^k}, \qquad \rho \in \{r, l\}. \tag{1}$$

Here δ_a denotes a unit point mass at response value a. Importantly, $\mathcal{D}_i^\rho$ is a distribution over query–history matching responses rather than an observation-independent distribution of historical features. The same identity history can therefore induce different matching distributions for different current queries. In this way, the distribution preserves the pattern of support across historical appearances before it is summarized by low-order statistics.

3.3 Moment-Generating Distributional Evidence

The distribution $\mathcal{D}_i^\rho(\cdot \mid \mathbf{x}_t)$ preserves all query–history matching responses within a temporal range, but the association model requires a compact representation of this evidence. We first characterize its first- and second-order statistics. Let $\bar{\mathbf{h}}_i^\rho$ and $\mathbf{\Sigma}_i^\rho$ denote the empirical mean and covariance of the

retained features in H_i^ρ. Since $d_i^k = \mathbf{x}_t^\top \mathbf{h}_i^k$, the mean and variance of the matching responses satisfy

$$\mu_i^\rho = \frac{1}{L_i^\rho} \sum_{k \in \mathcal{I}_i^\rho} d_i^k = \mathbf{x}_t^\top \bar{\mathbf{h}}_i^\rho,$$
$$v_i^\rho = \frac{1}{L_i^\rho} \sum_{k \in \mathcal{I}_i^\rho} (d_i^k - \mu_i^\rho)^2 = \mathbf{x}_t^\top \mathbf{\Sigma}_i^\rho \mathbf{x}_t. \tag{2}$$

Thus, μ_i^ρ measures the average historical support for the current query, while v_i^ρ describes how that support varies across the retained appearances of the same identity.

Rather than treating these statistics independently, we connect them through the exponential aggregation of the matching responses. For the empirical distribution in Equation 1, we define its *moment-generating function (MGF)* as

$$M_i^\rho(\beta \mid \mathbf{x}_t) = \mathbb{E}_{u \sim \mathcal{D}_i^\rho(\cdot \mid \mathbf{x}_t)}[e^{\beta u}] = \frac{1}{L_i^\rho} \sum_{k \in \mathcal{I}_i^\rho} e^{\beta d_i^k}. \tag{3}$$

This form aggregates the exponentiated pairwise responses over the identity history while assigning unit total mass to each retained distribution. Moreover, the first two derivatives of $\log M_i^\rho(\beta \mid \mathbf{x}_t)$ at $\beta = 0$ recover μ_i^ρ and v_i^ρ, respectively, showing that the log-MGF provides a common representation of the first two response moments.

To obtain a closed-form second-order descriptor, we introduce a moment-matched Gaussian (Li et al., 2021) surrogate for the retained feature statistics,

$$\widetilde{\mathbf{H}}_i^\rho \sim \mathcal{N}\left(\bar{\mathbf{h}}_i^\rho, \mathbf{\Sigma}_i^\rho\right). \tag{4}$$

This surrogate matches the empirical first and second moments of H_i^ρ; it is used only as an analytic device and does not assume that the retained tracking features themselves follow an exact Gaussian distribution. For the fixed query $\mathbf{x}_t$, linear projection gives

$$\widetilde{D}_i^\rho = \mathbf{x}_t^\top \widetilde{\mathbf{H}}_i^\rho \sim \mathcal{N}\left(\mathbf{x}_t^\top \bar{\mathbf{h}}_i^\rho, \mathbf{x}_t^\top \mathbf{\Sigma}_i^\rho \mathbf{x}_t\right) = \mathcal{N}(\mu_i^\rho, v_i^\rho). \tag{5}$$

Applying the Gaussian MGF identity then yields

$$\widetilde{M}_i^\rho(\beta \mid \mathbf{x}_t) = \exp\left(\beta \mu_i^\rho + \frac{\beta^2}{2} v_i^\rho\right) = \exp\left(\beta \mathbf{x}_t^\top \bar{\mathbf{h}}_i^\rho + \frac{\beta^2}{2} \mathbf{x}_t^\top \mathbf{\Sigma}_i^\rho \mathbf{x}_t\right). \tag{6}$$

Evaluating the corresponding log-MGF at $\beta = 1$ gives

$$m_i^\rho := \log \widetilde{M}_i^\rho(1 \mid \mathbf{x}_t) = \mu_i^\rho + \frac{1}{2} v_i^\rho = \mathbf{x}_t^\top \bar{\mathbf{h}}_i^\rho + \frac{1}{2} \mathbf{x}_t^\top \mathbf{\Sigma}_i^\rho \mathbf{x}_t. \tag{7}$$

The association model receives the empirical mean and variance together with their fixed log-MGF combination $m_i^\rho = \mu_i^\rho + \frac{1}{2} v_i^\rho$. This gives a second-order parameterization of query-conditioned historical evidence; m_i^ρ is a deterministic coordinate rather than an independent statistic. Recent and long-term descriptors are combined with prototype agreement and candidate context to construct the final identity score. All moments are evaluated from the scalar response vectors, avoiding Gaussian sampling and explicit feature-dimensional covariance matrices. Appendix B gives the projection and MGF derivations, and Table 7 evaluates the contribution of the additional parameterization.

3.4 Adaptive Temporal Distribution Association

The preceding section provides a distributional description of each candidate identity through the projected statistics $(\mu_i^\rho, v_i^\rho, m_i^\rho)$ for $\rho \in \{r, l\}$. These quantities characterize how the current query is supported by the retained history, but they do not by themselves define the final identity ranking. In tracking, temporal evidence must also account for direct appearance agreement, the consistency of historical support, and the relative reliability of recent and long-term observations. We therefore use the distributional evidence to construct a structured association score. For each temporal range ρ, let $q_i^\rho = \mathbf{x}_t^\top \mathbf{p}_{i,t-1}^\rho$ denote the query–prototype similarity. We combine it with the mean historical response and penalize response dispersion:

$$b_i^\rho = \frac{1}{2}(q_i^\rho + \mu_i^\rho) - \lambda_v v_i^\rho, \qquad \lambda_v = \mathrm{softplus}(\eta_v). \tag{8}$$

The branch score combines prototype agreement with average historical support and uses a learned positive coefficient to express a preference for concentrated responses when these agreement cues are comparable. Response dispersion can also reflect valid appearance changes, so it is interpreted jointly with the other candidate evidence rather than as an identity probability. The recent and long-term branch scores are combined through a candidate-specific gate, and the residual branch learns interactions with candidate-relative context. We construct the final score as follows:

$$\begin{aligned} \alpha_i &= G_\phi(\mathbf{g}_i), \qquad s_i^{\mathrm{str}} = \alpha_i b_i^r + (1-\alpha_i) b_i^l, \\ s_i &= \mathrm{softplus}(\gamma) s_i^{\mathrm{str}} + f_\theta\left([\mathrm{Norm}(\mathbf{z}_i), \alpha_i, s_i^{\mathrm{str}}]\right). \end{aligned} \tag{9}$$

The gate receives the raw temporal-state cues: query–prototype agreement, mutual prototype agreement, the recent and long-term response moments, normalized temporal gap, and history length. It determines how strongly each temporal range contributes to the current candidate. Let $\mathbf{z}_i \in \mathbb{R}^{33}$ denote the raw candidate descriptor. The residual receives its featurewise-standardized form $\mathrm{Norm}(\mathbf{z}_i)$ together with the gate output and structured score, allowing historical support to be interpreted relative to competing identities. All candidates share the same scoring parameters, and the scorer does not require a category-specific output head. This design permits newly encountered identities to be evaluated through their available histories without introducing new output parameters. For training, $s_{e,j}$ denotes this same final score for candidate j in association event e.

3.5 Association Learning and Identity Information

We train the final scores in Equation 9 directly for candidate ranking. For association event e, let $s_{e,j}$ denote the final score of candidate j. The objective encourages candidates consistent with the correct identity to receive higher scores than competing identities. Let $\mathcal{P}_e$ contain the reliably supervised candidate rows corresponding to the correct Base identity, let $\mathcal{N}_e$ contain the valid supervised candidates associated with competing identities, and define $\mathcal{V}_e = \mathcal{P}_e \cup \mathcal{N}_e$. We optimize

$$\begin{aligned} \mathcal{L}_e = & \underbrace{-\log \frac{\sum_{j\in\mathcal{P}_e} \exp(s_{e,j})}{\sum_{j\in\mathcal{V}_e} \exp(s_{e,j})}}_{\mathcal{L}_{\mathrm{list},e}} \\ & + \lambda_h \left[\Delta + \max_{j\in\mathcal{N}_e} s_{e,j} - \max_{j\in\mathcal{P}_e} s_{e,j}\right]_+ . \end{aligned} \tag{10}$$

Equation 10 jointly optimizes the gate parameters ϕ, residual parameters θ, variance coefficient η_v, and structured-score scale γ. The visual features and prototype update rates remain fixed. Each supervised event contains at least one positive and one negative Base candidate. The listwise term allocates normalized score mass to the correct identity, including multiple positive histories when present, while the hinge term enforces a margin between the strongest positive and negative candidates. We use $\lambda_h = 0.2$ and $\Delta = 0.2$. This supervision learns how distributional and candidate-context evidence determines identity rankings; it does not update the frozen visual representation.

The listwise objective connects temporal distribution scores to conditional identity information. Let X denote the current query, Y its correct identity, and C the candidate context, including the supervised identity partition $\{\mathcal{G}_y(C)\}_y$ of the valid candidate set $\mathcal{V}(C)$. Identity-group labels define supervision and are not inputs to the scorer. The scores induce the normalized identity posterior

$$q_\omega(y \mid X, C) = \frac{\sum_{j\in\mathcal{G}_y(C)} \exp(s_j)}{\sum_{j\in\mathcal{V}(C)} \exp(s_j)}. \tag{11}$$

Since $\mathcal{P} = \mathcal{G}_Y(C)$, the listwise term is $\mathcal{L}_{\mathrm{list}} = -\log q_\omega(Y \mid X, C)$, including events with multiple positive histories.

Proposition (Variational identity-information bound)**.** *For a fixed association-event distribution $p(X, Y, C)$ with finite conditional entropy and expected losses, and $\lambda_h \geq 0$,*

$$I(Y; X \mid C) \geq H(Y \mid C) - R_{\mathrm{list}}(\omega) \geq H(Y \mid C) - R_{\mathrm{assoc}}(\omega), \tag{12}$$

where $R_{\mathrm{list}} = \mathbb{E}[\mathcal{L}_{\mathrm{list}}]$ and $R_{\mathrm{assoc}} = \mathbb{E}[\mathcal{L}_e]$.

Proof sketch. The variational posterior decomposition (Poole et al., 2019) gives

$$R_{\text{list}}(\omega) = H(Y \mid X, C) + D_\omega, \qquad D_\omega = \mathbb{E}_{p(X,C)}[\mathrm{KL}(p(\cdot \mid X, C) \,\|\, q_\omega(\cdot \mid X, C))] \geq 0. \tag{13}$$

Substituting into $I(Y; X \mid C) = H(Y \mid C) - H(Y \mid X, C)$ yields the first inequality, with equality when the model and true posteriors agree almost surely. The non-negative margin term implies $R_{\text{assoc}} \geq R_{\text{list}}$, establishing the second inequality. Thus, reducing the population listwise risk tightens the variational identity-information bound, while the margin term encourages positive–negative separation. Appendix B provides the complete proof. This interpretation connects the grouped identity supervision to score learning; the contribution of the distributional scoring architecture is examined separately by the matched controls.

At each frame, the last-observation prefilter retains a context set and an eligible assignment subset from previously observed identities. TempoTrack evaluates the eligible candidates using Equation 9. For a query with at least two eligible candidates, the highest-scoring identity is proposed when both $s_{(1)} \geq \tau_s$ and $s_{(1)} - s_{(2)} \geq \tau_m$ hold. These acceptance thresholds define an operating point selected separately from the ranking objective. Competing proposals are resolved before frame-level identities are committed, and unmatched observations follow the frontend's track-initialization rules. The accepted observations then update the prototypes and retained histories.

4 Experiments

4.1 Experimental Setup

We evaluate TempoTrack at three complementary levels. First, Table 1 reports open-vocabulary TAO performance for Base and Novel categories, while Table 3 measures paired improvements within each tracking frontend. Second, Tables 5–7 examine historical evidence, temporal fusion, and scorer structure. Third, Appendix G examines matched-candidate identity ranking, while Appendix E reports training-seed robustness of the final model. We use AssocA as the primary association metric and report TETA, LocA, and ClsA to characterize the complete tracking output (Li et al., 2022). The overall TAO TETA and BDD MOT results are separate benchmark configurations. Base-only refers to the supervision used to optimize the association scorer, independently of each frontend's pretraining. Appendix A specifies the details.

4.2 Comparison with Existing Methods

Table 1 evaluates association beyond the supervised categories. TempoTrack improves Novel AssocA over COVTrack from 41.3 to 43.3 on TAO validation and from 32.6 to 33.6 on test, corresponding to gains of 2.0 and 1.0 percentage points. Novel TETA increases from 34.3 to 35.0 and from 28.9 to 29.5, respectively. These results support the use of temporal distributional evidence in the evaluated open-vocabulary setting. The complete LocA and ClsA values characterize accompanying changes in the tracker output rather than being assumed invariant under association changes. In the separate overall TAO TETA setting of Table 2, TempoTrack obtains 47.9 TETA and 47.6 AssocA, compared with 46.3 and 44.1 for MASA.

4.3 Applicability across Tracking Frontends

Table 3 tests whether the association formulation remains useful across different frozen tracking frontends. Adding TempoTrack to OVTrack, VOVTrack, and MASA improves Novel AssocA by 3.6, 3.1, and 3.1 percentage points, with Base improvements of 2.1, 2.3, and 2.0 points. TETer obtains gains of 1.1 and 1.0 points under its separately marked supervision setting, which includes both category groups. These paired results support modular applicability across the evaluated tracking pipelines. The experiment evaluates the applicability of the association design within each frontend configuration. The TETer setting is interpreted separately from the three configurations excluding Novel task-specific supervision.

Table 1: **Open-vocabulary tracking on the TAO validation and test sets (Li et al., 2023).** Results are separated into Novel and Base categories. Class flags indicate the category supervision of each listed configuration. Methods marked with † use the same detector. The best results are shown in **bold** and the second-best results are underlined.

Method	Venue (year)	Classes		Novel				Base			
		Novel	Base	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
Validation set											
QDTrack (2023)	TPAMI (2023)	✓	✓	22.5	42.7	24.4	0.4	27.1	45.6	24.7	11.0
TETer (2022)	ECCV (2022)	✓	✓	25.7	45.9	31.1	0.2	30.3	47.4	31.6	12.1
ByteTrack† (2022)	ECCV (2022)	-	✓	22.0	48.2	16.6	1.0	28.2	50.4	18.1	16.0
OC-SORT† (2023)	CVPR (2023)	-	✓	23.7	49.6	20.4	1.1	28.9	51.4	19.8	15.4
OVTrack† (2023)	CVPR (2023)	-	✓	27.8	48.8	33.6	1.5	35.5	49.3	36.9	20.2
MASA (R50)† (2024b)	CVPR (2024)	-	-	30.0	54.2	34.6	1.0	36.9	55.1	36.4	19.3
SLAck† (2024c)	ECCV (2024)	-	✓	31.1	54.3	37.8	1.3	37.2	55.0	37.6	19.1
OVSORT† (2025b)	TMM (2025)	-	✓	30.8	53.0	37.6	1.9	38.2	55.3	39.9	19.4
TRACT† (2025c)	ICCV (2025)	-	✓	31.3	52.7	37.8	3.4	38.5	55.0	39.0	**21.5**
OVTR (2025a)	ICLR (2025)	-	✓	31.4	54.4	34.5	**5.4**	36.6	52.2	37.6	20.1
COVTrack† (2025b)	ICCV (2025)	-	✓	34.3	58.2	41.3	3.5	**39.6**	**57.3**	42.0	19.6
SSGTrack† (2026)	TIP (2026)	-	✓	31.1	52.5	38.8	1.7	38.2	54.7	39.6	20.4
TempoTrack (Ours)†	-	-	✓	**35.0**	**58.6**	**43.3**	3.2	**39.6**	56.9	**42.2**	19.6
Test set											
QDTrack (2023)	TPAMI (2023)	✓	✓	20.2	39.7	20.9	0.2	25.8	43.2	23.5	10.6
TETer (2022)	ECCV (2022)	✓	✓	21.7	39.1	25.9	0.0	28.4	44.0	30.4	10.7
DeepSORT (ViLD)† (2017)	ICIP (2017)	-	✓	17.2	38.4	11.6	1.7	24.5	43.8	14.6	15.2
Tracktor++ (ViLD)† (2019)	ICCV (2019)	-	✓	18.0	39.0	13.4	1.7	26.0	44.1	19.0	14.8
OVTrack† (2023)	CVPR (2023)	-	✓	24.1	41.8	28.7	1.8	32.6	45.6	35.4	16.9
SLAck† (2024c)	ECCV (2024)	-	✓	27.1	49.1	30.0	2.0	34.7	52.5	35.6	16.1
OVSORT† (2025b)	TMM (2025)	-	✓	28.1	48.0	33.4	2.7	35.1	51.6	38.3	15.4
TRACT† (2025c)	ICCV (2025)	-	✓	27.3	48.2	30.7	3.1	36.2	52.3	39.1	**17.2**
OVTR (2025a)	ICLR (2025)	-	✓	27.1	47.1	32.1	2.1	34.5	51.1	37.5	14.9
COVTrack† (2025b)	ICCV (2025)	-	✓	28.9	50.9	32.6	**3.3**	37.9	54.5	42.1	**17.2**
SSGTrack† (2026)	TIP (2026)	-	✓	27.3	48.6	31.3	1.9	35.7	51.8	38.6	16.8
TempoTrack (Ours)†	-	-	✓	**29.5**	**51.8**	**33.6**	3.2	**38.0**	**55.0**	**42.3**	16.8

Table 2: **Comparison on the TAO TETA benchmark.** TETA is decomposed into association, localization, and classification accuracy. The best and second-best results are shown in **bold** and underline, respectively.

Method	Venue (year)	TETA↑	AssocA↑	LocA↑	ClsA↑
SORT Bewley et al. (2016)	ICIP 2016	24.9	14.3	48.1	12.1
Tracktor Bergmann et al. (2019)	ICCV 2019	24.2	13.0	47.4	12.1
Tracktor++ Bergmann et al. (2019)	ICCV 2019	28.0	22.8	49.0	12.1
DeepSORT Wojke et al. (2017)	ICIP 2017	26.0	17.5	48.4	12.1
UNINEXT Yan et al. (2023)	CVPR 2023	31.9	35.5	43.3	17.1
AOA Du et al. (2021)	arXiv 2021	25.3	30.6	23.4	21.9
ByteTrack Zhang et al. (2022)	ECCV 2022	27.6	20.2	48.3	14.4
OC-SORT Cao et al. (2023)	CVPR 2023	28.6	21.8	49.7	14.3
OVTrack Li et al. (2023)	CVPR 2023	34.7	36.7	49.3	18.1
QDTrack Fischer et al. (2023)	TPAMI 2023	30.0	27.4	50.5	12.1
TETer Li et al. (2022)	ECCV 2022	34.6	36.7	52.1	15.0
GLEE-Plus Wu et al. (2024)	CVPR 2024	41.5	40.9	52.9	**30.8**
GLATrack Li et al. (2024a)	ACMMM 2024	37.1	38.5	54.1	18.6
MASA Li et al. (2024b)	CVPR 2024	46.3	44.1	65.8	28.9
OVTR Li et al. (2025a)	ICLR 2025	35.7	36.3	52.3	18.4
SSGTrack Li et al. (2026)	TIP 2026	37.2	39.1	54.3	18.2
TempoTrack	Ours	**47.9**	**47.6**	**66.0**	30.1

Table 3: **Association improvements across tracking frontends.** TempoTrack is integrated into each tracker without retraining its frozen feature representation. Each improvement is calculated against the corresponding baseline under the same configuration.

Setting	Classes		Novel				Base			
	Base	Novel	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
TETer	✓	✓	25.7	45.9	31.1	0.2	30.3	47.4	31.6	12.1
TETer+Ours	✓	✓	26.3	46.1	32.2	0.7	31.0	47.6	32.6	12.8
vs. TETer			**(+0.6)**	(+0.2)	**(+1.1)**	(+0.5)	**(+0.7)**	(+0.2)	**(+1.0)**	(+0.7)
OVTrack	✓	-	27.8	48.8	33.6	1.5	35.5	49.3	36.9	20.2
OVTrack+Ours	✓	-	29.4	49.5	37.2	1.5	36.8	49.7	39.0	21.6
vs. OVTrack			**(+1.6)**	(+0.7)	**(+3.6)**	(+0.0)	**(+1.3)**	(+0.4)	**(+2.1)**	(+1.4)
VOVTrack	✓	-	35.3	58.5	40.9	6.5	39.6	58.9	40.9	19.1
VOVTrack+Ours	✓	-	36.5	59.1	44.0	6.5	40.9	59.4	43.2	20.0
vs. VOVTrack			**(+1.2)**	(+0.6)	**(+3.1)**	(+0.0)	**(+1.3)**	(+0.5)	**(+2.3)**	(+0.9)
MASA	✓	-	30.0	54.2	34.6	1.0	36.9	55.1	36.4	19.3
MASA+Ours	✓	-	31.3	55.0	37.7	1.1	38.2	55.8	38.4	20.5
vs. MASA			**(+1.3)**	(+0.8)	**(+3.1)**	(+0.1)	**(+1.3)**	(+0.7)	**(+2.0)**	(+1.2)

4.4 Evaluation on BDD MOT

Table 4 evaluates the association formulation in driving scenes. TempoTrack obtains 72.0 IDF1, 56.2 TETA, and 53.7 AssocA, the highest values in these three columns among the listed methods. Relative to MASA, the improvements are 0.7, 1.8, and 0.8 percentage points, respectively; mIDF1 and mMOTA also increase from 55.8 to 56.3 and from 44.6 to 45.3. The joint increase in identity and association measures supports the benefit of temporal evidence in this benchmark configuration. The complete five-metric comparison characterizes the tracking outcome under the reported BDD training and evaluation configuration.

Table 4: **Comparison on the BDD MOT benchmark (Yu et al., 2020).** We report identity, association, and overall tracking metrics in driving scenes. The best and second-best results are shown in **bold** and underline, respectively.

Method	IDF1↑	TETA↑	AssocA↑	mIDF1↑	mMOTA↑
QDTrack Fischer et al. (2023)	71.5	47.8	48.5	50.8	36.6
TETer Li et al. (2022)	71.1	50.8	52.9	53.3	39.1
MOTR Zeng et al. (2022)	65.8	-	-	54.0	32.3
Unicorn Yan et al. (2022)	71.3	-	-	54.0	41.2
UNINEXT-H Yan et al. (2023)	69.9	-	-	**56.7**	44.2
ByteTrack Zhang et al. (2022)	70.4	55.7	51.5	54.8	**45.5**
MASA Li et al. (2024b)	71.3	54.4	52.9	55.8	44.6
TempoTrack	**72.0**	**56.2**	**53.7**	56.3	45.3

Table 5: **Ablation of temporal matching evidence representation.** The variants share the frontend, candidate budget, history-construction policy, and training data. We vary the representation and scoring of historical evidence.

Method	Novel				Base			
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
Prototype Matching	32.9	57.6	38.2	2.9	37.2	55.8	37.3	18.4
Mean History	33.5	57.8	39.7	3.0	37.7	56.0	38.5	18.7
Max History	33.9	58.0	40.5	3.1	38.1	56.2	39.1	18.9
Mean + Variance	33.7	57.9	40.1	3.0	37.9	56.1	38.8	18.8
Single Distribution	34.2	58.2	41.2	3.1	38.6	56.5	40.2	19.1
TempoTrack (Full)	**35.0**	**58.6**	**43.3**	**3.2**	**39.6**	**56.9**	**42.2**	**19.6**

Table 6: **Multi-timescale temporal association ablation.** We compare recent-only and long-term-only evidence with equal-weight, fixed-weight, and learned adaptive fusion. The visual frontend and historical-evidence construction policy are shared.

Method	Novel				Base			
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
Recent only	33.6	57.9	39.8	3.0	38.0	56.1	38.9	18.9
Long-term only	34.2	58.2	41.2	3.1	38.6	56.5	40.2	19.1
Equal fusion	34.3	58.3	41.6	3.1	38.8	56.6	40.5	19.3
Fixed fusion	34.5	58.4	41.9	3.1	39.0	56.7	40.8	19.4
Adaptive fusion (Full)	**35.0**	**58.6**	**43.3**	**3.2**	**39.6**	**56.9**	**42.2**	**19.6**

Table 7: **Scorer-structure and parameterization controls.** The plain MLP controls and the full model use the same historical descriptors. The control without m_i^r, m_i^l removes the two deterministic log-MGF coordinates while retaining their constituent means and variances. Parameter counts cover the trainable association scorers.

Scorer	Parameters	TETA	LocA	AssocA	ClsA
Same-evidence MLP	4,289	34.5	58.5	41.9	3.1
Capacity-matched MLP	4,757	34.6	58.6	42.0	3.1
Core w/o m_i^r, m_i^l	4,596	34.9	58.6	43.0	3.2
TempoTrack (Full)	**4,724**	**35.0**	**58.6**	**43.3**	**3.2**

4.5 Ablation Study

Historical Matching Evidence Representation. Table 5 compares different representations of historical matching evidence under the same tracking configuration. Novel AssocA is 38.2 with Prototype Matching, 39.7 with Mean History, 40.5 with Max History, and 40.1 with Mean + Variance. Using a Single Distribution increases AssocA to 41.2, while the full model reaches 43.3. The corresponding Base AssocA values are 37.3, 38.5, 39.1, 38.8, 40.2, and 42.2. Mean + Variance does not uniformly outperform Max History, indicating that the observed gain is not explained by adding variance alone. The larger difference appears when historical distributional evidence is combined with the temporal association mechanism.

Multi-Timescale Temporal Association. Table 6 shows that adaptive fusion reaches 43.3 Novel and 42.2 Base AssocA, improving on fixed fusion by 1.4 points in both groups and on long-term evidence alone by 2.1 and 2.0 points. Candidate-dependent weighting therefore makes more effective use of the available temporal evidence than either a single range or a global weighting rule.

Scorer Structure and Moment Parameterization. Table 7 further separates scorer structure from parameter count and the deterministic MGF coordinates. TempoTrack reaches 43.3 Novel AssocA versus 42.0 for the capacity-matched MLP with slightly fewer parameters, while removing m_i^r, m_i^l gives 43.0. The larger observed separation therefore comes from the structured use of temporal and candidate-relative evidence rather than parameter count or the two MGF coordinates alone. Complementary analyses support distinct aspects of this interpretation: Appendix G reports improved Rank-1 identity discrimination under matched candidate evidence, while Appendix E shows little variation across the three examined training seeds. The former is a Base internal-holdout diagnostic and is not treated as evidence of Novel-category generalization.

5 Conclusion

TempoTrack offers a distributional perspective on OVMOT association, focusing on how retained histories support the current query rather than compressing them into a single reference. It represents each query–identity pair through query-conditioned temporal matching distributions, whose projected moments summarize support across recent and long-term ranges. Adaptive temporal fusion and candidate-conditioned scoring turn this evidence into a lightweight, frame-causal association score for frozen open-vocabulary frontends. Across multiple benchmarks and frontends, the results show the promise of organizing historical matching evidence in this way for improving identity association.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including TAO, BDD100K, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. We have also provided a full description of TempoTrack, to assist others in reproducing our experiments. Additionally, Types of public datasets used in the paper, such as TAO, BDD100K, are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

LLM USAGE

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

REFERENCES

Görkay Aydemir, Xiongyi Cai, Weidi Xie, and Fatma Güney. Track-On: Transformer-Based Online Point Tracking with Memory. In *International Conference on Learning Representations*, 2025.

Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 941–951, 2019.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*, pp. 3464–3468, 2016.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9686–9696, 2023.

Fei Du, Bo Xu, Jiasheng Tang, Yuqi Zhang, Fan Wang, and Hao Li. 1st place solution to ECCV-TAO-2020: Detect and represent any object for tracking. *arXiv preprint arXiv:2101.08040*, 2021.

Tobias Fischer, Thomas E. Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15380–15393, 2023.

Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11516–11525, 2021.

Guangyao Li, Yajun Jian, Yan Yan, and Hanzi Wang. GLATrack: Global and local awareness for open-vocabulary multiple object tracking. In *ACM International Conference on Multimedia*, pp. 2457–2466, 2024.

Jinyang Li, En Yu, Sijia Chen, and Wenbing Tao. OVTR: End-to-End Open-Vocabulary Multiple Object Tracking with Transformer. In *International Conference on Learning Representations*, 2025.

Run Li, Dawei Zhang, Yanchao Wang, Yunliang Jiang, Zhonglong Zheng, Sang-Woon Jeon, and Hua Wang. Open-vocabulary multi-object tracking with domain generalized and temporally adaptive features. *IEEE Transactions on Multimedia*, 27:3009–3022, 2025.

Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E. Huang, and Fisher Yu. Tracking every thing in the wild. In *European Conference on Computer Vision*, pp. 498–515, 2022.

Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. OVTrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5567–5577, 2023.

Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segù, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18963–18973, 2024.

Siyuan Li, Lei Ke, Yung-Hsu Yang, Luigi Piccinelli, Mattia Segù, Martin Danelljan, and Luc Van Gool. SLAck: Semantic, location, and appearance aware open-vocabulary tracking. In *European Conference on Computer Vision*, pp. 1–18, 2024.

Yunhao Li, Yifan Jiao, Dan Meng, Heng Fan, and Libo Zhang. Attention to trajectory: Trajectory-aware open-vocabulary tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14390–14398, 2025.

Zekun Qian, Ruize Han, Junhui Hou, Linqi Song, and Wei Feng. VOVTrack: Exploring the potentiality in raw videos for open-vocabulary multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7472–7482, 2025.

Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. COVTrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10054–10063, 2025.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*, pp. 3645–3649, 2017.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3795, 2024.

Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *European Conference on Computer Vision*, pp. 733–751, 2022.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15325–15336, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2636–2645, 2020.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pp. 659–675, 2022.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pp. 1–21, 2022.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5171–5180, 2019.

Jinyuan Qu, Hongyang Li, Shilong Liu, Tianhe Ren, Zhaoyang Zeng, and Lei Zhang. TAPTRv3: Spatial and Temporal Context Foster Robust Tracking of Any Point in Long Video. In *International Conference on Learning Representations*, 2026.

Bishoy Galoaa, Xiangyu Bai, Utsav Nandi, Sai Rangoju, Somaieh Amraee, and Sarah Ostadabbas. UniTrack: Differentiable Graph Representation Learning for Multi-Object Tracking. In *International Conference on Learning Representations*, 2026.

Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19004, 2024.

Guangyao Li, Siping Zhuang, Yajun Jian, Yan Yan, and Hanzi Wang. Spatial-temporal scene graph generation for open-vocabulary multiple object tracking. *IEEE Transactions on Image Processing*, 35:3437–3450, 2026.