TRACKRERANK: SEMANTIC ENHANCEMENT FOR TWO-STAGE OPEN-VOCABULARY TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

A two-stage open-vocabulary tracker can follow an object correctly while repeatedly assigning it the wrong category. We introduce TrackRerank, a causal post-association semantic head that revisits category decisions without changing boxes, detection confidence, or identities. It gathers candidates from detector predictions, current-frame vision–language retrieval, and past track observations. A category-shared scorer interprets each candidate through its source confidence, relative competition, uncertainty, cross-source support, and track context. On a custom video-disjoint OVT-B split, the complete pipeline raises YOLOE–ByteTrack from 13.307 to 25.300 ClsA (+11.993 points), exceeding causal voting and EMA baselines. Under shared candidates, supervision, and optimization, the full representation gains 1.050 points over a missingness-aware score-fusion MLP with nearly identical parameter count, with positive gains in all three paired seeds. The same trained head, connected through an observation-aligned adapter, improves native OVTR on TAO by 1.390 causal ClsA points without target-domain fitting. Its within-group accuracy gains are larger when current sources conflict or are uncertain. These results distinguish the benefit of gathering additional semantic evidence from the benefit of interpreting its candidate-level context.

1 INTRODUCTION

An object can be tracked correctly and still be named incorrectly. In a detection–association pipeline, an open-vocabulary detector proposes objects and an association method links them across frames. The resulting track may be continuous even when its category alternates between several wrong predictions. Figure 1 illustrates this mismatch. A stable identity says which observations belong together; it does not, by itself, resolve what the object is. This distinction matters for visually similar categories and blurred, occluded, or incomplete views.

Two-stage tracking offers a useful place to address this problem. Once observations have been associated, their shared identity provides context for interpreting the current prediction. A vision–language model can suggest categories absent from the detector's shortlist, while earlier views of the same object can preserve hypotheses that the current frame no longer supports strongly. These sources are available without changing the detector or rebuilding the tracks. The practical question is how to turn them into a better category decision.

A natural answer is score fusion. Yet the same score may be a clear winner in one source distribution and one of several close alternatives in another. A large memory score may reflect repeated support or a long sequence of consistent mistakes. A scorer that sees only one scalar per source does not see these surrounding distributions or how the track evolved. The question is therefore not just which source to trust, but what makes its support for this candidate credible.

We propose TrackRerank, an online semantic enhancement head for two-stage open-vocabulary tracking. It first forms a candidate union from detector predictions, current-frame CLIP retrieval, and causal track memory. It then describes each candidate with a compact evidence vector: source scores and presence, ranks and margins, distribution uncertainty, cross-source agreement, and observation and track statistics. A small shared network scores these vectors and selects the output category. Category identifiers are used for candidate alignment and memory indexing, but are not inputs to the network. Memory is updated only after the current prediction, using upstream evidence rather than reranked labels.

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.

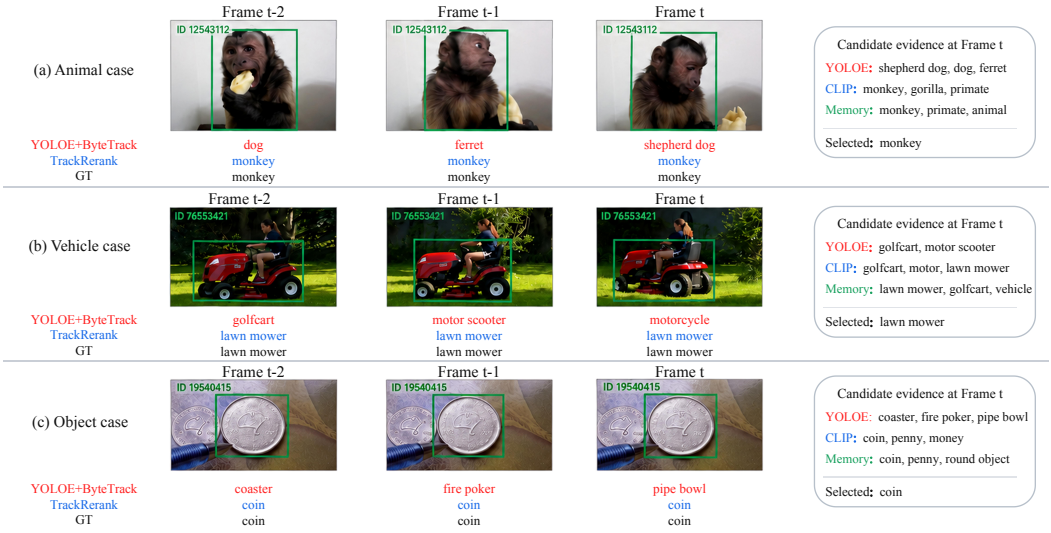


Figure 1: Continuous identities, inconsistent categories. Across three views, the detector assigns different incorrect labels to the same monkey, lawn mower, and coin. Alternative sources retain useful hypotheses for semantic correction. These original examples illustrate the motivation; their prediction labels come from an earlier system, not the current held-out evaluation.

The complete pipeline raises YOLOE–ByteTrack from 13.307 to 25.300 ClsA on our held-out OVT-B split without changing boxes, confidence, or identities. This 11.993-point gain includes additional sources and training. Under shared candidates and supervision, TrackRerank retains a 1.050-point advantage over an unrestricted score-and-mask MLP with nearly the same parameter count. The two comparisons distinguish the benefit of the complete enhancement from the additional value of candidate context.

Temporal semantic modeling is already an important part of open-vocabulary tracking. OVTR propagates category information across frames (Li et al., 2025a), and TRACT uses trajectory-level visual and language information for classification (Li et al., 2025b). We retain native OVTR and derive observation-aligned evidence from its two decoder branches. The source-trained head improves causal ClsA by 1.390 points, including 0.530 over the score-and-mask control. On these outputs, its matched-accuracy advantage is larger for conflicting sources and high-uncertainty predictions. This connects the method's purpose to where it helps: interpreting competing hypotheses, including those produced by an already trajectory-aware model.

Our contributions are threefold. First, we formulate semantic enhancement for two-stage tracking as candidate-level evidence resolution on fixed tracks. Second, we introduce a category-shared evidence interface and test it against causal aggregation and capacity- and missingness-controlled fusion heads. Third, paired-seed analysis, supervision controls, native OVTR transfer, and conditional diagnostics examine its reliability and scope. Neither a new visual backbone nor a new association algorithm is required.

2 Related Work

Open-vocabulary detection and tracking. Vision–language pretraining makes text-defined recognition possible (Radford et al., 2021). OWLv2, YOLO-World, Grounding DINO, and YOLOE extend this capability to object detection (Minderer et al., 2023; Cheng et al., 2024; Liu et al., 2024; Wang et al., 2025). Pairing such a detector with an association method such as ByteTrack (Zhang et al., 2022) gives a modular tracking pipeline, but leaves category quality closely tied to detector predictions. OVTrack learns open-vocabulary classification and association through vision–language knowledge transfer (Li et al., 2023); SLAck combines semantic, location, and appearance cues for tracking (Li et al., 2024). Our method uses upstream outputs as evidence and leaves the underlying detection and association results intact. TAO and OVT-B provide broad-vocabulary evaluation set-

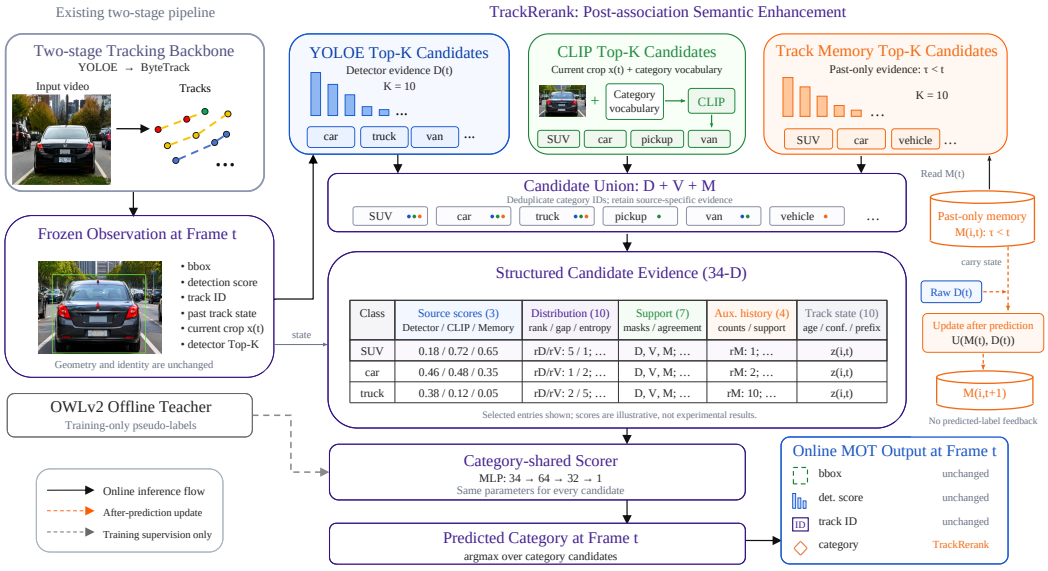


Figure 2: TrackRerank after a two-stage tracker. Current detector and CLIP predictions are combined with past-only memory to form a candidate union. A shared scorer reads 34-dimensional candidate evidence and changes only the category. The base implementation updates memory with raw detector evidence after prediction; reranked labels never feed back into memory or association. OWLv2 supplies offline training labels only. Selected feature values in the diagram are illustrative.

tings (Dave et al., 2020; Liang & Han, 2024), while TETA separates localization, association, and classification quality (Li et al., 2022).

Trajectory-aware semantics. OVTR uses category information propagation to carry semantic priors into subsequent frames (Li et al., 2025a). TRACT combines trajectory-enhanced association with TraCLIP, whose visual aggregation and language enrichment improve trajectory classification (Li et al., 2025b). COVTrack adapts the fusion of appearance, motion, and semantic cues to their estimated reliability (Qian et al., 2025b), and VOVTrack exploits object states and video-centric training (Qian et al., 2025a). These methods already establish the usefulness of temporal context. Our focus is the representation used by the final category decision: a candidate is described by evidence from several sources, rather than by a new trajectory embedding. The OVTR experiment tests this head while retaining the upstream semantic model; it is not a comparison between complete tracking systems.

Fusion and candidate scoring. Learned gates provide input-dependent control over source contributions (Arevalo et al., 2017); score MLPs can also learn nonlinear combinations. We distinguish the fusion function from the information supplied to it. Our controls include a masked gate and an unrestricted score-and-mask MLP with nearly the same parameter count as TrackRerank. They share candidates and supervision; the full representation additionally exposes competing-score statistics and track context (Appendix C).

3 TrackRerank

Figure 2 summarizes the base two-stage implementation.

3.1 Semantic enhancement on fixed tracks

Let $o_{i,t} = (b_{i,t}, p_{i,t}, \iota_i, \mathbf{d}_{i,t})$ denote an observation of track i at frame t, with box $b_{i,t}$, detection confidence $p_{i,t}$, identity ι_i, and detector scores $\mathbf{d}_{i,t}$ over a target vocabulary $\mathcal{V}$. The crop $x_{i,t}$ is

extracted from the predicted box. TrackRerank produces a category

$$\widehat{c}_{i,t} = \arg\max_{c \in \mathcal{C}_{i,t}} g_\theta\big(\phi_{i,t}(c)\big), \tag{1}$$

where $\mathcal{C}_{i,t}$ is the candidate pool and $\phi_{i,t}(c)$ is the evidence for candidate c. All non-category output fields are passed through unchanged. No observation is added, removed, or reassociated.

This interface is causal. The head reads the current observation and a state constructed from observations at $u < t$. It predicts first and updates memory second. Frozen tracks are a condition of comparison, not an assumption of perfect tracks: upstream localization and identity errors remain present and can affect the evidence available to the head.

3.2 Candidate discovery from current and past evidence

The detector contributes its top K_D categories. A frozen CLIP model supplies a second shortlist by matching the current crop to category text:

$$s_V(c) = \cos(f_{\text{img}}(x_{i,t}), f_{\text{text}}(c)), \qquad \mathcal{C}^V_{i,t} = \text{TopK}_{K_V}(s_V). \tag{2}$$

Here $f_{\text{text}}(c)$ encodes the category prompt. Entropy and margins use a full-vocabulary softmax at temperature 0.07. Only the current crop is encoded; text embeddings remain upstream of the reranker.

In the base two-stage implementation, memory accumulates raw detector evidence under each video and track identity. If $c^D_{i,u}$ is the detector's top category at an earlier observation, define

$$n_{i,t}(c) = \sum_{u<t} \mathbf{1}[c^D_{i,u} = c], \qquad w_{i,t}(c) = \sum_{u<t} p_{i,u} \mathbf{1}[c^D_{i,u} = c]. \tag{3}$$

The memory score is $s_M(c) = \log(1 + n_{i,t}(c)) + 0.5 w_{i,t}(c)$, and $\mathcal{C}^M_{i,t} = \text{TopK}_{K_M}(s_M)$. Memory also retains statistics of historical detector top-K evidence. It contains no teacher labels, ground truth, or previous reranker predictions. At a track's first observation, the memory source is absent.

The final pool is $\mathcal{C}_{i,t} = \mathcal{C}^D_{i,t} \cup \mathcal{C}^V_{i,t} \cup \mathcal{C}^M_{i,t}$, with $K_D = K_V = K_M = 10$. Duplicate category IDs are merged while source-specific support is retained. The current sources can recover hypotheses missed by each other; memory keeps earlier hypotheses available when the current view is less informative.

3.3 Candidate-level evidence resolution

For each candidate, we organize evidence into five groups:

$$\phi_{i,t}(c) = \big[\mathbf{s}(c); \mathbf{u}(c); \mathbf{a}(c); \mathbf{h}(c); \mathbf{z}_{i,t}(c)\big] \in \mathbb{R}^{34}. \tag{4}$$

The groups encode source scores $\mathbf{s}$, current-source distribution context $\mathbf{u}$, source support $\mathbf{a}$, auxiliary history $\mathbf{h}$, and observation/track context $\mathbf{z}$. The last includes both shared state and candidate-specific prefix consistency. Appendix B gives the 34-slot inventory and feature definitions.

These features expose information missing from a per-candidate score tuple. Two observations can supply the same candidate scores and masks while differing in the scores of competing categories or in past support; a score-and-mask head receives identical inputs in those cases. Ranks, margins, and prefix statistics distinguish the evidence states. Here, "structured" refers to organization by candidate, source, and track, not to a graph architecture or a new visual representation.

Ranks and margins describe a candidate relative to its competitors; entropy and track state provide shared context. For each current source, the margin is the candidate probability minus that of its strongest rival, and entropy is computed over the complete vocabulary. Shared context affects ranking through interactions: an additive track-state term would cancel under the argmax, whereas a nonlinear scorer can interpret memory support differently for a short, uncertain prefix and a well-observed track. The feature controls test the empirical value of this information without presupposing which interactions the model learns.

Continuous inputs are normalized using source-training statistics, which are frozen for validation, testing, and transfer. Missing normalized source scores are set to zero. The scorer is a $34 \to 64 \to 32 \to 1$ MLP with ReLU and 0.05 dropout, totaling 4,353 parameters including biases. We refer to this implementation as Full34. The same parameters score every category: no category ID, one-hot vector, or text embedding enters g_θ. Its output ranks the current candidates and is not assumed to be a calibrated probability.

3.4 TRAINING AND ONLINE EXECUTION

A frozen OWLv2 teacher (Minderer et al., 2023) supplies pseudo-labels on the training videos. Each frozen box is matched to the highest-IoU teacher detection if IoU is at least 0.5. Confidence-weighted teacher distributions are aggregated within a training track; the top category supplies c_i^T when its aggregate probability is at least 0.20. The teacher may use later training views, while student inputs remain causal. OWLv2 is absent during validation and test inference.

An observation is eligible only with a valid teacher match and its track pseudo-category naturally in the candidate pool. We retain that positive, sample up to three high-prior negatives and one random negative, and share sampling lists across heads. For the resulting set $\mathcal{S}_{i,t}$, we use

$$\ell_{i,t} = -\frac{1}{|\mathcal{S}_{i,t}|} \sum_{c \in \mathcal{S}_{i,t}} \left[4y_c \log \sigma(r_c) + (1 - y_c) \log(1 - \sigma(r_c)) \right], \tag{5}$$

where $r_c = g_\theta(\phi_{i,t}(c))$ and $y_c = \mathbf{1}[c = c_i^T]$. Losses are averaged within each eligible track and then across tracks, limiting the influence of long trajectories. Inference scores the entire candidate pool.

At each observation, we read past memory, construct and score the current candidate vectors, and select the category. Only then is raw upstream evidence written to the next memory state. Memory is reset between videos. The OVTR adapter uses CTI logits and projected OFA features instead of detector and crop-CLIP scores, with a past-only EMA memory; the scorer and source normalization remain fixed (Appendix G).

4 EXPERIMENTS

We first measure the semantic improvement over the original two-stage tracker, then test whether score fusion explains it. Candidate and feature controls examine the mechanism; native OVTR, TAO transfer, and supervision controls examine its scope.

4.1 EXPERIMENTAL SETUP

Data and frozen outputs. We use OVT-B (Liang & Han, 2024) with a custom video-disjoint split of 1,381 training, 296 validation, and 296 held-out test videos. These contain 603,678, 128,352, and 143,663 frozen observations, respectively. This is not an official OVT-B test-dev split, and we do not mix its scores with published full-benchmark results. The source pipeline is YOLOE–ByteTrack (Wang et al., 2025; Zhang et al., 2022). All semantic methods operate on the same predictions; field-level checks preserve video/frame IDs, boxes, confidence, track IDs, and associations. TAO (Dave et al., 2020) is used for transfer and the native OVTR evaluation.

Vocabulary and positive-supervision protocols. The main comparison uses a target-vocabulary protocol: names are available to the detector, CLIP, and teacher, and eligible Base and Novel pseudo-labels supervise the head. A separate Base-only control excludes Novel positive targets but retains the vocabulary and negative candidates. This tests transfer beyond positive training categories, not a category-disjoint pipeline.

Metrics and diagnostics. We report TETA and its classification component ClsA (Li et al., 2022), with $\mathrm{TETA} = (\mathrm{LocA} + \mathrm{AssocA} + \mathrm{ClsA})/3$. In the OVT-B comparison, LocA is 54.851 and AssocA is 63.761 throughout. Reachability and matched accuracy use a fixed set of 81,113 matched observations. History is computed on the complete causal prediction stream before this restriction, including unmatched earlier observations. These are observation-level diagnostics, not alternative implementations of ClsA (Appendix A).

Implementation and uncertainty. CLIP uses ViT-L/14, 224×224 crops, and FP16 encoding. Heads train for 20 epochs with AdamW, learning rate and weight decay 10^{-4}, and validation-ClsA checkpoint selection. Seeds are 42, 43, and 44; $\pm$ denotes sample standard deviation where reported. Paired-video intervals use 10,000 bootstrap samples, recomputing category/threshold aggregation and averaging paired differences over the three fixed seeds. These intervals quantify video-sampling uncertainty, not seed-to-seed variation.

Table 1: Semantic enhancement on the custom OVT-B split. LocA $= 54.851$ and AssocA $= 63.761$ throughout. Causal aggregation uses the observed prefix without head training. The learned lower block shares candidates and supervision. Parameters count the head only; $\pm$ is sample seed standard deviation.

Semantic decision	Parameters	ClsA ↑	TETA ↑
YOLOE–ByteTrack category	—	13.307	43.973
Current-frame CLIP	—	17.876	45.496
Causal majority vote	—	20.850	46.487
Detector EMA	—	21.620	46.744
Detector + CLIP EMA	—	22.980	47.197
Global affine fusion	4	17.500 ± 0.320	45.371
Score3-Small	2,369	23.470 ± 0.119	47.361
Score3-PM	4,305	23.700 ± 0.130	47.437
Gated-Score-PM	4,275	24.050 ± 0.140	47.554
Score+Mask-PM	4,348	24.250 ± 0.060	47.621
TrackRerank (Full34)	4,353	$\mathbf{25.300} \pm 0.145$	**47.971**

4.2 Improving the semantic output of a two-stage tracker

Table 1 shows that TrackRerank raises the original pipeline from 13.307 to 25.300 ClsA, a gain of **11.993 points**. TETA increases from 43.973 to 47.971 entirely through its classification component: localization and association do not change. Thus, the existing tracks support substantially better category predictions without modifying the upstream tracker.

Current-frame CLIP reaches 17.876. Causal majority voting reaches 20.850, detector EMA reaches 21.620, and detector–CLIP EMA reaches 22.980. Temporal aggregation is useful, but the strongest tested smoothing baseline remains 2.320 points below Full34. This difference compares complete methods with different training requirements; the learned controls below more closely test the value of candidate context (Appendix C).

4.3 Does the gain reduce to score fusion?

All learned comparisons share candidate pools, teacher labels, training observations, loss, optimization budget, and splits. Global fusion is an affine map of three source scores. Score3-Small is a score-only MLP; Score3-PM widens it to within 1.1% of Full34's parameter count. Gated-Score-PM also reads source masks and predicts candidate-dependent source weights.

The gate forms a convex combination of available normalized scores; absent sources receive zero weight. To separate that output constraint from access to missingness information, **Score+Mask-PM** takes the same six inputs and uses an unrestricted $6 \to 87 \to 42 \to 1$ MLP. It has 4,348 parameters, only five fewer than Full34. Neither control receives competing-category statistics or auxiliary track context (Appendix C).

Score fusion accounts for much of the overall gain. Score3-Small reaches 23.470 ClsA; increasing its size adds 0.230 points, and gating reaches 24.050. Score+Mask-PM improves further to 24.250. Yet Full34 reaches 25.300, retaining **1.050 points** over this missingness-aware, unrestricted control. The remaining difference is not explained by parameter count, source masks, or freedom of the nonlinear readout alone. It supports the value of the additional evidence under the tested architectures and training protocol, rather than a universal limit on score-based models.

The Full34–Score+Mask gain is positive for all three paired seeds, ranging from 0.960 to 1.130 points; its mean and sample standard deviation are 1.050 ± 0.085. Separately, the paired-video 95% interval is $[0.390, 1.700]$. A deletion test retains a 1.310-point Full34–Score3-PM gap after removing the five largest positive-influence videos (Appendix D).

Table 2: Leave-one-group-out ablation on OVT-B. Candidate pools, training observations, and the 34-slot network interface are fixed. Disabled features are zeroed after normalization, and each variant is retrained. Differences are conditional deletion effects, not additive component contributions.

Evidence available	ClsA ↑	Difference
Full34	**25.300**	—
Without rank / margin / uncertainty	24.200	-1.100
Without source presence / agreement	24.620	-0.680
Without auxiliary history (slots 18–21)	24.880	-0.420
Without memory score (slot 17)	25.050	-0.250
Without all history values (slots 17–21)	24.650	-0.650
Without observation / track state	25.080	-0.220

4.4 WHICH EVIDENCE HELPS?

Discovery versus selection. On the fixed matched diagnostic set, detector top-10 candidates contain the correct category in 76.0% of observations, and CLIP top-10 in 70.0%. Their union reaches 92.9%; memory increases this to 93.7%. The current sources are complementary, while memory adds relatively few previously unreachable categories. With the full pool, matched accuracy is 45.4%, so the presence of a correct candidate is far from sufficient for selecting it. This distinction concerns the diagnostic set and does not turn reachability into an upper bound on class-balanced ClsA.

The shared diagnostic denominator also separates two kinds of error: approximately 6.3% of observations lack the correct candidate, whereas 48.3% contain it but select another category. Conditional accuracy among reachable observations is about $45.4/93.7 = 48.5\%$. These values, derived from the rounded rates, locate substantial remaining error in selection rather than candidate discovery.

Distribution and support context. Table 2 removes feature groups while retaining the network and candidate interface. The largest drop, 1.100 points, occurs when ranks, margins, and uncertainty are removed. Removing source presence and agreement costs 0.680 points. Both results support conditioning a candidate's score on its competition and source support. Removing the memory score and auxiliary history together costs 0.650 points; removing observation and track state costs 0.220. These are effects in the presence of the remaining features, so redundancy and interactions preclude adding them as separate contributions.

The role of history. A separate 2×2 control distinguishes adding memory candidates from exposing memory values. The four models omit generic track state and use the same teacher-reachable training observations. Adding candidates alone yields $+0.250$ ClsA, with a 95% interval of $[-0.030, 0.520]$. Exposing historical values for current candidates yields $+0.470$, with $[0.110, 0.810]$, and improves matched accuracy by 4.4 percentage points. The interaction is -0.090, with $[-0.350, 0.170]$. Memory's more clearly supported role here is to inform selection, not simply to add category names. Candidate identities and source relations can still carry history, so these are conditional effects rather than a complete removal of temporal information. The score-fusion controls also receive the memory score: their comparison with Full34 is not temporal versus frame-only inference. Appendix E gives the full design.

4.5 COMPLEMENTING A TRAJECTORY-AWARE SEMANTIC MODEL

We retain OVTR's native category information propagation (CIP) and attach the head to its frozen outputs. The adapter maps the two current-source slots to the same query's native CTI distribution and an OFA-alignment distribution, obtained by projecting its aligned feature onto frozen CLIP text embeddings at temperature 0.07. OFA is not treated as another native classifier. CIP remains in the upstream recurrence and propagates information into the next frame's track queries. A past-only EMA with decay 0.90 supplies memory; no external boxes or IoU rematching are introduced. This adapter changes how evidence is constructed, while the OVT-B-trained scorer and normalization remain fixed. The experiment therefore tests transfer through a specified evidence interface, not transfer of unchanged detector scores (Appendix G).

Table 3: Semantic heads on the same frozen native OVTR outputs on TAO. All added heads transfer from OVT-B without target-domain fitting. The causal setting excludes TCC; the offline setting applies the same TCC after every head. A dash denotes an unreported result.

Semantic decision	Causal ClsA ↑	Offline TCC ClsA ↑
Native OVTR/CIP	15.760	17.352
Global affine fusion	16.050	—
Score3-Small	16.250	—
Score3-PM	16.400	18.350
Gated-Score-PM	16.550	18.550
Score+Mask-PM	16.620	—
Full34	**17.150**	**19.350**

Table 4: Conditional matched accuracy on OVTR. Corrected and harmed are percentages of each group relative to Score+Mask-PM; their difference equals the net accuracy gain. The two partitions are analyzed separately.

Current evidence	Score+Mask	Full34	Gain (pp)	Corrected	Harmed
Sources agree	61.0%	61.8%	+0.8	6.2%	5.4%
Sources conflict	36.8%	39.7%	+2.9	9.1%	6.2%
Low uncertainty	55.4%	56.1%	+0.7	4.6%	3.9%
High uncertainty	30.2%	34.6%	+4.4	10.7%	6.3%

Under causal inference without TCC, Full34 improves native ClsA from 15.760 to 17.150 (Table 3). It exceeds Score3-PM by 0.750, gated fusion by 0.600, and Score+Mask-PM by **0.530** points. The last comparison has a paired-video 95% interval of $[0.160, 0.910]$. Thus, evidence context remains useful after trajectory-aware representation learning, even against a missingness-aware, unrestricted nonlinear fusion head.

For offline evaluation, every head is followed by the same full-trajectory category consistency (TCC) step. Full34 reaches 19.350, compared with 17.352 for native OVTR and 18.550 for gated fusion. These are offline mean comparisons; they do not extend the causal claim to TCC. A single category per trajectory also makes category switching zero by construction, which is not evidence of additional online stability. The experiment supports complementarity at the semantic interface, rather than superiority over OVTR as a complete tracker.

4.6 When does candidate context help?

The score-and-mask control lets us examine where the additional context matters. On matched OVTR observations, we group predictions by agreement between the current CTI and OFA top categories, without using memory or ground truth to assign groups. We also separate high-uncertainty observations using the larger of the two normalized distribution entropies and a validation-derived 75th-percentile threshold. These are diagnostic partitions; they do not change the semantic head or its predictions (Appendix H).

Table 4 shows a larger within-group gain when sources conflict: 2.9 percentage points, versus 0.8 when they agree. The high-uncertainty group gains 4.4 points, compared with 0.7 for the low-uncertainty group. These patterns are consistent with the motivation: a candidate's surrounding evidence is especially useful when current scores do not give a clear decision. They are conditional associations, not interventions on uncertainty or proof that most total corrections come from either group.

The same analysis exposes a cost. In the conflict group, Full34 corrects 9.1% of observations but also changes 6.2% from correct to incorrect. Context improves the net decision without making every revision safe. A long but mistaken history or mutually reinforcing source errors can still favor the wrong hypothesis.

Table 5: Direct transfer to YOLOE–ByteTrack on TAO, without target-domain fitting. LocA $= 22.424$ and AssocA $= 23.993$ throughout. These are not the OVTR tracks of Table 3.

Semantic decision	ClsA ↑	TETA ↑
Detector category	15.726	20.714
Current-frame CLIP	14.300	20.239
Score3-PM	17.200	21.206
Gated-Score-PM	17.450	21.289
Full34	**18.000**	**21.472**

4.7 Transfer and supervision scope

Cross-dataset transfer. On frozen TAO YOLOE–ByteTrack outputs, TrackRerank improves ClsA from 15.726 to 18.000 without target-domain fitting (Table 5). Its gains over Score3-PM and gated fusion are 0.800 and 0.550 points, smaller than on OVT-B; CLIP alone is worse than the detector. Additional sources need not help by themselves, and their useful relationships can shift across datasets. Absolute scores are not directly comparable to the different OVTR tracks.

Base-only positive-supervision transfer. Let $\mathcal{T}$ be the teacher-eligible training observations and c_i^T the track pseudo-label. The Base-only condition trains on $\mathcal{T}_B = \{o_{i,t} \in \mathcal{T} : c_i^T \in \mathcal{C}_B\}$, excluding every Novel positive target. Novel names remain in the upstream vocabulary and can appear as negative candidates. All conditions use the same fixed Base and Novel evaluation class sets.

Without Novel positives, Full34's Novel ClsA changes from 25.820 to 23.400 and Base ClsA from 24.800 to 24.600. A size-matched Base+Novel control retains 25.480 Novel ClsA, supporting an effect of positive-label scope beyond sample count. This evaluates Full34's supervision dependence, not its advantage over fusion under Base-only training (Appendix F).

Training coverage. Teacher matching, confidence filtering, and candidate reachability retain 499,845 of 603,678 frozen observations (82.8%; Appendix B.5). This measures coverage under the three eligibility conditions, not whether retained views are representative: confidence-based selection may still favor easier observations.

Computational scope. The shared head has 4,353 parameters and scores at most 30 candidates per observation. This small readout does not make the complete enhancement inexpensive: crop extraction and current-frame CLIP add work beyond the original tracker. We distinguish head complexity from end-to-end latency and make no throughput claim from frozen-output accuracy (Appendix I).

5 Discussion and conclusion

TrackRerank raises two-stage tracking ClsA from 13.307 to 25.300 without changing localization or identity outputs. Causal aggregation explains part of this improvement; shared-evidence controls retain another 1.050 points over score-and-mask fusion across three paired seeds. Native OVTR transfer and conditional diagnostics show where context remains useful: stronger trajectory representations still leave competing hypotheses, especially when sources conflict or are uncertain.

The method cannot recover absent candidates or repair tracking errors; contaminated history may reinforce mistakes. Class-aware association feedback remains untested. Teacher filtering can bias the shared training population, and student-history controls do not isolate the benefit of future teacher views. These limits qualify, rather than erase, the practical distinction: producing semantic hypotheses and deciding which hypothesis the evidence supports are complementary problems.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization, Chinese-to-English translation, English drafting and editing, literature lookup, refinement of the conceptual framing, feedback on experimental design and interpretation, and LaTeX and vector-figure editing. Numerical results used in this manuscript were provided by the authors, not generated as part of this writing process. Automated checks covered table arithmetic, cross-references, compilation, and rendered layout. The authors are responsible for the final scientific content and disclosure of AI use.

ETHICS STATEMENT

Open-vocabulary tracking can support video analysis, but it can also enable intrusive monitoring. Category corrections should not be interpreted as evidence about a person's identity or sensitive attributes. Errors and biases inherited from vision–language models and pseudo-label teachers may persist or be amplified through memory. Deployment requires attention to video privacy, dataset and model licenses, and the consequences of incorrect labels. The experiments concern benchmark object categories rather than sensitive personal inference.

REPRODUCIBILITY STATEMENT

Section 3 describes candidate construction, the shared scorer, supervision, and causal memory updates. Appendix B enumerates the 34 input slots, and Appendices A–C specify the evaluation and fusion controls. Appendices D–I document paired seeds, resampling, history controls, supervision scope, the OVTR adapter, and computational boundaries. Custom-split results are not mixed with official benchmark scores, and frozen-output accuracy is not presented as an end-to-end speed measurement.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. URL https://arxiv.org/abs/1702.01992.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024. URL https://arxiv.org/abs/2401.17270.

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *European Conference on Computer Vision*, pp. 436–454, 2020. URL https://arxiv.org/abs/2005.10356.

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*, 2025a. URL https://arxiv.org/abs/2503.10616.

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. URL https://arxiv.org/abs/2207.12978.

Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. OVTrack: Open-vocabulary multiple object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5567–5577, 2023. URL https://arxiv.org/abs/2304.08408.

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. URL https://arxiv.org/abs/2409.11235.

Yunhao Li, Yifan Jiao, Dan Meng, Heng Fan, and Libo Zhang. Attention to trajectory: Trajectory-aware open-vocabulary tracking. In *IEEE/CVF International Conference on Computer Vision*, pp. 14390–14398, 2025b. URL https://arxiv.org/abs/2503.08145.

Haiji Liang and Ruize Han. OVT-B: A new large-scale benchmark for open-vocabulary multi-object tracking. In *Advances in Neural Information Processing Systems*, volume 37, pp. 14849–14863, 2024. URL https://arxiv.org/abs/2410.17534.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024. URL https://arxiv.org/abs/2303.05499.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72983–73007, 2023. URL https://arxiv.org/abs/2306.09683.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 7472–7482, 2025a. URL https://arxiv.org/abs/2410.08529.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *IEEE/CVF International Conference on Computer Vision*, pp. 24591–24602, 2025. URL https://arxiv.org/abs/2503.07465.

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. URL https://arxiv.org/abs/2110.06864.