OMNIGLIMPSE: ASYMMETRIC IDENTITY MEMORY FOR ONLINE MULTI-VIEW TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view tracking extends visual coverage beyond a single camera, providing complementary observations when targets are occluded or leave the view. This raises a question: how can observations from different views be matched to the correct identities without offline scene-specific training or camera calibration? Progress toward this goal remains limited because many offline formulations assume such deployment-specific knowledge, while online adaptation can amplify association errors. We study this problem as online multi-view identity maintenance (OMIM): targets are initialized once per view, with no further identity labels provided after each initialization. To this end, we introduce OmniGlimpse, an asymmetric propose–verify framework that separates adaptive association from stable identity consolidation. At the representation level, short-term local and cross-view evidence is mapped into target-specific adaptive spaces to propose associations, whereas long-term identity distributions are maintained for verification in a canonical space defined by a frozen encoder. At the update level, short-term evidence may guide current associations, whereas, after bootstrapping, only verified observations may update persistent memory or drive adaptation. Under a unified target-specified evaluation protocol, OmniGlimpse improves over the best evaluated online baseline for each metric by 9.1 IDF1 and 22.1 CVIDF1 points on M3Track. The corresponding gains on MvMHAT are 20.3 and 26.2 points. Despite using no offline scene-specific training, it outperforms offline-trained references in IDF1 and remains competitive in cross-view association. These results support an approach to online identity maintenance that adapts to changing appearances while consolidating reliable evidence into stable identity anchors.

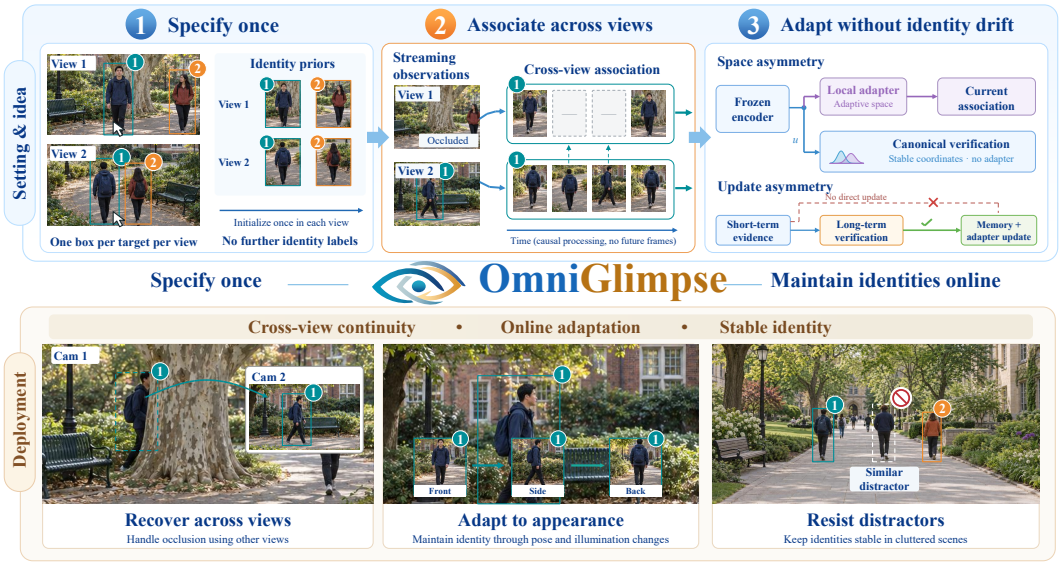

Figure 1: OmniGlimpse maintains initialized target identities across views through adaptive association and selective updates.

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.

1 INTRODUCTION

Multi-view tracking maintains identity continuity beyond a single camera and supports long-horizon perception in surveillance Xie et al. (2024), sports analysis Cui et al. (2023), and traffic understanding Liu et al. (2024). Multiple views provide complementary observations of the targets but also introduce new ambiguity in identity matching. The tracker must distinguish each target from visually similar distractors despite large viewpoint and appearance changes. Multi-view trackers associate observations through appearance, trajectories, or geometry Gan et al. (2021); Cheng et al. (2023a), with online variants updating tracks and appearance memories as observations arrive Xie et al. (2024); Specker (2024). A related line of target-specific online methods adapts the representation to changes in a designated target's appearance Ye et al. (2022; 2024).

Deploying these capabilities together in unseen multi-view scenes exposes three connected challenges. **(i) Deployment requirements.** Multi-view multi-target (MVMT) trackers often learn appearance models offline Gan et al. (2021); Fan et al. (2025), while geometry-based approaches additionally require calibration or world coordinates Specker (2024), restricting deployment when these geometric priors are unavailable. **(ii) Uncertain cross-view evidence.** Cross-view observations can compensate for unreliable local evidence, but view-dependent appearance gaps and asymmetric observation noise Amosa et al. (2023) make these cues themselves uncertain. **(iii) Self-reinforcing errors.** When recent identity assignments are reused for online adaptation, a wrong pseudo-label can enter memory, alter the representation, and bias subsequent matches. Together, these challenges make the problem more than online cross-view association: the tracker must also determine which observations are reliable enough to change its persistent identity state.

We refer to this as online multi-view identity maintenance (OMIM). Each target is initialized with a bounding box at its first valid appearance in each view. Its identity is then maintained across views without further identity labels. To address this, we introduce OMNIGLIMPSE, an asymmetric propose–verify framework that separates adaptive association from stable identity consolidation. Its design is asymmetric in both update permission and representation space. First, short-term evidence can only propose or guide associations. After initialization and bootstrapping, only observations consistent with the long-term identity distributions may update persistent memory and target-specific adaptive spaces. Second, short-term local and cross-view evidence is matched in the adapters' adaptive spaces to propose associations, while per-target identity distributions serve as anchors for verification in a canonical space defined by a frozen encoder. This asymmetric design enables OmniGlimpse to adapt to changing appearances while preserving stable verification and supervision. Experiments on M3Track Fan et al. (2025) and MvMHAT Gan et al. (2021) show improved identity maintenance over online baselines, with ablations supporting the design.

Our work makes three main contributions:

- We introduce OmniGlimpse for online multi-view identity maintenance, where targets are initialized once per view and maintained across views without further identity labels. The framework combines cross-view evidence with target-specific online adaptation to track changing appearances without offline scene-specific training or camera calibration.
- We develop an asymmetric propose–verify design with update permissions and representation spaces. Short-term evidence proposes associations. After bootstrapping, long-term verification controls which observations may update persistent memory and adaptive representations. A frozen canonical space anchors the evolving target-specific adapters.
- We evaluate OmniGlimpse on M3Track and MvMHAT under a unified target-specified protocol. It outperforms evaluated online baselines in IDF1 and CVIDF1 and remains competitive with offline-trained references. Ablations support the benefits of online adapter learning and verification-gated updates within the complete system.

2 RELATED WORK

Single-view online identity tracking. Online MOT maintains identities by associating detections with motion and appearance cues Maggiolino et al. (2023); Shim et al. (2025). ByteTrack Zhang et al. (2022b) exploits low-confidence detections to recover fragmented trajectories, while OC-SORT Cao et al. (2023) improves motion-based association under occlusion and non-linear motion.

Template updating Zhou et al. (2026) and test-time adaptation Zaveri et al. (2025); Segu et al. (2023) address changes in appearance or deployment conditions. Target-specific methods such as Following Closely Ye et al. (2022) and OCLReID Ye et al. (2024) update the target appearance model. OCLReID further uses short-term and long-term experiences to optimize the feature extractor online. Such adaptation improves responsiveness, but recent identity assignments also become supervision for subsequent updates. An incorrect assignment can therefore alter the appearance model used for future matching, allowing a temporary association error to persist through later updates.

Multi-view identity association. Multi-view tracking extends identity maintenance by using observations from cameras when local evidence becomes unreliable. One approach establishes cross-view correspondence from appearance and trajectory cues Huang et al. (2023); Yamane et al. (2025). MvMHAT Gan et al. (2021) considers tracking and cross-view association, while ReST Cheng et al. (2023a) models spatial-temporal relations through graph-based association. Benchmarks MT-MMC Woo et al. (2024) and M3Track Fan et al. (2025) extend evaluation to realistic scenes. When appearance is ambiguous, MUTR3D Zhang et al. (2022a), ADA-Track Ding et al. (2024), and OCMCTrack Specker (2024) exploit 3D queries, geometric cues, or world coordinates for cross-view matching. Although OCMCTrack performs association online, such methods may still depend on offline appearance learning Gan et al. (2021); Fan et al. (2025) or calibration and world-coordinate information Specker (2024), limiting deployment when these priors are unavailable.

Memory and Adaptation in Identity Tracking. Memory has been used in tracking and ReID systems to support identity decisions Videnović et al. (2026). Templates, tracklet states, or memories are maintained to recover targets after occlusion and stabilize appearance matching Aydemir et al. (2025). Person ReID stores identity features or prototypes, providing more stable references than the observation alone Dai et al. (2022); Cheng et al. (2023b). Online adaptation uses unlabeled observations to update caches or refine representations Karmanov et al. (2024); Segu et al. (2023). However, evidence useful for matching may not be reliable enough to supervise updates, while representation changes can make stored features inconsistent with observations. OmniGlimpse addresses these issues through an asymmetric propose–verify design that combines adaptive association with canonical-space verification to govern persistent memory updates and learning after bootstrapping.

3 METHOD

3.1 PROBLEM SETUP AND OVERVIEW

Our objective is to maintain the identities of a small set of targets of interest $\mathcal{T} = \{1, \dots, M\}$ across multiple views $\mathcal{V} = \{1, \dots, V\}$. Each target $m \in \mathcal{T}$ is initialized with a bounding box at its first valid appearance in each view. The target reference images $\mathcal{P}_m$ are crops extracted from these initialization boxes and become available only when the corresponding initialization occurs. No further frame-level identity annotations are provided after initialization in each view.

At time t, view v produces an observation set $\mathcal{O}_t^{(v)}$ containing $n_t^{(v)}$ observations. Using only observations and initialization information available up to t, the system must infer, for every target m and view v, the observation currently corresponding to that target:

$$\mathcal{O}_t^{(v)} = \{o_{t,i}^{(v)}\}_{i=1}^{n_t^{(v)}}, \qquad a_t^{(m,v)} \in \mathcal{O}_t^{(v)} \cup \{\varnothing\}. \tag{1}$$

Here, $\varnothing$ denotes that the target is absent, occluded, or cannot be identified with sufficient confidence. The objective is to maintain the identity consistency of each target across views and over time.

Given observations and target priors, OmniGlimpse maintains target identities through a closed-loop process of target-specific association, long-term verification, and online adaptation (Fig. 2).

3.2 ASSOCIATION AND VERIFICATION

Canonical and adaptive features. Online association must adapt to appearance changes, while cross-view evidence and long-term identity statistics need a consistent feature space across updates. At each time step, an initial detector–tracker produces person observations, each consisting of a bounding box, a person crop, and a local track identity. Following the part-based representation

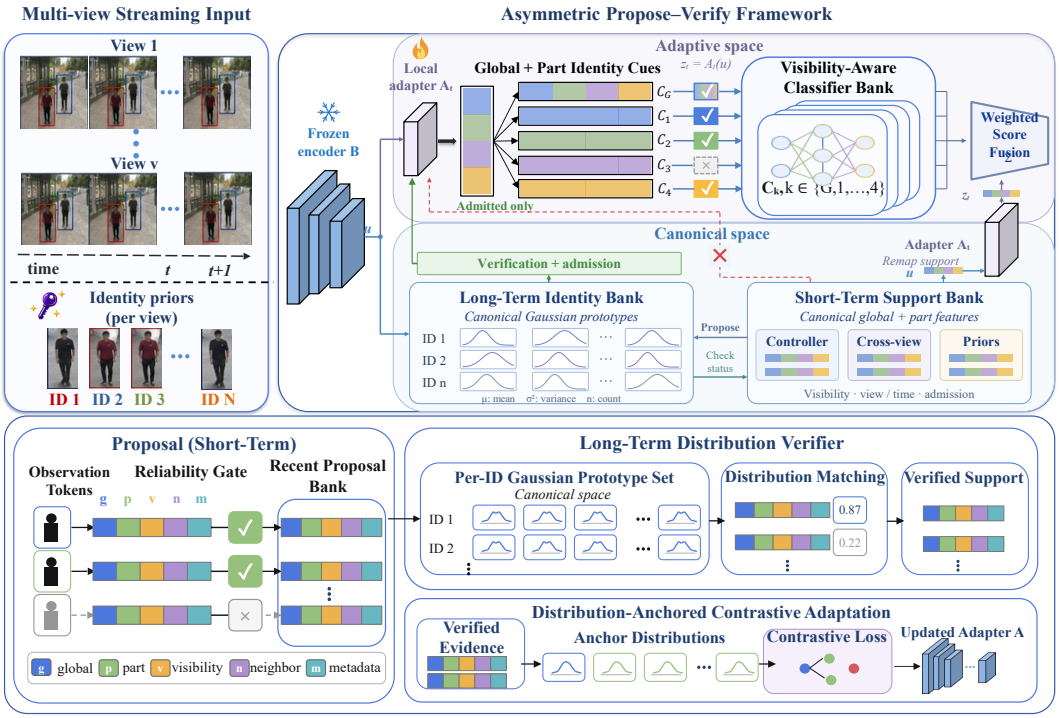

Figure 2: Overview of OmniGlimpse. Short-term evidence supports association in adaptive spaces, while long-term distributions provide canonical-space verification. After initialization and bootstrapping, only admitted observations may update persistent memory and local adapters.

design in OCLReID Ye et al. (2024), a ReID encoder B extracts a global embedding $g_0(o)$ and K part embeddings $\{g_k(o)\}_{k=1}^{K}$ from each observation o. All views use the same frozen encoder, while each target–view pair maintains a local adapter $A_t^{(m,v)}$:

$$u(o) = B(o) = \big(g_0(o), \ldots, g_K(o)\big), \qquad z_t^{(m,v)}(o) = A_t^{(m,v)}\big(u(o)\big). \tag{2}$$

Here, $u(o)$ denotes the canonical features and $z_t^{(m,v)}(o)$ the adaptive features, with $z_{t,k}^{(m,v)}(o)$ denoting the global or part cue k. The part features are accompanied by visibility scores $w_k(o) \in [0,1]$, and an orientation head predicts $\omega(o) \in \{\text{front}, \text{back}\}$ for orientation-aware scoring.

Memory banks store canonical features, and views exchange these features rather than local adapter outputs. For short-term association, both local and received cross-view features are mapped through the current local adapter. Prior references and long-term identity distributions instead use the canonical features for matching and verification. Online learning updates the local adapters without changing B, preserving the coordinate system used for communication and long-term verification. The identity distributions can still accumulate verified evidence within this fixed space.

Candidate association. For each target–view pair (m, v), candidate observations are evaluated using initialization references and adaptive ridge classifiers. The prior score measures global and part-level similarity in the canonical space, while the classifiers evaluate adaptive features under current appearance conditions. Prior evidence supports initialization and recovery; once the classifiers are available, their confidence provides the local association score $s_{\text{local}}^{(m,v)}(o)$.

The active support bank contains local target samples, temporally eligible cross-view samples, and initialization references. All support features are mapped through the current local adapter before comparison. The memory score $s_{\text{mem}}^{(m,v)}(o)$ aggregates top-L similarities, using part evidence when it meets a discriminativeness criterion and global evidence otherwise. Cross-view evidence therefore contributes through the support bank rather than a separate additive score.

Memory assistance is activated when the current binding is missing, uncertain, or contested. For these candidates, the association score combines local confidence and memory support:

$$s_{\text{assoc}}^{(m,v)}(o) = (1 - \lambda_t^{(m,v)}) s_{\text{local}}^{(m,v)}(o) + \lambda_t^{(m,v)} s_{\text{mem}}^{(m,v)}(o). \tag{3}$$

Both scores are bounded in $[0, 1]$, and $\lambda_t^{(m,v)} = \lambda_{\text{mem}} \in [0, 1]$ when assistance is active and support is available; otherwise $\lambda_t^{(m,v)} = 0$. These bounded scores are not calibrated identity probabilities.

For targets in the view, we first retain uncontested bindings that satisfy the tracking conditions. For the remaining targets, eligible target–observation pairs are ranked by s_{assoc} and assigned greedily, allowing each target and observation to participate in at most one pair. Candidates below the acceptance threshold are discarded, and overlapping assignments are suppressed. The associations remain proposals for long-term verification; selecting a candidate does not authorize memory consolidation or learning. Scoring, support eligibility, and conflict handling are detailed in Appendix A.1.

Long-term verification. The long-term memory verifies candidate associations by maintaining a consolidated identity distribution for each target. Since target appearance can be multi-modal under viewpoint changes, occlusion, and partial visibility, we represent this distribution with a small set of Gaussian prototypes. Each view v maintains a local target distribution

$$\mathcal{G}_t^{(m,v)} = \left\{ (\mu_j^{(m,v)}, \sigma_j^{2,(m,v)}, n_j^{(m,v)}) \right\}_{j=1}^{P_t^{(m,v)}}, \qquad P_t^{(m,v)} \le P, \tag{4}$$

where μ_j is the prototype mean, σ_j^2 its diagonal variance, and n_j its admitted-sample count. These distributions use normalized canonical global features and can incorporate evidence from other views. They share a feature coordinate system, not necessarily identical distribution parameters.

For a candidate observation o, let $\bar{u}(o) = g_0(o)/\|g_0(o)\|_2$. The verifier measures its agreement with the consolidated distribution using a dimension-normalized squared Mahalanobis distance:

$$d_j^{(m,v)}(\bar{u}) = \frac{1}{D} \sum_{r=1}^{D} \frac{(\bar{u}_r - \mu_{j,r}^{(m,v)})^2}{\sigma_{j,r}^{2,(m,v)} + \epsilon}, \qquad \ell_{\text{long}}^{(m,v)}(o) = \exp\left[-\frac{1}{2} \min_j d_j^{(m,v)}(\bar{u}(o)) \right]. \tag{5}$$

Here, D is the global feature dimension and $\epsilon > 0$ stabilizes the denominator. The diagonal variance accounts for uncertainty along different feature dimensions, while the minimum selects the best-matching appearance mode. The resulting consistency score is not a normalized identity probability.

Verification uses the distribution available before admitting the current observation. Once the count $n_t^{(m,v)} = \sum_j n_j^{(m,v)}$ reaches $N_{\min}$, a candidate passes if $\ell_{\text{long}}^{(m,v)}(o) \ge \tau_\ell$ and fails otherwise. Before this point, the result is unavailable, neither failed nor verified. Initialization references seed the distribution, which matures through conservative bootstrap admission (Appendix A.2).

A failed check prevents a new or recovered association from being committed. An established binding may remain an output if it still satisfies the tracking conditions, but the failed observation cannot update persistent memory or train the adapters or classifiers. Passing verification indicates consistency with the identity model; further admission rules govern updates. This separates evidence for current association from evidence allowed to modify persistent identity state.

3.3 Verification-Gated Updates

Association outputs and learning samples have different admission requirements. After assignment, an output observation may enter the short-term support bank to assist subsequent proposals, but its presence there does not authorize training or long-term consolidation. Each stored observation retains its admission status. Persistent identity memory and positive training buffers accept only evidence that satisfies the update rule below, apart from initialization and restricted bootstrap.

For a local observation o, let $V_t^{(m,v)}(o) = 1$ only when the mature long-term verifier returns pass, and zero when it returns fail or unavailable. A positive observation is admitted if

$$\begin{aligned} \mathcal{A}_{+,t}^{(m,v)}(o) = \mathbb{I}\big[\, & o = a_t^{(m,v)} \ \wedge\ V_t^{(m,v)}(o) = 1 \\ & \wedge\ s_{\text{assoc}}^{(m,v)}(o) \ge \tau_a \\ & \wedge\ \min\{\Delta_{\text{tgt}}(o), \Delta_{\text{obs}}(o)\} \ge \tau_\Delta \\ & \wedge\ T_{\text{stable}}^{(m,v)}(o) \ge \tau_T \big]. \end{aligned} \tag{6}$$

Here, $a_t^{(m,v)}$ is the final retained assignment, the two margins are computed before allocation, and T_{stable} counts consecutive observed frames with the same target–track binding. Missing observations or a changed binding reset this count. The rule controls long-memory admission and positive samples for both Ridge fitting and adapter learning; it is not a universal condition for emitting an association. Initialization labels provide initial supervision. Before maturity, conservative bootstrap can populate the long-term distribution but cannot authorize ordinary pseudo-label training. Remote evidence requires independent receiver-side admission (Appendix A.2).

All admission decisions are made before any memory or parameter update. We first fix the admitted sample set using the current assignments and pre-update distributions, then use that set for long-memory maintenance and online learning. The Gaussian parameters used as adaptation anchors are cached before these updates, so an observation cannot modify its own anchor within the same step. The updated distributions are used for verification only at the next step.

3.4 Distribution-Anchored Adaptation

The propose–verify memory maintains candidate-level identity consistency, but appearance changes can still weaken feature discriminability. We therefore refine the local adapters using admitted memory samples. Each adapter applies a residual transformation to the canonical cues,

$$z_{t,k}^{(m,v)}(o) = g_k(o) + h_{\phi_{t,k}^{(m,v)}}\big(g_k(o)\big), \qquad \widehat{z}_{t,k}^{(m,v)}(o) = \frac{z_{t,k}^{(m,v)}(o)}{\|z_{t,k}^{(m,v)}(o)\|_2}. \tag{7}$$

The residual branches initially output zero, so the normalized adaptive features start from their canonical counterparts. The frozen encoder defines the coordinates, while residual branches learn target-specific corrections. Architecture and sampling details are given in Appendix A.3.

For each target–view pair, the memory system constructs same-view and cross-view triplets (a, p, n) from admitted evidence. The anchor a and positive p represent the same target; negatives come from observations admitted for other specified targets. Unassigned detections and unresolved overlapping observations are excluded rather than assumed to be reliable negatives. All three samples are mapped through the same current local adapter, including those received from other views. Same-view triplets preserve local discrimination, while cross-view positives bring different appearances together within the receiver's adaptive space. For jointly visible cues, we use the loss

$$\ell_{\text{tri},k}(a,p,n) = \big[\cos(\widehat{z}_k(a), \widehat{z}_k(n)) - \cos(\widehat{z}_k(a), \widehat{z}_k(p)) + \delta_{\text{tri}}\big]_+. \tag{8}$$

Here, target, view, and time indices are omitted for readability. Averaging over available cues and triplets gives $\mathcal{L}_{\text{sv}}$ and $\mathcal{L}_{\text{cv}}$ for same-view and cross-view evidence.

Triplet learning separates target observations from distractors. To also regularize adaptation with accumulated identity evidence, we penalize deviations of admitted positive global features from the canonical Gaussian prototypes:

$$\begin{aligned} \mathcal{L}_{\text{anchor}} &= \frac{1}{|\mathcal{P}_+|} \sum_{o \in \mathcal{P}_+} \min_j d_{j,t}^{(m,v)}\big(\widehat{z}_{t,0}^{(m,v)}(o)\big), \\ \mathcal{L}_{\text{adapt}} &= \mathcal{L}_{\text{sv}} + \gamma_{\text{cv}}\mathcal{L}_{\text{cv}} + \beta_{\text{dist}}\mathcal{L}_{\text{anchor}}. \end{aligned} \tag{9}$$

Here, $\mathcal{P}_+$ contains the positive observations in the training batch, and $d_{j,t}^{(m,v)}$ uses the cached pre-update Gaussian parameters in Eq. 5. The residual initialization makes this penalty compare aligned coordinates at the start of adaptation; the penalty subsequently discourages departures from the canonical identity distribution. It constrains the adaptive global features, whereas inference-time verification continues to use the unmodified canonical features.

During adaptation, Gaussian parameters are detached and gradients update only the local residual branches. After an update, cached canonical features are remapped through the new adapter, and Ridge classifiers are refit using only samples with training permission. The refreshed features and classifiers support the next association step without another backbone pass on stored crops. Buffer-triggered refitting and missing-class handling are detailed in Appendix A.1.

Table 1: Selected-target results on M3Track. Offline references use target-selection information, including ground-truth-assisted trajectory selection for CrossMOT and AGW and full-sequence appearance matching for MvMHAT (Appendix B.2). Best and second-best in bold and underline.

Method	HOTA↑	DetA↑	AssA↑	IDF1↑	MOTA↑	MOTP↑	IDs↓	MDA↑	CVIDP↑	CVIDR↑	CVIDF1↑	CVMA↑
Offline-trained MVMT references												
MvMHAT	.5832	.4407	**.7723**	.6267	.4641	**.9335**	11	.1620	.2914	.1920	.2277	.3204
CrossMOT	.5970	.5540	.6445	.7945	.6628	.8182	116	**.5635**	.6150	.5499	.5786	.6335
AGW	.4748	.3928	.5773	.6347	.4630	.7936	9	.2350	.3374	.2556	.2864	.3717
Online baselines												
OCLReID	.5862	.4810	.7173	.7424	.5795	.7972	18	.3290	.4480	.3552	.3918	.4586
OC-SORT+RPF	.5071	.3904	.6614	.6508	.4603	.7960	**2**	.2678	.3771	.2892	.3237	.3984
ByteTrack	.4664	.3610	.6073	.5872	.3967	.8016	8	.2582	.3282	.2687	.2916	.3607
BoT-SORT+ReID	.4651	.3559	.6128	.5944	.4013	.7980	5	.2892	.4036	.2955	.3375	.3889
Ours (Full)	**.6438**	**.5861**	.7102	**.8333**	**.7217**	.7945	28	.5503	**.6576**	**.5792**	**.6129**	**.6677**

Table 2: Selected-target results on MvMHAT. Offline references use information for target selection: CrossMOT uses ground-truth-assisted trajectory selection; MvMHAT and AGW use full-sequence appearance matching (Appendix B.2). Best and second-best results are shown in bold and underline.

Method	HOTA↑	DetA↑	AssA↑	IDF1↑	MOTA↑	MOTP↑	IDs↓	MDA↑	CVIDP↑	CVIDR↑	CVIDF1↑	CVMA↑
Offline-trained MVMT references												
MvMHAT	.3646	.1786	.7443	.3012	.1448	.9556	**1**	.0500	.0522	.0505	.0513	.2354
CrossMOT	.5217	.4448	.6130	.6901	.4874	.8196	17	**.5626**	**.6380**	**.5819**	**.6078**	**.7190**
AGW	.5761	.4352	**.7628**	.5875	.3518	**.9993**	6	.3448	.4687	.3759	.4131	.5114
Online baselines												
OCLReID	.4644	.3060	.7133	.5292	.3290	.8403	3	.2280	.2975	.2454	.2676	.4245
OC-SORT+RPF	.3444	.2008	.6024	.3984	.2319	.7833	**1**	.0786	.1112	.0868	.0964	.2878
ByteTrack	.4516	.3112	.6651	.5130	.3104	.8448	9	.2719	.3330	.2875	.3068	.4494
BoT-SORT+ReID	.4198	.2742	.6509	.4661	.2754	.8514	4	.2114	.2764	.2278	.2481	.4007
Ours (Full)	**.5820**	**.4863**	.7135	**.7323**	**.5060**	.8168	36	.5189	.6053	.5408	.5685	.6601

4 EXPERIMENTS

Our experiments address two questions: (1) Can OmniGlimpse maintain target identities across views without identity labels after initialization? (2) Can identity verification support online adaptation while limiting the propagation of association errors? We examine its behavior in sequences.

4.1 EXPERIMENTAL SETUP

Benchmarks and protocol. M3Track Fan et al. (2025) is our main benchmark, providing synchronized RGB and infrared streams from two moving-camera views. We use only the visible-light streams and exclude extremely low-light nighttime sequences, yielding 36 single-view RGB sequences for training offline references and 38 for testing, corresponding to 19 paired-view test instances. MvMHAT Gan et al. (2021) provides 3–4 synchronized views per video group; following its original split, we use 13 groups for training offline references and 13 for testing. In the online protocol, each target is initialized with a bounding box at its first annotated appearance in each view, with no subsequent frame-level identity labels. All methods are evaluated on the same selected identities; offline target-selection procedures are detailed in Appendix B.2.

Baselines. We compare with online methods and offline-trained references. Online baselines are OCLReID Ye et al. (2024), OC-SORT+RPF Cao et al. (2023), ByteTrack Zhang et al. (2022b), and BoT-SORT+ReID Aharon et al. (2022). They receive initialization boxes but no scene-specific training and run independently per view without cross-view information exchange. Corresponding targets share an exported label only to normalize the identity namespace, providing no cross-view association or recovery cues. OCLReID and OC-SORT+RPF retain their target-identification mechanisms, whereas ByteTrack and BoT-SORT follow the tracker identity bound at initialization. Offline references are MvMHAT Gan et al. (2021), CrossMOT Hao et al. (2024), and AGW Ye et al. (2021). Since their target selection may use full-sequence information, including GT-assisted trajectory selection, we report them as references rather than identical-condition baselines.

Table 3: Ablation study on 19 non-dark M3Track test instances. The upper block shows incremental variants and the lower block full-model controls (Appendix B.3). V-Upd. denotes verification-gated updates; *Adaptive Query* uses adapted features. Best and second-best are in bold and underline.

Variant	Rel.	Cross	Adapt.	Gauss.	V-Upd.	HOTA↑	DetA↑	IDF1↑	MOTA↑	IDs↓	MDA↑	CVIDP↑	CVIDF1↑	CVMA↑
Base	–	–	–	–	–	.6240	.5548	.7999	.6801	**19**	.3995	.5602	.4881	.5395
+ Reliability Controller	√	–	–	–	–	.6386	.5802	.8246	.6989	46	.4769	.6271	.5608	.5805
+ Cross-view Assistance	√	√	–	–	–	.6389	<u>.5890</u>	.8280	.7100	45	<u>.5360</u>	<u>.6460</u>	<u>.6000</u>	.6250
+ Online Contrastive	√	√	√	–	–	**.6528**	**.5964**	**.8453**	**.7322**	83	.5255	.6271	.5850	.6398
Full w/o Rel.	–	√	√	√	√	.6178	.5482	.7938	.6710	32	.3924	.5507	.4796	.5333
w/o Online	√	√	–	√	√	.6250	.5650	.8050	.6900	<u>26</u>	.5140	.6210	.5710	.6210
w/o Verified Updates	√	√	√	√	–	.6390	.5880	.8240	.7090	57	.5190	.6220	.5830	.6410
Adaptive Query	√	√	√	√	√	.6370	.5810	.8250	.7110	44	.5330	.6410	.5960	<u>.6510</u>
Ours (Full)	√	√	√	√	√	<u>.6438</u>	.5861	<u>.8333</u>	<u>.7217</u>	28	**.5503**	**.6576**	**.6129**	**.6677**

Metrics. Single-view over-time identity maintenance is evaluated by aggregating results independently computed in each view, including HOTA, DetA, AssA, IDF1, MOTA, MOTP, and identity switches (IDs) Bernardin & Stiefelhagen (2008); Ristani et al. (2016); Luiten et al. (2021). IDs denotes TrackEval's standard IDSW count. Cross-view identity association is evaluated using CVIDP, CVIDR, CVIDF1, MDA, and CVMA Gan et al. (2021); Fan et al. (2025). Initialization fixes the selected targets and output labels; metric matching follows the evaluation rules in Appendix B.2.

Implementation. We use YOLOX Ge et al. (2021) and BoT-SORT Aharon et al. (2022) to generate frame-level observations for each view. Following OCLReID Ye et al. (2024), observations are encoded by a ResNet-18 based ReID backbone. Association, memory admission, and adapter fitting are detailed in Appendix A, with numerical settings in Appendix B.1.

4.2 Overall Results

M3Track. Table 1 reports the single-view and cross-view metrics on the M3Track Fan et al. (2025) benchmark. Compared with online baselines such as OCLReID, BoT-SORT, ByteTrack, and OC-SORT+RPF, our method achieves higher IDF1 and CVIDF1. In particular, it increases IDF1 from .7424 to .8333 and CVIDF1 from .3918 to .6129 over OCLReID, the strongest online baseline on both metrics. These gains indicate improved identity continuity within views and consistency across views under the reported evaluation measures. We further compare with offline-trained MVMT references. Relative to CrossMOT, our method improves IDF1 from .7945 to .8333 and CVIDF1 from .5786 to .6129, while its MDA is lower (.5503 versus .5635). AssA, MOTP, and IDs are not the best in the table. These results show competitive performance against offline references, whose additional target-selection information is detailed in Appendix B.2.

MvMHAT. Table 2 reports the comparison with online and offline methods on MvMHAT. The gains over online baselines extend to this benchmark: IDF1 increases from .5292 for OCLReID to .7323, while CVIDF1 increases from .3068 for ByteTrack to .5685. These are the strongest online baselines for the metrics. Compared with CrossMOT, our method achieves higher IDF1 (.7323 versus .6901), but lower CVIDF1 (.5685 versus .6078). Its cross-view metrics remain below CrossMOT, and it incurs more identity switches. Together, the benchmarks show IDF1 and CVIDF1 gains over the online baselines, while the comparison with offline references remains mixed.

4.3 Ablation Study

Table 3 examines component combinations and full-model controls (Appendix B.3). Reliability control improves IDF1 and CVIDF1, while cross-view assistance raises CVIDF1 from .5608 to .6000. Adding online contrastive adaptation improves HOTA and IDF1 but lowers CVIDF1 and increases IDs. Full adds Gaussian verification with its write and learning constraints, raising CVIDF1 to .6129 and reducing IDs from 83 to 28, at the cost of lower IDF1 than the Online Contrastive variant.

The controls show that online adapter learning remains useful within the system. Disabling adapter optimization while retaining Ridge refitting lowers IDF1 and CVIDF1 by 2.83 and 4.19 points. With adapter learning retained, removing verification-gated updates lowers CVIDF1 by 2.99 points and

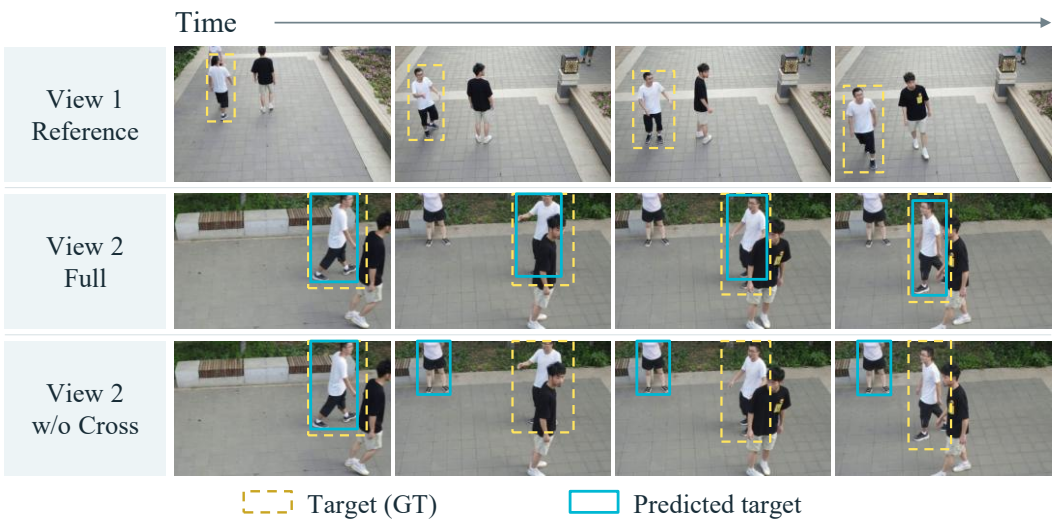

Figure 3: Identity maintenance in the M3Track garden scene. Columns show synchronized observations; the first row shows the reference view. Yellow dashed boxes mark the target, and cyan boxes mark confirmed predictions.

increases IDs from 28 to 57. These comparisons support learning from new observations while restricting which observations may update persistent memory and supervise adaptation.

Using adapted rather than canonical verification queries lowers CVIDF1 by 1.69 points and increases IDs from 28 to 44. This supports retaining canonical verification queries while adapting representations for association. Because Gaussian statistics remain canonical in this control, the comparison addresses query choice under canonical statistics, not the necessity of the entire two-space design.

4.4 Qualitative Analysis

Figure 3 shows an identity maintenance example on M3Track. As the target moves near other pedestrians, the configuration without cross-view assistance assigns the target identity to another person wearing a similar white shirt. This incorrect association persists through the displayed observations, while the selected target remains visible. With cross-view assistance, the controller maintains the selected identity through the same interval. The other view provides complementary observations of the target, and the recovery step records the use of cross-view memory. This example illustrates cross-view assistance, not the isolated effects of long-term verification or online adaptation. Appendix B.4 compares incorrect assignments and interrupted outputs in the wood scene.

5 Conclusion

In this paper, we study online multi-view identity maintenance, where targets are specified once and tracked across views without further identity supervision. To address this problem, we propose OmniGlimpse, which uses asymmetric memory roles to separate association proposals from verified updates, and asymmetric representation spaces to adapt appearance features while keeping identity verification in a stable shared space. Together, these choices support tracking and adaptation across views with only initialization supervision, without scene-specific training or camera calibration. Experiments on M3Track and MvMHAT show stronger identity maintenance than the evaluated online baselines and competitive performance against offline-trained references. The current evaluation focuses on a small set of designated targets with per-view initialization; scaling the framework to denser target sets and weaker initialization remains an important direction. We will release the code to support reproducibility and future research. We hope OmniGlimpse will advance multi-view tracking toward sustained identity maintenance in changing environments.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and to support literature retrieval and discovery during the preparation of this manuscript. All retrieved materials were independently checked by the authors, and all AI-assisted text was reviewed and revised to ensure that it accurately reflects the authors' intended meaning and the reported work. The authors take full responsibility for the final manuscript, including its technical content, claims, citations, and conclusions.

REPRODUCIBILITY STATEMENT

Implementation details are provided in Appendix A, with model configurations and online-update settings in Appendix B.1. Evaluation protocols and baseline adaptations are described in Appendix B.2, and ablation controls are specified in Appendix B.3. We will release the code and experiment configurations to facilitate reproducibility.

REFERENCES

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

Temitope Ibrahim Amosa, Patrick Sebastian, Lila Iznita Izhar, Oladimeji Ibrahim, Lukman Shehu Ayinla, Abdulrahman Abdullah Bahashwan, Abubakar Bala, and Yau Alhaji Samaila. Multi-camera multi-object tracking: A review of current trends and future advances. *Neurocomputing*, 552:126558, 2023.

Görkay Aydemir, Xiongyi Cai, Weidi Xie, and Fatma Guney. Track-on: Transformer-based online point tracking with memory. In *International Conference on Learning Representations*, volume 2025, pp. 57112–57134, 2025.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, 2008.

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.

Cheng-Che Cheng, Min-Xuan Qiu, Chen-Kuo Chiang, and Shang-Hong Lai. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10051–10060, 2023a.

De Cheng, Xiaolong Wang, Nannan Wang, Zhen Wang, Xiaoyu Wang, and Xinbo Gao. Cross-modality person re-identification with memory-based contrastive embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 425–432, 2023b. doi: 10.1609/aaai.v37i1.25116.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9921–9931, 2023.

Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian conference on computer vision*, pp. 1142–1160, 2022.

Shuxiao Ding, Lukas Schneider, Marius Cordts, and Juergen Gall. Ada-track: End-to-end multi-camera 3d multi-object tracking with alternating detection and association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15184–15194, 2024.

Huijie Fan, Yu Qiao, Yihao Zhen, Tinghui Zhao, Baojie Fan, and Qiang Wang. All-day multi-camera multi-target tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16892–16901, 2025.

Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 282–290, 2021.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Shengyu Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *International Journal of Computer Vision*, 132(4):1075–1090, 2024.

Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, Pyong-Kun Kim, Kyoungoh Lee, Kwangju Kim, Samartha Ramkumar, Chaitanya Mullapudi, In-Su Jang, Chung-I Huang, et al. Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency id re-assignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5239–5249, 2023.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Lihao Liu, Yanqi Cheng, Zhongying Deng, Shujun Wang, Dongdong Chen, Xiaowei Hu, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica Aviles-Rivero. Trafficmot: a challenging dataset for multi-object tracking in complex traffic scenarios. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1265–1273, 2024.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3025–3029. IEEE, 2023.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pp. 17–35. Springer, 2016.

Mattia Segu, Bernt Schiele, and Fisher Yu. DARTH: Holistic test-time adaptation for multiple object tracking. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9683–9693. IEEE, 2023.

Kyujin Shim, Kangwook Ko, Yujin Yang, and Changick Kim. Focusing on tracks for online multi-object tracking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11687–11696. IEEE, 2025.

Andreas Specker. Ocmctrack: Online multi-target multi-camera tracking with corrective matching cascade. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7236–7244, 2024.

Jovana Videnović, Matej Kristan, and Alan Lukežič. Distractor-aware memory-based visual object tracking. *International Journal of Computer Vision*, 134(5):211, 2026.

Sanghyun Woo, Kwanyong Park, Inkyu Shin, Myungchul Kim, and In So Kweon. Mtmmc: a large-scale real-world multi-modal camera tracking benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22335–22346, 2024.

Zhenyu Xie, Zelin Ni, Wenjie Yang, Yuang Zhang, Yihang Chen, Yang Zhang, and Xiao Ma. A robust online multi-camera people tracking system with geometric consistency and state-aware re-id correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7007–7016, 2024.

Taiga Yamane, Ryo Masumura, Satoshi Suzuki, and Shota Orihashi. Mvtrajecter: Multi-view pedestrian tracking with trajectory motion cost and trajectory appearance cost. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13270–13280, 2025.

H. Ye, J. Zhao, Y. Zhan, W. Chen, L. He, and H. Zhang. Person re-identification for robot person following with online continual learning. *IEEE Robotics and Automation Letters*, 2024.

Hanjing Ye, Jieting Zhao, Yaling Pan, Weinan Chen, and Hong Zhang. Following closely: A robust monocular person following system for mobile robot. *arXiv preprint arXiv:2204.10540*, 2022.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Ram Zaveri, Shivang Patel, Yu Gu, and Gianfranco Doretto. Improving accuracy and generalization for efficient visual tracking. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 9450–9460, 2025.

Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4537–4546, 2022a.

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. Springer, 2022b.

Jinglin Zhou, Tianyang Xu, Xuefeng Zhu, Xiao-Jun Wu, and Josef Kittler. SPD-Updater: Symmetric positive definite manifold geometry based temporal updating for visual object tracking. *Neural Networks*, 193:107985, 2026. doi: 10.1016/j.neunet.2025.107985.

A Method Details

A.1 Candidate Association Details

Prior identity reference. While the classifiers adapt online, initialization provides an identity reference before pseudo-labels are available. For target m in view v, let $\mathcal{P}^{\text{init}}_{m,v}$ contain the crops supplied at that view's initialization. A reference is constructed for each available cue:

$$r_k^{(m,v)} = \frac{1}{|\mathcal{P}^{\text{init}}_{m,v,k}|} \sum_{p \in \mathcal{P}^{\text{init}}_{m,v,k}} g_k(p), \qquad k \in \{0, \dots, K\}, \tag{10}$$

where $\mathcal{P}^{\text{init}}_{m,v,k}$ includes only initialization crops in which cue k is available. The global cue is always available. These canonical references remain fixed after the corresponding initialization; reference crops from another view can enter the support bank only after that view's initialization has occurred and its message has arrived. The prior score combines global cosine similarity with the mean cosine similarity over jointly available parts, using weights 0.7 and 0.3, respectively. When no part is jointly available, only the global similarity is used. The resulting score is clipped to $[0, 1]$.

Orientation-aware classifiers. Each target–view unit maintains orientation-aware ridge classifiers $\{c^{(m,v)}_{\omega,k}\}$, where $\omega \in \{\text{front}, \text{back}\}$ and $k = 0$ denotes the global cue. Each classifier is trained as a binary target/non-target regressor with labels one and zero. For observation o, only trained classifiers matching its predicted orientation and visible cues are evaluated. Let $\mathcal{K}^{(m,v)}_t(o)$ denote this set. Their mean response is

$$C^{(m,v)}_t(o) = \frac{1}{|\mathcal{K}^{(m,v)}_t(o)|} \sum_{k \in \mathcal{K}^{(m,v)}_t(o)} c^{(m,v)}_{\omega(o),k}\big(\hat{z}^{(m,v)}_{t,k}(o)\big). \tag{11}$$

The local association score is $\text{clip}(C^{(m,v)}_t(o), 0, 1)$. If no applicable classifier is trained, the prior score is used instead. This clipping gives a bounded ranking score, not a calibrated posterior probability. All target-specific classifiers use the same feature-normalization convention, ridge regularization, and equal total weights for positive and negative training samples. Fusion weights and

acceptance thresholds are shared across targets rather than fitted independently to each target. These common settings constrain score construction but do not guarantee cross-target calibration. The classifier buffer stores canonical features and their admitted labels; it does not treat every emitted association as a training sample. Refitting uses features mapped by the current local adapter.

For each orientation/cue classifier, let $\mathcal{B}_+$ and $\mathcal{B}_-$ contain admitted positive and negative samples with the required orientation and visible cue. Negatives carry an admitted identity different from the target, as specified in Appendix A.3. With x_i the normalized feature mapped through the current target-specific adapter, fitting solves

$$\min_{w,b} \frac{1}{2|\mathcal{B}_+|} \sum_{i\in\mathcal{B}_+} (w^\top x_i + b - 1)^2 + \frac{1}{2|\mathcal{B}_-|} \sum_{i\in\mathcal{B}_-} (w^\top x_i + b)^2 + \lambda_R \|w\|_2^2, \qquad \lambda_R > 0. \tag{12}$$

The intercept is unpenalized, and each class receives half the total sample weight. The response is $c(x) = w^\top x + b$. A classifier is fitted only when both sets are nonempty. Otherwise it is disabled and excluded from $\mathcal{K}_t^{(m,v)}(o)$; if that set is empty, scoring falls back to the prior. Initialization labels can supply either class, but bootstrap pseudo-labels cannot. A single-target deployment without admitted negatives therefore uses prior and memory scores rather than a one-class Ridge fit.

An affected classifier is refit whenever its admitted buffer changes, including insertion, eviction, or a change in training eligibility, even if the adapter does not change. An adapter update invalidates every affected orientation/cue model: its training features are remapped from canonical storage and it is refit before the next association step. Missing either class disables the model rather than retaining stale weights. If neither its buffer nor its adapter changes, the fitted model is reused. Newly available initialization labels are incorporated before scoring; other admitted samples affect subsequent steps only.

Support bank and temporal eligibility. The support bank combines recent local target observations, received target observations from other views, and available initialization references. Each sample retains its source view, timestamp, local track identity, canonical global/part features, and visibility mask. Remote samples are eligible only if their timestamps do not exceed the current processing time and their age is within a synchronization window Δ_{sync}. Features and masks are inserted and removed together. A missing part mask makes the sample unavailable for part comparison, but does not prevent global comparison. Tentative short-term support may help propose an association without being admitted for representation learning.

Global and part support. For brevity, suppress (m, v, t) and let $z_k(o)$ and $z_k(b)$ denote the query and support features produced by the same current local adapter. Global pair similarity is $h_g(o, b) = \cos(z_0(o), z_0(b))$. For each support sample b, define the jointly visible parts $\mathcal{J}(o, b) = \{k \in \{1, \dots, K\} : w_k(o) > 0,\ w_k(b) > 0\}$ and

$$h_p(o, b) = \frac{1}{|\mathcal{J}(o, b)|} \sum_{k\in\mathcal{J}(o,b)} \cos\big(z_k(o), z_k(b)\big). \tag{13}$$

Samples with $\mathcal{J}(o, b) = \varnothing$ are excluded from part aggregation. The global and part support scores, $G(o)$ and $P(o)$, are the means of the largest L valid pair similarities in their respective channels. Fewer than L samples use all available values; an empty channel returns zero.

Part scores are used where available when a discriminativeness criterion is met: among all current observations for the target, at least two have positive part support, the largest part score is at least τ_p, and its margin over the second largest is at least δ_p. For candidates with positive part support under this condition, $s_{\mathrm{mem}}(o) = \mathrm{clip}(P(o), 0, 1)$; otherwise it is $\mathrm{clip}(G(o), 0, 1)$. Thus missing or ambiguous part evidence falls back to global support. An empty support bank disables memory assistance rather than treating a zero memory score as negative evidence.

Binding retention and conflict handling. A current binding is eligible for retention when its observation remains present and its local score exceeds a tracking threshold τ_{keep}. It is retained only if no other target's existing binding meets the same retention condition for that observation. Such contested bindings enter reassignment rather than being resolved by target iteration order; a newly scored candidate alone does not count as a competing retained binding. Targets without a retained

binding activate memory assistance. Candidate pairs must exceed an association threshold τ_{assoc}. When entering recovery after losing a binding, they must also satisfy a prior-support threshold τ_{prior}. A target's verified initialization observation takes precedence when initialization occurs.

Eligible pairs are sorted by association score. Each pair is selected only if both its target and observation remain unassigned; ties are resolved by target index and then local track index. After allocation, pairs whose boxes exceed an overlap threshold τ_{IoU} cannot both remain: verified initialization pairs have priority, followed by retained bindings and then association score. A suppressed pair does not trigger another search in the same frame. These operations generate provisional associations, not permission to train on the selected observations. They do not impose universal target- and track-margin rejection thresholds.

A.2 Long-Term Distribution Details

Initialization and bootstrap. Each local distribution is seeded with the normalized global features of initialization crops available at that view. New prototypes start with diagonal variance σ^2_{init} within the clipping interval $[\sigma^2_{\min}, \sigma^2_{\max}]$. Before the total count reaches $N_{\min}$, lack of a mature distribution does not veto an association. Bootstrap observations must instead satisfy prior support, association confidence, competing-target and competing-observation margins, and consecutive binding stability. They may populate the long-term distribution but are not treated as verified adaptation samples. Thus, bootstrap admission establishes the reference needed for verification without authorizing ordinary pseudo-label training before that reference is mature.

Admission conditions. Both margins are computed from a pre-allocation score matrix $Q_t^{(v)}(m,i) = s_{\text{assoc}}^{(m,v)}(o_{t,i}^{(v)})$. Its rows include all targets already initialized in view v, including targets with retained bindings, and its columns include all current observations. Retention status and memory-assistance weights are determined first; all pair scores are then computed before acceptance filtering, greedy allocation, overlap suppression, and long-term verification. For a target m and observation o_i, the margins are

$$\begin{aligned} \Delta_{\text{tgt}}(m,i) &= Q_t^{(v)}(m,i) - \max_{m' \neq m} Q_t^{(v)}(m',i), \\ \Delta_{\text{obs}}(m,i) &= Q_t^{(v)}(m,i) - \max_{i' \neq i} Q_t^{(v)}(m,i'). \end{aligned} \tag{14}$$

If a competitor set is empty, that margin imposes no restriction. The matrix is not reduced after assignment, so selecting one target–observation pair does not erase its competitors from the admission test. Long-memory admission requires association confidence at least τ_a, both margins at least τ_Δ, and the same target–track binding for at least τ_T consecutive observed frames. Missing observations or a changed binding reset this count. Bootstrap additionally requires prior support at least τ_{boot}; after maturity, admission instead requires passing long-term verification. These are write conditions, not universal requirements for emitting a current association.

Prototype updates. An admitted canonical global feature x is assigned to the prototype with maximum cosine similarity. If this similarity is below τ_{new} and capacity remains, a new prototype is initialized with mean x, variance σ^2_{init}, and count one. Otherwise the nearest prototype is updated without replacing another mode. After bank maturity, a new prototype can be created only from evidence that already passes the existing verifier. Maturity is assessed at bank level: a newly created prototype participates in matching even before its individual count reaches $N_{\min}$. This permits additional modes within the accepted region but does not ensure recovery from abrupt appearance changes outside it. Let μ_{old} and σ^2_{old} be its parameters before the update and $\Delta = x - \mu_{\text{old}}$. The update is

$$\begin{aligned} \mu_{\text{new}} &= (1-\alpha)\mu_{\text{old}} + \alpha x, \\ \sigma^2_{\text{new}} &= \text{clip}_{[\sigma^2_{\min}, \sigma^2_{\max}]} \left((1-\alpha)\sigma^2_{\text{old}} + \alpha(1-\alpha)\Delta^{\odot 2} \right). \end{aligned} \tag{15}$$

The factor $(1-\alpha)$ in the residual term accounts for the moving mean. The prototype count is incremented by one for each newly admitted sample; it records accumulated support rather than the effective sample size of the exponential moving average. Variance clipping prevents degenerate uncertainty estimates. Matching uses minimum Mahalanobis distance, whereas update routing uses maximum cosine similarity.

Cross-view evidence. Each process maintains its own target distributions. A received long-memory event contains the canonical feature, source view, timestamp, and an observation identifier for deduplication. Sender-side admission does not automatically authorize a receiver-side update. At a mature receiver, an event may update the distribution only when the receiver has a local association for the same target at the current step that satisfies its own verification, confidence, margin, and stability conditions. The received feature must independently pass the receiver's pre-update verifier and meet the temporal eligibility rule in Appendix A.1. Stability and margins are measured on the receiver's local association, not inferred from the remote track identifier. Before maturity, the receiver instead requires a local association satisfying its bootstrap conditions, and the received feature must independently satisfy the bootstrap prior-support threshold in the canonical space. All checks use the receiver's statistics at the start of the step. Without qualifying local evidence, remote samples can support candidate association but cannot update the receiver's long-term distribution. Thus, cross-view evidence cannot directly mature that distribution while the target is unobserved locally. Repeated events are not counted twice. Local distributions can consequently differ while remaining comparable in the common canonical space.

Verification outcomes. A mature bank returns pass or fail according to the consistency threshold; an immature bank returns unavailable. Failed new or recovered proposals are removed without searching for a replacement in the same frame. A retained binding can remain visible in the output despite a failed check, but this does not admit its observation to long memory, Ridge training, or adapter training. An unavailable result does not by itself authorize learning; initialization labels and the restricted bootstrap procedure provide the starting evidence. All checks use pre-update statistics, so an observation cannot first change the reference against which it is verified.

A.3 Online Adaptation Details

Residual architecture. For each target–view pair and cue k, the residual branch has two linear layers with a ReLU nonlinearity:

$$h_{\phi_k}(g) = W_{2,k}\,\mathrm{ReLU}(W_{1,k}g + b_{1,k}) + b_{2,k}. \tag{16}$$

The input and output dimensions equal the canonical cue dimension D_k, and the hidden width is $\max(1, \lfloor D_k/4 \rfloor)$. The first layer uses a standard variance-scaled initialization; the second-layer weights and bias are zero-initialized. Thus the residual output is initially zero without making both layers identically zero. The output is normalized before similarity computation, classifier fitting, and distribution anchoring. The canonical encoder, visibility estimator, and orientation predictor remain frozen. Adapters are local to each target–view pair and their parameters are not exchanged.

Sample permissions and negatives. Initialization observations carry their supplied labels. Other positive samples require admission under Sec. 3.3, including receiver-side checks for cross-view samples. Training eligibility is stored separately from presence in a short-term proposal bank. A negative for target m must carry an admitted identity $m' \neq m$; not being selected for m is insufficient. Observations whose identity remains unresolved after conflict or overlap handling are excluded. Among eligible negatives, samples most similar to the anchor under the current target-m adapter are selected for hard-negative training. This procedure does not treat an arbitrary co-present detection as a negative. When no eligible negative is available, the contrastive term is skipped rather than assigning an unverified negative label.

Cue aggregation and triplet construction. The positive pool combines training-eligible observations in the short-term support bank with the Ridge positive buffer, deduplicated by source view and observation identifier. Tentative support without training permission is excluded. Within each same-view or cross-view allocation, valid anchor–positive pairs are sampled uniformly without replacement, subject to the time-separation requirement. An observation cannot serve as both anchor and positive in a pair. Negatives follow the admitted hard-negative rule above.

All features in a triplet are recomputed from their cached canonical features through the same local adapter. The cue set is

$$\mathcal{K}(a, p, n) = \{0\} \cup \{k \in \{1, \ldots, K\} : w_k(a) > 0,\ w_k(p) > 0,\ w_k(n) > 0\}. \tag{17}$$

Missing part masks exclude the corresponding parts. The triplet loss averages cue-level losses:

$$\ell_{\mathrm{tri}}(a,p,n) = \frac{1}{|\mathcal{K}(a,p,n)|} \sum_{k \in \mathcal{K}(a,p,n)} \ell_{\mathrm{tri},k}(a,p,n). \tag{18}$$

For the receiver view v, same-view triplets use a local anchor and a distinct local same-target positive; cross-view triplets use a local anchor and an admitted same-target positive from another view. Negatives may come from either source if their receiver-side training permission is valid. The losses are

$$\mathcal{L}_{\mathrm{sv}} = \frac{1}{|\mathcal{T}_{\mathrm{sv}}|} \sum_{\mathcal{T}_{\mathrm{sv}}} \ell_{\mathrm{tri}}(a,p,n), \qquad \mathcal{L}_{\mathrm{cv}} = \frac{1}{|\mathcal{T}_{\mathrm{cv}}|} \sum_{\mathcal{T}_{\mathrm{cv}}} \ell_{\mathrm{tri}}(a,p,n). \tag{19}$$

An empty triplet set contributes zero. The two sets are averaged independently before applying γ_{cv}, so sample counts do not implicitly determine the relative loss weight.

Anchor eligibility and update order. The set $\mathcal{P}_+$ comprises distinct admitted target observations used as anchors or positives in the batch; negative observations are never pulled toward the target's Gaussian distribution. If no valid triplet exists, no contrastive or anchor-only adapter step is performed. Adapter optimization begins only when the receiver's target distribution is mature. The distribution snapshot taken before the current step's memory writes is held fixed throughout all optimizer steps for that batch. The optimizer differentiates the anchor loss through the adaptive global feature but not through Gaussian statistics or the canonical encoder. Part features receive the contrastive loss without a separate Gaussian penalty.

After optimizing the adapter, the receiver invalidates adaptive-feature caches and recomputes them from stored canonical features. Orientation/cue Ridge classifiers with both admitted classes are refit in the new adaptive space using Appendix A.1; missing-class models are disabled, and old weights are never used with new features. Buffer changes also trigger refitting without an adapter update. Available classifiers are aggregated; if none applies, scoring uses the prior fallback. This refresh is completed before the next association step. No backbone forward pass on historical images and no synchronization of adapter parameters are required.

B Experimental Details

B.1 Implementation Settings

The canonical encoder uses the trained ResNet-18 checkpoint from OCLReID Ye et al. (2024).

Tables 4 and 5 summarize the configuration of the complete model. Ablation-specific changes are described in Appendix B.3. The canonical encoder and its feature output path, visibility estimator, and orientation predictor remain frozen; only the local residual adapters are optimized. Their architecture and initialization are specified above. Offline references are trained using two NVIDIA RTX A6000 GPUs, while evaluation uses a single RTX A6000 GPU.

Buffer and update conventions. Initialization references are retained separately from the bounded short support bank. Ridge buffers use FIFO eviction and retain the visibility information of each observation. Their capacity is per class, not per cue; each orientation/cue classifier uses the applicable subset. Contrastive negatives must have an admitted identity different from the target. Remote observations are used only after receipt, with timestamps no later than the current processing time and ages within Δ_{sync}.

The optimization interval counts newly admitted local observations with training permission, not elapsed video frames or received messages. Each eligible batch contains at least four valid triplets. The sampling proportions specify the intended allocation; unavailable triplets are not duplicated to fill a batch. The two loss groups are averaged separately as in Eq. 19. Long-memory writes and pseudo-label training use the same confidence, margin, and stability requirements after bank maturity. Bootstrap observations only establish the reference and do not count toward adapter-training eligibility. Adapter updates use pre-write Gaussian statistics, followed by feature remapping and Ridge refitting before the next association step.

Table 4: Association, verification, and admission settings. Capacities count observations or prototypes, not individual feature coordinates.

Setting	Value
Detector / canonical backbone	YOLOX-L / ResNet-18
Canonical checkpoint source	OCLReID-trained ResNet-18
Number of part cues K	4
Global and part dimensions D_k	512
Detection score / NMS IoU threshold	0.60 / 0.35
Support aggregation L	3
Active memory fusion weight λ_{mem}	0.80
Prior global / mean-part weights	0.70 / 0.30
Part-support threshold τ_p / margin δ_p	0.80 / 0.005
Binding retention τ_{keep}	0.60
Association acceptance τ_{assoc}	0.70
Recovery prior support τ_{prior}	0.72
Assignment overlap threshold τ_{IoU}	0.70
Maximum remote-observation age Δ_{sync}	0.10 s
Short support capacity per target and source view	64 observations
Gaussian capacity per target–view pair	4 prototypes
Prototype update rate α	0.05
Initial diagonal variance σ^2_{init}	$1/512$
Variance bounds $[\sigma^2_{\mathrm{min}}, \sigma^2_{\mathrm{max}}]$	$[10^{-3}, 1]$
Distance denominator stabilizer ϵ	10^{-6}
New-prototype cosine threshold τ_{new}	0.75
Bank maturity N_{min}	10 observations
Verification score threshold τ_ℓ	0.50
Admission score τ_a / competing margins τ_Δ	0.90 / 0.15
Consecutive binding stability τ_T	10 observed frames
Bootstrap prior-support threshold τ_{boot}	0.80

Table 5: Local adapter learning and classifier settings. Ridge regularization refers to the class-normalized objective in Eq. 12.

Setting	Value
Adapter hidden width	128
Optimizer	Adam
Learning rate / weight decay	10^{-4} / 10^{-4}
Optimization interval per target–view pair	5 newly admitted local observations
Optimizer steps per eligible batch	1
Maximum / minimum triplets per batch	16 / 4
Same-view / cross-view sampling proportion	0.5 / 0.5
Minimum same-view positive time separation	0.10 s
Triplet margin	0.20
Cross-view loss weight γ_{cv}	0.50
Distribution-anchor weight β_{dist}	0.004
Ridge regularization λ_R	0.01
Ridge buffer per target–view pair and class	32 observations
Contrastive negative pool per target–view pair	256 observations

B.2 Evaluation Protocol and Baseline Adaptation

Target selection and initialization. Each scene contains two designated target identities, with a fixed mapping to dataset identities shared across views and methods. M3Track reads this mapping from the target metadata; MvMHAT uses dataset identities 1 and 2. The four online baselines receive a bounding box at each target's first annotated occurrence in each view's test segment. Visible target frames remain in evaluation even if initialization fails. Only selected identities are retained in the evaluation ground truth; missed targets and false outputs remain subject to the evaluator's matching rules.

Metric computation and aggregation. Tracking metrics are computed with TrackEval, treating each view as a separate sequence and retaining selected-target GT. The selected-target exporter uses the initialized target label, mapped to its dataset ID, as the prediction ID rather than the underlying local track ID. This fixes exported labels but does not constrain metric matching by label equality: CLEAR uses spatial matching with its temporal preference, and Identity uses global trajectory assignment. Dataset-level results combine sequence statistics rather than averaging scene scores; IDs are summed, and MOTP measures matched-box IoU. Cross-view scores are computed for synchronized view pairs, then averaged over scenes on M3Track and over all scene–view pairs on MvMHAT. Thus, four-view scenes contribute six pairs and three-view scenes three pairs. Both benchmark wrappers use the same pairwise evaluation function described below. These metrics assess tracking and association consistency rather than enforcing the initialized identity name at every match.

Identity switches. The IDs columns use the CLEAR IDSW field returned by TrackEval, also used in MOTA. An identity switch is counted when a GT identity is matched to a prediction ID different from its previous matched prediction ID. A gap does not by itself reset that previous ID. This is not a count of wrong-identity frames or the separate selected-target diagnostic switch counter.

Cross-view matching and empty frames. Within each view, each GT box is matched to the prediction with the highest IoU among candidates with IoU at least 0.5; if no such prediction exists, the GT is treated as unmatched. The selected prediction label is then used to construct the GT-to-prediction identity map. A shared GT identity is cross-view correct when its matched prediction labels agree across the two views; equality to the GT identity label is not required.

Let G_t be the number of GT identities common to both views, R_t the number of distinct positive prediction labels shared by the two GT-to-prediction maps, and C_t the number of cross-view-correct GT identities. Let Q_t count GT entries in view a whose mapped prediction label occurs in both maps. Frame precision and recall are C_t/Q_t and C_t/G_t, respectively. Frame MDA is

$$\frac{C_t}{G_t + \max(G_t - C_t, 0) + \max(R_t - C_t, 0)}.$$

Zero-denominator values contribute zero. If either view has no GT, frame precision, recall, and MDA also contribute zero; such frames remain in the averaging denominator. With GT in both views but no predictions in one view, $C_t = R_t = Q_t = 0$, so these three frame scores are zero. CVIDP, CVIDR, and MDA are frame means; pair CVIDF1 is the harmonic mean of pair CVIDP and CVIDR before averaging across pairs.

For CVMA, let $E_t = \max(R_t - C_t, 0)$ and $M_t = 2\max(G_t - R_t, 0)$. Let F_t count distinct predicted labels present in both views but absent from one view's GT-to-prediction map, summed over the two directions. With N_t the total number of GT boxes in both views,

$$\mathrm{CVMA} = 1 - \frac{\sum_t (2E_t + F_t + M_t)}{\sum_t N_t}.$$

When either view has no GT, E_t, F_t, M_t are zero, but available GT boxes still contribute to N_t. A zero total GT-box count gives CVMA zero.

Baseline adaptation. The online baselines run independently for each target and view, without exchanging observations, features, trajectories, or association decisions across cameras. Since these baselines do not perform cross-view association, we assign outputs for the same initialized target a

common label across views. This supplies cross-view label correspondence without requiring the baselines to infer it. Their cross-view scores are then computed from these aligned outputs using the same evaluation rules.

B.3 Full-Model Ablation Controls

The full-model controls retain the same observations, target initialization, association procedure, and non-ablated settings. They differ from the cumulative component rows because each changes one specified aspect of Full. Reliability control denotes the non-Gaussian admission conditions: the confidence threshold, both competing margins, and consecutive binding stability. It is distinct from V-Upd., which makes long-term verification a prerequisite for persistent writes and training. Base disables these additional reliability conditions, cross-view assistance, adapter optimization, and Gaussian verification; it retains local association and classifier fitting. Detection filtering, association acceptance, one-to-one allocation, and overlap handling remain common to all variants.

Without reliability control. Full w/o Rel. removes the confidence, competing-margin, and binding-stability conditions from ordinary memory and training admission. It retains Gaussian verification and verification-gated writes, cross-view assistance, and adapter learning. Initialization, restricted bootstrap, bank maturity, and valid-sample requirements remain unchanged. Thus, this control removes non-Gaussian reliability screening rather than long-term verification or basic association safeguards.

Cumulative online adaptation. For the cumulative Online Contrastive variant without Gauss., training admission uses the remaining reliability conditions without requiring long-term verification or a mature Gaussian bank, and the distribution-anchor loss is omitted. Thus, adding Gauss. changes both verification and its associated learning constraints.

Without online adaptation. The local adapters remain at their identity initialization, and no contrastive or distribution-anchor optimization is performed. Cross-view assistance, canonical-space verification, and memory admission remain active. Ridge classifiers are still refit when their admitted training buffers change, so this control removes representation learning rather than all online state updates.

Without verified updates. Long-term verification remains available for association decisions, but its pass/fail outcome no longer gates persistent-memory admission or training eligibility for the adapters and Ridge classifiers. This removal applies to both local and received observations. Confidence, competing margins, binding stability, and the requirement for a locally admitted same-target observation at the receiver are retained. Bootstrap and the requirement for a mature anchor before adapter learning are unchanged. This control tests verification-dependent write permissions rather than removing the entire reliability controller.

Adaptive verification features. This control replaces only the normalized canonical global query in long-term verification with its normalized output from the current local adapter; Gaussian statistics remain in canonical space, and their update rule, write permissions, and adaptation losses are unchanged. Both variants use the same verification threshold $\tau_\ell = 0.50$. This comparison tests query substitution at a common threshold, not matched acceptance rates or the necessity of the entire two-space design.

B.4 Additional Qualitative Analysis

Figure 4 shows an additional identity maintenance example in the M3Track wood scene. The selected target and a nearby pedestrian have similar dark clothing. The full model maintains the target identity throughout the displayed interval, whereas the configuration without online contrastive adaptation assigns it to the other pedestrian. Without cross-view assistance, the controller also selects the distractor before and after a brief interval with no confirmed output. Thus, the two ablations differ in whether an incorrect candidate is emitted, rather than exhibiting two consistently distinct failure modes.

The absence of a cyan box in the third column of the last row corresponds to this interruption. It lasts seven consecutive dataset frames (1868–1874), during which an incorrect tentative candidate remains available internally. The interruption therefore does not indicate that the identity has been corrected. These examples illustrate tracking behavior rather than isolate the effects of representation spaces or verification-dependent update permissions.



Figure 4: Additional comparison in the M3Track wood scene. Columns correspond to dataset frames 1861, 1867, 1871, and 1875. The reference view is shown once; the other rows use identical images and crops for the three configurations. Yellow dashed boxes denote the target and cyan boxes denote confirmed outputs; tentative candidates are not drawn. Displayed View 1 and View 2 correspond to source view2 and view1, respectively.