UAV-DIVERSE: A BENCHMARK FOR MULTI-PERSON TRACKING IN DIVERSE UAV SCENARIOS

Anonymous authors
Paper under double-blind review

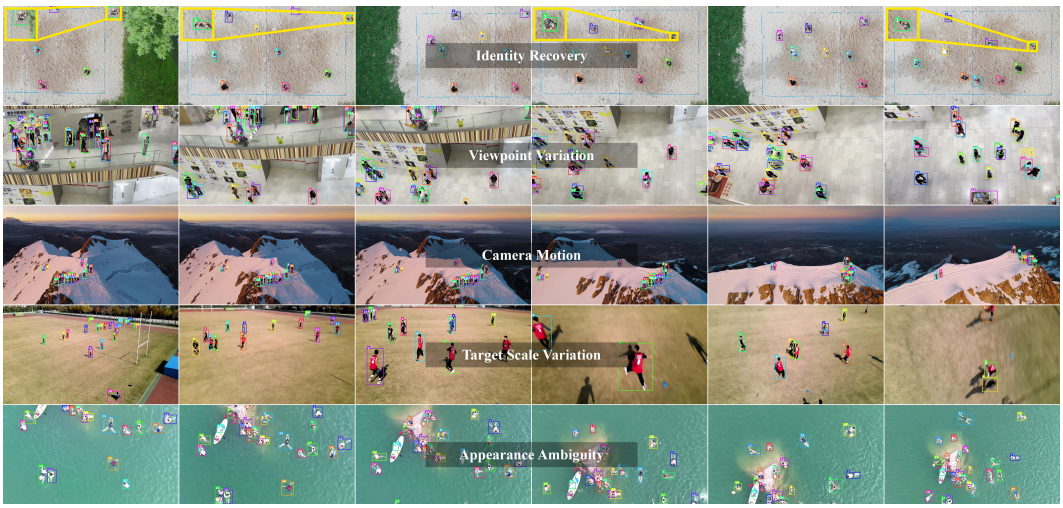


Figure 1: Representative sequences from UAV-Diverse illustrating five representative tracking challenges: identity recovery after target disappearance, viewpoint variation, camera motion, target scale variation, and appearance ambiguity. Bounding boxes with the same color indicate the same identity, while yellow insets highlight reappearing targets.

ABSTRACT

Robust multi-person tracking across diverse UAV scenarios is critical for real-world UAV applications. Existing UAV multi-object tracking (MOT) benchmarks are typically developed for specific application scenarios, providing limited insight into the cross-scene transfer capability of current trackers. To address this gap, we present UAV-Diverse, a benchmark for evaluating multi-person tracking across diverse UAV scenes. UAV-Diverse evaluates trackers without additional dataset-specific training, providing a unified protocol for assessing cross-scene transfer across three major scenario groups, including competitive sports, wilderness search and patrol, and urban and indoor public spaces, with nine fine-grained subscenes in total. The benchmark exposes key challenges including identity recovery after target disappearance, drastic viewpoint changes, dynamic camera motion, large target scale variations, and severe appearance ambiguity. Extensive evaluations demonstrate that representative MOT methods suffer substantial performance degradation, highlighting the significant challenges that remain in robust multi-person tracking in diverse UAV scenarios. We further provide a simple training-free baseline to facilitate future research.

1 INTRODUCTION

Multi-object tracking (MOT) aims to localize multiple targets and maintain consistent identities over time, serving as a fundamental component for surveillance, behavior analysis, and autonomous decision making (Dendorfer et al., 2020a;b). With their flexible deployment, wide-area coverage,

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.

and diverse viewpoints, unmanned aerial vehicles (UAVs) have become an important platform for multi-person tracking in applications such as public safety, search and rescue, wide-area patrol, and crowd behavior analysis (Zhu et al., 2021; Wen et al., 2021). Compared with ground-based videos, UAV videos exhibit substantially greater variations in camera motion, viewpoint, target scale, visibility, scene layout, and human appearance, making robust human localization and long-term identity preservation considerably more challenging.

Recent MOT methods have achieved remarkable progress through stronger object detection, motion estimation, appearance association, query propagation, and memory modeling (Zhang et al., 2022; Aharon et al., 2022; Cao et al., 2023; Zeng et al., 2022; Gao & Wang, 2023; Gao et al., 2025; Wang et al., 2026). UAV-oriented trackers further address challenges specific to aerial platforms, such as camera motion, rapid viewpoint changes, small targets, and frequent occlusions (Liu et al., 2022; Wang et al., 2024; Dang et al., 2024; Yang et al., 2023). Despite these advances, existing methods are typically developed and evaluated on benchmarks designed for individual UAV scenarios, making it difficult to assess their cross-scene transfer capability across diverse UAV environments.

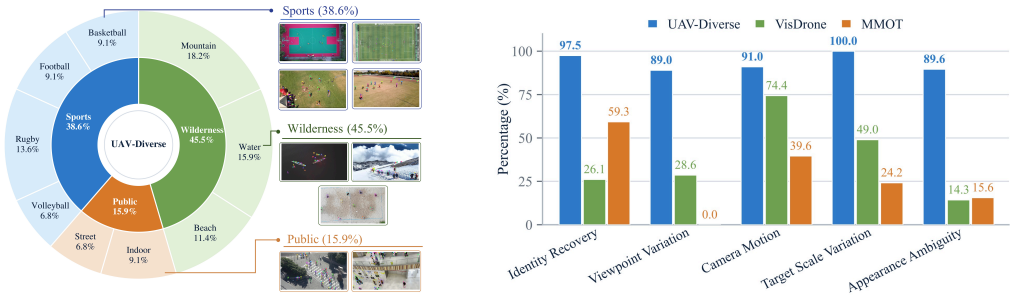
Existing UAV tracking benchmarks provide limited support for answering this question, as they are typically designed for specific application domains or sensing settings. For example, UAVDT (Du et al., 2018) focuses on traffic scenarios, SeaDronesSee-MOT (Varga et al., 2022) targets maritime search and rescue, MMOT (Li et al., 2025) emphasizes multispectral UAV tracking, and VisDrone (Zhu et al., 2021) is primarily collected in urban and suburban environments. Although these benchmarks have advanced UAV tracking within their respective domains, they provide limited evidence of whether trackers can transfer across substantially different UAV scenarios. Consequently, multi-person tracking across diverse UAV environments, such as competitive sports, wilderness search and patrol, urban public spaces, and indoor scenes, has yet to be systematically evaluated under a unified benchmark.

To address these limitations, we introduce UAV-Diverse, a benchmark for multi-person tracking in diverse UAV scenarios. As shown in Figure 2a, UAV-Diverse spans three major scenario groups, namely competitive **sports**, **wilderness** search and patrol, and urban and indoor **public** spaces, covering nine fine-grained subscenes in total. The benchmark contains 1,498 human trajectories and 821.9K bounding-box annotations with temporally consistent identity labels across 37,363 frames from 44 videos. By moving beyond a single application domain, UAV-Diverse captures substantial diversity in environments, activities, viewpoints, camera motion, target scales, and human appearance. Consequently, it introduces five representative tracking challenges, including identity recovery after target disappearance, viewpoint variation, camera motion, target scale variation, and appearance ambiguity, as shown in Figure 1.

Extensive experiments show that representative MOT and UAV-MOT methods exhibit limited cross-scene transferability when evaluated under their original inference pipelines. For example, the recent UAV-MOT method AMOT (Ma et al., 2026) achieves only 31.84 HOTA and 35.62 IDF1 on UAV-Diverse, compared with 35.01 HOTA and 44.93 IDF1 on VisDrone-MOT (Zhu et al., 2021) under the same human-only evaluation protocol. The larger drop in IDF1 suggests that identity preservation across heterogeneous UAV scenarios remains particularly challenging. To establish a reference for future benchmark evaluation, we further introduce UAV-Tracker, a training-free baseline built upon the frozen SAM3 foundation model. UAV-Tracker decouples internal memory maintenance from MOT trajectory generation and causally adapts its tracking policy during inference without dataset-specific training or fine-tuning.

Our main contributions are summarized as follows:

- We introduce UAV-Diverse, a benchmark for evaluating cross-scene transferability in UAV multi-person tracking. It covers three major scenario groups and nine fine-grained subscenes under a unified evaluation protocol.
- We establish a comprehensive benchmark evaluation by systematically assessing representative MOT and UAV-MOT methods, revealing their limited cross-scene transferability across diverse UAV scenarios.
- We present UAV-Tracker, a training-free baseline built upon frozen SAM3 that decouples memory maintenance from MOT trajectory generation and causally adapts its tracking policy during inference.



(a) Scenario distribution

(b) Challenge profile

Figure 2: Characterization of UAV-Diverse. (a) Scenario distribution of UAV-Diverse. The benchmark covers three major scenario groups, competitive sports, wilderness search and patrol, and urban and indoor public spaces, comprising nine fine-grained subscenes. (b) Challenge-oriented dataset profiles across identity recovery, viewpoint variation, camera motion, target scale variation, and appearance ambiguity. Identity recovery reports the conditional reappearance rate; viewpoint, camera-motion, and appearance scores report normalized entropy; and target-scale range ratios are normalized to UAV-Diverse (100%). All values in (b) are displayed as percentages.

2 Related Work

2.1 Human MOT Benchmarks

Human MOT benchmarks have played a central role in standardizing the evaluation of person localization, identity association, and trajectory consistency. The MOTChallenge series, including MOT15 (Leal-Taixé et al., 2015), MOT16 (Milan et al., 2016), and MOT20 (Dendorfer et al., 2020b), establishes widely used protocols for pedestrian tracking in crowded urban videos captured by fixed surveillance cameras. Subsequent benchmarks extend human MOT to more dynamic activities and interactions. DanceTrack (Sun et al., 2021) focuses on tracking dancers with similar appearances and diverse nonlinear motion, emphasizing association under ambiguous visual cues. SoccerNet-Tracking (Cioppa et al., 2022) then studies player tracking in soccer videos, while SportsMOT (Cui et al., 2023) broadens sports tracking to three sports with fast motion and frequent group interactions. TeamTrack (Scott et al., 2024) further introduces three team-sport videos with wide views covering the entire playing area, captured by fixed fisheye and UAV cameras. More recently, AVTrack (Wang et al., 2026) extends human-centric instance tracking to complex audio-visual scenes. These benchmarks broaden human MOT beyond fixed-camera surveillance, but provide limited coverage of UAV-specific observation conditions.

2.2 UAV MOT Benchmarks

UAVDT (Du et al., 2018) and VisDrone (Zhu et al., 2021) provide widely used benchmarks for UAV detection and tracking. More recently, MMOT (Li et al., 2025) introduces multispectral UAV MOT with oriented annotations. DynUAV (Ye et al., 2026) further benchmarks multi-class MOT in sequences purposefully captured under agile UAV maneuvers. These benchmarks have substantially advanced UAV perception, but are not centered on human identity association and remain concentrated on traffic, urban, and suburban scenes. Other UAV datasets focus more directly on human-centered perception. DroneCrowd (Wen et al., 2021) uses head-region annotations for crowd counting, localization, and tracking, which provide limited body-level appearance cues for identity association, while UAV-Human (Li et al., 2021) studies UAV-based human behavior understanding, but does not provide continuous multi-object tracking annotations. Meanwhile, SeaDronesSee-MOT (Varga et al., 2022) studies maritime search and rescue in domain-specific scenes. Existing datasets either address adjacent human-perception tasks or evaluate tracking within specialized UAV settings. In contrast, UAV-Diverse provides a unified benchmark for evaluating cross-scene transferability in human MOT.

Table 1: Comparison of UAV-Diverse with representative MOT benchmarks. **Test** denotes the percentage of all sequences assigned to the official test split, excluding validation sequences when a separate validation split is provided. **Frames**, **Ann.**, and **Length** report statistics over the full dataset. **Ann.** denotes the number of box annotations, and **Length** denotes the average sequence duration. † indicates statistics computed from publicly available annotations.

Dataset	Capture	Test	Frames	Ann.	Length	Target	Scene
MOT20 (2020b)	Surveillance	50.0	13.4K	2.1M	66.8s	Pedestrian	Crowded pedestrians
DanceTrack (2021)	Ground video	35.0	105K	877K	52.9s	Dancer	Group dance
SportsMOT (2023)	Sports video	62.5	150K	1.6M	25.1s	Player	Sports
TeamTrack (2024)	Fisheye+UAV	15.3	279.9K	4.37M	29.7s	Player	Sports
UAVDT-MOT (2018)	UAV	40.0	40.4K	763.8K	26.9s	Vehicle	Traffic
VisDrone-MOT† (2021)	UAV	20.3	33.6K	1.1M	–	Generic	Urban/suburban
MMOT (2025)	UAV	40.0	13.8K	488.3K	22.1s	Generic	Urban/rural
UAV-Diverse (ours)	UAV	**100.0**	37.4K	821.9K	28.1s	**Human**	**Multi-scenario**

3 UAV-DIVERSE BENCHMARK

UAV-Diverse is a benchmark for evaluating multi-person tracking across diverse UAV scenarios. To enable a clean assessment of cross-scene transferability, UAV-Diverse is released exclusively as a test set without an in-domain training split. This evaluation protocol prevents dataset-specific optimization and provides a standardized benchmark for comparing MOT methods under diverse UAV conditions.

3.1 SCENARIO COVERAGE AND ANNOTATION

UAV-Diverse contains 44 UAV sequences, 37,363 frames, 1,498 human trajectories, and 821.9K full-body bounding-box annotations with temporally consistent identities. Table 1 compares UAV-Diverse with representative MOT benchmarks. UAV-Diverse covers three major scenario groups: competitive sports, wilderness search and patrol, and urban and indoor public spaces. These groups are further divided into nine fine-grained subscenes, including basketball, football, rugby, volleyball, mountain, water, beach, indoor, and street. The distribution of scenario groups is shown in Figure 2a. A subset of competitive-sports videos is sourced from the TeamTrack (Scott et al., 2024) test split, and all included sequences are annotated and independently verified under a unified annotation protocol.

The annotations follow the standard MOT format, with full-body human boxes and sequence-specific identity labels. The same identity is retained when a person reappears after leaving the field of view, while fully invisible targets are not annotated. YOLO26 (Jocher et al., 2026) and SAM2 (Ravi et al., 2024) are used to generate initial annotations, which are manually refined for localization accuracy, detection completeness, and identity consistency. Further details on annotation refinement and identity verification are provided in Appendix A.1.

3.2 CHALLENGE-ORIENTED ANALYSIS

UAV-Diverse exposes five challenge dimensions for human identity association. Figure 1 illustrates representative examples, while Figure 2b compares their dataset-level statistics with VisDrone and MMOT. For viewpoint, camera motion, and appearance, we apply the same sequence-level taxonomies to all three datasets. Detailed category definitions and the blinded annotation protocol are provided in Appendix A.2.

Identity recovery after disappearance. Using a shared 15-frame protocol, we define a valid disappearance within a ground-truth trajectory as an internal absence gap of at least 15 frames following at least 15 consecutive visible frames. A valid reappearance further requires at least 15 subsequent visible frames. In UAV-Diverse, 7.9% of trajectories contain a valid disappearance, among which the conditional reappearance rate is 97.5%, compared with 26.1% and 59.3% for the person trajectories in VisDrone and MMOT. UAV-Diverse therefore places substantially greater emphasis on identity recovery after sustained visibility gaps.

Viewpoint variation. Motivated by the sequence-level camera-view attributes used in UAVDT (Du et al., 2018), we categorize each sequence as bird's-eye, oblique, or side view. UAV-Diverse contains 23, 15, and 6 sequences in these categories, yielding 89.0% normalized viewpoint entropy. In comparison, VisDrone reaches 28.6%, while all MMOT sequences are categorized as bird's-eye, yielding 0.0%. This broader viewpoint coverage tests whether trackers remain reliable across different UAV observation geometries.

Camera motion. Inspired by a prior video-level camera-motion taxonomy (Nasrabadi et al., 2019), we assign each sequence one of six mutually exclusive labels: static, horizontal translation, ascent, descent, rotation, or mixed. Mixed denotes sequences containing multiple sustained motion patterns for which no single pattern adequately characterizes the overall observation change. UAV-Diverse achieves 91.0% normalized camera-motion entropy, compared with 74.4% for VisDrone and 39.6% for MMOT, indicating broader coverage of camera-motion conditions.

Target scale variation. Target scale is measured by the per-sequence median human bounding-box height normalized by frame height. UAV-Diverse spans 1.3%–19.3%, corresponding to a 14.9× range ratio. In comparison, VisDrone and MMOT span 1.5%–10.6% (7.3×) and 0.8%–2.8% (3.6×). UAV-Diverse therefore covers both tiny aerial targets and larger close-range humans, challenging accurate localization and robust identity association.

Appearance ambiguity. To characterize appearance conditions that contribute to identity ambiguity, we group each sequence by its predominant human appearance category: everyday clothing, athletic attire, swimwear, or outdoor gear. UAV-Diverse reaches 89.6% normalized appearance entropy, compared with 14.3% for VisDrone and 15.6% for MMOT. In addition, UAV-Diverse combines diverse human appearances with visually challenging backgrounds, ranging from sand and water to snow-covered terrain and indoor clutter, further complicating localization and identity association.

Overall, these properties create three recurring challenges: identity discontinuities after sustained disappearance, unstable localization and visual cues under viewpoint, camera-motion, and scale changes, and ambiguous association under similar appearances and confounding backgrounds. UAV-Diverse therefore tests whether trackers can maintain identity consistency across heterogeneous UAV conditions. Condition-wise tracking results are provided in Appendix A.3.

4 UAV-TRACKER: A SIMPLE BASELINE

UAV-Diverse exposes three recurring difficulties for human tracking. Sustained disappearance and appearance ambiguity disrupt identity continuity, while viewpoint, camera-motion, and target-scale changes destabilize detector and propagation evidence. These factors also occur in different combinations across scenarios, making a fixed inference policy suboptimal. We therefore develop UAV-Tracker, a simple training-free baseline built on frozen SAM3 (Carion et al., 2025). SAM3 provides object representations and video-memory propagation, but propagated hypotheses may still fragment, drift, duplicate existing identities, or persist without detector support.

As shown in Figure 3, UAV-Tracker separates identity maintenance from frame-level trajectory emission. The ***Object Lifecycle Manager*** maintains identity continuity across temporary disappearance and validates recovery when targets reappear. The ***Context-Guided Track Stabilizer*** determines whether detector-supported or propagation-only evidence should be emitted in the current frame, suppressing drifting outputs. The ***Scene Fingerprint*** addresses multi-scenario heterogeneity by causally adapting their shared policy from past tracking diagnostics. No UAV-Diverse annotations or challenge labels are used for parameter selection or inference, no future frames are accessed, and SAM3 remains frozen throughout the entire inference process.

4.1 OBJECT LIFECYCLE MANAGER

As shown in the center panel of Figure 3, the Object Lifecycle Manager maintains an ID registry, anchor memory, and re-entry gallery to organize persistent identities and recovery. The manager maintains candidate, active, dormant, and released states, together with recovery transitions. A supported candidate becomes active, while identities losing detection support may become dormant

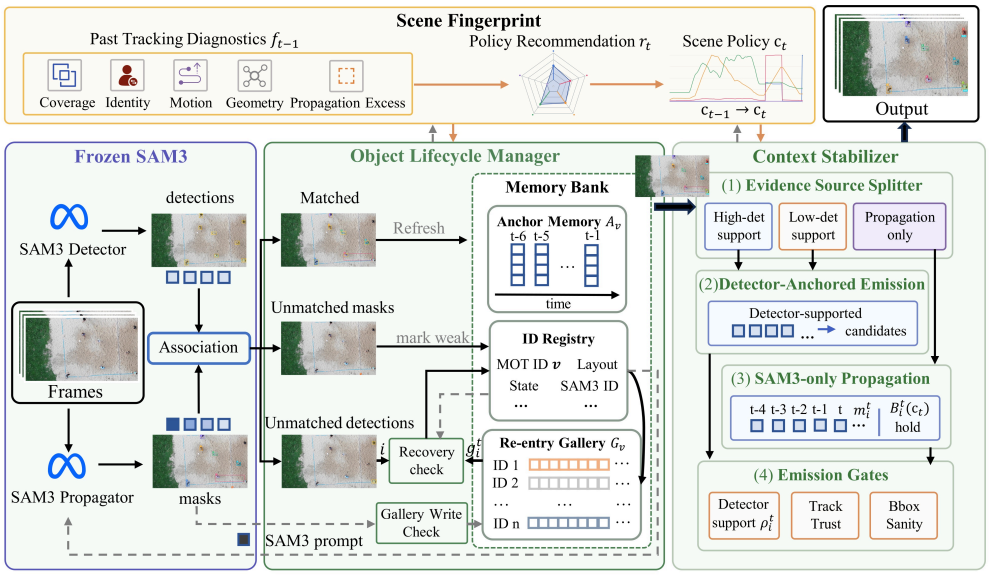


Figure 3: Overview of UAV-Tracker. Frozen SAM3 provides detector observations, object-focused features, and video-memory propagation. The Object Lifecycle Manager maintains persistent identities and validates recovery after disappearance, while the Context-Guided Track Stabilizer controls frame-level MOT emission. The Scene Fingerprint causally adapts their shared policy from past tracking diagnostics under heterogeneous UAV conditions.

but remain recoverable. Thus, temporarily absent targets can remain in memory for subsequent recovery without being emitted as trajectories.

UAV-Tracker reuses object features from frozen SAM3 for identity recovery and requires no additional ReID model. For candidate i and recoverable identity v, their memory similarity is

$$s_{\text{mem}}(i, v) = \max_{\mathbf{u} \in G_v \cup A_v} \cos\left(\mathbf{g}_i^t, \mathbf{u}\right), \tag{1}$$

where $\mathbf{g}_i^t$ is the candidate feature at frame t, and G_v and A_v denote the re-entry gallery and anchor memory of identity v. Only observations satisfying the corresponding memory-update conditions are written to these memories, limiting contamination from weak evidence.

Let $\mathcal{P}_{\text{rec}}$ be the recoverable identity pool. We denote the highest and second-highest similarities by s_1 and s_2. An appearance-based recovery proposal passes the initial gate only when

$$s_1 \geq \gamma_{\text{reid}}\left(|\mathcal{P}_{\text{rec}}|\right), \qquad s_1 - s_2 \geq \Delta, \tag{2}$$

where γ_{reid} is an adaptive similarity threshold that increases with the size of $\mathcal{P}_{\text{rec}}$, and Δ is the minimum matching margin. The scene policy further restricts identity recovery when the estimated ambiguity level increases.

Appearance may be ambiguous when nearby people look similar. A proposal passing the initial gate undergoes temporal confirmation and a layout veto. Let $\boldsymbol{\ell}_{i,a}^t = (\mathbf{p}_i^t - \mathbf{p}_a^t)/d_I$ denote the normalized displacement from common anchor a to candidate i, where $\mathbf{p}^t$ is the box center and d_I is the image diagonal. The layout inconsistency is

$$e_{\text{layout}}(i, v, t) = \underset{a \in \mathcal{A}_{i,v}^t}{\text{median}} \left\| \boldsymbol{\ell}_{i,a}^t - \boldsymbol{\ell}_{v,a}^{\text{old}} \right\|_2, \tag{3}$$

where $\mathcal{A}_{i,v}^t$ contains the common anchors and $\boldsymbol{\ell}_{v,a}^{\text{old}}$ is the stored layout of identity v. When valid layout evidence is available, a proposal is vetoed if $e_{\text{layout}}(i, v, t) > \epsilon_{\text{layout}}$. If the layout evidence is insufficient, the check is skipped. A proposal that passes all checks reactivates the dormant visible identity and rebinds it to the current SAM3 hypothesis. If recovery fails, the system retains the current SAM3 assignment. A new visible identity is committed only after newborn admission.

4.2 CONTEXT-GUIDED TRACK STABILIZER

The right panel of Figure 3 summarizes how the Context-Guided Track Stabilizer converts detector and propagation evidence into frame-level MOT outputs through evidence splitting, detector-anchored emission, propagation control, and final emission gates. SAM3 propagation bridges temporary detector misses but may drift or persist without detector support when prolonged. The Stabilizer therefore combines detector anchors with propagation evidence to determine whether each hypothesis should be emitted as a frame-level MOT output.

For each hypothesis i, UAV-Tracker distinguishes high-confidence and low-confidence detector support, and SAM3-only propagation. Their recent evidence is summarized as

$$\rho_i^t = \frac{1}{|\mathcal{H}_i^t|} \sum_{\tau \in \mathcal{H}_i^t} w_i^\tau, \qquad w_i^\tau = \begin{cases} 1, & \text{high-det,} \\ \omega_{\text{low}}, & \text{low-det,} \\ 0, & \text{SAM3-only,} \end{cases} \tag{4}$$

where $\mathcal{H}_i^t$ is the available support history and $0 < \omega_{\text{low}} < 1$. Normalizing by the available history length avoids penalizing newly initialized tracks.

High-confidence detections directly form detector-anchored candidates and bypass the recent-support gate. Low-confidence detections retain an observation anchor but require sufficient recent support. Let m_i^t denote the consecutive propagation length of a SAM3-only hypothesis. Such a hypothesis is emitted only when

$$m_i^t \le B_i^t(c_t) \qquad \text{and} \qquad \rho_i^t \ge \gamma_{\text{sup}}(c_t), \tag{5}$$

where $B_i^t(c_t)$ is its propagation budget and $\gamma_{\text{sup}}(c_t)$ is the required detector-support level. Low-confidence hypotheses use the same support gate but are not constrained by the propagation budget.

The shared policy state c_t controls the propagation budget $B_i^t(c_t)$ and support threshold $\gamma_{\text{sup}}(c_t)$. Its causal update is described in the next subsection. A final box-sanity check removes geometrically implausible rows before emission.

4.3 SCENE FINGERPRINT

The upper policy pathway in Figure 3 shows how the Scene Fingerprint causally maps past tracking diagnostics to the policy applied at the current frame. Since a fixed policy cannot balance target coverage and identity stability across heterogeneous UAV conditions, UAV-Tracker summarizes the diagnostics from a causal window W_{t-1} after finalizing frame $t-1$:

$$\mathbf{f}_{t-1} = [f_{\text{cov}}, f_{\text{id}}, f_{\text{mot}}, f_{\text{geo}}, f_{\text{prop}}]_{t-1}\,, \tag{6}$$

where the entries measure coverage deficiency, identity ambiguity, motion instability, geometric variation, and propagation excess. They are computed from past detector evidence, finalized MOT outputs, and identity diagnostics, without using scene or challenge labels.

The fingerprint determines the policy applied at frame t:

$$r_t = \Phi(\mathbf{f}_{t-1}), \qquad c_t = \mathcal{U}(c_{t-1}, r_t), \tag{7}$$

where Φ maps the diagnostics to a policy recommendation and $\mathcal{U}$ updates the previous policy only when the recommendation is temporally consistent. The resulting c_t controls the re-entry gate, dormant retention horizon, propagation budget, and detector-support requirement shared by the Object Lifecycle Manager and Context-Guided Track Stabilizer. After frame t is finalized, its diagnostics update $\mathbf{f}_t$ for the subsequent transition $c_t \to c_{t+1}$. Detailed fingerprint computation is provided in Appendix A.6.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Training protocol. We use official implementations and released checkpoints whenever available. For trainable methods without released checkpoints, we train them on the VisDrone-MOT (Zhu

Table 2: Tracking performance on UAV-Diverse. The original pipeline group evaluates direct transfer using each method's native inference pipeline. The shared SAM3 group compares tracking strategies using the same pre-generated SAM3 detections. AMOT+SAM3Feat additionally uses frozen SAM3 object features. All values are percentages.

Protocol	Method	HOTA ↑	DetA ↑	AssA ↑	IDF1 ↑	MOTA ↑	Pub.
Original pipeline	UAVMOT (2022)	26.46	17.55	40.16	30.08	21.20	CVPR'22
	MeMOTR (2023)	35.07	25.95	48.47	41.16	29.23	ICCV'23
	MMTracker (2025)	31.10	22.30	43.86	33.68	24.06	AAAI'25
	MOTIP (2025)	31.52	25.35	40.00	35.35	27.18	CVPR'25
	AMOT (2026)	31.84	21.46	47.58	35.62	26.20	AAAI'26
Shared SAM3 dets.	ByteTrack (2022)	54.56	50.27	59.83	65.40	55.45	ECCV'22
	BoT-SORT (2022)	54.51	49.89	60.14	63.64	55.17	arXiv'22
	OC-SORT (2023)	56.49	51.90	62.29	65.96	54.53	CVPR'23
	Native SAM3 (2025)	51.64	44.48	61.16	61.05	33.74	arXiv'25
	AMOT+SAM3Feat	<u>59.79</u>	<u>54.42</u>	<u>66.14</u>	<u>68.80</u>	**62.83**	–
	UAV-Tracker (Ours)	**61.90**	**55.88**	**69.43**	**71.97**	<u>58.20</u>	–

et al., 2021) training split using their official configurations and retain the final checkpoint from the prescribed training schedule. No UAV-Diverse annotations are used to train learned components. UAV-Tracker is training-free, and SAM3 remains frozen throughout inference. A single inference configuration is applied to all 44 sequences.

Evaluation metrics. We evaluate all methods using TrackEval (Luiten & Hoffhues, 2020) and report HOTA (Luiten et al., 2020), DetA, AssA, IDF1, and MOTA. HOTA is used as the primary metric because it jointly measures detection and association quality.

Comparison protocols. We report two complementary groups. The *Original Pipeline* group evaluates direct transfer to UAV-Diverse using each method's native detector and inference pipeline. We use released checkpoints when available and the VisDrone-trained checkpoints described above otherwise. End-to-end, query-based, and detector-coupled methods remain in this group because replacing their detections would alter their original designs.

For methods accepting external detections, the *Shared SAM3 Detections* group uses the same pre-generated boxes and confidence scores from frozen SAM3 with the text prompt "person." Each tracker retains its native association and trajectory-realization logic. AMOT+SAM3Feat additionally pools per-detection object features from the frozen SAM3 FPN for appearance association. Scores of at least 0.65 are high-confidence, scores in $[0.45, 0.65)$ are low-confidence, and lower-scoring detections are discarded before tracking.

5.2 Main Results on UAV-Diverse

Table 2 reports the main results on UAV-Diverse. Under the *Original Pipeline* group, all evaluated methods remain below 36 HOTA. UAV-oriented methods, including UAVMOT, MMTracker, and AMOT, remain below 32 HOTA. These results show that direct transfer to UAV-Diverse remains challenging across current MOT and UAV-oriented tracking methods.

Methods in the *Shared SAM3 Detections* group obtain higher scores overall, while clear performance differences remain when detector observations are controlled. UAV-Tracker achieves 61.90 HOTA and outperforms the strongest SAM3-assisted baseline, AMOT+SAM3Feat, by 2.11 HOTA, 3.29 AssA, and 3.17 IDF1. These improvements indicate that the proposed identity-maintenance and frame-level emission mechanisms provide additional benefits for association and trajectory continuity. While AMOT+SAM3Feat achieves the highest MOTA, UAV-Tracker leads on HOTA, DetA, AssA, and IDF1, with clear gains in identity association.

Results on human-only VisDrone-MOT using the same fixed configuration without VisDrone-specific retuning are reported in Appendix A.5.

5.3 Ablation Study

Table 3: Component ablation on UAV-Diverse. OLM, SF, and CS denote the Object Lifecycle Manager, Scene Fingerprint, and Context-Guided Track Stabilizer. Values are percentages.

ID	OLM	SF	CS	HOTA ↑	AssA ↑	IDF1 ↑	MOTA ↑
I	✗	✗	✗	51.64	61.16	61.05	33.74
II	✓	✗	✗	59.48	68.95	68.06	45.40
III	✓	✓	✗	61.03	69.23	70.17	52.46
IV	✓	✗	✓	61.35	69.07	71.17	56.20
V	✓	✓	✓	**61.90**	**69.43**	**71.97**	**58.20**

Table 3 evaluates the contribution of each component under the same frozen SAM3 observations and experimental setting. Compared with Native SAM3 (I), adding the Object Lifecycle Manager (II) yields the largest individual gain, improving HOTA by 7.84, IDF1 by 7.01, and MOTA by 11.66. Building on this configuration, the Scene Fingerprint (III) further improves HOTA by 1.55 and MOTA by 7.06, while the Context-Guided Track Stabilizer (IV) provides larger gains in IDF1 and MOTA, increasing them by 3.11 and 10.80. Combining all three components (V) achieves the best overall performance, reaching 61.90 HOTA, 69.43 AssA, 71.97 IDF1, and 58.20 MOTA.

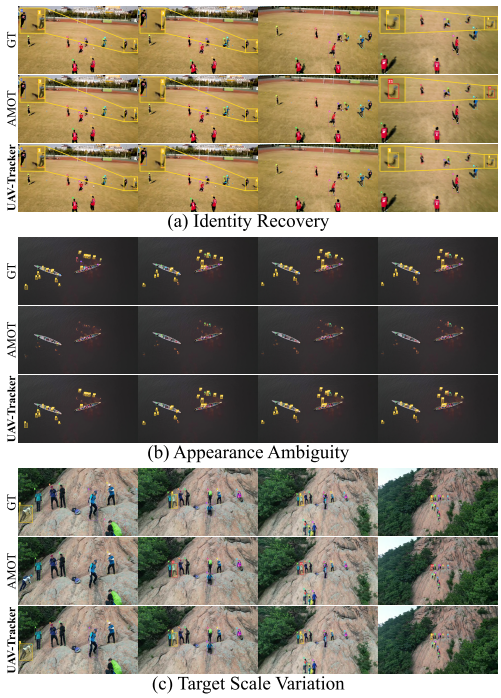


Figure 4: Qualitative comparison between AMOT+SAM3Feat and UAV-Tracker. AMOT denotes AMOT+SAM3Feat.

These results demonstrate the complementary effects of persistent identity management, frame-level trajectory realization, and causal policy adaptation.

5.4 Qualitative Analysis

As shown in Figure 4, AMOT+SAM3Feat assigns a new identity after the target re-enters the view in (a), while UAV-Tracker restores the original identity. UAV-Tracker also maintains more complete and consistent trajectories under partial submersion and substantial scale changes in (b) and (c). Additional comparisons across the five challenge dimensions are provided in Appendix A.4.

6 Conclusion

Existing UAV MOT benchmarks are often tied to specific application domains, limiting the evaluation of human tracking across diverse UAV scenes. To fill this gap, we introduce UAV-Diverse, a unified benchmark covering heterogeneous scenarios and observation conditions, with emphasis on identity recovery after disappearance, viewpoint variation, camera motion, target scale variation, and appearance ambiguity. Evaluations show limited cross-scene transferability for current MOT and UAV-oriented trackers, while clear differences in identity association remain under shared detector observations. We also present UAV-Tracker, a simple training-free baseline that separates identity maintenance from frame-level trajectory emission and causally adapts its tracking policy. We hope UAV-Diverse fosters future research on identity continuity and reliable cross-scene human tracking across diverse UAV environments.

Limitations. UAV-Diverse cannot cover all real-world UAV environments and sensing conditions. UAV-Tracker also inherits SAM3's limitations in localization, propagation drift, and efficiency. Broader benchmark coverage and more robust video models remain future directions.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of this paper. All scientific content, experimental design, results, analyses, and conclusions were produced and verified by the authors. We reviewed all AI-assisted edits and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

UAV-Diverse comprises publicly available online UAV videos, selected sequences from TeamTrack, and indoor UAV recordings collected by the authors. For the self-recorded indoor sequences, visually identifiable faces are anonymized prior to release. The benchmark provides bounding boxes and sequence-local trajectory labels only, without names, facial identity labels, demographic attributes, or cross-video identity associations. For externally sourced videos, we will release annotations and source references, while redistribution of the original video content will follow the applicable licenses and permissions. UAV-Diverse is intended for academic research on robust human tracking and scene understanding in diverse UAV environments.

REPRODUCIBILITY STATEMENT

We provide information to support the reproducibility of both UAV-Diverse and UAV-Tracker. Appendix A.1 describes the dataset construction, annotation pipeline, manual refinement, and quality-control procedure. Appendix A.2 specifies the challenge definitions and statistical protocols, while Appendix A.3 provides the condition-specific evaluation settings. Appendix A.5 describes the cross-dataset evaluation on VisDrone-MOT, and Appendix A.6 reports the fixed inference configuration, baseline training protocols, random seeds, and computing environment. Upon publication, we will publicly release the UAV-Diverse annotations, evaluation code, and the implementation of UAV-Tracker, together with documentation of data sources and access procedures. Original video content will be redistributed only where permitted by the applicable licenses and permissions.

REFERENCES

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

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.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3491–3502, 2022.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*, 2023.

Zhaoyang Dang, Xiaoyong Sun, Bei Sun, Runze Guo, and Can Li. Omctrack: Integrating occlusion perception and motion compensation for uav multi-object tracking. *Drones*, 8(9), 2024. ISSN 2504-446X. doi: 10.3390/drones8090480. URL https://www.mdpi.com/2504-446X/8/9/480.

Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking, 2020a. URL https://arxiv.org/abs/2010.07548.

Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes, 2020b. URL https://arxiv.org/abs/2003.09003.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking, 2018. URL https://arxiv.org/abs/1804.00518.

Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9901–9910, October 2023.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as id prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 27883–27893, June 2025.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: Unified real-time end-to-end vision models, 2026. URL https://arxiv.org/abs/2606.03748.

Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking, 2015. URL https://arxiv.org/abs/1504.01942.

Tianhao Li, Tingfa Xu, Ying Wang, Haolin Qin, Xu Lin, and Jianan Li. Mmot: The first challenging benchmark for drone-based multispectral multi-object tracking, 2025. URL https://arxiv.org/abs/2510.12565.

Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16266–16275, June 2021.

Shuai Liu, Xin Li, Huchuan Lu, and You He. Multi-object tracking meets moving uav. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8866–8875, 2022. doi: 10.1109/CVPR52688.2022.00867.

Jonathon Luiten and Arne Hoffhues. Trackeval. https://github.com/JonathonLuiten/TrackEval, 2020.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip H. S. Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, pp. 1–31, 2020.

Jianbo Ma, Hui Luo, Qi Chen, Yuankai Qi, Yumei Sun, Amin Beheshti, Jianlin Zhang, and Ming-Hsuan Yang. Tracking the unstable: Appearance-guided motion modeling for robust multi-object tracking in uav-captured videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7773–7781, 2026.

Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking, 2016. URL https://arxiv.org/abs/1603.00831.

Jeffri Murrugarra-Llerena, Lucas N Kirsten, and Claudio R Jung. Can we trust bounding box annotations for object detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4813–4822, 2022.

Afshin Taghavi Nasrabadi, Aliehsan Samiei, Anahita Mahzari, Ryan P McMahan, Ravi Prakash, Mylène CQ Farias, and Marcelo M Carvalho. A taxonomy and dataset for 360 videos. In *Proceedings of the 10th ACM Multimedia Systems Conference*, pp. 273–278, 2019.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Atom Scott, Ikuma Uchida, Ning Ding, Rikuhei Umemoto, Rory Bunker, Ren Kobayashi, Takeshi Koyama, Masaki Onishi, Yoshinari Kameda, and Keisuke Fujii. Teamtrack: A dataset for multi-sport multi-object tracking in full-pitch videos, 2024. URL https://arxiv.org/abs/2404.13868.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. *arXiv preprint arXiv:2111.14690*, 2021.

Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Seadronessee: A maritime benchmark for detecting humans in open water. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2260–2270, 2022.

Peng Wang, Yongcai Wang, and Deying Li. Dronemot: Drone-based multi-object tracking considering detection difficulties and simultaneous moving of drones and objects, 2024. URL https://arxiv.org/abs/2407.09051.

Yaoting Wang, Yun Zhou, Zipei Zhang, and Henghui Ding. AVTrack: Audio-visual tracking in human-centric complex scenes. In *International Conference on Machine Learning (ICML)*, 2026.

Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *CVPR*, 2021.

Cheng-Yen Yang, Hsiang-Wei Huang, Zhongyu Jiang, Heng-Cheng Kuo, Jie Mei, Chung-I Huang, and Jenq-Neng Hwang. Sea you later: Metadata-guided long-term re-identification for uav-based multi-object tracking, 2023. URL https://arxiv.org/abs/2311.03561.

Mufeng Yao, Jinlong Peng, Qingdong He, Bo Peng, Hao Chen, Mingmin Chi, Chao Liu, and Jon Atli Benediktsson. Mm-tracker: Motion mamba for uav-platform multiple object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9409–9417, 2025.

Jingtao Ye, Kexin Zhang, Xunchi Ma, Yuehan Li, Guangming Zhu, Peiyi Shen, Linhua Jiang, Xiangdong Zhang, and Liang Zhang. Breaking smooth-motion assumptions: A uav benchmark for multi-object tracking in complex and adverse conditions. *arXiv preprint arXiv:2603.05970*, 2026.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

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

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021.

A Appendix

A.1 Dataset Construction and Annotation Protocol

A.1.1 Video Collection and Screening

UAV-Diverse is constructed from three sources: UAV footage captured by the authors, UAV videos collected from publicly accessible web sources, and selected sequences from the official TeamTrack (Scott et al., 2024) release. We first collected candidate videos that contain visible humans

under aerial viewpoints, and then screened them according to their suitability for multi-person tracking. Videos were retained only when they contained trackable human instances, sufficient temporal continuity, and meaningful UAV observation conditions. Videos with severe compression artifacts, extremely low resolution, near-complete target invisibility, or insufficient human activity were discarded.

A subset of the competitive-sports videos, including football and basketball sequences, is sourced from the TeamTrack test split. The original TeamTrack annotations primarily cover athletes on the playing field and the ball, while other visible people in the scene are generally not annotated. For UAV-Diverse, we remove all ball annotations, refine nearly all existing human bounding boxes according to our full-body annotation standard, and add bounding boxes and sequence-local identities for previously unannotated visible people. All resulting annotations are subsequently reviewed under the same identity verification and quality-control protocol as the other UAV-Diverse sequences.

The selected videos cover three major scenario groups: competitive sports, wilderness search and patrol, and urban or indoor public areas. This screening process was designed to increase scenario diversity while keeping the annotation target consistent: full-body human bounding boxes and temporally consistent identities.

A.1.2 Automatic Pre-annotation

To reduce manual annotation cost and improve temporal consistency, we first generated coarse annotations using a model-assisted pre-annotation pipeline. The pipeline combines a detector with a SAM2-based (Ravi et al., 2024) video tracker. In our implementation, YOLO26 (Jocher et al., 2026) is used to detect human candidates, and SAM2 is used to propagate object masks and track hypotheses across frames. The detector provides frame-level human observations, while the SAM2 tracker provides temporal continuity between adjacent frames.

Given video frames $\{I_t\}_{t=1}^{T}$, detector $\mathcal{D}$, and SAM2 tracker $\mathcal{S}$, the pre-annotation procedure is summarized in Algorithm 1. At each frame, track boxes propagated by the SAM2 tracker are matched with detector boxes using Hungarian association. Matched objects inherit their previous track identities. Unmatched detections are treated as candidate new identities, while unmatched propagated tracks are discarded. The output of this stage is a set of coarse full-body boxes and preliminary identity trajectories, which are then passed to human annotators for refinement.

A.1.3 Manual Refinement and Identity Verification

After automatic pre-annotation, all sequences were imported into CVAT for manual refinement. Annotators corrected both full-body bounding boxes and identity assignments. For localization, each bounding box was adjusted to tightly enclose the visible full-body extent of the target person whenever the body boundary was visually distinguishable. Annotators were instructed to include the human body region belonging to the target while avoiding excessive background.

For identity annotation, the goal was to maintain temporally consistent identities throughout each sequence, including after temporary disappearance. When a person entered the frame, annotators checked whether the target corresponded to an earlier individual who had left the field of view. If the correspondence was sufficiently confident, the previous identity was reused. Otherwise, a new identity was assigned. This conservative rule avoids forcing uncertain reappearance matches while still allowing valid identity recovery when visual and temporal evidence is clear.

A.1.4 Multi-annotator Review and Quality Control

Ambiguous cases were further reviewed by an additional annotator. These cases include uncertain reappearance, visually similar nearby people, partial occlusion, severe scale changes, and unclear full-body boundaries. When the annotators disagreed, the case was discussed until a consensus decision was reached. For identity recovery, the same conservative principle was applied during consensus review: a previous identity was reused only when the match was sufficiently reliable; otherwise, the target was assigned a new identity.

The quality-control process checked temporal continuity, duplicate identities, missing visible humans, and implausible bounding boxes. The cases with detected errors were returned to CVAT for correction

Algorithm 1 Automatic Pre-annotation for UAV-Diverse

Require: Video frames $\{I_t\}_{t=1}^{T}$, detector $\mathcal{D}$, SAM tracker $\mathcal{S}$
Initialize active tracks $\mathcal{T}$
for $t = 1$ to $T - 1$ **do**
  Propagate active tracks from I_t to I_{t+1} using $\mathcal{S}$ and obtain predicted boxes $\hat{\mathcal{B}}_{t+1}$
  Detect human boxes $\mathcal{D}_{t+1}$ on frame I_{t+1}
  Match $\hat{\mathcal{B}}_{t+1}$ and $\mathcal{D}_{t+1}$ using Hungarian association
  for each matched prediction-detection pair **do**
    Inherit the previous track identity
  end for
  Discard unmatched propagated tracks
  for each unmatched detection **do**
    Create a candidate new identity
  end for
  Save coarse boxes and identity hypotheses for human verification
end for
return Pre-annotated human tracks for manual refinement

and then again reviewed prior to final confirmation. The final annotations therefore combine model-assisted temporal proposals, manual box and identity refinement, and consensus-based quality control, providing full-body human boxes and temporally consistent identity trajectories for UAV-Diverse.

A.2 Challenge Definitions and Additional Statistics

The main paper reports the dataset-level challenge profile of UAV-Diverse and compares it with representative UAV MOT benchmarks. This section provides the operational definitions and additional statistics used in that analysis. Identity disappearance and reappearance are measured at the trajectory level. Viewpoint, camera motion, and appearance conditions are assigned at the sequence level according to the dominant observation condition, while target scale is computed directly from the annotations. These attributes are used only for evaluation and are never provided to any tracker during inference.

Identity disappearance and reappearance. Because sequence-level frame rates are not consistently documented for the comparison benchmarks, we define disappearance and reappearance in the frame domain and apply the same protocol to UAV-Diverse, VisDrone (Zhu et al., 2021), and MMOT (Li et al., 2025). Given a threshold T, a qualifying visible segment contains at least T consecutive annotated frames. A trajectory contains a valid disappearance when a qualifying visible segment is followed by an *internal* absence interval of at least T consecutive frames. Terminal exits at the end of a sequence are not counted as valid disappearances. A valid reappearance additionally requires the post-gap visible segment to contain at least T consecutive annotated frames.

For trajectory-level statistics, each trajectory is counted at most once, regardless of how many qualifying disappearance intervals it contains. Let $N_{\mathrm{disp}}(T)$ denote the number of trajectories containing at least one valid disappearance under threshold T, and let $N_{\mathrm{reapp}}(T)$ denote the number of these trajectories for which at least one valid disappearance is followed by a qualifying reappearance. The conditional reappearance rate is defined as

$$R_{\mathrm{reapp}}(T) = \frac{N_{\mathrm{reapp}}(T)}{N_{\mathrm{disp}}(T)}. \tag{8}$$

A sequence is considered to contain reappearance if it includes at least one trajectory counted by $N_{\mathrm{reapp}}(T)$. We use $T = 15$ in the main analysis, corresponding to 0.5 seconds for UAV-Diverse at 30 FPS. Cross-dataset results are interpreted as a shared frame-domain comparison rather than an exact physical-time comparison.

Under $T = 15$, 118 of the 1,498 UAV-Diverse trajectories contain a valid disappearance, corresponding to 7.9% of all trajectories. Among these, 115 contain a valid reappearance, yielding a conditional reappearance rate of 97.5%. At the sequence level, at least one valid reappearance event occurs

Table 4: Sensitivity of the trajectory-level conditional reappearance rate to the frame-domain threshold T. Each entry reports the number of trajectories containing a valid reappearance over the number containing a valid disappearance.

Threshold T	**UAV-Diverse**	**VisDrone person**	**MMOT person**
10 frames	122/125 (97.6%)	88/289 (30.4%)	84/116 (72.4%)
15 frames	115/118 (97.5%)	71/272 (26.1%)	32/54 (59.3%)
30 frames	99/103 (96.1%)	49/211 (23.2%)	1/3 (33.3%)

in 24 of the 44 UAV-Diverse sequences. The corresponding rates are 26.1% for VisDrone person trajectories (71/272) and 59.3% for MMOT person trajectories (32/54).

As shown in Table 4, UAV-Diverse maintains a substantially higher conditional reappearance rate than VisDrone and MMOT throughout the tested range. The ordering remains unchanged for $T \in \{10, 15, 30\}$, indicating that the observed difference is not driven by the selected threshold. We report exact counts because the number of qualifying MMOT trajectories becomes small under stricter settings.

Blinded sequence-level labeling. Viewpoint, camera-motion, and appearance labels are independently assigned by three annotators who did not participate in the bounding-box annotation of UAV-Diverse. Dataset names and source identifiers are hidden during annotation to reduce dataset-specific bias. Uniformly sampled frames at one-second intervals provide an initial overview, after which annotators review the complete sequence before selecting its dominant condition. Disagreements are resolved through joint review and consensus. The resulting labels are mutually exclusive for evaluation grouping, but do not imply that other observation components are absent from a sequence.

For a taxonomy containing K categories, we report normalized Shannon entropy:

$$H_{\text{norm}} = -\frac{1}{\log K} \sum_{k=1}^{K} p_k \log p_k, \tag{9}$$

where p_k is the proportion of sequences assigned to category k, and terms with $p_k = 0$ are defined as zero. We use $K = 3$ for viewpoint, $K = 6$ for camera motion, and $K = 4$ for appearance. Entropy values are reported as percentages in the main paper.

Viewpoint condition. Viewpoint is defined relative to the local ground plane rather than absolute flight altitude. Each sequence is assigned one of three labels: *bird's-eye*, *oblique*, or *side view*. Bird's-eye denotes a near-nadir observation in which the viewing direction is approximately perpendicular to the local ground and the upper surfaces of targets dominate. Side view denotes a near-horizontal observation dominated by frontal or lateral human appearance. Oblique denotes the intermediate elevated geometry, in which both upper and lateral body structures remain visible. If viewpoint changes over time, the label occupying the largest temporal portion is used. UAV-Diverse contains 23 bird's-eye, 15 oblique, and 6 side-view sequences. The corresponding counts are 4, 74, and 2 for VisDrone, and 125, 0, and 0 for MMOT.

Camera-motion condition. Each sequence is assigned one dominant camera-motion label: *static*, *horizontal translation*, *ascent*, *descent*, *rotation*, or *mixed*. Static is defined conservatively and permits only minor jitter, slow drift, or brief camera adjustment without sustained visually meaningful motion. Horizontal translation denotes sustained translational motion without dominant vertical or rotational change. Ascent and descent denote observation changes dominated by increasing or decreasing camera height. Rotation denotes sustained turning or rotational motion. Mixed is used when multiple sustained and visually significant motion patterns are present and no single category adequately describes the sequence. It is not used as an uncertainty label. The category counts for UAV-Diverse are 16, 7, 2, 5, 8, and 6. The corresponding counts are 10, 39, 1, 1, 17, and 12 for VisDrone, and 0, 97, 4, 0, 4, and 20 for MMOT.

Target scale. For sequence q, target scale is measured by the median bounding-box height normalized by image height:

$$s_q = \underset{(t,i)\in\mathcal{B}_q}{\text{median}} \frac{h_i^t}{H_t}, \tag{10}$$

where $\mathcal{B}_q$ contains all annotated target boxes in sequence q, h_i^t is the height of target i's box in frame t, and H_t is the frame height. For MMOT, each oriented box is first converted to its enclosing axis-aligned box before its height is measured.

The dataset-level scale range ratio reported in the main paper is

$$R_{\text{scale}} = \frac{\max_q s_q}{\min_q s_q}. \tag{11}$$

For the binned robustness analysis, sequences are grouped into five fixed intervals: $[0, 3)$, $[3, 5)$, $[5, 7)$, $[7, 10)$, and ≥ 10, where values are percentages of image height.

Appearance condition. Each sequence is assigned one predominant human appearance label: *everyday clothing*, *athletic attire*, *swimwear*, or *outdoor gear*. Everyday clothing covers ordinary pedestrian attire. Athletic attire includes team uniforms and sportswear. Swimwear or minimal clothing includes swimming scenes, bare-torso observations, and partially submerged targets. Outdoor gear includes hiking, winter, and other activity-specific protective clothing. The label is determined from target appearance rather than background content. Backgrounds such as sand, water, vegetation, snow, and indoor clutter are analyzed as association confounders but do not contribute to the appearance entropy. UAV-Diverse contains 7, 22, 7, and 8 sequences in the four appearance categories. The corresponding counts are 76, 4, 0, and 0 for VisDrone, and 118, 7, 0, and 0 for MMOT.

A.3 Challenge-Specific Evaluation

In addition to the overall benchmark evaluation, we conduct analyses across the five challenge dimensions defined in the main paper. For HOTA, IDF1, and AssA, each metric is first computed independently for every sequence and then averaged over all sequences in the corresponding subset. As a result, each sequence contributes equally to the final result, regardless of its duration or number of annotations. IDSW, in contrast, is summed over all sequences in the subset.

For tables that report a macro score across condition categories, such as viewpoint or camera motion, the macro score is the unweighted average of the category-level results. These condition-specific values are therefore not directly comparable with the dataset-level `COMBINED_SEQ` results in the main paper, which are obtained by jointly accumulating predictions and annotations across the complete test set. All condition labels are used only to define evaluation subsets and are not provided to any tracker during inference.

A.3.1 Identity Recovery

Identity recovery after disappearance is most directly related to the Object Lifecycle Manager. Valid disappearance and reappearance are defined at the trajectory level in Appendix A.2, while the evaluation in Table 5 is conducted at the sequence level. A sequence is included in the subset if it contains at least one trajectory with a valid reappearance under $T = 15$. This criterion selects 24 of the 44 UAV-Diverse sequences.

UAV-Tracker achieves 63.93 HOTA, 76.30 IDF1, and 64.00 AssA. Relative to the strongest baseline, AMOT+SAM3Feat, these results represent improvements of 4.57 HOTA, 7.98 IDF1, and 7.79 AssA. UAV-Tracker also records the fewest identity switches, with 763 IDSW compared with 859 for the next-best method.

UAV-Tracker obtains similar sequence-level HOTA on the reappearance and non-reappearance subsets, with scores of 63.93 and 63.31. This result suggests that its performance remains stable on sequences in which identities leave the field of view and subsequently return.

Table 5: Tracking performance on the 24 UAV-Diverse sequences containing valid identity reappearance events. HOTA, IDF1, and AssA are macro-averaged over sequences, whereas IDSW is summed over the subset.

Method	**HOTA ↑**	**IDF1 ↑**	**AssA ↑**	**IDSW ↓**
UAVMOT (2022)	21.11	27.91	22.96	1444
MMTracker (2025)	26.18	31.48	27.49	1185
AMOT (2026)	25.56	32.78	26.99	859
AMOT+SAM3Feat	59.36	68.32	56.21	902
UAV-Tracker	**63.93**	**76.30**	**64.00**	**763**

A.3.2 TARGET-SCALE VARIATION

Following the sequence-level normalized target scale s_q defined in Appendix A.2, Table 6 reports HOTA across five fixed scale intervals. Each sequence is assigned to one interval according to s_q, and HOTA is averaged over all sequences in that interval.

Table 6: HOTA across target-scale regimes on UAV-Diverse. Sequences are grouped according to the normalized target scale s_q defined in Appendix A.2.

Method	**Scale (%)**				
	$[0, 3)$	$[3, 5)$	$[5, 7)$	$[7, 10)$	≥ 10
UAVMOT (2022)	4.35	14.40	18.87	31.98	46.00
MMTracker (2025)	0.31	21.72	22.65	39.11	56.76
AMOT (2026)	2.86	22.79	23.47	36.74	44.68
AMOT+SAM3Feat	18.35	71.02	61.62	62.70	73.18
UAV-Tracker	**38.33**	**73.17**	**64.49**	**66.69**	**73.62**

UAV-Tracker achieves the highest HOTA in every scale interval. Its largest margin occurs in the smallest-target regime, where $s_q < 3\%$. In this interval, UAV-Tracker reaches 38.33 HOTA and exceeds AMOT+SAM3Feat by 19.98 points. The corresponding margins in the remaining intervals are 2.15, 2.87, 3.99, and 0.44 HOTA.

Because s_q assigns a single scale regime to each sequence, the resulting subsets may also differ in viewpoint, camera motion, scene density, and appearance condition. These results should therefore be interpreted as condition-specific robustness rather than as an isolated causal effect of target scale.

A.3.3 VIEWPOINT VARIATION

A sequence may contain gradual viewpoint changes as the UAV moves. We assign each sequence the dominant viewpoint that best characterizes its overall observation geometry. Table 7 reports the mean sequence-level HOTA within each viewpoint category. The Macro column is the unweighted average of the bird's-eye, oblique, and side-view results.

Table 7: HOTA across UAV viewpoint conditions on UAV-Diverse.

Method	**Bird's-eye**	**Oblique**	**Side**	**Macro**
UAVMOT (2022)	8.10	37.35	39.41	28.29
MMTracker (2025)	10.54	43.27	50.10	34.64
AMOT (2026)	9.75	45.43	39.63	31.60
AMOT+SAM3Feat	53.83	61.43	**69.51**	61.59
UAV-Tracker	**62.85**	**62.65**	69.18	**64.89**

UAV-Tracker achieves the highest macro HOTA, reaching 64.89 compared with 61.59 for AMOT+SAM3Feat. Its largest advantage occurs under bird's-eye views, where it improves HOTA from 53.83 to 62.85. UAV-Tracker also leads under oblique views by 1.22 points. Under side views,

AMOT+SAM3Feat is marginally better, reaching 69.51 HOTA compared with 69.18 for UAV-Tracker. These results show that the largest viewpoint-conditioned gain occurs under bird's-eye observation, while the two methods perform similarly under side views.

A.3.4 Appearance Ambiguity

Table 8 evaluates 37 sequences containing non-everyday human appearance, including sports uniforms, swimwear in aquatic scenes, and outdoor gear. Each reported rate-based metric is macro-averaged over the sequences in this subset, while IDSW is summed.

Table 8: Sequence-level tracking performance on the 37 UAV-Diverse sequences containing non-everyday human appearance.

Method	**HOTA ↑**	**IDF1 ↑**	**AssA ↑**	**IDSW ↓**
UAVMOT (2022)	23.25	30.40	25.11	1612
MMTracker (2025)	27.65	33.48	29.34	1010
AMOT (2026)	26.70	33.23	29.17	<u>922</u>
AMOT+SAM3Feat	<u>57.23</u>	<u>66.26</u>	<u>55.67</u>	999
UAV-Tracker	**63.72**	**75.53**	**65.85**	**740**

UAV-Tracker achieves 63.72 HOTA, 75.53 IDF1, and 65.85 AssA, outperforming AMOT+SAM3Feat by 6.49, 9.27, and 10.18 points. It also records the fewest identity switches, with 740 IDSW compared with 922 for the next-best method. These results show that the gains extend beyond everyday-clothing sequences to athletic, aquatic, and outdoor-gear conditions.

The appearance categories represent coarse semantic conditions rather than an ordered measure of visual difficulty. Differences within this subset may also be influenced by target scale, viewpoint, scene density, and camera motion.

A.3.5 Camera Motion

UAV camera trajectories often combine translation, altitude changes, and rotation rather than exhibiting a single isolated motion component. We therefore group each sequence according to its *dominant camera-motion condition*. The category labels are mutually exclusive for evaluation, but this does not imply that other motion components are absent. A sequence is assigned to the mixed category when multiple motion patterns are visually prominent and no single pattern adequately characterizes the resulting observation changes. The Macro column is the unweighted average of the six category-level HOTA values.

Table 9: HOTA across dominant camera-motion conditions on UAV-Diverse.

Method	**Stat.**	**Horiz.**	**Asc.**	**Desc.**	**Rot.**	**Mix.**	**Macro**
UAVMOT (2022)	14.86	27.88	35.68	23.16	23.96	28.54	25.68
MMTracker (2025)	18.03	29.66	47.02	28.28	32.84	32.97	31.47
AMOT (2026)	18.15	28.17	42.99	28.68	28.45	33.16	29.93
AMOT+SAM3Feat	<u>55.17</u>	<u>60.50</u>	<u>67.59</u>	<u>57.92</u>	**61.40**	<u>59.06</u>	<u>60.28</u>
UAV-Tracker	**61.48**	**74.90**	**68.57**	**60.46**	<u>61.34</u>	**60.39**	**64.52**

As shown in Table 9, UAV-Tracker achieves the highest HOTA in five of the six camera-motion conditions. Its macro HOTA is 64.52, compared with 60.28 for AMOT+SAM3Feat. The largest margin occurs under horizontal motion, where UAV-Tracker improves HOTA from 60.50 to 74.90. It also leads under static, ascent, descent, and mixed motion by 6.31, 0.98, 2.54, and 1.33 points. Under rotation, the two methods are nearly tied, with AMOT+SAM3Feat reaching 61.40 HOTA and UAV-Tracker reaching 61.34.

Because camera motion may co-vary with viewpoint, target scale, and scene content, this analysis should be interpreted as condition-specific robustness rather than as evidence of an isolated causal effect. The near-identical rotation results also indicate that neither method has a clear advantage in this condition.

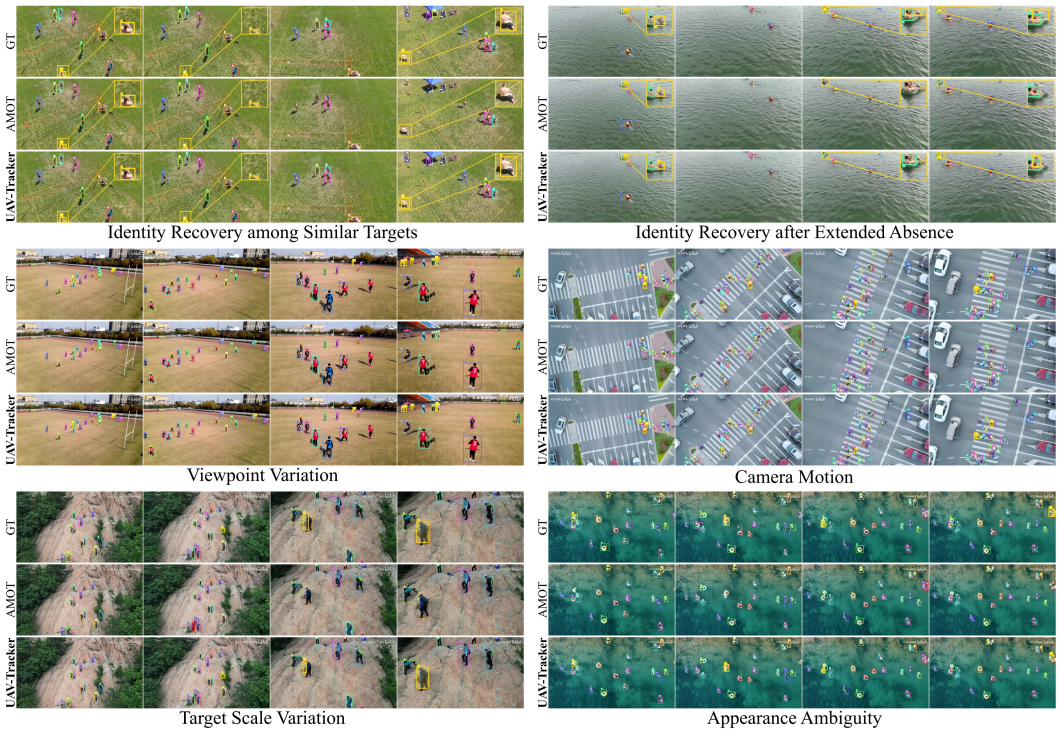


Figure 5: Qualitative comparison between UAV-Tracker and AMOT+SAM3Feat.

Overall observation. Across the condition-specific analyses, UAV-Tracker leads on the valid reappearance and non-everyday appearance subsets, in all five target-scale intervals, under bird's-eye and oblique viewpoints, and in five of the six camera-motion categories. Its largest HOTA margins occur in the smallest-target interval, under horizontal camera motion, and under bird's-eye views. These results show that the improvements are not confined to a single UAV condition. Because the sequence-level subsets may differ along multiple factors, they should not be interpreted as isolated causal effects of individual challenge dimensions.

A.4 Additional Qualitative Results

Figure 5 presents six qualitative examples covering the five representative challenges in UAV-Diverse, with identity recovery illustrated under both similar-target ambiguity and extended absence. Both methods use the same pre-generated SAM3 detections and frozen SAM3 object features, while retaining their respective identity-management and trajectory-realization strategies.

In the *Identity Recovery among Similar Targets* example, the highlighted target temporarily disappears and later returns near visually similar people. AMOT+SAM3Feat fails to reliably reconnect the returning observation to its earlier trajectory, whereas UAV-Tracker recovers the corresponding sequence-local identity. In the *Identity Recovery after Extended Absence* example, the highlighted target disappears for an extended interval between the selected frames. AMOT+SAM3Feat produces an inconsistent post-gap identity assignment, while UAV-Tracker reactivates the earlier identity and more closely matches the ground truth. Together, these two cases illustrate identity recovery under both appearance ambiguity and a prolonged visibility gap.

In the *Viewpoint Variation* example, the descending camera changes the observation from a distant oblique view to a closer, near-side view, substantially altering target scale, appearance, and geometry. AMOT+SAM3Feat produces less complete and less consistent trajectories during this transition, whereas UAV-Tracker more closely follows the ground-truth identities across the selected viewpoints. In the *Camera Motion* example, camera rotation changes the image orientation and the relative layout of the crowded scene. UAV-Tracker retains more complete target coverage and identity assignments that are more consistent with the ground truth, particularly within the dense pedestrian group.

Table 10: Tracking performance on VisDrone-MOT under the human-only setting. Detection, Query, and UAV denote tracking-by-detection, tracking-by-query, and UAV-oriented methods. ByteTrack, BoT-SORT, and OC-SORT use identical pre-generated detections from frozen SAM3. The remaining baselines retain their released configurations, while UAV-Tracker uses the fixed SAM3-based configuration described in the main paper. All values are percentages except IDSW.

Type	Method	HOTA ↑	DetA ↑	AssA ↑	IDF1 ↑	MOTA ↑	IDSW ↓
Detection	ByteTrack (2022)	33.15	28.97	38.92	41.83	24.32	1141
	BoT-SORT (2022)	33.10	30.15	37.48	40.74	20.58	3342
	OC-SORT (2023)	31.60	26.93	38.07	39.89	21.80	986
Query	MeMOTR (2023)	27.79	19.01	41.25	33.41	19.72	**379**
	MOTIP (2025)	28.12	24.17	33.49	34.17	20.44	1842
UAV	UAVMOT (2022)	22.52	14.81	34.61	25.33	15.92	793
	MMTracker (2025)	33.70	29.91	38.66	41.26	**29.20**	1105
	AMOT (2026)	35.01	28.56	43.65	44.93	27.54	853
	UAV-Tracker (ours)	**37.47**	**30.93**	**46.39**	**47.32**	18.21	674

In the *Target-Scale Variation* example, the apparent target size increases substantially as the camera approaches the climbing area. The highlighted target remains detected by AMOT+SAM3Feat after the scale transition, but its identity becomes fragmented, as indicated by the change in bounding-box color. UAV-Tracker instead preserves the same identity across the distant and close-range observations. This example therefore illustrates identity continuity under substantial changes in target scale and partial visibility.

Finally, the *Appearance Ambiguity* example contains multiple swimmers with similar appearances against repetitive water patterns. AMOT+SAM3Feat omits or fragments several trajectories across the selected frames, whereas UAV-Tracker produces more complete coverage and identity assignments that more closely match the ground truth.

These qualitative examples complement the quantitative results and show that UAV-Tracker better handles prolonged disappearance, changes in observation geometry, camera motion, target-scale variation, and ambiguous human appearance. These examples are consistent with the intended roles of lifecycle-based identity recovery and context-guided trajectory stabilization.

A.5 Cross-Dataset Evaluation on VisDrone-MOT

A.5.1 Evaluation Protocol.

We evaluate UAV-Tracker on VisDrone-MOT to assess its transferability beyond UAV-Diverse. For all methods, evaluation is restricted to the two human categories in VisDrone, namely pedestrian and people, corresponding to category IDs 1 and 2.

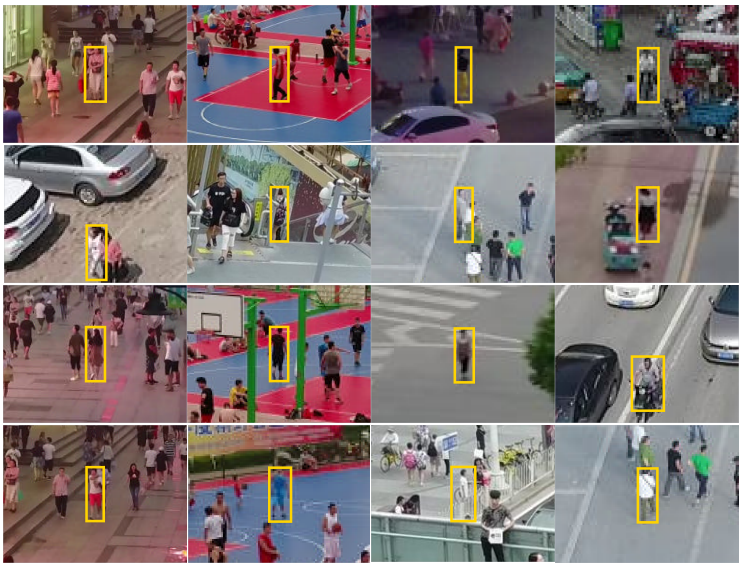
All metrics are computed using TrackEval. Before metric computation, we apply the ignore-region exclusion procedure implemented in the official VisDrone evaluation code. The remaining ground-truth annotations and tracking outputs are then passed to TrackEval. Thus, ignore regions are handled according to the official VisDrone implementation, while HOTA, DetA, AssA, IDF1, MOTA, and IDSW are computed with TrackEval.

The training setup of each method follows its publicly released configuration. MeMOTR and MOTIP are trained on the official VisDrone training split using their released training procedures and hyperparameters, with training restricted to categories 1 and 2. UAVMOT is trained according to its official VisDrone recipe, which retains all annotated object categories during training. For MMTracker and AMOT, we use the pretrained weights officially released by the authors. Regardless of the categories used during training, all methods are evaluated exclusively on categories 1 and 2.

ByteTrack, BoT-SORT, and OC-SORT use the same pre-generated SAM3 detections. This setting controls for detector differences within the tracking-by-detection group. The remaining methods retain their respective released training and inference configurations.



(a) Oversized GT Boxes Yield IoU < 0.5



(b) Predictions on Unannotated Targets

Figure 6: Representative annotation issues in the human-only VisDrone-MOT evaluation. Red boxes denote ground truth, while yellow boxes denote UAV-Tracker predictions. Panel (a) presents oversized ground-truth boxes that yield an IoU below 0.5; panel (b) contains predictions on visible but unannotated human targets outside the official ignore regions.

A.5.2 Cross-dataset Results

Table 10 reports the human-only VisDrone results after applying the official ignore-region filtering procedure. Among all compared methods, UAV-Tracker achieves the highest HOTA, DetA, AssA, and IDF1, reaching 37.47, 30.93, 46.39, and 47.32. Compared with AMOT, the strongest baseline in terms of HOTA, UAV-Tracker improves HOTA by 2.46 points, AssA by 2.74 points, and IDF1 by 2.39 points. MMTracker obtains the highest MOTA, whereas UAV-Tracker performs best on HOTA, DetA, AssA, and IDF1. These results indicate that the identity maintenance and track realization mechanisms of UAV-Tracker remain effective when transferred to a different UAV benchmark. In particular, the improvements in HOTA and the association-related metrics demonstrate stronger tracking quality and identity consistency under this cross-dataset setting.

A.5.3 Annotation Sensitivity

VisDrone remains an important and challenging benchmark for UAV perception. Human targets in VisDrone are often small, densely distributed, and weakly resolved. However, VisDrone was designed as a general-purpose, multi-category UAV dataset rather than as a benchmark dedicated specifically to human tracking. Consequently, some characteristics of its human annotations can have a particularly strong influence when evaluation is restricted to the two human categories.

Figure 6 illustrates two representative annotation cases. As shown in (a), some ground-truth bounding boxes contain substantially more background than the visible extent of the person. Prior work has shown that IoU is particularly sensitive to bounding box annotation discrepancies for small objects, even when the coordinate differences are minor (Murrugarra-Llerena et al., 2022). Consistent with this observation, the selected cases in Figure 6(a) have IoUs ranging from 0.482 to 0.4998. Consequently, these predictions fall below the commonly used 0.5 IoU criterion despite appearing visually well localized.

Figure 6 (b) shows predictions on clearly visible human targets without corresponding ground-truth annotations. We manually inspected the official ignore-region annotations for every highlighted example in panel (b). None of these targets lies inside an official ignore region. Their predictions are therefore not removed by the official VisDrone ignore-region exclusion procedure and remain in the inputs passed to TrackEval. Since no corresponding human annotation is available, these predictions may be counted as false positives.

These examples show that human-only evaluation in aerial imagery is particularly sensitive to annotation completeness and bounding-box accuracy. This sensitivity is further amplified for small targets, where a displacement of only a few pixels can substantially alter the matching outcome. Accurate, tight, and complete human annotations are therefore especially important for future benchmarks on small-target detection and multi-object tracking in UAV videos. This analysis does not alter the official evaluation results, but provides context for interpreting human-only tracking errors on small aerial targets.

A.6 UAV-Tracker Implementation Details

Development protocol. We use an *external development set* comprising five rugby videos to select the fixed inference parameters of UAV-Tracker. The development videos are disjoint from all UAV-Diverse evaluation sequences. The development set is used only for global parameter selection and is not used to train SAM3. After parameter selection, the configuration is frozen and applied unchanged to all UAV-Diverse and VisDrone-MOT sequences.

Fixed inference configuration. UAV-Tracker uses a single global inference configuration for all evaluation sequences. Detector observations are partitioned using $\tau_{\text{low}} = 0.45$ and $\tau_{\text{high}} = 0.65$. Appearance-based recovery uses a base similarity threshold of $\gamma_{\text{reid}} = 0.85$, a matching margin of $\Delta = 0.02$, and a layout-consistency threshold of $\epsilon_{\text{layout}} = 0.08$. The recovery threshold increases deterministically with the size of the recoverable identity pool, up to 0.94. SAM3-only propagation uses a base budget of $B_{\text{base}} = 5$ frames. As described in the main paper, the shared scene policy deterministically controls the effective recovery and emission gates. All settings are shared across sequences.

Scene fingerprint computation. The Scene Fingerprint summarizes recent tracking diagnostics into the five-dimensional state introduced in Eq. (6),

$$\mathbf{f}_{t-1} = [f_{\text{cov}}, f_{\text{id}}, f_{\text{mot}}, f_{\text{geo}}, f_{\text{prop}}]_{t-1}. \tag{12}$$

It is computed only after frame $t-1$ has been finalized and is therefore available for the policy applied at frame t, preserving the causal update in Eq. (7). All statistics are aggregated over the causal window

$$\mathcal{W}_{t-1} = \{\max(1, t-30), \ldots, t-1\}. \tag{13}$$

The fingerprint uses detector observations, frame-level MOT outputs, and association diagnostics only; no ground-truth identities, scene labels, challenge labels, or future frames are accessed. Bounding-box coordinates are normalized by image width and height. We write $[x]_0^1 = \min(1, \max(0, x))$ for clipping to $[0, 1]$.

The five components capture complementary tracking diagnostics. f_{cov} summarizes coverage deficiency from uncovered high-confidence detections and clear coverage gaps. f_{id} captures online identity ambiguity from detection-output conflicts, association ambiguity, and short-term identity turnover. f_{mot} measures motion instability from inter-frame displacements of matched detections. f_{geo} summarizes target-scale dispersion and spatial crowding, while f_{prop} measures output excess relative to high-confidence detector support.

Coverage statistics. Let $\mathcal{D}_s$ denote the retained detections with confidence at least 0.65 at frame s, and let $\mathcal{O}_s$ denote the frame-level MOT outputs. For each $d \in \mathcal{D}_s$, define

$$q_s(d) = \max_{o \in \mathcal{O}_s} \mathrm{IoU}(d, o), \tag{14}$$

with $q_s(d) = 0$ when $\mathcal{O}_s$ is empty. A detection is considered uncovered when $q_s(d) < 0.5$ and is counted as a clear coverage gap when $q_s(d) < 0.1$. Uncovered detections with $0.1 \le q_s(d) < 0.5$ are treated as potential conflicts with existing outputs.

Let N_s, U_s, G_s, and C_s denote the numbers of high-confidence detections, uncovered detections, clear coverage gaps, and potential conflicts, respectively. Let A_s count uncovered high-confidence detections that have a raw SAM3 association candidate before filtering but no retained matched object after association filtering. Over $\mathcal{W}_{t-1}$, define

$$\begin{aligned} u_{t-1} &= \frac{\sum_s U_s}{\max(1, \sum_s N_s)}, \quad g_{t-1} = \frac{\sum_s G_s}{\max(1, \sum_s N_s)}, \\ a_{t-1} &= \frac{\sum_s A_s}{\max(1, \sum_s N_s)}, \quad d_{t-1} = \frac{\sum_s C_s}{\max(1, \sum_s U_s)}, \end{aligned} \tag{15}$$

where all sums are over $s \in \mathcal{W}_{t-1}$. Here, d_{t-1} measures conflict risk conditional on a detection being uncovered. The coverage component is

$$f_{\mathrm{cov},t-1} = [0.55 u_{t-1} + 0.45 g_{t-1}]_0^1. \tag{16}$$

Geometric variation. For each frame, we compute four statistics from high-confidence detections: the coefficient of variation of bounding-box areas, the median nearest-neighbor center distance, the fraction of unordered detection pairs with IoU greater than 0.1, and the number of high-confidence detections. Center coordinates are normalized by image width and height, and center distances are divided by $\sqrt{2}$. Let v_{t-1}, δ_{t-1}, and o_{t-1} denote the window averages of the first three statistics, and let

$$n_{t-1} = \frac{1}{|\mathcal{W}_{t-1}|} \sum_{s \in \mathcal{W}_{t-1}} N_s \tag{17}$$

be the mean high-confidence detection count. The geometric component is

$$\begin{aligned} f_{\mathrm{geo},t-1} = \Bigg[0.30 \left[\frac{v_{t-1} - 0.30}{0.70} \right]_0^1 + 0.25 \left[\frac{0.08 - \delta_{t-1}}{0.08} \right]_0^1 \\ + 0.25 \left[\frac{o_{t-1}}{0.05} \right]_0^1 + 0.20 \left[\frac{n_{t-1}}{30} \right]_0^1 \Bigg]_0^1. \end{aligned} \tag{18}$$

This component summarizes scale dispersion and spatial crowding and does not explicitly estimate camera pose.

Identity ambiguity. Let $\mathcal{V}_s$ denote the set of identities emitted at frame s. We measure frame-level identity turnover as

$$\chi_s = \frac{|\mathcal{V}_s \setminus \mathcal{V}_{s-1}| + |\mathcal{V}_{s-1} \setminus \mathcal{V}_s|}{\max(1, |\mathcal{V}_{s-1}|)}, \tag{19}$$

with $\chi_1 = 0$. Let $\bar{\chi}_{t-1}$ denote its average over $\mathcal{W}_{t-1}$. The identity component is

$$f_{\mathrm{id},t-1} = [0.40 d_{t-1} + 0.25 a_{t-1} + 0.20 \min(1, \bar{\chi}_{t-1}) + 0.15 f_{\mathrm{geo},t-1}]_0^1. \tag{20}$$

This quantity is an online risk indicator derived from current tracking diagnostics rather than a ground-truth identity-error measure.

Motion instability. High-confidence detections in consecutive frames are greedily matched in descending IoU order, requiring an IoU of at least 0.3, with each detection participating in at most one match. For every matched pair, we compute the center displacement after normalizing coordinates by image width and height and dividing the distance by $\sqrt{2}$. Let μ_s denote the median displacement at frame s, set to zero when no valid match exists, and let $\bar{\mu}_{t-1}$ denote its average over $\mathcal{W}_{t-1}$. The motion component is

$$f_{\mathrm{mot},t-1} = [8 \bar{\mu}_{t-1} + 0.30 a_{t-1} + 0.15 f_{\mathrm{geo},t-1}]_0^1. \tag{21}$$

Propagation excess. Let

$$\ell_{t-1} = \frac{1}{|\mathcal{W}_{t-1}|} \sum_{s \in \mathcal{W}_{t-1}} |\mathcal{O}_s| \tag{22}$$

be the mean number of frame-level MOT outputs. The propagation component is

$$f_{\mathrm{prop},t-1} = \left[\frac{\ell_{t-1}/(n_{t-1}+1) - 1}{3} \right]_0^1. \tag{23}$$

This component is a proxy for output excess relative to detector support rather than a direct count of propagation-only tracks.

Initialization and implementation details. Before 30 frames are available, all statistics are computed from the available causal history without zero padding, and all running statistics are reset at the beginning of each sequence. The area coefficient of variation uses $\max(\mathrm{mean}(\mathrm{area}), 10^{-6})$ as its denominator and is set to zero when no detection exists. The nearest-neighbor statistic is set to one when fewer than two detections exist, and the pair-overlap fraction is set to zero when no pair exists. All count denominators above are lower-bounded by one. The coefficients and normalization constants are fixed in the implementation and are not updated during inference.

Baseline training protocol. For baseline methods without publicly released pretrained checkpoints, we train them exclusively on the official VisDrone-MOT training split using their publicly released training configurations. We retain the original optimization settings, training schedules, hyperparameters, and random seeds specified by the respective implementations. Specifically, the configured seeds are 42 for MOTIP and MeMOTR, and 317 for UAVMOT. No UAV-Diverse annotations are used for training, fine-tuning, hyperparameter selection, or seed selection.

Runs and computing environment. Experiments are conducted on a server running Ubuntu 22.04 with NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of VRAM. Each inference run uses a single GPU. UAV-Tracker inference and evaluation are performed using Python 3.12, PyTorch 2.10, and CUDA 12.8.

For each method, we perform one inference run on every evaluation sequence and compute the reported aggregate metrics from the resulting outputs. UAV-Tracker is deterministic given the fixed pre-generated SAM3 observations, and no test-time ensembling is used.