MOTION-AWARE TRANSFORMER FOR MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-object tracking with query propagation maintains object identities by carrying track queries across frames, but the propagated state can become misaligned with the current observation before it interacts with newly initialized detection queries. We show that this temporal state misalignment can create *query ownership conflicts*, where a detection query represents an existing target better than its designated track-query owner. To address this problem, we introduce MATR, an end-to-end tracking framework built around a Motion-Aware Transformer (MAT) module. MAT uses current-frame evidence to pre-align both the feature and spatial states of inherited track queries before joint decoding. We further derive a geometric ambiguity measure and show empirically that ownership-conflict risk increases sharply with ambiguity, while MAT applies larger corrections to more difficult states. Across DanceTrack, SportsMOT, and BDD100K, MATR consistently improves association while introducing only modest computational overhead. On DanceTrack, MATR reaches 71.3 HOTA with 43M parameters and 179 GFLOPs. On BDD100K, without additional tracking data or external detector proposals, MATR reaches 54.7 mTETA and 41.6 mHOTA, the strongest result we are aware of among query-propagation trackers trained under this native-data setting. Source code will be publicly released upon acceptance.

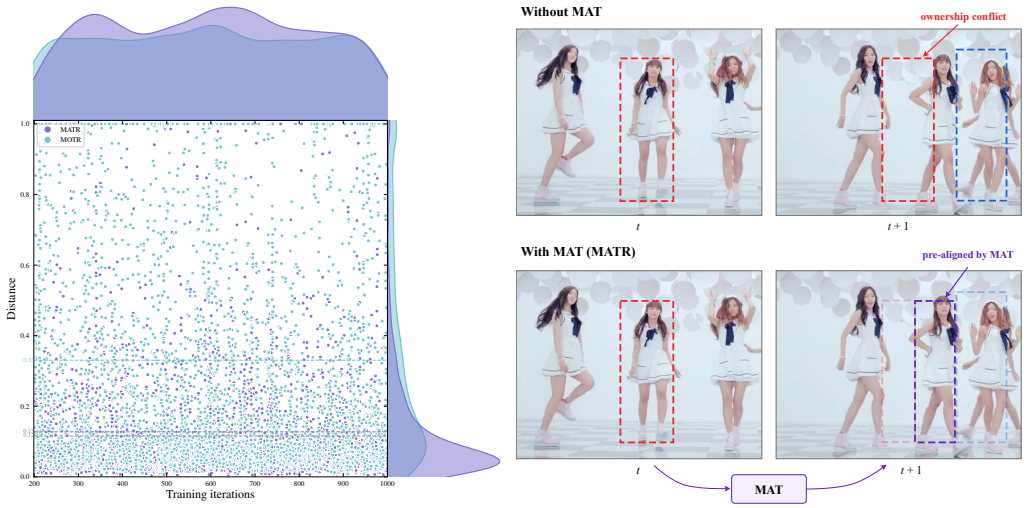
1 INTRODUCTION

Multi-object tracking (MOT) aims to localize multiple objects while preserving their identities throughout a video. Recent end-to-end trackers have increasingly replaced explicit association pipelines with query propagation: once an object is detected, its query is carried across frames as a track query and continuously updated to represent the same identity (Meinhardt et al., 2022; Zeng et al., 2022). This formulation is appealing because detection and association can be learned within a unified model. However, jointly decoding newly initialized detection queries and propagated track queries also creates nontrivial interactions between the two query types. Recent methods address this issue from different perspectives, for example by modifying label assignment (Yu et al., 2023; Yan et al., 2025) or separating detection and association (Wang et al., 2025). In this work, we study a complementary problem that arises *before* these queries interact: the temporal state carried by a track query can be misaligned with the current observation.

A propagated track query has two distinct responsibilities. Its identity must remain consistent with the trajectory inherited from the previous frame, while its feature and spatial state must adapt to the object in the current frame. These two requirements are not automatically satisfied together. Under target-consistent assignment, an existing target remains owned by its inherited track query, but object or camera motion can make the propagated state stale.

Figure 1 illustrates the problem. In MOTR, direct propagation yields a broad distribution of track-query localization errors ($1-\mathrm{IoU}$ with the ground-truth target; mean 0.33), with a heavy tail of states that barely overlap their targets (Fig. 1a); MATR reduces the mean to 0.13, close to the 0.12 reference obtained by Hungarian matching. Such misalignment matters because a stale track state can describe its target worse than a competing detection query (Fig. 1b). The identity is not reassigned—target-consistent assignment keeps the inherited track query as the legal owner—but the shared decoder must correct a poor owner representation in the presence of a stronger competitor. We call this

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.



(a) (b)

Figure 1: Motivation of MATR. (**a**) Track-query localization error ($1 - \mathrm{IoU}$) throughout training, together with its marginal distributions. MATR (purple) shifts the distribution toward lower error compared with MOTR (cyan). (**b**) Illustration of query ownership conflict: a propagated track state (red) can become stale, while a current-frame detection hypothesis (blue) provides a stronger competing representation. MAT pre-aligns the inherited track state (purple) before joint decoding.

a *query ownership conflict*; it becomes more likely when inter-frame motion is large or targets are close.

This motivates a simple question: *can the inherited track state be aligned with the current frame before detection and track queries compete in the shared decoder?* We propose **MATR**, built around a Motion-Aware Transformer (MAT) module that uses the current-frame encoder memory to update both the features and the spatial references of propagated track queries before joint decoding. Unlike a motion predictor that extrapolates from history, MAT observes the current frame; we therefore view it as *observation-conditioned pre-alignment*. MATR keeps target-consistent ownership and the shared decoder unchanged and adds no association procedure.

To examine this mechanism directly, we define an *ownership margin* that measures when a detection query represents an existing target better than its legal track-query owner, and a geometric ambiguity ratio $\rho = 2e/d$ (e: track-reference error; d: distance to the nearest other target), for which $\rho < 1$ guarantees that the reference is closer to its own target than to any other. Ownership-conflict risk rises sharply with ρ, and MAT applies its largest corrections to the most ambiguous states.

Our contributions are summarized as follows:

- We identify and formalize *temporal state misalignment* in query-propagation tracking, and show how it can produce query ownership conflicts under target-consistent assignment.
- We introduce MATR, which performs observation-conditioned pre-alignment of both feature and spatial states before detection and track queries enter the shared decoder, without changing the underlying ownership rule.
- We provide a geometric analysis of track-state ambiguity and directly connect it to ownership conflict, showing that conflict risk increases with ambiguity while MAT applies larger corrections to more difficult states. Controlled ablations and multiple benchmarks further validate the resulting association improvements at modest computational cost.

2 Related Work

2.1 Tracking-by-Detection

Tracking-by-detection (TbD) decomposes multi-object tracking into per-frame detection followed by temporal association. Early methods such as SORT (Bewley et al., 2016) combine Kalman filtering with bipartite matching, while DeepSORT (Wojke et al., 2017) augments motion with appearance embeddings. Later approaches improve association through stronger detection, re-identification, motion modeling, or track management (Zhang et al., 2021; 2022; Cao et al., 2023; Guo et al., 2024), and learned motion predictors replace the Kalman filter for non-linear motion (Lv et al., 2024). Although these methods remain highly competitive, their association stage is typically designed separately from the detector. In contrast, our work focuses on query-propagation trackers, where object identity and state are maintained directly inside an end-to-end Transformer.

2.2 Query-Based Tracking

Query-based trackers integrate detection and association by propagating object queries through time. TrackFormer (Meinhardt et al., 2022) introduced track queries into DETR (Carion et al., 2020), while MOTR (Zeng et al., 2022), built on Deformable DETR (Zhu et al., 2021), established target-consistent assignment and recurrent query propagation. Subsequent methods improve this paradigm from different directions. MeMOTR (Gao & Wang, 2023) strengthens temporal representation using long-term memory, SambaMOTR (Segu et al., 2025) jointly models track-query histories with synchronized state-space models, and MOTRv2 (Zhang et al., 2023) bootstraps tracking with a pre-trained detector. MOTRv3 (Yu et al., 2023) and CO-MOT (Yan et al., 2025) instead revisit training supervision and label assignment, providing detection queries with richer positive supervision.

Recent work has increasingly reconsidered how detection and association should interact. LA-MOTR (Wang et al., 2025) decouples the two tasks and learns association explicitly between independently produced object and tracklet representations. MOTIP (Gao et al., 2025) departs further from query propagation by formulating association as in-context identity prediction over current detections and historical trajectories. FDTA (Shao et al., 2026) refines insufficiently discriminative DETR object embeddings through spatial, temporal, and identity-aware adaptation, and SelfMOTR (Gülhan et al., 2026) shows that track queries suppress MOTR's latent detection ability and runs a detection-only decoder pass to generate priors. These approaches change the representation used for matching or the supervision that produces it, and they improve the detection side rather than the propagated state: SelfMOTR generates priors through a separate detection-only decoder pass, and CO-MOT rebalances label assignment between the two query types. MAT is complementary to both, since it leaves assignment, supervision, and the number of decoder passes unchanged.

MATR addresses a complementary point in the pipeline. Rather than changing target ownership, replacing query propagation with an explicit ID-prediction task, or decoupling the shared decoder, we retain the target-consistent query-propagation formulation and operate on the *temporal state presented to the decoder*. Unlike history-based state models such as MeMOTR and SambaMOTR, MAT uses the current observation to pre-align the inherited feature and spatial states before detection and track queries interact, which lets us isolate temporal state misalignment and its relationship with ownership conflict.

3 Approach

In MOTR-style trackers, *identity is propagated across time, while the state carried by the query may still describe the previous observation*. We formalize the resulting query conflict and then introduce MAT.

3.1 Temporal Misalignment of Track Queries

Let M^t denote the encoder memory extracted from frame I^t, and let $\mathcal{S}_{\mathrm{trk}}^{t-1} = (Q_{\mathrm{trk}}^{t-1}, R_{\mathrm{trk}}^{t-1})$ denote the propagated track states, where Q_{trk} and R_{trk} denote the track-query features and their spatial references, respectively.

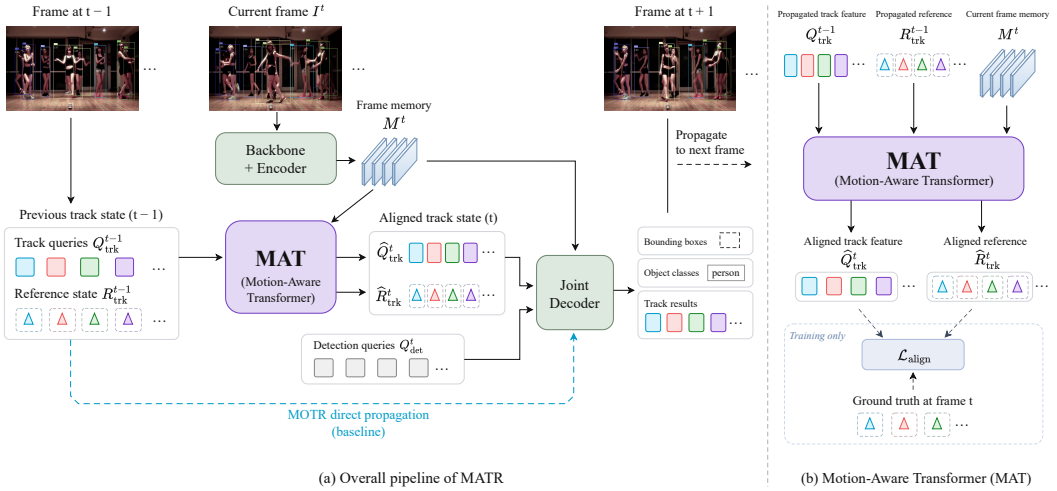


Figure 2: Overview of MATR. (**a**) At frame t, MAT uses the current-frame memory M^t to pre-align the propagated track state before it is combined with detection queries in the joint decoder. The dashed cyan path denotes the direct track-state propagation used by MOTR. (**b**) MAT updates both the feature and spatial-reference components of the inherited track state. During training, the intermediate aligned box is supervised by the auxiliary ℓ_1 alignment loss.

Under target-consistent assignment, an existing object retains its inherited track query, while detection queries are matched only to newborn objects. This preserves identity continuity but not alignment: after object or camera motion, the inherited query may enter frame t with a feature and spatial reference still biased toward frame $t-1$. To measure when a detection query describes an existing target better than its legal owner, for an existing ground-truth object y_i^t with a valid inherited track query $q^t_{\text{trk},i}$, we define the *ownership margin*

$$\Delta_i^t = C\big(q^t_{\text{trk},i}, y_i^t\big) - \min_{q^t_{\text{det},k} \in Q^t_{\text{det}}} C\big(q^t_{\text{det},k}, y_i^t\big)\,, \tag{1}$$

where C is the original matching cost composed of focal (Lin et al., 2017), ℓ_1, and GIoU terms. When $\Delta_i^t > 0$, at least one detection query has a lower matching cost to the target than its designated track query; we call this a *query ownership conflict*. This is not a matching error: the inherited track query remains the legal owner, but the shared decoder must correct a comparatively poor track state in the presence of a stronger competitor. Our goal is therefore to improve the state of the legal owner *before this competition occurs*, not to change the assignment rule.

3.2 Motion-Aware Transformer

Based on the observation above, we introduce the Motion-Aware Transformer (MAT), illustrated in Fig. 2. Rather than directly forwarding $\big(Q^{t-1}_{\text{trk}}, R^{t-1}_{\text{trk}}\big)$ to the main decoder, MAT first updates the inherited track states using the current-frame memory M^t.

Given the propagated track queries, MAT performs self-attention among existing tracks followed by cross-attention with the current observation:

$$\widehat{Q}^t_{\text{trk}} = Q^{t-1}_{\text{trk}} + \text{CrossAtt}\left(\text{SelfAtt}\left(Q^{t-1}_{\text{trk}}\right), M^t\right). \tag{2}$$

The residual connection preserves the inherited track representation, while cross-attention allows each track query to retrieve evidence from the current frame before interacting with detection queries.

Feature alignment alone, however, leaves the spatial reference unchanged. This creates an undesirable situation in which the updated feature may describe the current observation while its positional state still points toward the previous one. MAT therefore predicts an aligned box and converts it into the updated spatial reference,

$$\widehat{B}^t_{\text{align}} = f_{\text{box}}\big(\widehat{Q}^t_{\text{trk}}\big), \qquad \widehat{R}^t_{\text{trk}} = \text{PE}\left(\widehat{B}^t_{\text{align}}\right). \tag{3}$$

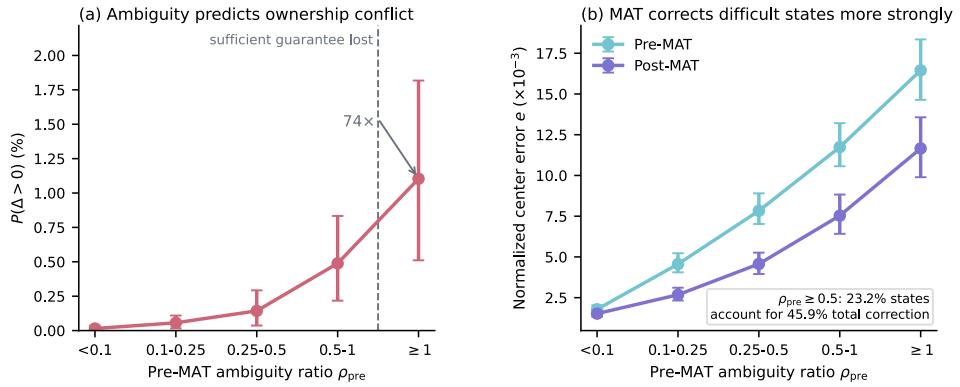


Figure 3: Mechanism analysis of MATR. (**a**) Ownership-conflict probability increases with the pre-MAT ambiguity ratio ρ_{pre}. (**b**) MAT reduces track-state localization error across ambiguity levels, with larger corrections on more ambiguous states. Error bars denote 95% confidence intervals obtained by video-level bootstrap.

The aligned state $\widehat{\mathcal{S}}^t_{\text{trk}} = (\widehat{Q}^t_{\text{trk}}, \widehat{R}^t_{\text{trk}})$ is concatenated with the detection queries and passed to the original joint decoder:

$$O^t = \mathcal{D}\left(\left[Q^t_{\text{det}}; \widehat{Q}^t_{\text{trk}}\right], \left[R^t_{\text{det}}; \widehat{R}^t_{\text{trk}}\right], M^t\right). \tag{4}$$

MAT thus refreshes the *feature state* with current-frame evidence and moves the *spatial state* toward the current target location (contributions ablated in Sec. 4). Unlike a history-only extrapolator such as a Kalman filter (Welch & Bishop, 1995), MAT has access to M^t: motion produces the mismatch, and the current observation corrects it.

Alignment supervision. The final tracking objective already supervises the output of the joint decoder, but provides no direct requirement that the intermediate MAT prediction itself be spatially aligned with the current target. We therefore add an auxiliary ℓ_1 alignment loss for valid inherited tracks:

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_t \sum_{i \in \Omega^t} \left\| \widehat{b}^t_{\text{align},i} - b^t_{\text{gt},i} \right\|_1, \tag{5}$$

where Ω^t contains ground-truth objects present at frame t that have a legal inherited track query, and $N = \sum_t |\Omega^t|$. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{track}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \tag{6}$$

where $\mathcal{L}_{\text{track}}$ is the original tracking objective. The auxiliary prediction introduces no additional association heuristic at inference time.

3.3 When Does Misalignment Become Ambiguous?

The ownership margin in Eq. (1) measures query competition using the complete matching cost. We next consider a simpler geometric question: *when is a propagated spatial reference guaranteed to remain closer to its own target than to any other target?*

Let c^t_i be the center of target i and $\widetilde{c}^t_i$ the center represented by its inherited track state. We define the normalized center error as

$$e^t_i = \frac{\|\widetilde{c}^t_i - c^t_i\|_2}{D_{\text{img}}}, \tag{7}$$

where D_{img} is the image diagonal. We further define the distance to the nearest other ground-truth target as

$$d^t_i = \frac{\min_{j \neq i} \left\|c^t_i - c^t_j\right\|_2}{D_{\text{img}}}. \tag{8}$$

Remark 1. If $e_i^t < d_i^t/2$, then $\widetilde{c}_i^t$ is closer to c_i^t than to the center of any other target. *Proof.* For any $j \neq i$, the triangle inequality gives $\|\widetilde{c}_i^t - c_j^t\|_2 \geq \|c_i^t - c_j^t\|_2 - \|\widetilde{c}_i^t - c_i^t\|_2 \geq D_{\text{img}}(d_i^t - e_i^t) > D_{\text{img}}\, e_i^t = \|\widetilde{c}_i^t - c_i^t\|_2$. □

This defines the *ambiguity ratio*

$$\rho_i^t = \frac{2e_i^t}{d_i^t}, \tag{9}$$

for which $\rho_i^t < 1$ is a sufficient guarantee that the propagated reference lies closer to its own target than to every neighboring target.

Denoting the error before and after MAT by e_{pre} and e_{post}, MAT aims to make $e_{\text{post}} < e_{\text{pre}}$, lowering ρ for fixed scene geometry. Note that ρ is a continuous, purely geometric measure of ambiguity: $\rho \geq 1$ only removes the guarantee and does not imply a conflict, whereas Δ also includes classification and box compatibility. ρ thus characterizes the ambiguity induced by misalignment, and Δ whether it manifests as an actual ownership conflict.

Empirical mechanism analysis. We next examine whether the geometric ambiguity characterized by ρ is reflected in the actual query competition measured by Δ. As shown in Fig. 3(a), ownership-conflict probability increases monotonically with ρ_{pre}. States with $\rho_{\text{pre}} \geq 1$ exhibit approximately $74\times$ the conflict probability of those with $\rho_{\text{pre}} < 0.1$, confirming that ρ provides a meaningful continuous measure of ownership ambiguity rather than merely a binary geometric threshold.

Figure 3(b) further shows that MAT responds preferentially to these difficult states. The normalized center error decreases across all ambiguity levels, with larger corrections as ρ_{pre} increases. Overall, MAT reduces the mean normalized center error by 34.9%; more notably, the 23.2% of states with $\rho_{\text{pre}} \geq 0.5$ account for 45.9% of the total correction. Thus, temporal ambiguity predicts ownership-conflict risk, and MAT concentrates its correction where that ambiguity is greatest. Additional per-video robustness and ambiguity-transition analyses are provided in the appendix.

Table 1: Comparison on the DanceTrack test set. QP indicates whether object identities are maintained through propagated track queries. Params are reported in millions (M); GFLOPs and FPS are provided where available. FPS is measured on a single GPU with batch size one at the evaluation resolution; values for other methods are taken from their publications and may use different hardware. Gray rows denote recent alternative formulations or additional-data results shown for reference. ‡ indicates training with the DanceTrack validation set.

Method	QP	Params (M)↓	GFLOPs↓	FPS↑	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
FairMOT (Zhang et al., 2021)	✗	–	–	–	39.7	66.7	23.8	82.2	40.8
CenterTrack (Zhou et al., 2020)	✗	–	–	–	41.8	78.1	22.6	86.8	35.7
TraDeS (Wu et al., 2021)	✗	–	–	–	43.3	74.5	25.4	86.2	41.2
ByteTrack (Zhang et al., 2022)	✗	–	–	–	47.7	71.0	32.1	89.6	53.9
GTR (Zhou et al., 2022)	✗	–	–	–	48.0	72.5	31.9	84.7	50.3
QDTrack (Pang et al., 2021)	✗	–	–	–	54.2	80.1	36.8	87.7	50.4
OC-SORT (Cao et al., 2023)	✗	–	–	–	55.1	80.3	38.3	**92.0**	54.6
C-BIoU (Yang et al., 2023)	✗	–	–	–	60.6	81.3	45.4	91.6	61.6
TransTrack (Sun et al., 2020)	✗	–	–	–	45.5	75.9	27.5	88.4	45.2
MOTR (Zeng et al., 2022)	✓	40	173	19	54.2	73.5	40.2	79.7	51.5
MOTRv3-R50 (Yu et al., 2023)	✓	40	–	–	63.9	76.7	53.5	86.8	67.2
MeMOTR (Gao & Wang, 2023)	✓	50	–	12	68.5	80.5	58.4	89.9	71.2
CO-MOT (Yan et al., 2025)	✓	40	173	19	**69.4**	**82.1**	58.9	91.2	71.9
MATR-R50 (**ours**)	✓	42	175	18	**69.4**	81.5	**59.1**	91.0	**72.6**
MOTRv2 (Zhang et al., 2023)	✓	139	455	14	69.9	83.0	59.0	91.9	71.7
MOTRv3 (Yu et al., 2023)	✓	103	980	10	70.4	**83.8**	59.3	**92.9**	72.3
MATR (**ours**)	✓	43	179	17	**71.3**	82.6	**61.6**	91.9	**75.3**
LA-MOTR (Wang et al., 2025)	✗	50	–	12	71.1	84.4	59.0	93.5	71.8
FDTA (Shao et al., 2026)	✗	63	–	13	71.7	81.0	63.5	91.3	77.2
MOTIP (Gao et al., 2025)	✗	59	408	13	72.0	81.8	63.5	91.9	76.8
MATR‡ (**ours**)	✓	43	179	17	73.9	84.5	64.8	92.9	76.5

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate MATR on DanceTrack (Sun et al., 2022), SportsMOT (Cui et al., 2023), and BDD100K (Yu et al., 2020). DanceTrack emphasizes association under similar appearance and complex motion, while SportsMOT contains fast object and camera motion in densely populated sports scenes. BDD100K further evaluates generalization to large-scale multi-category tracking. For DanceTrack and SportsMOT, we report HOTA (Luiten et al., 2020) together with DetA and AssA, as well as MOTA and IDF1. For BDD100K, we follow prior work and report the multi-category TETA metrics (Li et al., 2022), including mTETA, mLocA, and mAssocA, together with mHOTA and mAssA.

Implementation. We build MATR on MOTR (Zeng et al., 2022). Unless otherwise specified, we use a ResNet-50 (He et al., 2016) backbone and additionally evaluate Swin-Tiny (Liu et al., 2021) on DanceTrack. Images are resized to a short side of 800 pixels with a maximum long side of 1536 pixels. We use 300 detection queries, a fixed sequence length of five frames, and a track-query dropping probability of 0.1; consecutive training frames are sampled with intervals from 1 to 10. Training uses AdamW with learning rates of 2×10^{-4} for ResNet-50 and 1×10^{-4} for Swin-Tiny. The auxiliary alignment weight in Eq. (6) is $\lambda_{\text{align}} = 5$.

Our strengthened MOTR baseline, denoted as Bl-IMP, additionally uses box-coordinate propagation, fixed-length training sequences, and random track-query dropping. These changes are kept separate from MAT in the ablations below so that the gain from the proposed module can be measured independently. We report Bl-IMP separately throughout because it accounts for a substantial part of the improvement over the original MOTR results (54.2 to 58.8 HOTA in Table 2a). Comparisons with other published trackers therefore reflect both the strengthened baseline and MAT, while the ablations isolate the contribution of MAT itself. For DanceTrack, we follow common practice (Zeng et al., 2022; Gao & Wang, 2023; Yan et al., 2025) and augment the training set with CrowdHuman (Shao et al., 2018), where pseudo trajectories are generated by spatial perturbation following CenterTrack (Zhou et al., 2020). The final DanceTrack models are trained for 20 epochs. SportsMOT uses only its native training set under the same general setup. For BDD100K, we use an 800×1333 maximum resolution, an eight-class prediction head, and train for 13 epochs on the original training set. All models are trained on eight NVIDIA A5000 GPUs.

4.2 ABLATION STUDIES

Evaluation protocol. All ablation models are trained for 5 epochs under the same optimization schedule. The core ablations in Table 2a follow the DanceTrack test-set setting in order to retain the directly reported MOTR baseline under the same protocol; MOTR does not provide the corresponding validation baseline used here. The later fine-grained design ablations in Table 3 are instead conducted on the DanceTrack validation set. We therefore use each table only for comparisons within the same split and do not compare their absolute values across tables.

Contribution beyond the strengthened baseline. Table 2a separates the effect of the strengthened baseline from the contribution of MAT. The baseline modifications improve MOTR from 54.2 to 58.8 HOTA. Adding MAT on top of this stronger baseline further raises HOTA by 4.8 points, from 58.8 to 63.6. The gain is concentrated in association: AssA increases from 45.2 to 52.7 and IDF1

Table 2: Ablation studies on MATR under different components and settings.

(a) Effect of MATR components. Bl-IMP denotes baseline improvement; MAT denotes inclusion of the MAT module.

Bl-IMP	MAT	HOTA	DetA	AssA	MOTA	IDF1
MOTR		54.2	73.5	40.2	79.7	51.5
√		58.8	75.9	45.2	84.9	59.2
√	√	**63.6**	**77.1**	**52.7**	**86.4**	**66.4**

(b) Effect of alignment design and MAT depth. "Dec" indicates the number of Decoder layers. "KF" refers to Kalman filter.

Dec	HOTA	DetA	AssA	MOTA	IDF1
KF	59.6	72.9	49.5	82.2	59.0
1	**63.6**	77.1	**52.7**	**86.4**	**66.4**
3	63.5	**77.9**	51.8	85.1	65.7

Table 3: Additional ablation studies on the DanceTrack validation set.

(a) Effect of auxiliary supervision.

Aux.	HOTA	DetA	AssA	MOTA	IDF1
None	56.8	68.9	47.2	73.7	59.0
GIoU	56.9	**70.1**	46.5	**76.5**	58.5
L1	**58.1**	69.4	**48.8**	75.3	**60.5**

(b) Effect of feature and position updates.

Update	HOTA	DetA	AssA	MOTA	IDF1
Feature	57.9	**71.0**	47.4	**78.0**	59.9
Position	53.7	65.4	44.2	70.6	56.5
Both	**58.1**	69.4	**48.8**	75.3	**60.5**

from 59.2 to 66.4, while DetA improves more modestly from 75.9 to 77.1. The benefit of MATR is therefore not explained by the baseline modifications alone.

Is simple motion extrapolation sufficient? We next compare MAT with a Kalman-filter (KF) alternative in Table 2b. In this baseline, a constant-velocity Kalman filter (Welch & Bishop, 1995) is fitted to the box trajectory of each track and its prediction replaces the propagated spatial reference at frame t, while the query feature is propagated unchanged; it therefore represents a history-only alternative that does not observe the current frame. Relative to the strengthened baseline, KF increases AssA from 45.2 to 49.5 but reduces DetA from 75.9 to 72.9 and MOTA from 84.9 to 82.2, indicating that a purely predicted reference, uncorrected by current-frame evidence, can place track queries away from the observed target. In contrast, a one-layer MAT reaches 63.6 HOTA, 52.7 AssA, and 77.1 DetA. Increasing MAT from one to three decoder layers yields nearly identical HOTA (63.6 vs. 63.5), while the single-layer design provides higher AssA, MOTA, and IDF1. We therefore use one MAT layer throughout the final models, which also keeps the additional computational cost small.

Alignment supervision. Table 3a evaluates the auxiliary supervision used by MAT. Removing direct alignment supervision reduces HOTA from 58.1 to 56.8 and AssA from 48.8 to 47.2. GIoU supervision improves DetA and MOTA but does not improve association to the same extent. The ℓ_1 objective gives the strongest HOTA, AssA, and IDF1 and is adopted in the final model.

Feature versus spatial alignment. Table 3b isolates the two components of the MAT update. Updating only the query feature reaches 57.9 HOTA, whereas updating only the spatial state reaches 53.7. Combining both achieves the highest HOTA (58.1), AssA (48.8), and IDF1 (60.5). The feature update dominates, but the best association requires aligning both.

4.3 Benchmark Comparison

DanceTrack and efficiency. Table 1 compares MATR with representative tracking methods on the DanceTrack test set. With ResNet-50, MATR-R50 reaches 69.4 HOTA, matching CO-MOT while improving AssA from 58.9 to 59.1 and IDF1 from 71.9 to 72.6. Compared with MeMOTR, MATR-R50 improves HOTA by 0.9 and IDF1 by 1.4.

The full MATR model reaches 71.3 HOTA, 61.6 AssA, and 75.3 IDF1. Among query-propagation trackers of comparable size, this is the strongest HOTA and association performance, at 43M parameters and 179 GFLOPs. Recent trackers that modify the tracking formulation itself, such as FDTA (Shao et al., 2026), report higher absolute HOTA using larger models or different association mechanisms; our comparison isolates the effect of track-state pre-alignment under a fixed query-propagation formulation. The overhead relative to MOTR is small (42M vs. 40M parameters, 175 vs. 173 GFLOPs, 18 vs. 19 FPS for MATR-R50), whereas MOTRv2 and MOTRv3 use 139M/455 and 103M/980 GFLOPs, respectively.

We also include LA-MOTR (Wang et al., 2025), MOTIP (Gao et al., 2025), and FDTA (Shao et al., 2026) in gray as recent reference points. These methods alter the tracking formulation itself (explicit association, ID prediction, or discriminative association embeddings) and are shown for context rather than as controlled query-propagation baselines. The gray MATR‡ row additionally trains on the DanceTrack validation set.

Generalization to SportsMOT. On SportsMOT, we train only on the native training set and compare primarily with methods under the same no-extra-data setting. MATR obtains 72.7 HOTA (Table 4a), improving over MeMOTR by 2.7 HOTA, 2.9 AssA, and 3.2 IDF1; the gain is again concentrated in association under fast motion and frequent interactions.

Table 4: Comparison on SportsMOT and BDD100K.

(a) Results on the SportsMOT test set.

Methods	HOTA	DetA	AssA	MOTA	IDF1
w/o extra data					
FairMOT	49.3	70.2	34.7	86.4	53.5
QDTrack	60.4	77.5	47.2	90.1	62.3
ByteTrack	62.1	76.5	50.5	93.4	69.1
OC-SORT	68.1	84.8	54.8	93.4	68.0
MeMOTR	70.0	83.1	59.1	91.5	71.4
MATR (**ours**)	**72.7**	**85.3**	**62.0**	**95.1**	**74.6**

(b) Results on the BDD100K validation set.

Methods	mTETA	mHOTA	mLocA	mAssocA	mAssA
QDTrack	47.8	–	45.9	48.5	–
DeepSORT	48.0	–	46.4	46.7	–
MOTR	50.7	37.0	35.8	51.0	47.3
TETer	50.8	–	**47.2**	52.9	–
CO-MOT	52.8	–	38.7	56.2	–
MeMOTR	53.6	40.4	38.1	56.7	52.6
MATR (**ours**)	**54.7**	**41.6**	41.8	**59.0**	**53.0**

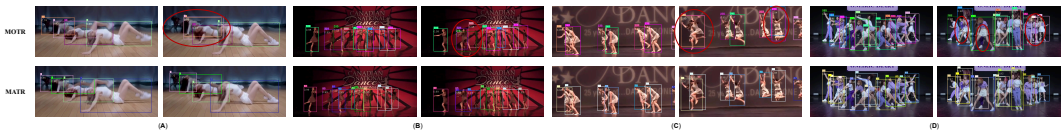


Figure 4: Qualitative comparison between MOTR and MATR under occlusion, crossing trajectories, and complex motion. MATR maintains object identities more consistently, whereas MOTR exhibits identity switches or lost tracks.

Generalization to multi-category tracking. BDD100K tests whether the same design transfers beyond single-class human tracking. Using only the BDD100K tracking training set, without additional data or external detector proposals, MATR achieves 54.7 mTETA and 41.6 mHOTA on the validation set (Table 4b), improving over MeMOTR by 1.1 mTETA, 1.2 mHOTA, and 2.3 mAssocA. TETer retains higher localization accuracy (47.2 vs. 41.8 mLocA), so MATR's advantage again lies in association rather than localization.

4.4 Qualitative Analysis

Figure 4 compares MOTR and MATR in representative challenging sequences. In case A, MATR preserves the identity of an occluded background target despite interference from a foreground object. Cases B and C illustrate crossing trajectories, where MATR maintains more consistent identities while MOTR exhibits identity switches. Case D shows a longer and more crowded sequence, in which MATR retains multiple tracks that are lost or switched by MOTR. These examples are consistent with our quantitative analysis: pre-aligning the propagated track state reduces the burden on the shared decoder when the inherited state is temporally misaligned.

5 Conclusion

We presented MATR, an end-to-end multi-object tracking framework that addresses temporal state misalignment in query-propagation trackers. Rather than changing target ownership or introducing a separate association stage, MATR uses current-frame evidence to pre-align the inherited feature and spatial states before joint decoding. We formalized the resulting query ownership conflict through the ownership margin and a geometric ambiguity ratio, and showed that conflict risk grows with ambiguity while MAT concentrates its correction on these states. Experiments on DanceTrack, SportsMOT, and BDD100K show consistent association gains at modest cost. Several limitations remain. MATR mitigates rather than eliminates ownership conflicts: pre-alignment reduces temporal misalignment but cannot recover a track whose target is fully occluded for an extended period, and it introduces no re-identification mechanism. MAT acts only on inherited track queries, so newborn objects and abrupt camera motion, where no reliable prior state exists, are unaffected. Our mechanism analysis is conducted on DanceTrack, and although the benchmark results transfer to SportsMOT and BDD100K, the ambiguity statistics may differ in scenes with different object density. Finally, part of the improvement over the original MOTR results comes from our strengthened baseline, which we therefore report separately. Applying MAT on top of other query-propagation trackers is a natural next step that we leave to future work.

AI USE STATEMENT

Generative AI tools were used to assist with improving the clarity and readability of the manuscript and, where applicable, to provide limited feedback during the research and writing process. All research ideas, methodological design, experiments, analyses, results, and scientific conclusions were developed, conducted, verified, and finalized by the authors.

REPRODUCIBILITY STATEMENT

We provide the implementation details, training settings, and evaluation protocols necessary to reproduce our reported results in the paper and appendix. Upon acceptance, we will publicly release the source code, including the complete training and evaluation scripts, configuration files, pre-trained model weights and checkpoints, and the scripts required to reproduce the main experiments and analyses reported in this paper.

REFERENCES

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468. IEEE, 2016.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9686–9696, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

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.

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*, pp. 9901–9910, 2023.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as id prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27883–27893. IEEE, 2025.

Fabian Gülhan, Emil Mededovic, Yuli Wu, and Johannes Stegmaier. Selfmotr: Revisiting motr with self-generating detection priors. In *European Conference on Computer Vision*, pp. 213–231. Springer, 2026.

Xu Guo, Yujin Zheng, and Dingwen Wang. Pmtrack: Multi-object tracking with motion-aware. In *Asian Conference on Computer Vision*, pp. 429–444. Springer, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2020.

Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19321–19330. IEEE, 2024.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8844–8854, 2022.

Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 164–173, 2021.

Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Luc Van Gool, and Bernt Schiele. Samba: Synchronized set-of-sequences modeling for multiple object tracking. In *International Conference on Learning Representations*, volume 2025, pp. 30057–30070, 2025.

Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

Yuqing Shao, Yuchen Yang, Rui Yu, Weilong Li, Xu Guo, Huaicheng Yan, Wei Wang, and Xiao Sun. From detection to association: Learning discriminative object embeddings for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6878–6888, 2026.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20993–21002, 2022.

Peng Wang, Yongcai Wang, Hualong Cao, Wang Chen, and Deying Li. La-motr: End-to-end multi-object tracking by learnable association. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12438–12448. IEEE, 2025.

Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, University of North Carolina at Chapel Hill, 1995.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649. IEEE, 2017. doi: 10.1109/ICIP.2017.8296962.

Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12352–12361, 2021.

feng Yan, Weixin Luo, Yujie Zhong, Yiyang Gan, and Lin Ma. Co-mot: Boosting end-to-end transformer-based multi-object tracking via coopetition label assignment and shadow sets. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53359–53371, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8428da31da191712130ce8cce265691a-Paper-Conference.pdf.

Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 4799–4808, 2023.

En Yu, Tiancai Wang, Zhuoling Li, Yuang Zhang, Xiangyu Zhang, and Wenbing Tao. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv preprint arXiv:2305.14298*, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

Yifu Zhang, Changbao Wang, Xia Wang, Wei Zeng, and Weiguo Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.*, 129(11):3069–3087, 2021. doi: 10.1007/s11263-021-01513-4.

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.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22056–22065, 2023.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pp. 474–490. Springer, 2020.

Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8771–8780, 2022.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations, ICLR 2021*, 2021.

A Appendix

A.1 Additional Mechanism Analysis

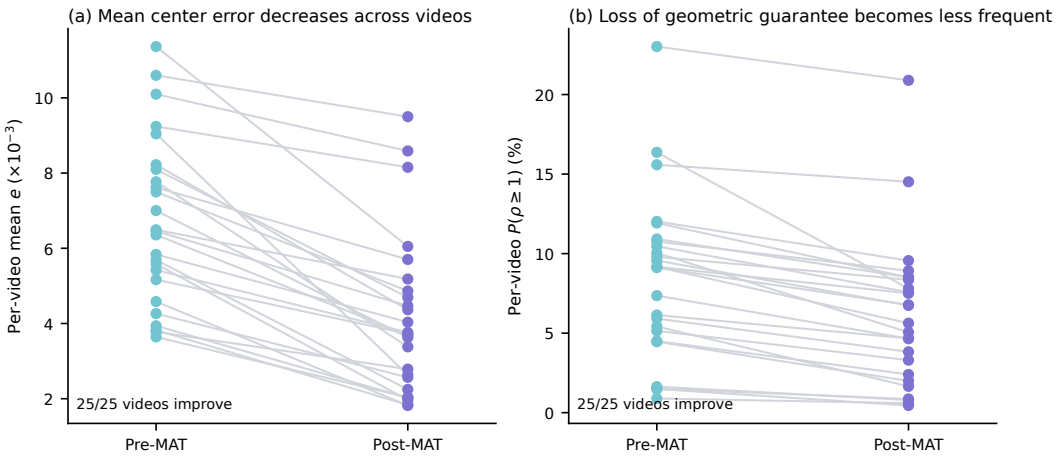


Figure 5: Per-video robustness of MAT on the DanceTrack validation set. (**a**) Mean normalized center error before and after MAT. (**b**) Fraction of states beyond the $\rho = 1$ sufficient-guarantee boundary. All 25 videos improve under both measures.

The mechanism observed in the main paper is consistent across individual DanceTrack validation sequences. As shown in Fig. 5, all 25 videos exhibit lower mean normalized center error after MAT. The fraction of states beyond the $\rho = 1$ sufficient-guarantee boundary also decreases in every sequence, indicating that the reduction in geometric ambiguity is not driven by a small subset of videos.

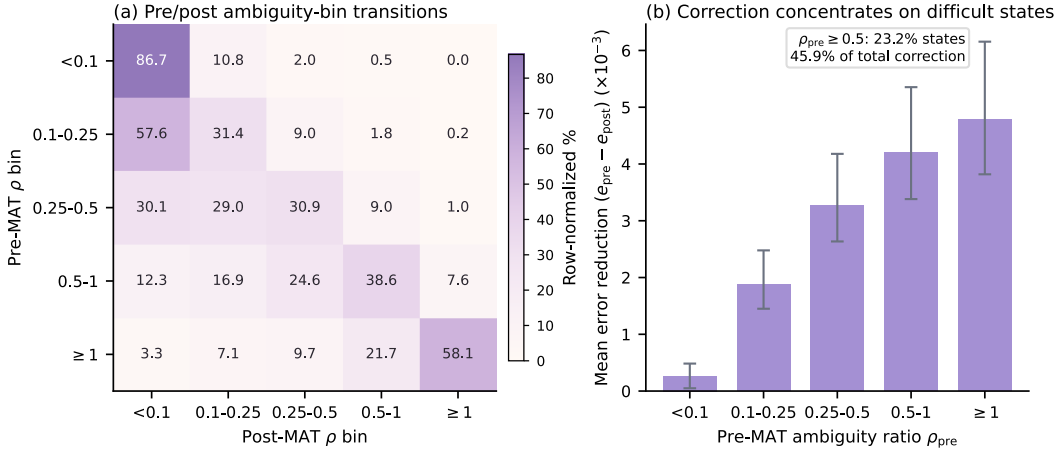


Figure 6: Detailed behavior of MAT across ambiguity levels. (**a**) Row-normalized transitions from pre-MAT to post-MAT ambiguity bins. (**b**) Mean center-error reduction as a function of ρ_{pre}. States with $\rho_{\text{pre}} \geq 0.5$ constitute 23.2% of samples but account for 45.9% of the total correction.

Figure 6 provides a finer view of how MAT changes the ambiguity distribution. Most transitions move difficult states toward lower-ambiguity bins, while the magnitude of the center-error correction increases with the initial ρ_{pre}. This complements the aggregate analysis in the main paper and shows that MAT acts most strongly on states with larger temporal misalignment.

A.2 Additional Implementation Details

Unless otherwise specified, training uses a batch size of one video clip per GPU. During training, inherited tracked targets with IoU below $\tau_{\text{IoU}} = 0.5$ are filtered. For DanceTrack and SportsMOT, the learning rate is decayed by a factor of 0.1 every 8 epochs; for BDD100K, it is decayed at epoch 11.

For inference, we follow MOTR (Zeng et al., 2022) and MOTRv2 (Zhang et al., 2023) without modifying their inference hyperparameters.