Mode-Conditioned Residual Calibration for Multi-Object Tracking

Muyu Li[1,2], Henan Hu[3⋆], Deepak Kumar Jain[1,2], and Xudong Zhao[1,2]

[1] Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, 116024, Dalian, Liaoning, China

[2] School of Control Science and Engineering, Dalian University of Technology, Dalian, 116024, Liaoning, China

{muyuli@dlut.edu.cn, dkj@dlut.edu.cn, xudongzhao@dlut.edu.cn}

[3] School of Mechanical Engineering, Dalian Jiaotong University, Dalian, 116028, Liaoning, China

{huhenan@djtu.edu.cn}

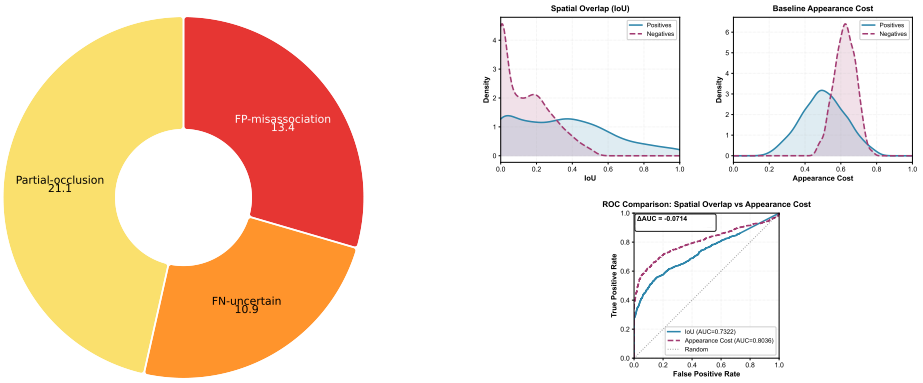
Abstract. Multi-object tracking degrades in crowded scenes. The reliability of geometry and appearance shifts under occlusion, uniform appearance, and fast motion. Consequently, association scores suffer from mode-dependent miscalibration. We observe that these failures concentrate into a small set of recurring error modes. This pattern renders single global thresholds ineffective. To address this issue, we propose a diagnose-then-correct strategy via Mode-Conditioned Residual Calibration (MCRC). It is a lightweight auxiliary module that identifies the dominant failure mode from pairwise and set-level geometric cues. The module then applies a mode-conditioned residual correction to the association cost matrix. For highly ambiguous tracking pairs, MCRC selectively activates an uncertainty-gated semantic refiner. This refiner incorporates simple acceptance checks to ensure representation stability, thereby avoiding global computational overhead. Extensive evaluations on DanceTrack, SportsMOT, MOT17, and MOT20 confirm its effectiveness. MCRC consistently improves identity preservation when inserted into diverse trackers, maintaining minimal added overhead.

Keywords: Multi-Object Tracking · Data Association · Error Modeling · Representation Refinement · Vector Quantization

1 Introduction

Multi-object tracking links detection hypotheses across consecutive frames to reconstruct continuous target trajectories. Most online trackers formulate data association by computing a pairwise cost matrix between existing tracks and incoming detections. They compute association costs from geometric, motion, confidence, or appearance cues, depending on the tracker design [1, 2, 30]. This

⋆ Corresponding author: huhenan@djtu.edu.cn



(a) Hard sample distribution **(b)** Spatial overlap comparison with appearance

Fig. 1: Audit of tracking failures on the DanceTrack dataset [25]. The left panel illustrates how hard cases responsible for tracker failures cluster into distinct error modes. These include partial occlusion and misassociation. The right panel compares spatial overlap and appearance cost, showing their distinct yet complementary discriminative behaviors. This motivates mode-conditioned correction instead of relying on a single global association threshold.

fixed scoring rule is generally effective in unconstrained scenes. However, it often degrades in crowded environments. Occlusions corrupt embeddings. Uniform appearance induces visual confusion. Fast motion weakens geometric continuity. Consequently, association scores become miscalibrated. High-confidence matches can be incorrect, and true matches may be erroneously penalized. Therefore, relying on a single global threshold is insufficient.

To understand this performance drop, we audit association failures across multiple baselines. We observe three consistent patterns. First, errors are not random. Hard cases cluster into a small set of recurring temporal modes like partial occlusion and crowd confusion, showcased in Fig. 1a. Second, cue-dependent miscalibration emerges in ambiguous cases. Distractors may receive overly favorable costs, while visually obstructed true matches are penalized, as suggested by Fig. 1b. Third, ambiguity is inherently a set-level issue. The correctness of a pair depends entirely on competing candidates, yet most scoring is computed in isolation. This insight dictates our diagnose-then-correct strategy. We must evaluate the set-level competitive context before altering association scores. Rather than globally tuning thresholds, we infer the dominant failure regime and apply targeted, mode-specific corrections.

Based on this perspective, we introduce Mode-Conditioned Residual Calibration (MCRC). It is an auxiliary module that refines the association cost matrix of a tracker without altering its detector or motion model. MCRC extracts structured cues for each candidate pair. These include set-level rank statistics that capture local competition. The module then performs prototype-based mode diagnosis in a normalized latent space. Conditioned on the inferred mode, a

lightweight regressor predicts an additive residual to calibrate the baseline cost. When the geometric correction remains uncertain, MCRC selectively activates a semantic refiner. This refiner serves to repair appearance embeddings. It applies the update only if simple acceptance tests indicate an actual similarity improvement. This selective and residual design preserves baseline behavior on easy pairs. Meanwhile, it focuses representational capacity on recurrent hard modes.

The primary contributions of this work are as follows.

- We revisit MOT association failures as mode-dependent miscalibration, and provide an empirical audit showing hard errors concentrate into recurring regimes.
- We propose MCRC, which couples prototype-based mode diagnosis with residual cost calibration using pairwise and set-level geometric cues, enabling targeted corrections while preserving baseline scoring on easy cases.
- We introduce an uncertainty-gated semantic repair pathway with simple acceptance checks, and demonstrate consistent identity-preservation gains when plugging MCRC into diverse trackers on challenging benchmarks.

2 Related Work

2.1 Multiple Object Tracking

Most online multi-object tracking systems resolve data association via bipartite matching with costs derived from geometry, motion, and appearance features. Heuristic trackers such as ByteTrack [30] and OC-SORT [2] improve robustness through confidence handling and motion smoothing. Recent trackers further improve association through track-focused matching, synchronized query/sequence modeling, discriminative embedding refinement, or multi-cue designs [18, 21, 22, 24, 28]. Despite these advances, association scores are typically produced by a single scoring rule with global thresholds. This implicitly assumes that cue reliability is stationary. In dense crowds, cue reliability shifts and hard errors recur in specific regimes. Pairwise scoring also ignores competition among candidates in these regimes. Our work addresses this gap. We introduce a mode-aware calibration layer to diagnose failure regimes from structured cues. This layer applies residual corrections while remaining adaptable to diverse baselines. Transformer-based trackers [29, 32] integrate detection and association through propagated object queries. However, they still rely on fixed matching objectives and do not explicitly model mode-specific miscalibration.

2.2 Association Calibration and Discretization

Calibration methods align predicted confidence with empirical accuracy. This is often achieved via global post-hoc scaling [9] or uncertainty modeling [20, 23]. Such techniques are broadly sample-agnostic and apply identical formulations

universally. Consequently, global refiners ignore the structured set-level ambiguity intrinsic to data association. We instead position data association as a mode-conditioned calibration problem. We bridge discrete representation learning with multi-object tracking. Drawing on vector quantization principles [15, 26, 33], our method learns a discrete vocabulary of recurring tracking error configurations. This enables us to calibrate association costs conditionally based on the inferred geometric neighborhood. It encourages consistent mode embeddings and reusable correction patterns across recurring association configurations. Permutation-invariant set encoders [11] further support our use of set-level statistics to capture candidate competition safely, deploying residual updates rather than overwriting the baseline completely.

2.3 Tracking Refiners and Feature Refinement

Feature refiners aim to repair corrupted representations under ambiguity. Diffusion-based formulations have been explored for object detection and MOT association [3, 12], while denoising representation models support robust feature learning [27]. Nevertheless, they typically apply uniform refinement globally. This pattern escalates latency and risks unnecessary changes on easy cases. In contrast, our semantic refiner operates selectively. It is explicitly gated by a mode-derived uncertainty score and constrained by deterministic acceptance checks. This mechanism ensures targeted repair only when the baseline representation is likely unreliable.

3 Method

3.1 Architectural Overview

As depicted in Fig. 2, our framework refines the baseline association cost matrix $\mathbf{S}$ produced by a tracker. Lower values in this matrix indicate higher match likelihood. Rather than replacing the tracker's scoring rule, MCRC predicts a residual correction. This preserves baseline behavior on easy pairs and corrects miscalibrated hard cases. For each track-detection pair i and j with baseline cost s_{ij}, the model encodes structured discrepancies into a tensor c_{ij} and projects it to a latent representation z_{ij}. This latent space drives two pathways. First, prototype-based mode diagnosis is followed by an additive geometric correction δ_{ij}. Second, a conditional semantic refinement repairs unstable appearance embeddings.

Operationally, MCRC follows a score-level calibration chain. For each native association pair, we construct the error cue c_{ij}, encode it into z_{ij}, estimate the spatial and motion mode probabilities $(\pi_{ij}^{C}, \pi_{ij}^{M})$, and predict an additive residual δ_{ij} to obtain $\hat{s}_{ij} = s_{ij} + \delta_{ij}$. For uncertain pairs, residual error coding further provides the VQ anchor e_q, and the semantic branch predicts $\tilde{d}_{ij}$. The final calibrated cost $\bar{s}_{ij}$ is integrated from $\hat{s}_{ij}$ and $\tilde{d}_{ij}$ and returned to the original association procedure.

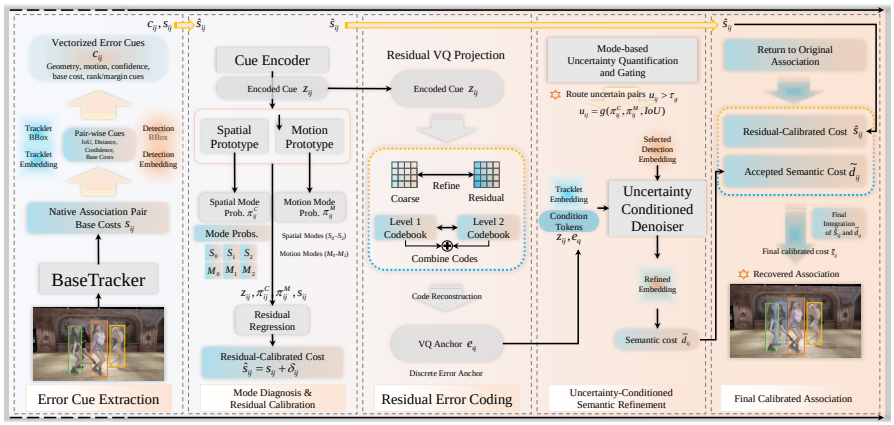


Fig. 2: Data flow of MCRC. The five columns correspond to error-cue extraction, mode diagnosis with residual calibration, residual error coding, uncertainty-conditioned semantic refinement, and final calibrated association. Given a native association pair and base cost s_{ij}, MCRC constructs vectorized error cues c_{ij}, encodes them into z_{ij}, predicts mode probabilities and an additive residual δ_{ij}, and obtains the residual-calibrated cost $\hat{s}_{ij} = s_{ij} + \delta_{ij}$. Uncertain pairs are further routed to the semantic branch, where residual error coding provides the VQ anchor e_q and the denoiser predicts $\tilde{d}_{ij}$. The final calibrated cost $\bar{s}_{ij}$ is integrated from $\hat{s}_{ij}$ and $\tilde{d}_{ij}$ and returned to the original association step.

3.2 Set-Level Error Cues Extraction

Robust mode classification requires evaluating competitive context beyond isolated pairs. Our method constructs a predictive error cue vector c_{ij} to capture the physical dynamics of consecutive bounding boxes. It concatenates discrete finite differences of position, velocity, scale, and aspect ratio. We then expand this pairwise geometry by integrating set-level affinity statistics. Specifically, we compute local tracking rank predictions. This is achieved by sorting the baseline spatial distances of all competing detections for a given target. We incorporate these discrete ranks alongside bounding box neighborhood margins. Subsequently, we concatenate them with baseline contextual confidence metrics to form the integrated tensor c_{ij}.

3.3 Mode Diagnosis and Residual Calibration

The mapping module translates the encoded geometric cues into a normalized continuous feature space via a dedicated mapping network f_{mlp}. We enforce a strict unit-norm constraint on this projection layer to emphasize directional orientation over raw numerical magnitudes. This is formulated as $z_{ij} = f_{\text{mlp}}(c_{ij}) / \|f_{\text{mlp}}(c_{ij})\|_2$. This normalization step stabilizes the subsequent dot-product similarity with cluster centers by reducing scaling variance. Consequently, the model can focus on structural error patterns. The framework maintains separate unit-norm geometric prototype matrices P_C for spatial clusters

and P_M for localized motion clusters. Initialized via spherical k-means on a pre-computed feature bank, these centers are updated during training using an Exponential Moving Average (EMA). This ensures stability against high-frequency gradient noise. Utilizing a scaled temperature parameter τ, the model computes softmax activation probabilities $\pi_{ij}^{C} = \text{softmax}(z_{ij}^{\top} P_C/\tau)$ and $\pi_{ij}^{M} = \text{softmax}(z_{ij}^{\top} P_M/\tau)$. This establishes a reliable mixture distribution over discrete association regimes. To formalize miscalibration modes, we categorize the physical meaning of these prototypes based on dataset auditing. We apply a heuristic pseudo-labeling constraint during early training to explicitly align these unsupervised clusters. High spatial overlap and low appearance similarity pairs supervise a designated crowd centroid in P_C. This auxiliary loss anchors specific softmax indices to physical regimes. It yields the crowd confusion score π_{ij}^{crowd} and motion displacement π_{ij}^{motion} from P_M.

Following mode identification, a regression multi-layer perceptron φ processes the latent feature, prototype probabilities, and baseline cost to predict an additive residual. This prediction is computed by $\delta_{ij} = \varphi(z_{ij}, \pi_{ij}^{C}, \pi_{ij}^{M}, s_{ij})$. The network subsequently updates the calibrated cost as $\hat{s}_{ij} = s_{ij} + \delta_{ij}$. We train all components jointly using the objectives detailed in Sec. 3.6.

3.4 Residual Error Coding

Complicated tracking errors frequently exhibit recurring geometrical patterns. Our architecture captures these configurations to prevent the classification model from overfitting to localized training noise. We discretize the continuous feature space to effectively learn a reusable vocabulary of tracking failure modes. We achieve this by imposing a multi-level Residual Vector Quantization structure over the normalized features. The primary codebook establishes general patterns. These act as stable contextual anchors to conditionally guide downstream semantic repair. Conversely, deeper hierarchical levels act as an auxiliary regularization bottleneck. By forcing the continuous space to align with these fine-grained anchors, the network learns robust representations. It resists feature deviation without overloading the denoiser with excessive input tokens. Instead of a single codebook, we deploy a hierarchical arrangement of L codebooks.

Continuous geometric features are directly projected onto dictionary centroids to compute quantization loss. The forward pass routes unquantized features directly to the residual estimator. This preserves geometric gradients across all hierarchical levels. Concurrently, the primary quantized code acts as the sole stable anchor for subsequent semantic refinement. We formulate the multi-level quantization loss as follows.

$$\mathcal{L}_{\text{vq}} = \sum_{l=1}^{L} \left(\beta_{\text{vq}} \left\| e_q^{(l)} - \text{sg}(r^{(l-1)}) \right\|_2^2 + \beta_{\text{commit}} \left\| \text{sg}(e_q^{(l)}) - r^{(l-1)} \right\|_2^2 \right), \quad (1)$$

where $r^{(0)} = z_{ij}$ and $e_q^{(l)}$ is the nearest centroid at level l. The first term updates the dictionary codebook, while the second commitment term forces the

continuous encoder to align with geometric anchors. The next level residual is $r^{(l)} = r^{(l-1)} - e_q^{(l)}$. Centroids update dynamically alongside prototypes.

3.5 Uncertainty Conditioned Semantic Refinement

While residual coding addresses geometric failures, physical occlusion corrupts detection representation. Therefore, the architecture selectively deploys an uncertainty-driven appearance refinement module for gated candidates. The gating logic filters false positives originating from geometric occlusion and abrupt motion divergence. These are identified via the crowd and motion displacement clusters respectively. While crowd confusion π_{ij}^{crowd} is intrinsically reliable due to overlapping distractors, temporal motion uncertainty π_{ij}^{motion} only signals failure when spatial continuity is compromised. We model this geometric breakdown as a spatial prior using an indicator function. The summation over these learned probability modes serves as an uncertainty score for geometric failure. Thus, we compute the uncertainty score according to the following relationship.

$$g_{ij} = \pi_{ij}^{\text{crowd}} + \pi_{ij}^{\text{motion}} \mathbf{1}\{\text{IoU}_{ij} < \tau_{\text{iou}}\}. \tag{2}$$

The framework dynamically routes candidate pairs demanding repair when their associated score g_{ij} exceeds the decision threshold τ_g. This processing step generates a binary routing gate $\gamma_{ij} = \mathbf{1}\{g_{ij} > \tau_g\}$. We utilize this routing mask exclusively to reduce the computational overhead of the transformer encoder during inference, ensuring operational efficiency.

For affirmatively routed candidates, an attention-guided transformer encoder $\mathcal{D}_\theta$ contextualizes the corrupted detection embedding z_j^d. We linearly project the target track embedding z_i^t, the geometric cue z_{ij}, the cluster centroid $e_q^{(1)}$, and z_j^d directly into a shared dimensional space. The network concatenates these four spatial vectors to construct a sequence of tokens. Unlike a Multi-Layer Perceptron which imposes static weighting, the self-attention layers within $\mathcal{D}_\theta$ dynamically weight geometric cues versus structural priors conditional on the specific corruption. This produces a structurally refined detection representation $\hat{z}_j^d = \mathcal{D}_\theta(z_j^d, z_i^t, z_{ij}, e_q^{(1)})$. Despite the short sequence length of four tokens, self-attention effectively models complex conditional dependencies among these disparate cues. The network synthesizes the final association cost by blending the prototype-corrected geometric cost $\hat{s}_{ij}$ with the refined semantic cost $\tilde{d}_{ij}$. We define $\tilde{d}_{ij} = d(z_i^t, \hat{z}_j^d)$, where $d(\cdot, \cdot)$ is cosine distance scaled to the unit interval. To enforce compatible numerical domains before integration, we apply a local min-max standardization over the active pairwise matrix for $\hat{s}_{ij}$. During training, we maintain a differentiable cost landscape to prevent discontinuous assignment jitter. This is achieved by substituting the hard routing gate with a continuous shifted sigmoid activation $\sigma(g_{ij} - \tau_g)$ during interpolation. We apply a fixed scalar $\beta \in (0, 1)$ according to the following blend operation.

$$\bar{s}_{ij} = (1 - \sigma(g_{ij} - \tau_g)\beta)\, \hat{s}_{ij} + \sigma(g_{ij} - \tau_g)\beta\, \tilde{d}_{ij}. \tag{3}$$

Unconstrained cost adjustments frequently suffer from out-of-distribution feature collapse. Therefore, we mandate deterministic safety filters to oversee the semantic representation. These post-processing checks secure both repair validity and empirical stability during inference. The latent block refines features without decoding coordinate geometries. Consequently, the repaired representation propagates only if it safely passes two pure-feature sequential criteria. First, an embedding margin requirement $d(z_i^t, \hat{z}_j^d) \leq d(z_i^t, z_j^d) - \tau_s$ verifies this semantic cost decreases by at least threshold τ_s. Second, an amplitude limit $\|\hat{z}_j^d - z_j^d\| \leq \tau_{\text{clip}}$ bounds the absolute magnitude of the vector update. This prevents excessive representation deviations.

3.6 Integrated Model Training

The training architecture concurrently optimizes the primary classification engine, residual codebook projection, and appearance refinement module into a single scalar loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rerank}} + \mathcal{L}_{\text{denoise}} + \mathcal{L}_{\text{vq}}$.

The core classification objective $\mathcal{L}_{\text{rerank}}$ combines a standard focal loss on match probability with structural prototype regularization, defined as $\mathcal{L}_{\text{rerank}} = \mathcal{L}_{\text{focal}}(\sigma(-\bar{s}_{ij}), y_{ij}) + \lambda_{\text{ortho}}\|P_C^\top P_C - I\|_F^2 + \lambda_{\text{compact}} \sum \min_k \|z_{ij} - p_k\|_2^2$, where $y_{ij} = 1$ if detection j matches track i under ground-truth identity and 0 otherwise, and σ is the sigmoid. The orthogonality penalty enforces semantic separation between distinct error clusters, while the compactness constraint pulls latent embeddings toward these defined discrete anchors. To supervise semantic refinement, we minimize $\mathcal{L}_{\text{denoise}} = d(z_i^t, \hat{z}_j^d) + \max(0, d(z_i^t, \hat{z}_j^d) - d(z_i^t, z_j^d) + m)$, enforcing a margin $m = 0.1$ over the pre-refinement embedding.

4 Experiments

4.1 Implementation Details

Our implementation builds on CAMELTrack [24] for backbone, detector, keypoints and re-identification embeddings. We insert MCRC only at the association stage. We train for 20 epochs using a OneCycleLR scheduler and batch size 8 on a single RTX 4090. To ensure consistency, we apply CAMELTrack [24] augmentations. For the refinement module, we systematically initialize the hyperparameter configuration across all tracking benchmarks. Specifically, we set the prototype temperature scaling $\tau = 0.07$ to sharpen distribution boundaries. For the appearance denoiser processing gates, we empirically define the intersection threshold $\tau_{\text{iou}} = 0.3$ and the uncertainty activation scalar $\tau_g = 0.65$. The deterministic semantic post-processing parameters utilize an embedding similarity margin $\tau_s = 0.05$ and an absolute magnitude clip $\tau_{\text{clip}} = 0.2$. The interpolating scalar β governing cost blending is uniformly fixed at 0.4.

Datasets and Metrics. We report results on DanceTrack [25] for uniform appearance, MOT17 [16] for street scenes, SportsMOT [4] for fast motion, and MOT20 [5] for extreme crowded environments. Metrics include HOTA, AssA, IDF1 and ID switches.

4.2 State-of-the-art Comparison

We compare MCRC against representative recent trackers across four demanding public datasets. We distinguish official standalone tracker results from reproduced plug-in baselines. Official results are reported for contextual comparison with recent trackers, whereas reproduced baselines are used only to measure the controlled before–after effect of inserting MCRC under the same tracker interface and inputs. Unless otherwise specified, the detector, motion model, appearance features, and base inference path remain unchanged.

Method	HOTA	AssA	DetA	IDF1	MOTA
ByteTrack [30]	47.7	32.1	71.0	53.9	91.3
MOTR [29]	54.2	40.2	73.5	51.5	79.7
OC-SORT [2]	55.1	38.0	80.3	54.2	89.4
GHOST [19]	56.7	39.8	81.1	57.7	89.6
Deep OC-SORT [14]	61.3	45.2	<u>82.2</u>	61.2	92.0
DiffMOT [13]	62.3	48.8	**82.5**	66.4	<u>92.7</u>
Hybrid-SORT [28]	65.7	–	–	67.4	91.8
TrackTrack [22]	66.5	52.9	–	67.8	**93.6**
SambaMOTR [18]	67.2	57.5	78.8	70.5	88.1
MeMOTR [8]	68.5	58.4	80.5	71.2	89.9
CAMELTrack [24]	69.3	58.9	81.8	74.9	91.4
MOTIP [7]	69.6	60.4	80.4	74.7	90.6
FDTA [21]	<u>71.7</u>	**63.5**	81.0	**77.2**	91.3
CAMEL+MCRC	**72.0**	<u>63.1</u>	82.1	<u>76.2</u>	91.6

Table 1: DanceTrack [25] test set. **Bold** and <u>underline</u> denote the best and second-best results, respectively.

Method	HOTA	AssA	DetA	IDF1	MOTA
MOTR [29]	57.8	55.7	60.3	68.6	73.4
MOTIP [7]	59.2	56.9	62.0	71.2	75.5
FairMOT [31]	59.3	–	–	72.3	73.7
OC-SORT [2]	61.7	–	–	76.2	76.0
MOTRv2 [32]	62.0	60.6	**63.8**	75.0	78.6
CAMELTrack [24]	62.4	<u>61.4</u>	<u>63.6</u>	76.5	78.5
ByteTrack [30]	<u>63.1</u>	–	–	<u>77.3</u>	<u>80.3</u>
GHOST [19]	62.8	–	–	77.1	78.7
CAMEL+MCRC	**64.5**	**64.8**	**63.8**	**79.4**	**80.5**

Table 2: MOT17 [16] test set. **Bold** and <u>underline</u> denote the best and second-best results, respectively.

DanceTrack Analysis. As shown in Table 1, on DanceTrack [25] test, MCRC reaches 72.0 HOTA, 63.1 AssA and 76.2 IDF1, improving over CAMELTrack [24] by 2.7 HOTA and 4.2 AssA with stable DetA. This gain arises from DanceTrack's uniform appearance and crowded scenarios, which generate high-cost false matches. To establish this motivation, an offline heuristic audit of baseline failures on the training set reveals that hard samples concentrate in specific regimes including 21% partial occlusion, 13.4% misassociation, and 10.9% false negatives. Consequently, these structural modes directly benefit from prototype-based correction and VQ pattern reuse across similar dance sequences, while the dynamic denoiser subsequently activates on approximately 12% of gated pairs to recover structural errors.

MOT17 Analysis. From Table 2, on MOT17 [16] test, MCRC obtains 64.5 HOTA. This improvement of 2.1 HOTA over the CAMELTrack baseline confirms MCRC's consistent impact, even though MOT17 is less crowded than DanceTrack and features fewer ambiguous occlusion events. Consistent with our primary objectives, MCRC yields reliable improvements in identity consistency, reaching 64.8 AssA and 79.4 IDF1 rather than demonstrating arbitrary detection

dominance. MCRC resolves Identity Switches, confirming it functions as an effective integration module that manages sporadic error modes without degrading isolated targets.

Method	HOTA	AssA	DetA	IDF1	MOTA
ByteTrack [30]	62.1	50.5	76.5	69.1	93.4
OC-SORT [2]	68.1	54.8	84.8	68.0	93.4
MeMOTR [8]	70.0	59.1	83.1	71.4	91.5
MOTIP [7]	71.9	62.0	83.4	75.0	92.9
MotionTrack [17]	74.0	61.7	88.8	74.0	<u>96.6</u>
MixSORT [4]	74.1	62.0	85.8	74.4	96.5
DiffMOT [13]	76.2	65.1	**89.3**	76.1	**97.1**
Deep-EIoU [10]	77.2	67.7	88.2	79.8	96.3
CAMELTrack [24]	<u>80.4</u>	<u>72.8</u>	88.8	<u>84.8</u>	96.3
CAMEL+MCRC	**81.6**	**74.0**	<u>89.2</u>	**85.6**	96.5

Table 3: SportsMOT [4] test set. **Bold** and <u>underline</u> denote the best and second-best results, respectively.

Method	HOTA	MOTA	AssA	DetA
FairMOT [31]	54.6	61.8	54.7	54.7
StrongSORT [6]	62.6	73.8	64.0	61.3
OC-SORT [2]	62.4	75.7	62.5	62.4
ByteTrack [30]	61.3	77.8	59.6	63.4
ByteTrack+MCRC	**63.7**	**78.6**	**64.3**	**63.4**

Table 4: Generalization to Extreme Crowds on MOT20 [5].

SportsMOT Analysis. As shown in Table 3, on SportsMOT [4] test, MCRC achieves 81.6 HOTA, 74.0 AssA and 85.6 IDF1, improving over CAMELTrack [24] by 1.2 HOTA and 1.2 AssA with stable DetA. This gain arises from SportsMOT's fast motion and camera panning, which emphasize geometry cues over appearance. When trained directly on the SportsMOT training split, the prototype refiner leverages velocity and scale changes to resolve high-speed encounters, while the VQ codebook captures sport-specific patterns like synchronized team tracking.

Generalization to Crowded Scenes on MOT20. To verify capabilities in densely crowded scenarios, we train and evaluate on MOT20 [5], which features up to 220 pedestrians per frame. Using ByteTrack [30] as the base model demonstrates our method generalizes to another baseline architecture. Table 4 shows MCRC yields gains of 2.4 HOTA and 4.7 AssA with stable DetA. This confirms the geometric refiner resolves intersecting trajectories in dense crowds where simple heuristics fail. It successfully enhances temporal consistency without degrading overall detection quality.

4.3 Ablation Studies

We ablate MCRC on DanceTrack [25] val set because uniform appearance and frequent occlusion expose targeted failure modes. We report HOTA, AssA and IDSW with fixed training schedule, varying a single factor per study.

Component Effectiveness. Starting from CAMELTrack [24], we sequentially add components. As shown in Table 5, prototype-based error refinement addresses miscalibration. It classifies pairs into discrete modes and applies targeted corrections. Vector quantization discretizes the feature space into reusable

Method	HOTA	AssA	IDSW
Baseline	66.8	57.1	1898
+ PR	67.6	57.7	1867
+ PR + VQ	67.9	57.9	1843
+ PR + VQ + Denoise	69.0	59.0	1812

Table 5: Component ablation on DanceTrack [25] val.

Clusters	HOTA	AssA	IDSW
3	68.2	58.3	1855
6	69.0	59.0	1812
9	69.1	59.1	1809
12	69.0	59.0	1815

Table 6: Prototype count ablation on DanceTrack [25] val.

patterns to treat similar configurations consistently. The uncertainty-gated denoiser operates selectively on hard slices where prototype classification fails. Each component targets a distinct failure mode and refines embeddings only when conditionally accepted. Therefore, they cooperatively yield steady gains in temporal consistency.

Prototype Count Ablation. The prototype count determines the granularity of error modes. Table 6 catalogs the effect of concurrently varying the combined prototype clusters. These are divided equally across the spatial geometry P_C and localized motion P_M domains. Using only 3 total clusters performs poorly. It yields 68.2 HOTA and 1855 IDSW. This configuration fails to distinguish distinct failure patterns like partial occlusion versus low overlap. Distributing 6 total clusters achieves optimal performance with 69.0 HOTA and 1812 IDSW. This successfully maps valid tracking and discrete complex errors. Expanding beyond 6 total clusters yields negligible gains. We observe 69.1 HOTA for 9 clusters and 69.0 HOTA for 12 clusters. This indicates that 6 total aggregated prototypes adequately cover the primary error taxonomy without fragmenting the training signal.

Impact of Cue Groups. Having confirmed component effectiveness, we isolate cue group contributions. To prevent co-adaptation artifacts, independent models were trained from random initialization with their respective input channels fully omitted. We avoid test-time zero-masking. From Table 7, motion cues are highly valuable for DanceTrack's complex articulation. Standard Kalman predictions fail here. Geometric cues provide complementary evidence through IoU computations. They partially compensate for motion uncertainty but degrade unconditionally under severe occlusion. Confidence cues offer meaningful associative regularization. This confirms that robust error mode classification mandates diverse multi-cue evidence. Single-cue feature streams inevitably collapse when isolated tracking contexts are physically corrupted.

Denoiser Acceptance. We determine which repairs to commit through post-processing acceptance rules. As shown in Table 8, these safety tests prevent harmful representation updates. They enforce a guaranteed embedding cost reduction alongside a vector drift boundary. The margin test rejects superficial score variations that inject unintended semantic noise. Simultaneously, the fixed-amplitude clipping constraint blocks feature space explosions. This enables semantic refinement safely upon exceptionally degraded tracking pairs.

Variant	HOTA	AssA	IDSW
MCRC full	69.0	59.0	1812
w/o motion	68.0	58.0	1889
w/o geometry	68.4	58.4	1858
w/o confidence	68.7	58.8	1837

Table 7: Cue group ablation.

Acceptance	HOTA	AssA	IDSW
No denoiser	67.6	57.7	1867
Margin only	68.5	58.6	1834
Margin + limit	69.0	59.0	1812

Table 8: Denoiser acceptance gates.

VQ Capacity and Code Alignment We analyze the residual codebook's structure on performance. Capacity scaling, as shown in Table 9, reveals diminishing returns beyond double 128 codes. The chosen double 256 structure, implementing two levels with 256 codes each, best balances expressiveness and generalization without overfitting to training patterns. Additionally, cluster alignment ties these discrete codes to semantic error modes, as outlined in Table 10. Unaligned codes waste capacity on spurious patterns. Conversely, our alignment strategy enforces specialization. This improves generalization when new sequences exhibit familiar failure modes.

Codebook levels × size	HOTA	IDSW
1 × 64	67.4	1917
1 × 128	67.7	1882
2 × 128	68.9	1827
2 × 256	69.0	1821

Table 9: VQ capacity.

Alignment	HOTA	IDSW
Unaligned codes	68.7	1849
Cluster aligned codes	69.0	1829

Table 10: Code alignment.

Computational Cost and Latency. We analyze both theoretical complexity and practical runtime FPS on an RTX 4090. MCRC adds 2.7M parameters. Most of these reside in the denoiser, which activates only on approximately 12 percent of pairs. As shown in Table 11, the theoretical overhead is merely 0.66 GFLOPs.

Experimentally, we evaluate end-to-end processing speed when integrating MCRC. Table 12 reports absolute frame rates under identical hardware conditions. This establishes relative overhead. The processing penalty remains minimal. ByteTrack drops slightly from 23.5 to 22.2 FPS. CAMELTrack shifts from 13.5 to 12.8 FPS. This confirms MCRC operates efficiently as an integration module.

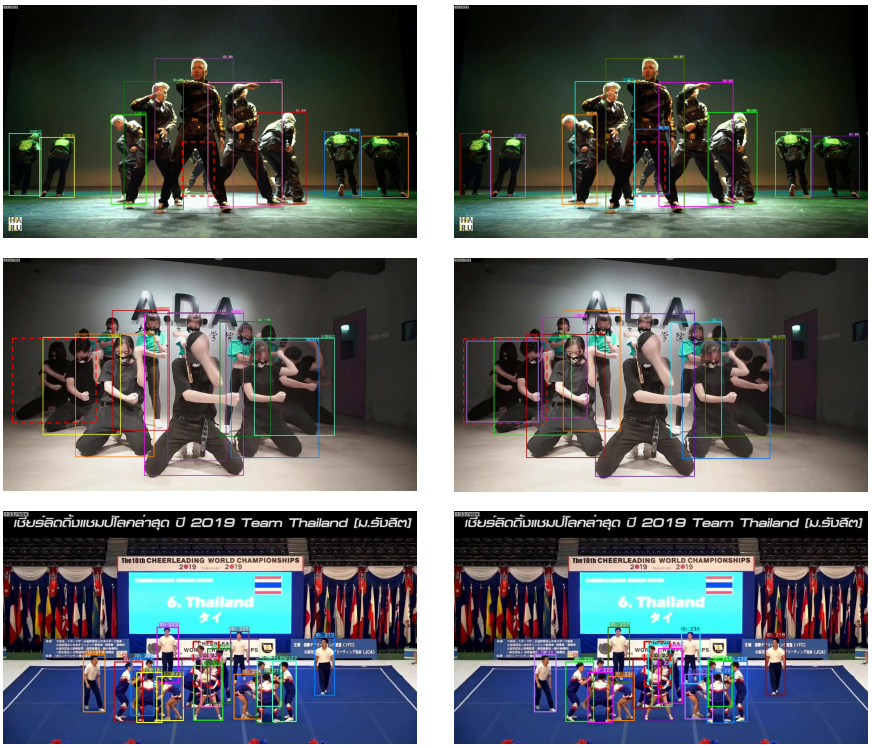
MCRC as an Adaptable Integration Module. We evaluate MCRC as an adaptable refiner across diverse trackers on DanceTrack [25]. As shown in Table 13, MCRC directly integrates with heuristic trackers like ByteTrack [30] and CAMELTrack [24]. The Prototype Refiner and Vector Quantization solely require standard geometrical bounding boxes. For transformer architectures like MOTIP [7] and MeMOTR [8], instances are persistent queries. MCRC extracts geometric cues by comparing decoded queries against detection hypotheses. The continuous residual δ_{ij} adds into the native matching cost matrix. Query embed-

Component	Params	GFLOPs
Prototype refiner	133K	0.12
Residual VQ	50K	0.03
Denoiser (gated)	2.5M	0.51
MCRC total	2.7M	0.66

Table 11: Cost breakdown of MCRC.

Method	FPS↑
ByteTrack [30]	23.5
+ MCRC	22.2
CAMELTrack [24]	13.5
+ MCRC	12.8
MOTIP [7]	20.3
+ MCRC	18.7

Table 12: FPS on DanceTrack val.



(a) CAMELTrack Visualization **(b)** MCRC Visualization

Fig. 3: Qualitative comparison on DanceTrack [25] val set. CAMELTrack [24] in the left panel suffers identity switches and fragmentations under occlusion and crowding. MCRC in the right panel maintains consistent IDs by correcting miscalibrated costs through mode-conditioned refinement and selective denoising. Red dashed boxes highlight recovered trajectories.

dings are highly sensitive to adjusting associations. Therefore, MCRC functions as an architectural extension requiring a brief 5-epoch joint fine-tuning. Strict controls confirm identical extra baseline epochs yield zero gain. Absolute gains peak on heuristic platforms. ByteTrack improves by 4.2 HOTA. Nonetheless,

Method	HOTA	AssA	DetA	IDF1	MOTA
ByteTrack [30]	47.7	32.1	71.0	53.9	91.3
ByteTrack + MCRC	51.9	36.5	70.8	57.8	91.4
CAMELTrack [24]	69.3	58.9	81.8	74.9	91.4
CAMELTrack + MCRC	72.0	63.1	82.1	76.2	91.6
MOTIP† [7]	67.5	57.6	79.4	72.2	90.3
MOTIP + MCRC	69.3	60.4	79.5	74.0	90.4
MeMOTR [8]	68.5	58.4	80.5	71.2	89.9
MeMOTR + MCRC	70.6	61.9	80.5	73.5	89.9

Table 13: Controlled plug-in evaluation across diverse baseline trackers on DanceTrack [25] test set. Baseline rows are reproduced under the same tracker interface and inputs. † denotes a reproduced baseline under our plug-in interface and is used only for controlled before–after comparison, not as the official standalone result.

relative gains remain robust across paradigms. MCRC advances MeMOTR by 2.1 HOTA. This suggests that transformer-based trackers can also benefit from geometric residual calibration.

4.4 Visualization Results

We present qualitative side-by-side tracking visualizations on the DanceTrack [25] val set comparing the CAMELTrack [24] baseline against our MCRC module across challenging sequences. As demonstrated in Fig. 3, MCRC yields fewer identity switches, recovers missed trajectories under severe occlusion, and preserves target geometry. Conversely, the baseline exhibits fragmented identities. These concrete visual trends directly corroborate the empirical ablations. Additional visualizations, including t-SNE embedding separability plots detailing absolute denoiser repair margins, are provided in the supplementary material.

5 Conclusion and Limitations

This work presents Mode-Conditioned Residual Calibration (MCRC). It is a lightweight integration module treating MOT association as mode-dependent miscalibration. It follows a diagnose-then-correct strategy. Applying prototype-based mode diagnosis with residual cost calibration and selectively gated semantic repair improves identity consistency in crowded scenes. The module preserves baseline behavior and incurs minimal overhead. Experiments demonstrate consistent gains across diverse trackers and benchmarks. Explicit mode-aware calibration is a practical complement to existing association scoring.

Limitations. A primary limitation of MCRC is its fixed prototype capacity and codebook entries. These discrete anchors can mismatch underlying data distributions when deployed across completely novel dataset geometries. The network does not dynamically adjust its learned error modes to unseen camera perspectives during zero-shot inference. In future work, we plan to address these

constraints by establishing online dynamic updates for prototypes. This approach will handle cross-domain tracking variations effectively.

Acknowledgements

This work is supported by Doctoral Start-up Foundation of Liaoning Natural Science Foundation under Grant No. 2025-BS-0024 and No. 2026-BS-0409, National Natural Science Foundation of China under Grant No. 62533004.

References

1. Aharon, N., Orfaig, R., Bobrovsky, B.Z.: BoT-SORT: Robust associations for multi-pedestrian tracking. arXiv preprint arXiv:2206.14651 (2022)
2. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9686–9696 (June 2023)
3. Chen, S., Sun, P., Song, Y., Luo, P.: DiffusionDet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19830–19843 (October 2023)
4. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9921–9931 (October 2023)
5. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: MOT20: A benchmark for multi-object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
6. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: StrongSORT: Make DeepSORT great again. IEEE Transactions on Multimedia **25**, 8725–8737 (2023)
7. Gao, R., Qi, J., Wang, L.: Multiple object tracking as ID prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27883–27893 (June 2025)
8. Gao, R., Wang, L.: 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)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017)
10. Huang, H.W., Yang, C.Y., Sun, J., Kim, P.K., Kim, K.J., Lee, K., Huang, C.I., Hwang, J.N.: Iterative Scale-Up ExpansionIoU and deep features association for multi-object tracking in sports. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 163–172 (January 2024)
11. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 97, pp. 3744–3753. PMLR (2019)

12. Luo, R., Song, Z., Ma, L., Wei, J., Yang, W., Yang, M.: DiffusionTrack: Diffusion model for multi-object tracking. In: Proceedings of the AAAI conference on artificial intelligence (AAAI). vol. 38, pp. 3991–3999 (2024)
13. Lv, W., Huang, Y., Zhang, N., Lin, R.S., Han, M., Zeng, D.: DiffMOT: A real-time diffusion-based multiple object tracker with non-linear prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19321–19330 (June 2024)
14. Maggiolino, G., Ahmad, A., Cao, J., Kitani, K.: Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification. In: 2023 IEEE International conference on image processing (ICIP). pp. 3025–3029 (2023)
15. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: VQ-VAE made simple. In: Proceedings of the International Conference on Learning Representations (ICLR). pp. 51772–51783 (2024)
16. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
17. Qin, Z., Zhou, S., Wang, L., Duan, J., Hua, G., Tang, W.: MotionTrack: Learning robust short-term and long-term motions for multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17939–17948 (June 2023)
18. Segu, M., Piccinelli, L., Li, S., Yang, Y.H., Van Gool, L., Schiele, B.: Samba: Synchronized set-of-sequences modeling for multiple object tracking. In: Proceedings of the International Conference on Learning Representations (ICLR). pp. 30057–30070 (2025)
19. Seidenschwarz, J., Brasó, G., Serrano, V.C., Elezi, I., Leal-Taixé, L.: Simple cues lead to a strong multi-object tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13813–13823 (June 2023)
20. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
21. Shao, Y., Yang, Y., Yu, R., Li, W., Guo, X., Yan, H., Wang, W., Sun, X.: 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 (CVPR). pp. 6878–6888 (2026)
22. Shim, K., Ko, K., Yang, Y., Kim, C.: Focusing on tracks for online multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11687–11696 (2025)
23. Solano-Carrillo, E., Sattler, F., Alex, A., Klein, A., Pereira Costa, B., Bueno Rodriguez, A., Stoppe, J.: UTrack: Multi-object tracking with uncertain detections. In: Computer Vision – ECCV 2024. pp. 219–236. Springer (2024)
24. Somers, V., Standaert, B., Joos, V., Alahi, A., De Vleeschouwer, C.: CAMELTrack: Context-aware multi-cue exploitation for online multi-object tracking. arXiv preprint arXiv:2505.01257 (2025)
25. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20993–21002 (2022)
26. Van Den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)

27. Xu, Z., Wang, G., Huang, X., Sang, J.: DenoiseRep: Denoising model for representation learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 40032–40056 (2024)
28. Yang, M., Han, G., Yan, B., Zhang, W., Qi, J., Lu, H., Wang, D.: Hybrid-SORT: Weak cues matter for online multi-object tracking. In: Proceedings of the AAAI conference on artificial intelligence (AAAI). vol. 38, pp. 6504–6512 (2024)
29. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: MOTR: End-to-end multiple-object tracking with transformer. In: European conference on computer vision (ECCV). pp. 659–675. Springer (2022)
30. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: European conference on computer vision (ECCV). pp. 1–21. Springer (2022)
31. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: FairMOT: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021)
32. Zhang, Y., Wang, T., Zhang, X.: 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 (CVPR). pp. 22056–22065 (2023)
33. Zheng, R.C., Du, H.P., Jiang, X.H., Ai, Y., Ling, Z.H.: ERVQ: Enhanced residual vector quantization with intra-and-inter-codebook optimization for neural audio codecs. IEEE Transactions on Audio, Speech and Language Processing **33**, 2539–2550 (2025)