

Table 4: The ablation experiments of applying CO-MOT to TrackFormer on the DanceTrack validation set. As components are added, the tracking performance improves gradually.

	COLA	Shadow	HOTA
(a) TrackFormer			41.4
(b)	✓		47.8(+6.4)
(c)	✓	✓	50.7(+9.3)

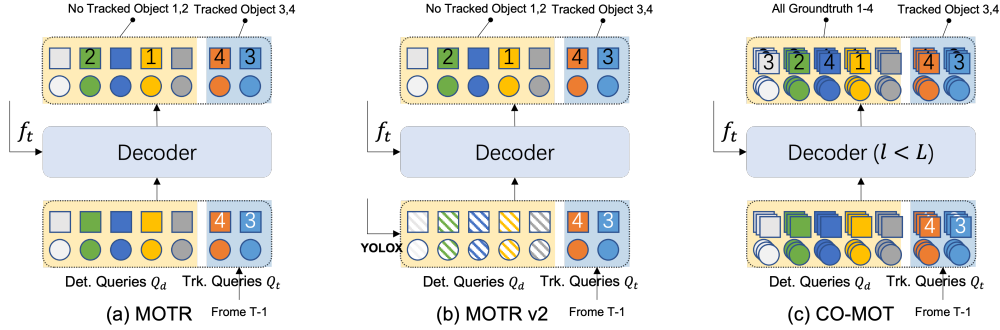


Figure 6: Comparison among MOTR, MOTRv2, and CO-MOT (ours).

A TRACKFORMER WITH COLA AND SHADOW SET

COLA and Shadow Set are model-independent methods that can be applied to any end-to-end multi-object tracking model, not just MOTR, but also TrackFormer. We do not make any modifications to the TrackFormer framework and used the default hyperparameters provided by TrackFormer. The result shown as Table 4 demonstrates the effectiveness of COLA (+6.4%) and Shadow Set (+9.3%) when applied to TrackFormer. Codes will be released upon acceptance.

B COMPARISON AMONG MOTR, MOTRv2, AND CO-MOT

As shown in the Figure 6, MOTR(v2) use TALA supervision for queries output, the output of detection queries can only match with newborn targets (targets 1 and 2). CO-MOT uses COLA supervision for $l < L$ decoder layers, allowing the output of detection queries to match not only with newborn targets (targets 1 and 2), but also with already tracked targets (targets 3 and 4). At the same time, CO-MOT uses the Shadow Set method, enabling multiple queries to match the same target.

C EXPLANATION OF THE EFFECT OF INTRODUCING COLA

As shown in Figure 3, the goal of COLA is not only to improve detection performance, but more importantly, to enhance tracking performance. During training, TALA forcibly matches the tracking query with the corresponding track object, even if the current prediction of the tracking query is far from the target. This strategy makes it more difficult for the network to converge compared to the optimal solution of the Hungarian matching, which in turn makes it difficult for the track query to reach the optimal solution. Introducing COLA can improve this convergence. The detect query recalls the tracked targets through Hungarian matching, and then passes the relevant information to the track query through the attention mechanism of the Transformer. With the prior knowledge from the detect query, the convergence of the track query becomes easier, and the target will be consistently assigned the same ID, resulting in a higher AssA.

D THE REASON FOR SAMPLE IMBALANCE IN E2E-MOT

The TALA matching mechanism mainly causes the imbalance between the detection query and the tracking query. TALA first matches the tracking query, and then the remaining ground truth (newborns) is matched with the detection query. In most scenarios, especially in closed scenarios, there are very few newborns in the video frames except for the first frame. We count the DanceTrack data and find that the ratio of newborns to tracked targets is 213: 25483. This easily result in insufficient training for the detection query.

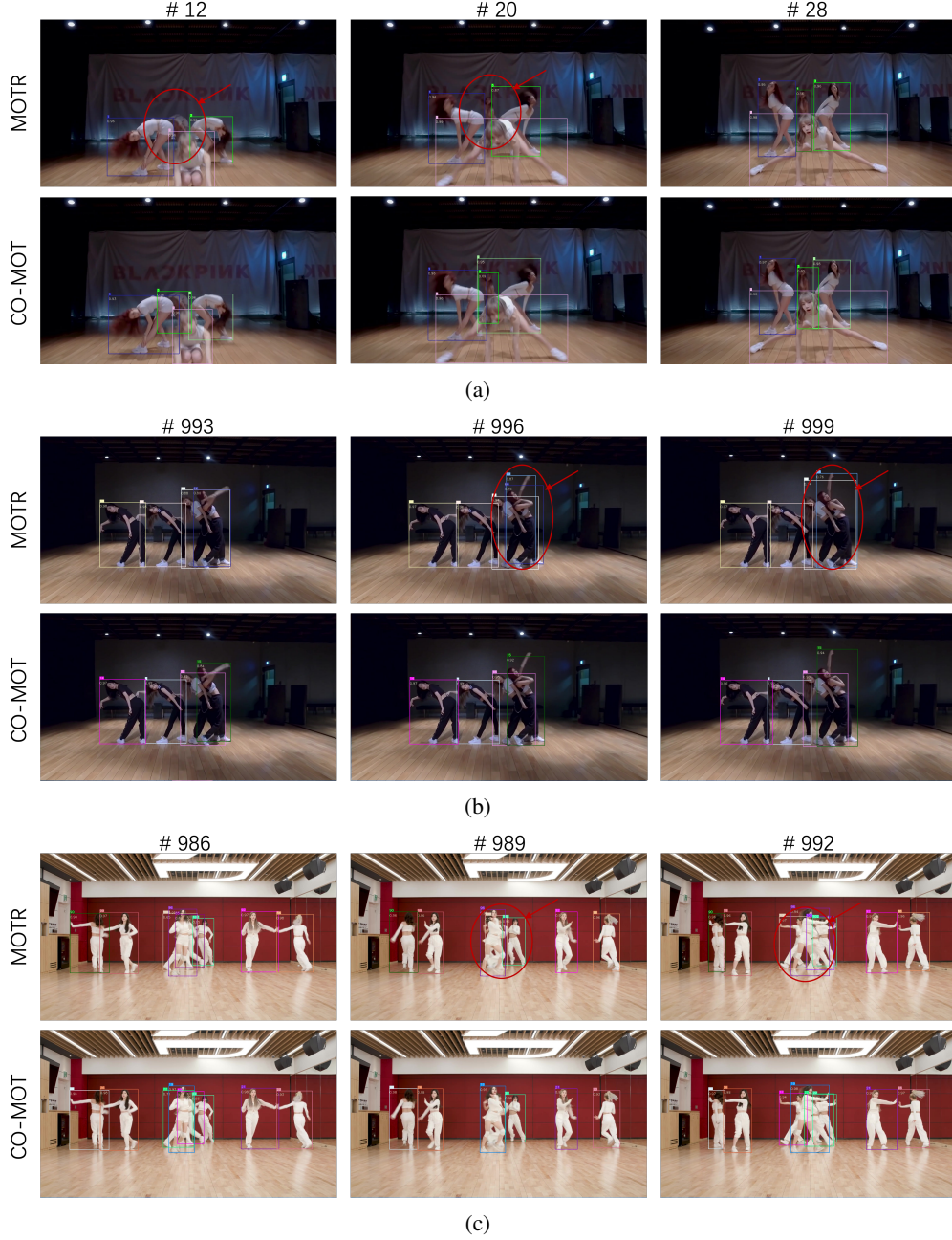


Figure 7: Failed cases cases of MOTR and denote which kind of case has been solved by CO-MOT.

E COMPARISON OF FAILURE CASES

The Figure 7 illustrates the failure cases of MOTR and denotes which kind of case has been solved by CO-MOT. MOTR has poor detection and tracking performance. First, as shown in Figure 7a, it fails to detect the person in time under the extreme case of bending over. Second, as shown in Figure 7b, due to the tiny visual features when a person stretches out their hand, the detection box is inaccurate, and the model misidentifies it as multiple people. Third, as shown in Figure 7c, the tracking identity switches after the human body is obscured. However, all of the above cases can be solved by CO-MOT, showing the extraordinary performance of MOT.