
Supplementary Material for

Dual-Path Temporal Decoder for End-to-End Multi-Object Tracking

1 A Analysis of sampling offset patterns between consecutive frames

2 To analyze sampling offset patterns, we construct a histogram of offset differences defined as
 3 $\delta^{(t)} = \Delta_T^{(t)} - \Delta_T^{(t-1)}$, which exploits the temporal variation of sampling offsets between consecutive
 4 frames. Figures S-1 (a) and (b) show the histograms of all offset differences $\delta^{(t)}$ across all frames for
 5 all objects and correctly matched objects, respectively. We observe that both histograms exhibit sharp
 6 peaks centered around zero, indicating that the same object typically exhibits stable sampling offset
 7 patterns across adjacent frames. Notably, the histogram for correctly matched objects in Figure S-1(b)
 8 is more tightly concentrated near zero, indicating that smaller offset differences tend to be associated
 9 with more stable and accurate tracking.

10 Figure S-2 compares histograms of sampling offset differences (a) without and (b) with the identity-
 11 preserving decoder layer (IDL). The histograms are constructed from frames where identity switches
 12 (IDSW) occurred under the configuration without IDL. In Figure S-2 (a), the offset differences are
 13 large and widely dispersed, representing unstable and inconsistent sampling offsets in the absence
 14 of IDL. In contrast, the result shows that the offset differences in Figure S-2 (b) are more tightly
 15 concentrated around zero, indicating that IDL yields more stable sampling offsets. This confirms that
 16 the proposed IDL design effectively mitigates abrupt offset shifts and helps reduce identity switches.

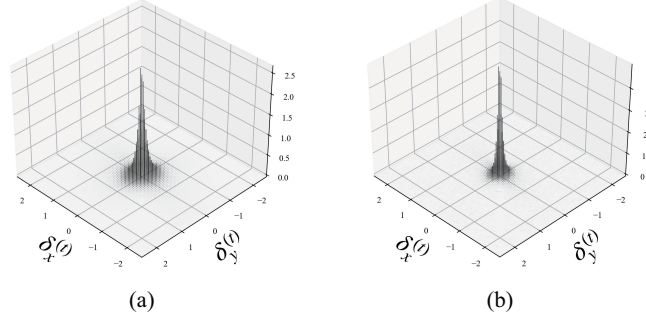


Figure S-1: Histograms of offset differences $\delta^{(t)}$ across all frames for (a) all objects and (b) correctly matched objects. $\delta_x^{(t)}$ and $\delta_y^{(t)}$ represent the offset differences along the x - and y -axes, respectively.

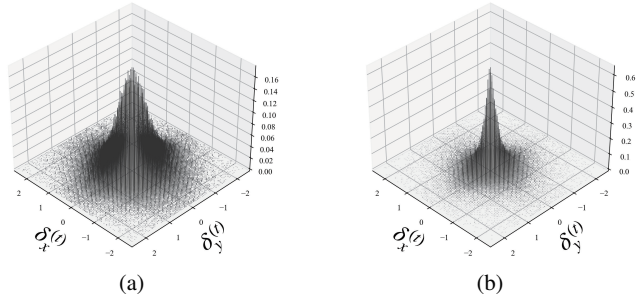


Figure S-2: Histograms of sampling offset differences (a) without and (b) IDL at IDSW frames.

17 B Efficiency Comparison

Table S-1: Comparison of transformer-based model efficiency on Dancetrack [26] validation set.

Method	FLOPs (G)	Params (M)	FPS
MOTRv2 [40]	431.8	41.7	25.7
ColTrack [16]	600.9	15.8	11.3
MOTIP [11]	-	58.9	12.0
Ours	623.0	11.3	12.7

18 Table S-1 presents the computational efficiency of the proposed MOT compared to representative
 19 transformer-based methods, evaluated in terms of FLOPs, the number of parameters, and frames
 20 per second (FPS). All FPS measurements are conducted on an NVIDIA GeForce 4090Ti GPU. We
 21 measure FPS the proposed MOT and other transformer-based approaches using NVIDIA GeForce
 22 4090Ti GPU. The proposed MOT has the smallest parameter count among all methods. Among
 23 end-to-end approaches (ColTrack [16] and MOTIP [11]), the proposed MOT achieves the highest
 24 FPS. Note that MOTRv2 [40] relies on precomputed bounding boxes from YOLOX [13] detector,
 25 and its speed measurement excludes the time required for detection inference.

26 C Experiments on the SportMOT dataset

27 Table S-2 lists a comparison of the proposed MOT with existing methods on the SportsMOT [9]
 28 test set, where all models are trained using both the official training and validation sets. Under this
 29 unified training setting, the proposed method achieves the best overall performance across multiple
 30 tracking metrics. In particular, it outperforms the previous state-of-the-art DiffMOT [19] in terms of
 31 association accuracy (AssA: 69.6 vs. 65.1) and identity preservation (IDF1: 82.5 vs. 76.1), while also
 32 attaining the highest HOTA score (76.7). These results demonstrate the effectiveness of the proposed
 33 dual-path temporal decoder in improving association quality and identity consistency in challenging
 34 MOT scenarios. Although DiffMOT [19] achieves the best results in detection-focused metrics such
 35 as DetA and MOTA, this advantage is partly attributable to its use of an external YOLOX detector, as
 36 it does not follow a fully end-to-end tracking framework.

Table S-2: Quantitative comparison on the SportsMOT [9] test set, where all models are trained using both training and validation data. Bold indicates the best.

Methods	HOTA	DetA	AssA	MOTA	IDF1
ByteTrack [38]	64.1	78.5	52.3	95.9	71.4
MixSort-Byte [9]	65.7	78.8	54.8	96.2	74.1
OC-SORT [6]	73.7	88.5	61.5	96.5	74.0
MixSort-OC [9]	74.1	88.5	62.0	96.5	74.4
DiffMOT [19]	76.2	89.3	65.1	97.1	76.1
Ours	76.7	84.5	69.6	94.8	82.5

37 D Qualitative results

38 Figures S-3 and S-4 illustrate qualitative comparison of the proposed MOT with MOTRv2 [40] and
 39 ColTrack [16] on Dancetrack [26] validation set. We observe that the proposed MOT consistently
 40 preserves identity even in challenging scenes with severe occlusions.

41 E Limitations

42 Despite the strong performance, our approach has limitations. The proposed MOT shows relatively
 43 lower performance in MOT17, which is scenarios with linear motion, compared to heuristic methods
 44 that incorporate heuristic algorithms. This issue is commonly observed across end-to-end MOT

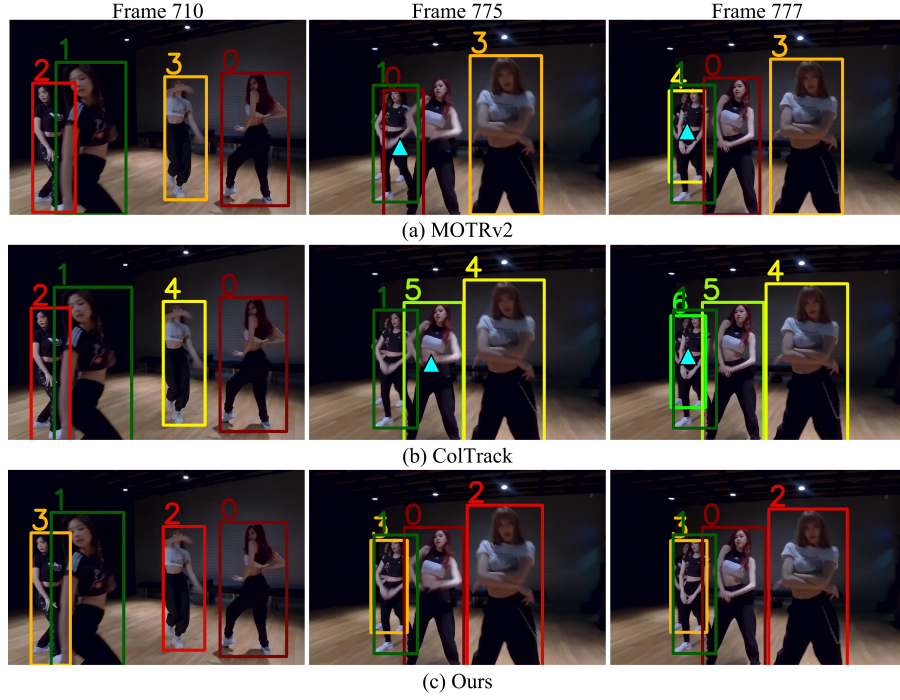


Figure S-3: Qualitative comparison for the Dancetrack0005 sequence in the validation set of Dance-track [26]. Failure examples are marked with blue triangles.



Figure S-4: Qualitative comparison for the Dancetrack0010 sequence in the validation set of Dance-track [26]. Failure examples are marked with blue triangles.

45 frameworks and was also partially present in the proposed method, highlighting an inherent limitation
 46 that calls for further investigation to fully address the performance gap.