# MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking: Supplementary Material

## A. Boosting Tracking Performance

For the tracking-by-detection paradigm, with the development of the Object Detection task, they upgraded the detector used in MOT from Faster R-CNN [4] to YOLOX [1] and obtained impressive detection performance. Although Deformable-DETR [8] has competitive detection performance, it still lags behind some popular detectors such as YOLOX [1]. This will impair the final tracking performance.

Recently, unlike the original Deformable-DETR [8], some methods [7] generate the position embeddings from the learnable anchors. On the one hand, this design will improve the model's detection performance, as discussed in many object detection studies [2]. On the other hand, the anchor-based position-prior is quite effective for tracking due to frame continuity.

Therefore, as discussed in Section 4.2, we built our MeMOTR upon DAB-Deformable-DETR [2] instead of Deformable-DETR [8]. We believe that better detection performance of DAB-Deformable-DETR will lead to better tracking performance, as shown in Table 1 (#2 *vs.* #5). We discuss that DAB-Deformable-DETR can be applied in future works as a technology development (like from Faster R-CNN [4] to YOLOX [1] in the tracking-by-detection paradigm). For a fair comparison with previous transformer-based methods [3, 6], we also provide the results of MeMOTR based on the standard Deformable-DETR in Table 1 (main page) and 1 (#2 and #3). This indicates our method still has impressive performance without DAB-Deformable-DETR. As ablation experiments in MOTRv2 [7], we further add the anchor-based position generation process to the standard Deformable-DETR in our method, thus slightly improving the tracking performance (Table 1 #3 *vs.* #4).

Moreover, we also add the YOLOX [1] proposal to our model following MOTRv2 [7]. As they concluded, this significantly improves the detection and tracking performance simultaneously (Table 1 #8). Since the proposals are generated from a frozen CNN-based model, it makes the whole model a non-fully-end-to-end method. For this reason, we list MOTRv2 [7] in Table 2 as a new hybrid architecture.

In summary, we provide the cumulative improvements over MOTR [6] on the val and test set of DanceTrack [5], as shown in Table 1. This further verifies the effectiveness of our various components and gives a more intuitive comparison.

## B. Comparison on Difficult Sequences

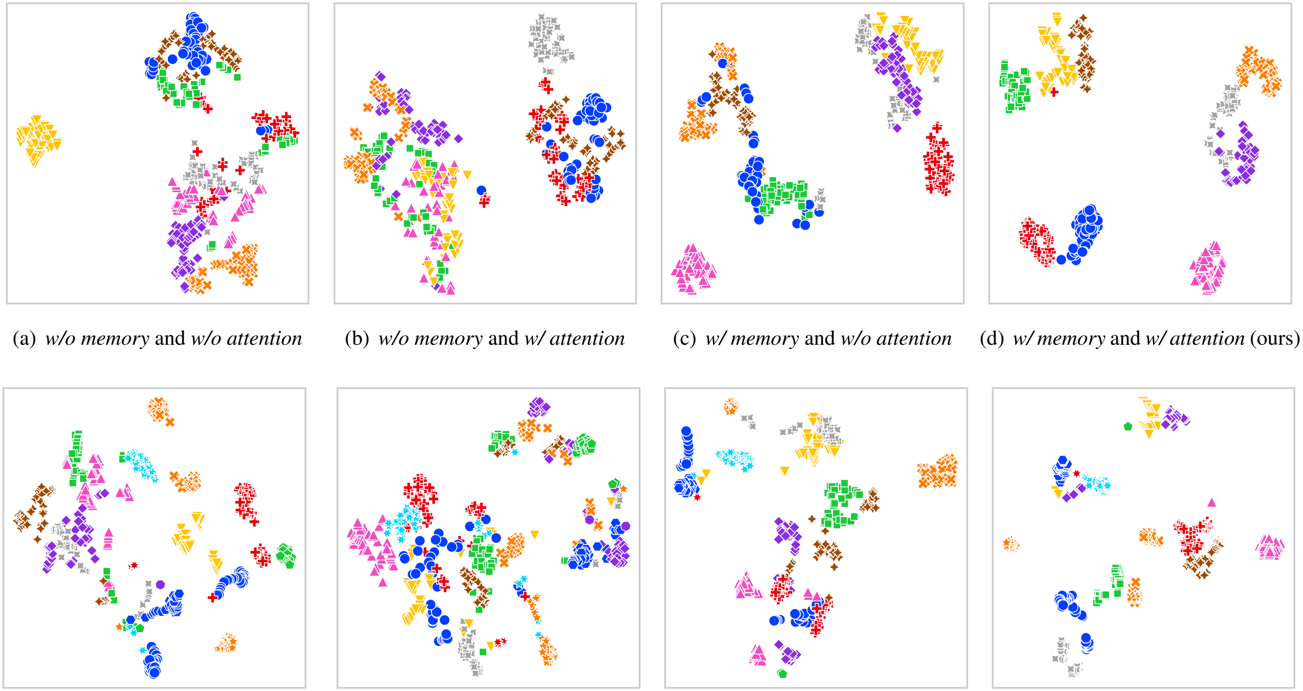
In order to further certify the improvement of our method on target association challenge, we list experimental metrics on some challenging sequences. We selected eight sequences with the lowest AssA metric of MOTR [6] on the DanceTrack [5] validation set. As shown in Table 2, the association results on these complex sequences are unsatisfactory (23.6 average AssA), although the detection performance is passable (65.8 average DetA). Our method substantially improves the performance of object association (35.5 *vs.* 23.6 AssA) while slightly improving detection performance (70.9 *vs.* 65.8 DetA). However, compared to the overall association performance (58.4 AssA of our method), there is still a significant deficiency in the results of these challenging sequences. Therefore, we suggest that improving the object association performance of multi-object tracking is still an unsolved problem that should not be ignored.

## C. More Visualizations

In this section, we supply additional visualization results. Same as Figure 4 in our main paper, we utilize t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize track embeddings. More visualizing results are provided in Figure 1, the upper (Figure 1(a) to 1(d)) is from *dancetrack0025*, and the lower (Figure 1(e) to 1(h)) is from *dancetrack0034* sequence. These results further verify that our *long-term memory* and *memory-attention layer* help learn a more stable and distinguishable representation for the tracked target.

| # Row | val set |  |  |  |  | test set |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | HOTA | DetA | AssA | MOTA | IDF1 | HOTA | DetA | AssA | MOTA | IDF1 |
| 1. MOTR (baseline) | 51.7 | 69.4 | 38.7 | 75.6 | 49.7 | 54.2 | 73.5 | 40.2 | 79.7 | 51.5 |
| 2. #1 + *memory-augment* | 56.5 | 70.4 | 45.5 | 78.4 | 58.8 | 62.5 | 77.0 | 50.9 | 85.1 | 63.5 |
| 3. #2 + $\mathcal{L}_d, \mathcal{L}_j = 1, 5$ | 61.0 | 71.2 | 52.5 | 79.2 | 64.1 | 63.4 | 77.0 | 52.3 | 85.4 | 65.5 |
| 4. #3 + Anchor | 61.1 | 73.0 | 51.3 | 81.3 | 63.8 | 64.6 | 78.4 | 53.4 | 87.6 | 67.3 |
| 5. #2 + DAB-D-DETR | 62.1 | 74.3 | 52.2 | 83.1 | 65.6 | 65.9 | 78.8 | 55.2 | 87.9 | 68.9 |
| **6. #5 + $\mathcal{L}_d, \mathcal{L}_j = 1, 5$** | **63.9** | **74.6** | **55.0** | **83.4** | **67.1** | **68.5** | **80.5** | **58.4** | **89.8** | **71.2** |
| 7. #5 + $\mathcal{L}_d, \mathcal{L}_j = 2, 4$ | 63.2 | 73.8 | 54.3 | 81.9 | 65.8 | 66.2 | 80.2 | 54.8 | 89.5 | 68.7 |
| 8. #6 + YOLOX [11] | 66.8 | 78.7 | 57.0 | 88.1 | 70.5 | 70.0 | 81.8 | 60.1 | 90.3 | 72.5 |

Table 1. Supplemental comparison on DanceTrack [5]. Best viewed in color. The same base color results represent using the same DETR framework (D-DETR [8] or DAB-D-DETR [2]). $\mathcal{L}_d$ and $\mathcal{L}_j$ are the numbers of detection and joint decoder layers in Figure 1 (main page), respectively. It should be noted that except for the baseline (#1), training augmentations (track query erasing and false positive inserting in MOTR [6]) are removed from other experiments.



(a) *w/o memory* and *w/o attention* (b) *w/o memory* and *w/ attention* (c) *w/ memory* and *w/o attention* (d) *w/ memory* and *w/ attention* (ours)

(e) *w/o memory* and *w/o attention* (f) *w/o memory* and *w/ attention* (g) *w/ memory* and *w/o attention* (h) *w/ memory* and *w/ attention* (ours)

Figure 1. Visualizing track embedding $E_{tck}^{t}$ from the first 50 frames of *dancetrack0025* (upper) and *dancetrack0034* sequences (lower). Track embeddings for different tracked targets (IDs) are marked in different colors and shapes. The visualizations of our method are shown in Figure 1(d) and 1(h).

## References

[1] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. *CoRR*, abs/2107.08430, 2021. 1

[2] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: dynamic anchor boxes are better queries for DETR. In *ICLR*. OpenReview.net, 2022. 1, 2

[3] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, pages 8834–8844. IEEE, 2022. 1

[4] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1

[5] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai,

| Sequence | MOTR [6] | | | MeMOTR *(ours)* | | |
|---|---|---|---|---|---|---|
| | HOTA | DetA | AssA | HOTA | DetA | AssA |
| dancetrack0041 | 30.6 | 50.9 | 18.8 | 32.9 | 49.4 | 22.4 |
| dancetrack0081 | 35.7 | 63.9 | 20.0 | 46.0 | 69.6 | 30.4 |
| dancetrack0063 | 30.5 | 44.9 | 20.7 | 43.2 | 57.4 | 32.6 |
| dancetrack0019 | 40.8 | 79.3 | 21.0 | 49.7 | 84.2 | 29.3 |
| dancetrack0014 | 43.7 | 79.4 | 24.2 | 47.6 | 80.0 | 28.4 |
| dancetrack0004 | 44.1 | 77.0 | 25.3 | 66.2 | 82.7 | 53.0 |
| dancetrack0034 | 42.3 | 65.7 | 27.3 | 56.9 | 73.2 | 44.3 |
| dancetrack0090 | 45.1 | 65.1 | 31.5 | 55.4 | 70.5 | 43.7 |
| *average* | 39.1 | 65.8 | 23.6 | **49.7** | **70.9** | **35.5** |

Table 2. Comparison on difficult sequences on DanceTrack [5] validation set. The results of MOTR [6] are obtained from the checkpoint provided by the official repo.

Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, pages 20961–20970. IEEE, 2022. 1, 2, 3

[6] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: end-to-end multiple-object tracking with transformer. In *ECCV (27)*, volume 13687 of *Lecture Notes in Computer Science*, pages 659–675. Springer, 2022. 1, 2, 3

[7] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors, 2022. 1

[8] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*. OpenReview.net, 2021. 1, 2