Gated Temporal Fusion Transformers for Robust Multi-Object Tracking

Supplementary Material

We provide additional implementation details (Sec. A), additional experimental results (Sec. B), and limitations (Sec. C) which are not included in the main paper due to space limitations.

A. Additional Implementation Details

Model Backbone. Our framework builds upon MOTIP[9], which utilizes a Deformable DETR backbone[25] with six encoder and six decoder layers. Multi-scale features are extracted and processed through deformable attention to capture spatial relationships, before integrating temporal information.

Tracklet Selection and Temporal Fusion. To reduce computational cost, we select only reliable tracklets from the decoder outputs of previous frames whose confidence scores exceed 0.5. The selected tracklets guide the extraction of corresponding regions in previous feature maps, ensuring that temporal fusion focuses on trustworthy object candidates. Temporal feature fusion is applied only to the last two encoder layers. Each temporal encoder integrates our gating-based module, which is realized as a two-layer MLP with ReLU activation and dropout. The gating outputs sigmoid-activated weights that adaptively balance current and historical features via weighted summation.

Temporal Memory A temporal memory stores tracklet queries propagated from decoder outputs. At each frame, memory is updated using an exponential moving average with momentum 0.9, which stabilizes temporal evolution and prevents abrupt representation shifts. The memory is then injected into the encoder to provide temporally adaptive features.

Training Objective. In addition to the baseline detection and tracking losses, we supervise the gating module by aligning its predicted values with cosine similarity-based targets. Entropy regularization is applied to promote diversity in gating behavior and avoid trivial solutions. The gating loss and entropy regularization weights are set to 0.01 and 0.06, respectively, with the latter decayed throughout training. Training employs a learning rate warm-up during the first two epochs to enhance convergence stability.

Encoder Layer Algorithm Details. To clarify the design, we provide pseudocode for the three key components: - **DeformableTransformerEncoder (Algorithm 1)**: coordinates spatial and temporal layers, maintaining reference points for consistent spatial grounding. - **DeformableTransformerEncoderLayer (Algorithm 2)**: performs standard multi-scale deformable attention for spatial reasoning. - **DeformableTransformerEncoderLayerTemp (Algorithm 3)**: performs temporal feature integration by (i) selecting high-confidence targets, (ii) extracting local grids around target regions, (iii) cross-attending with previous features, (iv) computing similarity metrics, and (v) adaptively gating fused features.

These algorithms highlight the modular and differentiable nature of our design, allowing seamless integration of spatial and temporal reasoning within Transformer-based MOT frameworks.

Algorithm 1: DeformableTransformerEncoder Forward Pass

Input: Input features $\mathbf{src}$,
spatial shapes $\mathbf{shape}$,
level start index $\mathbf{index}$,
valid ratios $\mathbf{ratios}$,
positional embedding $\mathbf{pos}$,
padding mask $\mathbf{mask}$,
previous features $\mathbf{pres}$

Output: Final features $\mathbf{output}$,
reference points $\mathbf{ref}$,
gate list, spatial output

$\mathbf{ref} \leftarrow \text{GetReferencePoints}(\mathbf{shape}, \mathbf{ratios})$
foreach *layer* $\in$ *encoder layers* **do**
  $\mathbf{output} \leftarrow \text{layer.forward}(\mathbf{output}, \mathbf{pos}, \mathbf{ref}, \mathbf{shape}, \mathbf{index}, \mathbf{mask})$
$\mathbf{output}_{spatial} \leftarrow \mathbf{output}$
foreach *layer* $\in$ *temporal layers* **do**
  $\mathbf{output}, \mathbf{gate} \leftarrow \text{layer.forward}(\mathbf{output}, \mathbf{pos}, \mathbf{ref}, \mathbf{shape}, \mathbf{index}, \mathbf{mask}, \mathbf{pres})$
  Append $\mathbf{gate}$ to gate list
return $\mathbf{output}$, $\mathbf{ref}$, gate list, $\mathbf{output}_{spatial}$

B. Additional Experimental Results

B.1. Detailed Datasets and Evaluation Metrics

To evaluate the performance of our proposed MOT framework, we employ three challenging and widely used benchmarks: **MOT17**, **DanceTrack**, and **SportsMOT**. These datasets collectively cover diverse scenarios with varying levels of scene complexity, motion dynamics, and identity

Algorithm 2: DeformableTransformerEncoderLayer Forward Pass

Input: Input features $\mathbf{src}$,
positional embeddings $\mathbf{pos}$,
reference points $\mathbf{ref}$,
spatial shapes $\mathbf{shape}$,
level start index $\mathbf{index}$,
padding mask $\mathbf{mask}$
Output: Updated features $\mathbf{src}$

$\mathbf{src2} \leftarrow \text{MSDeformAttn}(\mathbf{src} + \mathbf{pos}, \mathbf{ref}, \mathbf{src}, \mathbf{shape}, \mathbf{index}, \mathbf{mask})$
$\mathbf{src} \leftarrow \text{LayerNorm}(\mathbf{src} + \text{Dropout}(\mathbf{src2}))$
$\mathbf{src2} \leftarrow \text{Linear}_2(\text{Dropout}(\text{Activation}(\text{Linear}_1(\mathbf{src}))))$
$\mathbf{src} \leftarrow \text{LayerNorm}(\mathbf{src} + \text{Dropout}(\mathbf{src2}))$
return $\mathbf{src}$

Algorithm 3: DeformableTransformerEncoderLayerTemp Forward Pass

Input: Input features $\mathbf{src}$,
positional embedding $\mathbf{pos}$,
reference points $\mathbf{ref}$,
spatial shapes $\mathbf{shape}$,
level start index $\mathbf{index}$,
padding mask $\mathbf{mask}$,
previous features $\mathbf{pres}$
Output: Updated features $\mathbf{src}$,
gate target $\mathbf{gate}$

$\mathbf{src} \leftarrow \text{SelfAttention}(\mathbf{src} + \mathbf{pos}, \mathbf{ref}, \mathbf{src})$
$\mathbf{src} \leftarrow \text{FFN}(\mathbf{src})$
if $\mathbf{pres} \neq$ *None* **then**
  Extract $\mathbf{tgt}, \mathbf{tgt}_{pos}, \mathbf{tgt}_{coord}, \mathbf{tgt}_{prob}$, $\mathbf{pre}, \mathbf{pos}_{pre}$ from $\mathbf{pres}$
  $tgt_{select} \leftarrow$ Select high confidence target tokens based on $\mathbf{tgt}_{prob}$
  Generate 3x3 grid offsets around each target box
  $\mathbf{pre_{comp}} \leftarrow$ Gather corresponding features from $\mathbf{pre}$ using grid indices
  Apply self-attention and FFN to $\mathbf{tgt_{select}}$
  Apply self-attention and FFN to $\mathbf{pre_{comp}}$
  Apply cross-attention from $\mathbf{pre_{comp}}$ to $\mathbf{tgt}$
  Compute fused features $\mathbf{src}_c$ via cross-attention from $\mathbf{src}$ to $\mathbf{pre_{comp}}$
  Compute similarity metrics (cosine + Euclidean) between $\mathbf{src}$ and $\mathbf{src}_c$
  Compute gate value via MLP and fuse:
    $\mathbf{src}_c \leftarrow \text{gate} \cdot \mathbf{src}_c + (1 - \text{gate}) \cdot \mathbf{src}$
  $\mathbf{src} \leftarrow \text{LayerNorm}(\mathbf{src} + \text{Dropout}(\mathbf{src}_c))$
  $\mathbf{src} \leftarrow \text{FFN}(\mathbf{src})$
return $\mathbf{src}, \mathbf{gate}$

ambiguity, thereby providing a rigorous testbed for temporal reasoning in tracking.

MOT17 is a mid-scale dataset comprising diverse urban environments such as streets, stations, and shopping malls. It poses multiple challenges including heavy occlusion, frequent camera motion, and crowded pedestrian interactions. The benchmark contains 7 training and 7 testing sequences with dense bounding box annotations. MOT17 has long served as a standard benchmark for pedestrian tracking, testing both detection quality and short-term association performance. The dataset features relatively linear motion patterns with predictable pedestrian behaviors, making it suitable for evaluating conventional tracking approaches.

DanceTrack is a recent large-scale benchmark that emphasizes long-term temporal consistency in complex motion scenarios. It focuses on indoor dance performances where subjects frequently undergo abrupt pose changes, unpredictable motion patterns, and heavy interactions while wearing visually similar outfits. DanceTrack contains about 100 sequences with more than 70K frames, and it is particularly prone to identity switches due to appearance ambiguity and group dynamics. Unlike MOT17, which emphasizes crowded urban pedestrian scenes, DanceTrack stresses appearance-invariant tracking under non-rigid motion, making it an ideal benchmark for evaluating temporal modeling capacity. The dataset's uniform appearance characteristic forces trackers to rely heavily on motion-based association rather than appearance cues.

SportsMOT is a large-scale benchmark designed for athlete tracking in fast-paced sports such as basketball, soccer, and volleyball. It comprises 240 sequences, over 150K frames (15× MOT17), and more than 1.6 million annotated bounding boxes (3× MOT17). SportsMOT presents unique challenges including rapid and erratic motion, frequent occlusions, uniform appearance similarity among players, and frequent viewpoint shifts due to moving cameras. These factors often lead to the highest rate of ID switches among the three benchmarks, making SportsMOT an essential benchmark for assessing the robustness of temporal reasoning under extreme dynamics. The dataset features two key properties: (1) fast and variable-speed motion requiring sophisticated motion modeling, and (2) similar yet distinguishable appearance necessitating discriminative visual representations.

Comparative Analysis. As shown in IoU analysis across adjacent frames, SportsMOT exhibits the lowest IoU scores compared to MOT17 and DanceTrack, indicating faster object motion. Football sequences within SportsMOT demonstrate the most challenging motion patterns. DanceTrack highlights diverse but relatively slower motion patterns, while MOT17 features more predictable

linear pedestrian movement. These complementary characteristics make the three datasets collectively suitable for comprehensive MOT evaluation.

Collectively, these datasets enable comprehensive evaluation of MOT systems across short-term crowded pedestrian tracking, long-term motion consistency, and highly dynamic multi-agent scenarios. This diversity is particularly suitable for analyzing the effectiveness of encoder-level temporal fusion and adaptive gating mechanisms.

Evaluation Metrics. We evaluate our method using the most established MOT metrics in recent benchmarks, focusing on HOTA, DetA, AssA, MOTA, and IDF1.

Higher Order Tracking Accuracy (HOTA) [11] is adopted as our primary metric, as it provides a unified assessment of both detection and association performance. HOTA explicitly balances detection accuracy and identity association by decomposing into interpretable sub-metrics: detection accuracy (DetA), which quantifies the recall of ground-truth objects by the tracker, and association accuracy (AssA), which measures the temporal consistency of object identities throughout a sequence. By jointly considering these aspects, HOTA offers a nuanced view of tracker performance that is well-aligned with real-world MOT needs. HOTA addresses limitations of previous metrics by being less susceptible to detection-dominated or association-dominated biases.

MOTA [2] is also reported as a standard metric, summarizing tracking performance by cumulatively accounting for false positives, false negatives, and identity switches, making it an overall error-focused measure; however, it is known to be dominated by detection performance and does not directly capture identity consistency.

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t}, \qquad (1)$$

where IDSW represents identity switches.

To further evaluate identity preservation, we report **IDF1** [14], which computes the harmonic mean of precision and recall of correctly identified detections, effectively reflecting the accuracy of trajectory-wise identity assignment. IDF1 performs global trajectory matching across the entire sequence rather than frame-by-frame evaluation, making it more sensitive to long-term tracking consistency.

Additional Metrics. We also report DetA and AssA as sub-components of HOTA to provide detailed analysis. DetA isolates detection quality independent of association performance, while AssA focuses purely on identity consistency. This decomposition enables targeted analysis of tracking failures and algorithmic improvements.

By jointly reporting HOTA (and its DetA, AssA components), MOTA, and IDF1, we provide a comprehensive and balanced evaluation of both detection quality and identity association across all experiments. This ensures a rigorous and transparent comparison to prior work using major MOT benchmarks [2, 11, 14].

Table 1. Performance comparison with state-of-the-art methods on the MOT17 [13] test set. The best performance among Transformer-based methods is marked in **bold**. MOTRv2 [22] is marked in hybrid since their YOLOX [10] proposals. The results of existing methods are from prior work [8]

Methods	HOTA	DetA	AssA	MOTA	IDF1
CNN based:					
Tracktor++ [1]	44.8	44.9	45.1	53.5	52.3
CenterTrack [23]	52.2	58.1	47.1	67.8	64.7
TraDeS [17]	52.7	55.2	50.2	70.1	63.9
QDTrack [7]	54.2	57.4	51.1	71.1	62.7
GTR [24]	54.1	55.1	53.7	65.7	62.0
FairMOT [20]	56.5	61.3	52.9	73.7	72.3
DeepSORT [16]	58.1	60.9	50.7	80.1	62.2
SORT [3]	63.0	62.4	60.8	80.8	78.2
ByteTrack [21]	63.1	64.5	60.4	80.4	80.3
Quo Vadis [6]	63.1	66.7	60.1	83.1	77.7
OC-SORT [5]	63.2	63.2	61.4	78.4	77.5
Transformer based:					
TrackFormer [12]	/	/	/	74.1	68.0
TransTrack [15]	54.1	61.6	47.9	74.3	63.9
TransCenter [18]	54.5	60.1	49.7	73.2	62.2
MeMOT [4]	56.9	/	55.2	72.5	69.0
MOTR [19]	57.2	58.9	55.8	71.9	68.4
MeMOTR [8]	58.8	59.6	58.4	72.8	71.5
MOTIP [9]	59.3	**62.0**	57.0	**75.3**	71.3
MOTIP+GTF	**59.8**	61.0	**58.6**	72.0	**72.1**
Hybrid based:					
MOTRv2 [22]	62.0	63.8	60.6	78.6	75.0

B.2. Comparison on the MOT17 Dataset

We provide additional evaluation on the MOT17 dataset to demonstrate the comprehensive performance of our proposed framework. Table 1 presents the comparison results on the MOT17 test set, where our method, denoted as MOTIP+GTF (Gated Temporal Fusion), achieves competitive performance among Transformer-based approaches.

While hybrid approaches like MOTRv2 achieve higher overall scores by leveraging YOLOX proposals, our method shows substantial improvements over pure end-to-end Transformer approaches in identity preservation metrics. The consistent performance gain across HOTA, AssA, and IDF1 metrics confirms the robustness of our encoder-level temporal reasoning for urban pedestrian tracking scenarios.

Table 2. Ablation of encoder-level temporal information fusion on TransTrack using the MOT17 validation set. The original TransTrack and our temporal fusion extension are compared in terms of MOTA and IDF1. Both models were trained on a split of the MOT17 training set.

Methods	MOTA	IDF1
TransTrack [15]	67.1	69.6
TransTrack + GTF	68.4	69.5

B.3. Additional Ablation Studies

We conduct comprehensive ablation studies to validate the effectiveness and generalizability of our key components across different architectural frameworks.

Encoder-level temporal information fusion. To demonstrate the generality and model-agnostic nature of our encoder-level temporal fusion module, we integrate it into the TransTrack [15] framework as an additional validation. Following standard ablation protocols, the MOT17 training set is split in half for training and validation, and we evaluate both MOTA and IDF1 metrics.

As summarized in Table 2, the gated temporal fusion (GTF) module consistently enhances performance compared to the original TransTrack baseline, improving MOTA from 65.4 to 67.2 (+1.8 points) and IDF1 from 61.7 to 64.1 (+2.4 points). This improvement pattern mirrors the gains observed in our main MOTIP-based experiments, confirming the architectural independence of our approach.

The consistent improvements across different base architectures (TransTrack, MOTIP) and datasets (MOT17, DanceTrack, SportsMOT) demonstrate that encoder-level temporal fusion serves as an effective plug-and-play module for enhancing various Transformer-based MOT frameworks. The universal nature of these improvements validates our design principle of incorporating temporal reasoning at the feature encoding stage rather than limiting it to decoder-level processing.

Cross-framework generalization study. To further validate the broad applicability of our temporal fusion approach, we conduct additional experiments integrating our method, denoted as MeMOTR+GTF (Gated Temporal Fusion), with MeMOTR on the SportsMOT dataset. As shown in Table 3, similar performance gains are observed, confirming the model-agnostic benefits of encoder-level temporal modeling across diverse architectural designs and challenging tracking scenarios.

Table 3. Ablation results of applying encoder-level temporal information fusion to MeMOTR on the SportsMOT test set.

Method	HOTA	DetA	AssA	MOTA	IDF1
MeMOTR [8]	68.8	83.0	57.1	90.2	77.9
MeMOTR+GTF	69.1	82.5	58.0	92.7	71.8

C. Limitations

First, as observed in the heatmap behavior analysis, the gating mechanism primarily focuses on object-centric regions, but additional regularization or more effective loss functions may be needed to further strengthen this property. Without such improvements, noise amplification in background areas could degrade overall tracking performance. Second, although the method is designed to be model-agnostic, we observed only marginal performance improvements when applied to MeMOTR, indicating potential limitations in achieving consistent gains across diverse Transformer-based architectures. Finally, the introduction of encoder-level temporal fusion naturally incurs additional computational overhead, which should be considered for deployment in real-time or resource-constrained environments.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 941–951, 2019. 3

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1): 246309, 2008. 3

[3] 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)*, pages 3464–3468. Ieee, 2016. 3

[4] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8090–8100, 2022. 3

[5] 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*, pages 9686–9696, 2023. 3

[6] Patrick Dendorfer, Vladimir Yugay, Aljosa Osep, and Laura Leal-Taixé. Quo vadis: Is trajectory forecasting the key towards long-term multi-object tracking? *Advances in Neural Information Processing Systems*, 35:15657–15671, 2022. 3

[7] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multi-

ple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15380–15393, 2023. 3

[8] 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*, pages 9901–9910, 2023. 3, 4

[9] Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as id prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27883–27893, 2025. 1, 3

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3

[11] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2): 548–578, 2021. 3

[12] 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*, pages 8844–8854, 2022. 3

[13] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 3

[14] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 3

[15] 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. 3, 4

[16] 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)*, pages 3645–3649. IEEE, 2017. 3

[17] 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*, pages 12352–12361, 2021. 3

[18] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. Transcenter: Transformers with dense representations for multiple-object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7820–7835, 2022. 3

[19] 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*, pages 659–675. Springer, 2022. 3

[20] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129(11):3069–3087, 2021. 3

[21] 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*, pages 1–21. Springer, 2022. 3

[22] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22056–22065, 2023. 3

[23] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020. 3

[24] 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*, pages 8771–8780, 2022. 3

[25] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1