

RAM: Recover Any 3D Human Motion in-the-Wild

Supplementary Material

1. More Evaluation Details

1.1. Comparison with MOT Tracker

To further assess RAM’s tracking capability, we compare SegFollow with a wide range of state-of-the-art MOT methods on DanceTrack test set, a dataset tailored for tracking tasks. We evaluate four SegFollow variants (Tiny/Small/Base/Large) on this benchmark, each using a different SAM2 backbone to analyze the impact of model scale. Across the board, SegFollow consistently outperforms both traditional and modern supervised MOT approaches, even in a zero-shot setting.

As shown in Table 3, SegFollow-L achieves **83.7** HOTA, **90.8** IDF1, and **85.2** AssA, substantially outperforming recent supervised models like MOTIP and ColTrack. IDF1 and AssA show the most substantial improvements, underscoring SegFollow’s superior ability to maintain identity integrity and handle frame-to-frame associations, which are two core challenges in multi-object tracking, where it improves IDF1 and AssA by over **10%** compared to the leading supervised methods. Notably, even SegFollow-T, the smallest variant, exceeds all supervised baselines on HOTA and AssA, suggesting that SegFollow’s tracking strength is not solely dependent on the parameter size of the backbone.

These results highlight the strong zero-shot generalization capabilities and tracking identity stability of SegFollow, positioning it as a strong foundation for RAM’s long-term, in-the-wild 3D recovery pipeline.

Reconstruction Ablation Study on 3DPW. We conduct ablations on the 3DPW dataset to evaluate the contribution of each component in RAM. As shown in Table 4. Adding the T-HMR module substantially improves the overall performance, reducing MPJPE by **13.3** mm, which highlights the importance of temporal memory for stable reconstruction. Including the Predictor brings further gains, especially in frames with occlusion or uncertainty, by leveraging motion history. Then, the Combiner improves temporal smoothness by balancing predicted and observed cues. Without it, we simply concatenate the T-HMR and Predictor features, which lacks an adaptive mechanism to handle noisy inputs. Finally, training with the masking strategy improves the robustness of the RAM. Each component contributes meaningfully, and the full model achieves the best overall accuracy.

1.2. Parameter Selection Experiment

Effect of the Kalman Fusion Weight. We conduct ablation studies on the TrackID3x3 (Indoor) benchmark, which

Table 1. Ablation of the decay factor α in Kalman-based fusion on the TrackID3x3 (Indoor) set.

α	TI-HOTA \uparrow	TI-DetA \uparrow	TI-AssA \uparrow
0.50	72.91	61.76	86.07
0.75	75.07	62.87	89.66
0.85	73.88	62.10	87.90
1.00	72.73	61.6	85.92

features motion in the real-world, occlusion, and fast motions. This setting provides a solid foundation for evaluating fusion weight in Kalman fusion. We vary the decay factor α in our Kalman-based memory fusion to assess its effect on tracking accuracy. As shown in Table 1, setting $\alpha = 0.75$ yields the best performance, achieving the optimal balance between historical consistency and responsiveness to new observations. While lower ($\alpha = 0.5$) or higher ($\alpha = 1.0$) values lead to minor drops, the overall performance remains stable, demonstrating the robustness of our fusion scheme.

Table 2. Ablation of the gating threshold τ_γ in motion consistency fusion on the TrackID3x3 (Indoor) set.

τ_γ	TI-HOTA \uparrow	TI-DetA \uparrow	TI-AssA \uparrow
0.50	73.26	61.81	88.01
0.70	74.25	62.05	88.92
0.85	75.07	62.87	89.66
1.00	74.01	62.03	88.50

Effect of the Temporal Buffer Update Threshold. We further analyze the role of the gating threshold τ_γ , which determines the contribution of the current features based on its motion consistency score. The experiment is conducted on the TrackID3x3 (Indoor) benchmark, where frequent occlusions and complex motion make memory fusion especially critical. As shown in Table 2, setting $\tau_\gamma = 0.85$ achieves the highest performance across all metrics. This value strikes a balance between preserving reliable history and adapting to new consistent observations. Lower thresholds (e.g., $\tau_\gamma = 0.5$) result in overly conservative updates, while higher values (e.g., $\tau_\gamma = 1.0$) may introduce noisy information. Overall, the performance remains robust across a range of values, confirming the stability of our gating strategy.

Effect of the Selected frames k We evaluate the impact of the temporal selected frame size k in our T-HMR mod-

Table 3. Comparison with other popular MOT methods on DanceTrack test set.

Methods	Publication	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
<i>Supervised Methods</i>						
SORT[1]	ICIP2016	47.9	91.8	50.8	31.2	72.0
DeepSORT[9]	ICIP2017	45.6	87.8	47.9	29.7	71.0
FairMOT[12]	IJCV2021	39.7	82.2	40.8	23.8	66.7
CenterTrack[15]	ECCV2020	41.8	86.8	35.7	22.6	78.1
GTR[16]	CVPR2022	48.0	84.7	50.3	31.9	72.5
ByteTrack[13]	ECCV2022	47.3	89.5	52.5	31.4	71.6
MOTR[11]	ECCV2022	54.2	79.7	51.5	40.2	73.5
SUSHI[3]	CVPR2023	63.3	88.7	63.4	50.1	80.1
MOTRv2[14]	CVPR2022	69.9	91.9	71.7	59.0	83.0
ColTrack[6]	ICCV2023	72.6	92.1	74.0	62.3	-
FineTrack[8]	CVPR2023	52.7	89.9	59.8	38.5	72.4
OC-SORT[2]	CVPR2023	54.6	89.6	54.6	40.2	80.4
DiffMOT[7]	CVPR2024	62.3	92.8	63.0	47.2	82.5
Hybrid-SORT[10]	AAAI2024	65.7	91.8	67.4	-	-
AED[4]	TIP2025	66.6	92.2	69.7	54.3	82.0
MOTIP[5]	CVPR2025	73.7	92.7	79.4	65.9	82.6
<i>Zero-Shot</i>						
SegFollow-T	Ours	77.9	88.7	85.8	78.4	77.4
SegFollow-S	Ours	79.6	90.3	87.0	79.8	79.4
SegFollow-B	Ours	80.2	91.5	87.9	80.5	79.9
SegFollow-L	Ours	83.7	93.2	90.8	85.2	82.2

Table 4. Ablation study of each component in our framework on 3DPW. We report MPJPE and PA-MPJPE (in mm).

T-HMR	Predictor	Combiner	Masked Train	MPJPE↓	PA-MPJPE↓
✗	✗	✗	✗	81.3	54.3
✓	✗	✗	✗	68.0	44.5
✓	✓	✗	✗	59.5	37.0
✓	✓	✓	✗	56.5	35.8
✓	✓	✓	✓	53.0	34.1

074 ule on the 3DPW benchmark. We vary the number of top- k
075 selected in adjacent L frames window (set L as 30) and re-
076 port results in Table 5. The setting $k = 7$ yields the best
077 overall performance, with an MPJPE of 68.0 mm and PA-
078 MPJPE of 44.5 mm. A smaller k limits the temporal con-
079 text and thus is suboptimal, while overly large values may
080 introduce some noisy frames and degrade precision. These

results confirm that a moderate context window allows for
more reliable cross-frame fusion while maintaining tempo-
ral coherence.

Effect of the Window Size L To evaluate the effect of the
memory length L in our T-HMR module, we conduct exper-
iments on the 3DPW dataset, which features long-term real-

081
082
083

084
085
086

Table 5. Ablation of the Top-K selected frames parameter in T-HMR module on 3DPW.

k	MPJPE↓	PA-MPJPE↓
5	70.1	46.2
7	68.0	44.5
9	69.3	45.1
10	70.5	46.5

world motion. As shown in Table 6, using $L = 30$ yields the best results, achieving the lowest MPJPE and PA-MPJPE. A smaller window may not retain enough historical context for effective temporal modeling, while an overly long memory introduces noisy or redundant features, slightly degrading performance. These results confirm that moderate temporal horizons are critical for accurate human motion recovery in T-HMR module design.

Table 6. Ablation of the temporal window size L in T-HMR on the 3DPW dataset.

Window Size L	MPJPE↓	PA-MPJPE↓
20	71.5	47.2
26	69.1	46.0
30	68.0	44.5
36	68.9	45.3

Effect of Predictor Queue Length T . We evaluate the impact of the predictor queue length T on the 3DPW dataset to determine the temporal context length required for accurate motion prediction. As shown in Table 7, we vary T from 6 to 10. The results show that performance improves as T increases, reaching its best at $T = 8$ with the lowest MPJPE of 56.5 mm and PA-MPJPE of 35.8 mm. Further increasing the window size slightly degrades performance, possibly due to the inclusion of less relevant or noisy frames. These findings indicate that a moderate temporal window provides sufficient context for accurate motion refinement, while excessive history can dilute relevant information. These results demonstrate that the Predictor is robustness across window sizes, but benefits most from an appropriate queue length.

Effect of Masking Ratios During Training During stage-3 training, we apply temporal masking with a per-frame probability of 1/8 to simulate partial occlusion events, encouraging the model to rely on motion priors when visual cues are absent. We also performed ablation studies on the masking ratio applied in the training to assess its impact on temporal robustness. As shown in Table 8, a moderate

Table 7. Ablation of the predictor queue length T on the 3DPW dataset (without masking strategy).

Queue Length T	MPJPE↓	PA-MPJPE↓
6	58.2	37.0
7	57.0	36.3
8	56.5	35.8
9	56.9	36.0
10	57.3	36.1

amount of masking helps the model learn to recover missing temporal cues. Compared to the unmasked baseline, introducing 50% masking yields a clear performance gain. Increasing the masking ratio to 60% leads to the best results, achieving 53.0 mm MPJPE and 34.1 mm PA-MPJPE. However, pushing the masking further to 75% results in a slight degradation, suggesting that overly sparse temporal input may impair learning. These results indicate that masking serves as an effective regularization strategy for RAM.

Table 8. Ablation of the masking ratio during stage-3 training on the 3DPW dataset.

Masking Ratio	MPJPE↓	PA-MPJPE↓
None	56.5	35.8
50%	54.1	34.9
60%	53.0	34.1
75%	53.6	34.3

2. Training Set Up

We follow a three-stage training strategy as described in the main paper. All models are implemented in PyTorch and optimized using AdamW. For Stage 1 and 2, we use a base learning rate of 4e-4. In Stage 3, we keep the T-HMR and Predictor frozen, and fine-tune the full RAM model with simulated occlusion using masking strategy. The training is conducted on 4xA100 GPUs with a batch size of 64 unless otherwise specified. We adopt mixed-precision and distributed training for efficiency. Our training runs for up to 120 epochs.

3. Code Availability Statement

Due to current constraints, the code and models are not released at submission time. We commit to making the code and models publicly available upon publication.

References

- [1] 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. 2
- [2] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirrodar, 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. 2
- [3] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22877–22887, 2023. 2
- [4] Zimeng Fang, Chao Liang, Xue Zhou, Shuyuan Zhu, and Xi Li. Associate everything detected: Facilitating tracking-by-detection to the unknown. *IEEE Transactions on Image Processing*, 2025. 2
- [5] 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. 2
- [6] Yiheng Liu, Junta Wu, and Yi Fu. Collaborative tracking learning for frame-rate-insensitive multi-object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9964–9973, 2023. 2
- [7] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. 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*, pages 19321–19330, 2024. 2
- [8] Hao Ren, Shoudong Han, Huilin Ding, Ziwen Zhang, Hongwei Wang, and Faquan Wang. Focus on details: Online multi-object tracking with diverse fine-grained representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11289–11298, 2023. 2
- [9] Balaji Veeramani, John W Raymond, and Pritam Chanda. Deepsort: deep convolutional networks for sorting haploid maize seeds. *BMC bioinformatics*, 19:1–9, 2018. 2
- [10] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6504–6512, 2024. 2
- [11] 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. 2
- [12] 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*, 129:3069–3087, 2021. 2
- [13] 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. 2
- [14] 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. 2
- [15] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020. 2
- [16] 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. 2