STORM: End-to-End Referring Multi-Object Tracking in Videos

Supplementary Material

In the supplementary material, we provide additional visualizations of RMOT model performance and detailed statistics of our STORM-Bench dataset.

1. Qualitative Results

In this section, we present more qualitative examples comparing the referring multi-object tracking performance of our STORM with baseline models [2, 4, 5, 7] in Figures 3–7.

Figure 3 shows a setting where the model must track both a salient object (the child) and nearby small objects (the toys). Baseline models such as Qwen2.5-VL and LaMOT often miss small objects, while VisionLLMv2 over-detects them, producing imprecise grounding. These issues stem from the fact that most baseline models perform image-based grounding; thus, an object detected in one frame may be missed in subsequent frames, as seen in Qwen2.5-VL. In contrast, STORM provides stable predictions and consistently localizes both objects across all frames.

Figure 4 illustrates a case with heavy occlusion. The high chair beneath the baby is only partially visible throughout the video. Baselines either miss the chair entirely or ground only a visible sub-part. Occlusion is a common challenge for object detection and tracking, but our task-composition learning strategy allows STORM to learn robust occlusion handling from diverse image grounding and single-object tracking data. As a result, STORM predicts complete bounding boxes despite partial visibility.

Figure 5 shows an example where objects exit and re-enter the frame. Baselines incorrectly produce false-positive bounding boxes in the second frame, even though both the cup and the meat are absent. During training, we enforce that STORM outputs an empty bounding box when an object is not visible. Thus, STORM does not assume referred objects always appear in the video and can reliably detect their presence or absence.

Figure 6 presents a scenario with rapid changes in object location and scale. These variations challenge the model to maintain accurate bounding box predictions. GroundingDINO exhibits identity switches between the lady and the dog, whereas STORM tracks both objects with consistent identities and stable trajectories.

Finally, Figure 7 shows a failure case involving a partially visible table with a white cloth. The crowded scene significantly increases difficulty, leading all models to mislocalize the table. STORM misses the right portion of the table. This example highlights that heavy occlusion in dense scenes remains challenging for referring object trackers. We plan to address this limitation in future work by improving training strategies and refining data curation.

2. Statistics of the STORM-Bench Dataset

In Section 4 of the main paper, we compared the statistics of our dataset with existing referring object tracking datasets. Here, we provide additional visualizations. Figure 1 presents a word cloud of referring expressions in our dataset, showing that expressions commonly describe positional relations (e.g., "positioned", "left"), colors (e.g., "yellow", "pink"), age (e.g., "adult", "toddler"), appearance (e.g., "outfit", "dressed"), and temporal actions or movement (e.g., "resting", "holding"). Figure 2 visualizes the distribution of object categories, demonstrating that STORM-Bench covers a wide range of everyday objects, including people, animals, furniture, electronics, and food. This stands in contrast to prior MOT datasets [1, 3, 6] that primarily focus on pedestrians and vehicles. Together, these characteristics show that STORM-Bench provides diverse expressions and object types, enabling models to learn richer spatial–temporal grounding and improving generalization across varied real-world scenarios.

3. Efficiency Analysis.

We further measure inference cost as a function of video length and target count. With a fixed number of targets, decoding remains stable at around 3 FPS as the video grows from 32 to 256 frames. As the number of tracked targets increases, throughput drops roughly linearly because the model must auto-regressively emit longer output sequences.

4. Failure Analysis and Limitations.

The high AssA and low IDsw in Table 4 of the main paper indicate that STORM is already strong at maintaining cross-frame identity consistency. The remaining gap is primarily in grounding quality, as reflected by DetA: the model can still miss precise spatial alignment when prompts are highly ambiguous or when scenes are heavily crowded. We also observe that very long videos or videos with more than ten target objects become challenging because the autoregressive output sequence grows quickly.

References

[1] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object

#frames	32	64	128	256	64	64	64	64
#targets	1	1	1	1	1	2	4	8
time (s)	9.7	19.7	36.4	76.4	19.7	34.9	67.4	143.3
FPS	3.3	3.2	3.3	3.3	3.2	1.8	1.0	0.4

Table 1. Efficiency analysis of STORM with an 8B MLLM backbone. The first four columns vary video length with one target, while the last four vary the number of targets at 64 frames.

Figure 1. Word cloud of the referring expressions in the STORM-Bench dataset.

tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1

[2] Wu Jiannan, Zhong Muyan, Xing Sen, Lai Zeqiang, Liu Zhaoyang, Chen Zhe, Wang Wenhai, Zhu Xizhou, Lu Lewei, Lu Tong, Luo Ping, Qiao Yu, and Dai Jifeng. Vision-llm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*, 2024. 1

[3] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 1

[4] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 1

[5] Yunhao Li, Xiaoqiong Liu, Luke Liu, Heng Fan, and Libo Zhang. Lamot: Language-guided multi-object tracking. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6816–6822. IEEE, 2025. 1

[6] 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. 1

[7] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

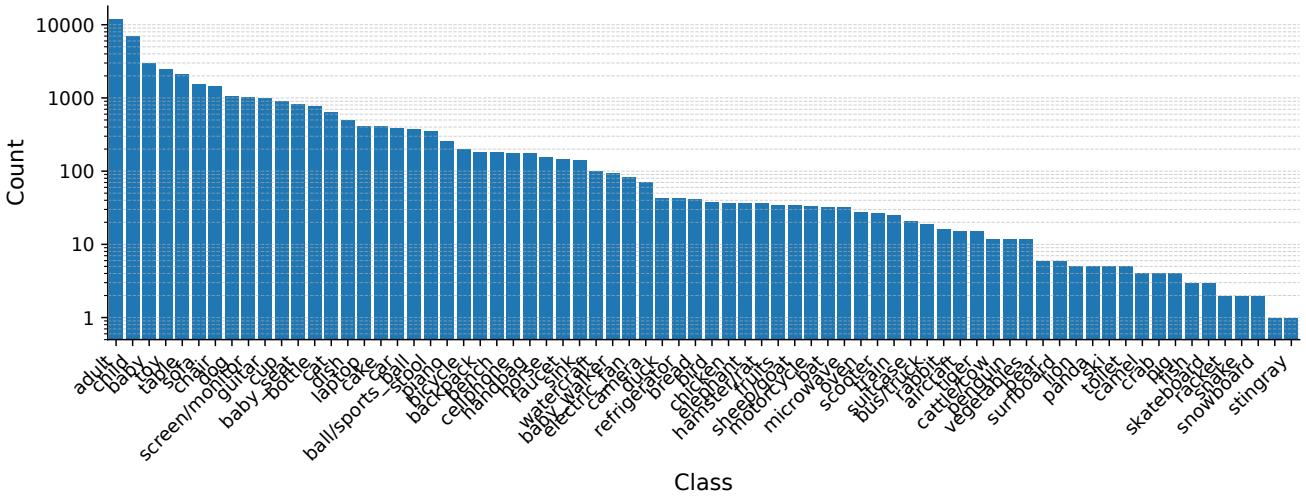

Figure 2. Distribution of object classes in STORM-Bench.

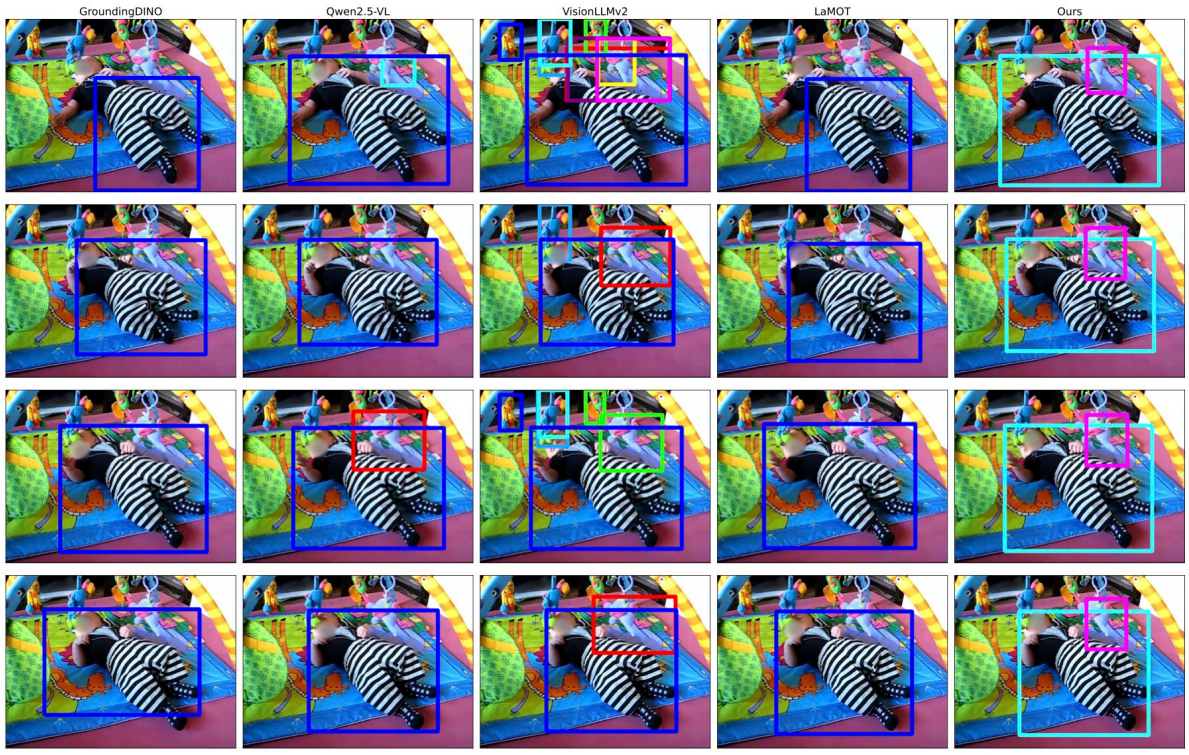

Figure 3. Model results with the referring expression **the baby and the pink toy above it**.

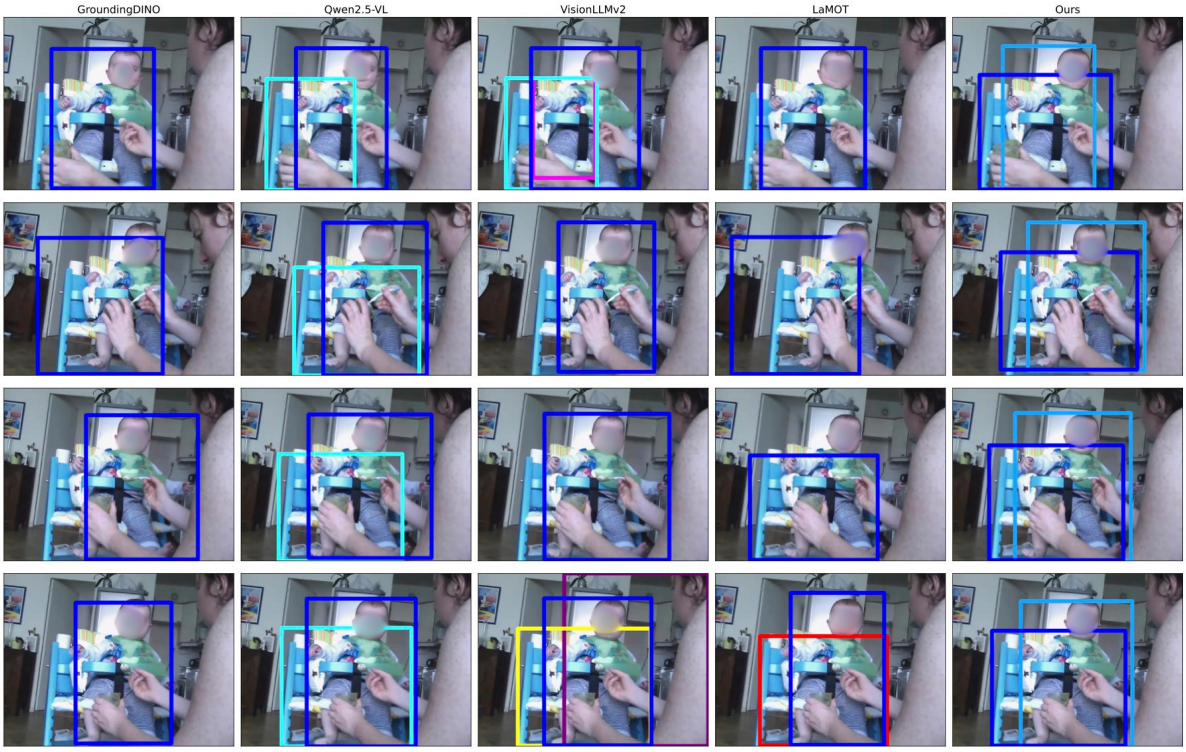

Figure 4. Model results with the referring expression **the baby and the blue high chair beneath the baby**.

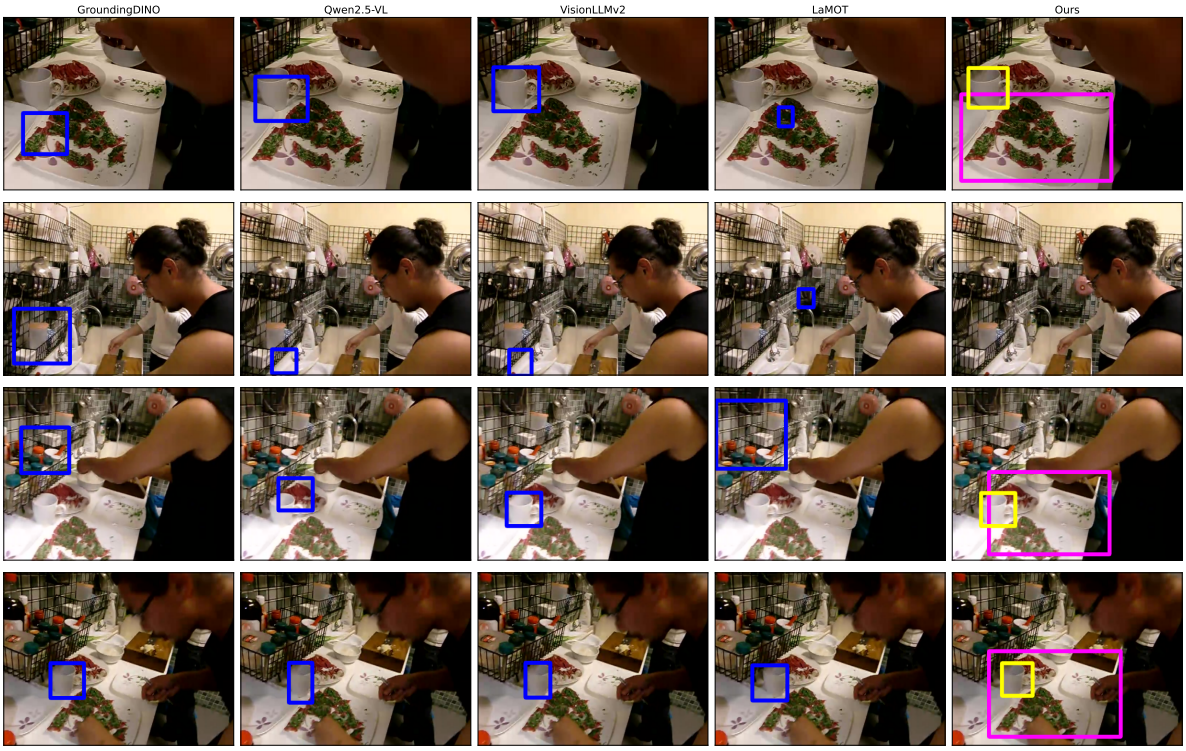

Figure 5. Model results with the referring expression **the cup next to the dish with fan-shaped meat arrangement**.

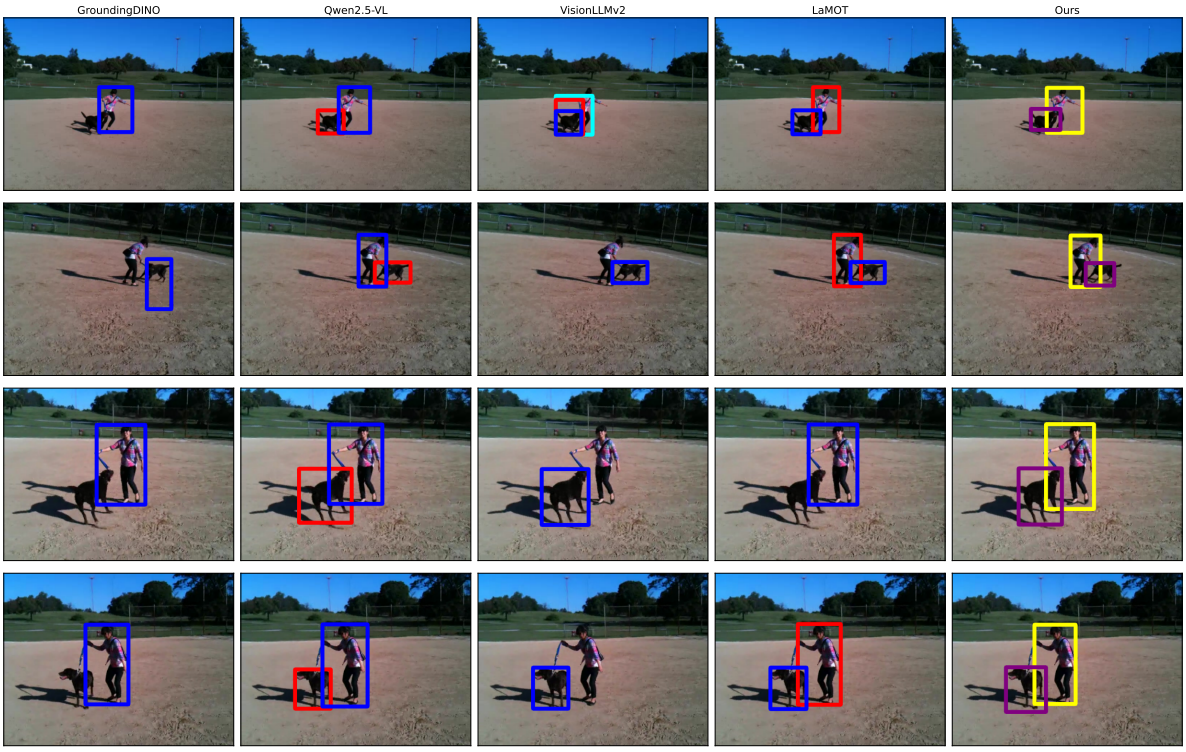

Figure 6. Model results with the referring expression **the adult wearing a pink and white patterned shirt and the black dog**.

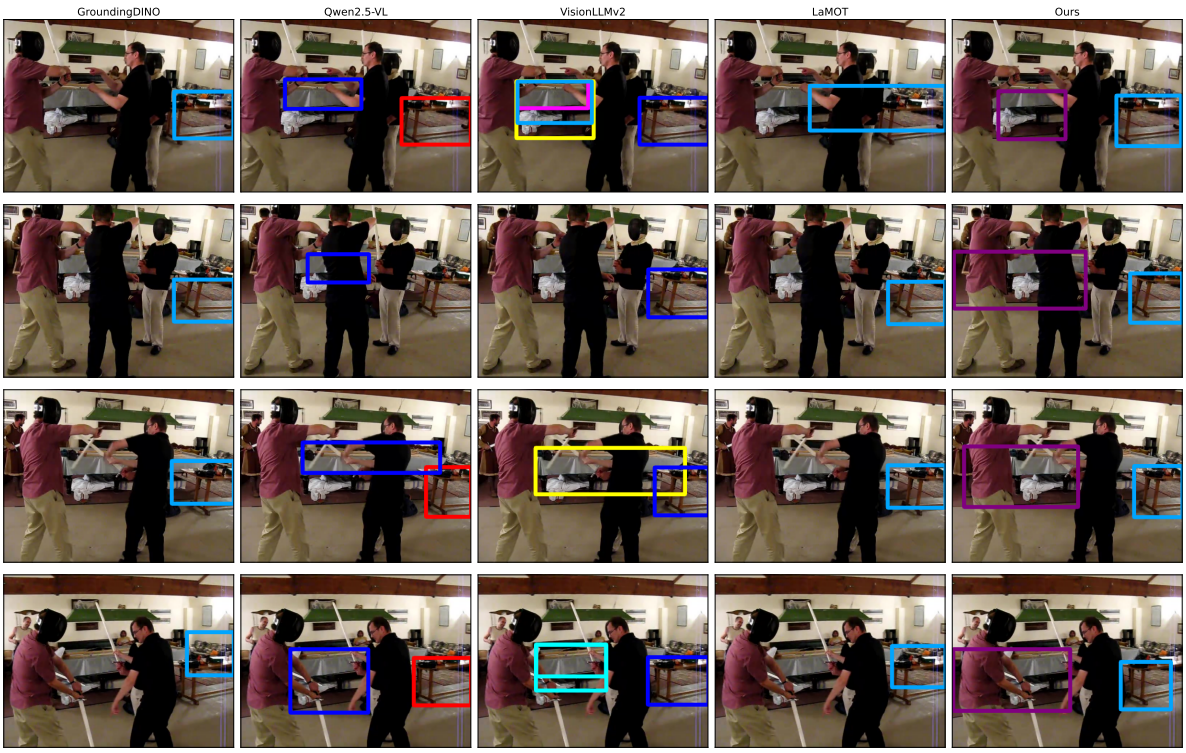

Figure 7. Model results with the referring expression **the table with white cloth and the wooden table**.