Interactive Tracking: A Human-in-the-Loop Paradigm with Memory-Augmented Adaptation

Supplementary Material

We provide additional details for the InteractTrack benchmark, including a thorough explanation of its motivation and construction protocol, the interaction dynamics that characterize the task, and scene-wise visualizations covering all six scenarios. Moreover, we present extended experimental results—encompassing precision analysis, challenging-case evaluations, and failure-case studies—that offer deeper insight into tracker behavior under complex and dynamic interaction conditions.

A. Additional Dataset Details

A.1. Motivation and Protocols

Motivation. Existing tracking benchmarks such as LaSOT [15], GOT-10k [24], and TrackingNet [46] assume a fully autonomous paradigm: the tracker is initialized once and must follow the target throughout the video without further human feedback. While this setting has enabled significant progress, it does not reflect real-world usage, where users frequently correct drift, refine the focus of tracking, or switch targets altogether.

In addition, existing benchmarks impose a fixed target throughout the entire sequence and provide no mechanism to model or evaluate such interactive behaviors. Vision–language tracking datasets (e.g., TNL2K [60]) allow natural-language initialization but remain strictly one-shot, offering no support for evolving or time-varying user intent. Multimodal datasets such as MGIT [23] introduce long videos and rich semantic annotations, yet their language descriptions serve as static global context rather than dynamic, incremental instructions.

Consequently, neither autonomous visual tracking nor language-initialized tracking captures the essential dimension of interaction. To bridge this gap, our benchmark is designed to explicitly model the missing interactive component—where users issue instructions over time, and trackers are expected to perceive, interpret, and adapt to these changing directives.

Protocol. InteractTrack is designed to simulate real-world scenarios in which human input plays an essential role in the tracking process. We intentionally select sequences that naturally create opportunities for user intervention, including severe occlusions, abrupt appearance changes, ambiguous object interactions, and long-term tracking segments that require re-detection via language cues. The videos span a diverse set of domains—sports, surveillance, UAV footage, daily activities, and wildlife—capturing a broad range of motion patterns and scene complexities.

Each video is annotated with dense, frame-level bounding boxes and timestamped language instructions marking explicit interaction points (e.g., initialization, drift correction, and target switching). To rigorously evaluate a tracker's ability to switch targets while maintaining situational awareness, we annotate both the new target and the previous target at every switch. This dual annotation provides clear ground truth for assessing whether the tracker transitions to the correct entity while appropriately rejecting distractors.

A.2. Interaction Dynamics

InteractTrack explicitly models the dynamic interaction process between the user and the tracker. Unlike traditional tracking tasks that assume a fixed target throughout the entire video, our dataset introduces multiple interaction points where human input is required. These interactions are driven by the evolving visual context and include:

- Changes in target appearance (e.g., a player switching roles or taking possession of the ball),
- Drift or loss due to occlusion (e.g., a player temporarily disappearing behind others),
- Role transitions (e.g., when a player passes the ball to a teammate),
- Shifts in user focus (e.g., switching from tracking the player to tracking the ball).

Each interaction point is accompanied by precise ground-truth annotations, with special care given to target-switching events. At every switch, we annotate both the new target and the previous target, enabling rigorous evaluation of whether the tracker transitions correctly while avoiding distractors. Between interaction points, the tracker is expected to operate autonomously, maintaining stable tracking until the next user instruction.

This protocol reflects realistic human-in-the-loop tracking scenarios and allows us to evaluate how well a tracker adapts to corrections, target switches, and evolving user guidance across multiple phases of interaction.

A.3. Scene-wise Sequence Visualizations

For each of the six scenarios in our dataset, we include three representative sequences (18 in total) in the supplementary material. Each scenario provides one sequence with detailed textual interaction, explaining the annotation context and highlighting the key qualitative aspects of the scene.

The remaining sequences are provided as visual-only examples, showcasing appearance diversity, intra-scenario variation, and overall scene complexity. Together, these visualizations offer a comprehensive overview of the environments and tracking challenges encompassed by the dataset.

Representative sequences from each scenario are shown in Figs. 5, 6, 7, 8, 9, and 10.

B. Additional Experimental Results

Table 4. Categorization of the evaluated tracking methods by task type, architectural family, and publication venue.

Type	Tracker	Architecture	Where
VLT	JointNLT	Trans.	CVPR'23
	DUTrack	Trans.	CVPR'25
	SUTrack	Trans.	AAAI'25
VOS	SAM2	Trans.	ICLR'25
	HIM2SAM	Trans.	PRCV'25
	SAMURAI	Trans.	arxiv'25
	VL-SAM2	Trans.	CVPRW'25
	DAM4SAM	Trans.	CVPR'25
VOT	ToMP	CNN-T	CVPR'22
	SeqTrack	Trans.	CVPR'23
	TaMOs	Trans.	WACV'24
	SimTrack	Trans.	ECCV'22
	CiteTracker	Trans.	ICCV'23
	STARK	CNN-T	ICCV'21
	MixFormer	Trans.	CVPR'22
	OSTrack	Trans.	ECCV'22
	DropTrack	Trans.	CVPR'23
	ARTrack	Trans.	CVPR'23
	GRM	Trans.	CVPR'23
	MixViT	Trans.	PAMI'24
	ROMTrack	Trans.	ICCV'23
	ODTrack	Trans.	AAAI'24
	MCITrack	Trans.	AAAI'25
RVOS	VideoLisa	Trans.	NeurIPS'24
	SA2VA	Trans.	arxiv'25

B.1. Evaluated Trackers

Tab. 4 summarizes the tracking methods evaluated in InteractTrack, categorized by task type (VLT, VOS, VOT, and RVOS), model architecture, and publication venue. This table highlights the breadth of contemporary tracking approaches, spanning vision–language, segmentation-based, and visual object tracking baselines. Such categorization provides a clear overview of the evaluated methods and reflects the diverse architectural designs and research directions represented in recent literature.

B.2. Challenging Case Visualizations

Fig. 11 presents visualizations of several challenging tracking cases, including severe occlusion, small-object tracking, rapid motion, and heavy background clutter. Across these difficult scenarios, our method consistently maintains accurate localization and robust target continuity compared with existing trackers.

For example, in sequences involving occlusion or temporary target disappearance—such as the first-row examples ("the child with the umbrella" or "the dice with a missing face")—IMAT reliably perceives and re-identifies the target once it reappears. Traditional trackers often struggle in these situations and fail to resume tracking after significant occlusion or off-screen movement.

These visual results highlight the strengths of our interaction-aware design. The IPM effectively interprets user-provided queries to ground the visual features, enabling the tracker to relocate the target even after it becomes temporarily lost. The MAVT maintains stability by adapting to appearance changes over time and filtering distractors, while the CAM ensures the tracker responds promptly to new instructions or shifts in user intent. Together, these components allow our model to handle both traditional tracking challenges and dynamic, real-time user interactions with greater robustness.

B.3. Scenario-Based Precision

As depicted in Fig. 12, we further provide precision plots for all six scenarios to complement the success plots presented in the main paper. The precision evaluation offers an additional perspective on localization accuracy across varying scene conditions, revealing consistent performance trends in daily activities, sports analysis, UAV tracking, surveillance, wildlife monitoring, and other scenarios. These results reinforce the robustness and adaptability of our method across diverse environments and interaction settings.

B.4. Failure Case Analysis

As illustrated in Fig. 13, we present a representative failure case that highlights the inherent challenges of interactive tracking in real-world scenarios. Although updated user instructions are issued at several key frames (e.g., "Track the red-billed leiothrix that is about to fly from the top left to the bottom right"), the target undergoes rapid motion, drastic scale variation, and partial occlusions within a highly cluttered background. These factors collectively make accurate localization extremely difficult for all trackers.

This example shows that, while textual instructions can provide additional guidance, interactive tracking remains fundamentally challenging—particularly in scenarios that require the system to fuse user feedback with rapidly changing visual information. The failure case demonstrates that there is still substantial room for improvement, especially in enhancing a tracker's adaptability to dynamic scenes, robustness to occlusion and clutter, and ability to maintain consistent contextual understanding throughout interaction-heavy sequences.

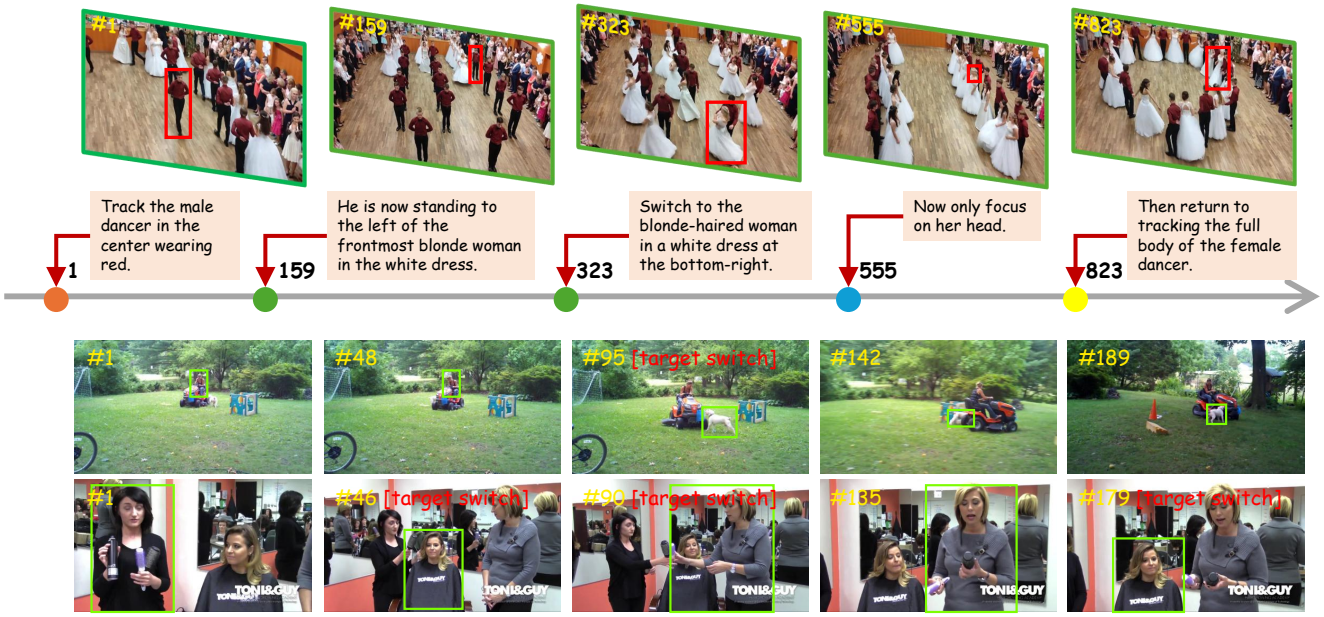

Figure 5. Representative sequences from the **daily activities** scenario. This scenario includes everyday indoor and outdoor activities captured from a third-person viewpoint. The sequences depict interactions such as playing with pets, handling household objects, and casual human motion under moderate viewpoint shifts. The highlighted sequence illustrates target-switch interaction cues.

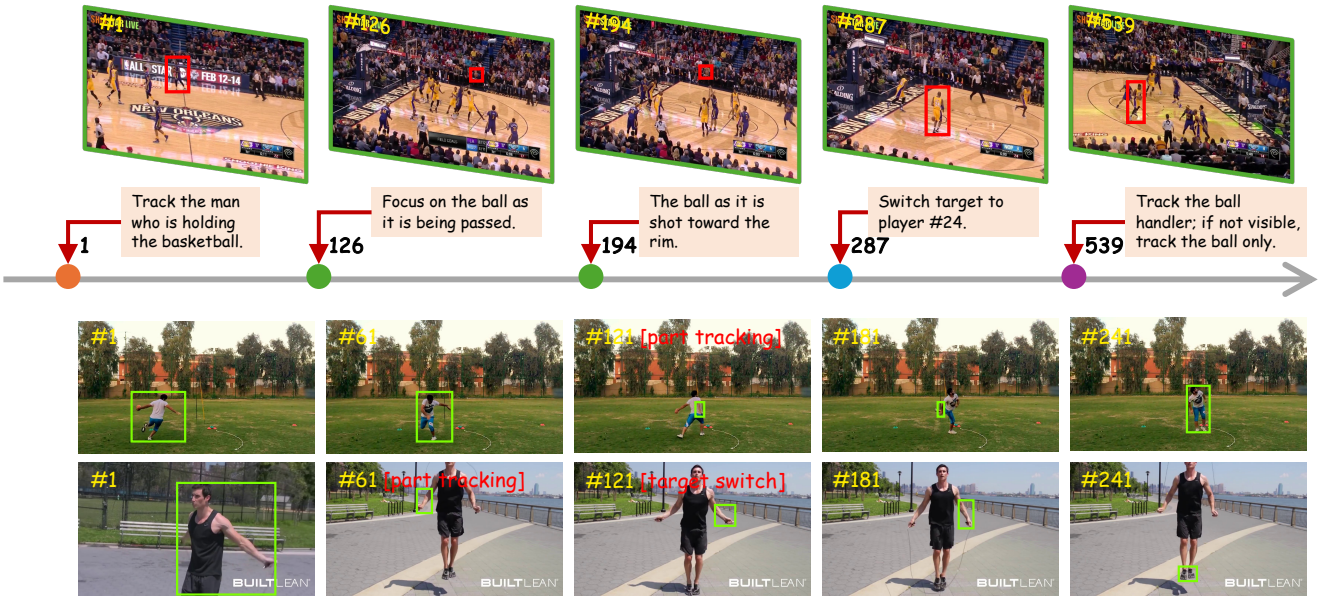

Figure 6. Representative sequences from the **sports analysis** scenario. The sequence illustrates multiple interaction points in a basketball game, including target initialization, ball-focused attention during passes, shot-following, and explicit target switching among players. At each key timestamp, the user provides updated instructions that reflect the evolving visual context—shifting attention from the ball handler to the ball itself, following the shot trajectory, and then redirecting focus to a specific player. The highlighted sequence illustrates target-switch and part-tracking interaction cues as user instructions evolve across multiple stages.

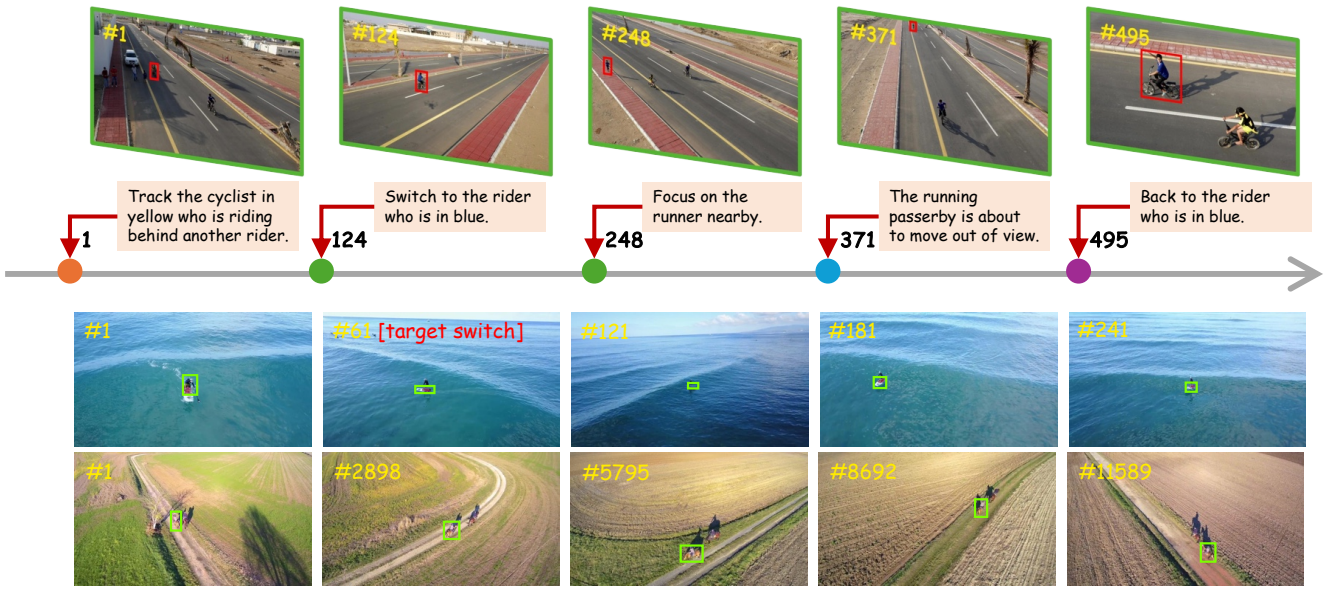

Figure 7. Representative sequences from the **UAV tracking** scenario. This scenario includes aerial-view videos characterized by large scale variation, long-range target motion, and perspective distortions typical of drone footage. The sequences feature vehicles, pedestrians, and small moving objects captured at varying altitudes and along diverse trajectories. The highlighted sequence illustrates challenges such as drastic scale changes, temporary target disappearance, and viewpoint transitions during UAV maneuvers. Other examples demonstrate additional difficulties common in aerial surveillance, including low-resolution targets, cluttered backgrounds, and complex scene geometries.

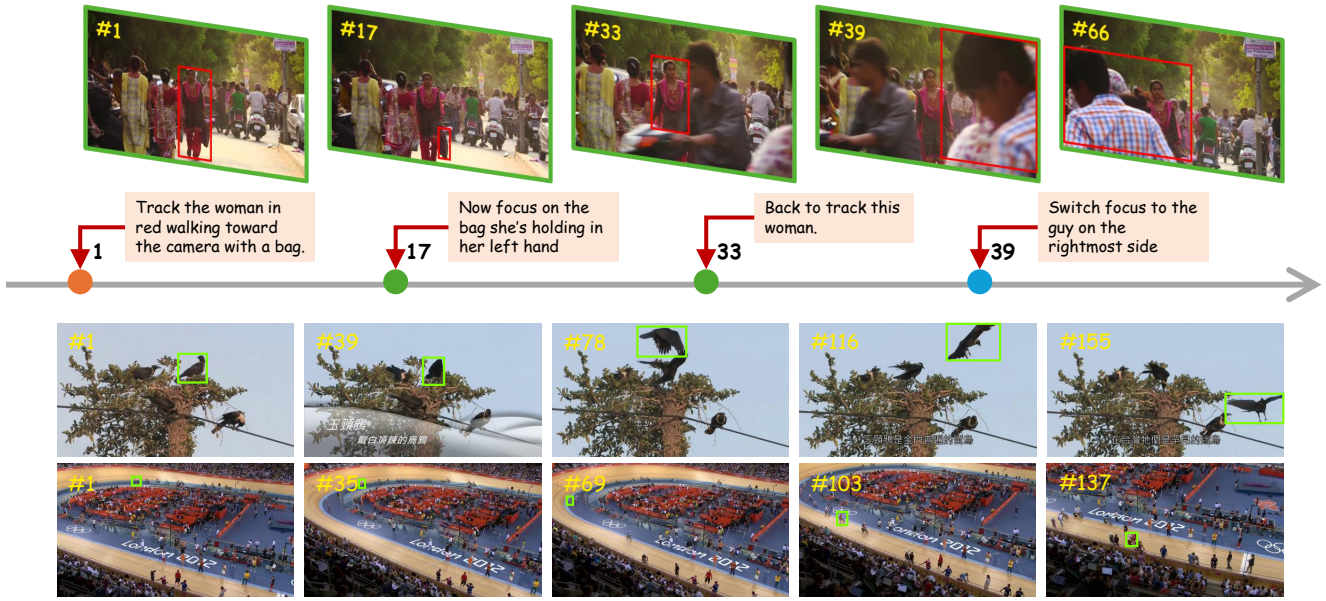

Figure 8. Representative sequences from the **surveillance** scenario. This scenario features crowded street environments captured from a third-person viewpoint. The sequences depict interactions such as monitoring individuals in dense crowds, shifting attention between people and carried objects, and handling frequent occlusions and complex motion patterns. The highlighted sequence illustrates target-switch interaction cues.

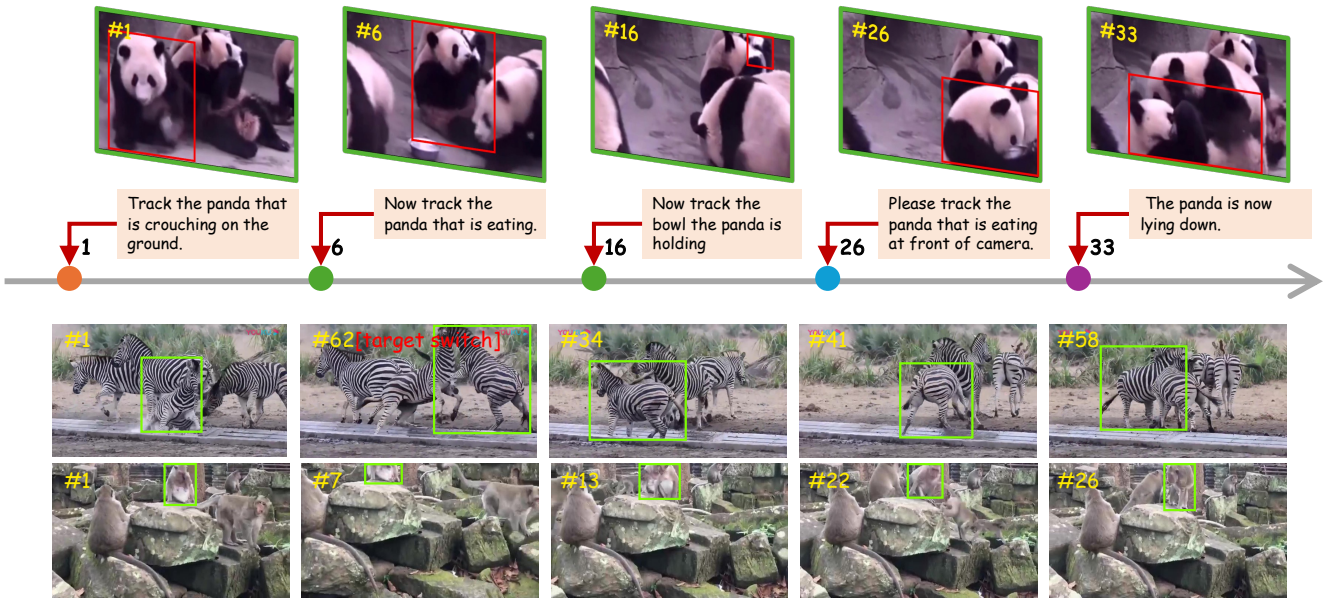

Figure 9. Representative sequences from the **wildlife monitoring** scenario. This scenario includes animal behaviors captured in natural environments. The sequences depict interactions such as observing feeding behaviors, shifting attention between different animals, and focusing on objects being manipulated by the animals.

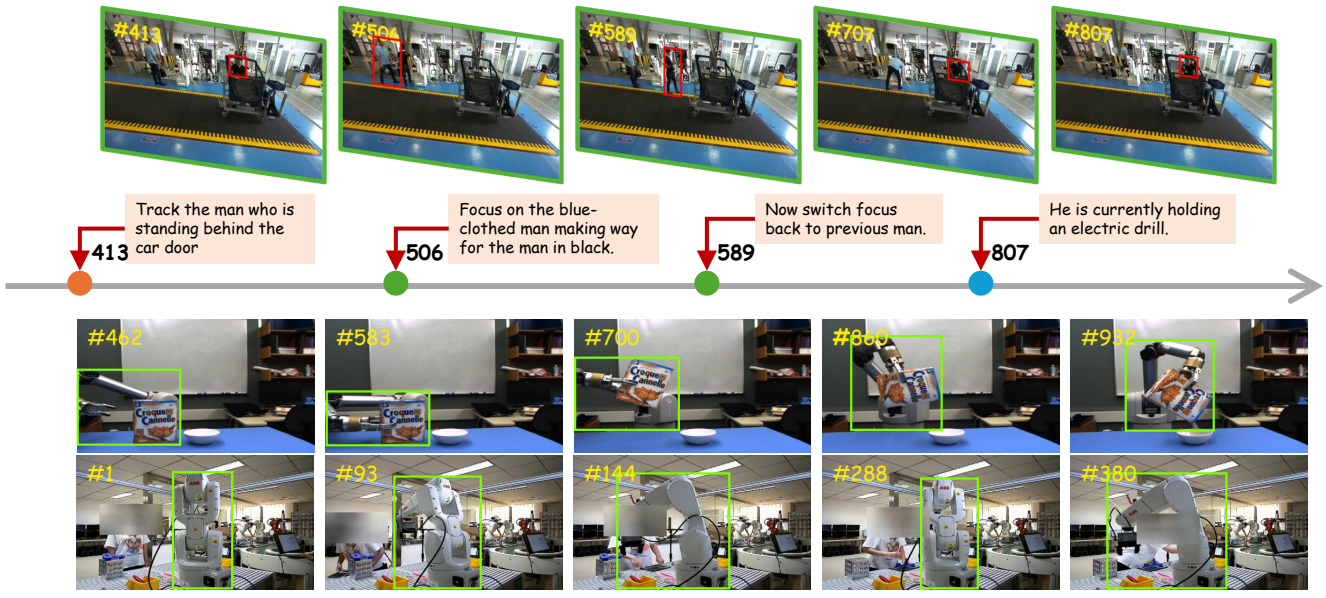

Figure 10. Representative sequences from the **other** scenario. This scenario contains diverse scenes that do not fall into the previous categories, including industrial, laboratory, and indoor environments. The sequences illustrate various interaction points where user instructions shift attention among different people, objects, and activities.

Figure 11. **Visualization of challenging cases in InteractTrack.** It demonstrates that our proposed method can effectively handle complex interactive scenarios and accurately track the target, whereas traditional trackers may lose the target.

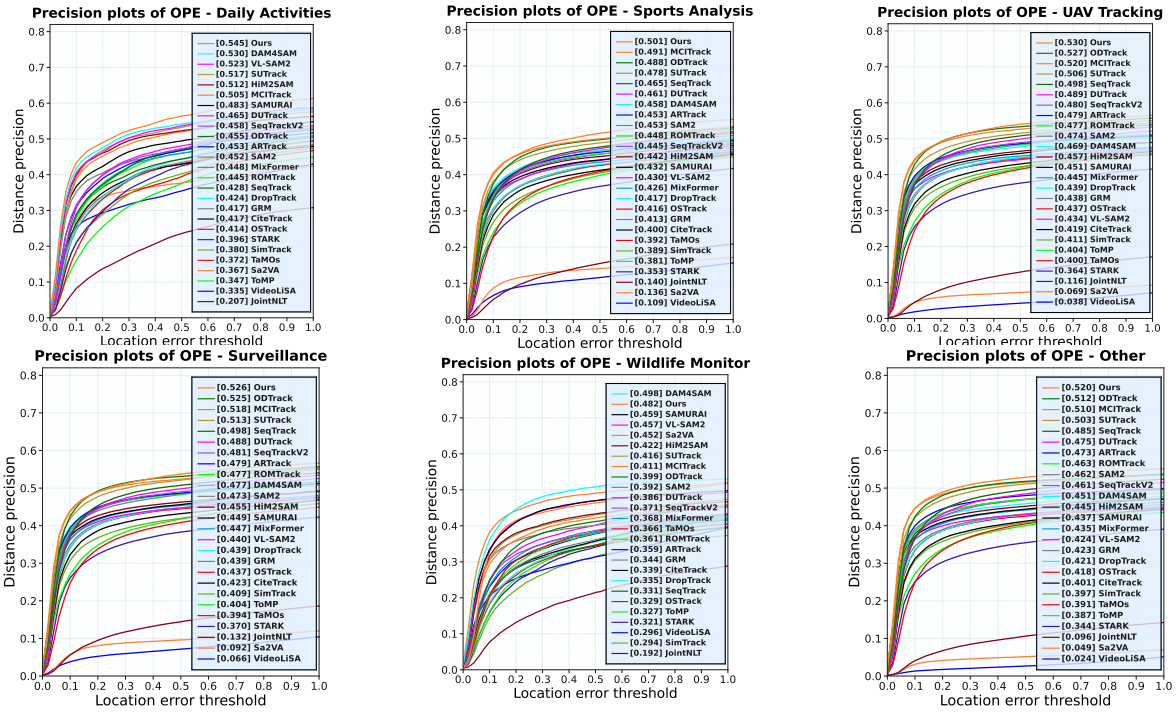

Figure 12. **Precision evaluation results across the six scenarios in the InteractTrack benchmark.** The plots show that the proposed baseline achieves strong and consistent precision performance in all scenarios.

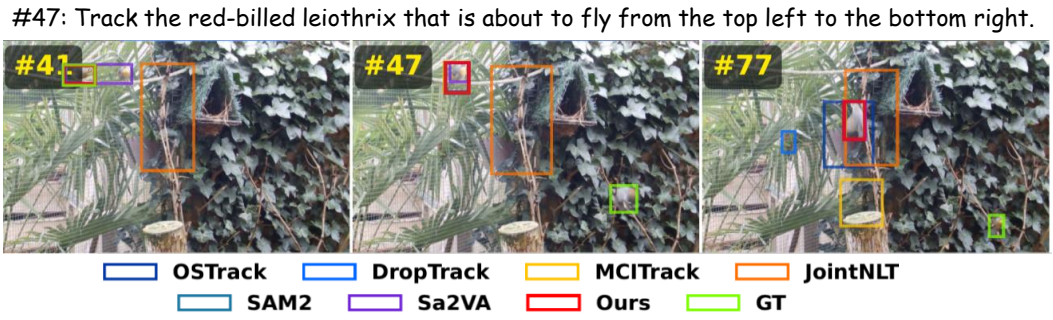

Figure 13. **Failure cases.** Even with updated user instructions, all trackers struggle with heavy occlusion and background clutter, ultimately losing the target despite interactive guidance.