QTRACK: LEARNING TO USE TEMPORAL EVIDENCE FOR REFERENCE-GROUNDED VIDEO TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Category-level multi-object trackers estimate trajectories for all detected instances, whereas some applications require propagating only targets specified by a user. We study offline, reference-grounded video tracking: given an ordered video clip, a natural-language query, and the reference-frame boxes of the specified targets, a model predicts their subsequent trajectories. We introduce **QTrack**, a vision-language policy trained with structured trajectory rewards and Temporal Perception-Aware Policy Optimization (TAPO). TAPO contrasts the likelihood of a generated response under the observed clip and a reference-frame repetition, encouraging the output to depend on visual evidence beyond the reference frame. We also introduce `RMOT26`, a controlled diagnostic benchmark constructed from existing tracking datasets. Under the reported protocol, QTrack obtains a motion-consistency score of 0.30, MOTP of 0.75, center-location error of 44.61 pixels, and normalized distance error of 0.39; its motion-consistency score ties the strongest comparison method. Component ablations suggest complementary effects from the motion reward and temporal contrast, while an exploratory four-clip, 50-frame evaluation examines transfer beyond the primary short-clip setting. The results support further study of temporal-evidence learning in trajectory-generating vision-language models. Code and data are available at https://anonymous.4open.science/r/QTrack-5548.

1 INTRODUCTION

Video tracking estimates the locations of persistent objects across frames. Category-level multi-object tracking usually detects and associates every instance of a predefined class, while single-object and referring trackers follow a selected target Bernardin & Stiefelhagen (2008); Kalake et al. (2021); Luo et al. (2021); Shim et al. (2025). We study a controlled vision-language setting between these formulations: the input provides the reference-frame boxes of one or more specified targets, and a natural-language query supplies descriptive context; the output is a separate trajectory for each grounded target. The scientific question is whether a trajectory-generating vision-language policy can use observations beyond the reference frame to propagate these identities accurately.

Consider *"a crowded clip containing similarly dressed pedestrians."* The reference boxes identify two requested people, while the query describes their clothing and relative positions. Predicting their later boxes requires the model to use changing spatial evidence, distinguish nearby distractors, and preserve the mapping between each supplied identity and its output trajectory. A model that copies the initial boxes or relies only on static appearance can produce plausible but incorrect trajectories as the people move or overlap.

We call this task *reference-grounded, query-conditioned tracking*. Target membership is supplied through the reference boxes in the present protocol; language can contribute attributes, relations, and motion descriptions that help maintain the correct mapping over time as shown in Figure 1. We therefore evaluate two questions separately: whether the policy propagates the grounded targets using temporal visual evidence, and whether the semantic content of the query improves that propagation when the boxes and video are held fixed. We do not treat this setting as language-only target discovery.

Vision-language models (`VLMs`) provide a natural foundation for this setting because they align visual content with free-form language and have demonstrated strong capabilities in multimodal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

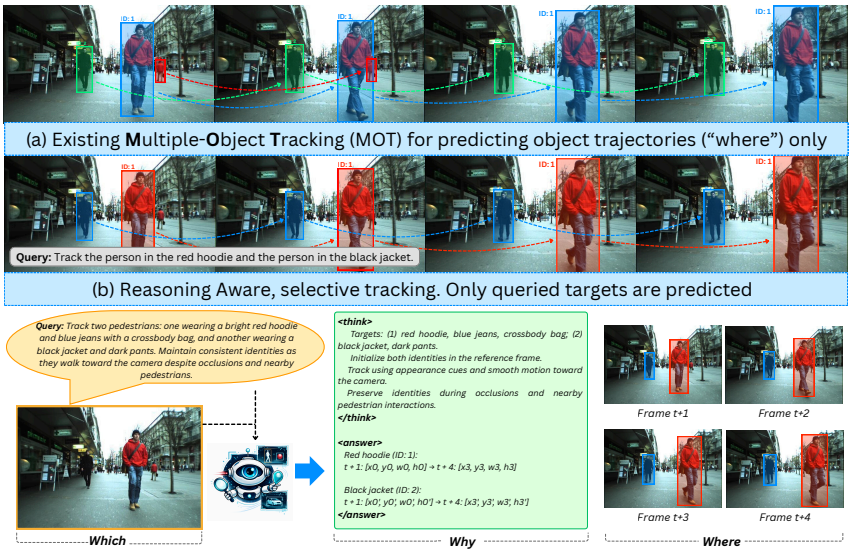


Figure 1: **Reference-grounded, query-conditioned tracking.** Category-level MOT outputs trajectories for all detected category instances. In our protocol, the input supplies reference-frame boxes for the targets specified by a query, and the model predicts one persistent trajectory per supplied target. The primary challenge is temporal propagation under distractors and visibility changes.

grounding Tong et al. (2024); Zhang et al. (2024a), visual reasoning Tan et al. (2026); Wei et al. (2026); Al-Tahan et al. (2024), and instruction following Ding et al. (2025); Zhou et al. (2023). A coordinate-generating model may ignore later frames, copy the reference box, rematch similar objects independently in every frame, or produce well-formed outputs with spatial drift. These failure modes motivate training signals that separately measure output validity, localization, identity-preserving trajectory structure, and dependence on temporal evidence.

Model: We introduce **QTrack**, an offline vision-language policy that generates a structured response containing a trajectory for each reference-grounded target. **QTrack** is initialized from VisionReasoner-3B and optimized with Group Relative Policy Optimization (GRPO) using output-format, localization, and motion rewards. We additionally introduce Temporal Perception-Aware Policy Optimization (TAPO), a temporal adaptation of corrupted-input likelihood contrast: for the same sampled output, the model compares its likelihood under the observed clip with its likelihood when all future frames are replaced by the reference frame. The corrupted input retains the reference image and initial grounding but removes subsequent changes in position, visibility, and appearance. Because this contrast measures input sensitivity rather than correct motion by itself, we evaluate trajectory quality and temporal interventions separately.

Data: We release RMOT26, a controlled diagnostic benchmark assembled from existing tracking datasets. Its primary examples are short, reference-grounded clips used to test localization, query use, and motion-sensitive propagation. Each instance contains a reference frame with explicit target grounding, a short video sequence, and a natural-language query specifying the tracking objective. RMOT26 includes single-target, multi-target, and occlusion-aware queries across diverse real-world scenarios, including crowded scenes, sports interactions, non-linear motion, and partial occlusion.

Under the current RMOT26 protocol, **QTrack** combines the strongest reported localization results with a motion-consistency score that ties the best comparison method. Matched component ablations indicate that spatial GRPO training, the motion reward, and TAPO have different effects. A fixed-box language intervention tests whether query content contributes beyond the supplied coordinates, and a trajectory-only ablation tests the cost and benefit of generating structured reasoning. We report these results with their protocols and avoid using rationale-quality correlation as evidence of causal faithfulness. In summary, our contributions are as follows.

- We define a controlled, reference-grounded tracking protocol for studying trajectory generation by a vision-language policy and release RMOT26 with predefined sequence-level splits, verified queries, and a documented evaluation protocol.

- We introduce **QTrack**, which combines structured trajectory rewards with a temporal corrupted-input contrast designed to increase dependence on observations beyond the reference frame.
- We separate the effects of task-specific training, motion reward, temporal contrast, language content, and structured reasoning through matched ablations, and evaluate the resulting model against general vision-language models and specialized tracking baselines.

2 Related Work

Tracking and language-guided grounding. Category-level multi-object tracking estimates trajectories for all instances of a predefined class, using detection, association, or joint tracking architectures (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2021; Dendorfer et al., 2021). Language-guided single-object tracking instead specifies a target with a natural language expression (Li et al., 2017; Wang et al., 2021), while referring multi-object tracking extends language guidance to multiple objects (Wu et al., 2023; Li et al., 2025c). **QTrack** studies a narrower reference-grounded setting: the target identities are supplied by initial boxes, and the model must generate their later trajectories. This protocol isolates localization and short-horizon propagation; it does not test language-only target discovery.

Reasoning-oriented tracking. ReasoningTrack combines supervised fine-tuning and `GRPO` for long-term single-object tracking with textual reasoning (Wang et al., 2025). ReaMOT addresses implicit social or causal instructions with a modular Thinking-LVLM–SAM2 pipeline and full-video association (Chen et al., 2025). These settings are complementary to **QTrack**: ReasoningTrack emphasizes long-term single-target tracking, ReaMOT emphasizes implicit target discovery, and **QTrack** evaluates direct coordinate generation after reference boxes identify the requested targets. Their published scores should not be compared directly because their inputs, temporal horizons, output protocols, and metrics differ.

Policy optimization for visual reasoning. `GRPO` and related reinforcement-learning objectives have been adapted to vision-language reasoning (Shao et al., 2024; Liu et al., 2025b; Zheng et al., 2025). Video-R1 applies reinforcement learning to video reasoning (Feng et al., 2025), and PAPO contrasts likelihoods under original and perceptually corrupted visual inputs (Wang et al., 2026). **QTrack** shares PAPO's corrupted-input likelihood-contrast principle and specializes the corruption to temporal evidence by repeating the reference frame. The intended distinction is therefore the temporal intervention and tracking.

3 Methodology

3.1 Preliminaries and Problem Setup

Let $\mathcal{F} = (F_0, F_1, \ldots, F_T)$ be an ordered video clip, let q be a natural-language query, and let

$$\mathcal{G}_0 = \{(j, b_j^0)\}_{j=1}^{N_q} \tag{1}$$

contain the reference-frame box $b_j^0 \in \mathbb{R}^4$ for every target requested by the query and N_q be number of query-relevant targets. Thus, $\mathcal{G}_0$ supplies target membership; the task is not language-only target discovery. The desired output is one trajectory per supplied identity,

$$\mathcal{Y}^* = \{\tau_j^*\}_{j=1}^{N_q}, \qquad \tau_j^* = (b_j^{*,0}, b_j^{*,1}, \ldots, b_j^{*,T}). \tag{2}$$

QTrack generates a structured response

$$o = (z, \widehat{\mathcal{Y}}), \qquad \widehat{\mathcal{Y}} = \{\widehat{\tau}_j\}_{j=1}^{\widehat{N}_q}, \qquad \pi_\theta(o \mid \mathcal{F}, q, \mathcal{G}_0), \tag{3}$$

where z is a textual rationale and $\widehat{\tau}_j$ is a predicted trajectory. Missing or extra outputs are penalized during spatial matching. The current benchmark contains positive queries with one or more supplied targets and does not evaluate an empty-target response.

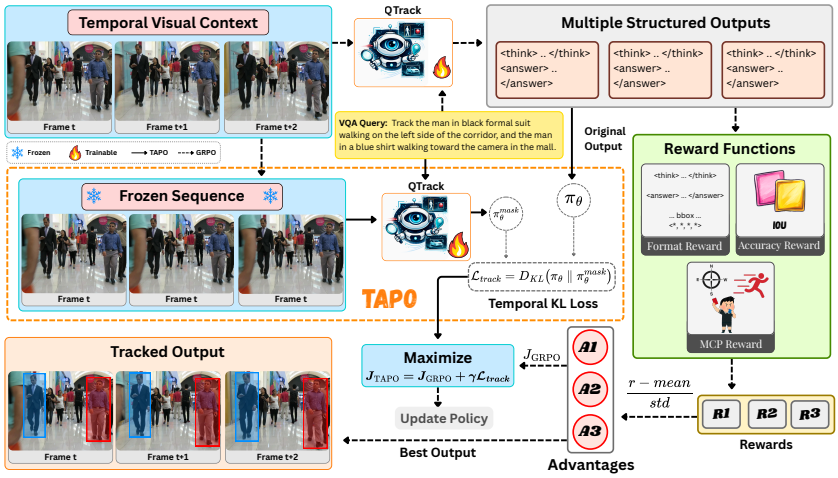


Figure 2: **Overview of the QTrack framework.** Given a video sequence $\mathcal{F} = (F_0, \ldots, F_T)$, reference grounding $\mathcal{G}_0$, and query q, **QTrack** generates a structured reasoning trace and predicts bounding-box trajectories only for the query-specified target(s). The model jointly performs query-conditioned grounding, temporal association, and trajectory prediction.

Why a vision-language policy? The policy can condition trajectory generation on free-form descriptions of attributes, relations, and motion while emitting coordinates in a single structured response. The experiments therefore test whether this formulation improves reference-grounded propagation; they do not establish open-vocabulary target discovery. We train **QTrack** using `GRPO` style optimization, which compares multiple sampled outputs for the same input and updates the policy using relative rewards. For each input x, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G rollouts: $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$. Each rollout receives a scalar reward $R_i = R(o_i, \mathcal{Y}^*)$ measuring output validity, spatial accuracy, and temporal consistency.

3.2 QTrack Framework

Given $x = (\mathcal{F}, q, \mathcal{G}_0)$, QTrack produces

$$(z, \widehat{\mathcal{Y}}) = F_\theta(\mathcal{F}, q, \mathcal{G}_0), \qquad \widehat{\mathcal{Y}} = \{\widehat{\tau}_j\}_{j=1}^{\widehat{N}_q}. \tag{4}$$

The same serialization supports single-target and multi-target examples. Because the current protocol supplies every requested identity in the reference frame, the model is evaluated on propagation, localization, and preservation of the identity-to-output mapping.

3.3 Unified Reward Mechanism

For each rollout $o = (z, \widehat{\mathcal{Y}})$ and ground truth $\mathcal{Y}^*$, we use a sequence-level reward

$$R(o, \mathcal{Y}^*) = \lambda_{\text{think}} R_{\text{think}} + \lambda_{\text{fmt}} R_{\text{fmt}} + \lambda_{\text{iou}} R_{\text{iou}} + \lambda_{\text{mcp}} R_{\text{MCP}}. \tag{5}$$

R_{think} verifies that the output contains valid `<think>` and `<answer>` fields, while R_{fmt} checks that bounding boxes are parsable in the required JSON format. These rewards stabilize structured generation but do not directly measure tracking quality.

Spatial reward. At frame t, predicted boxes $\widehat{B}_t$ and ground-truth boxes B_t^* are matched using Hungarian assignment with cost $1 - \text{IoU}$. Let $\mathcal{M}_t$ denote the matched pairs. We define

$$R_{\text{iou}} = \frac{1}{T+1} \sum_{t=0}^{T} \frac{1}{\max(\widehat{N}_q, N_q)} \sum_{(i,j) \in \mathcal{M}_t} \mathbb{I}\left[\text{IoU}(\widehat{b}_i^t, b_j^{*,t}) > \eta\right], \tag{6}$$

with $\eta = 0.5$.

Motion reward. Let $c(b)$ be the center of box b, and define predicted and ground-truth motion vectors $\Delta\widehat{c}_j^t = c(\widehat{b}_j^t) - c(\widehat{b}_j^{t-1}), \qquad \Delta c_j^{*,t} = c(b_j^{*,t}) - c(b_j^{*,t-1})$. For each target and timestep, we

compute direction alignment, speed consistency, and anti-static penalty:

$$A_j^t = \frac{\langle \Delta\widehat{c}_j^t, \Delta c_j^{*,t} \rangle}{\|\Delta\widehat{c}_j^t\|_2 \|\Delta c_j^{*,t}\|_2 + \epsilon}, \quad S_j^t = \exp\left(-\frac{(\|\Delta\widehat{c}_j^t\|_2 - \|\Delta c_j^{*,t}\|_2)^2}{2\sigma_v^2}\right), \quad P_j^t = \begin{cases} 1, & \|\Delta c_j^{*,t}\|_2 \le \epsilon_v, \\ 1 - \exp\left(-\frac{\|\Delta\widehat{c}_j^t\|_2^2}{2\sigma_s^2}\right), & \text{otherwise.} \end{cases} \tag{7}$$

The Motion Consistency Precision (MCP) reward is

$$R_{\text{MCP}} = \frac{1}{N_q T} \sum_{j=1}^{N_q} \sum_{t=1}^{T} \max(0, A_j^t) S_j^t P_j^t. \tag{8}$$

This term penalizes static boxes, incorrect motion direction, and unstable speed. The anti-static term prevents the optimizer from discovering a degenerate solution in which it predicts nearly stationary boxes for targets that are moving. During training, MCP is only one component of the reward in Eq. 5; evaluation reports MCP together with MOTP, CLE, and NDE to avoid optimizing a single metric in isolation.

3.4 Temporal Perception-Aware Policy Optimization

Although GRPO improves structured output quality, vision-language trackers can still overrely on static appearance cues. In crowded scenes, this can produce trajectories that are plausible frame-by-frame but insensitive to motion, leading to drift under occlusion or similar-looking distractors. We address this with Temporal Perception-Aware Policy Optimization (TAPO), which explicitly encourages the policy to distinguish motion-bearing videos from temporally corrupted ones.

Algorithm 1: TAPO: Temporal Perception-Aware Optimization

Input: Video $\mathcal{F}$, query q, grounding $\mathcal{G}_0$, policy π_θ, weight γ
Sample rollout $o \sim \pi_\theta(\cdot \mid q, \mathcal{F}, \mathcal{G}_0)$
Construct $\mathcal{F}^{\text{mask}} = (F_0, F_0, \ldots, F_0)$
Compute original log-probability
    $\log p = \log \pi_\theta(o \mid q, \mathcal{F}, \mathcal{G}_0)$
Compute corrupted log-probability
    $\log p^{\text{mask}} = \log \pi_\theta(o \mid q, \mathcal{F}^{\text{mask}}, \mathcal{G}_0)$
Estimate temporal log-ratio
    $\widehat{\mathcal{L}}_{\text{track}} = \log p - \log p^{\text{mask}}$
$J_{\text{TAPO}} = J_{\text{GRPO}} + \gamma \widehat{\mathcal{L}}_{\text{track}}$
Update θ using J_{TAPO}

Temporal corruption. Given a video sequence $\mathcal{F} = (F_0, F_1, \ldots, F_T)$, we construct a motion-corrupted sequence

$$\mathcal{F}^{\text{mask}} = (F_0, F_0, \ldots, F_0), \qquad F_t^{\text{mask}} = F_0, \ \forall t. \tag{9}$$

This preserves static appearance and reference-frame evidence while removing temporal motion. A motion-aware policy should therefore assign different likelihoods to trajectory rollouts under $\mathcal{F}$ and $\mathcal{F}^{\text{mask}}$.

Temporal log-ratio estimator. We define the temporal sensitivity objective as a forward KL divergence from the policy conditioned on the original video to the policy conditioned on the corrupted video:

$$\mathcal{L}_{\text{track}} = D_{\text{KL}}\left(\pi_\theta(\cdot \mid q, \mathcal{F}, \mathcal{G}_0) \,\middle\|\, \pi_\theta(\cdot \mid q, \mathcal{F}^{\text{mask}}, \mathcal{G}_0)\right). \tag{10}$$

In practice, we estimate this term using sampled rollouts. For a rollout o generated from the original-video policy, the single-sample Monte Carlo estimator is

$$\widehat{\mathcal{L}}_{\text{track}} = \log \pi_\theta(o \mid q, \mathcal{F}, \mathcal{G}_0) - \log \pi_\theta(o \mid q, \mathcal{F}^{\text{mask}}, \mathcal{G}_0). \tag{11}$$

Thus, Algorithm 1 implements a sampled log-ratio estimate of the KL in Eq. 10, rather than a separate full-distribution KL computation. Maximizing this term discourages motion-invariant behavior: a policy should not assign the same likelihood to a trajectory when all future frames have been frozen.

Overall objective. Since GRPO is written as a maximization objective, the final objective is

$$J_{\text{TAPO}}(\theta) = J_{\text{GRPO}}(\theta) + \gamma \mathbb{E}_x\left[\widehat{\mathcal{L}}_{\text{track}}\right], \tag{12}$$

where γ controls the strength of temporal regularization. In loss-minimization form, this is equivalent to $\mathcal{L}_{\text{TAPO}} = -J_{\text{GRPO}} - \gamma \widehat{\mathcal{L}}_{\text{track}}$.

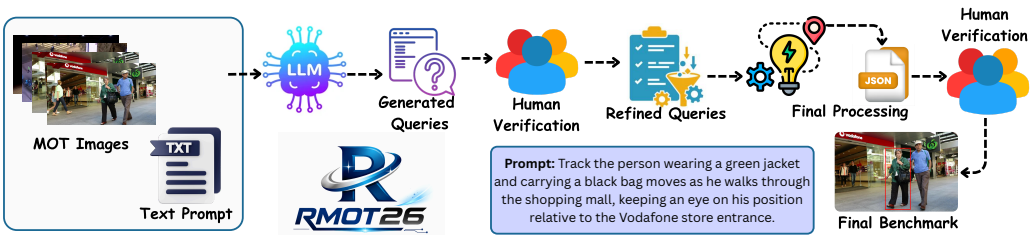


Figure 3: **RMOT26 construction pipeline.** We sample short video windows from existing tracking datasets, select query-relevant targets using reference-frame annotations, generate grounded natural-language queries, verify target identities and trajectories, and convert each instance into a structured query-conditioned tracking sample.

Relation to perception-aware regularization. Unlike generic corrupted-input consistency and PAPO Wang et al. (2026), TAPO removes temporal evidence while preserving appearance. Its log-ratio regularizer is applied to trajectory rollouts within GRPO, explicitly penalizing motion-invariant policies for identity-preserving video tracking.

4 RMOT26 Benchmark

Dataset design. RMOT26 is a diagnostic benchmark for query-conditioned spatiotemporal grounding in videos. Each instance is defined as $\mathcal{I} = (\mathcal{F}, \mathcal{G}_0, q, \mathcal{Y}^*)$, where $\mathcal{F} = (F_0, F_1, \ldots, F_T)$ is a short video clip, $\mathcal{G}_0$ denotes explicit reference-frame grounding, q is a natural-language query, and $\mathcal{Y}^* = \{\tau_j^*\}_{j=1}^{N_q}$ is the set of ground-truth trajectories for the query-relevant target(s). Each trajectory is $\tau_j^* = (b_j^{*,0}, b_j^{*,1}, \ldots, b_j^{*,T})$, where $b_j^{*,t}$ is the bounding box of target j in frame F_t. We curate clips from seven tracking datasets: DanceTrack Sun et al. (2022), MOT16/17 Milan et al. (2016); Leal-Taixé et al. (2015), MOT20 Dendorfer et al. (2020), SportsMOT Cui et al. (2023), DAMUNT Abeysinghe et al. (2023) and PETS Ferryman & Shahrokni (2009). These sources cover diverse motion regimes, including non-linear pedestrian motion, fast sports interactions, crowded surveillance scenes, heavy overlaps, and static-camera settings. This diversity allows RMOT26 to test whether models can move beyond frame-level grounding toward motion-aware query-conditioned tracking. Benchmarks such as LaSOT (Fan et al., 2019), TNL2K (Wang et al., 2021), and TNLLT (Wang et al., 2025) focus on single-target language-guided tracking, whereas RMOT26 supports single- or multi-target queries with identity-consistent trajectories, penalizes missing and extra trajectories, jointly evaluates spatial localization and motion consistency, and uses explicit reference grounding to isolate temporal propagation, making the settings complementary: LaSOT/TNL2K-style benchmarks emphasize long-horizon single-object tracking, while RMOT26 targets controlled, query-conditioned multi-target trajectory generation. The claims supported by RMOT26 are: (1) propagation of one or more targets supplied by reference-frame boxes; (2) localization over the benchmark's controlled horizon; (3) short-horizon preservation of the supplied identity-to-output mapping under crowding, interaction, and partial occlusion; and (4) sensitivity to temporal evidence under the reported intervention. RMOT26 does not evaluate language-only target discovery, online tracking, or long-term re-identification.

Query categories. Each query is grounded in the reference frame and maps to one or more target trajectories. We define three evaluation categories: *Single-object queries* specify one target in the reference frame and require predicting its trajectory across future frames. *Multi-object queries* specify multiple targets and require producing separate identity-consistent trajectories for each queried object. *Occlusion-aware queries* involve targets that are partially occluded or visually ambiguous, requiring identity preservation under visibility changes. Because every requested identity is spatially anchored by a reference-frame box, errors primarily reflect future localization, temporal propagation, output parsing, or preservation of the identity-to-output mapping.

Dataset scale. The test split contains 900 single-object and 100 multi-object instances. Although the multi-object subset is smaller, it is substantially more annotation-intensive: each instance requires separate reference grounding, target-specific language descriptions, and consistent identity labels for every queried object. Each multi-object sequence contains 2–5 tracked targets, yielding more than 300 annotated trajectory-query pairs in the multi-target subset.

Table 1: **Main results on the RMOT26 benchmark.** Higher is better for MCP and MOTP, while lower is better for CLE and NDE.

Model Name	Params	MCP↑	MOTP↑	CLE (px)↓	NDE↓
Open-Source Models					
Qwen3-VL-Instruct Yang et al. (2025)	8B	0.25	0.64	96.0	0.97
Gemma 3 Team et al. (2025)	27B	0.24	0.56	<u>58.4</u>	0.88
Gemma 3 Team et al. (2025)	12B	0.18	<u>0.73</u>	172.9	0.95
Llama 3.2 Vision-Instruct Grattafiori et al. (2024)	11B	0.19	0.15	552.1	2.67
DeepSeek Wu et al. (2024)	16B	0.11	0.27	989.13	4.7
Mistral-3-Instruct Liu et al. (2026)	8B	0.15	0.54	225.3	0.98
Molmo 2 Clark et al. (2026)	8B	<u>0.28</u>	0.54	271.87	1.05
Ovis 2.5 Lu et al. (2025)	9B	0.10	0.08	986.7	4.74
Reasoning Models					
VisionReasoner Liu et al. (2025a)	7B	0.23	0.24	428.9	2.24
VisionReasoner Liu et al. (2025a)	3B	0.21	0.44	416.58	2.32
Migician Li et al. (2025b)	7B	0.25	0.22	658.39	3.28
InternVL Chen et al. (2024)	8B	0.21	0.66	117.44	0.64
Closed-Source Models					
gpt-4o-mini Achiam et al. (2023)	-	0.20	0.57	130.48	0.67
gpt-5.2 Singh et al. (2025)	-	0.25	0.61	94.2	<u>0.55</u>
Other Models					
TempRMOT Zhang et al. (2024b)	-	**0.30**	0.19	510	3.02
DKGTrack Li et al. (2025a)	-	0.28	0.16	580	3.35
QTrack (Ours)	3B	**0.30**	**0.75**	**44.61**	**0.39**

Annotation protocol. Trajectory supervision is derived from verified tracking annotations. For each selected target, we extract its bounding boxes across the reference and future frames to form an ordered trajectory. Single-object samples are represented as one trajectory array, while multi-object samples use a structured identity-to-trajectory mapping: $\mathcal{Y}^* = \{(id_j, \tau_j^*)\}_{j=1}^{N_q}$. Each sample contains: (1) a natural-language description of the target(s), (2) explicit reference-frame bounding boxes, and (3) a query prompt specifying spatial, temporal, or relational constraints. To improve linguistic diversity, query templates are generated with Qwen-2.5-VL-Instruct Team (2025), conditioned on scene context, object attributes, and motion cues. All generated queries are verified against the reference grounding and original identity annotations, ensuring that the benchmark remains visually grounded and does not rely on external commonsense assumptions. Additional details are provided in Appendix §B.

5 Experiments

We evaluate **QTrack** along four axes: spatial localization, short-horizon motion consistency, contributions of the reported training components, and rationale-trajectory consistency. All experiments use RMOT26 unless otherwise stated.

Evaluation metrics. All stored boxes use the $[x, y, w, h]$ convention. Center-based metrics first convert a box to $c(b) = (x + w/2, y + h/2)$. We report MOTP, center-location error (CLE), normalized displacement error (NDE), and Motion Consistency Precision (MCP).

Implementation details. All experiments are run on NVIDIA H100 GPUs. **QTrack** is initialized from the 3B VisionReasoner backbone Liu et al. (2025a) and trained with `GRPO` using $G = 8$ rollouts per input. Each output is serialized as a structured response containing a reasoning field and a JSON trajectory field: `<think>...</think><answer>[id, frame, bbox_2d]</answer>`. Invalid or unparsable outputs receive zero format reward and zero tracking reward. For multi-target queries, predictions and ground truth are matched using Hungarian assignment; missing and extra trajectories are penalized through the $\max(\widehat{N}_q, N_q)$ normalization in Eq. 6. Closed-source and open-source baselines receive identical prompts and one decoded output per sample. Additional prompts, hyperparameters, and parsing rules are provided in Appendix §C.

5.1 Main Results on RMOT26

Table 1 compares **QTrack** with open-source, closed-source, and reasoning-oriented vision-language models. **QTrack** attains an MCP score of 0.30, tying TempRMOT for the highest reported MCP.

Among the evaluated general-purpose vision-language models, the strongest competing MCP is 0.28. **QTrack** additionally obtains the highest MOTP and lower center-localization and normalized-distance errors in this evaluation. These gains show that temporal consistency is not obtained by scale alone: for example, Gemma-3 27B reaches MCP 0.24, whereas Qwen-3-VL 8B reaches MCP 0.25, indicating that motion-aware supervision and temporal regularization are critical. For completeness, we also evaluate **QTrack** with standard tracking metrics such as `IDF1` and `HOTA` on query-conditioned subsets of MOT17 Milan et al. (2016) and DanceTrack Sun et al. (2022) in Appendix Table 7, where it remains competitive, suggesting that selective language-guided tracking preserves trajectory quality. Compared with long-horizon language-guided trackers such as ReasoningTrack Wang et al. (2025), `RMOT26` targets a different regime: short-horizon query-conditioned tracking with explicit target grounding and motion consistency. ReaMOT Chen et al. (2025) is also complementary, using a training-free Thinking-LVLM+SAM2 pipeline for implicit full-video reasoning, whereas **QTrack** is an end-to-end RL-optimized model trained with structured rewards for explicit reference-grounded tracking.

5.2 Ablation Studies

Component ablation. Fig. 4(a) shows that the components have complementary effects rather than monotonic gains on every metric. GRPO improves localization substantially, increasing MOTP from 0.44 to 0.65, but MCP remains low at 0.22, indicating limited motion consistency. Adding the MCP reward increases MCP from 0.22 to 0.25, while MOTP decreases from 0.65 to 0.61 and NDE increases from 0.96 to 1.06. This indicates that motion-oriented reward shaping introduces a trade-off with frame-level localization accuracy when used in isolation.

Longer-sequence evaluation. We additionally evaluated **QTrack** on four 50-frame DanceTrack sequences (Table 2) featuring crowding, similar-looking pedestrians, and occlusion. **QTrack** reduces identity switches from 29 to 21 versus VisionReasoner and achieves the best HOTA, MOTA, MOTP, IDF1, and MCP. All frames were provided simultaneously, following the clip-level protocol in Table 1, so the evaluation is offline/non-causal. As the sequences were not curated for prolonged disappearance and re-entry, we view these results as evidence of longer-horizon robustness rather than a complete evaluation of online re-identification.

Table 2: Comparison of QTrack on longer sequences.

Model	HOTA↑	MOTA↑	MOTP↑	IDF1↑	IDSW↓	MCP↑
Qwen2.5-VL-7B	0.281	0.166	0.708	0.331	**18**	0.334
VisionReasoner-7B	0.272	0.168	0.722	0.318	29	0.337
QTrack-3B	**0.304**	**0.190**	**0.751**	**0.352**	<u>21</u>	**0.371**

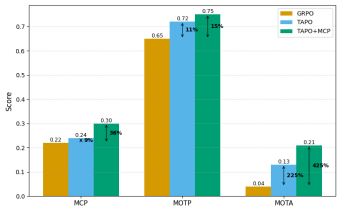
GRPO vs. TAPO. Fig. 4(b) isolates the effect of `TAPO`. GRPO encourages higher-reward structured outputs, but the resulting trajectories can remain motion-invariant, especially in crowded scenes with similar-looking objects. `TAPO` addresses this failure mode by comparing model behavior under original and temporally corrupted clips, encouraging predictions that depend on true motion cues. This yields smoother trajectories, lower drift, and higher MCP.

Comparison with traditional MOT methods. We conduct a restricted comparison against traditional `MOT` methods on query-specific clip segments from MOT17 Milan et al. (2016) and DanceTrack Sun et al. (2022). Results are reported in Appendix Table 7, with protocol details in Appendix §D. This comparison is not intended as a full-video benchmark race: all methods are evaluated on the same short query-specific clips, while **QTrack** is constrained to output only query-specified targets. Under this controlled protocol, **QTrack** achieves competitive or superior `HOTA` and `MOTP`, suggesting that selective language-conditioned tracking does not sacrifice trajectory quality. Full-video evaluation requires a different inference protocol and is left for future work.

Reasoning Quality Analysis. We score 200 stratified cases to assess whether **QTrack**'s generated reasoning aligns with its trajectories. Each trace is rated on object identification, temporal reasoning, causal justification, and prediction alignment, producing a normalized Reasoning Quality Score (RQS). RQS is positively but modestly correlated with IoU ($r = +0.158$, $p = 0.025$), suggesting that better reasoning traces are associated with better tracking but do not by themselves establish

Model (3B)	MCP↑	MOTP↑	NDE↓
VisionReasoner	0.21	0.44	2.32
GRPO	0.22	0.65	0.96
MCP-Reward	0.25	0.61	1.06
TAPO	0.24	0.72	0.82
QTrack (Ours)	**0.30**	**0.75**	**0.39**

(a) Component-wise



(b) GRPO vs TAPO comparison.

Figure 4: **Ablation analysis.** GRPO, MCP, and TAPO provide complementary gains, with the full combination performing best.

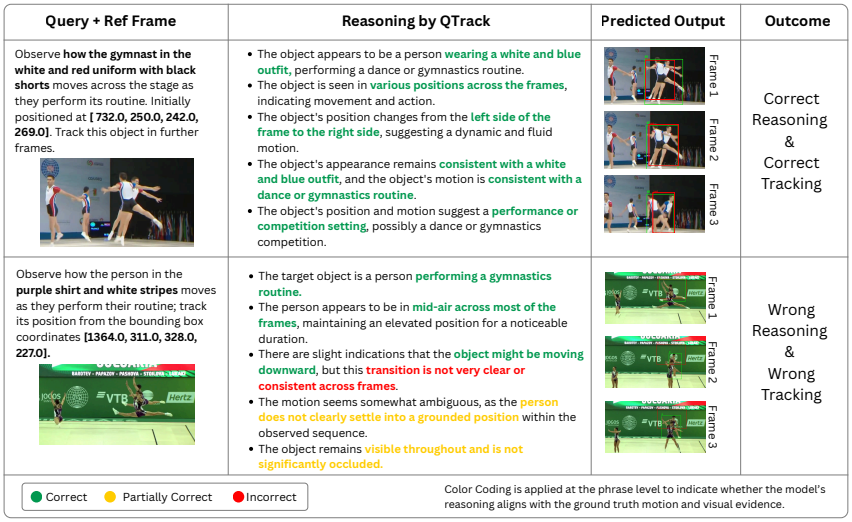


Figure 5: **Reasoning trace analysis.** Phrase-level annotations show whether the generated reasoning is consistent with visual evidence and the predicted trajectory.

causality. Faithfulness decreases from 0.661 in easy cases to 0.457 in failure cases, indicating that low-quality predictions more often contain unsupported motion claims. The dominant failure mode is weak causal justification (49%), followed by directional hallucination and motion failure (18% each). Full scoring protocol and qualitative examples are provided in Appendix §I.

6 Conclusion

We presented **QTrack**, a vision–language framework for reference-grounded, query-conditioned video tracking. We introduced RMOT26 as a diagnostic benchmark for evaluating short-horizon target propagation, localization, identity preservation, and motion consistency. We also proposed Temporal Perception-Aware Policy Optimization (TAPO), which uses temporally corrupted inputs and a sampled likelihood-ratio objective to encourage predictions that depend on evidence beyond the reference frame. Under the reported protocol, **QTrack** attains an MCP score of 0.30, and obtains the highest MOTP and the lowest CLE and NDE in this evaluation. The component ablations indicate that the motion reward and temporal contrast provide complementary benefits, although the motion reward alone introduces a trade-off between motion consistency and frame-level localization. These results support further investigation of temporal-evidence learning in trajectory-generating vision–language models.

Limitations. The primary RMOT26 evaluation uses short clips and therefore does not measure long-term re-identification, re-entry after prolonged absence, or full-video identity drift. Future work should consider longer videos, source-video-disjoint splits, broader object categories, and multilingual queries.

ETHICS STATEMENT

RMOT26 is constructed from seven existing multi-object tracking benchmarks that were released for academic research. We do not collect new video recordings or introduce new personally identifying annotations. Nevertheless, some source videos depict people in public or semi-public environments, and tracking technology can create privacy risks or be misused for surveillance, profiling, or monitoring without consent. QTrack and RMOT26 are intended solely for research on reference-grounded visual reasoning and short-horizon object tracking; they are not designed or validated for identifying individuals, inferring sensitive attributes, or making high-stakes decisions. The benchmark may also inherit geographic, demographic, environmental, and behavioral biases from its source datasets, while generated language descriptions may contain appearance-related biases despite human review. These limitations should be considered when interpreting results or transferring models to new domains. Redistribution of source material will follow the applicable licenses and terms of the original datasets, and valid removal requests from the original dataset maintainers will be honored.

REPRODUCIBILITY STATEMENT

We provide an anonymized repository containing the QTrack implementation, RMOT26 benchmark resources, evaluation code, and instructions for reproducing the reported experiments. The paper specifies the model initialization, GRPO and TAPO objectives, reward functions, training hyperparameters, hardware configuration, prompt and output formats, evaluation metrics, and baseline protocols. RMOT26 includes a data card, source-dataset composition, sequence-level train–test splits, and dataset statistics. The appendix additionally provides metric definitions, implementation details, prompt templates, traditional-tracker evaluation procedures, qualitative results, and failure analyses. The reported experiments were conducted on a single NVIDIA H100 GPU with 100 GB of memory. Because training uses the reference frame and one future frame whereas evaluation uses sequences of five to six frames, we explicitly document this train–test sequence-length mismatch. The code and data are available at `https://anonymous.4open.science/r/QTrack-5548`.

AI USE STATEMENT

Generative AI tools were used in both the research workflow and the preparation of this manuscript. As part of the RMOT26 construction pipeline, Qwen2.5-VL-Instruct was used to generate initial natural-language queries, reasoning steps, and candidate outputs from the underlying tracking examples. These generated materials were subsequently reviewed, corrected, and refined by human annotators before inclusion in the benchmark. Large language models were also used to assist with language editing, organization, and presentation of the manuscript. AI-generated suggestions were not accepted without author review. The authors verified the dataset annotations, experimental procedures, numerical results, citations, claims, and conclusions and take full responsibility for the final submission.

REFERENCES

Chamath Abeysinghe, Chris Reid, Hamid Rezatofighi, and Bernd Meyer. Tracking different ant species: an unsupervised domain adaptation framework and a dataset for multi-object tracking. *arXiv preprint arXiv:2301.10559*, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. *Advances in Neural Information Processing Systems*, 37:82411–82437, 2024.

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.

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)*, pp. 3464–3468. Ieee, 2016.

Sijia Chen, Yanqiu Yu, En Yu, and Wenbing Tao. Reamot: a benchmark and framework for reasoning-based multi-object tracking. *arXiv preprint arXiv:2505.20381*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9887–9897. IEEE, 2023.

Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking: P. dendorfer et al. *International Journal of Computer Vision*, 129(4):845–881, 2021.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1099–1109. IEEE, 2025.

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 tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5369–5378. IEEE, 2019.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. URL https://arxiv.org/abs/2503.21776.

James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*, pp. 1–6. IEEE, 2009.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lesole Kalake, Wanggen Wan, and Li Hou. Analysis based on recent deep learning approaches applied in real-time multi-object tracking: a review. *IEEE Access*, 9:32650–32671, 2021.

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.

Guangyao Li, Siping Zhuang, Yajun Jian, Yan Yan, and Hanzi Wang. Language decoupling with fine-grained knowledge guidance for referring multi-object tracking. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23626–23635. IEEE, 2025a.

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9845–9867, 2025b.

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)*, pp. 6816–6822. IEEE, 2025c.

Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7350–7358. IEEE, 2017.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv e-prints*, pp. arXiv–2505, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044. IEEE, 2025b.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

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.

Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial intelligence*, 293:103448, 2021.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Kyujin Shim, Kangwook Ko, Yujin Yang, and Changick Kim. Focusing on tracks for online multi-object tracking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11687–11696. IEEE, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20961–20970. IEEE, 2022.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *Advances in neural information processing systems*, 38:5772–5822, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13758–13768. IEEE, 2021.

Xiao Wang, Liye Jin, Xufeng Lou, Shiao Wang, Lan Chen, Bo Jiang, and Zhipeng Zhang. Reasoningtrack: Chain-of-thought reasoning for long-term vision-language tracking. *arXiv preprint arXiv:2508.05221*, 2025.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 80502–80530, 2026.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, et al. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *Advances in Neural Information Processing Systems*, 38:4463–4494, 2026.

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)*, pp. 3645–3649. IEEE, 2017.

Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14633–14642. IEEE, 2023.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Yani Zhang, Dongming Wu, Wencheng Han, and Xingping Dong. Bootstrapping referring multi-object tracking. *arXiv preprint arXiv:2406.05039*, 2024b.

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.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. Easyr1: An efficient, scalable, multi-modality rl training framework, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

APPENDIX

Appendix Section	Description
Section A	**RMOT26 Data Card**
Section B	**RMOT26 Dataset Statistics**
Section C	**Metrics and Implementation Details**
Section D	**Evaluation Protocol for Traditional MOT Comparison**
Section E	**Comparison with Fine-Tuned Models**
Section F	**Additional Qualitative Results**
Section G	**Failure Cases**
Section H	**Prompt Templates for Baseline Models**
Section I	**Reasoning Quality Evaluation Details**

A DATACARD FOR RMOT26

A.1 MOTIVATION

- **For what purpose was the dataset created?**
  The RMOT26 dataset is designed to evaluate *query-driven multi-object tracking*, where models track only objects specified by a natural language query rather than all objects in a scene. Each task requires identifying the correct object(s) in a reference frame and maintaining consistent trajectories across subsequent frames. The dataset evaluates language-conditioned spatiotemporal reasoning, including query understanding, spatial grounding, temporal identity consistency, and robustness under challenging conditions such as occlusion or object interactions.
- **Who created the dataset?**
  The dataset was created by the authors of this paper.
- **Who funded the creation of the dataset?**
  Funding information will be disclosed after the anonymous review process.

A.2 COMPOSITION

- **What do the individual instances represent?**
  Each instance represents a query-conditioned tracking task consisting of a reference frame with annotated bounding boxes, a sequence of subsequent frames, a natural language query describing the target object(s), and ground-truth trajectories.
- **How many instances are there in the benchmark?**
  The benchmark dataset contains 1000 instances: 900 single-object tracking tasks and 100 multi-object tracking tasks. Among the single-object tasks, 20 queries specifically focus on occlusion scenarios.
- **Is the dataset a complete set or a sample?**
  The dataset is a curated collection of tracking scenarios rather than an exhaustive set.
- **What data does each instance contain?**
  Each instance includes a reference frame, subsequent video frames, a natural language query, and ground-truth bounding box trajectories with object IDs.
- **Is there a label associated with each instance?**
  Yes. Ground-truth bounding boxes and object IDs define the target trajectories.
- **Is any information missing from instances?**
  No.

- **Are there recommended data splits?**
  Yes. The dataset provides predefined training and test splits described in Section B.
- **Are there errors or noise in the dataset?**
  The dataset is generated using a semi-automated pipeline and verified by humans. Minor annotation noise may occur due to human error.
- **Does the dataset rely on external resources?**
  Yes. The video sequences are derived from existing multi-object tracking datasets.
- **Does the dataset contain confidential data?**
  No.
- **Does the dataset contain potentially offensive content?**
  No.

A.3 Collection Process

- **How was the data acquired?**
  Language queries and reasoning steps were generated using Qwen2.5-VL-Instruct Team (2025) and subsequently reviewed and refined by human annotators. The visual data was collected from publicly available datasets.
- **What methods were used to generate and curate the data?**
  A semi-automated pipeline was used. Queries, reasoning steps, and candidate outputs were first generated by Qwen2.5-VL-Instruct and then manually verified for correctness and clarity.
- **Who was involved in the data collection process?**
  Researchers and student annotators.
- **Over what timeframe was the data collected?**
  The dataset was created between 2025 and 2026.
- **Were ethical considerations taken into account?**
  Yes. All visual data comes from datasets released for academic research. Removal requests from original dataset authors will be honored.

A.4 Preprocessing/Cleaning/Labeling

- **Was preprocessing or annotation performed?**
  Yes. Automatically generated queries and reasoning outputs were manually reviewed and corrected before inclusion.
- **Was the raw data saved?**
  No separate raw data is stored. The dataset contains curated subsets of images and videos from existing datasets along with generated annotations.

A.5 Uses

- **Has the dataset already been used for tasks?**
  Yes. It is used in this paper to evaluate query-driven multi-object tracking.
- **Is there a repository listing papers using the dataset?**
  No.
- **What other tasks could this dataset support?**
  The dataset can support research in vision-language reasoning, multimodal grounding, and agent-based visual decision-making.
- **Could the dataset composition impact future uses?**
  No known limitations are expected.
- **Are there potential negative social impacts?**
  No direct negative social impacts are anticipated.

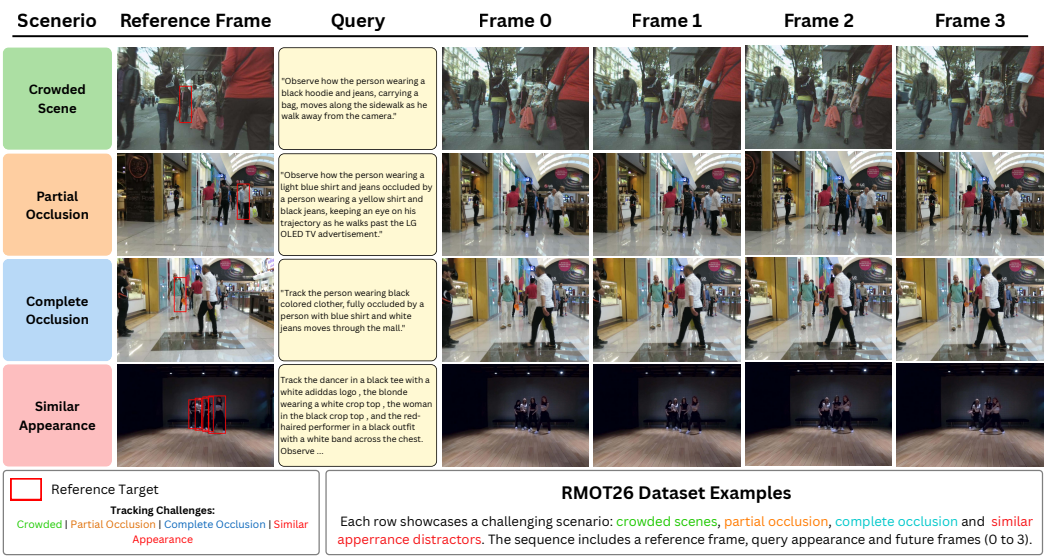


Figure 6: Overview of the RMOT26 dataset components. The dataset contains four types of query scenarios: (1) crowded scenes, (2) partial occlusion, (3) complete occlusion, and (4) similar-appearance distractors. These cases require models to perform robust language-guided tracking under challenging visual conditions.

A.6 Distribution

- **Will the dataset be distributed externally?**
  Yes.
- **How will the dataset be distributed?**
  The dataset will be released through a public GitHub repository.
- **Under what license will it be released?**
  The dataset will be released under the Apache License.
- **Are there any third-party restrictions?**
  No.

A.7 Maintenance

- **Who will maintain the dataset?**
  The authors of this paper.
- **How can the maintainers be contacted?**
  Contact details are provided in the paper.
- **Will the dataset be updated?**
  Yes. Updates and corrections will be made through the GitHub repository.
- **Can others contribute to the dataset?**
  Yes. Contributions and suggestions can be proposed through GitHub or by contacting the authors.

B RMOT26 Dataset Statistics

The RMOT26 dataset is constructed by curating video sequences from seven existing multi-object tracking (MOT) benchmarks. To ensure fair evaluation and avoid identity leakage between training and testing sets, all splits are performed at the sequence level rather than the frame level. Figure 6 illustrates the key components and challenging scenarios represented in the dataset. In particular, the

Table 3: Composition of the RMOT26 dataset derived from existing MOT benchmarks.

Dataset	Train		Test	
	Sequences	Frames	Sequences	Frames
MOT17	253	1058	242	995
MOT16	122	535	66	284
MOT20	73	305	46	190
DanceTrack	962	3864	574	2275
SportsMOT	73	297	2	8
PETS	21	81	2	6
DAMUNT	132	538	68	250

dataset includes four types of query conditions: (1) crowded scenes, (2) partial occlusion, (3) complete occlusion, and (4) similar-appearance distractors. These scenarios require models to perform both spatial grounding and temporal reasoning when tracking objects specified by natural language queries.

B.1 Dataset Composition

Each instance in the dataset consists of a reference frame followed by a short future sequence of 5–6 frames. The dataset is constructed using sequences from multiple existing MOT benchmarks to ensure diversity in scene types, motion patterns, and tracking difficulty. Table 3 summarizes the dataset composition, showing the contribution of each source dataset to the training and test sets in terms of sequences and frames.

B.2 Train / Test Splits

The overall dataset statistics for the training and test splits are presented in Table 4. The training set contains 1,636 sequences and 6,678 frames, while the test set includes 1,000 sequences and 4,008 frames. These splits ensure a balanced distribution of scenes and tracking conditions while preventing sequence overlap between the training and evaluation sets.

Table 4: Overall RMOT26 dataset statistics for training and test splits.

Split	Sequences	Frames
Train	1636	6678
Test	1000	4008

C Metrics and Implementation Details

Table 5 summarizes the evaluation metrics used throughout our experiments. **Hardware setup.** All model evaluations were conducted on a single NVIDIA H100 GPU (100GB), ensuring consistent hardware conditions across experiments. During evaluation, visual inputs and corresponding queries are processed by the selected `VLMM` agent, which produces structured outputs in a predefined `JSON` format.

Training hyperparameters. Table 6 summarizes the hyperparameters used to train `QTrack` on the `RMOT26` benchmark. Due to memory constraints, training was performed using only two frames per query. Training with the full sequence length may lead to different performance characteristics.

Train/test sequence length mismatch. Due to GPU memory constraints, **QTrack** is trained using only two frames per query (the reference frame plus one future frame), while evaluation is performed on sequences of 5-6 frames. This mismatch means the model must generalize its learned motion reasoning to longer horizons than seen during training. Despite this constraint, **QTrack** achieves the best performance across all metrics on `RMOT26`, suggesting that the structured reward design

Table 5: Evaluation metrics for RMOT26.

Metric	Description
IoU	Spatial overlap between B_t^{pred} and B_t^{gt} per frame.
Center Distance	Euclidean distance between predicted and ground-truth centers (x_t, y_t).
Direction Consistency (A_t)	Cosine similarity between predicted and ground-truth motion vectors Δ_t^{pred} and Δ_t^{gt}.
Speed Consistency (S_t)	Gaussian penalty on motion magnitude difference $\|\Delta_t^{pred}\|$ and $\|\Delta_t^{gt}\|$.
MCP	$\text{MCP} = \frac{1}{T-1}\sum_{t=2}^{T} A_t \cdot S_t.$

Table 6: Training hyperparameters for QTrack.

Hyperparameter	Value	Hyperparameter	Value
max_prompt_length	1300	*max_response_length*	2000
max_pixels	12845056	*use_kl_loss*	true
kl_loss_coef	5.0e-3	*fsdp.param_offload*	false
fsdp.torch_dtype	null	*offload.param_offload*	true
offload.optimizer_offload	true	*rollout.n*	8
limit_images	2	*learning_rate*	1.0e-6
weight_decay	1.0e-2	*max_grad_norm*	1.0
global_batch_size	1	*adv_estimator*	grpo
kl_coef	0.0	*kl_penalty*	kl
tapo_strategy	freeze	*tapo_keep_prob*	0.7
tapo_interval	2	*tapo_kl_weight*	0.1
rollout_batch_size	1	*temperature*	1.0
tensor_parallel_size	1	*total_episodes*	4
n_gpus_per_node	1		

and TAPO regularization provide generalizable motion-aware representations rather than overfitting to the two-frame training regime. Training with full sequence length remains an important direction for future work.

C.1 Motion Consistency Precision (MCP)

Traditional tracking metrics primarily focus on spatial alignment (e.g., IoU) or identity preservation across frames. However, such metrics do not explicitly evaluate whether predicted trajectories follow physically consistent motion patterns over time. To address this limitation, we introduce **Motion Consistency Precision (MCP)**, a metric designed to measure the temporal coherence between predicted and ground-truth object trajectories.

MCP evaluates whether the predicted motion follows the same trajectory dynamics as the ground truth by jointly considering both the *direction* and the *speed* of motion. Intuitively, a reliable tracker should not only localize objects accurately in each frame but should also maintain motion patterns consistent with the underlying object dynamics.

Bounding-box convention. Ground-truth annotations are stored as

$$\mathbf{b}^{xywh} = (x_{\text{tl}}, y_{\text{tl}}, w, h),$$

where $(x_{\text{tl}}, y_{\text{tl}})$ is the upper-left corner. Model outputs use the corner-coordinate representation

$$\widehat{\mathbf{b}}^{xyxy} = (\widehat{x}_1, \widehat{y}_1, \widehat{x}_2, \widehat{y}_2).$$

Before computing rewards or evaluation metrics, a stored ground-truth box is converted according to

$$\phi(x_{\text{tl}}, y_{\text{tl}}, w, h) = (x_{\text{tl}}, y_{\text{tl}}, x_{\text{tl}} + w, y_{\text{tl}} + h).$$

Table 7: Comparison with specialized multi-object trackers and reasoning-based trackers on MOT17 and DanceTrack. The results illustrate the relative strengths of QTrack across localization, association, and motion-consistency metrics.

(a) MOT17 Dataset

Model	MOTA	MOTP	HOTA	MCP
MOTR	0.61	0.81	0.22	0.44
BoostTrack++	0.63	0.76	0.38	**0.44**
TrackTrack	**0.75**	0.50	0.23	0.29
VisionReasoner	0.64	0.86	0.60	0.21
QTrack	0.69	**0.87**	**0.69**	0.26

(b) DanceTrack Dataset

Model	MOTA	MOTP	HOTA	MCP
MOTR	0.42	0.70	0.35	0.51
MOTRv2	0.49	0.73	0.37	0.52
TrackTrack	0.36	0.73	0.40	**0.55**
VisionReasoner	0.59	**0.85**	0.61	0.26
QTrack	**0.63**	0.83	**0.66**	0.35

After conversion, all IoU, matching, center-distance, and motion calculations operate on the common $xyxy$ representation. For a box $\mathbf{b} = (x_1, y_1, x_2, y_2)$, its center is

$$c(\mathbf{b}) = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right).$$

Motion Consistency Precision. The frame-level motion consistency score is defined as the product of direction and speed consistency:

$$\text{MCP}_t = A_t \cdot S_t. \tag{13}$$

For a sequence of T frames, the overall Motion Consistency Precision is computed as

$$\text{MCP} = \frac{1}{T-1} \sum_{t=2}^{T} \text{MCP}_t. \tag{14}$$

By construction, MCP lies within the range $[0, 1]$, where higher values indicate stronger agreement between predicted and ground-truth motion trajectories. Unlike purely spatial metrics, MCP explicitly measures the *temporal realism* of object motion, making it particularly suitable for evaluating tracking systems that must maintain coherent trajectories across time.

Relation to existing temporal metrics. Existing association-based metrics such as HOTA Luiten et al. (2021) decompose tracking performance into detection accuracy (DetA) and association accuracy (AssA). However, AssA measures identity consistency via overlap-based matching across frames and does not penalize physically implausible motion patterns. Similarly, TETA measures localization, association, and classification but does not enforce direction or speed coherence. MCP fills this gap by directly measuring whether predicted trajectories follow the same motion dynamics as ground-truth trajectories, making it uniquely suited to short-horizon query-conditioned tracking where identity switches are rare but spatial drift and motion inconsistency are the dominant failure modes.

Why not IDF1 or HOTA as primary metrics? Standard association metrics such as IDF1 and HOTA are designed for long-video, all-object tracking where identity switches and fragmentation are meaningful failure modes. In `RMOT26`, each instance spans only 5-6 frames with explicit reference-frame grounding, meaning all models begin with a known target location. Under these conditions, identity switches are structurally rare across all evaluated models (exceeding 95% IDF1 for most baselines), making association-based metrics uninformative for model discrimination. The dominant failure mode in our setting is spatial drift and motion inconsistency, which MCP directly quantifies. We therefore use MCP, MOTP, CLE, and NDE as primary metrics, and report HOTA and MOTA only in the traditional MOT comparison (Section D), where longer segments make association failures meaningful.

Table 8: Comparison of VLLM models after fine-tuning.

Model	Params	MCP↑	MOTP↑	MOTA↑	NDE↓
Qwen 2.5 VL Instruct Team (2025)	3B	0.14	**0.76**	-0.51	3.41
Gemma3 Team et al. (2025)	4B	0.18	0.73	-0.16	0.95
Visionreasoner Liu et al. (2025a)	3B	<u>0.22</u>	0.65	<u>0.01</u>	<u>0.76</u>
QTrack (Ours)	**3B**	**0.30**	<u>0.75</u>	**0.21**	**0.39**

D Evaluation Protocol for Traditional MOT Comparison

To compare **QTrack** with traditional multi-object tracking (MOT) methods, we design a query-based evaluation protocol derived from standard MOT benchmarks. Specifically, we use sequences from the MOT17 Milan et al. (2016) and DanceTrack Sun et al. (2022) datasets, adapting them to reflect the query-driven nature of our task. Traditional MOT benchmarks evaluate trackers that detect and track *all* objects across an entire video. In contrast, **QTrack** is designed to track only objects specified by a natural-language query. Direct evaluation on full MOT videos would therefore be inconsistent with the intended task. To address this mismatch, we construct *query-specific tracking segments* from benchmark sequences so that both **QTrack** and conventional trackers can be evaluated under comparable conditions. For each sequence, we first select a reference frame containing the target object and generate a natural-language query describing that object. Starting from this reference frame, the tracker is required to follow only the queried target across subsequent frames. The predicted trajectories are then converted to the standard MOT annotation format, enabling the computation of common tracking metrics such as HOTA, MOTA, and MOTP.

This protocol allows **QTrack** to be evaluated on widely used MOT datasets while preserving the query-based nature of the task. At the same time, it enables meaningful comparison with traditional trackers by isolating the trajectory of the queried object within a multi-object scene. Notably, this evaluation setting introduces a stricter requirement for **QTrack**. Instead of relying on global associations among all objects in the scene, the tracker must reason about the identity of the queried target and maintain its trajectory over time. Despite this constraint, **QTrack** achieves competitive or superior performance on standard MOT metrics, demonstrating that query-aware reasoning can improve robustness in complex tracking scenarios. Overall, this protocol bridges traditional MOT evaluation with query-based tracking, enabling fair and practical comparisons between the two paradigms.

Table 7 presents the quantitative comparison between **QTrack** and traditional MOT methods on query-specific segments from MOT17 and DanceTrack. On MOT17, **QTrack** achieves the best MOTP (0.87) and HOTA (0.69) while maintaining competitive MOTA (0.69). On DanceTrack, **QTrack** achieves the best MOTA (0.63) and HOTA (0.66). These results confirm that query-conditioned reasoning does not sacrifice trajectory quality relative to methods that track all objects indiscriminately.

E Comparison with Fine-Tuned Models

Table 8 evaluates fine-tuned VLLMs under identical training settings. QTrack achieves the highest MCP (0.30), highest MOTA (0.21), and lowest NDE (0.39), highlighting the importance of reward design and temporal regularization beyond supervised fine-tuning.

F Additional Qualitative Results

We present additional qualitative comparisons to further analyze the behavior of **QTrack** under different tracking scenarios. Figures 7, 8, and 9 compare the predictions of several VLLM-based baselines with our proposed method across multiple frames. As illustrated, **QTrack** produces more stable and accurate trajectories across time, particularly in challenging situations involving motion, occlusion, and appearance similarity. In contrast, baseline models often exhibit drifting predictions or inconsistent localization across frames. Figure 8 presents a qualitative comparison for multi-object tracking. VisionReasoner fails to detect and track all objects across frames, whereas **QTrack** successfully localizes and tracks all queried objects in every frame.

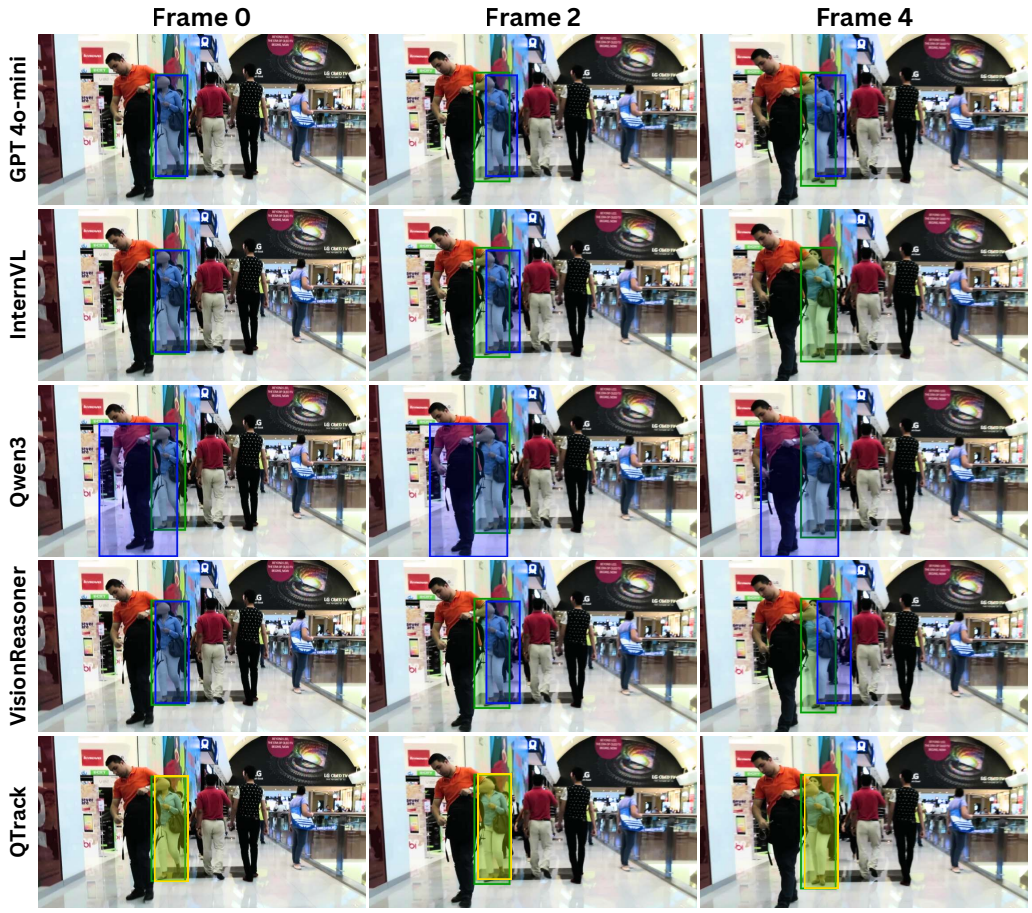
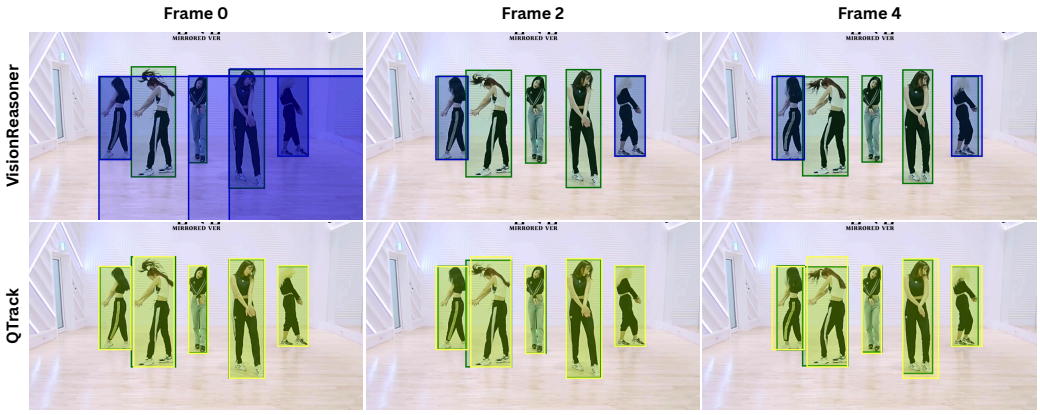


Figure 7: Qualitative comparison for single-object tracking across multiple frames. **QTrack** produces more accurate and stable bounding boxes over time, while baseline VLLM models often exhibit drifting or inconsistent localization as the sequence progresses.



"Track the dancer in the black crop top with yellow side stripes at [402.0, 256.0, 183.0, 477.0], the red-haired performer in a white top and black pants at [601.0, 201.0, 232.0, 631.0], and the center dancer in a long-sleeve black top with blue jeans at [915.0, 253.0, 112.0, 501.0]. Also keep an eye on the woman in a sleeveless black top with white-stripe pants at [1149.0, 213.0, 204.0, 684.0] and the blonde in an all-black outfit at [1428.0, 252.0, 175.0, 462.0]; how do their trajectories shift and their spacing and alignment evolve relative to each other over the next frames?"

Figure 8: In this figure, Visionreasoner is not able to detect and track all the objects in the frames, but QTrack can detect and track all objects in all frames



Track the man wearing a red shirt and beige pants as he walks through the mall corridor. Observe how his position changes as he moves past the storefronts and along the railing

Figure 9: Additional qualitative comparison of different VLLM models on challenging tracking scenarios. Our method demonstrates stronger temporal consistency and better object localization compared to baseline models.

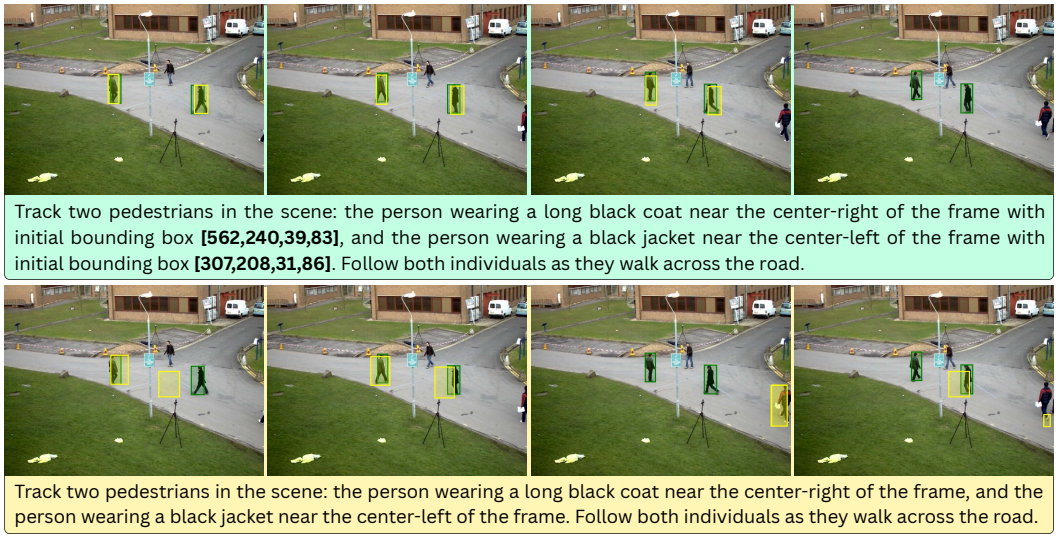


Figure 10: Effect of spatial initialization. Tracking is reliable with an initial bounding box, but localization becomes difficult when relying solely on language queries.

Figure 10 investigates the impact of providing an initial bounding box in the reference frame. We evaluate two settings: (1) the model receives both a natural-language query and the initial bounding box of the target object, and (2) the model receives only the natural-language query without any spatial initialization. When the initial bounding box is provided, the model can reliably maintain the trajectory of the queried object across subsequent frames. However, when the model relies solely on the textual query, it occasionally struggles to precisely localize the target object in the first frame. This highlights an important limitation of current vision-language models: while they can reason about object descriptions, precise spatial grounding from language alone remains challenging in complex scenes.

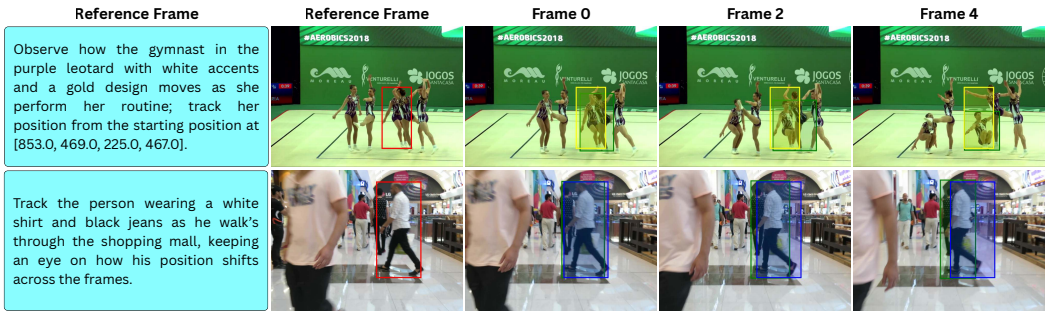


Figure 11: This figure shows failure cases of QTrack, we can see that for some sequences QTrack is unable to predict next bounding box for frames, it repeatedly predicts the same initial bounding box for all future frames.

G Failure Cases

Figure 11, shows failure cases of QTrack; we can see that for some sequences, QTrack is unable to predict the next bounding box for frames; it repeatedly predicts the same initial bounding box for all future frames.

H Prompt Templates for Baseline Models

To ensure a fair comparison, all baseline models are evaluated using the same input information and prompt structure.

Input format. Each model receives the following inputs:

- Video frames from the sequence
- A natural-language query describing the target object
- A reference bounding box indicating the initial location of the object in the reference frame

Prompt template. We use a unified prompt template for all baseline models. The template instructs the model to track the queried object across the provided video frames and output the predicted bounding boxes in a structured JSON format.

Prompt Template

```
{textual_query}

Initially, object {obj_id} is located at {initial_bbox}
in the reference image.

The following images correspond to consecutive
frames of a video.

Please output the bounding box for the object in
each frame using the following format:

[{"frame": N, "bbox": [x1, y1, x2, y2]}, ...]

Do not include any additional text.
```

This prompt ensures that all models receive identical instructions and output constraints. By enforcing a structured JSON response format, we simplify downstream parsing and ensure consistent evaluation across different baseline models.

I Reasoning Quality Evaluation Details

Why Reasoning? To check if reasoning is adding any contribution to results we added a direct trajectory-only ablation in Table 9. Removing structured reasoning degrades every metric: MCP decreases from 0.30 to 0.28, MOTP decreases from 0.75 to 0.72, CLE increases from 44.61 to 129.5, approximately a 2.9× increase, NDE increases from 0.39 to 0.81, approximately a 2.1× increase.

Table 9: Ablation for reasoning

Model	MCP↑	MOTP↑	CLE↓	NDE↓
QTrack with structured reasoning	**0.30**	**0.75**	**44.61**	**0.39**
QTrack without structured reasoning	0.28	0.72	129.5	0.81

The structured reasoning field gives the policy an intermediate representation for binding: language-specified attributes, selected target identity, motion direction, visibility changes, object interactions, final trajectory coordinates. The ablation suggests that this structure is functionally useful rather than purely post-hoc. We do not claim that every generated rationale is perfectly faithful. This is why the paper separately analyzes 200 reasoning traces and reports unsupported-motion and weak-causal-justification failures. The two analyses answer different questions: (1) the correlation analysis measures whether better individual reasoning traces are associated with better predictions (2) the trajectory-only intervention tests whether the structured reasoning-output formulation contributes to tracking performance.

Scoring protocol. We evaluate 200 test instances sampled with stratification across four IoU tiers: easy ($\geq$0.7, n=93), medium ($[0.4, 0.7)$, n=56), hard ($[0.1, 0.4)$, n=28), and failure (< 0.1, n=23). Each reasoning trace is scored on four dimensions (0–2 each): (A) *Object identification* explicit

reference to target identity or initial location; (B) *Temporal reasoning* mention of motion direction, continuity, or occlusion; (C) *Causal justification* explanation of *why* the target matches the query beyond appearance; (D) *Faithfulness* geometric consistency between stated motion direction and predicted trajectory. RQS is defined as:

$$\text{RQS} = \frac{1}{8N} \sum_{i=1}^{N} \big(r_i^{\text{rel}} + r_i^{\text{temp}} + r_i^{\text{causal}} + r_i^{\text{faith}} \big). \tag{15}$$

Table 10: RQS and faithfulness by IoU tier (n=200).

Tier	n	IoU	RQS	Faithfulness
Easy ($\geq$0.7)	93	0.882	0.593	0.661
Medium ($[0.4, 0.7)$)	56	0.599	0.554	0.580
Hard ($[0.1, 0.4)$)	28	0.278	0.585	0.643
Failure (< 0.1)	23	0.021	0.489	0.457
All	200	0.619	0.569	0.280†

†Fraction with faithfulness ≥ 1.5. RQS $\leftrightarrow$ IoU: r=+0.158, p=0.025. RQS $\leftrightarrow$ MCP: r=$-$0.10, p=0.16 (n.s.).

Failure mode distribution. Among 51 low-IoU samples (hard + failure tiers): weak causal justification 25 (49%), directional hallucination 9 (18%), motion failure 9 (18%), appearance bias 8 (16%).