

CoMaTrack: Competitive Multi-Agent Game-Theoretic Tracking with Vision-Language-Action Models

Li Gao, Liu Liu, Mingyang Lv, and Yang Cai

Amap, Alibaba Group, Beijing, China

liangliang.gl@alibaba-inc.com, diana.ll@alibaba-inc.com,
lvmingyang2024@ia.ac.cn, yangcai.cy@autonavi.com

Abstract. Embodied Visual Tracking (EVT) requires an agent to continuously follow a language-specified target from egocentric observations. Most existing methods adopt single-agent imitation learning, which depends on expert demonstrations and can struggle to generalize to scenarios beyond the training distribution. Inspired by competition-driven capability evolution in nature, we propose CoMaTrack, a competitive multi-agent reinforcement learning (RL) framework that trains agents in a dynamic adversarial setting with competitive subtasks, yielding stronger adaptive planning and interference-resilient strategies. We further propose CoMaTrack-Bench, a Habitat-based benchmark for language-conditioned competitive EVT, featuring dynamic dueling scenarios between a tracker and adaptive opponents across diverse environments and instructions, enabling standardized robustness evaluation under active adversarial interactions. Experiments show that CoMaTrack achieves state-of-the-art results on both standard benchmarks and CoMaTrack-Bench. Notably, a 3B VLM trained with our framework surpasses previous single-agent imitation learning methods based on 7B models on the challenging EVT-Bench, achieving 92.1% in STT, 74.2% in DT, and 57.5% in AT.

Keywords: Embodied Visual Tracking · Visual Language Navigation · Reinforcement Learning

1 Introduction

In recent years, embodied agents driven by Large Language Models (LLMs) [1, 5, 24, 26, 31, 57] and vision transformers [15, 17, 40] have made remarkable progress, allowing perception, reasoning, and decision-making in complex 3D environments and moving from merely understanding instructions to executing tasks. Among them, Embodied Visual Tracking (EVT) [43, 61, 76, 77, 79] is a long-studied embodied task that requires an agent to continuously follow a dynamic target from egocentric observations, while operating under environmental uncertainty, changing target behaviors, frequent occlusions, and partial observability

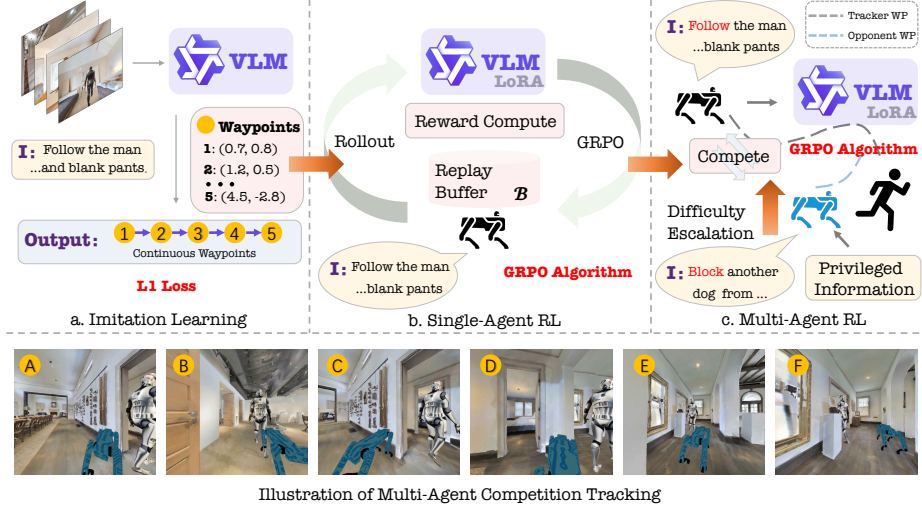


Fig. 1: CoMaTrack frames EVT as a competitive multi-agent game rather than a single-agent pursuit in a static environment. (a) IL: the agent learns from offline demonstrations in static scenes, with limited exposure to rare failures. (b) Single-Agent RL: the agent improves through interaction, but the target and environment remain largely fixed, leading to slow exploration and overfitting to predefined behaviors. (c) Multi-Agent RL: the agent trains against adaptive opponents that evade or block on purpose, dynamically increasing difficulty and producing diverse adversarial trajectories, encouraging anticipation, relocalization, and robustness under interference.

over long horizons. Unlike static recognition or short-range navigation, EVT emphasizes persistent closed-loop control and online error correction, inherently demanding target understanding, spatiotemporal reasoning, memory maintenance, and robustness to interference—making it a critical stepping stone towards general embodied intelligence and real-world robotic applications.

However, existing research remains largely in a single-agent setting and is predominantly trained through Imitation Learning (IL) [13, 18, 22, 61, 72], leading to limited generalization and robustness in out-of-distribution scenes. A central limitation of single-agent IL is its dependence on high-quality expert trajectories: data collection are expensive, and the task distribution is manually designed, making it difficult to cover the rich variations encountered in the real world [7, 14, 44].

More importantly, although reinforcement learning (RL) [10, 32, 59, 79] provides closed-loop interaction and exploration, and in principle can yield more general policies through trial-and-error [73], EVT still lacks systematic RL studies and effective training paradigms. Conventional RL training [25, 55, 56] typically relies on relatively static environments and fixed target generators, which struggle to continuously produce hard enough and progressively more challenging supervision signals. Without an automatically escalating learning pressure,

agents are prone to inefficient exploration and overfitting, resulting in limited sample efficiency and generalization gains.

We draw inspiration from competition in nature: under the Red Queen dynamic [58], co-evolving species continually escalate selective pressure on one another, and evolutionary game theory [45] formalizes how such competitive interactions shape adaptive strategies. Analogously, changes in an opponent’s strategy continuously raise task difficulty, forcing the agent to improve anticipation, planning, and counteraction. Inspired by this, we introduce CoMaTrack, a competitive multi-agent game-theoretic learning mechanism into EVT and build a co-evolving training loop among agent–opponent–environment. The agent no longer faces a static environment and a fixed data distribution; instead, it repeatedly encounters stronger, more cunning, and more diverse behaviors through adversarial interactions, yielding an automatic curriculum with progressive difficulty. Any improvement by one agent becomes a new challenge that others must overcome, ultimately forming a self-reinforcing arms race. Different from conventional multi-agent cooperation that primarily optimizes system-level efficiency, we focus on sharpening the core competence of a single tracking agent via competitive games, enabling robust behaviors under typical tracking difficulties, and improving transferability to unseen environments and target behaviors.

As shown in Fig. 1, we shift the paradigm from static data-driven training to adversarial interaction-driven learning. Critically, existing EVT benchmarks (*e.g.*, EVT-Bench [61], TPT-Bench [66], Gym-UnrealCV [52], Unreal-Zoo [81]) exacerbate this limitation: they evaluate tracking under idealized or passively disturbed conditions, lacking scenarios with active adversarial competition that mirror real-world challenges like intentional evasion, strategic deception, or contested pursuit. To bridge this gap, we propose CoMaTrack-Bench, the first benchmark for competitive EVT, featuring dynamic dueling scenarios between a tracker agent and adaptive opponent across diverse environments and language instructions. This benchmark establishes a standardized protocol for rigorously assessing robustness under intentional adversarial interactions. Our main contributions are as follows:

- First integration of competitive multi-agent game-theoretic RL with modern VLA architectures for language-conditioned embodied visual tracking: We devise a co-evolving adversarial training loop where opponent strategies dynamically escalate task difficulty.
- Proposing CoMaTrack-Bench: We present the first benchmark suite that standardizes competitive EVT evaluation in Habitat with language instructions, featuring dynamic dueling scenarios with adaptive opponents to enable rigorous, real-world-aligned evaluation of tracking robustness—addressing the idealized limitations of prior benchmarks.
- Extensibility to broader VLA embodied tasks: While this paper focuses on tracking, our approach offers a general paradigm of opponent-driven difficulty generation and competitive-game-driven generalization, and can be naturally extended to other VLA embodied tasks to address out-of-distribution generalization.

2 Related Work

2.1 Visual Language Navigation

Recent VLN [2, 3, 11, 39] research has increasingly shifted from modular pipelines to Vision-Language-Action (VLA) [7, 27, 47] that map egocentric observations and language instructions to actions in continuous environments. Uni-NaVid [72] extends video VLMs with online token merging for long-context processing and scales multi-task data for unified navigation. DV-VLN [30] introduces navigational chain-of-thought with dual verification signals to re-rank candidate actions. SpatialVLN [69] injects explicit spatial awareness through multi-sensor fusion and multi-expert reasoning. DeepVLN [19] adapts LLMs to VLN via SFT followed by online RL to reduce error accumulation.

These advances collectively demonstrate that stronger VLM backbones, scalable multi-task data, explicit spatial priors, and closed-loop RL can significantly improve VLN performance. Our approach is complementary: we use multi-agent game-based RL as the core mechanism to produce robust VLA policies that can transfer to navigation-style tasks when needed, while being designed and evaluated primarily around tracking-centric requirements.

2.2 Embodied Visual Tracking

Embodied Visual Tracking (EVT), demanding agents to pursue language-specified targets amid dynamic occlusions, crowd interference, and environmental uncertainty, has witnessed rapid progress through large-scale imitation learning (IL) pipelines [80]. TrackVLA [61] pioneers a unified VLA architecture, coupling target identification and waypoint planning via supervised token prediction. TrackVLA++ [36] enhances robustness through Polar Chain-of-Thought spatial reasoning and a confidence-gated Target Identification Memory module—yet remains fundamentally anchored to static demonstration datasets. Similarly, LOVON [48] integrates LLM-based planning, open-vocabulary perception, and a language-to-motion model for legged robot tracking; while innovating in motion-stabilized perception [11], its policy derives entirely from demonstration-driven training. Collectively, these works epitomize the field’s paradigm: scaling EVT performance hinges on expanding expert trajectory repositories, while policies remain brittle under distribution shift and active interference.

Beyond imitation-centric pipelines, reinforcement learning and game-theoretic approaches [46, 82] have long been explored for active visual tracking. Luo *et al.* [41] demonstrated end-to-end active object tracking via RL with real-world deployment. Pioneering work AD-VAT [77] introduced the asymmetric dueling mechanism in which a tracker and a target agent engage in adversarial competition, demonstrating that co-evolutionary training yields more robust policies than single-agent paradigms. This was further analyzed in AD-VAT+ [76] and extended to distractor-robust tracking via mixed cooperative-competitive multi-agent games [78], and subsequent studies validated adversarial training for

tracking generalization across diverse environments [75, 79]. However, these earlier frameworks operated in non-language-conditioned settings with traditional CNN/A3C perception backbones and discrete action spaces. HIEVT [62] proposes a hierarchical instruction-aware framework using LLM-based goal alignment, while [63] introduces a self-improving VLM framework for embodied tracking with memory-augmented reflection. CoMaTrack builds directly on the asymmetric-dueling principle of AD-VAT; our contribution is not the dueling principle itself but its *adaptation to modern language-conditioned VLA models*, predicting continuous waypoints under natural-language instructions and large-scale VLM priors.

2.3 Reinforcement Learning in VLN

Reinforcement learning (RL) [12, 16, 37] has recently re-emerged as an effective tool for improving VLN agents, where supervision is often sparse, long-horizon credit assignment is difficult [54], and generalization under distribution shift remains challenging. OpenVLN [33] proposes an open-world aerial VLN framework that uses a rule-based RL procedure and a value-reward mechanism to fine-tune VLMs when training data are scarce. Inspired by DeepSeekR1 [20], VLN-R1 [51] advances RL-based alignment for continuous navigation by directly operating on egocentric video streams with large VLMs. Bundeale *et al.* [8] establish one of the first offline RL benchmarking suites for VLN by generating suboptimal datasets through perturbations of expert rollouts. ActiveVLN [74] performs multi-turn RL training that treats VLN as a sequential decision process, enabling agents to learn optimal policies through direct interaction with the simulator. While these studies demonstrate that RL can improve VLN through better reward shaping, long-horizon credit assignment, and efficient model adaptation, they remain largely confined to single-agent training regimes. In contrast, prior VLA research has not systematically adopted online RL as a primary training paradigm for EVT, although Zhong *et al.* [79] propose an efficient EVT framework that integrates visual foundation models with offline RL and demonstrates strong sim-to-real transfer on a real robot.

We present the first GRPO-based online RL fine-tuning framework for VLA models in language-conditioned EVT, and further move beyond single-agent RL by formulating tracking as a competitive multi-agent game-theoretic training process. By leveraging competitive interactions, agents become each other’s adaptive environment, creating an automatically escalating curriculum that is difficult to obtain with static reward designs or offline data alone.

3 Methods

3.1 Task Formulation

The EVT task can be formulated as: an agent receives a natural language instruction $\mathcal{I}$ describing the visual characteristics of the specific target together

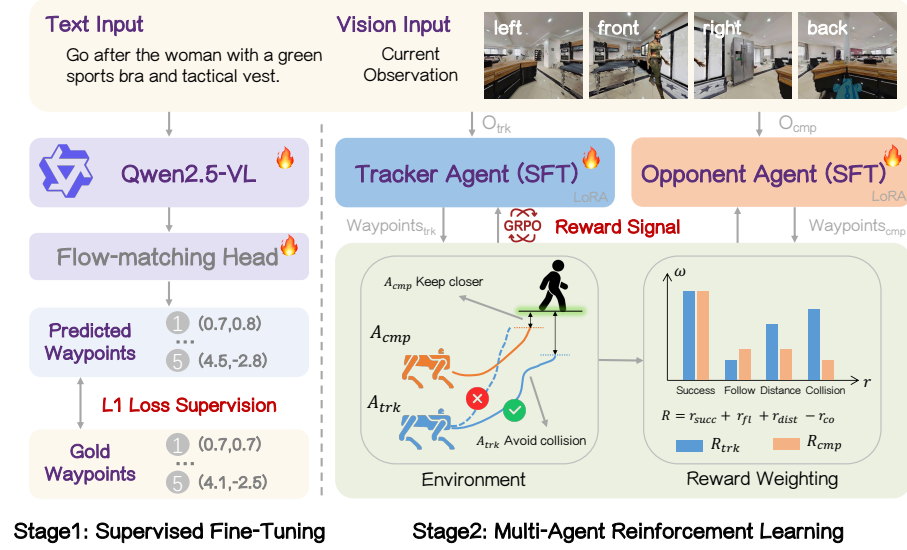


Fig. 2: Overview of the CoMaTrack framework. The system employs an end-to-end VLA architecture built upon the Qwen2.5VL-3B. During the SFT phase, the model learns to predict future trajectories from multi-view observations and historical visual sequences. In the RL phase, the tracker and opponent agents engage in competitive training within a dynamic adversarial environment, co-evolving robust tracking policies through the GRPO algorithm.

with a sequence of egocentric RGB visual observations from N cameras over the time indices $\{1, \dots, t\}$. Conditioned on these observations and the instruction, the agent must predict a sequence of five continuous tracking waypoints $\mathcal{W}_T = \{w_1, w_2, \dots, w_5\}$, where each waypoint $w_i = (x, y, \theta) \in \mathbb{R}^3$ specifies a relative motion in the agent’s coordinate frame, consisting of a planar displacement (x, y) and a heading angle θ . An episode is deemed successful when the agent follows the target at a safe distance of 1–3m, keeps the target in front of its view, and avoids collisions throughout the process.

3.2 CoMaTrack Overview

As illustrated in Fig. 2, CoMaTrack is an end-to-end VLA model built upon a video-centric VLM backbone Qwen2.5VL-3B [64]. On top of this backbone, we incorporate a flow-matching action module [34] that produces accurate distributions over future trajectories, enabling joint language generation and motion planning within a single framework. Text tokens are generated autoregressively in the standard manner, while trajectory planning is performed by the action module, which conditions on the backbone’s visual-linguistic representations to predict the agent’s action sequence.

3.3 Supervised Fine-Tuning

At time T , CoMaTrack processes multimodal inputs comprising current multi-view egocentric observations (front, rear, left, right) and temporally extended historical front-view sequences managed via a sliding window memory, together with a language instruction $\mathcal{I}$. Visual features are extracted using the vision backbone of Qwen2.5VL-3B [64] and compressed through multi-scale grid pooling: fine-grained tokens preserve spatial fidelity of the current observation for precise target grounding, while coarse-grained tokens compactly encode long-range temporal context. The navigation-specific visual sequence is structured as $\mathbf{V}_{\text{track}} = \{\mathbf{V}_{\text{coarse}}^{T-k}, \dots, \mathbf{V}_{\text{coarse}}^{T-4}, \mathbf{V}_{\text{fine}}^{T-3}, \dots, \mathbf{V}_{\text{fine}}^T\}$, where k denotes the memory window size. These tokens are fused with language embeddings containing a $\langle \text{Nav} \rangle$ special token and processed by the Qwen2.5VL-3B large language model. The LLM output adaptively branches: recognition tasks employ standard autoregressive decoding with cross-entropy loss $\mathcal{L}_{\text{text}}$; navigation tasks condition a Flow Matching-based action model [34] to directly regress a 5-step trajectory $w_i = \{(x_i, y_i, \theta_i)\}_{i=1}^5$, with tracking loss $\mathcal{L}_{\text{track}}$ defined as the mean squared error to ground-truth waypoints. The unified training objective is:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{track}} + \alpha \mathcal{L}_{\text{text}} = \sum_{i=1}^K \|\hat{w}_i - w_i\|_2^2 + \alpha \sum_j \text{CE}(\hat{t}_j, t_j),$$

where $\mathcal{L}_{\text{track}}$ is a waypoint regression loss, $\mathcal{L}_{\text{text}}$ is the token-level cross-entropy loss between predicted token distributions $\hat{t}_j$ and ground-truth text tokens t_j , and α balances the two objectives.

3.4 Single-Agent RL

To overcome the limitations of IL and enable autonomous policy improvement through environmental interaction, we employ Group Relative Policy Optimization (GRPO) [56] as our single-agent RL framework.

Reward Design. The reward combines dense, step-wise shaping with sparse terminal rewards. The dense term is a distance reward r_{distance} that follows a Gaussian centered at an optimal following distance to encourage safe proximity to the target:

$$r_{\text{distance}} = \exp\left(-\frac{1}{2}\left(\frac{d - d_{\text{opt}}}{\sigma}\right)^2\right) \cdot w_{\text{distance}}$$

where d is the current distance to the target, $d_{\text{opt}} = 2.25$, and $\sigma = 0.75$. Additionally, we provide a facing reward for maintaining proper orientation towards the target and a tracking persistence bonus for keeping the target within a safe following zone over consecutive steps. To provide clear long-term objectives, these dense signals are complemented by sparse terminal rewards: a large positive reward for successfully completing an episode, and significant penalties for failures such as losing the target or collisions.

GRPO Optimization. The policy π_θ is optimized using the GRPO objective, which clips the probability ratio to prevent overly large policy updates:

$$L_{\text{GRPO}} = \mathbb{E} \left[\min \left(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)} A^{\text{group}}, \text{clip} \left(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) A^{\text{group}} \right) \right]$$

To stabilize training and prevent catastrophic forgetting of the knowledge acquired during SFT, we integrate a KL divergence constraint L_{KL} that penalizes deviation from a *frozen* reference SFT policy π_{SFT} , estimated with the unbiased k_3 estimator:

$$L_{\text{KL}} = \lambda_{\text{KL}} \cdot \text{KL}(\pi_\theta \| \pi_{\text{SFT}})$$

The complete optimization objective is a weighted sum of the policy loss, an entropy bonus to encourage exploration, and the regularization terms:

$$L_{\text{total}} = L_{\text{GRPO}} - \lambda_{\text{ent}} \mathcal{H}(\pi_\theta) + L_{\text{KL}}$$

3.5 Multi-Agent RL

To cultivate robust tracking policies under active adversarial interference, we extend the single-agent RL framework into a competitive multi-agent game-theoretic RL paradigm. Specifically, we deploy two quadrupedal agents: the tracker agent A^{trk} , which is required to follow the language-specified human target, and the opponent agent A^{cmp} , which competes for the same target and implicitly interferes with the tracker through occlusion, path crossing, and physical blocking. Both agents adopt the identical VLA architecture and optimization pipeline. However, their reward functions are asymmetrically designed to reflect distinct strategic objectives, driving co-evolutionary policy adaptation.

Asymmetric Reward Design. The opponent should aggressively approach the target to create contested tracking, while the tracker should maintain stable following and avoid collisions with the opponent. Let d_{trk} and d_{cmp} denote the distances from A^{trk} and A^{cmp} to the target, and let d_{int} be the inter-agent distance between the two robots.

Opponent reward R^{cmp} . We encourage the opponent to track at a closer distance than the nominal safe-following range. We use a dense Gaussian distance reward but shift the optimum to a nearer value $d_{\text{opt}}^{\text{cmp}} = 1.25m$ to produce more direct contention.

Tracker reward R^{trk} . The tracker retains the standard EVT objective in Sec.3.4, but we additionally incorporate an opponent-aware safety term to discourage collisions and unsafe proximity to A^{cmp} .

Multi-Agent GRPO Optimization. Given simultaneous rollouts under the joint environment dynamics, we optimize each agent’s policy with the same GRPO algorithm, but using its own trajectory returns. In practice, the multi-agent setup is as stable as single-agent RL: both agents share the same on-policy update schedule and the KL anchor to the frozen SFT reference, which bounds per-update policy drift and mitigates the non-stationarity introduced by the co-evolving opponent.

3.6 CoMaTrack Benchmark

We propose CoMaTrack-Bench, a benchmark for competitive EVT in Habitat with language instructions. Unlike prior benchmarks that evaluate tracking under idealized or passively disturbed conditions, CoMaTrack-Bench places a tracker against an adaptive opponent that actively interferes through occlusion, path blocking, and competitive pursuit.

Benchmark Construction. CoMaTrack-Bench is built upon the Single-Target Tracking (STT) data from EVT-Bench, inheriting its diverse environments from HM3D [53] and MP3D [9] and natural language target descriptions. To create competitive scenarios, we augment each STT episode by introducing a second robotic dog agent initialized 0.5 meters ahead of the tracker’s starting position, ensuring immediate interaction from episode initialization.

Opponent Behavior Taxonomy. We design three progressively challenging opponent behaviors to comprehensively assess tracking robustness: (1) Static Obstacle: The opponent remains fixed at its initial position, serving as a static obstacle that occludes the target or blocks pursuit paths. This tests the tracker’s ability to handle persistent spatial occlusions and path planning around stationary obstacles. (2) Random Interference: The opponent executes random patterns within the environment, creating unpredictable dynamic occlusions and path crossings. This evaluates robustness to stochastic disturbances and the tracker’s capacity to maintain target identity amid moving distractors. (3) Competitive Tracking: The opponent loads the same SFT-trained policy as the tracker and actively competes to follow the same target, creating direct adversarial pursuit scenarios where both agents vie for optimal tracking positions.

4 Training Recipe and Data Collection

4.1 Data Collection

To improve the model’s versatility and generalization, and to facilitate transfer to other VLN-style tasks, we train not only on EVT data but also on navigation data collected from diverse synthetic environments (*e.g.*, HM3D [53]). In addition, we incorporate large-scale real-world VQA data [35] to further strengthen cross-domain generalization.

Tracking Data Collection. Following TrackVLA [61], we collect EVT training data using EVT-Bench [61] built on Habitat 3.0 [50] with the scenes from HM3D [53] and MP3D [9]. In total, we gather 6913 Single-Target Tracking (STT) episodes, 6685 Distracted Tracking (DT) episodes, and 6524 Ambiguity Tracking (AT) episodes.

Question Answering Data Organization. To equip the model with real-world 3D spatial reasoning, we incorporate ScanQA [4]. We further leverage LLaVA-Pretrain [35] to enhance general visual-language understanding. To broaden and diversify representation learning across wider domains, we additionally incorporate SYNTH-PEDES [83]. Finally, we employ RefCOCO [68] and Flickr30k [49]

to jointly strengthen grounded language understanding and image captioning capabilities.

Multi-Task Navigation Data Collection. We collect multi-task navigation data from diverse sources. First, we incorporate instruction-following data from YouTube videos like NaViLA [11] and use a shortest-path follower in the Habitat simulator [50] to generate expert action sequences for R2R-CE [28] and RxR-CE [29]. Second, we collect object-goal navigation data from HM3D ObjectNav [53] and HM3D OVON [67].

4.2 Training Recipe

Our training framework adopts a principled two-phase optimization strategy: First, following established VLM alignment protocols [35], we perform supervised pre-training for only one epoch optimization of the projector, vision encoder, and LLM parameters using the curated navigation, tracking and VQA datasets. Second, we then fine-tune the model for one epoch using GRPO [56] under our multi-agent competitive training setup. In this stage, we keep the backbone largely fixed and apply LoRA [23] adapters to the LLM, which are the only trainable parameters during RL fine-tuning.

5 Experiments

5.1 Experiment Setups

We describe the evaluation benchmarks and metrics, training details for both the SFT and RL stages, and the real-world deployment platform.

Benchmarks. To substantiate the robustness and generalization capability of our framework, we perform rigorous validation across two evaluation protocols: three tasks in EVT-Bench [61], and CoMaTrack-Bench, a novel benchmark introduced in this work. In parallel, an extensive ablation study positions our method against current state-of-the-art approaches, such as IBVS [21], DiMP [6], SARL [42], AD-VAT [77], AD-VAT+ [76], TS [78], EVT [79], PoliFormer [70], Uni-NaVid [72], VLingNav [60], TrackVLA [61], NavFoM [71] and TrackVLA++ [36]. This multi-faceted comparison ensures a holistic assessment of performance across diverse algorithmic paradigms and task formulations.

Metrics. Tracking performance is quantitatively assessed using the canonical metric suite endorsed by the EVT-Bench [61]. Specifically, we report the success rate (SR), tracking rate (TR), and collision rate (CR).

SFT Training Details. CoMaTrack is supervised fine-tuned for one epoch on 96 NVIDIA H20 GPUs for about 15 hours with the full multi-task dataset. The optimizer is AdamW with a learning rate of 10^{-5} , cosine decay schedule, and weight decay of 10^{-3} . Following Uni-NaVid [72], we prepend a task indicator token $\langle \text{NAV} \rangle$ for tracking and navigation tasks. Both the LLM backbone and the vision encoder merger are trainable; the target modules cover all linear layers.

RL Training Details. The RL stage is fully on-policy, with the policy updated after every 20 rollouts. A single epoch over the EVT-Bench tracking split consists of about seven thousand rollouts, amounting to roughly one million environment steps. Only the LoRA adapters of the LLM are trainable during RL; the vision encoder and flow-matching head are frozen. We optimize with AdamW at a learning rate of 10^{-5} , a clipping ratio of $\epsilon_{\text{clip}}=0.2$, and an entropy weight of $\lambda_{\text{ent}}=0.01$. The KL weight λ_{KL} follows a curriculum that increases from 0.02 to 0.10, with an adaptive target KL of 0.01. The RL stage runs on 8 L20 GPUs.

Real World Deployment. CoMaTrack operates on a Unitree GO2 X robot equipped with four Sensing ISX031 cameras during real-world deployment. The video stream is transmitted to a remote server with an NVIDIA RTX 4090 GPU for processing, and the results are transmitted back to the local device, as shown in Fig. 3.

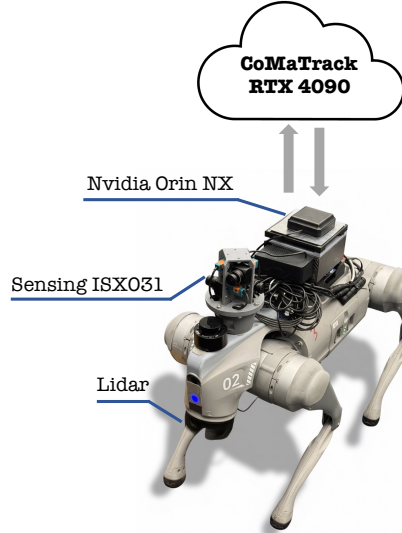


Fig. 3: Hardware Platform. Our deployment platform is built on a Unitree Go2 X quadrupedal robot equipped with four monocular RGB cameras, a Unitree 4D LiDAR L2.

5.2 Benchmark Results

We report quantitative results on EVT-Bench and CoMaTrack-Bench, followed by ablation studies that isolate the contributions of each training stage and key RL design choices.

Performance on EVT-Bench. We first evaluate our method on the public benchmark EVT-Bench [61], with results reported in Tab. 1. Compared with the strongest prior baseline TrackVLA++, CoMaTrack improves SR from 90.9 to 92.1 on STT, from 74.0 to 74.2 on DT, and from 55.9 to 57.5 on AT. In addition, CoMaTrack achieves consistently higher TR and lower CR across the three tasks, indicating improved tracking persistence and safety under distractors and ambiguity. Notably, our 3B model achieves state-of-the-art results against all existing 7B baselines, underscoring the effectiveness of the proposed multi-agent competition RL training paradigm.

Performance on CoMaTrack-Bench. As shown in Tab. 2, CoMaTrack substantially outperforms Uni-NaVid [72]—the only baseline with publicly available weights—on CoMaTrack-Bench across all metrics.

CoMaTrack-Bench is notably harder than EVT-Bench due to dynamic adversarial opponents that actively interfere through occlusion, path blocking, and

Table 1: Performance on EVT-Bench. Bold denote the best results. * means using GroundingDINO [38] as the detector. ‡ means using SoM [65] and GPT-4o [24] as the visual foundation model. Models annotated with † were fine-tuned on a 7B VLM, whereas our method is built upon a **3B VLM**.

Methods	STT			DT			AT		
	SR↑	TR↑	CR↓	SR↑	TR↑	CR↓	SR↑	TR↑	CR↓
IBVS* [21]	42.9	56.2	3.8	10.6	28.4	6.1	15.2	39.5	4.9
PoliFormer* [70]	4.7	15.5	40.1	2.6	13.2	44.5	3.0	15.4	41.5
EVT [79]	24.4	39.1	42.5	3.2	11.2	47.9	17.4	21.1	45.6
EVT† [79]	32.5	49.9	40.5	15.7	35.7	53.3	18.3	21.0	44.9
Uni-NaVid† [72]	53.3	67.2	12.6	31.9	50.1	21.3	15.8	41.5	26.5
TrackVLA† [61]	85.1	78.6	1.7	57.6	63.2	5.8	50.2	63.7	17.1
VLingNav† [60]	88.4	81.2	2.1	67.7	73.5	5.5	—	—	—
NavFoM† [71]	88.4	80.7	—	62.0	67.9	—	—	—	—
TrackVLA++† [36]	90.9	82.7	1.5	74.0	73.7	3.5	55.9	63.8	15.1
Ours	92.1	90.3	0.9	74.2	80.5	2.1	57.5	73.4	12.0

Table 2: Zero-shot Performance on CoMaTrack-Bench.

Methods	SR↑	TR↑	CR↓
Uni-Navid	42.4	56.5	23.8
Ours	85.0	82.9	5.5

Table 3: Ablation Study of Multi-Agent on EVT-Bench STT.

Methods	SR↑	TR↑	CR↓
SFT Model	88.2	85.4	3.1
Single-Agent RL	89.5	88.0	2.2
Multi-Agent RL	92.1	90.3	0.9

competitive pursuit, resulting in lower absolute metrics that reflect the increased difficulty of contested tracking.

5.3 Ablation Study of Multi Agents

To quantify the contribution of our proposed multi-agent RL competitive training paradigm, we conduct a comprehensive ablation study comparing three training configurations: (1) the SFT-only baseline, (2) single-agent RL fine-tuning, and (3) our full multi-agent RL framework. As presented in Tab. 3, each component progressively enhances performance across all metrics.

The SFT Model trained solely on expert demonstrations achieves 88.2% SR, 85.4% TR, and 3.1% CR, establishing a strong behavioral prior but lacking mechanisms for handling distribution shift or adversarial interference. Introducing Single-Agent RL improves SR to 89.5%, TR to 88.0%, and reduces CR to 2.2%, demonstrating that closed-loop policy optimization through environmental interaction yields moderate gains in robustness and safety. However, the improvements remain limited.

In contrast, Multi-Agent RL achieves 92.1% SR, 90.3% TR, and notably reduces CR to just 0.9%. These results confirm that competitive multi-agent

training, where adaptive opponents continuously escalate task difficulty through strategic interference, creates a self-reinforcing curriculum that drives the emergence of robust, anticipatory tracking behaviors. The dramatic collision reduction particularly highlights that the agent learns sophisticated spatial reasoning and safe maneuvering strategies when forced to navigate around intelligent, non-cooperative opponents—capabilities that are notably difficult to acquire through purely static demonstration data or single-agent exploration alone.

To better understand how competitive training improves robustness, we further analyze the success rate as a function of the number of multi-agent RL training episodes under the three opponent settings in CoMaTrack-Bench. As shown in Fig. 4, Static Obstacle brings only marginal gains, while Random Interference even leads to a slight performance degradation. The reason is that a random opponent injects motion that is uncorrelated with the tracker’s behavior, so it provides no directed learning signal: the resulting gradient is effectively noise and does not push the tracker toward interference-resilient strategies. In contrast, Competitive Tracking provides a directed adversarial gradient—the opponent actively exploits the tracker’s weaknesses—which yields the most informative curriculum and the best performance, confirming that stronger, strategic opponents more effectively drive policy improvement.

Effect of RL design choices. We further study the key design choices of the RL stage under the same recipe and EVT-Bench protocol, as reported in Tab. 4. Performance is stable with respect to the group size around our default of $K=10$: reducing it to $K=5$ yields a noisier advantage estimate and a small drop of 1.0 point, whereas increasing it to $K=20$ brings no further gain. The regularization terms have a larger impact. Removing the KL constraint lowers the success rate by 1.4 points, and removing the entropy bonus lowers it by 3.7 points as exploration collapses; removing both the KL constraint and the SFT anchor degrades it by 4.5 points. These results indicate that the KL term and the frozen-SFT anchor together maintain the trust region that keeps RL training stable.

5.4 Qualitative Results in Real-World

Fig. 5 presents qualitative real-world evaluations of our method under challenging conditions, including (A) similar distractors scenario, (B) obstacle scenario, and (C) dark and constrained environments. The results show that CoMaTrack

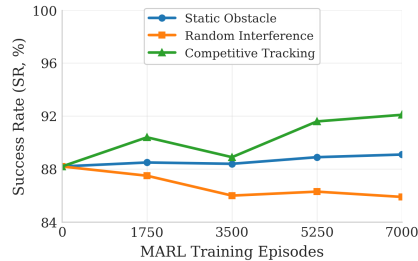


Fig. 4: Diagram Illustrating the Impact of Opponent Strength on Outcomes.

Table 4: Ablation of the multi-agent RL design choices, evaluated under the same recipe and EVT-Bench protocol. The full model corresponds to the configuration reported in Tab. 1, and ΔSR is measured relative to it.

Configuration	SR $\uparrow$	ΔSR
Full model	92.1	–
Group size $K=5$	91.1	–1.0
Group size $K=20$	92.1	+0.0
w/o KL constraint	90.7	–1.4
w/o entropy bonus	88.4	–3.7
w/o KL constraint and SFT anchor	87.6	–4.5

transfers effectively from simulation to the real world for EVT, supporting zero-shot deployment in highly dynamic settings.

6 Conclusion

This paper addresses two core challenges in EVT: the weak generalization of single-agent IL and the lack of effective RL training paradigms. We propose CoMaTrack, the first EVT training framework that integrates multi-agent competitive game-theoretic with RL. By constructing a co-evolution loop between a tracker and adaptive opponents, CoMaTrack shifts EVT training from a static, data-driven regime to an opponent-driven adversarial learning process, substantially improving tracking success. We further propose CoMaTrack-Bench, the first competitive EVT benchmark, pushing evaluation beyond idealized static settings toward realistic adversarial scenarios.

Our experiments demonstrate both effectiveness and practicality. Multi-agent strength-controlled studies show a clear step-wise performance improvement as the opponent is upgraded from a static obstacle to random interference and further to a competitive tracking agent. Under the same evaluation protocol, a 3B-parameter VLM trained with CoMaTrack significantly outperforms all prior single-agent methods relying on 7B-parameter models, indicating that competition-driven policy evolution, rather than model scale, is the key driver of robust generalization. Crucially, we successfully deploy the learned policy on a real quadruped robot, achieving stable, continuous following of a moving target in complex indoor and outdoor dynamic environments, which validates the sim-to-real transferability and engineering value of our approach.

Overall, CoMaTrack provides an efficient and transferable solution for EVT and establishes a new embodied learning paradigm of game-theoretic generalization, whose central idea naturally extends to broader VLA embodied tasks such as instruction following and object-goal navigation.

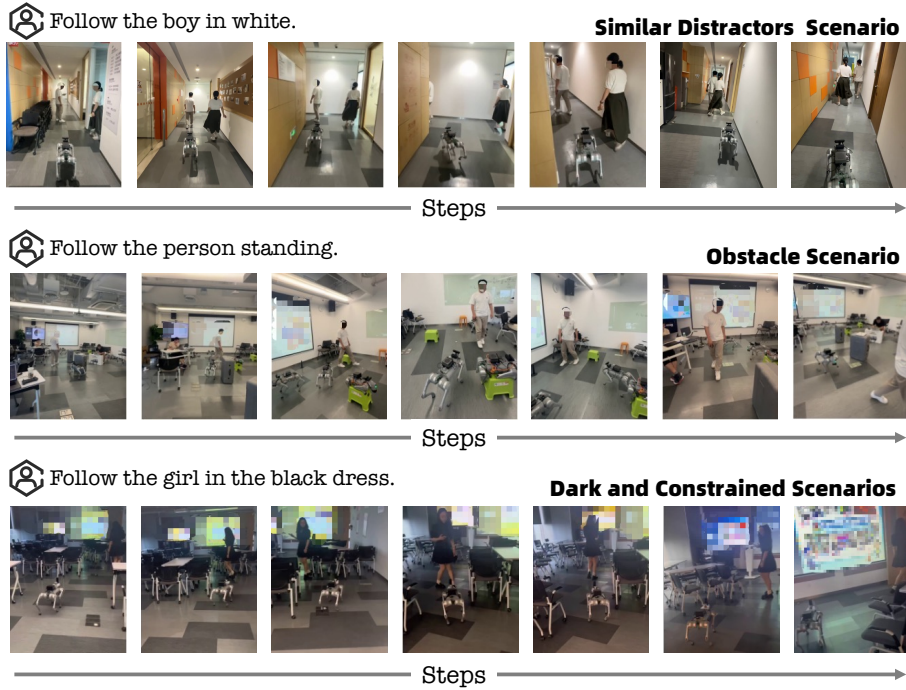


Fig. 5: Qualitative real-world results demonstrating CoMaTrack’s zero-shot deployment capabilities.

7 Limitations

While our multi-agent game-theoretic RL framework shows strong performance on EVT, several limitations remain. The current validation focuses on EVT and its competitive setting, without large-scale evaluation on broader VLN tasks such as instruction following and object navigation, limiting demonstrated task generality. Opponent strategies, though diverse, are bounded by simulated priors and may not reflect real-world dynamics, risking distribution shift. Training is computationally costly and unstable due to multi-agent non-stationarity, requiring further optimizations in sampling and stabilization. Future work will extend the framework to more tasks, enhance opponent generation, improve training efficiency, and refine evaluation to better reflect long-term performance.

Acknowledgements

We thank Youzhi Liu for his contributions to the SFT training, multi-agent reinforcement learning experiments, and the construction of the CoMaTrack-Bench opponent settings. His work on designing and implementing the competitive opponent behaviors was instrumental in validating the framework.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. An, D., Wang, H., Wang, W., Wang, Z., Huang, Y., He, K., Wang, L.: Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3674–3683 (2018)
4. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 19129–19139 (2022)
5. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
6. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 6182–6191 (2019)
7. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. arXiv preprint arXiv:2307.15818 (2024)
8. Bundele, V., Bhupati, M., Banerjee, B., Grover, A.: Scaling vision-and-language navigation with offline rl. arXiv preprint arXiv:2403.18454 (2024)
9. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
10. Chen, J., Gao, C., Meng, E., Zhang, Q., Liu, S.: Reinforced structured state-evolution for vision-language navigation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15450–15459 (2022)
11. Cheng, A.C., Ji, Y., Yang, Z., Gongye, Z., Zou, X., Kautz, J., Bıyık, E., Yin, H., Liu, S., Wang, X.: Navila: Legged robot vision-language-action model for navigation. arXiv preprint arXiv:2412.04453 (2024)
12. Choi, J., Kwon, J., Lee, K.M.: Visual tracking by reinforced decision making. arXiv preprint arXiv:1702.06291 2 (2017)
13. Dai, Y., Lee, J., Fazeli, N., Chai, J.: Racer: Rich language-guided failure recovery policies for imitation learning. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 15657–15664. IEEE (2025)
14. Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., Finn, C.: Robonet: Large-scale multi-robot learning. arXiv preprint arXiv:1910.11215 (2019)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021)
16. Gao, C., Jin, L., Peng, X., Zhang, J., Deng, Y., Li, A., Wang, H., Liu, S.: Octonav: Towards generalist embodied navigation. arXiv preprint arXiv:2506.09839 (2025)

17. Gao, L., Nie, D., Li, B., Ren, X.: Doubly-fused vit: Fuse information from vision transformer doubly with local representation. In: European Conference on Computer Vision. pp. 744–761. Springer (2022)
18. Gao, L., Zhang, J., Zhang, L., Tao, D.: Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 2825–2833 (2021)
19. Gao, P., Wang, P., Wang, F., Fujita, H., Aljuaid, H., Shang, J.L.: Deepvln: Vision-and-language navigation via deep reasoning and collaborative mechanisms based on large language models. *IEEE Journal of Selected Topics in Signal Processing* (2026)
20. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025)
21. Gupta, M., Kumar, S., Behera, L., Subramanian, V.K.: A novel vision-based tracking algorithm for a human-following mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **47**(7), 1415–1427 (2016)
22. Han, J., Liu, J., Wang, J., Peng, B., Yuan, Z.: Generative refinement networks for visual synthesis. *arXiv preprint arXiv:2604.13030* (2026)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *ICLR* **1**(2), 3 (2022)
24. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
25. Intelligence, P., Amin, A., Aniceto, R., Balakrishna, A., Black, K., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., DiCarlo, J., et al.: $\pi_{0.6}^*$: a VLA that learns from experience. *arXiv preprint arXiv:2511.14759* (2025)
26. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020)
27. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024)
28. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
29. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954* (2020)
30. Li, Z., Li, S., Zhang, Z., Li, B., Zhou, S.: Dv-vln: Dual verification for reliable llm-based vision-and-language navigation. *arXiv preprint arXiv:2601.18492* (2026)
31. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
32. Lin, B., Zhu, Y., Long, Y., Liang, X., Ye, Q., Lin, L.: Adversarial reinforced instruction attacker for robust vision-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(10), 7175–7189 (2021)
33. Lin, P., Sun, G., Liu, C., Li, F., Ren, W., Cong, Y.: Openvln: Open-world aerial vision-language navigation. *arXiv preprint arXiv:2511.06182* (2025)
34. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022)

35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
36. Liu, J., Qi, Y., Zhang, J., Li, M., Wang, S., Wu, K., Ye, H., Zhang, H., Chen, Z., Zhong, F., et al.: Trackvla++: Unleashing reasoning and memory capabilities in vla models for embodied visual tracking. *arXiv preprint arXiv:2510.07134* (2025)
37. Liu, Q., Huang, T., Zhang, Z., Tang, H.: Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884* (2025)
38. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: *European conference on computer vision*. pp. 38–55. Springer (2024)
39. Liu, Y., Yao, F., Yue, Y., Xu, G., Sun, X., Fu, K.: Navagent: Multi-scale urban street view fusion for uav embodied vision-and-language navigation (2024)
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10012–10022 (2021)
41. Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., Wang, Y.: End-to-end active object tracking and its real-world deployment via reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence* **42**(6), 1317–1332 (2019)
42. Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., Wang, Y.: End-to-end active object tracking and its real-world deployment via reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence* **42**(6), 1317–1332 (2019)
43. Maalouf, A., Jadhav, N., Jatavallabhula, K.M., Chahine, M., Vogt, D.M., Wood, R.J., Torralba, A., Rus, D.: Follow anything: Open-set detection, tracking, and following in real-time. *IEEE Robotics and Automation Letters* **9**(4), 3283–3290 (2024)
44. Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., Martín-Martín, R.: What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298* (2021)
45. Maynard Smith, J., Price, G.R.: The logic of animal conflict. *Nature* **246**(5427), 15–18 (1973)
46. Nowé, A., Vrancx, P., De Hauwere, Y.M.: Game theory and multi-agent reinforcement learning. In: *Reinforcement learning: State-of-the-art*, pp. 441–470. Springer (2012)
47. O’Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 6892–6903. IEEE (2024)
48. Peng, D., Cao, J., Zhang, Q., Ma, J.: Lovon: Legged open-vocabulary object navigator. *arXiv preprint arXiv:2507.06747* (2025)
49. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2641–2649 (2015)
50. Puig, X., Undersander, E., Szot, A., Cote, M.D., Yang, T.Y., Partsey, R., Desai, R., Clegg, A.W., Hlavac, M., Min, S.Y., et al.: Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724* (2023)

51. Qi, Z., Zhang, Z., Yu, Y., Wang, J., Zhao, H.: Vln-r1: Vision-language navigation via reinforcement fine-tuning. arXiv preprint arXiv:2506.17221 (2025)
52. Qiu, W., Zhong, F., Zhang, Y., Qiao, S., Xiao, Z., Kim, T.S., Wang, Y.: Unrealcv: Virtual worlds for computer vision. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1221–1224 (2017)
53. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)
54. Richards, B.A., Lillicrap, T.P.: Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology* **54**, 28–36 (2019)
55. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
56. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
57. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
58. Van Valen, L.: A new evolutionary law. *Evolutionary Theory* **1**, 1–30 (1973)
59. Wang, J., Wang, T., Cai, W., Xu, L., Sun, C.: Boosting efficient reinforcement learning for vision-and-language navigation with open-sourced llm. *IEEE Robotics and Automation Letters* (2024)
60. Wang, S., Luo, Y., Chen, X., Luo, A., Li, D., Liu, C., Chen, S., Zhang, Y., Yu, J.: Vlingnav: Embodied navigation with adaptive reasoning and visual-assisted linguistic memory. arXiv preprint arXiv:2601.08665 (2026)
61. Wang, S., Zhang, J., Li, M., Liu, J., Li, A., Wu, K., Zhong, F., Yu, J., Zhang, Z., Wang, H.: Trackvla: Embodied visual tracking in the wild. arXiv preprint arXiv:2505.23189 (2025)
62. Wu, K., et al.: Hierarchical instruction-aware embodied visual tracking. arXiv preprint arXiv:2505.20710 (2025)
63. Wu, K., et al.: Vlm can be a good assistant: Enhancing embodied visual tracking with self-improving vision-language models. In: IROS (2025)
64. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2.5 technical report. arXiv e-prints pp. arXiv-2412 (2024)
65. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
66. Ye, H., Zhan, Y., Situ, W., Chen, G., Yu, J., Zhao, Z., Cai, K., Ajoudani, A., Zhang, H.: Tpt-bench: A large-scale, long-term and robot-egocentric dataset for benchmarking target person tracking. arXiv preprint arXiv:2505.07446 (2025)
67. Yokoyama, N., Ramrakhya, R., Das, A., Batra, D., Ha, S.: Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5543–5550. IEEE (2024)
68. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023)
69. Yue, L., Fan, Y., Lian, S., Zhao, Y., Yu, J., Xie, L., Zhang, F.: Spatial-vln: Zero-shot vision-and-language navigation with explicit spatial perception and exploration. arXiv preprint arXiv:2601.12766 (2026)

70. Zeng, K.H., Zhang, Z., Ehsani, K., Hendrix, R., Salvador, J., Herrasti, A., Girshick, R., Kembhavi, A., Weihs, L.: Poliformer: Scaling on-policy rl with transformers results in masterful navigators. arXiv preprint arXiv:2406.20083 (2024)
71. Zhang, J., Li, A., Qi, Y., Li, M., Liu, J., Wang, S., Liu, H., Zhou, G., Wu, Y., Li, X., et al.: Embodied navigation foundation model. arXiv preprint arXiv:2509.12129 (2025)
72. Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. arXiv preprint arXiv:2412.06224 (2024)
73. Zhang, W., Song, K., Rong, X., Li, Y.: Coarse-to-fine uav target tracking with deep reinforcement learning. *IEEE Transactions on Automation Science and Engineering* **16**(4), 1522–1530 (2018)
74. Zhang, Z., Zhu, W., Pan, H., Wang, X., Xu, R., Sun, X., Zheng, F.: Activevln: Towards active exploration via multi-turn rl in vision-and-language navigation. arXiv preprint arXiv:2509.12618 (2025)
75. Zhong, F., Bi, X., Zhang, Y., Zhang, W., Wang, Y.: Rspt: reconstruct surroundings and predict trajectory for generalizable active object tracking. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 37, pp. 3705–3714 (2023)
76. Zhong, F., Sun, P., Luo, W., Yan, T., Wang, Y.: Ad-vat+: An asymmetric dueling mechanism for learning and understanding visual active tracking. *IEEE transactions on pattern analysis and machine intelligence* **43**(5), 1467–1482 (2019)
77. Zhong, F., Sun, P., Luo, W., Yan, T., Wang, Y.: Ad-vat: An asymmetric dueling mechanism for learning visual active tracking. In: *International Conference on Learning Representations* (2019)
78. Zhong, F., Sun, P., Luo, W., Yan, T., Wang, Y.: Towards distraction-robust active visual tracking. In: *International Conference on Machine Learning*. pp. 12782–12792. PMLR (2021)
79. Zhong, F., Wu, K., Ci, H., Wang, C., Chen, H.: Empowering embodied visual tracking with visual foundation models and offline rl. In: *European Conference on Computer Vision*. pp. 139–155. Springer (2024)
80. Zhong, F., Wu, K., Ci, H., Wang, C., Chen, H.: Empowering embodied visual tracking with visual foundation models and offline rl. In: *European Conference on Computer Vision*. pp. 139–155. Springer (2024)
81. Zhong, F., et al.: Unrealzoo: Enriching photo-realistic virtual worlds for embodied ai. In: *ICCV* (2025)
82. Zhou, Y., Kantarcioglu, M., Xi, B.: A survey of game theoretic approach for adversarial machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**(3), e1259 (2019)
83. Zuo, J., Hong, J., Zhang, F., Yu, C., Zhou, H., Gao, C., Sang, N., Wang, J.: Plip: Language-image pre-training for person representation learning. *Advances in Neural Information Processing Systems* **37**, 45666–45702 (2024)