2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

TrackVLA++: Unleashing Reasoning and Memory Capabilities in VLA Models for Embodied Visual Tracking

Jiahang Liu[1,2,*] Yunpeng Qi[3,4,*] Jiazhao Zhang[1,2,*] Minghan Li[2] Shaoan Wang[1] Kui Wu[5] Hanjing Ye[6]
Hong Zhang[6] Zhibo Chen[3] Fangwei Zhong[7] Zhizheng Zhang[2,4,†] He Wang[1,2,4,†]
[1]Peking University [2]Galbot [3]USTC [4]BAAI [5]Beihang University [6]SUSTech [7]Beijing Normal University
Project Page: https://pku-epic.github.io/TrackVLA-plus-plus-Web/

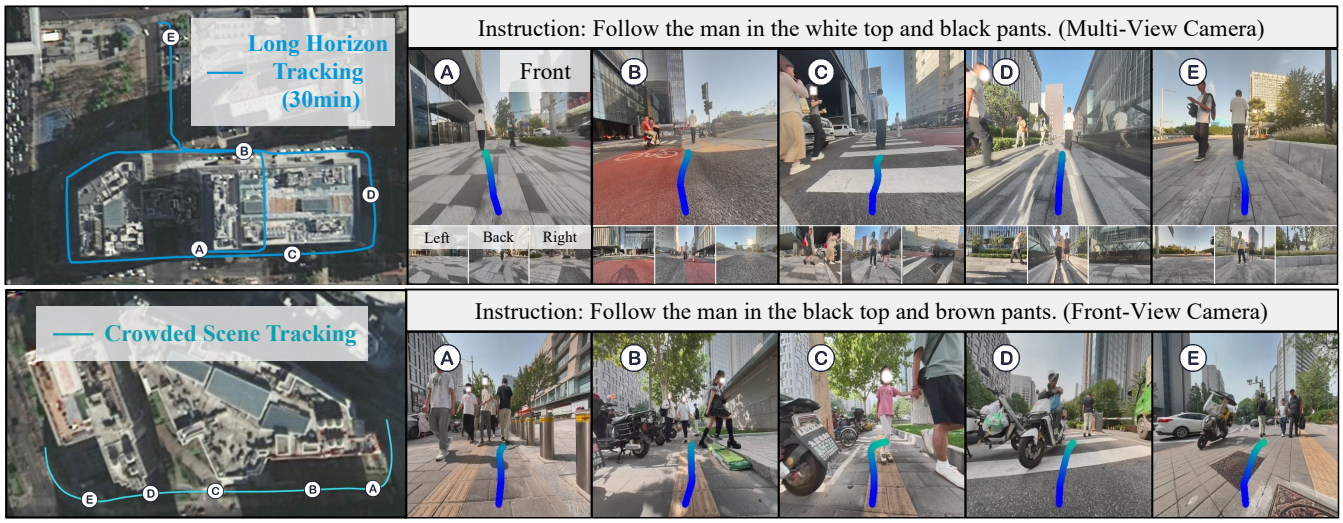

Fig. 1: **Real-world demonstration of TrackVLA++**. TrackVLA++ is a novel Vision-Language-Action model that incorporates spatial reasoning and target identification memory, enabling superior performance in both long-horizon and highly crowded tracking scenarios.

***Abstract*— Embodied Visual Tracking (EVT) is a fundamental ability that underpins practical applications, such as companion robots, guidance robots and service assistants, where continuously following moving targets is essential. Recent advances have enabled language-guided tracking in complex and unstructured scenes. However, existing approaches lack explicit spatial reasoning and effective temporal memory, causing failures under severe occlusions or in the presence of similar-looking distractors. To address these challenges, we present TrackVLA++, a novel Vision–Language–Action (VLA) model that enhances embodied visual tracking with two key modules: a spatial reasoning mechanism and a Target Identification Memory (TIM). The reasoning module introduces a Chain-of-Thought paradigm, termed Polar-CoT, which infers the target's relative position and encodes it as a compact polar-coordinate token for action prediction. Guided by these spatial priors, the TIM employs a gated update strategy to preserve long-horizon target memory, ensuring spatiotemporal consistency and mitigating target loss during extended occlusions. Extensive experiments show that TrackVLA++ achieves state-of-the-art performance on public benchmarks across both egocentric and multi-camera settings. On the challenging EVT-Bench DT split, TrackVLA++ surpasses the previous leading approach by 5.1% and 12% respectively. Furthermore, TrackVLA++ exhibits strong zero-shot generalization, enabling robust real-world tracking in dynamic and occluded scenarios.**

* Equal Contribution, † Equal Advising

I. Introduction

Embodied Visual Tracking (EVT) is a fundamental yet challenging task, where an agent navigates in dynamic physical environments and continuously track a specified moving target based on visual perception. Recent methods have shown remarkable progress in this task [1]–[6]. Recent advancements in EVT increasingly leverage the powerful generalization capability of pre-trained Visual Foundation Models (VFMs) [7]–[9] to enhance target identification from visual inputs. Building on this perceptual foundation, agents employ policy learning techniques, such as imitation learning [10] or reinforcement learning [3], [6], [11], to generate actions that enable effective target pursuit.

More recently, leveraging large language models (LLMs) has introduced a promising new paradigm for the EVT task. Pioneering works, notably TrackVLA [12] and LOVON [13], exemplify this trend by integrating powerful Vision-Language Models (VLMs) to handle complex, language-guided tracking tasks. TrackVLA, for instance, introduces a unified, end-to-end Vision-Language-Action (VLA) framework that learns a holistic tracking policy. It processes visual-language inputs using a VLM, with the latent representations decoded into tracking trajectories through an anchor-based diffusion policy. This design not only demonstrates strong

979-8-3315-8160-2/26/$31.00 ©2026 IEEE

sim-to-real generalization and real-time performance but also benefits from the tight coupling of perception and planning, which effectively mitigates the information loss and error propagation inherent in decoupled pipelines. In contrast, LOVON adopts a hierarchical strategy, using LLM as a high-level planner to decompose instructions into simpler sub-tasks, which are then executed by a low-level motion model to predict immediate tracking actions. Despite their advancements, these state-of-the-art (SOTA) methods lack explicit reasoning capability and robust mechanism for long-horizon target identification.As a result, performance degrades in complex, unstructured scenes, especially under severe occlusion or multiple similar distractors.

To address these challenges, we propose TrackVLA++, a novel VLA framework for the EVT task that is empowered with explicit spatial reasoning capability and effective temporal memory to enable long-horizon target identification. At the core of our approach is the Polar Chain-of-Thought (Polar-CoT) mechanism, which enables spatial reasoning by inferring the target's relative position, expressed as angle and distance in agent-centric polar coordinate system. In contrast to prior CoT mechanisms in robot manipulation, which generate verbose textual plans or auxiliary visual intermediates (*e.g.,* bounding boxes or subgoal images) [14]–[17], our Polar-CoT introduces a compact design that maintains inference efficiency by predicting only **one** reasoning token, which serves as the basis for the Target Identification Memory (TIM) module. TIM is specifically designed to preserve a persistent and robust representation of the target's visual identity over long horizons, even under challenging conditions such as prolonged occlusions. To this end, TIM employs a confidence-aware gating mechanism that strictly regulates memory updates: the memory state is refreshed only when Polar-CoT predicts the target's presence with high confidence. During each update, TIM integrates its historical state with newly extracted visual features from the region specified by Polar-CoT's spatial prediction, where the contribution of new observations is weighted in proportion to the confidence score. Furthermore, all the aforementioned techniques naturally extend to multi-view settings, where they not only retain compatibility but also deliver enhanced tracking performance.

We conducted extensive experiments to evaluate the effectiveness and generalization ability of TrackVLA++ across both simulated benchmarks and real-world scenarios. Our method achieves SOTA performance in both egocentric and multi-camera settings. Specifically, on the highly challenging EVT-Benchmark [12] `DT split`, TrackVLA++ outperforms previous leading methods by 5.1% and 12% in success rate for egocentric and multi-camera settings, respectively. Additionally, TrackVLA++ accomplishes new SOTA results on the Gym-UnrealCV benchmark [18], which further demonstrates its superiority over existing methods. Beyond these benchmarks, TrackVLA++ exhibits remarkable zero-shot generalization, demonstrating robust performance in real-world environments, as highlighted in Fig. 1, Fig. 5 and our supplementary video. The contributions of this work can be summarized as follows:

- We propose a novel Polar-CoT mechanism for the EVT task, which equips the model with explicit spatial reasoning capability, achieving significant performance improvements while maintaining computational efficiency.
- We propose the Target Identification Memory (TIM), a robust module for long-horizon target identification that leverages reasoning guided memory update to achieve resilience against severe occlusions and distractors.
- We conduct extensive evaluations, showing that TrackVLA++ achieves state-of-the-art performance across multiple simulation benchmarks and demonstrates remarkable generalization to real-world scenarios.

II. Related Works

Vision-Language-Action Models. The paradigm of extending pre-trained Vision-Language Models (VLMs) [19]–[21] with action-generation capabilities has established Vision-Language-Action (VLA) models as a cornerstone of modern embodied AI. This approach has yielded significant success in manipulation [14], [22]–[26] and navigation [10], [27], [28]. Recently, the VLA paradigm was extended to the dynamic task of Embodied Visual Tracking (EVT), with models like TrackVLA [12] achieving impressive results. In this work, we propose TrackVLA++, which enhances its predecessor with reasoning ability and long-horizon memory.
Embodied Visual Tracking (EVT) [29]–[31] requires an agent to continuously pursue a dynamic target based on its visual observations, relying on accurate target recognition and optimal trajectory planning. Early works [6], [11], [32]–[39] adopted a decoupled paradigm, pairing visual foundation models [7] for perception with reinforcement learning for planning. Recently, the field has shifted towards end-to-end VLA models to support natural language inputs [10], [12], [13]. Uni-NaVid [10] pioneered this direction with large-scale imitation learning, though its discrete action space limited real-world adaptability. Building on this, TrackVLA [12] made significant advances by integrating recognition and planning into unified frameworks, showing strong performance in real-world tracking tasks. Similarly, LOVON [13] employs a hierarchical approach, where a high-level LLM planner breaks complex instructions into simpler sub-goals, executed by a low-level controller for navigation and tracking. Despite their success, both models still lack explicit reasoning capabilities and robust long-horizon target identification. In this work, we introduce TrackVLA++, a novel framework that enhances embodied visual tracking by incorporating a reasoning module and target identification memory.
Chain-of-Thought Reasoning for Embodied AI. Chain-of-Thought (CoT) reasoning, which prompts models to think step-by-step, has proven effective for complex tasks [40] and is increasingly adopted in VLA models to enhance reasoning and generalization ability [14]–[16], [41], [42]. A common strategy in these works, primarily focusing on robotic manipulation, is to generate explicit and computationally intensive intermediate representations (*e.g.,* such as high-level plans,

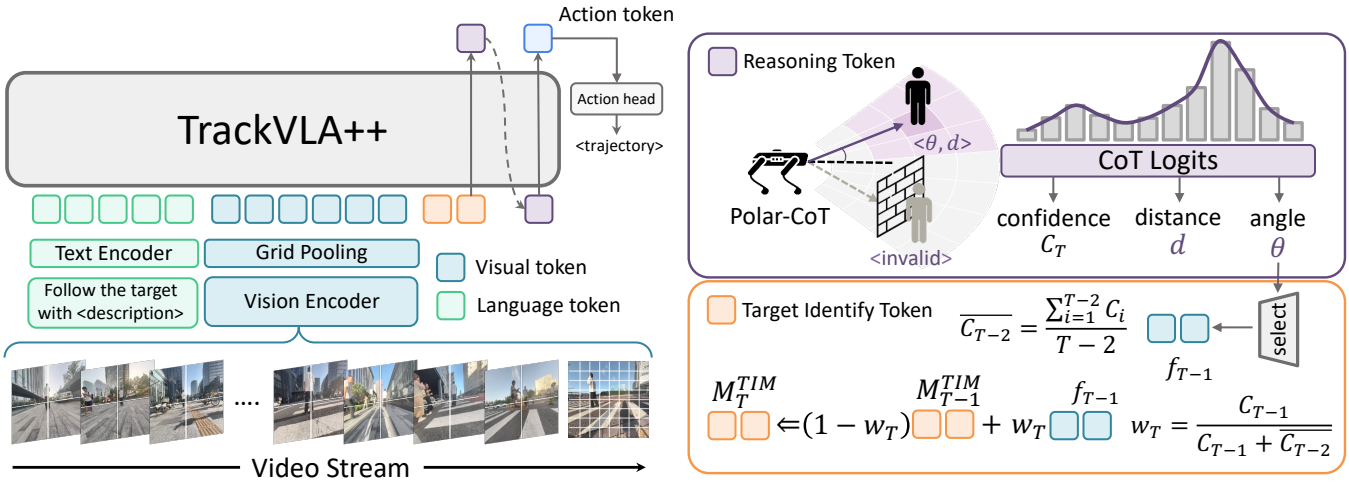

Fig. 2: **The pipeline of TrackVLA++.** Given a video stream and a language instruction, TrackVLA++ predicts a tracking trajectory by utilizing Polar-CoT reasoning to infer the target's position and continuously updating the Target Identification Memory with CoT-based predictions for long-horizon tracking.

object coordinates, or subgoal images) as prerequisites for final actions. These can include high-level textual plans, object bounding boxes, grasping coordinates, subgoal images, or coarse-grained discrete directions [14]–[17]. While effective for manipulation tasks, these approaches can introduce significant inference overhead, making them unsuitable for highly dynamic scenarios like EVT. In contrast, our method introduces an efficient CoT process especially designed to satisfy the dynamic demands of EVT, achieving robust reasoning while maintaining high inference efficiency.

III. Overview

Task Formulation. The task of Embodied Visual Tracking (EVT) can be formulated as: At each time step T, given a language description $\mathcal{L}$ of the target object and a set of on-the-fly captured RGB observations $\{\mathcal{O}_T^N \mid t = 1, \ldots, T,\ n = 1, \ldots, N\}$ from N cameras, the agent is required to predict a continuous tracking trajectory $\mathcal{W}_T = \{w_1, w_2, \ldots\}$. Each waypoint $w_i = (x, y, \theta) \in \mathbb{R}^3$ defines a target displacement (x, y) and a heading change θ within the agent's egocentric coordinate. The task is deemed successful if the agent maintains a predefined following distance D from the target.

Model Overview. As shown in Fig. 2, TrackVLA++ is an end-to-end VLA model built upon the navigation foundation model NavFoM [43]. To enhance tracking intelligence, TrackVLA++ introduces two key improvements: a CoT-based spatial reasoning mechanism Polar-CoT and a long-horizon Target Identification Memory (TIM). Given an online-captured video stream, TrackVLA++ extracts visual features from historical and current observations and predicts the reasoning token through the proposed Polar-CoT mechanism. Based on this prediction, the TIM tokens are adaptively updated to maintain a robust representation of the target's identity over time. The reasoning token, updated TIM tokens, visual tokens and language tokens are then concatenated to form the input sequence for the large language model (implemented with Qwen2-7B [44]). Leveraging this comprehensive context, the model predicts an action token, which is finally decoded by a MLP-based action head to predict the tracking trajectory.

IV. Architecture

A. TrackVLA++ Architecture

Observation Encoding. We process the on-the-fly video stream $\mathcal{O}_{1:T}^{1:N}$ by a dual-encoder architecture, extracting and concatenating visual features $\{V_t^n | t = 1, ..., T, n = 1, ..., N\}$ from SigLIP [45] and DINOv2 [46]. To mitigate the computational cost of long-horizon inputs, we then apply the grid pooling strategy [27], generating a different resolution representation: $V^{\text{fine}} \in \mathbb{R}^{64 \times C}$, which consists of high-resolution features for the fine-grained details of the current observation and low-resolution features $V^{\text{coarse}} \in \mathbb{R}^{4 \times C}$ summarizing the coarse-grained historical context, where C denotes the channel dimension.

To effectively manage the trade-off between long-range context and inference speed, our model employs a dual-memory architecture. For long-term memory, we introduce a fixed-size Target Identification Memory (TIM) to represent the target's history concisely. For short-term memory, we preserve the sliding window approach from TrackVLA, utilizing $k = 32$ frames to form the current visual feature sequence, $V_T^{\text{track}} = \{V_{T-k}^{\text{coarse}}, \ldots, V_{T-1}^{\text{coarse}}, V_T^{\text{fine}}\}$. The short-term visual sequence V_T^{track} and the long-horizon TIM features M_T^{TIM} are jointly projected into the LLM's latent space by a 2-layer MLP projector $\mathcal{P}(\cdot)$:

$$E_T^V = \mathcal{P}(V_T^{track}), \quad E_T^M = \mathcal{P}(M_T^{TIM}), \tag{1}$$

Polar-CoT Reasoning Forwarding. To equip the model with spatial reasoning capability, we propose a Polar Chain-of-Thought (Polar-CoT) mechanism for embodied visual tracking. Unlike conventional CoT approaches that rely on multi-step reasoning such as bounding box prediction, Polar-CoT adopts a lightweight, agent-centric design based on

polar coordinates. In contrast to bounding box-based methods, which often suffer from computational inefficiency and ambiguity—especially in multi-camera settings with overlapping FoVs—Polar-CoT avoids redundant or conflicting predictions and enables more consistent spatial reasoning.

As demonstrated in Fig. 2, Polar-CoT discretizes the annular FoV of the agent into structured sectors, where each sector is uniquely identified by a quantized combination of relative angle (θ) and distance (d). This discrete combination is then encoded as a unique vocabulary token, forming a compact and unified spatial representation. This representation naturally supports multi-camera setups by avoiding bounding box prediction, eliminating ambiguity and enabling consistent spatial reasoning across views.

The reasoning process is structured as follows. First, the projected visual embeddings (E_T^V) and long-term memory embeddings (E_T^M) are concatenated with the language tokens (E^L) to form the input sequence for the LLM. The model then generates a reasoning token, E_T^{CoT}, which encodes the target's spatial information (direction and proximity) into a compact representation. To further enhance robustness, the vocabulary is extended with a dedicated `<invalid>` token, allowing the model to explicitly signal when the target is occluded or outside the agent's FoV. This reasoning process is formally defined as:

$$E_T^{\text{CoT}} = \text{LLM}(\text{Concat}[E_T^M, E_T^V, E^L]), \tag{2}$$

Reasoning Feedback Memory Update. To maintain a stable target identity across occlusions, we introduce the Target Identification Memory (TIM), a confidence-gated mechanism that prevents memory corruption from distractors or drift during target absence. At each timestep T, the TIM state M_T^{TIM} is updated from its previous state M_{T-1}^{TIM} via a weighted average with a new candidate feature f_{T-1}:

$$M_T^{\text{TIM}} = (1 - w_T) \cdot M_{T-1}^{\text{TIM}} + w_T \cdot f_{T-1}, \tag{3}$$

where the candidate feature f_{T-1} represents the visual embedding from the predicted target view, identified from fine-grained features V_{T-1}^{fine} by the reasoning token E_{T-1}^{CoT}. An `<invalid>` token signifies that the target is occluded or absent.

The weight w_T modulates the update based on prediction certainty. It is formulated by normalizing the confidence score C_{T-1} against the historical average confidence $\overline{C_{T-2}}$:

$$w_T = \frac{C_{T-1}}{\overline{C_{T-2}} + C_{T-1}}, \quad \text{with} \quad \overline{C_{T-2}} = \frac{1}{T-2} \sum_{i=1}^{T-2} C_i, \tag{4}$$

The confidence score C_{T-1} itself quantifies the certainty of the reasoning token E_{T-1}^{CoT} and is calculated using the normalized entropy of its logits $\mathbf{P}$:

$$C_{T-1} = 1 - \frac{H(\text{softmax}(\mathbf{P}))}{\log K}, \tag{5}$$

where $H(p) = -\sum p_i \log p_i$ is the entropy over the K-sized reasoning vocabulary. Consequently, a confident, one-hot-like distribution yields $C_{T-1} \approx 1$ and a larger weight w_T, while an uncertain distribution results in $C_{T-1} \approx 0$, effectively suppressing the memory update.

The TIM is initialized to a null state ($M_1^{\text{TIM}} = \emptyset$) and adopts the first valid feature f_1 as its state at $T = 2$. Subsequently, the update process is governed by confidence: a high score ($C_{T-1} \to 1$) allows the memory to integrate the new feature f_{T-1}, whereas a low score ($C_{T-1} \to 0$) preserves the previous state M_{T-1}^{TIM}. Crucially, an `<invalid>` token at timestep t forces its confidence C_t to zero. This freezes the memory during the next update at $T = t + 1$, thereby preserving the last reliable representation until the target is confidently re-detected and ensuring robust long-term tracking.

Action Forwarding. After generating the reasoning token E_T^{CoT} and updating the TIM memory M_T^{TIM}, the model predicts an action token E_T^{pred}. E_T^{pred} is then decoded by an MLP-based action head into a sequence of waypoints $\mathcal{W}_T$. The action prediction process is formally defined as:

$$E_T^{\text{pred}} = \text{LLM}(\text{Concat}[E_T^M, E_T^V, E^L, E_T^{\text{CoT}}]), \tag{6}$$

$$\mathcal{W}_T = \text{ActionHead}(E_T^{\text{pred}}), \tag{7}$$

The overall training objective is defined as a weighted sum of three loss terms: the trajectory planning loss $\mathcal{L}_{\text{traj}}$, reasoning loss $\mathcal{L}_{\text{reason}}$, and vanilla text prediction loss $\mathcal{L}_{\text{text}}$:

$$\mathcal{L} = \mathcal{L}_{\text{traj}} + \alpha \mathcal{L}_{\text{reason}} + \beta \mathcal{L}_{\text{text}}, \tag{8}$$

where α and β are balancing hyperparameters, empirically set to 0.2 and 0.5, respectively. $\mathcal{L}_{\text{traj}}$ is defined as the Mean Squared Error (MSE) between the predicted waypoints $\hat{w}_i$ and the ground truth waypoints w_i^{gt}:

$$\mathcal{L}_{\text{traj}} = \sum_{i=1}^{M} \text{MSE}(\hat{w}_i, w_i^{\text{gt}}), \tag{9}$$

where M denotes the number of waypoints to predict and $\hat{w}_i$ and w_i^{gt} denote the predicted and ground truth trajectory waypoints, respectively. $\mathcal{L}_{\text{reason}}$ is formulated as the log-likelihood term over the reasoning token E_T^{CoT}, conditioned on the concatenated inputs:

$$\mathcal{L}_{\text{reason}} = -\log \mathbf{P}(E_T^{\text{CoT}} \mid \text{Concat}[E_T^M, E_T^V, E^L]). \tag{10}$$

In alignment with the established practices from VLM training [47], the model is trained for a single epoch on the combined dataset, as detailed in Sec. IV-B.

B. Dataset Construction

Polar-CoT Tracking Data Collection. We constructed a one-million-sample multi-view embodied visual tracking dataset from the EVT-Bench [12] training split using the Habitat 3.0 [32] simulator. Each sample contains multi-view RGB tracking history, a target description, Polar-CoT annotations, and the expert trajectory $\mathcal{W}_{\text{gt}}$. Polar-CoT annotations were generated by recording the target's relative angle (θ) and distance (d) to the robot at each timestep. Semantic masks were extracted from all views, and samples with fewer than 2,500 target pixels were marked as `<invalid>` to

TABLE I: **Performance on EVT-Bench.** The evaluation metrics are defined as follows: **Success Rate (SR)**, the proportion of episodes that the agent ends correctly oriented within 1–3m of the target; **Tracking Rate (TR)**, the proportion of timesteps with successful target tracking; and **Collision Rate (CR)**, the proportion of episodes terminated due to collisions. †: Uses GroundingDINO as the detector. ‡: Uses SoM [48] + GPT-4o [49] as the visual foundation model. **Bold** and underline denote the best and second-best results, respectively.

Methods	*Single-Target Tracking (STT)*			*Distracted Tracking (DT)*			*Ambiguity Tracking (AT)*		
	SR↑	TR↑	CR↓	SR↑	TR↑	CR↓	SR↑	TR↑	CR↓
IBVS† [50]	42.9	56.2	3.75	10.6	28.4	6.14	15.2	39.5	4.90
PoliFormer† [35]	4.67	15.5	40.1	2.62	13.2	44.5	3.04	15.4	41.5
EVT [6]	24.4	39.1	42.5	3.23	11.2	47.9	17.4	21.1	45.6
EVT‡ [6]	32.5	49.9	40.5	15.7	35.7	53.3	18.3	21.0	44.9
Uni-NaVid [10]	25.7	39.5	41.9	11.3	27.4	43.5	8.26	28.6	43.7
TrackVLA [12]	85.1	78.6	**1.65**	57.6	63.2	5.80	50.2	**63.7**	17.1
NavFoM [43] (Single view)	85.0	80.5	-	61.4	68.2	-	-	-	-
Ours (single view)	**86.0**	**81.0**	2.10	**66.5**	**68.8**	**4.71**	**51.2**	63.4	**15.9**
NavFoM [43] (Four views)	88.4	80.7	-	62.0	67.9	-	-	-	-
Ours(Four views)	**90.9**	**82.7**	**1.50**	**74.0**	**73.7**	**3.51**	**55.9**	**63.8**	**15.1**

indicate occlusion or excessive distance. To improve generalization, we randomized camera parameters (position, height, FoV) and augmented view configurations by always retaining the front camera while randomly sampling additional views.

QA Data Organization. In line with the TrackVLA [12], we co-trained the model by balancing tracking data with question-answering (QA) data in a 1:1 ratio. This approach was designed to enhance the model's open-world recognition capabilities. Specifically, we incorporated 294K person identification samples from SYNTH-PEDES [51], 205K image-based QA samples, and 501K video-based QA samples from publicly available datasets [19], [47]. In total, the QA data contributed one million samples, bringing the combined training dataset to two million samples. This comprehensive dataset enables the model to effectively integrate trajectory tracking and open-world recognition capability.

V. Experiments

In this section, we present a series of experiments designed to answer the following questions:

- How does TrackVLA++ perform in comparison to SOTA EVT models?
- What is the practical performance and robustness of TrackVLA++ in challenging, real-world scenarios?
- What are the individual contributions of our core components: the Polar-CoT mechanism and the TIM module, to the overall performance?

A. Experiment Setups

Benchmarks. We evaluate our method using the EVT-Bench [12] and Gym-UnrealCV [18] benchmarks. EVT-Bench is a comprehensive benchmark for embodied tracking in complex indoor scenes with lots of distractors, including visually identical appearances and ambiguous instructions. Gym-UnrealCV evaluation focuses on tracking in unseen, high-fidelity environments, providing a robust test for generalization. Additionally, we utilize the visual recognition benchmark from [12] to evaluate fine-grained, zero-shot recognition accuracy and efficiency.

Metrics. To evaluate tracking performance, we use the standard evaluation metrics from Gym-UnrealCV [18] and

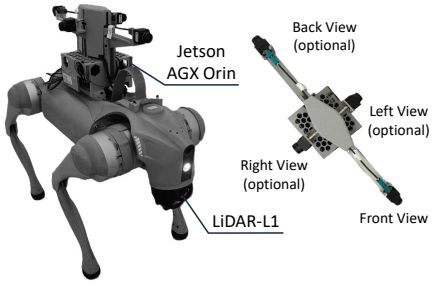

Fig. 3: **Real-world system architecture.**

EVT-Bench, including success rate (SR), average episode length (EL), tracking rate (TR), and collision rate (CR).

Implementation Details. TrackVLA++ is built upon NavFoM [43], with the Polar-CoT module discretizing the agent's perceivable space (an annular region between 0.6m and 5.0m) into 60 angular and 30 distance slices, each represented as a unique special token. The TIM state M_t^{TIM} is encoded by 4 tokens, while the predicted trajectory $\mathcal{W}_t$ comprises 8 future waypoints. The model is trained on 8 NVIDIA H100 GPUs for about one day, resulting in a total of 192 GPU hours. For deployment, as illustrated in Fig. 3, TrackVLA++ operates on a Unitree GO2 quadruped robot equipped with four SG3S11AFxK cameras for multi-view RGB streaming. The video stream is sent to a remote server with an NVIDIA RTX 4090 GPU for processing, where Polar-CoT tokens and trajectory waypoints are generated.

B. Simulation Benchmark Results

Performance on EVT-Bench. As shown in Table I and Fig. 4, we first evaluate our method on the challenging EVT-Bench benchmark. TrackVLA++ demonstrates substantial improvements over existing approaches across all three subtasks in both egocentric and multi-view camera settings, establishing a new SOTA. Notably, TrackVLA++ achieves particularly strong gains in the most challenging categories. For example, on the `DT` (Distracted Tracking) task, TrackVLA++ improves the Success Rate (SR) to 74.0%, representing a significant leap from the 62.0% achieved by NavFoM. The notable improvements in all metrics highlight the strengths of TrackVLA++ in robust recognition, long-horizon following and effective collision avoidance. Impor-

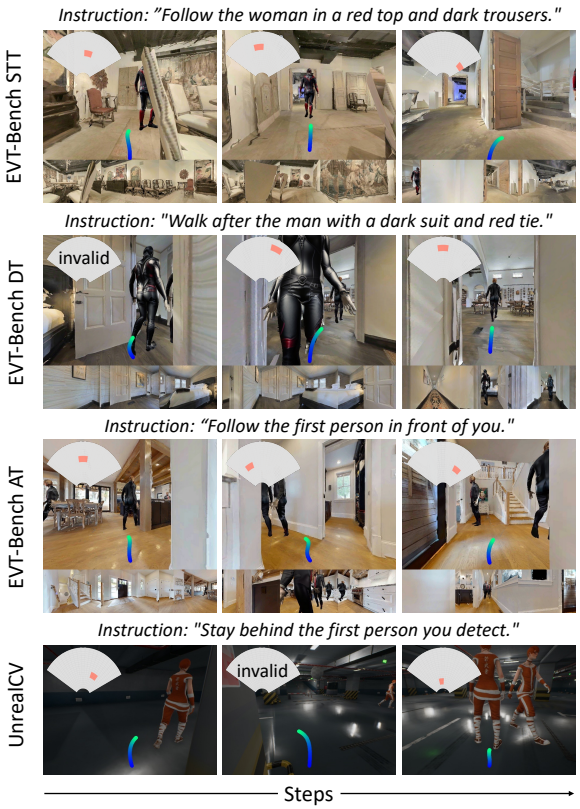

Fig. 4: **Visualizations of the Simulation Experiments.** TrackVLA++ performs well under occlusion and interference conditions. The upper-left inset displays the Polar-CoT prediction, with the red area indicating the predicted target position, and the visualization on EVT-Bench is cropped to a front sector for conciseness. Zoom in for a better view.

tantly, despite NavFoM being trained on a massive dataset of 10 million trajectories, TrackVLA++ achieves superior performance with significantly less training data, underscoring its data efficiency and advanced modular design.

Zero-shot performance on Gym-UnrealCV. Beyond EVT-Bench, we evaluate the model's generalization ability on the Gym-UnrealCV benchmark in a zero-shot manner, using a front-view camera for fair comparison. As shown in Table II and Fig.4, TrackVLA++ achieves SOTA performance across all sub-tasks. In the `Single Target` and `Unseen Objects` categories, our method, like TrackVLA, achieves the perfect scores (EL=500, SR=1.00), successfully tracking the target for the maximum episode duration. Crucially, in the more challenging `Distractor` task, where the agent must differentiate the target from identical distractors, TrackVLA++ outperforms the previous best method, TrackVLA, with a higher SR and longer EL.

Performance on Visual Recognition. To further evaluate the fine-grained recognition ability of TrackVLA++, we compare it with SOTA VLMs and tracking VLAs [12], [49], [52], [53] on a zero-shot human recognition task involving distinguishing between two unseen human images from the SYNTH-PEDES dataset. As shown in Table III, TrackVLA++ achieves a SOTA accuracy of 87.5%, outperforming

TABLE II: **Zero-shot Performance on Gym-UnrealCV.** The evaluation metrics are defined as follows: **Episode Length (EL)**, the average number of steps before episode termination (maximum is 500); and **Success Rate (SR)**, the proportion of episodes completed for the full 500-step duration. †: TrackVLA++ evaluated using only a single front-view camera for fair comparison. **Bold** and underline denote the best and second-best results, respectively.

Methods	Single Target		Distractor		Unseen Objects	
	EL↑	SR↑	EL↑	SR↑	EL↑	SR↑
DiMP [54]	367	0.58	309	0.27	-	-
SARL [33]	394	0.57	240	0.14	-	-
AD-VAT [3]	416	0.62	220	0.12	-	-
AD-VAT+ [55]	454	0.76	224	0.12	-	-
TS [36]	474	0.86	371	0.48	-	-
EVT [6]	490	0.95	459	0.81	480	0.96
TrackVLA [12]	**500**	**1.00**	474	0.91	**500**	**1.00**
Ours†	**500**	**1.00**	**484**	**0.92**	**500**	**1.00**

TABLE III: **Comparison of Different Methods on Recognition Ability.** †: Evaluation is restricted to the front-view setting for fair comparison.

Methods	ACC (%) ↑	FPS ↑
RexSeek [52]	54.3	1.1
LISA++ [53]	78.2	0.6
SoM [48]+GPT-4o [49]	82.4	0.1
TrackVLA [12]	80.7	10
NavFoM [43]	84	5.1
Ours† w/o Polar-CoT	83	5.2
Ours†	**87.5**	4.8

strong baselines such as SoM + GPT-4o (82.4%), TrackVLA (80.7%), and NavFoM (84.0%).

In terms of computational efficiency, TrackVLA++ maintains an inference speed of 4.8 FPS, which is comparable to NavFoM (5.1 FPS) and approximately **48× faster** than GPT-based baselines (SoM + GPT-4o). Despite a slight decrease in speed due to the Polar-CoT module (4.8 FPS vs. 5.2 FPS without Polar-CoT), it delivers a notable improvement in recognition accuracy (87.5% vs. 83.0%). This demonstrates the effectiveness of the Polar-CoT module in enhancing the model's reasoning capabilities while maintaining a strong balance between accuracy and efficiency.

C. Real World Results

We evaluated TrackVLA++ in three challenging real-world scenarios, with quantitative results shown in Fig. 5: (A) **Obstacle**: The target is temporarily occluded by large obstacles, testing the model's robustness to target disappearance and its ability to re-identify the target. (B) **Winding Path**: The target follows a complex, winding trajectory, evaluating the tracking fidelity amidst continuous changes in direction. (C) **Distractor**: The target is challenged by a human distractor, which serves to evaluate the model's robustness in recognition and the ability to recover from interference.

Across these tasks, TrackVLA++ outperforms TrackVLA by 14%, 7%, and 17% respectively, demonstrating substantially improved robustness in real-world conditions.

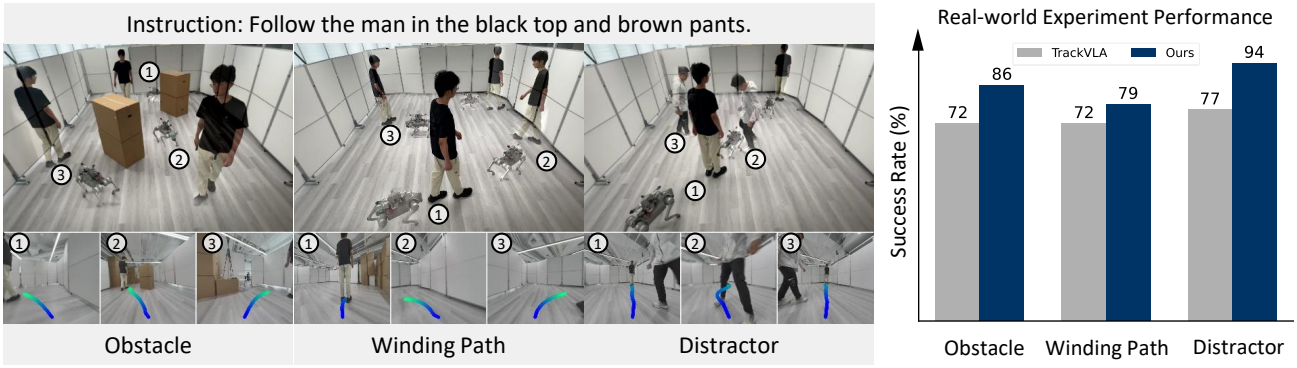

Fig. 5: **Visualizations of the Real World Experiments.** We evaluate TrackVLA++ on three different tasks: Obstacle, Winding Path, and Distractor, showcasing the tracking performance during target disappearance and occlusion. The bar chart provides a quantitative comparison of success rate between TrackVLA and TrackVLA++, highlighting the improved performance of our method.

TABLE IV: **Ablation Study of Proposed Designs.** We analyze the contributions of individual components on EVT-Bench `DT split`.

Methods	*Distracted Tracking (DT)*		
	SR ↑	TR ↑	CR ↓
TrackVLA [12]	57.6	63.2	5.80
NaVFoM (Four views)	<u>62.0</u>	67.9	-
TrackVLA++ (Ours)	**74.0**	**73.7**	**3.51**
w/o Polar-CoT & TIM	65.2	64.8	8.17
w/o TIM	71.2	69.8	4.74
w TIM (16 tokens)	74.2 (+0.2)	73.4 (-0.3)	3.27 (-0.24)

D. Ablation Study

We conduct an ablation study on the `DT` split of EVT-Bench (four views), with results shown in Table IV. Performance gains are mainly attributed to the proposed modules: the CoT module improves SR by 6.0%, and the TIM module (4 tokens) provides an additional 2.8%, demonstrating their complementary effects. Furthermore, we investigate the effect of varying the number of TIM tokens. To our surprise, increasing the token number from 4 to 16 does not result in a noticeable performance improvement, suggesting that the model can achieve robust tracking with concise token representations. This finding emphasizes the efficiency of our design in maintaining high performance with minimal computational overhead.

VI. Conclusion

We propose TrackVLA++, a Vision-Language-Action (VLA) model for embodied visual tracking that integrates explicit spatial reasoning and long-horizon target memory. Through the polar Chain-of-Thought (Polar-CoT) mechanism and the Target Identification Memory (TIM) module, TrackVLA++ achieves robust spatiotemporal consistency and effectively handles severe occlusions and similar distractors. Extensive experiments validate its effectiveness, establishing state-of-the-art results on simulation benchmarks in both egocentric and multi-camera settings, with strong generalization to real-world scenarios.

References

[1] A. Maalouf, N. Jadhav, K. M. Jatavallabhula, M. Chahine, D. M. Vogt, R. J. Wood, A. Torralba, and D. Rus, "Follow anything: Open-set detection, tracking, and following in real-time," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3283–3290, 2024. 1

[2] W. Zhang, K. Song, X. Rong, and Y. Li, "Coarse-to-fine uav target tracking with deep reinforcement learning," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 4, pp. 1522–1530, 2018. 1

[3] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "Ad-vat: An asymmetric dueling mechanism for learning visual active tracking," in *International Conference on Learning Representations*, 2019. 1, 6

[4] F. Zhong, X. Bi, Y. Zhang, W. Zhang, and Y. Wang, "Rspt: reconstruct surroundings and predict trajectory for generalizable active object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3705–3714. 1

[5] J. Li, J. Xu, F. Zhong, X. Kong, Y. Qiao, and Y. Wang, "Pose-assisted multi-camera collaboration for active object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 759–766. 1

[6] F. Zhong, K. Wu, H. Ci, C. Wang, and H. Chen, "Empowering embodied visual tracking with visual foundation models and offline rl," in *European Conference on Computer Vision*. Springer, 2024, pp. 139–155. 1, 2, 5, 6

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026. 1, 2

[8] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. 1

[9] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023. 1

[10] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," *Robotics: Science and Systems*, 2025. 1, 2, 5

[11] W. Luo, P. Sun, F. Zhong, W. Liu, T. Zhang, and Y. Wang, "End-to-end active object tracking via reinforcement learning," in *International conference on machine learning*. PMLR, 2018, pp. 3286–3295. 1, 2

[12] S. Wang, J. Zhang, M. Li, J. Liu, A. Li, K. Wu, F. Zhong, J. Yu, Z. Zhang, and H. Wang, "Trackvla: Embodied visual tracking in the wild," *arXiv pre-print*, 2025. [Online]. Available: http://arxiv.org/abs/2505.23189 1, 2, 4, 5, 6, 7

[13] D. Peng, J. Cao, Q. Zhang, and J. Ma, "Lovon: Legged open-vocabulary object navigator," *arXiv preprint arXiv:2507.06747*, 2025. 1, 2

[14] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, H. Cui *et al.*, "Graspvla: a grasping foundation model

pre-trained on billion-scale synthetic action data," *arXiv preprint arXiv:2505.03233*, 2025. 2, 3

[15] J. Zhang, S. Wu, X. Luo, H. Wu, L. Gao, H. T. Shen, and J. Song, "Inspire: Vision-language-action models with intrinsic spatial reasoning," *arXiv preprint arXiv:2505.13888*, 2025. 2, 3

[16] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," *arXiv preprint arXiv:2407.08693*, 2024. 2, 3

[17] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1702–1713. 2, 3

[18] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, and Y. Wang, "Unrealcv: Virtual worlds for computer vision," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1221–1224. 2, 5

[19] X. Shen, Y. Xiong, C. Zhao, L. Wu, J. Chen, C. Zhu, Z. Liu, F. Xiao, B. Varadarajan, F. Bordes *et al.*, "Longvu: Spatiotemporal adaptive compression for long video-language understanding," *arXiv preprint arXiv:2410.17434*, 2024. 2, 5

[20] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long *et al.*, "Paligemma 2: A family of versatile vlms for transfer," *arXiv preprint arXiv:2412.03555*, 2024. 2

[21] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 2

[22] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024. 2

[23] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025. 2

[24] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024. 2

[25] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025. 2

[26] Y. Zhong, X. Huang, R. Li, C. Zhang, Y. Liang, Y. Yang, and Y. Chen, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," *arXiv preprint arXiv:2502.20900*, 2025. 2

[27] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang, "Navid: Video-based vlm plans the next step for vision-and-language navigation," *Robotics: Science and Systems*, 2024. 2, 3

[28] A.-C. Cheng, Y. Ji, Z. Yang, X. Zou, J. Kautz, E. Biyik, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," in *RSS*, 2025. 2

[29] H. Ye, J. Zhao, Y. Zhan, W. Chen, L. He, and H. Zhang, "Person re-identification for robot person following with online continual learning," *IEEE Robotics and Automation Letters*, 2024. 2

[30] H. Ye, K. Cai, Y. Zhan, B. Xia, A. Ajoudani, and H. Zhang, "Rpf-search: Field-based search for robot person following in unknown dynamic environments," *arXiv preprint arXiv:2503.02188*, 2025. 2

[31] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–65, 2025. 2

[32] X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min *et al.*, "Habitat 3.0: A co-habitat for humans, avatars and robots," *arXiv preprint arXiv:2310.13724*, 2023. 2, 4

[33] W. Luo, P. Sun, F. Zhong, W. Liu, T. Zhang, and Y. Wang, "End-to-end active object tracking and its real-world deployment via reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 6, pp. 1317–1332, 2019. 2, 6

[34] A. Devo, A. Dionigi, and G. Costante, "Enhancing continuous control of mobile robots for end-to-end visual active tracking," *Robotics and Autonomous Systems*, vol. 142, p. 103799, 2021. 2

[35] K.-H. Zeng, Z. Zhang, K. Ehsani, R. Hendrix, J. Salvador, A. Herrasti, R. Girshick, A. Kembhavi, and L. Weihs, "Poliformer: Scaling on-policy rl with transformers results in masterful navigators," *arXiv preprint arXiv:2406.20083*, 2024. 2, 5

[36] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "Towards distraction-robust active visual tracking," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 782–12 792. 2, 6

[37] A. Bajcsy, A. Loquercio, A. Kumar, and J. Malik, "Learning vision-based pursuit-evasion robot policies," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 9197–9204. 2

[38] L. Scofano, A. Sampieri, T. Campari, V. Sacco, I. Spinelli, L. Ballan, and F. Galasso, "Following the human thread in social navigation," *arXiv preprint arXiv:2404.11327*, 2024. 2

[39] D. Shah, A. Bhorkar, H. Leen, I. Kostrikov, N. Rhinehart, and S. Levine, "Offline reinforcement learning for visual navigation," *arXiv preprint arXiv:2212.08244*, 2022. 2

[40] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022. 2

[41] Y. Cao, J. Zhang, Z. Yu, S. Liu, Z. Qin, Q. Zou, B. Du, and K. Xu, "Cognav: Cognitive process modeling for object goal navigation with llms," *arXiv preprint arXiv:2412.10439*, 2024. 2

[42] J. Zhang, L. Dai, F. Meng, Q. Fan, X. Chen, K. Xu, and H. Wang, "3d-aware object goal navigation via simultaneous exploration and identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6672–6682. 2

[43] J. Zhang, A. Li, Y. Qi, M. Li, J. Liu, S. Wang, H. Liu, G. Zhou, Y. Wu, X. Li, Y. Fan, W. Li, Z. Chen, F. Gao, Q. Wu, Z. Zhang, and H. Wang, "Embodied navigation foundation model," 2025. [Online]. Available: https://arxiv.org/abs/2509.12129 3, 5, 6

[44] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025. 3

[45] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986. 3

[46] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023. 3

[47] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023. 4, 5

[48] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023. 5, 6

[49] OpenAI, "Introducing 4o image generation," https://openai.com/index/introducing-4o-image, 2024, accessed: 2025-04-29. 5, 6

[50] M. Gupta, S. Kumar, L. Behera, and V. K. Subramanian, "A novel vision-based tracking algorithm for a human-following mobile robot," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1415–1427, 2016. 5

[51] J. Zuo, J. Hong, F. Zhang, C. Yu, H. Zhou, C. Gao, N. Sang, and J. Wang, "Plip: Language-image pre-training for person representation learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 45 666–45 702, 2024. 5

[52] Q. Jiang, L. Wu, Z. Zeng, T. Ren, Y. Xiong, Y. Chen, Q. Liu, and L. Zhang, "Referring to any person," *arXiv preprint arXiv:2503.08507*, 2025. 6

[53] S. Yang, T. Qu, X. Lai, Z. Tian, B. Peng, S. Liu, and J. Jia, "Lisa++: An improved baseline for reasoning segmentation with large language model," *arXiv preprint arXiv:2312.17240*, 2023. 6

[54] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6182–6191. 6

[55] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "Ad-vat+: An asymmetric dueling mechanism for learning and understanding visual active tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1467–1482, 2019. 6