Supplementary Materials for "Instance-level Visual Active Tracking with Occlusion-Aware Planning"

We organize the supplementary materials as follows. Section A reviews related work on visual active tracking. Section B presents the complete theoretical analysis and proofs of our instance prototype. Section C provides implementation details and hyperparameter analysis. Section D reports additional results from both simulation and real-world experiments.

A. Related Work

Visual Active Tracking. Most VAT methods fall into two categories. RL-based methods [4, 6, 12, 16, 19] learn end-to-end policies that map observations to actions, enabling low-latency control. However, they rely on sparse rewards that lead to slow convergence in complex scenes. To address this, EVT [19] uses offline RL to improve sample efficiency, and GC-VAT [12] employs curriculum learning [21, 22] to progressively guide policy optimization. Despite these advances, RL-based trackers still suffer from poor real-world performance due to the sim-to-real gap.

Pipeline methods [1, 7, 14, 18] decouple tracking into perception and control stages. FAn [7] uses foundation models like SAM [5] and DINOv2 [9] to detect the target and feed the predicted bounding box to a PID controller [8] for tracking. Although benefiting from strong visual models, these methods struggle to distinguish the target instance from distractors because most visual models are trained for category-level recognition. Besides, the controllers often struggle to recover tracking under occlusions. Recent extensions like TrackVLA [14] enhance the perception stage by incorporating language instructions based on an LLM, achieving SOTA performance. However, this improvement comes at high computational cost, which degrades performance in high-dynamic scenarios. Drawing inspiration from prototype-based historical representation [7, 20] and diffusion-based planning methods [2, 14], we propose an instance prototype for precise target matching and a diffusion-based planner to recover lost targets under occlusions.

B. Theoretical Analysis on Instance Prototype

Notations. For any target instance T_k, we define its true feature manifold M_k as the set of all features extracted by $\texttt{Desc}(\cdot)$ under arbitrary imaging conditions. From this manifold, the normalized reference features of T_k are obtained via Eq. (2) (in the main text) and denoted by the set F_k^*. For each reference feature $f_k^* \in F_k^*$, we generate a corresponding set of multi-view augmented features $\{f_{k,i}\}_{i=1}^N$. The mean of the augmented features, denoted $F_{\text{avg},k}$, is given by: $F_{\text{avg},k} = \frac{1}{N}\sum_{i=1}^N f_{k,i}$. The instance-aware prototypes are then derived via Eq. (3) (in the main text) and form the set $\hat{F}_k$. Finally, for any manifold M, we use $\mathbb{E}_{g\sim M}[\cdot]$ to denote expectation over M.

Assumption 1 ***(Multi-View Global Coverage)*** *For any target T_k, when the number of augmented views N is sufficiently large, the multi-view augmented features $\{f_{k,i}\}_{i=1}^N$ cover the true feature manifold M_k well enough such that the average squared distance from any point on the manifold to the augmented features is bounded by that of f_k^*:*

$$\mathbb{E}_{g\sim M_k}\left[\frac{1}{N}\sum_{i=1}^{N}\|f_{k,i}-g\|_2^2\right] \le \mathbb{E}_{g\sim M_k}\left[\|f_k^*-g\|_2^2\right]. \tag{1}$$

Assumption 1 captures that multi-view augmentation better covers the global structure of M_k, thereby reducing average distance to the manifold.

Assumption 2 ***(Manifold Cohesion and Separation)*** *There exist constants $\delta \in (0,1)$, $\eta \in (-1,1)$ such that for any features g_1, g_2 belonging to the same target manifold M_k, their cosine similarity satisfies $S(g_1,g_2) \ge \delta$ (intra-manifold cohesion), while for any features $g_k \in M_k$ and $g_j \in M_j$ from distinct targets $T_k \ne T_j$, the similarity satisfies $S(g_k,g_j) \le \eta$ (inter-manifold separation).*

Assumption 2 ensures cohesive features within the same target and well-separated features across different targets.

Lemma 1 *For any target T_k, the expected value of the squared Euclidean distance between $F_{avg,k}$ and manifold M_k is no larger than that between f_k^* and manifold M_k:*

$$\mathbb{E}_{g\sim M_k}\left[\|F_{avg,k}-g\|_2^2\right] \le \mathbb{E}_{g\sim M_k}\left[\|f_k^*-g\|_2^2\right]. \tag{2}$$

Proof 1 *For any fixed $g \in M_k$, the squared Euclidean distance $\|a-g\|_2^2$ is a convex function of a. By Jensen's inequality, for the multi-view augmented features set $\{f_{k,i}\}$:*

$$\left\| \frac{1}{N}\sum_{i=1}^{N} f_{k,i} - g \right\|_2^2 \leq \frac{1}{N}\sum_{i=1}^{N} \|f_{k,i}-g\|_2^2, \tag{3}$$

$$\|F_{avg,k}-g\|_2^2 \leq \frac{1}{N}\sum_{i=1}^{N} \|f_{k,i}-g\|_2^2. \tag{4}$$

Take the expectation of both sides of Eq. (4) over $g \sim M_k$. For brevity, we denote this expectation by $\mathbb{E}[\cdot]$.

$$\mathbb{E}\left[\|F_{avg,k}-g\|_2^2\right] \leq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \|f_{k,i}-g\|_2^2\right]. \tag{5}$$

By Assumption 1, RHS of Eq. (5) is bounded by:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \|f_{k,i}-g\|_2^2\right] \leq \mathbb{E}\left[\|f_k^*-g\|_2^2\right]. \tag{6}$$

Combining Eq. (5) and Eq. (6):

$$\mathbb{E}\left[\|F_{avg,k}-g\|_2^2\right] \leq \mathbb{E}\left[\|f_k^*-g\|_2^2\right]. \tag{7}$$

Lemma 2 *For any target T_k, the expected value of the squared Euclidean distance between $\hat{f}_k$ and manifold M_k is smaller than that between f_k^* and manifold M_k:*

$$\mathbb{E}_{g\sim M_k}\left[\|\hat{f}_k-g\|_2^2\right] \leq \mathbb{E}_{g\sim M_k}\left[\|f_k^*-g\|_2^2\right]. \tag{8}$$

Proof 2 *For unit vectors a, b, the squared Euclidean distance can be rewritten using inner product:*

$$\|a-b\|_2^2 = 2-2a\cdot b. \tag{9}$$

Applying Eq. (9) to $\hat{f}_k$ and f_k^*, Lemma 2 reduces to proving:

$$\mathbb{E}_{g\sim M_k}\left[\hat{f}_k \cdot g\right] \geq \mathbb{E}_{g\sim M_k}\left[f_k^* \cdot g\right]. \tag{10}$$

For brevity, we denote this expectation by $\mathbb{E}[\cdot]$. Substituting the definition of $\hat{f}_k$ into the left-hand side of Eq. (10) gives:

$$\mathbb{E}\left[\hat{f}_k \cdot g\right] = \frac{\mathbb{E}\left[(f_k^*+F_{avg,k})\cdot g\right]}{\|f_k^*+F_{avg,k}\|_2}. \tag{11}$$

By linearity of expectation, the expectation term expands to:

$$\mathbb{E}\left[(f_k^*+F_{avg,k})\cdot g\right] = \mathbb{E}\left[f_k^*\cdot g\right] + \mathbb{E}\left[F_{avg,k}\cdot g\right]. \tag{12}$$

By Lemma 1 and Eq. (9), we have:

$$2-2\mathbb{E}\left[F_{avg,k}\cdot g\right] \leq 2-2\mathbb{E}\left[f_k^*\cdot g\right], \tag{13}$$

simplifying gives:

$$\mathbb{E}\left[F_{avg,k}\cdot g\right] \geq \mathbb{E}\left[f_k^*\cdot g\right] = \mu. \tag{14}$$

Substitute Eq. (14) into Eq. (13):

$$\mathbb{E}\left[(f_k^*+F_{avg,k})\cdot g\right] \geq 2\mu. \tag{15}$$

We now bound $\|f_k^*+F_{avg,k}\|_2$ (denoted as L). By Assumption 2, $f_k^*\cdot F_{avg,k} \geq \delta$, and $\|F_{avg,k}\|_2^2 \geq \delta$. Thus,

$$L \geq \sqrt{1+3\delta}. \tag{16}$$

By the triangle inequality, $L \leq \|f_k^*\|_2 + \|F_{avg,k}\|_2$. Thus,

$$L \leq 2. \tag{17}$$

Substitute Eq. (15), Eq. (16) and Eq. (17) into (11):

$$\mathbb{E}\left[\hat{f}_k \cdot g\right] \geq \mu. \tag{18}$$

This proves Eq. (10), and thus the lemma.

Proposition 1 *For any two distinct targets $T_k \neq T_j$, the minimum squared distance between any pair of prototype features $(\hat{f}_k, \hat{f}_j)$ sampled from $\hat{F}_k$ and $\hat{F}_j$ is larger than the minimum squared distance between any pair of reference features (f_k^*, f_j^*) sampled from F_k^* and F_j^*:*

$$\min_{\hat{f}_k\in\hat{F}_k, \hat{f}_j\in\hat{F}_j} \|\hat{f}_k-\hat{f}_j\|_2^2 \geq \min_{f_k^*\in F_k^*, f_j^*\in F_j^*} \|f_k^*-f_j^*\|_2^2. \tag{19}$$

Proof 3 *Since all features are unit vectors, proving the inequality is equivalent to proving:*

$$\max_{\hat{f}_k,\hat{f}_j} \hat{f}_k\cdot\hat{f}_j \leq \max_{f_k^*,f_j^*} f_k^*\cdot f_j^*. \tag{20}$$

By Assumption 2 and Lemma 2, there exists a constant $\epsilon > 0$ such that:

$$\max_{\hat{f}_k,\hat{f}_j} \hat{f}_k\cdot\hat{f}_j \leq \eta+\epsilon \leq \max_{f_k^*,f_j^*} f_k^*\cdot f_j^*. \tag{21}$$

Therefore:

$$\min_{\hat{f}_k,\hat{f}_j} \|\hat{f}_k-\hat{f}_j\|_2^2 \geq \min_{f_k^*,f_j^*} \|f_k^*-f_j^*\|_2^2. \tag{22}$$

C. More Details of OA-VAT

C.1. Implementation Details

We train OA-VAT for 60 epochs using a batch size of 64 on a single RTX 3090 GPU, with the entire training process taking about 15 hours. We extract image features using the pre-trained `dinov3_vitl16` (300M parameters) [11], which processes 384×384 input images and output 768-dimensional features. Moreover, we generate candidate

Table 1. Comprehensive ablation studies on key hyperparameters. We analyze the EMA momentum β, Kalman filter parameters (γ, λ), and the feature descriptor model size. Models `B`, `H+` and `L` refer to `dinov3_vitb16`, `dinov3_vith+16` and `dinov3_vitl16`, respectively. **Bold** indicates the best performance, underline indicates the second best. Default settings are marked with a light blue background.

Parameter	Value	Parking Lot (2D)			UrbanCity (4D)			ComplexRoom (4D)			Avg.		
		$AR\uparrow$	$EL\uparrow$	$SR\uparrow$	$AR\uparrow$	$EL\uparrow$	$SR\uparrow$	$AR\uparrow$	$EL\uparrow$	$SR\uparrow$	$AR\uparrow$	$EL\uparrow$	$SR\uparrow$
EMA Momentum (β)	0.6	378	**485**	0.91	381	485	0.92	**402**	**487**	**0.95**	387	**486**	**0.93**
	0.7	381	484	0.91	378	483	0.91	400	480	0.93	386	482	0.92
	0.8	**392**	482	**0.93**	**385**	**486**	**0.95**	392	481	0.92	**390**	483	**0.93**
	0.9	383	472	0.90	366	471	0.89	393	479	0.91	381	474	0.90
Kalman Filter Center (γ)	0.3	385	469	0.91	381	471	0.92	397	488	0.93	388	476	0.92
	0.4	**392**	482	**0.93**	**385**	**486**	**0.95**	392	481	0.92	**390**	483	**0.93**
	0.5	381	**490**	**0.93**	365	467	0.90	**416**	**495**	**0.97**	387	**484**	**0.93**
Kalman Filter Slope (λ)	13	387	477	0.91	361	470	0.87	**418**	**495**	**0.96**	389	481	0.91
	15	**392**	**482**	**0.93**	**385**	**486**	**0.95**	392	481	0.92	**390**	**483**	**0.93**
	17	375	472	0.90	381	483	0.94	414	490	0.94	**390**	482	**0.93**
Feature Descriptor (DINOv3) Size	`B`	389	**485**	**0.93**	**395**	479	0.89	**406**	**484**	**0.95**	**397**	483	0.92
	`L`	**392**	482	**0.93**	385	**486**	**0.95**	392	481	0.92	390	483	**0.93**
	`H+`	388	484	**0.93**	377	485	0.92	400	**484**	0.94	388	**484**	**0.93**

Table 2. Results on UnrealCV benchmark. **Bold** represents the best while underline represents the second.

Trackers	Publication	SimpleRoom			Parking Lot			UrbanCity			UrbanRoad			Snow Village		
		AR	EL	SR	AR	EL	SR	AR	EL	SR	AR	EL	SR	AR	EL	SR
DiMP [1]	ICCV 2019	336	**500**	**1.00**	166	327	0.48	239	401	0.66	168	308	0.33	110	301	0.43
SARL [6]	TPAMI 2019	368	**500**	**1.00**	92	301	0.22	331	471	0.86	207	378	0.48	203	318	0.31
AD-VAT [15]	ICLR 2019	356	**500**	**1.00**	86	302	0.20	335	484	0.88	246	429	0.60	169	364	0.44
AD-VAT+ [16]	TPAMI 2019	373	**500**	**1.00**	267	439	0.60	389	497	0.94	326	471	0.80	182	365	0.44
TS [17]	ICML 2021	**412**	**500**	**1.00**	265	472	0.89	341	496	0.94	308	480	0.84	234	424	0.63
RSPT [18]	AAAI 2023	398	**500**	**1.00**	314	480	0.80	341	**500**	**1.00**	346	**500**	**1.00**	248	410	0.80
EVT [19]	ECCV 2024	374	**500**	**1.00**	274	484	0.92	306	**500**	**1.00**	300	496	0.96	229	471	0.87
FAn [7]	RAL 2024	318	**500**	**1.00**	215	481	0.96	193	466	0.90	152	409	0.76	306	456	0.90
FAn+SAM2 [10]	ICLR 2025	329	**500**	**1.00**	215	491	0.96	217	470	0.92	207	442	0.90	317	465	0.94
TrackVLA [14]	CoRL 2025	-	**500**	**1.00**	-	**500**	**1.00**	-	**500**	**1.00**	-	**500**	**1.00**	-	**500**	**1.00**
Ours	CVPR 2026	389	**500**	**1.00**	**382**	**500**	**1.00**	**390**	**500**	**1.00**	**401**	**500**	**1.00**	**391**	**500**	**1.00**

masks using the `yoloe-11l-seg` model [13], and select the candidate with the highest similarity exceeding the matching threshold $\eta_s = 0.5$ as the target. During tracking, the visual prototype is updated online via an exponential moving average (EMA):

$$\tilde{\mathbf{f}}' \leftarrow \beta\tilde{\mathbf{f}}' + (1-\beta)\hat{\mathbf{f}}_{\text{tar}}, \tag{23}$$

where the EMA momentum β is set to 0.8. Additionally, we employ a confidence-aware Kalman filter, where the measurement noise covariance $\mathbf{R}_t$ is modeled as a function of the detection confidence c_t:

$$\mathbf{R}_t = \sigma^2(c_t)\mathbf{I}, \quad \sigma^2(c_t) = \frac{1}{1+e^{\lambda\cdot(c_t-\gamma)}}, \tag{24}$$

where the hyperparameters are set to $\lambda = 15.0$ and $\gamma = 0.4$.

C.2. Hyperparameter Analysis

We perform ablation experiments to analyze the influence of key hyperparameters in the OA-VAT method, including the EMA momentum β of the online visual prototype enhancement module, the parameters (λ, γ) of the confidence-aware Kalman filter, and the model size of the feature extractor.

EMA Momentum. We first analyze the effect of the momentum coefficient β in the EMA update of the online prototype enhancement module, as shown in Eq. (23). We conducted experiments with β set to 0.6, 0.7, 0.8, and 0.9. As shown in Tab. 1 (rows 1-4), OA-VAT exhibits consistently strong performance across all settings, and our default choice of $\beta = 0.8$ achieves the best results, providing a good balance between historical and current observations.

Kalman Filter Parameters. We then analyze the parameters of the confidence-aware Kalman filter in Eq. (24).

To evaluate the effect of λ, we compare the default setting $\lambda = 15$ with $\lambda = 13$ and $\lambda = 17$ (Tab.1 rows 5-7). We then compare the default $\gamma = 0.4$ with $\gamma = 0.3$ and $\gamma = 0.5$, as shown in Tab.1 (rows 8-10). The results demonstrate that OA-VAT is robust to hyperparameter variations.

Feature Extractor Size. To demonstrate the effectiveness of OA-VAT across models of different sizes, we evaluate its performance by comparing the default feature extractor `dinov3_vitl16` (300M parameters), with `dinov3_vitb16` (86M) and `dinov3_vith+16` (840M) variants. As shown in Tab. 1 (Rows 11–13), replacing the 300M extractor with the smaller 86M variant results in only a 1.1% relative performance drop, indicating that OA-VAT is highly robust across model sizes.

D. More Experimental Results

D.1. Comparison Experiments

Details of Baselines. We compare OA-VAT against 12 baselines: DiMP [1] combines a pre-trained passive tracker with a PID [8] controller. SARL [6] is an end-to-end RL tracker with a Conv-LSTM backbone. AD-VAT [15] and AD-VAT+ [16] use an asymmetric dueling RL framework with adversarial learning. RSPT [18] leverages RGB-D input for structure-aware tracking. Cross-modal Teacher-Student (TS) [17] employs a cross-modal teacher-student strategy for distraction-robust tracking. EVT [19] integrates visual foundation models with offline RL for efficient embodied tracking. Follow Anything (FAn) [7] enables open-vocabulary tracking by combining foundation models, and FAn+SAM2 replaces its segmentation module with the latest SAM2 [10] model. D-VAT [4] maps RGB observations directly to continuous control signals through reinforcement learning. GC-VAT [12] designs a goal-centered reward for effective tracking in complex environments. TrackVLA [14] unifies target recognition and tracking within a single VLA framework built upon an LLM backbone.

Additional Results on UnrealCV. We evaluate all methods on the UnrealCV benchmark for single-target tracking in five distinct virtual scenes: `SimpleRoom`, `Parking Lot`, `UrbanCity`, `UrbanRoad`, and `Snow Village`, with detailed results presented in Tab. 2. OA-VAT achieves the highest success rate ($SR = 1.00$) across all five scenes, with zero target loss ($EL = 500$) in every episode.

D.2. Experiments in Real-world Scenarios

To evaluate the real-world applicability and robustness of OA-VAT, we deploy it on a *DJI Tello* drone. We use the `DJITelloPy` library [3] to capture video streams (at a resolution of 320×240) from the drone, transmit them over the network to a ground station equipped with an NVIDIA RTX 3090 GPU, and return the control signals generated by OA-VAT. The drone operates in velocity control mode, with control commands including linear and angular velocities. The entire pipeline runs at approximately 35 FPS in our experiments. We then evaluate OA-VAT on two challenging scenarios: long-term tracking with occlusions, and specific instance discrimination against same-category distractors.

Effectiveness in Long-Term Tracking. OA-VAT shows superior performance over all baselines in long-term real-robot tracking. It can actively plan collision-free paths and recover the occluded target, as shown in Fig. 6 of the main article. However, as illustrated in Fig. 1(a), FAn [7] lacks a planning module, causing the robot to stop when the target remains occluded for an extended period, leading to tracking failure. Moreover, as shown in Fig. 1(b), EVT [19] method, which adopts an offline RL-based planner, still produces incorrect plans under occlusion and fails to navigate around obstacles to recover the target. *Full long-term tracking videos of OA-VAT, FAn, and EVT methods are provided in the supplementary video.*

Effectiveness Under Distractors. As shown in Fig. 2, we evaluate OA-VAT under distractors, where the person in black is the true target and the one in red is the distractor. During online tracking, OA-VAT robustly tracks the correct target and remains unaffected when only the distractor is visible, as illustrated in the bottom-left subfigure of Fig. 2.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019. 1, 3, 4

[2] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024. 1

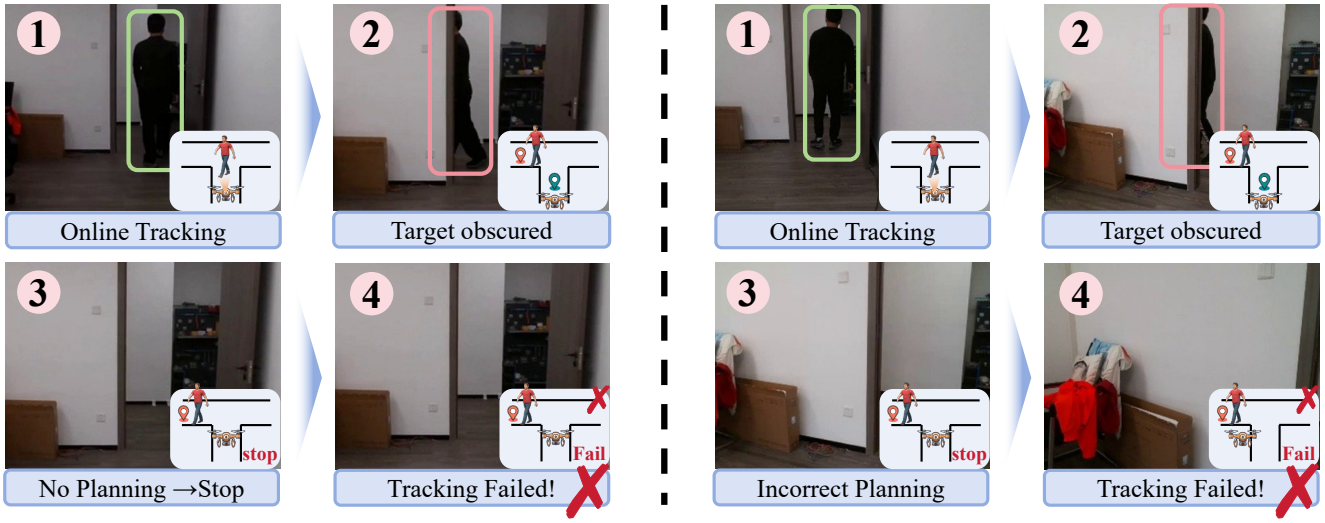
[3] damiafuentes. Djitellopy. https://github.com/damiafuentes/DJITelloPy, 2025. 4

[4] Alberto Dionigi, Simone Felicioni, Mirko Leomanni, and Gabriele Costante. D-vat: End-to-end visual active tracking for micro aerial vehicles. *IEEE Robotics and Automation Letters*, 9(6):5046–5053, 2024. 1, 4

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[6] Wenhan Luo, Peng Sun, Fangwei Zhong, Wei Liu, Tong Zhang, and Yizhou Wang. 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. 1, 3, 4

[7] Alaa Maalouf, Ninad Jadhav, Krishna Murthy Jatavallabhula, Makram Chahine, Daniel M Vogt, Robert J Wood, Antonio Torralba, and Daniela Rus. Follow anything: Open-set detection, tracking, and following in real-time. *IEEE*



(a) FAn fails to plan when the target is occluded.

(b) EVT performs incorrect planning under target occlusion.

Figure 1. Failure cases of the baseline method FAn [7] and EVT [19] on real drone *DJI Tello*.

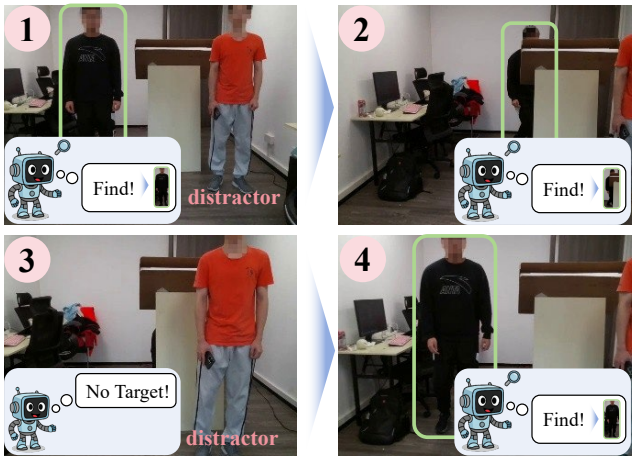


Figure 2. OA-VAT accurately detects the target against distractors.

Robotics and Automation Letters, 9(4):3283–3290, 2024. 1, 3, 4, 5

[8] Nicolas Minorsky. Directional stability of automatically steered bodies. *Journal of the American Society for Naval Engineers*, 34(2):280–309, 1922. 1, 4

[9] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[10] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 4

[11] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[12] Haowei Sun, Jinwu Hu, Zhirui Zhang, Haoyuan Tian, Xinze Xie, Yufeng Wang, Xiaohua Xie, Yun Lin, Zhuliang Yu, and Mingkui Tan. Open-world drone active tracking with goal-centered rewards. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 4

[13] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24591–24602, 2025. 3

[14] Shaoan Wang, Jiazhao Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. Trackvla: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*, 2025. 1, 3, 4

[15] Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. Ad-vat: An asymmetric dueling mechanism for learning visual active tracking. In *International Conference on Learning Representations*, 2019. 3, 4

[16] Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. 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. 1, 3, 4

[17] Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. Towards distraction-robust active visual tracking. In *International Conference on Machine Learning*, pages 12782–12792. PMLR, 2021. 3, 4

[18] Fangwei Zhong, Xiao Bi, Yudi Zhang, Wei Zhang, and Yizhou Wang. Rspt: reconstruct surroundings and predict

trajectory for generalizable active object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3705–3714, 2023. 1, 3, 4

[19] Fangwei Zhong, Kui Wu, Hai Ci, Churan Wang, and Hao Chen. Empowering embodied visual tracking with visual foundation models and offline rl. In *European Conference on Computer Vision*, pages 139–155. Springer, 2024. 1, 3, 4, 5

[20] Kai Zhou, Shuhai Zhang, Zeng You, Jinwu Hu, Mingkui Tan, and Fei Liu. Zero-shot skeleton-based action recognition with prototype-guided feature alignment. *IEEE Transactions on Image Processing*, 34:4602–4617, 2025. 1

[21] Yuwei Zhou, Hong Chen, Zirui Pan, Chuanhao Yan, Fanqi Lin, Xin Wang, and Wenwu Zhu. Curml: A curriculum machine learning library. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7359–7363, 2022. 1

[22] Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang, Zhixiao Xiong, Fangzhou Xiong, Peiyang Xu, Wenwu Zhu, et al. Curbench: curriculum learning benchmark. In *Forty-first International Conference on Machine Learning*, 2024. 1