TRACKEVT: DUAL-STATE TARGET MAINTENANCE FOR EMBODIED VISUAL TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied visual tracking (EVT) requires an agent to continuously follow a designated moving target from egocentric visual observations. However, high-confidence localization alone cannot confirm that the tracker still follows the designated target. An identity error can steer the agent away and be reinforced by later observations. We refer to this closed-loop failure as *target-state takeover*. To address it, we propose **TrackEVT**, which maintains two complementary states. A dense pixel state provides target geometry for visual servoing. In parallel, persistent target and distractor trajectories check whether the propagated pixels still belong to the selected target. When the states disagree, identity trajectories cannot overwrite the control state directly. Instead, *identity-constrained rebinding* verifies a current-frame target mask before reinitializing pixel propagation. TrackEVT combines pretrained visual components with explicit control and requires no task-specific training. Experiments cover several simulated scenes, target categories, and a real-world deployment. On established benchmarks, TrackEVT achieves the highest mean scores among the compared methods: a mean success rate of 0.97 on Gym-UnrealCV and a mean tracking success rate of 0.98 for aerial vehicle tracking on DAT. A targeted identity stress test further shows that persistent trajectories and verified rebinding improve identity preservation and recovery. Code: https://anonymous.4open.science/r/CodeReviewM8Q2/.

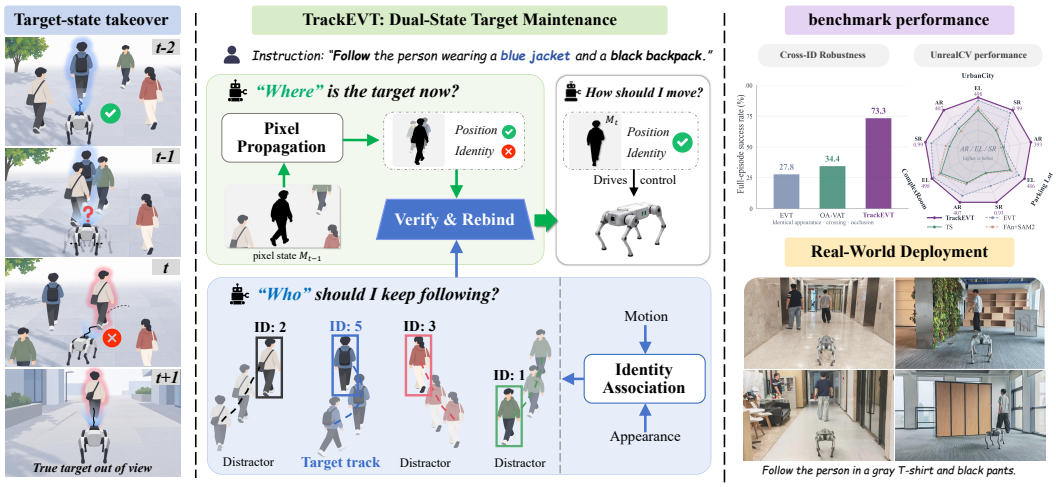


Figure 1: **TrackEVT at a glance.** Identity errors can redirect the agent toward a distractor. TrackEVT couples pixel-based control with persistent identity memory to guard against target-state takeover. The asterisk marks our OA-VAT reproduction under the shared Cross-ID protocol.

1 INTRODUCTION

Embodied visual tracking (EVT) requires a mobile agent to follow a designated moving target by controlling its motion from egocentric visual observations. In language-guided EVT, a natural-

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

language instruction specifies the target (Wang et al., 2025b; Ye et al., 2026). This capability supports assistive robotics, aerial filming, and surveillance (Zhong et al., 2019; 2024). Unlike passive video tracking, EVT couples perception and action: the estimated target state guides motion, which changes the next observation.

Existing approaches have evolved from reinforcement-learning policies (Zhong et al., 2019; Luo et al., 2020; Zhong et al., 2021b; 2024; Sun et al., 2025) to vision–language–action models that integrate target instructions, visual history, and action prediction (Zhang et al., 2024; Wang et al., 2025b; Liu et al., 2025; Ye et al., 2026). In parallel, modular systems combine pretrained perception with explicit tracking, recovery, and planning (Maalouf et al., 2024; Wu et al., 2025; Sun et al., 2026). These advances improve recognition and following, but a locally plausible target estimate can still belong to the wrong instance.

After a crossing or partial occlusion, a high-confidence box or pixel region may continue smoothly on a distractor. Frame-wise confidence and short-term continuity cannot guarantee that the estimate still belongs to the originally specified target. In an embodied loop, the resulting identity error also changes future observations: the agent moves toward the distractor and may push the true target out of view. Later observations can then reinforce the wrong estimate. We refer to this closed-loop failure as *target-state takeover* (Figure 1).

The controller needs a temporally continuous estimate of the visible target. Pixel propagation provides this estimate without selecting a new detection in every frame (Cheng & Schwing, 2022; Cheng et al., 2024; Ravi et al., 2024), but it can also carry an identity error forward. Maintaining target identity requires a longer history that distinguishes the selected target from other visible instances. Multi-object trajectories provide such a history by associating appearance and motion observations over time (Wojke et al., 2017; Zhang et al., 2022; Cao et al., 2023). Rather than using one representation for both roles, we separate the state used for control from the state used to verify target identity.

We introduce **TrackEVT**, a dual-state target-maintenance framework that separates an executable pixel state for control from a persistent trajectory state for identity maintenance. The two states operate in parallel. The pixel state provides dense target geometry and is the only state read by the controller, while trajectories audit target ownership even when propagation remains confident. Crucially, the trajectory state may propose a correction but cannot directly overwrite the control state. *Identity-constrained rebinding* instead recovers and verifies a current-frame target mask before reinitializing propagation memory, so corrections change which instance is propagated in subsequent frames rather than substituting a box estimate into control.

Notably, TrackEVT supports both person and vehicle following without task-specific training by combining pretrained visual components with explicit visual servoing. We evaluate it on Gym-UnrealCV, DAT, and EVT-Bench across ground-level and aerial viewpoints. We additionally use Cross-ID, a controlled identity stress test combining shared target–distractor clothing with varying orientations, repeated crossings, foreground occlusion, and post-occlusion identity recovery within each episode. Together with module ablations, these evaluations test both overall following performance and the benefits of identity-constrained state maintenance. Our contributions are summarized as follows:

- We characterize *target-state takeover*, a closed-loop failure in which an identity error redirects the agent and is reinforced by its subsequent observations. This failure motivates decoupling the control state from identity verification.
- We propose **TrackEVT**, a dual-state framework with asymmetric update permissions. The pixel state alone drives control, while persistent trajectories audit target ownership even when propagation remains confident. *Identity-constrained rebinding* is the sole cross-state update, requiring a verified current-frame mask to reinitialize propagation.
- Across person following, aerial vehicle tracking, identity stress tests, and real-robot deployment, TrackEVT provides accurate closed-loop tracking without task-specific training. The results establish persistent target identity as a practical component of embodied tracking.

2 Related Work

Embodied visual tracking. EVT has progressed from online, adversarial, and distraction-aware reinforcement-learning policies (Zhong et al., 2019; Luo et al., 2020; Zhong et al., 2021a;b; Devo et al., 2021; Xi et al., 2022; Dionigi et al., 2024; Sun et al., 2025) to offline and in-context policy learning (Zhong et al., 2024; Wu et al., 2026). General-purpose and tracking-oriented VLA or prompt-conditioned models further combine target instructions, visual history, reasoning, and waypoint or action prediction (Brohan et al., 2023; Kim et al., 2024; Zhang et al., 2024; Wang et al., 2025b; Liu et al., 2025; 2026; Xie et al., 2026; Ye et al., 2026). In parallel, modular systems connect pretrained perception or passive trackers to visual servoing, trajectory prediction, failure recovery, and occlusion-aware planning (Bhat et al., 2019; Zhong et al., 2023; Maalouf et al., 2024; Zhong et al., 2024; Wu et al., 2025; Sun et al., 2026). These systems already employ text-conditioned masks as policy observations, descriptor-based mask re-detection, or adaptive instance prototypes with motion and planning (Zhong et al., 2024; Maalouf et al., 2024; Sun et al., 2026). FAn stores historical descriptors for mask-based re-detection after target loss, while OA-VAT combines instance prototypes, online enhancement, and motion-aware tracking with occlusion recovery (Maalouf et al., 2024; Sun et al., 2026). These methods mainly address target loss or occlusion recovery. TrackEVT instead uses persistent trajectories to audit target ownership during confident pixel propagation, without waiting for a low-confidence failure.

Pixel propagation and identity association. Open-set perception provides language-conditioned or category-agnostic proposals (Radford et al., 2021; Caron et al., 2021; Kirillov et al., 2023; Liu et al., 2024), while tracking–segmentation and video object segmentation propagate dense target masks through matching and memory (Wang et al., 2019; Lu et al., 2020; Cheng et al., 2021; Yang et al., 2021; Yang & Yang, 2022; Cheng & Schwing, 2022; Cheng et al., 2023b; Yang et al., 2023b; Cheng et al., 2023a; 2024; Ravi et al., 2024). These methods provide fine geometry and temporal continuity, but normally update segmentation identity and pixel estimates within the same propagation process. Conversely, tracking-by-detection maintains competing box-level trajectories using motion, appearance, and global association (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2021; 2022; Cao et al., 2023). Such trajectories preserve identity over longer periods, but detector-dependent boxes are too coarse for pixel-level control. TrackEVT therefore assigns the two representations asymmetric roles and update permissions. The propagated mask alone drives control, while trajectories only propose corrections. The pixel state changes only after a verified current-frame target mask is used to reinitialize propagation.

3 TrackEVT: Dual-State Target Maintenance

3.1 Problem and Overview

At time t, an agent receives an egocentric RGB image $I_t \in \mathbb{R}^{H\times W\times 3}$ and outputs an action u_t that changes its next observation. The instruction ℓ constrains instance detection. After one instance is selected as the target, the agent must continue following that same instance. Its initial mask M_0 initializes pixel propagation, and its initial appearance provides an immutable identity anchor a_0. We address identity maintenance during continuous tracking and correct the tracked identity when current visual evidence supports the target. General-purpose search after the target has left the field of view is outside our scope.

TrackEVT separates two forms of temporal state:

$$\mathcal{W}_t = (M_t, \mathcal{H}_t^{\text{pix}}), \qquad \mathcal{Y}_t = (\mathcal{Z}_t, i_t^\star, a_0). \tag{1}$$

The pixel state $\mathcal{W}_t$ contains the binary target region $M_t \in \{0,1\}^{H\times W}$ and the propagator's internal memory $\mathcal{H}_t^{\text{pix}}$. The identity state $\mathcal{Y}_t$ contains target and distractor tracklets $\mathcal{Z}_t = \{z_t^i : i \in \mathcal{I}_t\}$, where $\mathcal{I}_t$ is the set of active track IDs. The identifier $i_t^\star$ marks the trajectory assigned to the originally selected target. This ID remains unassigned until the target trajectory is established. In our implementation, Cutie maintains the pixel state, while YOLOE detections are associated into persistent identity trajectories using FastReID/DINOv3 appearance cues and Kalman motion predictions. Figure 2 summarizes how the two states interact. The controller reads only the accepted mask M_t. The trajectory state separately compares the target track with the visible candidates. If its assignment conflicts with the propagated mask, it may propose a correction. That proposal cannot directly

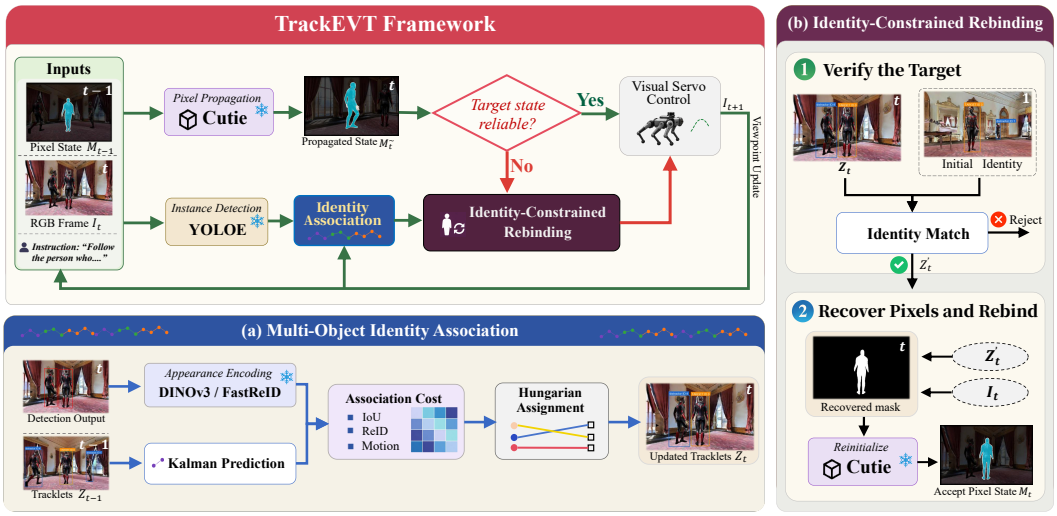


Figure 2: **TrackEVT framework.** Cutie propagates the pixel state for visual servoing. (a) Appearance, motion, and overlap associate detections into persistent identity trajectories. (b) A trajectory correction must be converted into a verified current-frame mask before reinitializing Cutie.

control the agent or reset the pixel state with a box. Rebinding first requires a verified mask in the current image. All visual networks are pretrained and frozen, while their memories and estimator states are updated online. Appendix A provides the scheduling, association, recovery, and platform-specific details.

3.2 Executable Pixel State

The pixel path maintains the visible region of the selected instance rather than choosing a new detection in every frame. Pretrained Cutie (Cheng et al., 2024) propagates the target mask through the current image:

$$(\widetilde{M}_t, \widetilde{\mathcal{H}}_t^{\text{pix}}) = \mathcal{P}_\theta(I_t; \mathcal{H}_{t-1}^{\text{pix}}). \tag{2}$$

The memory is initialized from M_0 and reset only after successful rebinding. We threshold the foreground probability and retain the largest connected component. Low foreground confidence or an implausible area marks the propagation as unreliable. These tests evaluate the predicted region, not the target identity: a plausible $\widetilde{M}_t$ may still cover a distractor.

For reliable propagation, let $B(M_t)$ be the tight box of the accepted mask, $c_x(B)$ be its horizontal center, and $|B|$ denote its area. The front-view control geometry is

$$\phi(M_t) = \left[\frac{c_x(B(M_t)) - W/2}{W/2}, \quad s^\star - \frac{|B(M_t)|}{HW}\right]^\top, \tag{3}$$

where $s^\star$ is the desired relative target size. The normalized errors are converted to platform-specific control units before bounded visual servoing maps horizontal error to yaw and scale error to forward velocity. The identity path never sends an independent control command.

3.3 Persistent Identity Trajectories

If an adaptive target template is updated after an identity switch, it may absorb features from a distractor. TrackEVT therefore retains competing instances as separate tracklets. At each scheduled identity check, YOLOE (Wang et al., 2025a) detects candidates in I_t according to ℓ. Each tracklet stores a Kalman motion estimate, an identity-feature gallery, an auxiliary appearance prototype, and its observation history. The target also retains the initial anchor a_0, which online updates never replace. We refer to each scheduled check as an *identity audit*. Audits run asynchronously during normal tracking, even while Cutie remains confident.

For an audit of frame t, t^- denotes the state before incorporating its detections. Let $\mathcal{I}_{t^-}$ be the existing track identifiers and $\mathcal{D}_t$ the current candidates. The compatibility of track i and candidate j is

$$\psi_{ij} = w_{\mathrm{id}} s_{ij}^{\mathrm{id}} + w_{\mathrm{aux}} s_{ij}^{\mathrm{aux}} + w_{\mathrm{mot}} s_{ij}^{\mathrm{mot}} + w_{\mathrm{iou}} \,\mathrm{IoU}(\widehat{b}_t^i, b_t^j). \tag{4}$$

Here $\widehat{b}_t^i$ is the motion-predicted box, and b_t^j is the detection box. The identity term compares a candidate with the track's feature gallery. For the target track, it also uses the immutable anchor a_0. The remaining terms measure auxiliary appearance similarity, Kalman motion consistency, and spatial overlap. All weights are nonnegative and sum to one.

Instead of matching only the target, TrackEVT jointly assigns current detections to the target and distractor tracklets:

$$\begin{aligned} X_t^\star &\in \arg\max_{X \in \mathfrak{M}_t} \sum_{i,j} X_{ij} \psi_{ij}, \\ \mathfrak{M}_t &= \{X \in \{0,1\}^{n^- \times m} : X\mathbf{1} \le \mathbf{1},\ X^\top \mathbf{1} \le \mathbf{1},\ X_{ij} \le g_{ij}\}, \end{aligned} \tag{5}$$

where $n^- = |\mathcal{I}_{t^-}|$, $m = |\mathcal{D}_t|$, and g_{ij} excludes incompatible appearance–motion pairs. The Hungarian assignment permits unmatched tracks and detections. Before the track states are updated or expired, target-anchor reservation gives an unmatched target a candidate when the identity evidence is sufficiently strong.

The association identifies both the candidate assigned to the target trajectory and the candidate covered by the propagated mask. If they differ, the system can detect a possible target-state takeover even when pixel confidence remains high. An unmatched target alone, however, does not constitute evidence for assigning its identity to a different instance.

3.4 Identity-Constrained Rebinding

Error decomposition. Let $\iota^\star$ denote the original target, and let $\mathbf{g}_t^\iota = \phi(M_{t,\mathrm{true}}^\iota) \in \mathbb{R}^2$ denote the ideal visible geometry of physical instance ι. The true identities and masks are used only for analysis and are not supplied to the tracker. For a visible instance and a source-frame estimate $\widehat{\mathbf{g}}_\tau^\iota$, $\tau \le t$,

$$\widehat{\mathbf{g}}_\tau^\iota - \mathbf{g}_t^{\iota^\star} = \underbrace{\widehat{\mathbf{g}}_\tau^\iota - \mathbf{g}_\tau^\iota}_{\text{localization}} + \underbrace{\mathbf{g}_\tau^\iota - \mathbf{g}_t^\iota}_{\text{temporal lag}} + \underbrace{\mathbf{g}_t^\iota - \mathbf{g}_t^{\iota^\star}}_{\text{identity mismatch}}. \tag{6}$$

No statistical independence is assumed. For a motion-predicted hint, the first two terms become its current-frame prediction error. Even an accurately localized or predicted region may belong to an instance $\iota \neq \iota^\star$.

Cross-instance fusion. At a fixed frame t, suppose the pixel estimate follows a distractor $\iota \neq \iota^\star$ and a recovery estimate corresponds to the true target: $\widetilde{\mathbf{g}}_t = \mathbf{g}_t^\iota + \mathbf{e}_p$ and $\widehat{\mathbf{g}}_t^r = \mathbf{g}_t^{\iota^\star} + \mathbf{e}_r$. Here the subscripts p and r denote propagation and recovery, respectively. The errors satisfy $\|\mathbf{e}_p\| \le \epsilon_p$ and $\|\mathbf{e}_r\| \le \epsilon_r$. All norms are Euclidean. A convex mixture $\mathbf{g}_t^\lambda = (1-\lambda)\widetilde{\mathbf{g}}_t + \lambda \widehat{\mathbf{g}}_t^r$, $0 \le \lambda \le 1$, has error

$$\mathbf{g}_t^\lambda - \mathbf{g}_t^{\iota^\star} = (1-\lambda)(\mathbf{g}_t^\iota - \mathbf{g}_t^{\iota^\star}) + (1-\lambda)\mathbf{e}_p + \lambda \mathbf{e}_r. \tag{7}$$

Applying the reverse triangle inequality and the error bounds gives

$$\|\mathbf{g}_t^\lambda - \mathbf{g}_t^{\iota^\star}\| \ge \big[(1-\lambda)\|\mathbf{g}_t^\iota - \mathbf{g}_t^{\iota^\star}\| - (1-\lambda)\epsilon_p - \lambda\epsilon_r\big]_+, \tag{8}$$

with $[x]_+ = \max(x, 0)$. When both estimates are exact,

$$\mathbf{e}_p = \mathbf{e}_r = \mathbf{0} \quad \Longrightarrow \quad \|\mathbf{g}_t^\lambda - \mathbf{g}_t^{\iota^\star}\| = (1-\lambda)\|\mathbf{g}_t^\iota - \mathbf{g}_t^{\iota^\star}\|. \tag{9}$$

For $\lambda < 1$, geometrically distinct instances therefore leave a nonzero residual even without localization noise. This result concerns coordinate mixing across instances, not fusion in general. Taking $\lambda = 1$ replaces the current output but leaves the old propagation memory unchanged unless that memory is also reinitialized.

Correction proposals. A target trajectory proposes a correction only after it has sufficient observation history and association support and is distinguishable from the candidate covered by the pixel mask. Sustained low pixel confidence instead activates reference-based recovery, which does not require a trajectory. Each proposal p_t records a spatial hint b_t^{hint}, an evidence timestamp $\tau \le t$, and a source $\nu \in \{\text{trajectory}, \text{reference}\}$. A recovery proposal may also include a source-frame mask.

Current-frame verification. The proposal specifies where to search, but it is not itself a valid control state. The recovery operator must produce a usable mask in the current frame:

$$\widehat{M}_t = \mathcal{R}(I_t; p_t, \mathcal{F}_t^+), \tag{10}$$

where $\mathcal{F}_t^+$ contains positive-reference descriptors that are kept separate from a_0. A source mask may be reused only when $\tau = t$. Otherwise, YOLOE extracts fresh instances, the spatial hint restricts the candidates, and reference similarity selects the mask. If no candidate has sufficient support, the operator returns $\varnothing$. It never converts a proposal box directly into a mask. Let G_τ^ν indicate that source checks pass and V_t^{seed} indicate accepted, nonempty current-frame support. The binary eligibility gate is

$$E_t(p_t) = G_\tau^\nu(p_t)\, \mathbb{I}[0 \le t - \tau \le \Delta_{\max}^\nu]\, V_t^{\text{seed}}(\widehat{M}_t), \tag{11}$$

where $\Delta_{\max}^\nu$ bounds the age of the proposal. The gate accepts p_t only if its source checks pass, its timestamp is recent, and the recovered mask has valid current-frame support. This seed check is separate from the propagation-area test V_t^{pix}.

Propagation after rebinding. For a successfully executed step, let $A_t = 1$ denote completed reinitialization from an eligible candidate. The pixel state is

$$\mathcal{W}_t = \begin{cases} \text{Init}_\theta(I_t, \widehat{M}_t), & A_t = 1, \\ (\widetilde{M}_t, \widetilde{\mathcal{H}}_t^{\text{pix}}), & A_t = 0. \end{cases} \tag{12}$$

Writing Seed_θ for the memory component of Init_θ, successful rebinding sets

$$\mathcal{H}_t^{\text{pix}} = \text{Seed}_\theta(I_t, \widehat{M}_t). \tag{13}$$

The next propagation, before any further rebinding, is then

$$(\widetilde{M}_{t+1}, \widetilde{\mathcal{H}}_{t+1}^{\text{pix}}) = \mathcal{P}_\theta\Big(I_{t+1}; \text{Seed}_\theta(I_t, \widehat{M}_t)\Big). \tag{14}$$

Once a candidate mask is accepted, it reinitializes Cutie's pixel memory. Subsequent masks are therefore propagated from the corrected target rather than from the previous pixel state. The trajectory and reference histories still determine which candidate mask is selected. If verification fails, no external mask is supplied, and ordinary propagation or low-confidence recovery continues. Algorithm 1 summarizes this procedure. This is the only interface through which TrackEVT reinitializes the propagator.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We use three established benchmarks and one identity-stress test. **Gym-UnrealCV** contains UrbanCity and ComplexRoom with four distractors and Parking Lot with two. We run 100 episodes in each scene and report accumulated reward (AR), episode length (EL), and success rate (SR) (Zhong et al., 2024). **DAT** contains six daytime scenes for aerial vehicle tracking. We evaluate four relative tracking directions, using 40 episodes per scene for each reported metric: cumulative reward (CR) and tracking success rate (TSR) (Sun et al., 2025). **EVT-Bench** tests indoor person following from the front RGB view. We report SR, tracking rate (TR), and collision rate (CR) for single-target tracking (STT) and distracted tracking (DT) (Wang et al., 2025b). **Cross-ID** combines similar target–distractor clothing, repeated crossings, foreground occlusion, and post-occlusion recovery. We run 90 episodes and report full-budget completion (Hard SR), distractor switches at reached crossings (Crossing IDSW), and correct-identity recovery within 30 frames (PN-Rec@30). Appendix D gives the complete protocol and metric definitions.

Table 1: Distractor-rich Gym-UnrealCV results. D denotes the number of distractors.

Method	Publication	Training time	UrbanCity (4D)			Parking Lot (2D)			ComplexRoom (4D)			Average		
			AR↑	EL↑	SR↑	AR↑	EL↑	SR↑	AR↑	EL↑	SR↑	AR↑	EL↑	SR↑
AD-VAT (Zhong et al., 2019)	ICLR'19	12h	32	204	0.06	43	232	0.13	16	223	0.16	30	220	0.12
DiMP (Bhat et al., 2019)	ICCV'19	–	170	348	0.32	111	271	0.24	97	307	0.26	126	309	0.27
SARL (Luo et al., 2020)	TPAMI'20	24h	74	221	0.16	53	237	0.12	22	263	0.15	50	240	0.14
AD-VAT+ (Zhong et al., 2021a)	TPAMI'21	24h	89	245	0.11	35	166	0.08	35	262	0.18	53	224	0.12
TS (Zhong et al., 2021b)	ICML'21	26h	227	381	0.51	186	331	0.39	250	401	0.54	221	371	0.48
FAn (Maalouf et al., 2024)	RA-L'24	–	167	334	0.30	126	301	0.28	189	351	0.29	161	329	0.29
EVT (Zhong et al., 2024)	ECCV'24	1h	272	472	0.92	192	425	0.63	354	479	0.88	273	459	0.81
FAn+SAM2 (Ravi et al., 2024)	ICLR'25	–	201	407	0.55	170	349	0.40	262	422	0.61	211	393	0.52
TrackVLA (Wang et al., 2025b)	CoRL'25	15h	–	476	0.92	–	467	0.90	–	479	0.91	–	474	0.91
OA-VAT (Sun et al., 2026)	CVPR'26	15h	385	486	0.95	392	482	**0.93**	392	481	0.92	390	483	0.93
TrackEVT (ours)	–	**0h**	**407**	**498**	**0.99**	**393**	**486**	**0.93**	**407**	**498**	**0.99**	**402**	**494**	**0.97**

Table 2: Vehicle tracking in six daytime DAT scenes. Each Average column summarizes the three scenes in its block.

Method	Publication	CityStreet		Desert		Village		Average	
		CR↑	TSR↑	CR↑	TSR↑	CR↑	TSR↑	CR↑	TSR↑
D-VAT (Dionigi et al., 2024)	RA-L'24	48	0.260	47	0.260	44	0.220	46	0.247
GC-VAT (Sun et al., 2025)	NeurIPS'25	279	0.800	307	0.840	239	0.730	275	0.790
OA-VAT (Sun et al., 2026)	CVPR'26	310	0.830	311	0.830	307	0.830	309	0.830
TrackEVT (ours)	–	**360**	**0.978**	**364**	**0.998**	**335**	**0.936**	**353**	**0.971**

Method	Publication	Downtown		Lake		Farmland		Average	
		CR↑	TSR↑	CR↑	TSR↑	CR↑	TSR↑	CR↑	TSR↑
D-VAT (Dionigi et al., 2024)	RA-L'24	9	0.060	46	0.260	13	0.070	23	0.130
GC-VAT (Sun et al., 2025)	NeurIPS'25	203	0.650	181	0.610	243	0.680	209	0.647
OA-VAT (Sun et al., 2026)	CVPR'26	**370**	0.990	318	0.840	310	0.830	333	0.887
TrackEVT (ours)	–	353	**0.998**	**370**	**0.998**	**359**	**0.998**	**361**	**0.998**

Implementation and comparisons. TrackEVT combines pretrained Cutie and YOLOE with appearance–motion association and bounded visual servoing. It uses neither task-specific training nor a learned action head. The target-maintenance procedure is shared across platforms. Only the appearance backend and control adapter change. We compare representative learned policies and modular trackers. Bold and underlined entries mark the best and second-best results, respectively. A dash indicates an unavailable measurement.

4.2 Main Results

Gym-UnrealCV. TrackEVT obtains a three-scene mean SR of 0.97, compared with OA-VAT's 0.93 (Table 1). The gain concentrates in UrbanCity and ComplexRoom, where four moving distractors repeatedly compete for the tracked state and TrackEVT reaches 0.99 SR. Parking Lot has two distractors and both methods reach 0.93. This scene dependence indicates that the advantage grows with sustained target–distractor interaction rather than uniformly inflating easier cases. Higher mean AR and EL (402 and 494 versus 390 and 483) are also consistent with fewer early failures. Figure 3 shows the corresponding pixel estimates near moving distractors.

DAT. Moving from ground-level person following to aerial vehicle tracking changes the target category, viewpoint, and action mapping. Without DAT-specific training, TrackEVT leads TSR in all six scenes and CR in five (Table 2). Even its lowest TSR is 0.936. The consistent transfer across streets, desert, village, lake, and farmland shows that the tracking formulation is not tied to pedestrians or a single camera geometry: the same state-maintenance principle adapts naturally once paired with the platform's generic identity backend and control adapter. Downtown, where TSR is slightly higher than OA-VAT's but CR is lower, further distinguishes robust target retention from uniformly maximizing every reward component.

Table 3: EVT-Bench results (%). Baseline scores follow Ye et al. (2026). Gray scores are not ranked. †: GroundingDINO (Liu et al., 2024) detector. ‡: SoM (Yang et al., 2023a) and GPT-4o (OpenAI, 2024) for visual grounding.

Method	Params.	STT			DT		
		SR↑	TR↑	CR↓	SR↑	TR↑	CR↓
Vision–Language–Action (VLA):							
Uni-NaVid (Zhang et al., 2024)	7B	53.3	67.2	12.6	31.9	50.1	21.3
TrackVLA (Wang et al., 2025b)	7B	85.1	78.6	1.7	57.6	63.2	5.8
Qwen-RobotNav (Zhang et al., 2026)	8B	78.6	89.7	5.7	–	–	–
IBVS† (Gupta et al., 2017)	–	42.9	56.2	3.8	10.6	28.4	**6.1**
PoliFormer (Zeng et al., 2024)	–	4.7	15.5	40.1	2.6	13.2	44.5
EVT (Zhong et al., 2024)	0.7B	24.4	39.1	42.5	3.2	11.2	47.9
EVT‡ (Zhong et al., 2024)	–	32.5	49.9	40.5	15.7	35.7	53.3
TrackEVT	**0.4B**	**53.7**	**68.1**	**3.4**	**28.9**	**52.2**	14.5

Table 4: Cross-ID identity stress test. Repeated crossings and occlusion by identically dressed distractors test recovery of the original target. All baseline results are reproduced by us under the same evaluation protocol.

Method	AR↑	Hard SR↑
AD-VAT (Zhong et al., 2019)	49	0.089
FAn (Maalouf et al., 2024)	113	0.211
FAn+SAM2 (Ravi et al., 2024)	151	0.289
EVT (Zhong et al., 2024)	143	0.278
OA-VAT (Sun et al., 2026)	171	0.344
TrackEVT	**321**	**0.733**

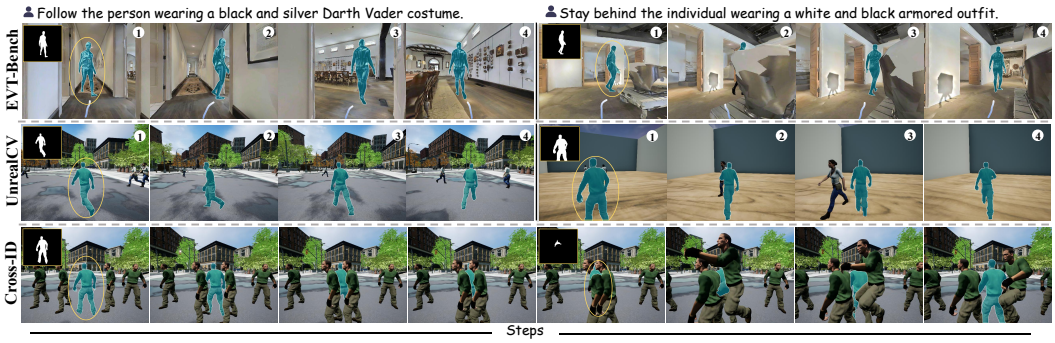


Figure 3: **Qualitative tracking across three environments.** Cyan overlays show target pixel estimates, with frames ordered temporally within each sequence. EVT-Bench illustrates indoor turns and partial occlusion. Gym-UnrealCV shows following amid distractor motion. Cross-ID highlights target continuity through same-clothing crossings.

EVT-Bench. TrackEVT leads the non-VLA methods in SR and TR on both STT and DT (Table 3). Its 53.7% and 28.9% SR exceed the strongest non-VLA baselines by 10.8 and 13.2 percentage points. The larger DT gain is consistent with identity maintenance becoming more valuable during pedestrian crossings. DT TR reaches 52.2%, but its 14.5% collision rate exceeds IBVS's 6.1%, showing that identity maintenance does not replace safety reasoning. TrackVLA scores higher, but uses a 7B policy trained on 1.7M samples (Wang et al., 2025b), versus TrackEVT's 0.4B frozen components and no task-specific training.

Cross-ID. We use Cross-ID as a targeted diagnostic of identity takeover rather than a general benchmark. Its repeated crossings test whether recovery preserves the originally selected identity. Under the same evaluation protocol, TrackEVT improves completion and identity metrics over the reproduced baselines (Table 4). Relative to OA-VAT, it lowers Crossing IDSW from 0.792 to 0.307 and raises PN-Rec@30 from 0.323 to 0.683 (Appendix D), connecting better completion to identity preservation and recovery. The matched ablations below test where these gains arise.

4.3 Real-World Deployment

We deploy TrackEVT on a Unitree Go2 Pro across three indoor locations with different backgrounds and ambient illumination. A local computer relays its RGB stream to an RTX 4090 server and converts returned target estimates into forward and yaw commands (Appendix C.1).

Across 20 trials per method and task, success rates for EVT/OA-VAT/TrackEVT are 20%/55%/85% in Distractor and 65%/80%/100% on Winding Path (Figure 4).

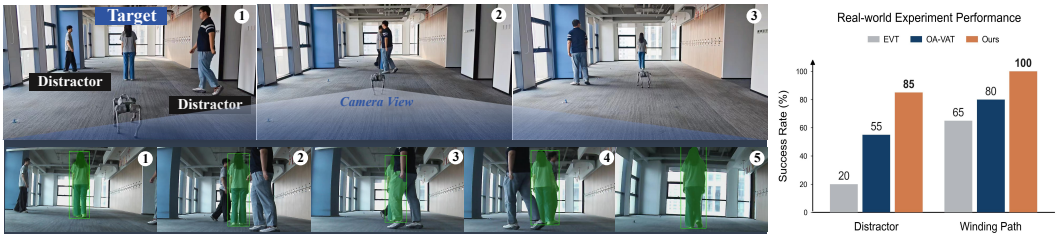


Figure 4: **Real-world robot evaluation.** Left: tracking through pedestrian crossings. Right: success rates over 20 trials per method and task across three indoor locations.

4.4 Ablation Studies

Table 5 compares P, T, and R under the same detector, audit interval, controller, seeds, and protocol. Cutie-safe uses the same immutable-anchor verification, current-frame mask recovery, and Cutie reset as the full TrackEVT system but omits competing trajectories, forming a matched initial-anchor verifier (Appendix E).

Table 5: Component ablations on UrbanCity and Cross-ID. P/T/R denote pixel propagation, persistent identity trajectories, and verified current-frame rebinding.

Configuration	Components			UrbanCity			Cross-ID				
	P	T	R	AR↑	EL↑	SR↑	AR↑	EL↑	Hard SR↑	IDSW↓	PN-Rec @30↑
Box baseline	–	–	–	367	469	0.86	153	236	0.322	0.863	0.268
Cutie-only	✓	–	–	383	479	0.92	238	329	0.511	0.611	0.470
Box + T	–	✓	–	374	474	0.88	207	293	0.467	0.702	0.319
Cutie + T	✓	✓	–	388	496	0.92	269	379	0.589	0.524	0.517
Cutie-safe	✓	–	✓	393	489	0.95	283	381	0.656	0.429	0.588
TrackEVT (full)	✓	✓	✓	**407**	**498**	**0.99**	**321**	**420**	**0.733**	**0.307**	**0.683**

Matched evidence for T and R. Three paired comparisons change one component at a time. Cutie-only versus Cutie + T isolates T without R and raises Hard SR from 0.511 to 0.589. Comparing Cutie-safe with the full system isolates T with R fixed. Hard SR rises from 0.656 to 0.733, IDSW falls from 0.429 to 0.307, and PN-Rec@30 rises from 0.588 to 0.683, showing that competing trajectories add beyond the matched initial-anchor verifier. Comparing Cutie + T with the full system isolates R with T fixed, raising Hard SR from 0.589 to 0.733 and PN-Rec@30 from 0.517 to 0.683. Verified rebinding therefore turns identity evidence into a persistent correction by reinitializing propagation. Together, the pairs attribute the gain to persistent auditing and gated executable-state updates.

Effect of pixel propagation. Cutie-only outperforms the box baseline in both scenes, and Box + T shows that identity tracks also help box tracking. The full system leads every Cross-ID metric and reaches 0.99 UrbanCity SR. A fixed-step rerun preserves the ordering (Appendix E), reducing the possibility that runtime-dependent sampling explains the gain.

5 Conclusion

TrackEVT couples dense pixel propagation for continuous control with persistent identity trajectories, admitting corrections only through a verified current-frame target mask that reinitializes propagation. Without task-specific training, it improves person following in distractor-rich environments, transfers to aerial vehicle tracking, and operates on a real Unitree Go2. The results show that target identity should constrain the state used for control rather than remain a by-product of localization. More broadly, separating fast action-oriented states from persistent semantic states may benefit other embodied perception–action tasks. A natural next step is to incorporate semantic world models for predictive scene modeling and more robust closed-loop tracking.

Ethics Statement

This work follows the ICLR Code of Ethics. The real-world trials were conducted in a controlled environment with volunteers. All participants provided informed consent for participation and publication of the recorded images, and the host institution approved both the experiments and image use. The robot was supervised throughout the trials, and the recordings were used only to evaluate target-following behavior.

Reproducibility Statement

Sections 3 and 4 describe the method and evaluation protocol. Appendices A–E provide pseudocode, inference settings, metric definitions, deployment details, and timing-controlled ablations. Source code, released configurations, and setup instructions are available in the anonymous repository at https://anonymous.4open.science/r/CodeReviewM8Q2/. The visual backbones use publicly released checkpoints, and benchmark resources for the released evaluations are obtainable from their original projects.

References

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.

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*, pp. 6182–6191, 2019.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9686–9696, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/iccv48922.2021.00951.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an Atkinson–Shiffrin memory model. In *European Conference on Computer Vision*, pp. 640–658, 2022.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5555–5564, 2021. doi: 10.1109/cvpr46437.2021.00551.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1316–1326, 2023a.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161, 2024.

Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023b.

Alessandro Devo, Alberto Dionigi, and Gabriele Costante. Enhancing continuous control of mobile robots for end-to-end visual active tracking. *Robotics and Autonomous Systems*, 142:103799, 2021. doi: 10.1016/j.robot.2021.103799.

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. doi: 10.1109/LRA.2024.3385700.

Meenakshi Gupta, Swagat Kumar, Laxmidhar Behera, and Venkatesh K. Subramanian. A novel vision-based tracking algorithm for a human-following mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(7):1415–1427, 2017. doi: 10.1109/TSMC.2016.2616343.

Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. FastReID: A PyTorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9664–9667, 2023. doi: 10.1145/3581783.3613460.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023. doi: 10.1109/iccv51070.2023.00371.

Jiahang Liu, Yunpeng Qi, Jiazhao Zhang, Minghan Li, Shaoan Wang, Kui Wu, Hanjing Ye, Hong Zhang, Zhibo Chen, Fangwei Zhong, et al. TrackVLA++: Unleashing reasoning and memory capabilities in VLA models for embodied visual tracking. *arXiv preprint arXiv:2510.07134*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Youzhi Liu, Li Gao, Liu Liu, Mingyang Lv, and Yang Cai. CoMaTrack: Competitive multi-agent game-theoretic tracking with vision-language-action models. *arXiv preprint arXiv:2603.22846*, 2026.

Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *European Conference on Computer Vision*, pp. 661–679, 2020.

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, 2020.

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 Robotics and Automation Letters*, 9(4):3283–3290, 2024.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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.

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 *Advances in Neural Information Processing Systems*, volume 38, pp. 147712–147747, 2025. doi: 10.52202/085713-4940.

Haowei Sun, Kai Zhou, Hao Gao, Shiteng Zhang, Jinwu Hu, Xutao Wen, Qixiang Ye, and Mingkui Tan. Instance-level visual active tracking with occlusion-aware planning. *arXiv preprint arXiv:2604.21453*, 2026.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24591–24602, 2025a.

Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1328–1338, 2019.

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*, 2025b.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649, 2017.

Kui Wu, Shuhang Xu, Hao Chen, Churan Wang, Zhoujun Li, Yizhou Wang, and Fangwei Zhong. VLM can be a good assistant: Enhancing embodied visual tracking with self-improving vision-language models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13154–13161, 2025. doi: 10.1109/IROS60139.2025.11246600.

Kui Wu, Hao Chen, Jinzhu Han, Haijun Liu, Churan Wang, Yizhou Wang, Zhoujun Li, Si Liu, and Fangwei Zhong. AdaTracker: Learning adaptive in-context policy for cross-embodiment active visual tracking. *IEEE Robotics and Automation Letters*, 2026. doi: 10.1109/LRA.2026.3692075.

Mao Xi, Yun Zhou, Zheng Chen, Wengang Zhou, and Houqiang Li. Anti-distractor active object tracking in 3d environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3697–3707, 2022. doi: 10.1109/TCSVT.2021.3107153.

Yuchen Xie, Xinyu Zhou, Kuangji Zuo, Yanshuo Lu, Fengrui Huang, Boyu Ma, and Jianfei Yang. USS: Unified spatial-semantic prompts for embodied visual tracking with latent dynamics learning. *arXiv preprint arXiv:2606.25880*, 2026.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023a. URL https://arxiv.org/abs/2310.11441.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023b.

Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36324–36336, 2022.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2491–2502, 2021.

Hanjing Ye, Tianle Zeng, Jiazhao Zhang, Shaoan Wang, Zibo Zhang, Weisi Situ, Yuchen Zhou, Yonggen Ling, and Hong Zhang. ReferTrack: Referring then tracking for embodied visual tracking. *arXiv preprint arXiv:2607.20061*, 2026.

Kuo-Hao Zeng, Zichen Zhang, Kiana Ehsani, Rose Hendrix, Jordi Salvador, Alvaro Herrasti, Ross Girshick, Aniruddha Kembhavi, and Luca Weihs. PoliFormer: Scaling on-policy RL with transformers results in masterful navigators. *arXiv preprint arXiv:2406.20083*, 2024.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024.

Jiazhao Zhang, Gengze Zhou, Hale Yin, Yiyang Huang, Zixing Lei, Qihang Peng, Haoqi Yuan, Jie Zhang, Xudong Guo, Xiaoyue Chen, et al. Qwen-RobotNav technical report: A scalable navigation model designed for an agentic navigation system. *arXiv preprint arXiv:2606.18112*, 2026.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. doi: 10.1007/s11263-021-01513-4.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pp. 1–21, 2022.

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. URL https://openreview.net/forum?id=HkgYmhR9KX.

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, 2021a. doi: 10.1109/TPAMI.2019.2952590.

Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. Towards distraction-robust active visual tracking. In *International Conference on Machine Learning*, pp. 12782–12792, 2021b. URL https://proceedings.mlr.press/v139/zhong21b.html.

Fangwei Zhong, Xiao Bi, Yudi Zhang, Wei Zhang, and Yizhou Wang. RSPT: Reconstruct surroundings and predict trajectory for generalizable active object tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3705–3714, 2023. doi: 10.1609/aaai.v37i3.25482.

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*, pp. 139–155, 2024.

LLM USAGE STATEMENT

Generative AI tools were used as general-purpose assistants to improve grammar and clarity while preserving the authors' intended meaning, to assist in drafting portions of several sections, to generate portions of the implementation code, and to provide feedback on experimental design and the interpretation and presentation of results. The research direction, hypotheses, method design, data collection, experimental execution, and final decisions and interpretations were determined by the authors. The authors reviewed and revised all AI-assisted text, and the relevant code was inspected and tested before use. Generative AI was not used to generate datasets, experimental observations, reported numerical results, or synthetic image content. Any AI-assisted code or figure formatting was reviewed against the author-provided data. The authors take full responsibility for the final text, claims, code, and artifacts.

A TRACKEVT IMPLEMENTATION DETAILS

This appendix specifies the inference procedure underlying Section 3. The distinction is between ordinary updates inside the pixel propagator and externally requested reinitialization: the latter, not every internal memory write, is constrained by identity-supported current-frame recovery. Network weights remain frozen throughout inference.

A.1 EVENT-LEVEL INFERENCE

Algorithm 1 summarizes one control step. An audit uses a saved image and its corresponding tracker box, not whichever box is current when the worker finishes. It updates trajectory memory in the worker and publishes a timestamped correction when warranted. The main loop consumes completed proposals without assuming that enqueueing and completion occur in the same step. A separate worker handles reference-based recovery after the lost-frame budget is reached. This path uses reference matching, while trajectory audits address identity competition during ongoing tracking.

Algorithm 1 TrackEVT: propagation and externally guarded rebinding

Require: Current RGB I_t, pixel memory $\mathcal{H}^{\text{pix}}_{t-1}$, identity worker, positive references $\mathcal{F}^{+}_{t}$, controller state

$(\widetilde{M}_t, \widetilde{\mathcal{H}}^{\text{pix}}_t) \leftarrow \mathcal{P}_\theta(I_t; \mathcal{H}^{\text{pix}}_{t-1})$
$c_t \leftarrow$ confidence returned by the pixel adapter (Appendix A.2)
$p_t \leftarrow$ CONSUMETRAJECTORYCORRECTION(t) ▷ Age check and box prediction at t
if a correction proposal p_t is available **then**
    $\widehat{M}_t \leftarrow \mathcal{R}(I_t; p_t, \mathcal{F}^{+}_{t})$
    if $\widehat{M}_t \neq \varnothing$ **then**
        $\mathcal{W}_t \leftarrow \text{Init}_\theta(I_t, \widehat{M}_t)$ and reset the lost counter
    else
        Retain the propagated state and do not reset from the proposal box
    end if
else
    Retain the propagated state
end if
Update the confidence-based tracking/recovery mode
if normal non-lost tracking and an audit is due **then**
    ENQUEUEAUDIT(I_t, current tracker box, t)
end if
if recovery is active **then**
    Enqueue the latest recovery image when due and poll a completed proposal
    Discard an over-age proposal. Otherwise, obtain $\widehat{M}_t$ via $\mathcal{R}$ and reinitialize only if valid
end if
Obtain the control box from the current pixel state or bounded fallback
Send bounded visual-servo commands, or zero velocity while recovery is pending

Normal tracking schedules full-frame audits every $K = 10$ environment steps with offset 5. Low-confidence and recovery stages do not enqueue new audits. An already queued audit may still finish. The audit queue has capacity one. The recovery queue has capacity two but discards superseded requests. Source timestamps are retained for both streams. Frames with no detections do not invoke trajectory association in this implementation.

A.2 Initialization and Pixel Reliability

An initially selected detector instance supplies the seed mask M_0. The identity descriptor of its crop initializes the immutable anchor a_0. Positive-reference descriptors initialize the separate set $\mathcal{F}_0^+$ used by reference matching. The pixel adapter retains the largest connected component of the seed and rejects empty support. Reinitialization creates a new Cutie inference processor, rather than clearing memory while retaining old object identifiers.

Let $P_t(x)$ be Cutie's foreground probability at pixel x. Define

$$M_t^{\mathrm{raw}}(x) = \mathbb{I}[P_t(x) \geq \tau_M], \qquad L_t = \mathrm{LCC}(M_t^{\mathrm{raw}}), \qquad \alpha_t = |L_t|/(HW). \tag{15}$$

The propagation-area test requires both nonempty support and a valid area ratio: $V_t^{\mathrm{pix}} = \mathbb{I}[|L_t| > 0 \text{ and } \alpha_{\min} \leq \alpha_t \leq \alpha_{\max}]$. With $r_t^{\mathrm{cc}} = |L_t|/|M_t^{\mathrm{raw}}|$ (zero for an empty raw mask), the returned confidence is

$$c_t = \begin{cases} \mathrm{clip}_{[0,1]}\left[\dfrac{1}{|L_t|}\displaystyle\sum_{x \in L_t} P_t(x)\,\dfrac{1 + r_t^{\mathrm{cc}}}{2}\right], & V_t^{\mathrm{pix}} = 1, \\ 0, & V_t^{\mathrm{pix}} = 0. \end{cases} \tag{16}$$

When $V_t^{\mathrm{pix}} = 1$, the adapter returns $\widetilde{M}_t = L_t$ and updates its cached mask box. Otherwise, it returns an empty mask and retains its previous box. A confidence below τ_{lost} increments the lost-frame counter. Geometric validity and confidence are distinct: a nonempty valid-area mask can still have low confidence.

These tests evaluate propagated outputs after Cutie's inference step. External reinitialization separately requires nonempty seed support. Ordinary internal memory updates follow the pretrained propagator.

A.3 Identity Association

Appearance and motion. Person tracking uses FastReID (He et al., 2023) for identity features and DINOv3 (Siméoni et al., 2025) for auxiliary appearance and reference matching. DAT uses DINOv3 for the identity interface without duplicating its auxiliary contribution. Identity descriptors are extracted from instance crops. Reference matching uses mask-suppressed backgrounds when available. Tracklets retain appearance history and a Kalman motion state, while the original-target anchor a_0 remains fixed. Together, these cues maintain target ownership across observations rather than selecting an instance independently in each frame.

Assignment and target reservation. The partial matching in Equation 5 is implemented with Hungarian assignment and zero-utility dummy matches for unassigned tracks and detections. Target reservation is a separate post-assignment step: an unmatched target may claim a candidate supported by its original anchor, displacing a competing assignment. After reservation, the affinity is the pre-update pairwise score between the target track and the reserved detection. Matched tracklets update their motion and appearance histories. Unmatched detections create tracklets, and inactive tracklets expire.

A.4 Correction Proposals and Current-Frame Rebinding

Identity-supported proposals. An audit compares the detection associated with the propagated region against the instance assigned to the target trajectory. Appearance and motion provide the association evidence. Proposed ownership changes are screened against the original-target anchor before requesting rebinding. This path addresses identity competition while a target candidate remains observable. Absence alone is not a rebinding request. Sustained low pixel confidence activates the separate reference-based recovery path.

Each proposal retains its source timestamp. A trajectory proposal's motion state is advanced to the current step to obtain a spatial hint, rather than an independent control target. The recovery worker instead matches visible candidates to positive-reference descriptors, first within a predicted region and then over the full image when needed.

Current-frame instantiation. Both proposal sources pass an age check and use the same current-frame reinitialization interface. A trajectory correction obtains a fresh instance mask from the latest observation. A recovery mask is reused only when it already belongs to that observation. Otherwise, it is extracted again. Spatial compatibility with the proposal and reference appearance select the candidate. Its largest connected component supplies the seed. Rebinding proceeds only with accepted, nonempty support, initializing a new pixel processor from the current image. Thus, verification specifies the evidence and image support required for this state update.

A.5 Control and Fallback

During reliable propagation, control uses $B(M_t)$ from Equation 3. Immediately after a successful correction, the implementation uses the detection box paired with the verified current mask. Subsequent propagation again returns the tight mask box. Thus, the constraint rules out an *unverified or stale proposal box* as a standalone replacement, not all use of detector geometry.

Let B_t^{ctrl} be this selected control box. Front-view errors are $e_t^x = c_x(B_t^{\mathrm{ctrl}}) - W/2$ and $e_t^s = s^\star HW - |B_t^{\mathrm{ctrl}}|$. Equation 3 gives the normalized representation. For the same control box, the horizontal and area errors are scaled by $W/2$ and HW, respectively. The controller gains are defined for these pixel-coordinate and pixel-area units. For either error, the controller uses

$$\mathrm{PID}(e_t) = K_p e_t + K_i \sum_{k \le t} e_k \Delta t + K_d (e_t - e_{t-1})/\Delta t. \tag{17}$$

Yaw follows e_t^x, and forward speed follows e_t^s with a scale dead zone and a minimum nonzero output outside that zone.

Before the low-confidence budget is exhausted, the adapter may retain its last box when the propagated mask is invalid, or use the current low-confidence box with control-side confidence-aware filtering. This grace period is not an identity confirmation. On entering recovery, the default implementation sends zero velocity while a short-term prediction guides candidate search. A bounded timeout returns control to re-detection. The optional yaw-only prediction controller is not part of the default reported configuration. The control-side filter is distinct from the per-tracklet Kalman filters.

A.6 Implementation Configuration

Table 6 lists the principal settings of the frozen implementation. Candidate masks use YOLOE-11L-seg and pixel propagation uses Cutie-base-mega. The simulator person identity configuration requests the FastReID SBS-S50 MOT17 checkpoint. DAT uses the shared DINOv3 ViT-L/16 extractor at input size 384. The selected backend is recorded per run and is not inferred from a dataset name. Tiny control images are upscaled to a minimum side of 480 for Cutie inference. Scene and platform adapters are separate from the state-maintenance logic. The Go2 deployment uses DINOv3 identity features and a different control adapter, detailed in Appendix C.1. The accompanying project includes the corresponding runtime configuration files.

Table 6: Principal TrackEVT inference settings. Scene-specific simulator and action-adapter settings are reported with the corresponding experiments.

Component	Shared / Gym-UnrealCV	DAT override
Input / detector	160×120, YOLOE conf. 0.10	320×240, YOLOE conf. 0.001
Pixel state	$\tau_M = 0.50$, area ratio $[0.001, 0.65]$, $\tau_{\text{lost}} = 0.55$, memory every 5 frames	same
Bounded fallback	$N_{\text{lost}} = 3$, max prediction 15	same
Identity backend	FastReID (person)	DINOv3 (generic object)
Identity audit	$K = 10$, offset 5, queue capacity 1	same
Lost worker	newest-only queue (capacity 2)	same
Tracklet memory	capacity 12, TTL 160, gallery size 5	same
Current-mask write	required for trajectory correction and lost recovery	same

B External Reinitialization Interface

The reported Cutie configuration requires a current-frame instance mask for both trajectory corrections and reference-based recovery. A proposal must pass its source checks and age limit, and any earlier-frame mask is re-extracted in the initialization frame. Accepted, nonempty support seeds a new inference processor. A rejected proposal provides no external seed. Ordinary pixel propagation continues according to the tracking mode. This interface property concerns proposal-triggered reinitialization, distinct from Cutie's internal memory updates.

Parameter transfer. The settings in Table 6 are fixed inference defaults and safety or scheduling guards, not parameters fitted separately to each scene. Mask thresholds, recovery budgets, audit timing, queues, tracklet memory, and rebinding checks are shared across Gym-UnrealCV, Cross-ID, and DAT without scene-specific tuning. Transfer changes only the observation scale, detector operating point, category-appropriate identity backend, and platform control adapter. The number of listed constants therefore reflects explicit interfaces and safeguards rather than per-benchmark hyperparameter optimization.

C Additional Experimental Details

Evaluation and reporting. Gym-UnrealCV uses 100 episodes per scene, with moving targets in ComplexRoom. Its reported SR follows the benchmark evaluator. It is not substituted for Cross-ID's full-budget completion criterion. DAT uses six daytime scenes and four relative directions, with 10 episodes per direction *for each metric*. CR and TSR are collected in separate reward-mode runs, giving $6 \times 4 \times 10 \times 2 = 480$ runs. DAT's CR denotes cumulative reward (higher is better), whereas EVT-Bench's CR denotes collision rate (lower is better). DAT TSR sums binary target-in-view rewards over the rollout and normalizes by the maximum episode horizon (Sun et al., 2025). We report it on $[0, 1]$, not as the fraction of successful episodes. Table 3 presents STT and DT results. Unavailable measurements are shown as dashes.

Qualitative recordings. Figure 3 uses selected closed-loop rollouts, not paired baseline comparisons. RGB is rendered at 640×480. ComplexRoom retains 160×120 tracking input with masks resized for display. Other shown sequences use 640×480 input and DINOv3 identity features. The EVT-Bench recording adapter additionally permits bounded rotational search during target loss. These visualization settings are distinct from the quantitative protocols. The images do not isolate the contribution of rebinding or establish aggregate success rates.

Precision. Rewards and episode lengths are displayed as integers. Gym-UnrealCV SR uses two decimals, DAT and Cross-ID rates use three, and EVT-Bench percentages use one. Main-text gains and parenthesized gains in Table 8 are computed from displayed values. Unavailable entries are excluded from rankings.

C.1 Real-World System and Control Interface

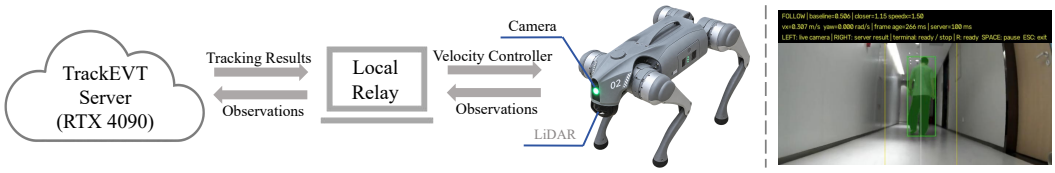


Figure 5: **Real-world system and control flow.** The local relay sends robot observations to the TrackEVT server, receives tracking results, and converts them into velocity commands for the Go2 Pro. The right panel shows the live interface with the returned target mask and bounding box.

Hardware and communication. As shown in Figure 5, we use a Unitree Go2 Pro (firmware 1.1.7), its built-in forward-facing RGB camera, a local Windows computer, and an NVIDIA RTX 4090 server. The local computer connects to the robot Wi-Fi and receives the 1280×720 camera stream over WebRTC through the community GitHub driver `unitree_webrtc_connect` (version 2.2.0). In use, the client establishes the WebRTC connection, enables the video channel, and registers a frame callback. The latest frame is sent to the server over SSH, and the server returns the target box, confidence, mask-validity state, and annotated image to the local relay.

Tracking and control. YOLOE-11L-seg initializes the person instance, Cutie-base-mega propagates the pixel mask, and DINOv3 ViT-L/16 supplies identity features. The operator places the designated target near the image center and enters `ready`. The local controller maps horizontal center error to yaw and relative box-height error to forward velocity. Over 3,641 deployment updates, the RTX 4090 server sustained 9.8 processed frames/s, while latest-frame, single-request relaying yielded a median frame-to-reply latency of 172 ms without accumulating stale observations. The accompanying project also includes the corresponding relay and control-interface implementation.

D Cross-ID: Additional Metrics and Protocol

Cross-ID is a purpose-built identity-stress scenario in UrbanCity with four distractors. Each episode combines four consecutive challenges: (i) the target and the designated crossing distractor share the same clothing, while their orientations and visible regions vary; (ii) the distractor repeatedly crosses the target's path; (iii) these foreground crossings produce strong occlusion; and (iv) after occlusion, the agent must recover the originally designated instance and continue closed-loop following, rather than switch to the distractor. The reference protocol uses 90 episodes (seeds 801–890), with a 500-step budget per episode. To make this purpose-built protocol auditable, the accompanying project provides its executable scenario profile, seeds, episode horizon, and termination rules. The metric definitions are given below. Table 7 contains the available extended summaries for the two listed configurations, not a complete expansion of the main table. Other methods are omitted where these records are unavailable.

Table 7: **Additional Cross-ID metrics for the two available summaries.** $\pm$: 95% confidence-interval half-width.

Method	Publication	AR↑	EL↑	Hard SR↑	Crossing IDSW↓	PN-Rec @30↑
OA-VAT (Sun et al., 2026)	CVPR 2026	171 ± 41	264	0.344	0.792	0.323
TrackEVT	Ours	321 ± 32	420	0.733	0.307	0.683

Configuration and data scope. The OA-VAT row uses our complete reproduction of its released code. OA-VAT and TrackEVT use the same Cross-ID evaluation configuration, with 90 episodes each.

Metrics. AR is accumulated reward and EL is episode length. Hard SR is the fraction of episodes completing all 500 steps. Crossing IDSW is the proportion of reached crossing opportunities at

which the executable state switches to a distractor. PN-Rec@30 is the fraction of scheduled crossing-recovery opportunities for which the correct identity is recovered for five consecutive frames within 30 frames. Opportunities not reached after early termination count as failures.

E Ablation Protocols

Table 8: Extended trajectory ablation. Parentheses: gains over Cutie-safe in the same scene, computed from displayed values (SR in percentage points).

Scene	Configuration	AR↑	EL↑	SR↑
UrbanCity	Cutie-safe	393	489	0.95
	TrackEVT	**407** (+14)	**498** (+9)	**0.99** (+4.0)
Parking Lot	Cutie-safe	363	456	0.86
	TrackEVT	**393** (+30)	**486** (+30)	**0.93** (+7.0)
ComplexRoom	Cutie-safe	297	475	0.92
(dynamic)	**TrackEVT**	**407** (+110)	**498** (+23)	**0.99** (+7.0)

Component configurations. Table 5 uses 100 UrbanCity episodes (seeds 101–200) and 90 Cross-ID episodes (seeds 801–890), each with a 500-step budget. UrbanCity reports release-protocol SR. Cross-ID reports Hard SR, $n_{\text{full}}/90$, rounded to three decimals. These criteria are not interchanged. The box baseline combines ORTrack with YOLOE/appearance recovery rather than using a detector alone. Cutie-only removes safe rebinding but retains legacy proposal-based resets. Cutie + T adds persistent trajectories without R. Cutie-safe uses the same scheduled, immutable-anchor-based current-frame recovery as the full system but omits persistent multi-object trajectories. The full system combines P, T, and R.

Timing-controlled Cross-ID check. We rerun the three decisive P-based configurations on the same 90 seeds and 500-step budget with a fixed simulator/control interval. The detector, controller, audit schedule, and implementation snapshot are shared. Only T and R vary. Table 9 shows that the full configuration remains best without runtime-dependent sampling.

Table 9: Cross-ID timing control. Release-timing values are the main ablation results. Fixed-step values use one shared simulator/control interval.

Configuration	Components			Release timing			Fixed step		
	P	T	R	AR↑	EL↑	Hard SR↑	AR↑	EL↑	Hard SR↑
Cutie + T	✓	✓	–	269	379	0.589	281	398	0.611
Cutie-safe	✓	–	✓	283	381	0.656	306	393	0.667
TrackEVT (full)	✓	✓	✓	**321**	**420**	**0.733**	**338**	**419**	**0.722**

Box + T reproduction. This configuration retains the ORTrack box state, detector/reference-based recovery, and visual servoing, and enables persistent identity association and trajectory-supported box correction. It uses neither Cutie nor the current-frame pixel-rebinding mechanism. Thus, "T-only" refers to adding only T to the box baseline, not removing the underlying visual tracker.

Extended trajectory comparison. Table 8 compares Cutie-safe and the full system across three scenes. Values for the full system match the main benchmark table. Cutie-safe averages 100 seeds (101–200) per scene and uses the EVT-native evaluator for dynamic ComplexRoom.

Gym-UnrealCV uses 10-Hz adaptive time dilation, while DAT advances through an action–observation handshake. The fixed-step Cross-ID rerun preserves the key ablation ordering with simulator evolution held constant during inference.