EMBODIED DRONE TRACKING IN THE WILD: BENCHMARK AND BASELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied drone tracking requires an agent to follow a moving target while actively controlling its viewpoint and avoiding obstacles. Drones offer broad views and flexible altitude adjustment, yet target motion and interrupted visibility make this coupling challenging. Existing benchmarks provide limited coverage of substantial target elevation changes and complex occlusions under a common promptable tracking protocol. We introduce the Embodied Drone Tracking Benchmark (EDT-BENCH), built on UnrealZoo to address this gap. It supports language, point, bounding box, mask, and template initialization across diverse environments. The training corpus contains 5,670 expert trajectories and their perturbed counterparts from 81 scenes, providing aligned supervision for tracking and recovery. We further propose EDT, a simple baseline built on frozen SAM2 representations. Its action head estimates relative target pose from current observations and temporal memory, then uses uncertainty-aware fusion to predict drone waypoints. The memory branch retains target information during occlusion, while spatial prompts reduce ambiguity when similar distractors are present. A unified closed-loop protocol evaluates tracking robustness and generalization across scenes and target categories. Together, EDT-Bench and EDT provide a reproducible closed-loop framework for evaluating prompt-conditioned tracking, spatial tracking accuracy, recovery from occlusion and viewpoint perturbations, and generalization across scenes and target categories.

1 INTRODUCTION

Drones can follow moving targets from viewpoints that are difficult for ground agents to reach. Their broad field of view, flexible motion, and ability to change altitude make them useful for inspection, search and rescue, and dynamic monitoring. Realizing these advantages requires an agent to choose its own observations: each flight command changes the target's scale, visibility, and position in the next image. Embodied drone tracking therefore couples instance recognition with three-dimensional motion and collision avoidance.

This coupling creates several challenges. A distant target may occupy only a few pixels, while changes in target speed require rapid control adjustments. Similar moving objects can induce identity switches, and buildings, vegetation, or other actors can hide the target. Flying closer can improve visual detail but also increase exposure to obstacles. Changing altitude can restore visibility or provide clearance, yet alters target resolution and viewing geometry. DeTrack highlights this interaction between altitude, perception, and flight behavior (Hu et al., 2026). A useful evaluation must consequently measure both whether the correct target is followed and whether the drone maintains an appropriate spatial relationship to it.

Existing benchmarks address complementary aspects of tracking. Passive aerial datasets evaluate localization along recorded camera trajectories (Zhang et al., 2023). Embodied visual tracking benchmarks couple recognition and navigation, often through language-conditioned policies (Zhong et al., 2024; Wang et al., 2025; Liu et al., 2025). Drone active-tracking benchmarks further introduce aerial control and three-dimensional scene interaction (Sun et al., 2024; Hu et al., 2026; Zhang et al., 2026). We focus on a shared setting in which target-initialization modality, scene variation, and recovery from displaced viewpoints can be studied with aligned observations and action supervision. This is particularly relevant when language describes several similar instances: a point, box, or mask

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.

can identify the intended instance directly, but its effect on subsequent closed-loop tracking needs to be measured.

We introduce the Embodied Drone Tracking Benchmark (EDT-BENCH), built on UnrealZoo (Zhong et al., 2025). Each episode supports language, point, bounding-box, mask, and visual-template initialization for the same target. The benchmark combines varied target appearance and speed with similar distractors, altitude changes, complex occlusions, and changes in illumination and weather. Its training corpus contains 5,670 expert episodes and 5,670 corresponding perturbed episodes across 81 scenes. The latter preserve target motion while changing the drone's position or viewpoint, with recovery labels expressed in the perturbed frame. A common action interface predicts ten four-dimensional waypoints over two seconds.

Evaluation separates single-target tracking (STT), distractor tracking (DT), occlusion tracking (OT), and unseen-target-category tracking (UTT). The 1,024 test episodes comprise 512 episodes in training-seen scenes and 512 in 20 training-unseen scenes. Target trajectories are fixed across methods, including manually collected OT trajectories that preserve the intended encounters with occluders. The drone still chooses and executes its actions online; only the prescribed target motion is replayed. We additionally define mean relative position error (MRPE), which measures three-dimensional deviation from a category-dependent following position and complements tracking and collision rates.

We provide EDT, a simple baseline built on frozen SAM2 representations (Ravi et al., 2025). A prompt adapter initializes the target, and an Uncertainty-Aware Measurement–Memory Fusion head (UMF-HEAD) estimates relative pose from the current observation and temporal memory. The head combines these estimates using dimension-wise uncertainty and decodes future waypoints. Our contributions are:

- EDT-BENCH, a five-prompt benchmark that combines diverse drone-tracking conditions with 11,340 paired training episodes and 1,024 test episodes across seen and unseen scenes.
- A reproducible evaluation protocol with fixed target trajectories, manually collected occlusion cases, four task subsets, and MRPE for category-dependent spatial tracking accuracy.
- EDT, a frozen-SAM2 baseline with uncertainty-aware measurement–memory fusion, together with controlled comparisons of initialization prompts and recovery supervision.

2 RELATED WORK

Embodied visual tracking. EVT combines foundation-model perception with offline reinforcement learning and uses a privileged controller to collect demonstrations (Zhong et al., 2024; Kumar et al., 2020). TrackVLA couples target recognition with trajectory generation and introduces EVT-Bench for language-conditioned tracking (Wang et al., 2025). TrackVLA++ further incorporates spatial reasoning and target-identification memory (Liu et al., 2025). These studies motivate the use of temporal representations for maintaining identity during active observation. EDT-BENCH extends the comparison of target specification to five prompt modalities under a common drone-control interface.

Drone tracking benchmarks. WebUAV-3M provides large-scale real aerial video for passive tracking (Zhang et al., 2023). DAT evaluates drone active tracking in Webots and SUMO, with discrete and continuous controls (Sun et al., 2024). DeTrack uses Unreal Engine and AirSim to study altitude-aware tracking under occlusion and changing target scale (Hu et al., 2026; Shah et al., 2018). UAV-Track VLA predicts continuous four-degree-of-freedom trajectories from language and visual history, with relative-pose supervision (Zhang et al., 2026). Related work also considers occlusion-aware planning and anomaly-aware assistance (Sun et al., 2026; Wu et al., 2025; Li et al., 2026). Our emphasis is the combination of multi-prompt initialization, paired expert and perturbed observations, and curated evaluation of altitude changes and complex occlusions. These individual challenges are shared with prior work; the benchmark organizes them within one data and evaluation interface.

Promptable video segmentation. SAM2 segments objects in images and videos through spatial prompts and streaming memory (Ravi et al., 2025). EDT uses its mask, image, memory, and object-

Table 1: Representative tracking benchmarks. Scale values use each source's reporting unit and are not directly comparable. L: language; P: point; B: bounding box; M: mask; T: visual template. The last column indicates paired expert/perturbed observations with recovery labels.

Benchmark	Reported scale	Target specification	Paired
WebUAV-3M	4,500 videos	Visual, L, audio	No
DAT	24 environment conditions	Environment-selected	No
DeTrack	11,368 target trajectories	T	No
EVT-Bench	25,986 episodes	L	No
UAV-Track	892,756 frames	L	No
EDT-BENCH	11,340 train; 1,024 test episodes	L, P, B, M, T	Yes

pointer representations without updating the backbone. Point, box, and mask are native spatial inputs; language and template inputs require an initialization adapter that resolves them to a spatial prompt. This distinction separates the benchmark's five-prompt interface from the inputs directly supported by SAM2. The learned action head then maps the resulting representations to drone waypoints.

3 EMBODIED DRONE TRACKING BENCHMARK

3.1 TASK AND ACTION INTERFACE

At time t, a policy receives an RGB observation I_t, a target prompt π supplied at initialization, and its online memory. It predicts

$$\mathcal{A}_t = \{a_t^k\}_{k=1}^{10}, \qquad a_t^k = [\Delta x_t^k, \Delta y_t^k, \Delta z_t^k, \mathrm{wrap}(\Delta\psi_t^k)] \in \mathbb{R}^4. \tag{1}$$

Each waypoint describes a future pose at $0.2k$ seconds in the current drone frame, using meters for translation and degrees for yaw. The policy runs at 5 Hz with a two-second prediction horizon. A shared low-level controller converts neighboring waypoint differences into translation velocity and yaw rate.

The data also include the target's relative pose,

$$r_t = [x_t^{\mathrm{fwd}}, y_t^{\mathrm{right}}, z_t^{\mathrm{up}}, \mathrm{wrap}(\psi_t^{\mathrm{target}} - \psi_t^{\mathrm{drone}})]. \tag{2}$$

Translation is expressed in the drone's yaw-aligned frame. The yaw label is the difference between actor orientations, rather than the line-of-sight bearing used by a controller to face the target. These labels supervise training and are not inputs to the learned policy.

3.2 SCENE COVERAGE AND MULTI-PROMPT INITIALIZATION

Diversity and flight geometry. EDT-BENCH comprises 101 UnrealZoo scenes: 81 training-seen scenes and 20 held-out scenes. It covers changes in target appearance and speed, same-category or visually similar distractors, lighting, and weather. We emphasize two interacting factors: *altitude variation*, which changes target scale and the visibility of ground structure, and *complex occlusion*, which requires maintaining identity while visibility is interrupted by static geometry or moving actors. Their combination tests whether perception and control remain consistent as the viewpoint changes. The camera provides 256×256 RGB observations; the collection configuration fixes aircraft roll and pitch and uses a downward-looking camera.

Human, robot-dog, motorcycle, and car targets appear in the training corpus. Animals are held out for UTT, with small and large animals assigned different reference following positions. Category novelty and scene novelty are separate axes: an animal can appear in a seen scene without having appeared in the training data.

Five ways to specify an instance. Every episode associates the same target with a language description, a point, a bounding box, a mask, and a visual template. The point lies inside the initial target mask, the box encloses that mask, and the template is a crop from the initial observation.

Table 2: Corpus and evaluation specification. Seen/unseen denotes scene exposure during policy training. The test episodes are distinct from the training demonstrations, including in seen scenes.

Component	Specification
Training scene pool	81 seen scenes
Training episodes	5,670 expert + 5,670 paired perturbed
Test episodes in seen scenes	512, drawn from the 81-scene pool
Test episodes in unseen scenes	512, drawn from 20 held-out scenes
Training target categories	Human, robot dog, motorcycle, car
Held-out target category	Animal (small and large)
Evaluation subsets	STT, DT, OT, UTT; each seen and unseen
Prompt modalities	Language, point, bounding box, mask, template
Action interface	10×4 waypoints; 2 s horizon; 5 Hz
Evaluation measures	SR ↑, TR ↑, CR ↓, MRPE ↓

A language prompt describes the intended instance. Point, box, and mask provide explicit spatial localization, which is useful when several visible objects share the same appearance or semantic description. Templates convey appearance; when their original image coordinates are withheld, they require matching the crop to the current image. All prompts are initialization-only in the primary protocol, so later ground-truth masks or corrective clicks are not supplied. Valid prompt combinations are supported as supplementary settings, but the primary comparisons evaluate each modality separately.

3.3 Expert Data and Paired Recovery Supervision

The training corpus contains $81 \times 70 = 5{,}670$ expert episodes and one perturbed counterpart per episode. During collection, a privileged expert follows the target's navigation route using four PID controllers for forward, lateral, vertical, and yaw motion. Depth-based visibility checks and local obstacle handling guide collection: the expert slows near obstacles and can select lateral or vertical escape directions. These signals are used to collect and validate demonstrations, rather than provided to the learned policy.

The paired pass keeps target and distractor motion time aligned while applying smooth offsets to the drone's position, yaw, or both. Observations are rendered again from the perturbed camera poses. If $T_{\mathrm{pert}}(t)$ is the perturbed drone pose and $T_{\mathrm{exp}}(t+0.2k)$ a future expert pose, recovery supervision is

$$a_{\mathrm{rec}}(t,k) = \mathrm{pose}_4\left[T_{\mathrm{pert}}(t)^{-1} T_{\mathrm{exp}}(t+0.2k)\right], \quad k = 1, \ldots, 10, \tag{3}$$

where pose_4 extracts translation and wrapped yaw in the convention of Eq. 1. Thus, the label expresses the corresponding expert future in the current perturbed frame. It does not assume an independently replanned recovery path. RGB, mask, box, visibility, actor poses and velocities, executed motion, and future waypoints share a timestamp, supporting perception, relative-pose estimation, and action learning from the same episodes. Collection details are given in Appendix A.

3.4 Test Subsets and Fixed Target Trajectories

The test set contains 1,024 episodes: 512 in the 81 training-seen scenes and 512 in 20 training-unseen scenes. Within both scene splits, **STT** measures tracking without a similar dynamic distractor; **DT** introduces same-category or visually similar moving distractors; **OT** focuses on occlusion by scene geometry or dynamic actors; and **UTT** evaluates the held-out animal category. The seen split measures transfer to new episodes in familiar environments, whereas the unseen split additionally tests scene generalization.

We use fixed target trajectories instead of resampling target navigation on every evaluation run. This choice is particularly important for OT: manually collected target routes preserve encounters with selected occluders, avoiding runs that accidentally omit the intended challenge. Replaying the same routes also supports matched comparisons between policies and prompts. Fixed target motion is compatible with closed-loop evaluation because the drone's trajectory, observations, and actions depend on the evaluated policy. In particular, policies can experience different occlusion durations or

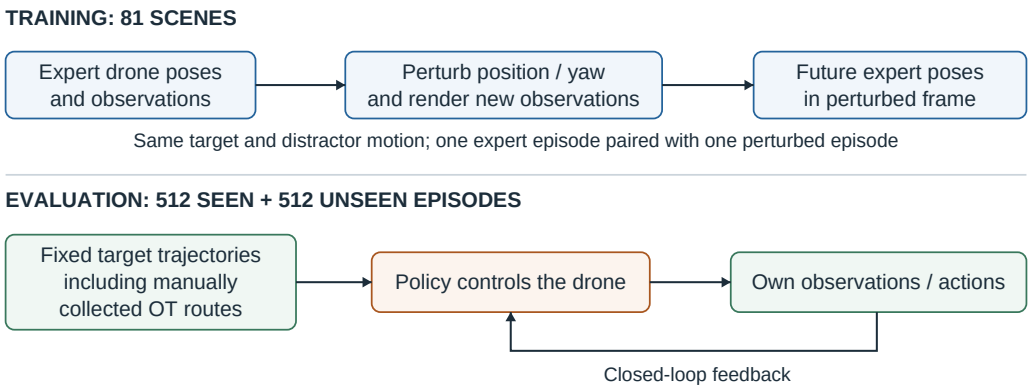


Figure 1: Paired supervision and closed-loop evaluation. Training changes the drone viewpoint while preserving target motion and transforms future expert poses into the perturbed frame. Test target trajectories are replayed across methods, while each policy controls its own drone and receives its own observations.

avoid an occlusion by changing viewpoint; fixed routes do not force identical visibility. Prescribed target paths are also used in DeTrack (Hu et al., 2026); here, they serve the quality control of the manually collected OT set and the consistency of prompt comparisons.

3.5 Evaluation Metrics

We report four complementary measures. **Success rate (SR)** is the fraction of successful tracking episodes. **Tracking rate (TR)** measures the fraction of evaluated time steps satisfying the correct-target tracking condition. **Collision rate (CR)** is the fraction of episodes with a collision or collision-induced termination. SR and TR are higher-is-better; CR is lower-is-better. Operational success and visibility thresholds must be fixed in the evaluation configuration and shared across methods.

Mean relative position error (MRPE). Let (x_i, y_i, z_i) denote the target's forward distance, lateral offset, and the drone's height above the target at time step i. Forward and lateral axes follow the drone's yaw. The height is defined as $z_i = h_i^{\text{drone}} - h_i^{\text{target}}$, so it is the negative of the upward target-coordinate component in Eq. 2. For category c, the desired relative position is $(x_c^*, 0, z_c^*)$. We compute

$$e_i = \sqrt{(x_i - x_c^*)^2 + y_i^2 + (z_i - z_c^*)^2}, \qquad \text{MRPE} = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{4}$$

Table 3 specifies the reference positions. MRPE is measured in meters; zero means that the drone maintains the desired relative position at every evaluated step. It uses the true simulator geometry, not the error of the policy's predicted relative pose, and does not include yaw. Measurements include occluded steps for which simulator states are available. We first compute MRPE per episode and then average across episodes within each task and scene split. For an early-terminated episode, N is its number of executed evaluation steps; MRPE is therefore reported with SR, TR, and CR to expose survivorship effects.

4 EDT: A SAM2-Based Tracking Baseline

4.1 From Target Prompts to Streaming Representations

EDT separates target initialization from action learning (Figure 2). An initialization adapter converts the selected modality into a spatial prompt accepted by SAM2: points, boxes, and masks are passed directly; language and templates require grounding or matching in the initial image. These adapters operate on the supplied prompt and image without access to the privileged target identifier. After initialization, frozen SAM2 provides mask logits M_t, image features F_t, past mask-memory features $\{H_{t-i}\}$, and the latest valid object pointer o_t. The same action head is used for all prompt modalities.

Table 3: Category-dependent desired relative positions for MRPE, in meters. The desired lateral offset is zero for all targets. Small and large animals are evaluation groups within the held-out animal category.

Target group	Forward distance x_c^*	Lateral offset y_c^*	Height z_c^*
Human	5	0	5
Robot dog	5	0	5
Small animal	5	0	5
Motorcycle	7.5	0	7.5
Car	10	0	10
Large animal	10	0	10

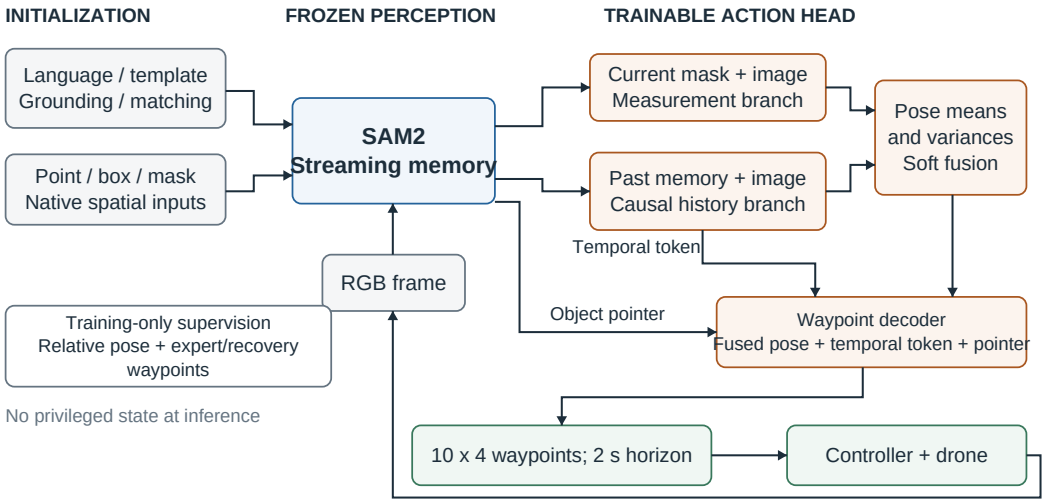


Figure 2: EDT overview. The five initialization modalities enter through a prompt adapter; language and template require conversion to spatial inputs. Frozen SAM2 supplies streaming representations. The trainable UMF-HEAD combines current and historical relative-pose estimates and predicts ten waypoints. Privileged pose and waypoint labels supervise training only.

4.2 MEASUREMENT–MEMORY FUSION

The measurement branch encodes the current mask and image features into a 256-dimensional token. The memory branch encodes the six most recent mask-memory maps together with the current image feature. A two-layer Transformer processes these tokens with temporal embeddings and validity masks, using only past and current observations. Each branch $b \in \{\text{obs}, \text{hist}\}$ predicts a four-dimensional relative-pose mean μ_t^b and log variance s_t^b, with s_t^b clamped to $[-6, 6]$.

For each dimension d, the head converts log variances to normalized precision weights,

$$[w_{t,d}^{\text{obs}}, w_{t,d}^{\text{hist}}] = \text{softmax}([-s_{t,d}^{\text{obs}}, -s_{t,d}^{\text{hist}}]). \tag{5}$$

Translation estimates are fused by weighted averaging. For yaw, we use a wrapped difference to avoid averaging across the $-180°/180°$ boundary:

$$\mu_{t,d}^{\text{fuse}} = w_{t,d}^{\text{obs}} \mu_{t,d}^{\text{obs}} + w_{t,d}^{\text{hist}} \mu_{t,d}^{\text{hist}}, \quad d \in \{x, y, z\}, \tag{6}$$

$$\mu_{t,\psi}^{\text{fuse}} = \text{wrap}\big(\mu_{t,\psi}^{\text{obs}} + w_{t,\psi}^{\text{hist}} \, \text{wrap}(\mu_{t,\psi}^{\text{hist}} - \mu_{t,\psi}^{\text{obs}})\big) . \tag{7}$$

This soft weighting allows current evidence and memory to contribute differently along each dimension without a separate visibility selector. It does not presume that predicted uncertainties are calibrated; this is an empirical question for the ablations.

4.3 Waypoint Prediction and Learning

The fused pose is normalized using training-set statistics and embedded into a 256-dimensional vector. Concatenating it with the temporal token and object pointer produces a 768-dimensional input to a waypoint MLP, which outputs the 10×4 action chunk. All SAM2 parameters remain frozen.

Let $\delta(\hat{y}, y)$ be the ordinary residual for translation and the wrapped residual for yaw. Denote training-set standard deviations for relative pose and horizon-specific waypoints by c_r and c_w. The branch loss is a normalized heteroscedastic Gaussian negative log likelihood,

$$\mathcal{L}_b = \frac{1}{2}\,\mathrm{mean}\left[e^{-s^b}\left(\frac{\delta(\mu^b, r)}{c_r}\right)^2 + s^b\right], \qquad b \in \{\mathrm{obs}, \mathrm{hist}\}. \tag{8}$$

The fused-pose and waypoint terms use Smooth L1 on the corresponding normalized residuals. The total objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{wp}} + 0.4\left(\mathcal{L}_{\mathrm{fuse}} + 0.25\mathcal{L}_{\mathrm{obs}} + 0.25\mathcal{L}_{\mathrm{hist}}\right). \tag{9}$$

Expert observations use expert futures; perturbed observations use Eq. 3. We train with AdamW at learning rate and weight decay 10^{-4}, clip length eight, head batch size four, mixed precision when available, and gradient-norm clipping at 1.0. Appendix B provides the encoder and decoder specifications.

5 Experiments

5.1 Protocol and Baselines

The evaluation is organized around three questions: how policies generalize across scene and target changes; how initialization prompts affect identity preservation under distraction; and how temporal fusion and paired perturbations affect recovery and spatial accuracy. All comparisons use the same test episode identifiers, target trajectories, initial drone states, controller, and termination rules. The primary learned-policy comparison fixes the prompt modality; five-way prompt comparisons are reported separately. Seen and unseen scores are retained for each of STT, DT, OT, and UTT.

We compare EDT with **SAM2-Mask PID**, which controls image centroid and scale from predicted masks; **Frame-BC MLP**, which predicts waypoints from a current-frame representation; **CNN-GRU BC** and **Temporal Transformer BC**, which add temporal context; and **ACT**, an action-chunking baseline (Zhao et al., 2023). A **PPO waypoint policy** tests online reinforcement learning under the same observation and action interface (Schulman et al., 2017); its interaction budget must be reported separately from offline imitation data. **GT-Box PID/IBVS** uses ground-truth boxes as a privileged diagnostic. It isolates some perception errors but is not a guaranteed control upper bound.

Status of numerical evaluation. The current manuscript specifies the comparisons but does not yet contain completed closed-loop results. Dashes in the tables denote pending measurements, not zero scores. No ranking, prompt advantage, or ablation gain is claimed without the corresponding evaluation logs. Table 4 separates all four tasks and both scene splits.

5.2 Prompt Initialization under Distractors

DT directly tests whether the initial prompt identifies the intended instance among similar moving objects. Spatial prompts provide a precise initial location, whereas language can remain ambiguous between similar instances and template matching can confuse similar appearances. This motivates the hypothesis that point, box, and mask initialization yield more stable and reliable tracking in DT. Establishing that advantage requires the five-way comparison in Table 5; it does not follow from prompt semantics alone.

We hold the action-head checkpoint, scene, target route, and controller fixed and vary only the initialization modality and its required adapter. Spatial annotations are supplied only in the first frame. The template condition receives the crop without the source bounding-box coordinates. Scores are compared on matched episodes, with seen and unseen results kept separate. SR and TR measure

Table 4: Closed-loop comparisons on STT, DT, OT, and UTT. Each cell is SR / TR / CR / MRPE. All results are pending. Privileged GT-Box PID/IBVS is a diagnostic.

Method	STT	DT	OT	UTT
Seen scenes (512 test episodes in total)				
GT-Box PID/IBVS	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
SAM2-Mask PID	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
Frame-BC MLP	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
CNN-GRU BC	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
Temporal Transformer BC	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
ACT	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
PPO waypoint policy	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
EDT	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
Unseen scenes (512 test episodes in total)				
GT-Box PID/IBVS	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
SAM2-Mask PID	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
Frame-BC MLP	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
CNN-GRU BC	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
Temporal Transformer BC	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
ACT	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
PPO waypoint policy	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–
EDT	–/–/–/–	–/–/–/–	–/–/–/–	–/–/–/–

Table 5: DT prompt comparison. Each entry is a *seen / unseen* pair for the same EDT action head. Dashes mark pending results. SR, TR, and CR are percentages; MRPE is in meters.

Initialization	SR ↑	TR ↑	CR ↓	MRPE ↓
Language	— / —	— / —	— / —	— / —
Point	— / —	— / —	— / —	— / —
Bounding box	— / —	— / —	— / —	— / —
Mask	— / —	— / —	— / —	— / —
Visual template	— / —	— / —	— / —	— / —

sustained correct-target tracking, CR measures flight failures, and MRPE measures the resulting spatial relationship. Initial target assignment and later identity switches should also be inspected to distinguish an initialization failure from subsequent memory drift. Any difference measures the complete prompt-initialization pipeline, including grounding or matching, rather than an intrinsic property of SAM2 alone.

5.3 Fusion, Memory, and Recovery Ablations

The ablation protocol changes one component at a time. **Equal Average** replaces precision weights with equal weights; **Hard Selection** chooses the lower-log-variance branch for each dimension. **History Length** varies the memory from the default six maps to zero, three, or nine. **No Pose Supervision** removes all explicit relative-pose losses. **Deterministic Branch Loss** replaces Gaussian NLL with Smooth L1 or MSE; because this removes variance supervision, its gating rule must be specified and its uncertainty outputs cannot be assumed calibrated. **Expert Only** omits paired perturbed episodes. These variants retain the waypoint interface and are evaluated with all four metrics on all task/scene splits.

OT and altitude-changing episodes are useful for analyzing whether the current and historical branches provide complementary information. Perturbation recovery is assessed by comparing Expert Only with paired-data training. Per-group MRPE should accompany the aggregate score because reference distances differ across target groups. Paired confidence intervals over test episodes and variation across training seeds separate evaluation uncertainty from optimization variance. Qualitative analysis should examine target masks, relative-pose estimates, branch weights, and executed

trajectories, with failures separated into identity switches, extended occlusion, mask drift, obstacle conflicts, and unrecovered viewpoint changes.

6 LIMITATIONS

EDT-BENCH evaluates simulated environments; the phrase "in the wild" refers to diverse scene conditions within this setting, not evidence of real-world deployment. Rendering, flight dynamics, and target behavior limit transfer conclusions. Fixed target paths improve repeatability and preserve curated OT cases, but do not cover all possible target responses to a pursuing drone. Privileged experts and transformed recovery labels provide supervision without guaranteeing dynamically feasible or optimal recovery from every perturbation. Numerical results and complete prompt-adapter specifications remain necessary before empirical conclusions can be drawn.

7 CONCLUSION

We introduced EDT-BENCH for studying embodied drone tracking with five initialization modalities, paired expert and perturbed training data, and diverse flight and visibility conditions. Its 1,024 fixed-target-trajectory test episodes separate four tracking tasks across seen and unseen scenes, while MRPE complements tracking and collision metrics with category-dependent spatial accuracy. The SAM2-based EDT baseline connects promptable segmentation to continuous waypoint control through measurement–memory fusion. This formulation supports controlled investigation of target identity, viewpoint selection, and recovery under challenging drone-tracking conditions.

AI USE STATEMENT

Generative AI tools assisted with language editing, LaTeX restructuring, figure preparation, and consistency checks. They were not used to generate experimental measurements. The authors are responsible for verifying the technical content, references, and final submission.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the task interface, data pairing, MRPE, and learning objective; the appendices provide implementation details and reporting layouts. Reproduction additionally requires scene and episode manifests, target trajectories, prompt-adapter checkpoints, evaluator thresholds, and training and evaluation configurations. The current draft does not represent pending numerical measurements as completed experiments.

REFERENCES

Guyue Hu, Haoming Liu, Siyuan Song, Chenglong Li, Feng Chen, and Jin Tang. DeTrack: A benchmark and altitude-aware dual world model for drone-embodied tracking. arXiv preprint arXiv:2605.17451, 2026.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-Learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Jiahao Li, Fuhui Zhou, and Qihui Wu. A novel expert-assisted anomaly-aware embodied learning framework for UAV active target tracking. *IEEE Transactions on Mobile Computing*, 25(7): 11063–11077, 2026. doi: 10.1109/TMC.2026.3666656.

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

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. In *International Conference on Learning Representations*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, pp. 621–635, 2018.

Haowei Sun, Jinwu Hu, Zhirui Zhang, Haoyuan Tian, Xinze Xie, Yufeng Wang, Zhuliang Yu, Xiaohua Xie, and Mingkui Tan. A cross-scene benchmark for open-world drone active tracking. arXiv preprint arXiv:2412.00744, 2024.

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.

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.

Qihui Wu, Jiahao Li, Fuhui Zhou, Jiahuan Ji, Haoyang Wang, Hongtao Liang, and Kai-Kuang Ma. Cognitive embodied learning for anomaly active target tracking. *Communications Engineering*, 4:224, 2025. doi: 10.1038/s44172-025-00556-6.

Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yinan Yang, Xiang Wan, Shiming Ge, and Dacheng Tao. WebUAV-3M: A benchmark for unveiling the power of million-scale deep UAV tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9186–9205, 2023. doi: 10.1109/TPAMI.2022.3232854.

Qiyao Zhang, Shuhua Zheng, Jianli Sun, Chengxiang Li, Xianke Wu, Zihan Song, Zhiyong Cui, Yisheng Lv, and Yonglin Tian. UAV-Track VLA: Embodied aerial tracking via vision-language-action models. arXiv preprint arXiv:2604.02241, 2026.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705, 2023.

Fangwei Zhong, Kui Wu, Hai Ci, Churan Wang, and Hao Chen. Empowering embodied visual tracking with visual foundation models and offline RL. arXiv preprint arXiv:2404.09857, 2024.

Fangwei Zhong, Kui Wu, Churan Wang, Hao Chen, Hai Ci, Zhoujun Li, and Yizhou Wang. UnrealZoo: Enriching photo-realistic virtual worlds for embodied AI. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

A Collection Details and Data Schema

The expert uses privileged target and drone poses during collection. A depth comparison between target-shown and target-hidden renders identifies visible target regions; mask, box, and pixel-count checks support quality control. Depth along the motion corridor is used for obstacle proximity, and initial clearance is checked in multiple directions. Forward speed is reduced near obstacles, with lateral or vertical escape directions selected when required. Neither privileged actor state nor collection-only depth checks are inputs to the learned action policy.

Perturbation segments apply smooth $\sin^2$ envelopes over 2.4–3.2 seconds. Mild, medium, and hard yaw offsets use ranges of 6–10, 10–16, and 16–24 degrees, respectively; position limits are specified by the collection configuration. Re-rendering each perturbed pass is necessary because a changed camera pose changes the image and visibility. The target and distractor states remain synchronized with the corresponding expert episode.

Each timestamp associates RGB, target mask, bounding box, visibility, drone and target poses and velocities, relative pose and velocity, executed action and motion, perturbation metadata, and future waypoints. With timestamps and executed velocities, the records also support continuous $[v_x, v_y, v_z, \dot{\psi}]$ supervision. Future waypoints use the current yaw-aligned drone frame and must not be confused with world-frame coordinates.

The training corpus contains 11,340 episodes in 5,670 pairs. Both members of a pair must remain together in any training/validation subdivision to prevent paired-observation leakage. Seen test episodes use the training scene pool but distinct episode trajectories; unseen episodes use the 20 held-out scenes. Animals are excluded from training and validation, including in seen scenes. The established test totals are 512 seen and 512 unseen episodes. Per-task counts, route identifiers, occlusion-event durations, and weather/illumination frequencies require the finalized manifest; equal allocation among the four tasks is not assumed.

B Implementation Details

Measurement encoder. A mask downsampler applies two stride-two convolutions, LayerNorm2d, and GELU to sigmoid mask logits. The resulting map is concatenated with current image features and normalized two-dimensional coordinates. Three convolutional blocks, adaptive 8×8 pooling, layer normalization, and a linear projection produce a 256-dimensional token. A two-layer MLP with hidden width 512 predicts the relative-pose mean and log variance.

Memory encoder. Separate spatial encoders convert each of the six past mask-memory maps and the current image feature into 256-dimensional tokens. Learnable temporal embeddings and memory validity masks are used by a two-layer Transformer with eight attention heads, feed-forward width 1,024, and dropout 0.1. The final token predicts the history branch's mean and log variance. Causality follows from restricting inputs to past and current observations.

Decoder and normalization. A $4 \to 128 \to 256$ MLP embeds the normalized fused pose. Its output is concatenated with the 256-dimensional temporal token and 256-dimensional object pointer. A $768 \to 512 \to 512 \to 40$ MLP predicts ten four-dimensional waypoints. Relative-pose and horizon-specific waypoint statistics are computed only from training data. The remaining losses in Eq. 9 are

$$\mathcal{L}_{\text{fuse}} = \text{SmoothL1}\left(\frac{\delta(\mu^{\text{fuse}}, r)}{c_r}, 0\right), \tag{10}$$

$$\mathcal{L}_{\text{wp}} = \text{SmoothL1}\left(\frac{\delta(\widehat{\mathcal{A}}, \mathcal{A})}{c_w}, 0\right). \tag{11}$$

Yaw residuals use a 360-degree period. Branch log variances represent uncertainty in normalized coordinates, so both branches use the same dimension-wise scale when computing fusion weights.

Prompt adapters. The common interface does not imply native SAM2 language or template encoding. A language condition needs a specified grounding model and checkpoint; a template condi-

tion needs a specified matching procedure. Their spatial outputs initialize the same segmenter and action head used by the other conditions. Adapter versions, checkpoint identifiers, preprocessing, and initialization failure handling must be recorded before the prompt comparison is reproducible. Initialization failure must not be silently corrected with a privileged mask or target identifier.

C Expanded Experimental Reporting

Table 4 separates task and scene axes for every baseline. Each cell reports four values in the fixed order SR / TR / CR / MRPE, with rates in percent and MRPE in meters. The prompt modality used for this comparison must be stated with the completed results. Tables 5 and 6 isolate prompt and design comparisons, respectively. A dash denotes a missing measurement throughout.

Table 6: Ablation reporting layout. Each variant is evaluated on all four task subsets and both scene splits; these columns are repeated for each task/split. Rates are percentages and MRPE is in meters.

Variant	SR ↑	TR ↑	CR ↓	MRPE ↓
EDT (six memory maps)	—	—	—	—
Equal Average	—	—	—	—
Hard Selection	—	—	—	—
History Length: 0 / 3 / 9	—	—	—	—
No Pose Supervision	—	—	—	—
NLL to Smooth L1 / MSE	—	—	—	—
Expert Only	—	—	—	—