AERIALWAM: RECONSTRUCTIVE AERIAL WORLD ACTION MODEL FOR OPEN-WORLD SEMANTIC-LEVEL OBJECT-GOAL SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world semantic-level object-goal search is an important yet challenging task for aerial agents. Existing approaches rely on spatio-semantic representations and multimodal large models to reason over historical observations, failing to maintain semantic-object alignment during long-horizon search in open-world environments. To solve this problem, this paper presents AerialWAM, the first reconstructive aerial world action model, which can autonomously search for semantic-level object-goals with semantic-object alignment. Specifically, AerialWAM first constructs a object-conditioned aerial world action model that couples semantic understanding and action generation using a Mixture-of-Transformers to achieve long-horizon prediction and action execution. We further design a semantic-object reconstructor that generates an goal-centric object image from current observations and the semantic description. During search, each new observation refines this object image and feeds it back to the world action model for the next action, forming a closed loop to achieve semantic-object alignment. We conducted extensive experiments on AerialDojo-200K. The result shows AerialWAM achieves average success rates 5.77× and 23.98× those of existing methods in in-distribution and out-of-distribution settings, respectively.

1 INTRODUCTION

Open-world Semantic-level Object-Goal Search (SemanticOGS) (Xiao et al., 2025) requires aerial agents to autonomously search for objects through semantic goal-driven long-horizon exploration in an unknown, large-scale, and unstructured three-dimensional environment, and to actively stop while onboard observations match the object description. Unlike aerial Vision-and-Language Navigation (VLN), which relies on detailed, step-by-step route instructions that constrain how the aerial agent should travel (Zhao et al., 2026), SemanticOGS provides only a high-level, open-ended semantic goal and requires the agent to autonomously explore. SemanticOGS has broad potential applications in search and rescue, infrastructure inspection, and autonomous delivery, where operators may specify the object without knowing its exact location or providing a complete flight route.

On the one hand, existing aerial agents have made remarkable progress by utilizing spatio-semantic representations, such as semantic maps (Zhou et al., 2026; Kim et al., 2025), scene graphs (Gai et al., 2026), and value maps (Zhang et al., 2026), to organize memories and plan over explored regions, frontiers, and object candidates. On the other hand, recent studies demonstrate that multimodal large models can leverage high-level semantic reasoning to interpret semantic goals (Xiao et al., 2025), assess object candidates (Pardyl et al., 2025), and guide search decisions (Wang et al., 2024). In essence, spatio-semantic representations provide persistent spatial memory and value-aware exploration, whereas vision-language reasoning enables open-vocabulary goal grounding and object-conditioned decision making.

Nevertheless, existing approaches rely on historical observations, failing to maintain semantic-object alignment during long-horizon search in open-world environments. This problem leads to two key challenges: 1) *inefficient exploration*: aerial agents must decide where to explore and how to fly while the object remains out of view, which leaves long-horizon search unguided; and 2) *unreliable stopping*: Aerial agents must determine whether an observed target is the requested object, and

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.

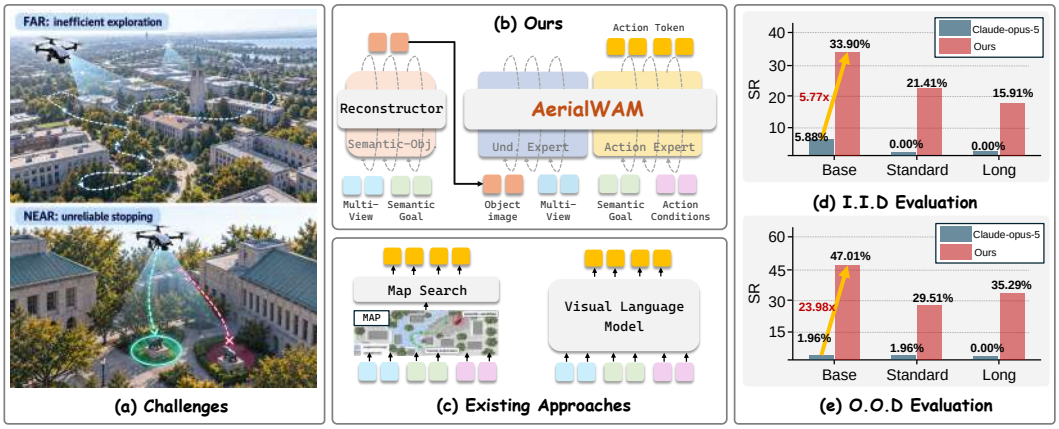


Figure 1: Overview of the task, framework, and navigation results. (a) The challenges of existing approaches. (b) The framework of AerialWAM.(c): Existing approaches, the left is spatio-semantic representations-based approaches, and the right is vision-language reasoning-based approaches. (d) In-distribution evaluation. (e) Out-of-distribution evaluation.

ambiguous semantic grounding hinders precise object identification with nearby lookalikes. Long-horizon search and nearby lookalikes make this task particularly difficult.

To solve these challenges, this paper presents AerialWAM, the first reconstructive aerial world action model, which can autonomously search for semantic-level object goals with semantic-object alignment. Specifically, AerialWAM first constructs an object-conditioned aerial world action model that couples semantic understanding and action generation using a Mixture-of-Transformers to achieve long-horizon prediction and action execution. We further design a semantic-object reconstructor that generates a goal-centric object image from current observations and the semantic description. During search, each new observation refines this object image and feeds it back to the world action model for the next action. This iterative observation-reconstruction-action-execution closed-loop maintains semantic-object alignment during exploration and stopping. We train the model in three stages: optimizing the world action model for action success, minimizing object-image reconstruction loss, and jointly optimizing both components for closed-loop search success.

We evaluate AerialWAM's performance on AerialDojo-200K (Feng et al., 2026a), an large-scale open-world aerial SemanticOGS benchmark. Extensive experiments demonstrate AerialWAM's superiority: as shown in Figure 1, the result shows AerialWAM achieves average success rates 5.77× and 6.0× those of existing methods in in-distribution and out-of-distribution settings, respectively. The contributions are summarized as follows:

- We propose AerialWAM, the first reconstructive aerial world action model, which couples semantic understanding and action generation using a Mixture-of-Transformers to achieve long-horizon prediction and action execution.
- We design a semantic-object reconstructor for AerialWAM, building observation-reconstruction-action-execution closed-loop dynamic to autonomously search with semantic-object alignment during exploration and stopping.
- We evaluate AerialWAM on AerialDojo-200K. the result shows AerialWAM achieves average success rates 5.77× and 23.98× those of existing methods in in-distribution and out-of-distribution settings, respectively.

2 Related Work

Open-world aerial Semantic-level Object-Goal Search (SemanticOGS). Open-world SemanticOGS asks an aerial agent to infer where to fly, which evidence to retain, and when to stop from a semantic description alone. Existing methods fall into two complementary groups. Spatio-semantic approaches maintain persistent spatial evidence: OctMem-Agent aggregates RGB-D observations in

an octree memory (Zhou et al., 2026); APEX separates attraction, exploration, and obstacle memories (Zhang et al., 2026); RAVEN constructs a semantic voxel-ray representation for behavior adaptation (Kim et al., 2025); USS-Nav builds a unified spatio-semantic scene graph (Gai et al., 2026); and GoalSwarm shares occupancy and value information across aerial agents (James et al., 2026). These representations support frontier selection, spatial coverage, and reuse of observations after a target leaves the field of view. Their performance, however, depends on pose, depth, semantic association, and temporal fusion, and a memory of past evidence does not explicitly predict how a candidate action will alter future target visibility. Vision-language approaches address the semantic side: UAV-ON establishes the aerial object-goal setting and its AOA baseline (Xiao et al., 2025); FlySearch studies VLM-controlled exploration (Pardyl et al., 2025); UAV-Need-Help studies assistant-guided search (Wang et al., 2025); PRPSearcher combines multimodal reasoning with spatial search (Ji et al., 2025); and AirHunt couples VLM semantics with continuous planning (Chen et al., 2026). These approaches make open-vocabulary goal interpretation and candidate assessment practical, but suffer from inference latency, storage overhead, and ineffective exploration. AerialWAM addresses this interface by returning an observation-conditioned semantic-object image to a predictive policy while retaining the real observation stream for exploration and stopping.

World action models. World action models select actions using predicted consequences rather than the current view alone. UWM couples video and action diffusion (Zhu et al., 2025); Motus separates specialized experts for multimodal action generation (Bi et al., 2026); and MobileWAM extends a world-action interface to mobile manipulation (Fan et al., 2026). EvolvingAgent studies continual planner-controller-reflector updates in an embodied world model (Feng et al., 2026b), while RoG uses reflection and feedback to improve embodied video prediction (Chi et al., 2025). ReconVLA is especially relevant to target grounding because it reconstructs an instructed manipulation region as an auxiliary visual signal (Song et al., 2026); AerialWAM differs in returning the reconstructed semantic-object image to the online policy at every observation update. In aerial settings, ANWM predicts action-conditioned latent dynamics (Zhang et al., 2025), WorldFly couples future visual prediction with navigation actions (Zheng et al., 2026), and WorldVLN performs autoregressive world-action prediction for aerial vision-language navigation (Zhao et al., 2026). These models use routes, waypoints, or movement instructions, whereas SemanticOGS requires open-ended exploration and an active decision to stop at the requested object among nearby lookalikes. Predictive models therefore provide the appropriate mechanism for long-horizon control but may accumulate rollout or grounding error. AerialWAM solves this problem through a two-expert Mixture-of-Transformers and an observation-reconstruction-execution loop.

3 METHOD

This section defines AerialWAM as a closed-loop policy for open-world SemanticOGS. The semantic-object reconstructor converts the current multi-view observation and language goal into a goal-centric object image; the reconstructed aerial world action model uses that reconstructed object image together with real observations and stop conditions to predict future latent states and decode the next action; joint training connects the two modules through the observation-reconstruction-action-execution loop. The target coordinate and the ground-truth object crop are used only to construct supervision and evaluator labels.

3.1 PRELIMINARIES

At step t, the aerial agent receives a multi-view RGB-D observation I_t, an available vehicle state s_t, and a semantic goal g. We write $o_t = (I_t, s_t)$ and $h_t = (o_{0:t}, a_{0:t-1})$ for the history available before the next decision. The action space is $\mathcal{A} = \mathcal{A}_{\text{fly}} \cup \{\textsc{Stop}\}$, where $\mathcal{A}_{\text{fly}}$ contains the parameterized flight commands. The environment transition is $o_{t+1} \sim P(\cdot \mid h_t, a_t)$. The policy is evaluated by the terminal active-stop objective

$$J(\pi) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{T} r_t\right], \qquad r_t = 0 \ (t < T), \qquad r_T = \mathbf{1}\{\text{active STOP at the requested object}\}. \tag{1}$$

Collision and timeout terminations are failures. The true target coordinate and the training crop are never provided as policy inputs at test time.

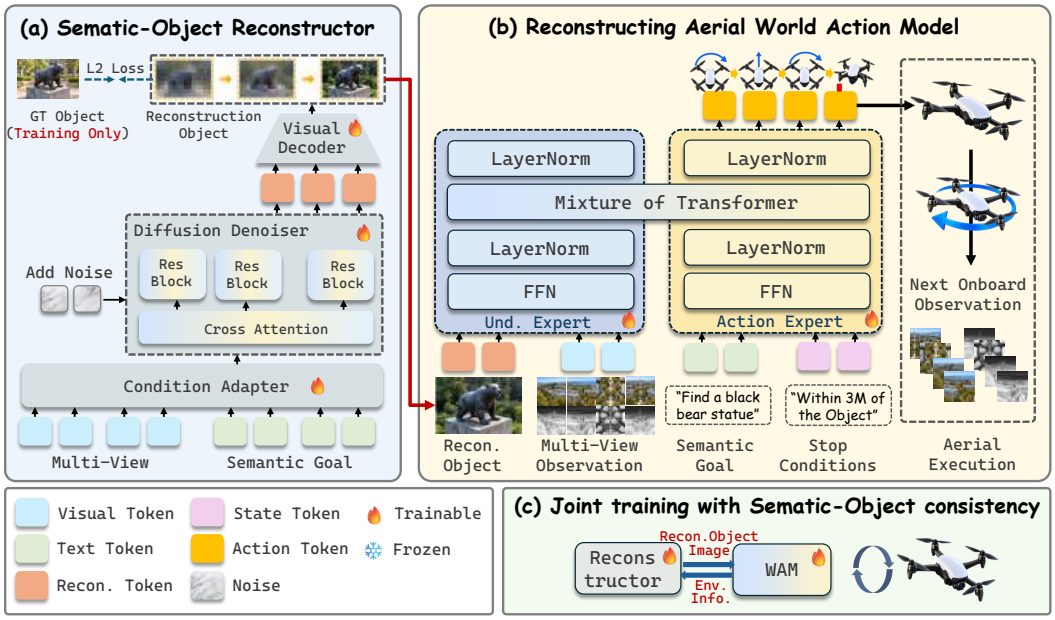


Figure 2: The overview of AerialWAM. (a) The semantic-object reconstructor generates a goal-centric object image from the semantic goal and current multi-view observation; the ground-truth object crop is used only for training. (b) The reconstructing aerial world action model couples an understanding expert and an action expert, predicts action-conditioned future features, and decodes the next flight or Stop action. (c) Joint training feeds each reconstructed object image back to the WAM, forming an observation-reconstruction-action-execution loop with semantic-object alignment.

3.2 Semantic-Object Reconstructor

The reconstructor in Figure 2(a) receives only information available at the current step: the multi-view observation and the semantic goal. Its output $\hat{x}_t$ is the goal-centric object image, which becomes the reconstructed object image supplied to the WAM. Let $x^\star$ denote an object-centric crop constructed from the annotated target during training, and let $z_0^\star = \mathcal{E}_{\text{obj}}(x^\star)$ be its visual-token representation. The crop $x^\star$ is not available to the policy at inference. The condition adapter combines observation and language tokens,

$$c_t^R = \mathcal{R}_\phi(\mathcal{E}_I(I_t), \mathcal{T}(g)), \tag{2}$$

where $\mathcal{R}_\phi$ includes the cross-attention condition adapter shown in Figure 2(a). For diffusion step n, the forward process is

$$z_n = \sqrt{\bar{\alpha}_n}\, z_0^\star + \sqrt{1-\bar{\alpha}_n}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

and the denoiser predicts the injected noise (Ho et al., 2020; Rombach et al., 2022):

$$\mathcal{L}_{\text{noise}} = \mathbb{E}_{n,\epsilon}\left[\left\|\epsilon_\phi(z_n, c_t^R, n) - \epsilon\right\|_2^2\right]. \tag{4}$$

The visual decoder $\mathcal{D}_\phi$ converts the denoised latent into the reconstructed object image,

$$\hat{x}_t = \mathcal{D}_\phi(\hat{z}_{0,t}), \qquad \hat{z}_{0,t} = \frac{z_n - \sqrt{1-\bar{\alpha}_n}\,\epsilon_\phi(z_n, c_t^R, n)}{\sqrt{\bar{\alpha}_n}}. \tag{5}$$

The image-level supervision shown in Figure 2(a) is

$$\mathcal{L}_{\text{img}} = \mathbb{E}\left[\left\|\hat{x}_t - x^\star\right\|_2^2\right]. \tag{6}$$

The adapter, denoiser, and visual decoder are the trainable reconstruction components indicated by the flame symbols in Figure 2(a); the annotated crop is a training target only. At inference, $\hat{x}_t$ is the

reconstructed object image generated by the complete reverse-diffusion sampler from the current I_t and g, with no access to $x^\star$. Because I_t changes after every executed action, the reconstructed object image is refined at every loop iteration without introducing a privileged target crop or target coordinate.

3.3 Reconstructing Aerial World Action Model

The WAM in Figure 2(b) receives four typed input groups: real multi-view observation tokens, semantic-goal tokens, reconstructed object-image tokens, and vehicle-state/stop-condition tokens. Define

$$h_t^I = \mathcal{E}_I(I_t), \quad h^g = \mathcal{T}(g), \quad h_t^r = \mathcal{E}_{\text{ref}}(\hat{x}_t), \quad h_t^s = \mathcal{P}_s(s_t), \quad h^{\text{stop}} = \mathcal{T}_{\text{stop}}(\text{STOP rule}), \tag{7}$$

and $c_t = [h_t^I; h^g; h_t^r; h_t^s; h^{\text{stop}}]$. The stop-condition token specifies the evaluator's terminal rule and collision/timeout constraints; it does not contain the target coordinate or target distance. The understanding expert and action expert have separate parameters and exchange information through the Mixture-of-Transformers. For expert $e \in \{u, a\}$,

$$\text{Attn}_e(X_u, X_a) = \text{softmax}\left(\frac{Q_e[K_u; K_a]^\top}{\sqrt{d}} + M_e\right)[V_u; V_a], \tag{8}$$

where $Q_e = X_e W_e^Q$, $K_e = X_e W_e^K$, $V_e = X_e W_e^V$, and M_e masks unavailable future observations. The action expert proposes a nominal flight sequence for prediction,

$$\bar{a}_{t:t+H-1} = \mu_\theta(c_t) \in \mathcal{A}_{\text{fly}}^H, \tag{9}$$

while the latent transition model predicts action-conditioned future features. With $z_t = \mathcal{E}_{\text{cur}}(o_t)$ and $\hat{z}_t = z_t$,

$$\hat{z}_{t+k+1} = \mathcal{M}_\theta(\hat{z}_{t+k}, \bar{a}_{t+k}; c_t), \qquad k = 0, \ldots, H-1, \tag{10}$$

and $\hat{Z}_t = (\hat{z}_{t+1}, \ldots, \hat{z}_{t+H})$. The final action distribution is

$$a_t \sim \pi_{\theta,\psi}(\cdot \mid h_t, g, \hat{x}_t, \hat{Z}_t), \tag{11}$$

where the action head outputs distribution parameters for the continuous/parameterized flight command and the discrete STOP action. Only the first selected action is executed. The nominal sequence is an internal prediction probe, not an executed open-loop plan, and predicted features are not replaced by recorded future observations at test time. During prediction training, recorded controls are used as aligned supervision actions for the observed future features; this does not expose those future observations to the policy during evaluation.

3.4 Joint Training with Semantic-Object Consistency

The closed loop in Figure 2(c) is trained in three stages. Stage I trains the two-expert WAM with a null reconstructed image while the latent predictor is supervised by recorded future features,

$$\mathcal{L}_{\text{I}} = -J(\pi_{\theta,\psi}; \hat{x} = \varnothing) + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}}, \qquad \mathcal{L}_{\text{pred}} = \mathbb{E}\left[\frac{1}{H}\sum_{k=1}^{H} \|\hat{z}_{t+k} - \mathcal{E}_{\text{obs}}(o_{t+k})\|_2^2\right], \tag{12}$$

where $\mathcal{E}_{\text{obs}}$ is a distinct fixed target encoder unless the checkpoint manifest explicitly documents parameter sharing. Stage II freezes the WAM and trains the reconstructor on paired observations, language goals, and training-only crops,

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{\text{noise}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}}. \tag{13}$$

Stage III unfreezes both modules and optimizes the terminal search objective under the closed-loop update

$$\hat{x}_t \sim q_\phi(\cdot \mid I_t, g), \qquad a_t \sim \pi_{\theta,\psi}(\cdot \mid h_t, g, \hat{x}_t, \hat{Z}_t), \qquad o_{t+1} \sim P(\cdot \mid h_t, a_t), \tag{14}$$

with $\hat{x}_t$ regenerated after every new observation. The joint objective is

$$\mathcal{L}_{\text{III}} = -J(\pi_{\theta,\psi,\phi}). \tag{15}$$

Here $J(\pi_{\theta,\psi,\phi})$ is the same terminal success return in Equation 1 under the policy induced by the WAM and reconstructor. If the reverse sampler is stochastic, the terminal return can be assigned to the sampled action and diffusion paths with a score-function estimator (Williams, 1992); no gradient is taken through the environment transition. This training schedule makes the interface in Figure 2 explicit: the reconstructor supplies the reconstructed object image, the WAM supplies the next action, and the next real observation closes the loop.

4 EXPERIMENTS

We evaluate AerialWAM on two AerialDojo-200K (Feng et al., 2026a) scenes and report Base, Standard, and Long tasks separately. Section 4.1 specifies the dataset, simulator setting, model configuration, evaluation protocol, and metrics. Section 4.2 reports In-Distribution (ID) results on $N_{\text{Island_0}}$, including success-radius sensitivity. Section 4.3 reports results on $N_{\text{Island_1}}$ under out-of-distribution setting. We provide visual validation in Section 4.4.

4.1 IMPLEMENTATION

Dataset and simulation environment. AerialDojo-200K (Feng et al., 2026a) is an open-world aerial SemanticOGS benchmark in which the agent receives multi-view RGB-D observations, available vehicle state, and a semantic object goal, and executes flight commands or an explicit STOP action. We organize the experiments into In-Distribution (ID) evaluation on $N_{\text{Island_0}}$ and Out-Of-Distribution (OOD) evaluation on $N_{\text{Island_1}}$. $N_{\text{Island_0}}$ testset contains 295 Base, 341 Standard, and 176 Long episodes, which are reported separately. The $N_{\text{Island_1}}$ testset contains 490 episodes, comprising 334 Base, 122 Standard, and 34 Long episodes. Results are reported at success radii of 3 m and 5 m by rescoring the same recorded trajectories for each run. The $N_{\text{Island_1}}$ checkpoint is trained using 137 training trajectories from that scene. Accordingly, the OOD-scene designation identifies the second evaluation scene in our experimental organization; these results characterize navigation after scene-specific training with supplied target references, rather than zero-shot transfer to a training-unseen environment.

Model configuration. AerialWAM consists of a semantic-object reconstructor and a two-expert Mixture-of-Transformers world action model. The reconstructor maps the current multi-view observation and semantic goal to a goal-centric object image. The WAM receives the real observation, semantic goal, reconstructed object image, vehicle-state tokens, and stop-condition tokens; it predicts action-conditioned future latent features and decodes one flight command or STOP. The model is trained in three stages: WAM training with latent prediction, object-image reconstruction, and joint closed-loop optimization.

Evaluation protocol. All methods in a matched comparison use the same episode manifest, observation and action interface, success radius, collision rule, stop semantics, and decision budget. The IID evaluation uses no test-time adaptation. The transfer evaluation keeps the trained checkpoint fixed. A trajectory terminates at a valid STOP, collision, timeout, or the simulator limit. One action from the policy readout is executed before the next onboard observation is acquired.

Baselines. We evaluate three open-source Multimodal Large Language Models (MLLMs), Qwen3-VL-4B, InternVL3.5-8B, and STEP3-VL, and four closed-source MLLMs, GPT-5.6 Sol, Grok-4.5, Gemini 3.8 Flash, and Claude Opus 5. We test these models under the Base, Standard, and Long-Horizon settings. For brevity, the main tables report model-setting combinations with nonzero overall success rates, covering all 7 MLLMs on Base tasks.

Metrics. We evaluate navigation performance using five complementary metrics. Success rate (SR) measures the fraction of episodes in which the agent explicitly stops within 3 m of the target anchor without collision and within the action budget. Oracle success rate (OSR) measures whether the agent ever enters this region, regardless of its final stopping decision or subsequent failure. Distance to success (DTS) is the mean final Euclidean distance to the target anchor, measured in meters without subtracting the success radius. Success weighted by path length (SPL) evaluates navigation efficiency:

$$\text{SPL} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i\frac{L_i}{\max(L_i,\ell_i)}, \tag{16}$$

where N is the number of evaluated episodes, σ_i indicates episode success, L_i is the reference trajectory length and ℓ_i is the executed trajectory length. This formulation uses the planned reference for normalization. Collision rate (CR) is the fraction of episodes containing a collision. All metrics are computed over both successful and failed episodes. We define the false stop rate (FSR) as the difference between the oracle success rate (OSR) and the success rate (SR), i.e., FSR = OSR SR.

Table 1: Base-task performance comparison on $N_{\text{Island_0}}$. SR, OSR, SPL, CR, and FSR are percentages; DTS is in meters. Arrows indicate the preferred direction; bold denotes the best available value.

Radius	Model	SR ↑	OSR ↑	SPL ↑	DTS ↓	CR ↓	FSR ↓
3 m	Qwen3-VL-4B	1.36	1.69	0.92	13.80	22.37	0.34
	InternVL3.5-8B	0	2.03	0	14.96	47.80	2.03
	STEP3-VL	0.50	3.52	0.09	13.71	38.19	3.02
	GPT-5.6 Sol	6.12	6.12	4.40	16.11	40.82	0
	Grok-4.5	0	0	0	15.19	14.29	0
	Gemini 3.8 Flash	4.17	12.50	3.99	29.06	37.50	8.33
	Claude Opus 5	5.88	13.73	5.45	10.90	64.71	7.84
	AerialWAM	**33.90**	**49.15**	**33.40**	**6.56**	**66.10**	**15.25**
5 m	Qwen3-VL-4B	7.46	11.19	2.94	13.80	22.37	3.73
	InternVL3.5-8B	5.08	14.24	2.87	14.96	47.80	9.15
	STEP3-VL	3.52	15.58	1.86	13.71	38.19	12.06
	GPT-5.6 Sol	12.24	14.29	10.52	16.11	40.82	2.04
	Grok-4.5	2.04	2.04	2.04	15.19	14.29	0
	Gemini 3.8 Flash	8.33	20.83	8.16	29.06	37.50	12.50
	Claude Opus 5	5.88	21.57	5.45	10.90	64.71	15.69
	AerialWAM	**33.90**	**54.92**	**33.40**	**6.56**	**66.10**	**21.02**

Table 2: AerialWAM on Standard and Long tasks in $N_{\text{Island_0}}$. Results are reported at two evaluation radii. Rates are percentages; DTS is in meters.

Radius	Task	N	SR ↑	OSR ↑	SPL ↑	DTS ↓	CR ↓	FSR ↓
3 m	Standard	341	21.41	26.98	21.09	18.91	77.13	5.57
	Long	176	15.91	20.45	15.63	38.38	84.09	4.55
5 m	Standard	341	21.41	29.91	21.09	18.91	77.13	8.50
	Long	176	15.91	21.02	15.63	38.38	84.09	5.11

4.2 In-Distribution Evaluation

We evaluate AerialWAM on the $N_{\text{Island_0}}$ scene using the Base, Standard, and Long task groups. Table 1 reports the navigation results at 3 m and 5 m evaluation radii. The two radii retrospectively rescore the same recorded rollouts and therefore do not correspond to separate policy executions. For each task group, we report the number of evaluated episodes together with six navigation metrics: success rate (SR), oracle success rate (OSR), success weighted by path length (SPL), distance to success (DTS), collision rate (CR), and FSR.

Performance across task groups. At the 3 m evaluation radius, AerialWAM achieves SRs of 33.90%, 21.41%, and 15.91% on the Base, Standard, and Long task groups, respectively. The corresponding SPL values are 33.40%, 21.09%, and 15.63%. Collisions account for a substantial fraction of unsuccessful episodes across all three groups, with CR increasing from 66.10% on Base to 77.13% on Standard and 84.09% on Long. The mean final distance to success also increases from 6.56 m on Base to 18.91 m on Standard and 38.38 m on Long.

These differences are reported descriptively rather than as a controlled comparison of task difficulty. In particular, the historical Base evaluation and the recovered Standard and Long evaluations use different initial-heading conventions. We therefore report the three task groups separately and do not aggregate them into a single overall score for $N_{\text{Island_0}}$.

Table 3: Base-task comparison on $N_{\text{Island_1}}$. Rates are percentages; DTS is in meters.

Radius	Model	SR ↑	OSR ↑	SPL ↑	DTS ↓	CR ↓	FSR ↓
3 m	Qwen3-VL-4B	0.94	3.77	0.57	18.14	28.30	2.83
	InternVL3.5-8B	0	2.83	0	45.12	65.09	2.83
	STEP3-VL[§]	2.83	2.83	2.31	20.35	39.62	0
	GPT-5.6 Sol	2.04	2.04	1.70	18.24	12.24	0
	Grok-4.5	0	0	0	14.61	10.42	0
	Gemini 3.8 Flash	2.56	2.56	2.56	29.29	17.95	0
	Claude Opus 5	1.96	3.92	1.85	28.49	41.18	1.96
	AerialWAM	**47.01**	**49.40**	**45.65**	**9.81**	**38.92**	**2.40**
5 m	Qwen3-VL-4B	3.77	12.26	2.45	18.14	28.30	8.49
	InternVL3.5-8B	0	14.15	0	45.12	65.09	14.15
	STEP3-VL[§]	6.60	14.15	4.31	20.35	39.62	7.55
	GPT-5.6 Sol	6.12	12.24	5.78	18.24	12.24	6.12
	Grok-4.5	1.96	5.88	1.96	14.61	10.42	3.92
	Gemini 3.8 Flash	2.56	10.26	2.56	29.29	17.95	7.69
	Claude Opus 5	7.84	13.73	7.49	28.49	41.18	5.88
	AerialWAM	**47.01**	**58.68**	**45.65**	**9.81**	**38.92**	**11.68**

Table 4: AerialWAM on Standard and Long tasks in $N_{\text{Island_1}}$. Results are reported at two evaluation radii. Rates are percentages; DTS is in meters.

Radius	Task	N	SR ↑	OSR ↑	SPL ↑	DTS ↓	CR ↓	FSR ↓
3 m	Standard	122	29.51	29.51	27.77	23.28	64.75	0
	Long	34	35.29	35.29	34.56	33.54	58.82	0
5 m	Standard	122	29.51	31.97	27.77	23.28	64.75	2.46
	Long	34	35.29	38.24	34.56	33.54	58.82	2.94

Evaluation-radius sensitivity. Increasing the evaluation radius from 3 m to 5 m leaves SR, SPL, DTS, and CR unchanged for all three task groups. This follows from the evaluation protocol: both radii rescore the same trajectories without changing the policy, checkpoint, or internal stopping threshold.

OSR increases under the larger scoring radius. For Base tasks, OSR increases from 49.15% to 54.92%; for Standard tasks, from 26.98% to 29.91%; and for Long tasks, from 20.45% to 21.02%. FSR correspondingly increases from 15.25% to 21.02%, from 5.57% to 8.50%, and from 4.55% to 5.11%, respectively. Thus, additional trajectories enter the larger target region, but no additional episode satisfies the complete success criterion, which additionally requires a policy STOP and collision-free termination. The radius comparison therefore characterizes the recorded trajectories under different scoring tolerances rather than the effect of changing the policy's stopping behavior.

4.3 OOD-Scene Evaluation with Scene-Specific Training

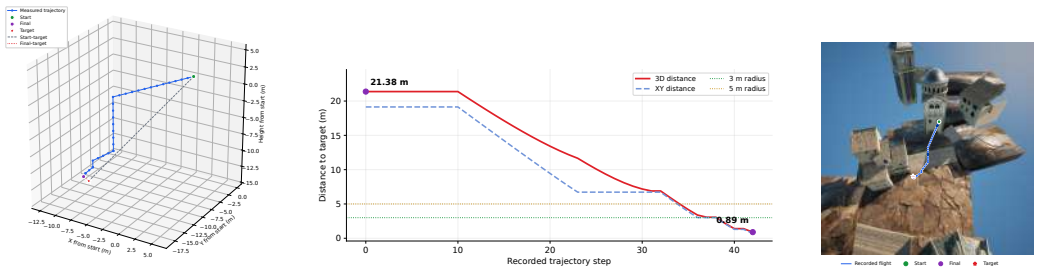
We report results on $N_{\text{Island_1}}$ under scene-specific training. The checkpoint uses 137 training trajectories from this scene and is evaluated on the complete 490-episode test set. The OOD-scene designation preserves the organization of our two-scene evaluation; this setting is not zero-shot transfer from a checkpoint trained only on $N_{\text{Island_0}}$.

Performance across task groups. AerialWAM succeeds in 157/334 Base, 36/122 Standard, and 12/34 Long episodes, corresponding to SRs of 47.01%, 29.51%, and 35.29%, respectively, at both radii. Across all 490 episodes, 205 succeed, giving an episode-weighted SR of 41.84%. This overall score is not an unweighted average of the three task-specific rates. At 3 m, the corresponding task-

specific SPL values are 45.65%, 27.77%, and 34.56%. Collision rates are 38.92%, 64.75%, and 58.82%, indicating substantial remaining navigation failures.

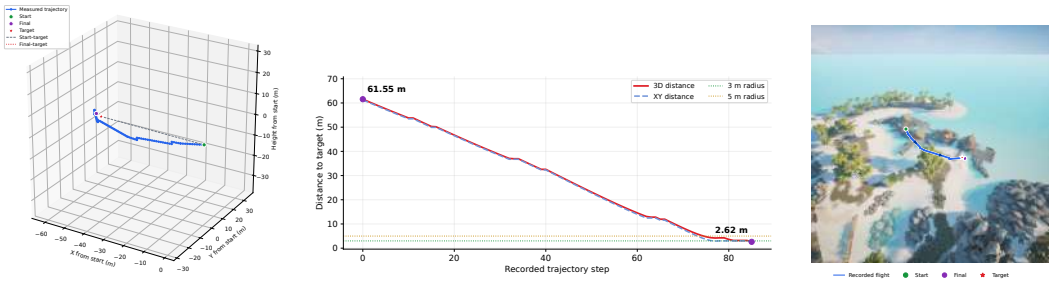
4.4 VISUALIZATION

Figures 3 and 4 visualize two representative successful search episodes on AerialDojo. On $N_{\text{Island_0}}$, the agent reduces its target distance from 21.38 m to 0.89 m. After an initial plateau, the distance decreases over most recorded steps. The gap between the 3D and horizontal distances also narrows, indicating reduced vertical separation from the target. The corresponding trajectory and oblique scene view illustrate the flight path within the environment. On $N_{\text{Island_1}}$, the long-task episode starts 61.55 m from the target and ends 2.62 m away. This episode involves a greater initial distance and more recorded steps than the first example. Despite local variations, its target distance exhibits a sustained downward trend. Both trajectories terminate within the 3 m target radius. These visualizations complement the quantitative evaluation by illustrating progressive target approach across search episodes of different lengths.



(a) 3D trajectory (b) Target-distance progression (c) Oblique scene view

Figure 3: Representative successful trajectory on $N_{\text{Island_0}}$. The panels show the 3D trajectory, target-distance progression, and the corresponding scene-level trajectory visualization.



(a) 3D trajectory (b) Target-distance progression (c) Oblique scene view

Figure 4: Representative successful Long-task trajectory on $N_{\text{Island_1}}$. The panels show the 3D trajectory, target-distance progression, and the corresponding scene-level trajectory visualization.

5 CONCLUSION

AerialWAM defines open-world SemanticOGS as a closed-loop interaction between a semantic-object reconstructor and a two-expert aerial world action model. At each decision, the reconstructor maps the current multi-view observation and semantic goal to a goal-centric object image; the WAM combines the reconstructed object image with the real observation, state and stop-condition tokens, and predicted future latent features to decode one flight action or STOP. The next observation updates the reconstruction and closes the observation-reconstruction-action-execution loop. Future work will explore collision-aware planning, reliable stopping, and transfer to unseen environments.

AI-USE STATEMENT

We used generative AI tools to assist with translation, literature search and reference checking.

REFERENCES

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Xuecheng Chen, Zongzhuo Liu, Jianfa Ma, Bang Du, Tiantian Zhang, Xueqian Wang, and Boyu Zhou. AirHunt: Bridging vlm semantics and continuous planning for efficient aerial object navigation, 2026.

Xiaowei Chi, Chun-Kai Fan, Hengyuan Zhang, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-Min Chan, Wei Xue, Qifeng Liu, Shanghang Zhang, and Yike Guo. Empowering world models with reflection for embodied video prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10383–10410, 2025.

Zehua Fan, Junjie He, Wenxuan Song, Xi Wang, Wenqi Lyu, Linge Zhao, Fuhao Li, Zihan You, Yifei Yang, Kaiming Xu, Qi Jiang, Yue Jiang, Haoang Li, Cheng Chi, Feng Gao, Bailin Li, and Yan Wang. MobileWAM: Bridging world action models to mobile manipulation with chain-of-foresight. arXiv preprint arXiv:2608.04657, 2026.

Tongtong Feng, Xin Wang, Haoran Hou, Ren Wang, Weiran Wang, Shaokai Zhu, Ziqi Jia, Hao Wang, Yu-Wei Zhan, Zongyuan Wu, Jinghao Cui, and Wenwu Zhu. AerialDojo-200K: A large-scale benchmark suite for open-world aerial object-goal search, 2026a.

Tongtong Feng, Xin Wang, Zekai Zhou, Ren Wang, Yuwei Zhan, Guangyao Li, Qing Li, and Wenwu Zhu. EvolvingAgent: Curriculum self-evolving agent with continual world model for long-horizon tasks. arXiv preprint arXiv:2502.05907, 2026b. Version 3; first posted in 2025.

Weiqi Gai, Yuman Gao, Yuan Zhou, Yufan Xie, Zhiyang Liu, Yuze Wu, Xin Zhou, Fei Gao, and Zhijun Meng. USS-Nav: Unified spatio-semantic scene graph for lightweight uav zero-shot object navigation, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

MoniJesu Wonders James, Amir Atef Habel, Aleksey Fedoseev, and Dzmitry Tsetserokou. GoalSwarm: Multi-uav semantic coordination for open-vocabulary object navigation, 2026.

Yatai Ji, Zhengqiu Zhu, Yong Zhao, Beidan Liu, Chen Gao, Yihao Zhao, Sihang Qiu, Yue Hu, Quanjun Yin, and Yong Li. Towards autonomous UAV visual object search in city space: Benchmark and agentic methodology. arXiv preprint arXiv:2505.08765, 2025.

Seungchan Kim, Omar Alama, Dmytro Kurdydyk, John Keller, Nikhil Keetha, Wenshan Wang, Yonatan Bisk, and Sebastian Scherer. RAVEN: Resilient aerial navigation via open-set semantic memory and behavior adaptation, 2025.

Adam Pardyl, Dominik Matuszek, Mateusz Przebieracz, Marek Cygan, Bartosz Zieliński, and Maciej Wolczyk. Flysearch: Exploring how vision-language models explore. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5594.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. ReconVLA: Reconstructive vision-language-action model as effective robot perceiver. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(22):18549–18557, 2026. doi: 10.1609/aaai.v40i22.38921.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology, 2024.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *International Conference on Learning Representations*, 2025.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjin Wu, Weili Guan, and Xiang Deng. UAV-ON: A benchmark for open-world object goal navigation with aerial agents, 2025.

Daoxuan Zhang, Ping Chen, Xiaobo Xia, Xiu Su, Ruichen Zhen, Jianqiang Xiao, and Shuo Yang. APEX: A decoupled memory-based explorer for asynchronous aerial object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15232–15242, 2026.

Weichen Zhang, Peizhi Tang, Xin Zeng, Fanhang Man, Shiquan Yu, Zichao Dai, Baining Zhao, Hongjin Chen, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, Xin Wang, Yong Li, and Wenwu Zhu. Aerial world model for long-horizon visual generation and navigation in 3D space. arXiv preprint arXiv:2512.21887, 2025.

Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, Weichen Zhang, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. WorldVLN: Autoregressive world action model for aerial vision-language navigation, 2026.

Shengtao Zheng, Kai Li, Weichen Zhang, Yu Meng, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. WorldFly: A world-model-based vision-language-action model for uav navigation, 2026.

Jiacong Zhou, Jiaxu Miao, Yourun Lin, Xianyun Wang, Jun Xiao, and Jun Yu. Memory-augmented scene understanding and exploration for open-world aerial object-goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21616–21626, 2026.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.015.