SELF-EVOLVING GROUNDING WORLD MODEL FOR DYNAMIC LONG-HORIZON OBJECT-GOAL AERIAL NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object-goal aerial navigation has witnessed rapid emergence with the development of vision–language–action models. However, dynamic long-horizon object-goal aerial navigation remains challenging: with goals arriving dynamically one after another, the model's current decision basis is repeatedly invalidated and a fresh direction must be determined entirely from its own observations, and as the chain grows long, the scene state and the value of earlier observations dynamically shift along the progressively revealed sequence and must be maintained rather than reset. Existing methods follow transient per-step observations with no direction signal and discard observation memory across goals, and therefore fail to handle the changing goals and to sustain success across the chain. To address these, we propose the Self-Evolving Grounding World Model (SE-GWM), a unified framework that couples goal grounding with memory evolution. Specifically, we introduce three collaborative modules: i) a **multi-view video grounding** module that performs *visual sequence grounding* over the current observations and remembered keyframes, converting each described object into an actionable 3-D goal estimate together with target, exploration, and avoidance judgments; ii) an **exploitation–exploration navigation** module that turns the grounded state into discrete flight actions through an evidence-bound planner and a goal-conditioned policy; and iii) **priority-gated memory evolution**, which preserves grounded evidence under a decaying admission threshold and feeds remembered keyframes back for re-grounding, letting the world model self-evolve with its own flight experience. Extensive experiments on our newly constructed UAV-ON-MG benchmark (1,000 episodes, 5,744 goals, 14 open-world scenes) show that SE-GWM consistently outperforms state-of-the-art methods.

1 INTRODUCTION

Object-goal aerial navigation has witnessed rapid emergence in recent years (Xiao et al., 2025; Liu et al., 2021; Chen et al., 2025). However, the shift to **dynamic long-horizon object-goal aerial navigation** poses two challenges: dynamically arriving goals repeatedly overturn whatever decision basis the model has built, and the direction to the active goal has to be re-derived from its own observations, and while the chain unrolls over a long horizon, earlier observations dynamically gain or lose relevance, so the accumulated observation memory must be carried forward and revised rather than reset, all within a finite step budget from egocentric multi-view observations alone.

Existing methods primarily focus on single-goal aerial ObjectNav (Xiao et al., 2025) and aerial vision–language navigation (Liu et al., 2021; Lee & Kim, 2024; Chen et al., 2025; Ding et al., 2026b). Single-goal ObjectNav methods locate one target per episode, relying on stopping heuristics or caption-based LLM reasoning that carries no signal about where the goal lies; aerial VLN and VLA methods (Xu et al., 2026; Sun et al., 2026; Zhang et al., 2025a; Lou et al., 2026) follow externally specified guidance without locating the goal from egocentric observations. Both are episodic: state resets at every goal, and the observation memory accumulated in earlier searches is discarded across goals. Consequently, they fail to sustain success once goals emerge dynamically as a chain. The primary difficulties of **dynamic long-horizon object-goal aerial navigation** lie in the following two aspects:

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.

- It is challenging to locate object goals from the model's own observations as goals arrive dynamically one after another: goals are specified only by category and never reveal where they lie, so without converting the described object into an actionable 3-D location, the model has no direction signal at all and its action distribution collapses onto constant maneuvers.
- It is challenging to keep pace with the dynamically changing goals and scenes over the long horizon: observations accumulate only during flight and earlier ones gradually go stale, so without maintaining and updating an internal state, the model either forgets what it has seen or acts on outdated evidence, and glimpses from earlier goals never benefit later ones.

To address these challenges, we propose the **Self-Evolving Grounding World Model** (SE-GWM), a novel framework for dynamic long-horizon object-goal aerial navigation. Our method differs from traditional episodic methods by explicitly grounding the scene state at each goal onset and by evolving a persistent memory across the whole chain, so that its world model keeps pace with its own flight experience. Specifically, our framework consists of three collaborative modules: a **Multi-View Video Grounding** (MVG) module that performs *visual sequence grounding* over current multi-view observations and remembered keyframes, lifting the active goal into world coordinates together with target, exploration, and avoidance judgments; an **Exploitation–Exploration Navigation** module whose **Evidence-bound Planner** (EBP) and goal-conditioned flight policy (EGP) turn the grounded state into discrete flight actions; and **Priority-gated Memory Evolution** (PAME), which preserves grounded evidence under a decaying admission threshold and feeds remembered keyframes back for re-grounding and memory-first goals to both the planner and the policy. Extensive experiments on our newly constructed **UAV-ON-MG** benchmark, which chains 1,000 episodes and 5,744 goals over the 14 open-world scenes of UAV-ON (Xiao et al., 2025) with every goal labeled by its cross-goal search context, demonstrate that SE-GWM consistently outperforms state-of-the-art methods.

In summary, we make the following contributions:

- We propose SE-GWM, a novel framework for dynamic long-horizon object-goal aerial navigation, which unifies grounded scene-state estimation, evidence-bound planning, goal-conditioned execution, and self-evolving memory in a closed loop.
- We introduce a multi-view video grounding module that performs visual sequence grounding over current multi-view observations and remembered keyframes, lifting targets observed earlier in the episode into world coordinates for later reuse, together with an evidence-bound planner that turns judgments into prioritized guidance.
- We introduce priority-gated memory evolution with a decaying admission threshold that retains and re-admits grounded evidence across the goal chain.
- We construct UAV-ON-MG, a 5,744-goal long-horizon benchmark with search-context stratification; extensive experiments demonstrate that SE-GWM significantly outperforms goal-conditioned, LLM/VLM, and migrated VLA baselines.

2 Related Work

Object goal navigation. ObjectNav (Anderson et al., 2018a) asks an agent to locate an object category from egocentric observations. Modular pipelines couple semantic priors with frontier-style exploration (Chaplot et al., 2020; Yokoyama et al., 2024; Zhou et al., 2023; Yu et al., 2023), recent work maps goal semantics into multimodal embeddings for zero-shot search (Majumdar et al., 2022), grounds open-vocabulary detection into online maps for efficient target search (Chabal et al., 2025), builds queryable multi-layered semantic feature maps (Raychaudhuri et al., 2025), accelerates reasoning over hierarchical scene graphs (Zhou et al., 2025), and couples world models or persistent memory with the search (Nie et al., 2025; Liu et al., 2025; Dong et al., 2026; Ge et al., 2026; Li et al., 2026; Torne et al., 2026). These benchmarks are single-goal, indoor, and near-planar, evaluated in embodied simulators such as Habitat (Savva et al., 2019); chain-level error accumulation and cross-goal memory are outside their scope. UAV-ON (Xiao et al., 2025) extends ObjectNav to aerial navigation in 14 open-world scenes with instance-level semantic goals and reports that all baselines (random, CLIP heuristic, and the LLM-based AOA) succeed on at most 7.3% of *single-goal* episodes. We build on UAV-ON's scenes and annotations but change the task: chained goals, per-goal budgets, and memory-stratified evaluation.

Aerial vision-language navigation. AerialVLN (Liu et al., 2021) extends grounded instruction-following navigation (Anderson et al., 2018b; Ku et al., 2020; Hong et al., 2020) to flight, with CityNav (Lee & Kim, 2024), TRAVEL (Wang et al., 2024), and OpenFly (Chen et al., 2025) typically using dense step-by-step language or privileged geographic context; HETT (Ding et al., 2026b) augments a two-stage transformer with navigation history, and SeqWalker (Han et al., 2026) handles sequential multi-task instructions with hierarchical planning. The 2026 generation moves toward end-to-end vision–language–action: AeroVLA (Xu et al., 2026) maps coarse instructions straight to continuous control, AutoFly (Sun et al., 2026) pairs a pseudo-depth encoder with large-scale flight data for navigation in the wild, Uni-LaViRA (Ding et al., 2026a) covers four embodied navigation tasks with pretrained MLLMs and no trajectory training, and WorldFly (Zheng et al., 2026) couples a world model with flow-matching action prediction for occlusion-robust urban flight, alongside video-VLA models (Zhang et al., 2024; 2025a; Lou et al., 2026; Hu et al., 2025; Zhang et al., 2025b; Xie et al., 2026) that translate observations into actions or waypoints. LLM planners grounded by value functions or affordances (Ahn et al., 2022; Driess et al., 2023) share our modular philosophy. When such methods lack an explicit goal-direction signal on our task, they collapse to constant actions; SE-GWM instead routes grounded goal evidence into both the planner and the policy state.

3 METHOD

In this section, we introduce SE-GWM. As shown in Figure 1, the proposed framework consists of three collaborative modules: a Multi-View Video Grounding (MVG) module, an Exploitation–Exploration Navigation module, and a Priority-gated Memory Evolution (PAME) module, which operate jointly to enable self-evolving navigation over long goal chains. Given the current four-view observation and the active goal, the MVG module first performs two-branch grounding: a trained discriminative read-out fuses the four views into a grounded scene state, while a zero-shot generative read-out re-grounds the goal over remembered keyframes, together producing a 3-D goal estimate and target, exploration, and avoidance judgments. The navigation module then turns the grounded state into behavior: an evidence-bound planner decomposes it into prioritized sub-goal guidance, and a goal-conditioned flight policy executes discrete flight actions. PAME in turn preserves grounded evidence into dual memory banks under a decaying admission threshold and samples keyframes along the trajectory, feeding remembered keyframes back to the MVG module and memory-first goals to the policy. This ground–plan–act–remember loop ensures that the goal signal stays anchored to visual evidence while the memory grows with the framework's own flight experience, turning targets observed in earlier searches into actionable goals for later ones. The memory is reset at episode end, so evolution accumulates only within a chain.

3.1 PROBLEM FORMULATION

An episode $\mathbf{e} = (c_1, \dots, c_K)$ specifies a chain of K object goals, each an instance-level semantic description $c_k = (\text{name}, \text{size}, \text{description})$ in the style of UAV-ON. The UAV starts at pose $p_0 = (x, y, z, \psi)$ and, at step t, observes four synchronized egocentric RGB views $O_t = \{o^{\text{fr}}, o^{\text{lf}}, o^{\text{rt}}, o^{\text{dn}}\}_t$ (front/left/right/down), without GPS, maps, depth, or privileged object poses. The action space is seven discrete maneuvers whose kinematics define the transition model. With heading unit vector $u(\psi) = (\cos\psi, \sin\psi, 0)^\top$, planar step length ℓ, vertical step h, and rotation increment ω,

$$p_{t+1} = \begin{cases} p_t + \ell\, u(\psi_t), & a_t = \texttt{fwd}, \\ p_t + \ell\, u\big(\psi_t \pm \frac{\pi}{2}\big), & a_t \in \{\texttt{left}, \texttt{right}\}, \\ p_t \pm h\, \mathbf{e}_z, & a_t \in \{\texttt{up}, \texttt{down}\}, \\ p_t, & a_t \in \{\texttt{rot}, \texttt{rot}'\}, \end{cases} \quad \psi_{t+1} = \psi_t \pm \omega\, \mathbf{1}\big[a_t \in \{\texttt{rot}, \texttt{rot}'\}\big]. \tag{1}$$

Here $\mathbf{1}[\cdot]$ denotes the indicator function. Each goal c_k is granted a budget of T steps collected in the window $\mathcal{T}_k$; a goal is *found* when the UAV comes within radius r of the annotated position x_k, and memory carries over across goals within the episode. Writing Θ for all learned parameters, the long-horizon control objective is the expected number of goals found per episode,

$$\mathcal{J}(\Theta) = \mathbb{E}_{\mathbf{e}\sim\mathcal{D}}\left[\sum_{k=1}^{K} \mathbf{1}\Big\{\min_{t\in\mathcal{T}_k} \|p_t - x_k\| < r\Big\}\right], \tag{2}$$

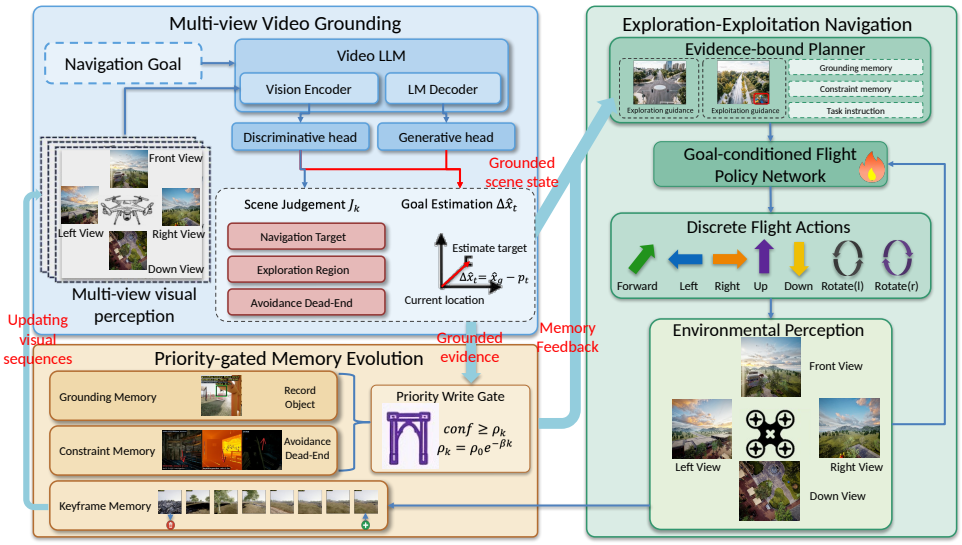

Figure 1: Framework of our Self-Evolving Grounding World Model. Given the current four-view observation and the active goal, multi-view video grounding performs two-branch grounding over current views and remembered keyframes; exploitation–exploration navigation turns the grounded state into discrete flight actions through an evidence-bound planner and a goal-conditioned policy; priority-gated memory evolution preserves grounded evidence under a decaying admission threshold, feeds remembered keyframes back to the grounding module for re-grounding, and returns memory feedback to the navigation module as memory-first goals and summaries.

which upper-bounds the probability of completing the full chain and makes explicit the two coupling sources studied in this paper: the dependence of each goal's search context on the observations accumulated in earlier windows, and the persistence of memory across windows $\mathcal{T}_1, \ldots, \mathcal{T}_K$.

3.2 Multi-View Video Grounding (MVG)

The MVG recasts visual grounding for long-horizon chains as *visual sequence grounding*: rather than grounding the goal in a single snapshot of the scene, it grounds the goal in a temporally extended visual sequence that concatenates the current four egocentric views with keyframes remembered over the episode, so that targets glimpsed while pursuing earlier goals remain first-class evidence for later ones. Formally, let E denote the visual encoder shared by all views, inherited from a pre-trained vision–language backbone and kept frozen; applied to each view $v \in \mathcal{V} = \{\text{front}, \text{left}, \text{right}, \text{down}\}$, it yields per-view token features $F_v = E(o^v) \in \mathbb{R}^{N \times d}$. Because no single view observes the goal-relevant geometry of an open-world scene, a trainable **cross-view attention** (CVA) module reconciles the views: each view's tokens attend bidirectionally to the concatenation of all views, and the attended features are aggregated with learned view weights,

$$\tilde{F}_v = \mathrm{MHA}_v\big(F_v,\ [F_{\mathrm{fr}}; F_{\mathrm{lf}}; F_{\mathrm{rt}}; F_{\mathrm{dn}}],\ [F_{\mathrm{fr}}; F_{\mathrm{lf}}; F_{\mathrm{rt}}; F_{\mathrm{dn}}]\big), \qquad F^{\mathrm{fus}} = \mathrm{proj}\Big(\textstyle\sum_v \alpha_v \tilde{F}_v\Big), \tag{3}$$

where $\alpha = \mathrm{softmax}(a)$ and the block is stacked L times. On the fused state, a grounding head regresses the metric goal position $\hat{x}_g \in \mathbb{R}^3$, and a set of learnable query tokens cross-attending to F^{fus} decode three categories of *judgments*,

$$J_k = \big\{\mathcal{T}_i(\ell_i, s_i, c_i)\big\} \cup \big\{\mathcal{E}_j(\ell_j, \hat{g}_j, c_j)\big\} \cup \big\{\mathcal{A}_m(\ell_m, c_m)\big\}, \tag{4}$$

each entry carrying a 3-D location ℓ, a source-view provenance token, and a confidence c; exploration entries additionally regress an information-gain estimate $\hat{g}$. The three categories partition the scene into *confirmed goal evidence*, *under-observed regions of prospective value*, and *exhausted regions*, exactly the triad an explore–exploit agent must balance. The MVG is trained end to end on

goal-annotated demonstrations with the multi-task objective

$$\mathcal{L}_{MVG} = \lambda_g \big\|\hat{x}_g - x_g\big\|_1 + \sum_{j \in J} \Big[\mathrm{BCE}\big(s_j, y_j\big) + \lambda_\ell \,\mathrm{SmoothL}_1\big(\ell_j, \ell_j^*\big) + \lambda_{\mathrm{ig}}\big(\hat{g}_j - g_j^*\big)^2\Big], \quad (5)$$

where y_j are demonstrated admission labels, (ℓ_j^*, g_j^*) are pseudo-targets derived from geodesic proximity, and the backbone is adapted parameter-efficiently while CVA_θ and all heads train from scratch.

Generative read-out: sequence grounding over memory. The temporal axis of the MVG is an *updating visual sequence*, i.e., an episode-level keyframe memory, sampled every m steps and sparsified by decimation so coverage spans the whole trajectory, with ξ_i the camera pose of dominant view v^* at step i:

$$\begin{aligned} \mathcal{K}_t &= \mathrm{sparsify}\big(\{(o_i^{v^*}, \xi_i) : i \le t,\ i \bmod m = 0\},\ N_{\max}\big), \\ (f^*, b^*, c^*) &= \mathrm{VLM}\big(\{o_f^{v^*}\}_{f \in \mathcal{K}_t} \cup O_t,\ c_k\big), \end{aligned} \quad (6)$$

where the second line, invoked at task boundaries and every n steps, asks the video LLM which remembered or current frame shows the goal and where: grounding thus ranges over the whole visual history of the episode, not merely the present observation. The detected box is lifted to world coordinates through the inverse projection of that frame under a depth prior d_0,

$$\hat{x}_k^{\mathrm{mem}} = T^{w}_{c(f^*)} \Big(d_0\, K_{v^*}^{-1}\, \tilde{b}^*\Big), \quad (7)$$

where $\tilde{b}^* = (u^*, v^*, 1)^\top$ is the homogeneous box center, K_{v^*} the view intrinsics, and $T^{w}_{c(f^*)}$ the camera-to-world pose of frame f^*. Responses whose frame identifier lies outside $\mathcal{K}_t \cup \{\text{current}\}$ are discarded; accepted estimates enter the grounding memory as *grounded evidence* under the memory admission gate. Eq. 6 together with Eq. 7 closes the visual loop from memory back into the world model: what was glimpsed at step i can be re-localized, in metric space, at any later goal. This is the mechanism that turns past observations into usable goal evidence, and it distinguishes visual sequence grounding from single-image grounding even when both produce 3-D estimates. The two read-outs place incompatible demands on one backbone: metric state estimation needs the calibrated supervision of Eq. 5, while open-world re-grounding benefits from the untouched prior of the native VLM, so the discriminative stack is trained and the generative stack runs zero-shot.

Two commitments complete the design: grounding is invoked *once per goal onset* and cached (a stable frame within a goal, an order of magnitude fewer backbone calls than per-step grounding), and the estimate $\hat{x}_g$ is routed into the policy state as $\Delta\hat{x}_t$, closing the loop between grounding and control.

3.3 Evidence-bound Planner (EBP)

The *grounded scene state* flows from the MVG into both the planner and the policy: it is the interface that module ii consumes. At each goal onset it consumes the goal specification, the judgments J_k, and confidence-ranked summaries $(\Sigma_{\mathcal{M}}, \Sigma_{\mathcal{C}})$ of the memory banks, and produces a sub-goal decomposition together with an execution directive,

$$\hat{c}_k \ \sim\ \pi_{\mathrm{LLM}}\big(\hat{c} \,\big|\, c_k,\ J_k,\ \Sigma_{\mathcal{M}},\ \Sigma_{\mathcal{C}}\big) \quad \text{s.t.} \quad \mathrm{supp}(\hat{c}_k) \ \subseteq\ \mathrm{supp}(J_k) \cup \mathrm{supp}(\mathcal{M}) \cup \mathrm{supp}(\mathcal{C}), \quad (8)$$

under the fixed priority ordering *Target* $\succ$ *Avoidance* $\succ$ *Exploration*, which totally orders directives by category first and confidence second. The planner is a frozen instruction-following LLM operated zero-shot, so all task-specific competence enters through the structured evidence rather than through weights. The support constraint in Eq. 8 makes the interface *evidence-bound by construction*: the planner may decompose, re-rank, or defer among the supplied judgments, but it cannot introduce evidence outside them, which confines the LLM to its comparative advantage, compositional reasoning over grounded variables, and away from hallucinating geometry. A deterministic fallback π_{rule} that enforces the same ordering on raw judgments guarantees graceful degradation.

3.4 Exploitation–Exploration Flight Policy (EGP)

The EGP closes the loop from grounded state to motor commands. It follows the vision–language–action recipe (Brohan et al., 2023; Black et al., 2024; Kim et al., 2024; Octo Model Team et al.,

2024) of a shared multi-view visual encoder, a state encoder, and a fused discrete action head, with a state representation that is deliberately minimal and **goal-conditioned**:

$$s_t = \big[\, \gamma_p(p_t),\ \psi_t,\ \gamma_\Delta(\Delta\hat{x}_t) \,\big], \qquad \Delta\hat{x}_t = \hat{x}_{g_k} - p_t, \tag{9}$$

where γ_p, γ_Δ are fixed normalizations and $\Delta\hat{x}_t$ is the goal-relative displacement induced by the MVG, the point-goal-style signal (Anderson et al., 2018a) that carries the goal direction to navigation. Conditioned on s_t, the averaged view embedding $z_t = \frac{1}{|\mathcal{V}|}\sum_v f_\phi(o_t^v)$, and the planner's sub-goal embedding $e(\hat{c}_k)$, the action head defines a categorical policy

$$\begin{aligned} \pi_\phi(a \mid O_t, s_t, \hat{c}_k) &= \mathrm{softmax}\big(W_a\, \sigma(h_t) + b_a\big), \\ c_t = \sigma\big(w_c^\top \sigma(h_t)\big), &\qquad h_t = \sigma\big(W_f[z_t; s_t; e(\hat{c}_k)] + b_f\big), \end{aligned} \tag{10}$$

where the auxiliary scalar c_t is an action-confidence estimate that later gates memory writes.

Two regimes. The name of the policy is deliberate: it runs in the two regimes the planner's prioritized guidance arbitrates. When target evidence is grounded, the resolved goal supplies $\Delta\hat{x}_t$ and the policy *exploits* it on approach; when nothing is grounded yet, $\Delta\hat{x}_t$ is absent and the agent *explores* under avoidance constraints while sequence grounding re-examines the keyframe memory, until a target observed earlier is lifted into $\Delta\hat{x}_t$ and the regime flips back to exploitation. Stagnation recovery safeguards both regimes; it does not arbitrate between them.

Stagnation recovery. Imitation policies occasionally enter self-reinforcing loops (e.g., perpetual rotation); we complement the policy with a stagnation operator on the distance trace $d_t = \|p_t - \hat{x}_g\|$, operationalizing the MVG's "target → backtrack" directive,

$$S_t = \mathbf{1}\Big\{ \min_{i\in(t-W,\,t]} d_i \ > \ \min_{i\in(t-W-\delta,\,t-W]} d_i - \varepsilon \Big\}, \tag{11}$$

and, when $S_t = 1$, overrides the commanded action by the closed-form re-aiming law

$$a_t \ \leftarrow\ \begin{cases} \mathrm{rot}\big(\mathrm{sign}(\Delta\psi_t^*)\big), & |\Delta\psi_t^*| > \tau_\psi, \\ \texttt{forward}, & \text{otherwise}, \end{cases} \qquad \Delta\psi_t^* = \mathrm{wrap}\big(\mathrm{atan2}(\hat{x}_{g,y} - p_{t,y},\ \hat{x}_{g,x} - p_{t,x}) - \psi_t\big), \tag{12}$$

with $\mathrm{wrap}(\cdot)$ folding the angle difference to $(-\pi, \pi]$. This single rule yields large chain-level gains, identifying divergence, rather than perception or step budget, as the dominant intra-goal failure mode.

Curricular training. Jointly training both pathways from random initialization collapses to majority actions, since an untrained visual encoder injects variance that overwhelms the goal signal early. We therefore first fit the state-to-action mapping with the visual pathway masked, then admit vision at a strongly reduced learning rate as a residual correction (schedules in Sec. 5.1).

3.5 Priority-Gated Memory Evolution (PAME)

Memory is what turns a per-goal navigator into a self-evolving one. A lightweight spatio-temporal scorer consumes the current experience window (multi-view observations, the recent trajectory, and the executed actions) and regresses an information gain together with a dead-end probability,

$$\hat{g}_t = \sigma\big(w_g^\top \phi_\theta(W_t)\big), \qquad \hat{d}_t = \sigma\big(w_d^\top \phi_\theta(W_t)\big), \qquad W_t = \big[\{o_{t-\tau:t}^v\}_{v\in\mathcal{V}};\ p_{t-\tau:t};\ a_{t-\tau:t}\big], \tag{13}$$

where ϕ_θ is a convolutional trunk over the stacked window. The information-gain head is a deliberately predictive component of the world model: it estimates where future observations will be most informative, in the spirit of curiosity-driven world models (Nie et al., 2025). These signals, combined with the policy's action confidence c_t from Eq. 10, drive a rule-based write gate (the full rule table is given in Appendix D).

Two complementary banks persist across the whole episode. The **grounding memory** $\mathcal{M}$ retains high-value target judgments, i.e., fused visual features with metric locations and confidences, deduplicated in space and queryable by object identity and proximity. The **constraint bank** $\mathcal{C}$ retains dead-end regions, consulted by the planner as context and by the policy as a proximity-triggered avoidance signal. Admission is governed by a **decaying threshold**

$$\rho_k = \rho_0\, e^{-\beta k}, \tag{14}$$

which lowers the bar for writing as the chain progresses: evidence glimpsed early but judged marginal at the time can still be admitted late in the episode, when the marginal value of any retained cue is higher. Memory is read back through a *memory-first* resolution of the active goal signal,

$$\hat{x}_g = \begin{cases} \arg\max_{m \in \mathcal{M}(c_k)} \ \mathrm{sim}(c_k, m) \ \text{ s.t. } \ \varsigma_m \geq \tau_m, & \mathcal{M}(c_k) \neq \varnothing, \\ \hat{x}_k^{\mathrm{plan}}, & \text{planner directive available}, \\ \hat{x}_k^{\mathrm{task}}, & \text{otherwise}, \end{cases} \tag{15}$$

where $\mathcal{M}(c_k)$ are entries matching the goal identity, sim compares goal and entry embeddings, ς_m is the stored confidence, and the last two branches form the fallback chain; entries of $\mathcal{C}$ act on the action side by vetoing maneuvers whose heading re-enters a stored dead-end region. This *memory feedback* closes the loop: at the next goal onset, summaries of $\mathcal{M}$ and $\mathcal{C}$ re-enter the EBP through Eq. 8, while re-grounded estimates from Eq. 7 keep refreshing the banks; this is the feedback loop that lets the world model accumulate and reorganize its own grounded state across goals, i.e., *self-evolve*. Unlike reflection-driven write-back (Ge et al., 2026) or fixed-scale retrieval buffers (Torne et al., 2026), the decaying threshold makes the write policy time-dependent: the same marginal cue is rejected early and admitted late, matching the rising marginal value of memory as the chain progresses.

4 The UAV-ON-MG Benchmark

Construction. UAV-ON-MG is built on UAV-ON's 14 held-out test scenes and its 1,270 instance-annotated objects (Xiao et al., 2025). We compose goals from 291 object categories into **chains** of $K = 3{-}8$ goals (mean 5.74) with inter-goal distances of 8–55 m, yielding 1,000 episodes and 5,744 goals, with difficulty tiers assigned by chain length (Fig. 2c). Every goal also carries a size label from the UAV-ON annotations based on the simulator footprint of the object mesh; we group goals into *small*, *medium*, and *large* tiers (64.5%/18.9%/16.6% of goals), and stratify all per-goal results by this label. Evaluation follows Sec. 3.1; all methods share the identical kinematic execution protocol (local pose tracking; the simulator renders observations), so no method benefits from privileged physics. Unlike UAV-ON, whose baselines we adopt and diagnose, UAV-ON-MG changes the task itself: chained goals, per-goal budgets, and memory-stratified evaluation that no existing aerial benchmark controls for.

Search-context stratification. The defining statistic of UAV-ON-MG is that every goal is labeled by how observable it was from vantage points earlier in the same chain (Fig. 2b): *initial* (first goal, 17.4%), *blind* (never visible before, 23.6%), *partial recall* (glimpsed once, 39.0%), and *memory-guided* (clearly visible 2+ times, 19.9%). This stratification converts "does the agent exploit cross-goal memory?" from an aggregate impression into a measurable contrast: a memoryless agent's per-goal success should be flat across contexts, while a self-evolving one should show a monotone advantage from initial through memory-guided goals.

5 Experiments

5.1 Implementation Details

Setup. Episodes use $K \in [3, 8]$ goals ($T{=}30$ steps per goal, success radius 20 m) from UAV-ON-MG (Sec. 4). All modules are trained on goal-directed heuristic demonstrations collected from the 10 UAV-ON *train* scenes (36,145 steps); the MVG builds on Qwen3-VL-8B-Instruct (Bai et al., 2025), the policy is a 4-layer MLP, and the scorer is ResNet-18.

Baselines. We compare against AirVLN (Liu et al., 2021), VoLN (Lou et al., 2026), See-Point-Fly (Hu et al., 2025), TravelUAV (Wang et al., 2024), AeroVLA (Xu et al., 2026), AOA-F (Xiao et al., 2025), and UniNaVid (Zhang et al., 2025a).

Metrics. We report **chain SR** (all K goals found) and **avg. goals found** (tf, the empirical estimate of Eq. 2), the latter as primary since chain SR is conservative for long chains. Training details, demonstration collection, hyperparameters, and per-component configurations are given in Appendix A.

Table 1: Total per-goal results on UAV-ON-MG. SR = success rate (%, within 20 m); OSR = ever entering the 25 m observation ring; SPL = success weighted by path length (%); DCA = median closest approach (m). **Bold** = best, underline = second.

Method	SR↑	OSR↑	SPL↑	DCA↓
AirVLN (Liu et al., 2021)	7.0	13.4	7.0	40.1
VoLN (Lou et al., 2026)	6.7	13.4	6.7	39.6
See-Point-Fly (Hu et al., 2025)	23.1	34.9	22.9	31.8
TravelUAV (Wang et al., 2024)	18.1	25.8	18.1	36.1
AeroVLA (Xu et al., 2026)	9.8	18.0	9.8	38.8
AOA-F (Xiao et al., 2025)	19.5	28.3	19.4	35.0
UniNaVid (Zhang et al., 2025a)	25.9	36.7	25.9	32.4
SE-GWM (ours)	**27.5**	**38.6**	**27.3**	**31.1**

Table 2: Per-goal results stratified by object size tier (Small/Medium/Large) on UAV-ON-MG.

	Small				Medium				Large			
Method	SR↑	OSR↑	SPL↑	DCA↓	SR↑	OSR↑	SPL↑	DCA↓	SR↑	OSR↑	SPL↑	DCA↓
AirVLN	7.9	13.9	7.9	39.5	4.5	11.0	4.5	44.0	6.4	14.6	6.4	38.5
VoLN	7.3	13.9	7.3	39.4	3.7	8.9	3.7	45.3	7.9	16.6	7.9	38.5
See-Point-Fly	20.5	35.2	20.3	**30.9**	23.2	28.9	23.2	34.2	**33.2**	**40.9**	**32.8**	30.6
TravelUAV	18.3	25.3	18.3	35.7	15.1	21.8	15.1	39.3	20.7	31.9	20.7	35.6
AeroVLA	10.3	17.0	10.3	38.8	6.4	15.8	6.4	41.6	11.8	24.1	11.8	33.3
AOA-F	20.1	28.4	20.1	34.4	14.0	21.4	13.9	40.1	23.3	35.7	23.3	31.7
UniNaVid	24.3	36.2	24.3	34.0	28.2	37.6	28.2	30.0	29.1	37.7	29.1	27.8
SE-GWM (ours)	**25.4**	**37.4**	**25.4**	33.0	**30.6**	**40.6**	**30.4**	**28.8**	31.5	40.8	31.2	**27.6**

5.2 Main Results

We provide the total per-goal results in Table 1 and their size-tier stratification in Table 2; all numbers are per-goal over the same covered subset, with higher better for SR/OSR/SPL and lower better for DCA.

Headline. As shown in Table 1, SE-GWM attains the best total value on all four metrics (SR 27.5%, OSR 38.6%, SPL 27.3%, DCA 31.1 m), and leads the approach-quality columns of the Medium and Large tiers as well (DCA 28.8/27.6 m vs. 30.0/27.8 m for the strongest baseline, Table 2). It is the only system in the table whose goal direction is grounded from its own observations; the strongest baseline reaches 25.9% total SR (UniNaVid, the straight-flight line to which direction-conditioned models collapse without a direction signal; Appendix E), and instruction- or caption-driven models drop to 7–10% (AirVLN 7.0%, VoLN 6.7%, AeroVLA 9.8%) and do not steer flight without an explicit goal-direction signal.

Size tiers. SE-GWM success rises monotonically with object size (SR 25.4→30.6→31.5%; DCA 33.0→27.6 m): larger objects remain re-identifiable in the keyframe memory at greater range, so the re-grounded goal signal stays usable, whereas the episodic baselines are flat or non-monotone across tiers (e.g., UniNaVid 24.3/28.2/29.1, AOA-F 20.1/14.0/23.3).

Small objects are the most competitive tier (the top two methods within 1.1 points); zero-shot pointing (See-Point-Fly) leads the Large tier (33.2% SR) but is evaluated on 48% of the goal set and falls behind on the two smaller tiers (SR 20.5/23.2%).

Approach quality. OSR exceeds SR by 11.1 points for SE-GWM (38.6% vs. 27.5%): nearly a third of the goals reached within the 25 m ring are not closed to 20 m, which we attribute to the fixed depth prior in back-projection; DCA of 31.1 m against a 20 m success radius quantifies the same bottleneck and marks depth as the next lever.

Table 3: Component ablations of the memory pathway (per-goal, exploration protocol, same covered subset). Dose–response statistics are per-episode means of pipeline counters.

Write gate	Keyframes	Results				Memory throughput (per episode)		
		SR↑	OSR↑	SPL↑	DCA↓	writes	goal hits	re-grounds
0.6	**every 2**	**26.9**	**38.1**	**26.8**	**31.4**	**0.72**	**176.2**	**0.91**
0.9	every 2	26.4	37.5	26.4	31.4	0.51	67.3	0.63
0.6	every 4	25.9	37.0	25.9	31.8	0.44	61.2	0.63

Where the chain slows. Per-goal success decays mildly with chain position (27.8% at position 1 to 22.4% at position 8): later goals sit more often in memory-dependent contexts and inherit the accumulated depth-prior error of earlier re-groundings, but the decay is far from a cliff, indicating that the memory write–read loop keeps functioning deep into the chain.

Search contexts. Blind goals are the easiest context (32.1%) and memory-guided goals the hardest (22.4%). This ordering is confounded by chain position: memory-guided goals appear late by construction, where budgets are thinner and errors have accumulated. The position-controlled contrast that would isolate the memory contribution is part of the ongoing analysis.

5.3 Ablations

The exploration contrast of Table 4 doubles as the memory ablation: every configuration that expects goal direction as an input degenerates into near-constant straight flight in the 26% band, whereas SE-GWM, whose only direction source is its own visual memory, holds 27.5% overall. The memory pathway is the difference between having a grounded direction and having none.

Two component ablations probe the design of this pathway (Table 3). Raising the write-confidence gate from 0.6 to 0.9 costs 0.5 points of SR, validating the default admission threshold.

Halving the keyframe sampling density (every 4 instead of every 2 steps) costs 1.4 points, with all four metrics degrading consistently, confirming that re-grounding depends on dense visual history.

The remaining component ablations (planner, grounding frequency, recovery) are being rerun; the fixed-prior depth bottleneck identified by the SR–OSR gap is their first target.

Dose–response of the memory pathway. Both ablations cut memory throughput sharply while leaving the VLM call budget unchanged (7.97 calls per episode in all three configurations): the stricter gate reduces writes by 29% and memory goal hits by 62%, the sparser sampler reduces writes by 39% and hits by 65%, and success falls accordingly. Search capability thus tracks the throughput of the write–read–re-ground loop rather than model capacity, quantitatively confirming that the memory pathway, not perception alone, carries the exploration gain.

6 Conclusion

In this paper, we propose the Self-Evolving Grounding World Model (SE-GWM), a unified framework for dynamic long-horizon object-goal aerial navigation that decomposes the task into grounded goal estimation, evidence-bound flight, and gated memory evolution. Our multi-view video grounding module accurately lifts each described goal into an actionable 3-D estimate together with target, exploration, and avoidance judgments, while priority-gated memory evolution effectively preserves grounded evidence under a decaying admission threshold and feeds remembered keyframes back for re-grounding. We further construct UAV-ON-MG, a 5,744-goal benchmark stratified by cross-goal search context. Extensive experiments on UAV-ON-MG demonstrate that SE-GWM consistently outperforms state-of-the-art baselines across size tiers, with ablation studies validating the memory write gate and keyframe sampling density. These results highlight the importance of self-evolving memory for dynamic long-horizon object-goal aerial navigation, and suggest that our framework can work orthogonally with advancing vision–language–action models to achieve more capable aerial navigation.

AI USE STATEMENT

In this work we used generative AI coding assistants for boilerplate engineering (data conversion and evaluation scripts), all of which were reviewed and tested by the authors. We used a VLM (as reported in the text) as a *component* of the benchmark annotation pipeline to verify object visibility; its outputs were manually audited, and objects failing verification were removed. All scientific claims, experimental design, and analysis were conducted and are owned by the authors.

ETHICS STATEMENT

This work uses simulation only; no human subjects or personal data are involved. Aerial navigation research is intended for beneficial applications (inspection, search and rescue, delivery); dual-use risks (surveillance) are noted, and our benchmark does not include any person-identifying targets beyond the synthetic figures shipped with the simulator.

REPRODUCIBILITY STATEMENT

All scripts, configurations, and result files are being prepared for anonymous release; hyperparameters and protocols are specified in Secs. 3.1 and 5.1, with prompts for the EBP in the appendix.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018a.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

Thomas Chabal, Shizhe Chen, Jean Ponce, and Cordelia Schmid. FOM-Nav: Frontier-object maps for object goal navigation. *arXiv preprint arXiv:2512.01009*, 2025.

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yunlong Chen et al. OpenFly: A unified platform for aerial vision-and-language navigation. *arXiv preprint arXiv:2501.02643*, 2025.

Hongyu Ding, Sizhuo Zhang, Ziming Xu, Jinwen Guo, Hongxiu Liu, Xingzhi Cheng, Zixuan Chen, Haifei Qi, Duo Wang, Hao Xu, Jieqi Shi, Yifan Zhang, Jian Huo, Jian Cheng, Yang Gao, and Jiebo Luo. Uni-LaViRA: Language-vision-robot actions translation for unified embodied navigation. *arXiv preprint arXiv:2605.27582*, 2026a.

Xichen Ding, Jianzhe Gao, Cong Pan, Wenguan Wang, and Jie Qin. History-enhanced two-stage transformer for aerial vision-and-language navigation. In *AAAI Conference on Artificial Intelligence*, 2026b.

Yifei Dong, Fengyi Wu, Guangyu Chen, Lingdong Kong, Qiyu Hu, Yuxuan Zhou, Xu Zhu, Jingdong Sun, Jun-Yan He, Qi Dai, Alexander G Hauptmann, and Zhi-Qi Cheng. Towards unified world models for visual navigation via memory-augmented planning and foresight. In *ECCV*, 2026.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, et al. PaLM-E: An embodied multimodal language model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Zuhao Ge, Xiaosong Jia, Chao Wu, Yuchen Zhou, Zuxuan Wu, and Yu-Gang Jiang. EvoMemNav: Efficient self-evolving fine-grained memory for zero-shot embodied navigation. *arXiv preprint arXiv:2606.03509*, 2026.

Zebin Han, Xudong Wang, Baichen Liu, Qi Lyu, Zhenduo Shang, Jiahua Dong, Lianqing Liu, and Zhi Han. SeqWalker: Sequential-horizon vision-and-language navigation with hierarchical planning. In *AAAI*, 2026.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language BERT for navigation. *arXiv preprint arXiv:2011.13922*, 2020.

Chih Yao Hu, Yang-Sen Lin, Yuna Lee, Chih-Hai Su, Jie-Ying Lee, and Shr-Ruei Tsai. See, point, fly: A learning-free VLM framework for universal unmanned aerial navigation. In *Conference on Robot Learning (CoRL)*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Jungdae Lee and Gunhee Kim. CityNav: Language-goal aerial navigation dataset with geographic information. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Ji Li, Bo Wang, Jing Xia, Mingyi Li, and Shiyan Hu. HIMM: Human-inspired long-term memory modeling for embodied exploration and question answering. *arXiv preprint arXiv:2602.15513*, 2026.

Fei Liu, Shichao Xie, Minghua Luo, Zedong Chu, Junjun Hu, Xiaolong Wu, and Mu Xu. NavForesee: A unified vision-language world model for hierarchical planning and dual-horizon navigation prediction. *arXiv preprint arXiv:2512.01550*, 2025.

Shubo Liu, Hongsheng Zhang, Kai Yu, and Aixin Zheng. AerialVLN: Vision-and-language navigation for UAVs in city environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Jiabin Lou, Haopeng Wang, Yuanshuai Wang, Xinyu Liu, Xuxin Lv, Yuxin Guo, Lei Huang, Rongye Shi, and Wenjun Wu. VoLN: Vision-only long-horizon navigation—paradigm, benchmark, and method. *arXiv preprint arXiv:2607.21400*, 2026.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. *arXiv preprint arXiv:2206.12403*, 2022.

Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. WMNav: Integrating vision-language models into world models for object goal navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Sonia Raychaudhuri, Enrico Cancelli, Tommaso Campari, Lamberto Ballan, Manolis Savva, and Angel X Chang. MLFM: Multi-layered feature maps for richer language understanding in zero-shot semantic navigation. *arXiv preprint arXiv:2507.07299*, 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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

Xiaolou Sun, Wufei Si, Wenhui Ni, Yuntian Li, Dongming Wu, Fei Xie, Runwei Guan, He-Yang Xu, Henghui Ding, Yuan Wu, Yutao Yue, Yongming Huang, and Hui Xiong. AutoFly: Vision-language-action model for UAV autonomous navigation in the wild. In *International Conference on Learning Representations (ICLR)*, 2026.

Marcel Torne, Karl Pertsch, Homer Walke, Kyle Vedder, Suraj Nair, Brian Ichter, Allen Z. Ren, Haohuan Wang, Jiaming Tang, Kyle Stachowicz, Karan Dhabalia, Michael Equi, Quan Vuong, Jost Tobias Springenberg, Sergey Levine, Chelsea Finn, and Danny Driess. MEM: Multi-scale embodied memory for vision language action models. *arXiv preprint arXiv:2603.03596*, 2026.

Xiangyu Wang et al. TRAVEL: UAV vision-and-language navigation with platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087*, 2024.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjing Wu, Weili Guan, and Xiang Deng. UAV-ON: A benchmark for open-world object goal navigation with aerial agents. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, 2025.

Zhipeng Xie, Zongyi Han, Xiangyi Wei, Shiliang Sun, Yang Li, and Jing Zhao. S^2-VLA: State-space guided vision-language-action models for long-horizon manipulation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2026.

Peng Xu, Zhengnan Deng, Jiayan Deng, Zonghua Gu, and Shaohua Wan. AeroVLA: A vision-language-action model for UAV navigation via minimalist end-to-end control. In *European Conference on Computer Vision (ECCV)*, 2026.

Naoki Yokoyama, Dhruv Batra, and Sehoon Ha. VLFM: Vision-language frontier maps for zero-shot semantic navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Bangguo Yu, Haolan Zhang, Wangyi Tan, Shangdong Xia, Yu Mao, Hongyu Liu, Guan Wang, and Shu Wang. L3MVN: Leveraging large language models for visual target navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. NaVid: Video-based VLM plans the next step for vision-and-language navigation. In *Robotics: Science and Systems (RSS)*, 2024.

Jiazhao Zhang, Hongjie Wang, Xiaomeng He, Rongtao Wang, Yicong Xie, Jung-woo Hwang, Gaoang Wang, Meng Ye, Zhizheng Liu, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Robotics: Science and Systems (RSS)*, 2025a.

Yuhang Zhang, Haosheng Yu, Jiaping Xiao, and Mir Feroskhan. Grounded vision-language navigation for UAVs with open-vocabulary goal understanding. *arXiv preprint arXiv:2506.10756*, 2025b.

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. *arXiv preprint arXiv:2606.06147*, 2026.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. ESC: Exploration with soft commonsense constraints for zero-shot object navigation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Xiaolin Zhou, Tingyang Xiao, Liu Liu, Yucheng Wang, Maiyue Chen, Xinrui Meng, Xinjie Wang, Wei Feng, Wei Sui, and Zhizhong Su. FSR-VLN: Fast and slow reasoning for vision-language navigation with hierarchical multi-modal scene graph. *arXiv preprint arXiv:2509.13733*, 2025.

A Additional Implementation Details

Action space. forward/left/right translate 5 m along the yaw-aligned or lateral axis; ascend/descend translate 2 m vertically; rotate left/right turn 15°. All methods use the same fixed-magnitude discrete action set and the same kinematic executor with local pose tracking; the simulator (AirSim (Shah et al., 2018) over Unreal Engine) renders the four observations at each pose.

Cross-view attention. Each of the four views has a dedicated multi-head attention block whose queries are its own tokens and whose keys/values are the concatenation of all views' tokens (Eq. 3); outputs are combined with softmax view weights and projected; two such layers are stacked, with the layer output broadcast to views between layers.

Judgment heads. Learnable query tokens (10 target, 10 exploration, 5 avoidance) attend to fused features; classification heads emit confidences (sigmoid); localization heads regress 3-D offsets; the exploration head additionally regresses an information-gain estimate. At inference, target entries require confidence > 0.7, exploration entries confidence > 0.5 and $\hat{g} > 0.6$, avoidance entries confidence > 0.5.

Training protocol. All modules are trained sequentially on one A100-80GB with AdamW (batch 8, lr 10^{-4}; images resized to 224^2 from the 512^2 renders). MVG: LoRA (Hu et al., 2021) over Qwen3-VL-8B-Instruct (rank 32, α=64, dropout 0.1) with the causal-LM objective, 10 epochs ($\sim$6.5 h); L=2 CVA layers (d=1024, 8 heads). EGP: hidden 512, 4 layers, horizon 16, cross-entropy on demonstrated actions with the visual encoder jointly fine-tuned, 10 epochs ($\sim$4.7 h); training uses a flat learning rate of 10^{-4} with no curriculum or explicit state-normalization schedule. Scorer: MSE on information gain plus $0.5\times$ BCE on dead-end labels derived from demonstration returns, 5 epochs ($\sim$1.8 h). The sub-goal embedding $e(\hat{c}_k)$ is held at a fixed index, so the goal vector alone carries the directive signal. Sequence grounding runs at task boundaries under a per-episode budget of 8 calls (8.0 on average), uses a depth prior d_0=25 m in Eq. 7, and discards out-of-range frame identifiers. Recovery: window W=8, tolerance ε=1 m over a 12-step buffer, rotation trigger 10°. EBP: the deterministic rule fallback is active in this instantiation; the LLM branch and its temperature/JSON settings are reported for completeness and were not invoked in evaluation.

Demonstration collection. The expert is a goal-directed greedy heuristic (no path planner or occupancy grid): if the heading error exceeds 15°, rotate toward the goal (15° per step); otherwise mix forward 5 m (60%), altitude alignment within 3 m of the target height (15°), lateral side-steps 5 m that simulate detours (10%), and near-distance slow advance under 25 m (15°). Four 512^2 RGB views (FOV 90°, f=256) are rendered per step; no sensor noise is injected.

Latency. End-to-end step time including rendering and pose synchronization is 2.29 s per step for SE-GWM (8 visual-memory calls per episode on average) and 3.36 s for the oracle-grounded variant.

EBP prompt. The prompt contains: role and execution rules; the goal triplet; the formatted judgments (target first); top-5 entries of $\mathcal{M}$ and $\mathcal{C}$; and an instruction to output strict JSON with `sub_instructions` (each with priority and optional target object/location), `action_priority_order` fixed to [target, avoidance, exploration], and a short `reasoning`. Parse failures fall back to the rule planner.

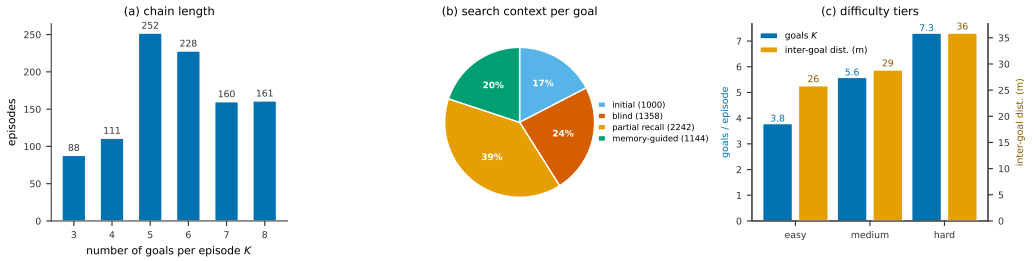

Figure 2: UAV-ON-MG statistics. (a) Chain lengths (3–8 goals; mean 5.74). (b) Search context per goal: 17.4% initial, 23.6% blind, 39.0% partial recall, 19.9% memory-guided; the last two (59%) are goals whose reward depends on retaining earlier observations. (c) Difficulty tiers vs. chain length and inter-goal distance.

Table 4: Per-goal results on UAV-ON-MG, stratified by object size tier. SR = success rate (%, within 20 m); OSR = ever entering the 25 m observation ring (1.25× the success radius); SPL = success weighted by path length (%); DCA = median closest approach (m). Higher is better for SR/OSR/SPL; lower is better for DCA.

	Small				Medium				Large				Total			
Method	SR↑	OSR↑	SPL↑	DCA↓	SR↑	OSR↑	SPL↑	DCA↓	SR↑	OSR↑	SPL↑	DCA↓	SR↑	OSR↑	SPL↑	DCA↓
AirVLN	7.9	13.9	7.9	39.5	4.5	11.0	4.5	44.0	6.4	14.6	6.4	38.5	7.0	13.4	7.0	40.1
VoLN	7.3	13.9	7.3	39.4	3.7	8.9	3.7	45.3	7.9	16.6	7.9	38.5	6.7	13.4	6.7	39.6
See-Point-Fly	20.5	35.2	20.3	**30.9**	23.2	28.9	23.2	34.2	**33.2**	**40.9**	**32.8**	30.6	23.1	34.9	22.9	31.8
TravelUAV	18.3	25.3	18.3	35.7	15.1	21.8	15.1	39.3	20.7	31.9	20.7	35.6	18.1	25.8	18.1	36.1
AeroVLA	10.3	17.0	10.3	38.8	6.4	15.8	6.4	41.6	11.8	24.1	11.8	33.3	9.8	18.0	9.8	38.8
AOA-F	20.1	28.4	20.1	34.4	14.0	21.4	13.9	40.1	23.3	35.7	23.3	31.7	19.5	28.3	19.4	35.0
UniNaVid	24.3	36.2	24.3	34.0	28.2	37.6	28.2	30.0	29.1	37.7	29.1	27.8	25.9	36.7	25.9	32.4
SE-GWM (ours)	**25.4**	**37.4**	**25.4**	33.0	**30.6**	**40.6**	**30.4**	**28.8**	31.5	40.8	31.2	**27.6**	**27.5**	**38.6**	**27.3**	**31.1**

Ongoing and planned experiments. We audited 20+ open-source systems; transferable methods and per-system reasons are tabulated in the migration audit below. In flight: eight migrated VLA/learned baselines; position-controlled per-search-context analysis; planner and grounding-frequency ablations; no-recovery and ungrounded-state reruns; confidence-gated takeover and depth from object-size priors for the generative read-out; backbone scaling of the MVG.

B UAV-ON-MG Construction Details

Overview. UAV-ON-MG contains 1,000 episodes and 5,744 goals (3–8 per episode) over 14 held-out Unreal Engine scenes: seven urban (DownTown, CityStreet, NYC, UrbanJapan, Venice, Slum, BrushifyUrban), six natural or suburban (Barnyard, CabinLake, CityPark, WesternTown, WinterTown, ModularNeighborhood), and one mountain-road scene (BrushifyRoad). Every goal carries a budget of T=30 steps and a success radius of 20 m.

Chain construction. Chain lengths follow the distribution K=3: 88, 4: 111, 5: 252, 6: 228, 7: 160, 8: 161 episodes (mean 5.74). The annotated geodesic distance to each goal has a median of 60 m (range 6–174 m), with a euclidean median of 40 m.

Train/test split. The 14 test scenes never appear in training. Demonstrations are collected in 10 disjoint training-scene variants (DownTown_train, Slum_train, Venice_train, WesternTown_train, WinterTown_train, UrbanJapan_TrainSets, BrushifyUrban_TrainSets, CabinLake_TrainSets, CityPark_TrainSets, ModularNeighborhood_TrainSets), yielding 36,145 raw steps and 20,557 steps after filtering (the label of an observation is the action taken after it; the last step of each segment and boundary samples at target switches are dropped).

Size tiers. follow the per-asset footprint annotations of the benchmark (asset length × width); the tier label is attached per instance, and 36 category names span more than one tier. Median footprints

are 0.75 m^2 for small (range 0.06–16), 6.0 m^2 for medium (0.5–16), and 23.6 m^2 for large (3.75–100) — an order of magnitude apart at the median; the ranges overlap across tiers, so tiers indicate scale rather than hard boundaries.

Search contexts. Every goal is labeled by how observable it was from the vantage points of earlier goals in the same chain; each visibility record stores its source goal, angle difference, observation distance, and quality (clear/partial/distant). The four contexts are *initial* (chain head; 1,000 goals, 17.4%), *blind* (no prior view; 1,358, 23.6%), *partial recall* (a single prior view of partial or distant quality; 2,242, 39.0%), and *memory-guided* (a clear prior view; 1,144, 19.9%, on average the deepest position in the chain at 4.49). The benchmark annotates 4,849 visibility records in total (896 clear, 2,531 partial, 1,422 distant); 59% of goals (partial recall and memory-guided) are goals whose success depends on retaining earlier observations.

Difficulty tiers. Episodes are tiered easy/medium/hard (200/500/300) with mean chain lengths 3.8/5.6/7.3 and mean geodesic distance to the goals of 60.6/61.2/63.2 m: spacing stays flat across tiers and the size mix is balanced, so difficulty is driven by chain length and the accumulated budget. The three stratifications are decoupled by design: difficulty controls the step budget, search context controls memory availability, and size tier controls perception difficulty.

Per-goal fields. Each goal record stores its instance name and description, object position, raw size string and size category, search context, the full list of historical visibility records, number of prior views, geodesic and euclidean distances, final-target flag, and target index.

C Per-Scene Results

Table 5: Per-scene per-goal SR (%) of SE-GWM on UAV-ON-MG, with goals per scene (n).

Scene	n	Per-goal SR
WinterTown	420	12.1%
Venice	395	15.4%
ModularNeighborhood	284	20.1%
BrushifyRoad	467	22.1%
DownTown	354	24.9%
CityStreet	670	25.7%
UrbanJapan	482	26.8%
NYC	579	27.8%
Slum	572	30.9%
Barnyard	437	33.4%
BrushifyUrban	400	34.8%
WesternTown	326	35.6%
CabinLake	330	38.2%
CityPark	28	39.3%

Scene variance. Per-goal success spans 12.1% (WinterTown, snow-covered and low-contrast) to 39.3% (CityPark) across the 14 scenes, tracking visual clutter and texture distinctiveness rather than scene novelty; the full per-scene table is in the appendix.

D Memory Gating Rules

The rule-based write gate of PAME (Sec. 3.5) operates as follows:

Action conf.	Info. gain	Decision
high / medium	high	write grounding cue (high / medium priority)
any	low, dead-end	write constraint (dead-end region)
any	low, no dead-end	no write
low	high	hold (low-confidence retention)

Table 6: Action-collapse diagnosis on UAV-ON-MG. "Dominant action" = most frequent action and its share over an evaluation sweep. Removing the goal signal from the policy state, or handing the task to LLM/VLM models whose prompts carry no goal-direction information, collapses the action distribution to a constant maneuver, independent of backbone modality.

Agent	Dominant action	Share	Result
EGP w/o goal signal	forward	const.	chain 0.1%, per-goal 25.9%, tf 1.46
AOA w/ text LLM (DeepSeek)	forward	81.7%	SR 0.0% (5,744 subtasks)
EGP w/ goal signal	forward	64%	SR 4.3%, tf 2.86

E Action-Collapse Diagnosis

Diagnosis: the grounding gap. Table 6 (Appendix E) isolates the failure mode that motivates SE-GWM. A policy identical to ours except that $\Delta\hat{x}_t$ is removed from its state settles into constant *forward* flight (chain SR $\approx$0.1%, per-goal 25.9%, tf 1.46), the same degenerate line that every direction-conditioned configuration shares in Table 4. The LLM method AOA (Xiao et al., 2025) degrades the same way: its text-LLM variant issues *forward* on 81.7% of steps because captions of four views contain no directional information about the goal, whereas its vision-LLM variant does explore and reaches 27.2% per-goal SR in the full run, still without a metric goal direction. We call this the **grounding gap**: general-purpose models, lacking navigation training, cannot convert egocentric images into goal-relative direction, and no prompt engineering closes the gap. SE-GWM addresses it structurally: the MVG grounds $\hat{x}_g$ in metric space and the policy state carries $\Delta\hat{x}_t$, which is why our grounded policy distributes its action mass (64% forward, the remainder spread over rotations and lateral motion) and succeeds.

F Baseline Migration Audit

We audited 20+ open-source aerial/VLN/VLA systems under our exact protocol; most are not transferable for structural reasons (mismatched task definitions, missing weights or dependencies, paid-API cores, or indoor bindings). Cleanly transferable methods span VLA models with image-to-action structure, non-VLA learned baselines, and zero-shot VLM pointing; the full audit table with per-system citations is in Appendix F.

We screened 20+ repositories across aerial VLN, UAV VLA, and VLM navigation. Transferable (same data, same protocol, retrained or zero-shot): AeroVLA (Xu et al., 2026), Uni-NaVid (Zhang et al., 2025a), AirVLN-CMA (Liu et al., 2021), VoLN (Lou et al., 2026), TRAVEL (Wang et al., 2024), and See-Point-Fly (Hu et al., 2025). Non-transferable with reasons: pipeline-bound geographic priors (HETT (Ding et al., 2026b): orthophoto+elevation+city coordinates; CityNav (Lee & Kim, 2024)), image-goal scoring paradigms (UA-NWM), missing sub-dependencies or private weights (3DG-VLN, OVER-NAV, GRaD-Nav++), empty/staged-release code (OnFly, UAV-VLN), paid-API cores (several LLM-planner stacks), and indoor Habitat bindings (IndoorUAV-Agent). All transferable baselines are being trained/evaluated under the protocol of Sec. 3.1; results will populate Table 4.