EVINAV: EVIDENCE-GROUNDED SEARCH AND ACTIVE VERIFICATION FOR TRAINING-FREE AERIAL OBJECT NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale Aerial Object Navigation (Aerial ObjectNav) requires a UAV to locate semantically relevant regions and determine whether an observed candidate is the requested target instance under partial and viewpoint-dependent observations. Existing aerial ObjectNav methods primarily strengthen spatial memory, exploration, and target grounding, while candidate-level evidence is not typically maintained and reused as a unified persistent state across search, verification, and stopping. We present EviNav, a training-free framework that treats candidate evidence as a persistent navigation state connecting evidence-grounded search, persistent target hypotheses, and active verification. EviNav constructs a shared evidence specification that distinguishes search-supporting cues from acceptance-critical evidence, accumulates candidate evidence across observations, and actively acquires complementary views when the available evidence is insufficient. Unresolved evidence guides further observation, contradictory evidence supports candidate rejection, and STOP is authorized when the available evidence provides sufficient support for the target identity under the applicable acceptance conditions. On the complete 1,000-episode, 14-scene UAV-ON test protocol, EviNav achieves a Success Rate (SR) of 27.00%, an Oracle Success Rate (OSR) of 36.60%, and a Success weighted by Path Length (SPL) of 22.98%, improving the previous best full-benchmark SR by 7.50 percentage points without task-specific navigation training. These results indicate that persistent evidence and active verification improve target-acceptance reliability while maintaining effective search and path efficiency in large-scale aerial ObjectNav.

1 INTRODUCTION

Unmanned aerial vehicles (UAVs) are increasingly deployed in applications such as emergency response, environmental monitoring, and urban inspection, where autonomously searching for task-relevant objects is a fundamental capability (Shakhatreh et al., 2019). Aerial Object Navigation (Aerial ObjectNav) formulates this target-search problem: given a high-level semantic description of a target, the UAV must autonomously locate the specified object without relying on a predefined route or fine-grained navigation instructions (Anderson et al., 2018; Xiao et al., 2025). Compared with conventional ObjectNav settings that are predominantly studied in indoor or ground-level environments (Savva et al., 2019; Kolve et al., 2017; Yadav et al., 2023; Chaplot et al., 2020; Ramakrishnan et al., 2022; Gadre et al., 2023), aerial ObjectNav requires long-horizon exploration in large-scale, partially observed outdoor 3-D scenes, where target appearance can vary substantially with distance, altitude, and viewpoint. The UAV must therefore answer two coupled questions throughout navigation: *where should it search next, and when has it collected enough evidence to stop?* Figure 1 illustrates this distinction for a goal such as "the yellow pickup truck with a rounded cab and a wooden cargo bed": coarse category cues can be visible from afar, whereas instance-discriminative attributes, contextual relations, and reliable geometric evidence often require a closer or complementary viewpoint (Bajcsy, 1988; Potthast et al., 2016; Jayaraman & Grauman, 2018).

This viewpoint-dependent observability creates an evidence-sufficiency challenge: target discovery does not by itself imply target acceptance. A distant or downward-facing observation can redirect

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: EviNav turns target-like observations into persistent hypotheses and acquires missing evidence before a verified STOP.

exploration while remaining insufficient to identify the correct instance, whereas a visible attribute conflict should prevent both termination and repeated pursuit of the same candidate. Evidence must therefore persist across observations to guide both search and stopping. UAV-ON's separate OSR and SR criteria reflect this distinction between reaching a target region and terminating successfully (Xiao et al., 2025). More generally, reliable long-horizon embodied navigation requires partial observations to be accumulated and reused throughout search (Wani et al., 2020; Huang et al., 2023; Fukushima et al., 2022; Wang et al., 2023; Li et al., 2024).

Recent aerial ObjectNav methods have substantially improved autonomous exploration, spatial representation, and target grounding in large-scale environments. AOA combines semantic goal information with egocentric observations to support long-horizon goal-directed exploration (Xiao et al., 2025). APEX further introduces dynamic spatio-semantic memory, hierarchical action decision, and open-vocabulary target grounding, while OctMem-Agent develops adaptive hierarchical 3-D memory for large-scale scene understanding and spatial exploration (Zhang et al., 2026; Zhou et al., 2026). These advances improve how the UAV discovers promising regions and localizes target-like observations. However, candidate-level evidence is not typically maintained and reused as a unified persistent state across search, verification, and stopping. Evidence that is sufficient to redirect exploration can still be insufficient for accepting a candidate, while failed, contradictory, or inconclusive interactions should in turn influence subsequent search. Without this persistent evidence state, the agent may revisit rejected candidates or regions associated with repeated unsuccessful searches, abandon promising targets before acquiring an informative view, or accept a semantically plausible instance before its defining attributes are sufficiently resolved.

We propose EviNav, a training-free framework that treats candidate evidence as a persistent navigation state linking search, verification, and stopping. A shared evidence specification distinguishes search-supporting cues from acceptance-critical evidence. Spatial and interaction history guide search while discouraging revisits to rejected candidates and repeatedly unproductive regions. Target-like observations form persistent hypotheses, whose evidence gaps guide further observation acquisition. Selected verification outcomes inform subsequent search, and terminal STOP is authorized only when the applicable acceptance conditions are satisfied. Persistent evidence thus determines where to search, what to observe next, and when to stop.

Our contributions are summarized as follows:

- We present EviNav, a training-free framework for large-scale outdoor aerial ObjectNav that treats evidence as a persistent navigation state connecting evidence-grounded search, persistent target hypotheses, and active verification.
- We introduce a shared evidence specification and persistent target hypotheses that accumulate requirement-level evidence across observations, while selected interaction outcomes are reused to guide subsequent search.

- We develop evidence-gap-driven active verification, which uses unresolved acceptance requirements to guide targeted evidence acquisition and determine whether a candidate should be further inspected, deferred, rejected, or accepted.

2 Related Work

2.1 Aerial Object Navigation

Aerial ObjectNav requires a UAV to autonomously locate a semantically specified target in large-scale outdoor environments without relying on fine-grained route instructions. UAV-ON establishes this setting using instance-level semantic goals and introduces the Aerial ObjectNav Agent (AOA) as a multimodal baseline for long-horizon goal-directed exploration (Xiao et al., 2025). Compared with conventional ground-based ObjectNav, aerial ObjectNav involves substantially larger search spaces, three-dimensional motion, and stronger variations in target scale and viewpoint.

Adjacent aerial VLN research studies language-conditioned flight in synthetic and real-world environments, including instruction following, destination grounding, and multimodal long-horizon reasoning (Liu et al., 2023; Wang et al., 2025; Gao et al., 2026; Lee et al., 2025; Shao et al., 2026). Training-free aerial navigation has further explored structured spatial memory, coarse-to-fine semantic search, and explicit candidate confirmation (Gai et al., 2026; Qi et al., 2026).

Recent methods have further advanced aerial ObjectNav through stronger spatial representation, decision making, and long-term scene understanding. APEX constructs dynamic spatio-semantic memory and combines hierarchical action decision with open-vocabulary target grounding (Zhang et al., 2026). OctMem-Agent incrementally aggregates RGB-D observations into an adaptive hierarchical 3-D memory and performs instruction-guided memory querying to support large-scale scene understanding, spatial exploration, and target localization (Zhou et al., 2026).

EviNav operates in the same aerial ObjectNav setting but focuses on how evidence accumulated throughout navigation can persist across search and target verification. Persistent evidence jointly informs where to search, which candidate to inspect, and when the current target is sufficiently supported for final acceptance.

2.2 Semantic Search and Target Verification

Open-vocabulary vision-language models and language-aligned 3-D representations support querying previously unseen concepts (Radford et al., 2021; Huang et al., 2023; Peng et al., 2023; Gu et al., 2024). Long-horizon ObjectNav extends this capability with semantic maps, frontier exploration, persistent memory, and structured scene context (Chaplot et al., 2020; Ramakrishnan et al., 2022; Yu et al., 2023; Zhou et al., 2023; Yokoyama et al., 2024; Yin et al., 2024; 2025; Kuang et al., 2024; Majumdar et al., 2022). Generative, mapless, and language-model agents further guide zero-shot navigation through predicted observations or structured high-level actions (Zhang et al., 2024b; Wang et al., 2026b; Huang et al., 2022; Chen et al., 2024).

Target-like observations may guide exploration without identifying the correct instance. Recent navigation and active-verification work therefore incorporates persistent candidate reasoning, contextual evidence, or additional viewpoints before acceptance (Lei et al., 2024; Jang & Kim, 2026; Wang et al., 2026a; Jiang, 2026; Potthast et al., 2016). Candidate-level evidence accumulation, however, remains comparatively underexplored in aerial ObjectNav, where large search spaces and substantial viewpoint changes amplify observation ambiguity.

EviNav unifies regional search and instance verification through persistent evidence. Spatial memory supports exploration, target hypotheses accumulate cross-observation evidence, and selected interaction outcomes influence subsequent decisions. Final termination is separated from any single detector or vision-language model (VLM) recommendation through an explicit target-acceptance process.



Figure 2: Overview of EviNav. Evidence-grounded search and persistent target verification form a closed loop through observation acquisition and retained interaction outcomes. Prompt excerpts and search-location markers are illustrative; overhead views are offline scene visualizations.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

We formulate Aerial ObjectNav as a sequential decision process. At the beginning of each episode, the UAV receives an instance-level semantic goal g specifying the target category, visual attributes, and optional contextual relations. At time step t, the UAV observes synchronized RGB-D images (I_t, D_t) together with its pose p_t, and selects a navigation action. The episode terminates when the UAV issues STOP or reaches the maximum navigation budget.

The key idea of EviNav is to maintain evidence about candidate target instances as a persistent navigation state, coupling regional search, candidate verification, and stopping through evidence accumulated over time. EviNav maintains

$$S_t = (M_t, H_t, R_t), \tag{1}$$

where M_t denotes the spatial-search memory, H_t contains persistent hypotheses for candidate target instances, and R_t retains selected interaction outcomes that influence subsequent search. H_t provides the candidate-level state for target verification, while M_t and R_t provide spatial and historical context.

Figure 2 summarizes the execution loop. Each episode begins with an evidence specification C_g and acceptance requirements $\mathcal{Q}_g$. RGB-D observations update world-referenced spatial memory, which supplies grounded search locations for a large language model (LLM) to select and a local planner to reach. Target-like observations initialize or update persistent hypotheses associated with the same candidate instance. Current observations and hypothesis history jointly update evidence states and gaps, guiding candidate-specific approach or viewpoint acquisition until the candidate is rejected, deferred, or sufficiently supported. Terminal STOP requires sufficient identity evidence, no explicit contradiction, and satisfaction of the applicable acceptance conditions. Support may come from a decisive current observation or accumulated hypothesis evidence, closing the search–hypothesis–verification–stopping loop.

3.2 TARGET EVIDENCE SPECIFICATION

At the beginning of each episode, EviNav constructs a shared evidence specification

$$C_g = (A_g, F_g, L_g, V_g), \tag{2}$$

which remains fixed throughout navigation. A language model parses A_g, F_g, and L_g from the semantic goal: A_g describes the target category and its semantic aliases, F_g contains instance-discriminative visual attributes, and L_g represents relevant contextual relations. V_g collects the observation conditions used for reliable verification, including the benchmark-provided target-scale prior and runtime conditions such as valid range, reliable depth, and an informative viewpoint. This representation converts the free-form semantic goal into a common structure for regional search and candidate verification.

From C_g, EviNav derives a set of candidate-level acceptance requirements $\mathcal{Q}_g = \{r_1, \ldots, r_m\}$, where each r_j corresponds to a target property or verification condition whose evidence is tracked during candidate evaluation. $\mathcal{Q}_g$ provides a common reference for organizing evidence in persistent hypotheses, identifying unresolved evidence, and guiding subsequent active verification. Candidate acceptance is determined by whether the available evidence provides sufficient support for the target identity and satisfies the applicable semantic, contextual, and observation conditions.

EviNav further distinguishes *search-supporting cues* from *acceptance-critical evidence*. Generic contextual cues increase the relevance of search regions. Identity-bearing attributes, explicitly required relations, and applicable observation conditions contribute to candidate acceptance. This separation allows C_g to support both regional search and instance-level verification. A concrete example is provided in Appendix C.1.

3.3 EVIDENCE-GROUNDED SEARCH

During regional search, EviNav uses the search-supporting cues in C_g together with accumulated spatial evidence to determine where to search next. At each time step, RGB-D observations are projected using the current pose p_t into a common world-referenced representation and continuously update the spatial-search memory M_t. The memory preserves local geometric and semantic information together with visitation history and longer-range topological exploration context, allowing spatial evidence acquired at earlier steps to remain available for subsequent decisions.

Given the current memory state, EviNav aggregates multiple search sources into a set of world-referenced, spatially grounded search locations $\mathcal{C}_t = \{c_1, \ldots, c_K\}$. These search locations include regions supported by target or contextual evidence, insufficiently explored regions, and longer-range priors derived from semantic memory and topological exploration history. The locations are filtered and prioritized according to reachability, local safety, repeated visitation, and retained search or verification outcomes, producing a compact set of physically executable search alternatives. The resulting set of search locations is therefore grounded in the current spatial and semantic evidence.

The navigation LLM selects the next high-level search location $c_t^\star$ from $\mathcal{C}_t$, conditioned on the semantic goal and a compact summary of the current spatial and navigation context. The selected world location is passed to a local planner for executable motion. The resulting RGB-D observation updates the navigation state and initiates the next search or verification cycle. A STOP proposal from the navigation LLM remains provisional, with terminal stopping authorized through the target-acceptance process described in Sec. 3.5. Additional details of the spatial memory, RGB-D geometry, and navigation interface are provided in Appendices B.1, B.3, and C.2.

3.4 PERSISTENT TARGET HYPOTHESES

A target-like observation initializes a persistent hypothesis h_i, or is associated with an existing hypothesis when the spatial location, semantic identity, and appearance evidence are sufficiently consistent. Observations acquired from different positions and viewpoints can therefore be associated with the same candidate target instance, allowing evidence to accumulate across multiple observations. Each hypothesis maintains a persistent candidate location together with its associated observation, visual, and verification history for subsequent candidate-level reasoning.

For each requirement $r_j \in \mathcal{Q}_g$, EviNav summarizes the current observation together with the evidence accumulated in h_i using a requirement-level evidence state

$$z_j(h_i) \in \{\text{supported}, \text{unknown}, \text{contradicted}\}. \tag{3}$$

Here, *supported* denotes visible evidence supporting the corresponding target property or verification condition. *Unknown* indicates that the relevant evidence remains unresolved because of factors such as distance, occlusion, insufficient target scale, or viewpoint limitations. *Contradicted* denotes a directly observable target-specific conflict. The evidence states integrate semantic verifier outputs, geometric evidence, and the accumulated hypothesis history. As new observations are associated with h_i, the corresponding states are updated accordingly.

We define the unresolved evidence gap of a hypothesis as

$$G(h_i) = \{r_j \in \mathcal{Q}_g \mid z_j(h_i) = \text{unknown}\}. \tag{4}$$

$G(h_i)$ summarizes the discriminative evidence that remains unresolved and provides the basis for selecting additional observations. Missing or unobservable attributes remain *unknown*, leaving them open for further evidence acquisition. Directly observed conflicts are marked as *contradicted* and can support candidate rejection. Sec. 3.5 describes how the evidence gap guides active verification and subsequent candidate decisions.

3.5 Evidence-Gap-Driven Active Verification

Once a target hypothesis h_i enters verification, EviNav uses its unresolved evidence gap $G(h_i)$ to acquire more discriminative observations. Figure 3 summarizes this evidence-to-action loop. Evidence limited by target scale or range motivates approaching the candidate, while view-dependent attributes or contextual relations motivate a complementary viewpoint. Verification can be deferred when the current observation conditions provide insufficient information, and resumed when a closer or more informative observation becomes available. Each newly acquired observation is associated with the persistent hypothesis and updates its evidence state, forming a candidate-level active evidence acquisition loop.

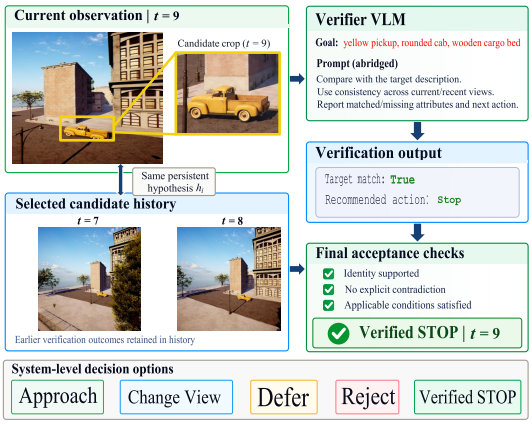


Figure 3: Target verification in the NYC truck example. Current and historical observations support a stopping recommendation, with terminal STOP subject to final acceptance checks. The prompt is abridged; the bottom row lists alternative system-level decisions.

The updated evidence determines the subsequent candidate decision. Directly observable conflicts can lead to rejection. Incomplete evidence triggers further observation acquisition or deferral. Terminal STOP is authorized when the available evidence provides sufficient support for the target identity, no explicit contradiction is present, and the applicable acceptance conditions are satisfied. Evidence sufficiency can be established by a decisive current observation that clearly resolves the target identity or by consistent evidence accumulated across the persistent hypothesis when the current observation remains incomplete. The acceptance decision therefore combines available identity evidence with the verification conditions relevant to the candidate.

Verification outcomes further influence subsequent search. Rejected candidates and selected verification outcomes are retained in R_t to reduce repeated search around previously disproven candidates or regions. Deferred hypotheses preserve their accumulated evidence for later verification, while other locally exhausted cases return to regional search. Stronger evidence emerging in a previously suppressed region can initiate a new hypothesis. These feedback mechanisms couple regional exploration with candidate-level verification throughout navigation. The evidence-gap-to-verification mapping is summarized in Appendix B.2, Table 8; additional details of hypothesis maintenance and the target-verification interface are provided in Appendices B.2 and C.3.

4 EXPERIMENTS

We evaluate EviNav around three research questions: **RQ1 (effectiveness and generalization),** whether gains hold across the complete benchmark, target scales, and seen/unseen environments (Tables 1–2); **RQ2 (component contribution and model sensitivity),** how each main component and navigation-LLM choice affects performance (Tables 3 and 4); and **RQ3 (successful acceptance behavior),** how persistent hypotheses support final STOP decisions (Table 5 and Figure 4).

4.1 EXPERIMENTAL SETUP

Benchmark. We evaluate EviNav on UAV-ON, a large-scale aerial ObjectNav benchmark with 14 high-fidelity outdoor environments (Xiao et al., 2025). Our primary result covers its complete 1,000-episode test set under a maximum budget of 100 navigation steps per episode. Benchmark annotations identify 629 small, 210 medium, and 161 large target instances.

We also identified reproducible environment issues in four scenes. We retain all 14 scenes for direct comparison and report the remaining 10-scene subset only as a diagnostic; details appear in Appendix A.

Evaluation metrics. Following UAV-ON (Xiao et al., 2025), we report Success Rate (SR), Oracle Success Rate (OSR), and Success weighted by Path Length (SPL). SR requires a successful stop within 20 m of the target; OSR records whether this neighborhood is reached at any time. The official evaluator invalidates both after a collision. SPL weights successful episodes by path length, measuring path efficiency; higher is better for all metrics.

Baselines and training setting. We compare EviNav with the UAV-ON baselines Random, CLIP-H, and AOA-F (Xiao et al., 2025), as well as aerial navigation methods including AerialVLN (Liu et al., 2023), NaVid (Zhang et al., 2024a), OpenFly (Gao et al., 2026), APEX (Zhang et al., 2026), and OctMem-Agent (Zhou et al., 2026). Baseline values follow the OctMem-Agent comparison (Zhou et al., 2026), with each method cited to its original paper; APEX† follows its reported evaluation setting (Zhang et al., 2026). TF denotes no task-specific UAV-ON navigation-policy training. EviNav uses neither such optimization nor benchmark trajectory supervision; model details appear in Appendix A.2.

Table 1: Comparison on UAV-ON across the complete test set and different target scales. All values are percentages.

Method	TF	Overall			Small			Medium			Large		
		SR↑	OSR↑	SPL↑	SR↑	OSR↑	SPL↑	SR↑	OSR↑	SPL↑	SR↑	OSR↑	SPL↑
Random (Xiao et al., 2025)	✓	3.70	8.00	2.66	4.14	7.80	2.80	3.33	8.10	3.05	2.48	8.07	1.62
CLIP-H (Xiao et al., 2025)	✓	6.20	11.90	4.15	2.86	8.43	1.51	10.95	16.67	7.17	13.04	19.25	10.53
AOA-F (Xiao et al., 2025)	✓	7.30	17.50	4.06	4.45	16.38	1.61	10.48	17.62	6.36	14.29	21.74	10.66
AerialVLN (Liu et al., 2023)	×	3.90	19.30	3.72	4.13	19.87	4.01	4.76	17.62	4.90	1.86	19.25	1.86
NaVid (Zhang et al., 2024a)	×	11.50	29.10	6.44	10.02	29.55	4.41	11.43	26.44	5.95	17.39	30.73	15.03
OpenFly (Gao et al., 2026)	×	12.00	25.90	6.09	12.40	26.23	6.35	10.00	20.48	4.19	13.04	31.65	7.56
APEX† (Zhang et al., 2026)	×	13.33	20.00	10.14	–	–	–	–	–	–	–	–	–
OctMem-Agent (Zhou et al., 2026)	×	19.50	29.30	6.37	18.91	29.57	6.03	18.09	26.66	6.27	23.60	31.67	7.80
EviNav (ours)	✓	**27.00**	**36.60**	**22.98**	**20.35**	**31.96**	**16.12**	**38.10**	**45.71**	**33.84**	**38.51**	**42.86**	**35.64**

4.2 MAIN RESULTS

EviNav establishes the strongest overall performance among the compared methods without task-specific navigation training. Table 1 reports 27.00% SR, 36.60% OSR, and 22.98% SPL, improving over OctMem-Agent by 7.50 SR points and 16.61 SPL points. It also converts 73.8% of reported oracle successes into successful termination, compared with 66.6% for OctMem-Agent. The SPL margin indicates that the gain is accompanied by substantially stronger path efficiency on successful episodes.

Performance across target scales. Gains extend across all target scales, with the largest SR improvement on medium targets (18.09% to 38.10% over OctMem-Agent). Despite improved SPL, small targets remain most challenging at 20.35% SR. They constitute 62.9% of the test set, making long-range acquisition of discriminative evidence a primary bottleneck.

Effect of benchmark scene issues. Performance increases on the diagnostic 10-scene subset. After excluding the four affected scenes, EviNav reaches 32.74% SR, 39.02% OSR, and 27.73% SPL; complete scene diagnostics are given in Appendix A.

Seen and unseen environments. EviNav maintains its advantage across both splits. Table 2 reports 29.14%/25.72% SR and 24.54%/22.05% SPL on the 374 and 626 episodes from seen and unseen environments, respectively; the corresponding OSR values are nearly identical. Baseline split values follow OctMem-Agent (Zhou et al., 2026).

Table 2: Performance on benchmark-seen and unseen environments.

Method	Seen			Unseen		
	SR↑	OSR↑	SPL↑	SR↑	OSR↑	SPL↑
AerialVLN	4.55	16.84	4.24	3.51	20.77	3.41
NaVid	14.17	30.08	7.19	10.70	28.51	5.99
OpenFly	12.29	26.20	6.83	11.82	25.71	5.64
OctMem-Agent	22.72	30.48	8.40	17.57	28.59	5.15
EviNav (ours)	**29.14**	**36.63**	**24.54**	**25.72**	**36.58**	**22.05**

4.3 Ablation and Robustness

We ablate the three main components on the same fixed 200 UAV-ON episodes while holding perception, planning, action space, and evaluation fixed. The search variant uses generic spatial exploration; the persistent-evidence variant removes cross-observation evidence and interaction outcomes; and the verification variant disables targeted approach and viewpoint acquisition after hypothesis formation.

Table 3: Module-level ablation on the same fixed 200 UAV-ON episodes.

Variant	SR ↑	OSR ↑	SPL ↑
w/o Evidence-Grounded Search	20.00	39.00	18.13
w/o Persistent Evidence	24.50	41.50	22.27
w/o Active Verification	11.00	21.50	9.92
Full EviNav	**27.50**	**43.00**	**24.70**

Active verification contributes most strongly: its removal reduces SR from 27.50% to 11.00%, OSR from 43.00% to 21.50%, and SPL from 24.70% to 9.92% (Table 3); removing search or persistent evidence also degrades performance. The OSR decrease also reflects the navigation role of active verification, since this ablation removes candidate-specific approach and viewpoint acquisition.

Navigation-LLM sensitivity. On a separate fixed 200-episode subset, we replace only the navigation LLM; the verifier and remaining components stay fixed. Across navigation LLMs, SR ranges from 19.5% to 25.0%, with `deepseek-v4-flash` highest on this subset. The `gpt-4.1-mini` row evaluates the primary configuration on this subset, which differs from the ablation subset.

4.4 Acceptance and Qualitative Analysis

Successful acceptance pathways. Of the 270 successful episodes, 187 (69.3%) use decisive-current-observation acceptance, while 83 (30.7%) use persistent-hypothesis-assisted acceptance. Among the 187 decisive-current-observation cases, 156 contain an earlier incomplete verification state, showing that active evidence acquisition often precedes the final decisive observation.

Table 4: Navigation-LLM sensitivity on a fixed 200-episode UAV-ON subset.

Main navigation LLM	SR ↑	OSR ↑	SPL ↑
`gpt-4.1-mini`	23.00	37.50	20.86
`qwen3.6-plus`	19.50	34.00	17.40
`deepseek-v4-flash`	25.00	39.00	22.06

Table 5: Successful acceptance pathways over all 270 successful episodes.

Acceptance pathway	Episodes	Share
Decisive current observation	187	69.3%
Persistent-hypothesis-assisted	83	30.7%
Total	270	100.0%

Qualitative examples. Figure 4 complements the pathway statistics with representative UAV-ON trajectories showing candidate formation, approach, viewpoint acquisition, and verified stopping.

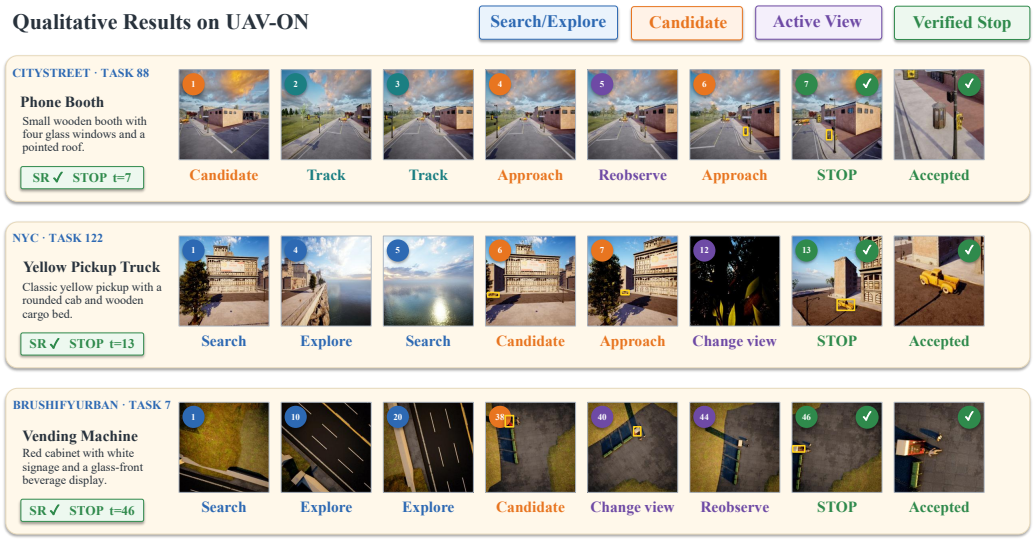


Figure 4: Candidate formation, active evidence acquisition, and verified stopping in representative UAV-ON trajectories.

5 Conclusion and Limitations

We presented EviNav, a training-free framework for large-scale aerial ObjectNav that connects evidence-grounded search, persistent target hypotheses, and active verification within a closed navigation loop. EviNav maintains candidate evidence as a persistent navigation state that guides both where to search and when a detected candidate has sufficient evidence for acceptance. On the complete 1,000-episode, 14-scene UAV-ON evaluation, EviNav achieves 27.00% SR, 36.60% OSR, and 22.98% SPL, improving SR by 7.50 percentage points over OctMem-Agent without task-specific navigation training. EviNav also converts 73.8% of oracle successes into successful terminations, indicating a stronger ability to translate target-neighborhood reachability into successful task completion.

Several limitations remain. EviNav is currently evaluated in simulation, and its robustness to real-world sensing noise, localization error, dynamic obstacles, and environmental disturbances remains to be validated. Reliable evidence acquisition also remains challenging when targets are distant, occupy only a small portion of the image, or are observed under degraded visual conditions. In such cases, target hypotheses can remain unresolved even after the UAV reaches a relevant region. Future work will investigate more robust evidence acquisition and reasoning under uncertain or conflicting observations and validate the complete search–verification loop on physical UAV platforms.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, software development, and debugging. All AI-assisted code was reviewed and tested by the authors. The evaluated system itself also invoked language and vision-language models during the experiments, as described in the paper. The authors take responsibility for the final content, implementation, and reported results.

REPRODUCIBILITY STATEMENT

The evaluation protocol is described in Sec. 4 and Appendix A, with foundation-model configuration in Appendix A.2. Additional implementation details are provided in Appendices B and C, and the complete prompts are reported in Appendix D. Code will be included in the supplementary material.

REFERENCES

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*, 2018. URL https://arxiv.org/abs/1807.06757.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. doi: 10.1109/5.5968.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4247–4258, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2c75cf2681788adaca63aa95ae028b22-Abstract.html.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. MapGPT: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9796–9810, 2024. doi: 10.18653/v1/2024.acl-long.529. URL https://aclanthology.org/2024.acl-long.529/.

Rui Fukushima, Kei Ota, Asako Kanezaki, Yoko Sasaki, and Yusuke Yoshiyasu. Object memory transformer for object goal navigation. In *2022 IEEE International Conference on Robotics and Automation*, pp. 11288–11294, 2022. doi: 10.1109/ICRA46639.2022.9812027. URL https://doi.org/10.1109/ICRA46639.2022.9812027.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. CoWs on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23171–23181, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gadre_CoWs_on_Pasture_Baselines_and_Benchmarks_for_Language-Driven_Zero-Shot_Object_CVPR_2023_paper.html.

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. *arXiv preprint arXiv:2602.00708*, 2026. URL https://arxiv.org/abs/2602.00708.

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, Dong Wang, Xuelong Li, Zhigang Wang, and Bin Zhao. OpenFly: A comprehensive platform for aerial vision-language navigation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OKm3w71ymP.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation*, pp. 5021–5028, 2024. doi: 10.1109/ICRA57147.2024.10610243.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation*, pp. 10608–10615, 2023. doi: 10.1109/ICRA48891.2023.10160969. URL https://ieeexplore.ieee.org/document/10160969.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9118–9147, 2022. URL https://proceedings.mlr.press/v162/huang22a.html.

Won Shik Jang and Ue-Hwan Kim. Context-Nav: Context-driven exploration and viewpoint-aware 3d spatial reasoning for instance navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9626–9636, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jang_Context-Nav_Context-Driven_Exploration_and_Viewpoint-Aware_3D_Spatial_Reasoning_for_Instance_CVPR_2026_paper.html.

Dinesh Jayaraman and Kristen Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1238–1247, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jayaraman_Learning_to_Look_CVPR_2018_paper.html.

Yuhang Jiang. PInVerify: An offline embodied benchmark for active instance verification. *arXiv preprint arXiv:2605.30639*, 2026. URL https://arxiv.org/abs/2605.30639.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3d environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. URL https://arxiv.org/abs/1712.05474.

Yuxuan Kuang, Hai Lin, and Meng Jiang. OpenFMNav: Towards open-set zero-shot object navigation via vision-language foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 338–351, 2024. doi: 10.18653/v1/2024.findings-naacl.24. URL https://aclanthology.org/2024.findings-naacl.24/.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. CityNav: A large-scale dataset for real-world aerial navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5912–5922, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lee_CityNav_A_Large-Scale_Dataset_for_Real-World_Aerial_Navigation_ICCV_2025_paper.html.

Xiaohan Lei, Min Wang, Wengang Zhou, Li Li, and Houqiang Li. Instance-aware exploration-verification-exploitation for instance imagegoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16329–16339, 2024. doi: 10.1109/CVPR52733.2024.01545. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lei_Instance-aware_Exploration-Verification-Exploitation_for_Instance_ImageGoal_Navigation_CVPR_2024_paper.html.

Hongxin Li, Zeyu Wang, Xu Yang, Yuran Yang, Shuqi Mei, and Zhaoxiang Zhang. MemoNav: Working memory model for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17913–17922, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MemoNav_Working_Memory_Model_for_Visual_Navigation_CVPR_2024_paper.html.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. AerialVLN: Vision-and-language navigation for UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15348, 2023. doi: 10.1109/ICCV51070.2023.01411. URL https://openaccess.thecvf.com/content/ICCV2023/html/Liu_AerialVLN_Vision-and-Language_Navigation_for_UAVs_ICCV_2023_paper.html.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32340–32352, 2022. doi: 10.52202/068431-2343.

Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Peng_OpenScene_3D_Scene_Understanding_With_Open_Vocabularies_CVPR_2023_paper.html.

Christian Potthast, Andreas Breitenmoser, Fei Sha, and Gaurav S. Sukhatme. Active multi-view object recognition: A unifying view on online feature selection and view planning. *Robotics and Autonomous Systems*, 84:31–47, 2016. doi: 10.1016/j.robot.2016.06.013.

Yu Qi, Hongyu Li, Shaofei Huang, Tianrui Hui, Yaxiong Wang, Lechao Cheng, Zhun Zhong, Si Liu, and Meng Wang. Parse, search, and confirmation: Training-free aerial vision-and-dialog navigation with chain-of-thought reasoning and structured spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23859–23868, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Qi_Parse_Search_and_Confirmation_Training-Free_Aerial_Vision-and-Dialog_Navigation_with_Chain-of-Thought_CVPR_2026_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18890–18900, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Ramakrishnan_PONI_Potential_Functions_for_ObjectGoal_Navigation_With_Interaction-Free_Learning_CVPR_2022_paper.html.

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*, pp. 9339–9347, 2019. URL

https://openaccess.thecvf.com/content_ICCV_2019/html/Savva_Habitat_A_Platform_for_Embodied_AI_Research_ICCV_2019_paper.html.

Hazim Shakhatreh, Ahmad H. Sawalmeh, Ala Al-Fuqaha, Zuochao Dou, Eyad Almaita, Issa Khalil, Noor Shamsiah Othman, Abdallah Khreishah, and Mohsen Guizani. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. *IEEE Access*, 7:48572–48634, 2019. doi: 10.1109/ACCESS.2019.2909530.

Dian Shao, Zhengzheng Xu, Peiyang Wang, Like Liu, Yule Wang, Jieqi Shi, and Jing Huo. FineCog-Nav: Integrating fine-grained cognitive modules for zero-shot multimodal UAV navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 1325–1334, 2026. URL https://arxiv.org/abs/2604.16298.

Haosen Wang, Zhenyang Li, Yinqiang Zhang, Zongqi He, Lutao Jiang, Kai Li, Yizhou Zhao, Liaoyuan Fan, Wenjian Hou, Tingbang Liang, Yibin Wen, and Defeng Gu. ConsistNav: Closing the action consistency gap in zero-shot object navigation with semantic executive control. *arXiv preprint arXiv:2605.09869*, 2026a. URL https://arxiv.org/abs/2605.09869.

Teng Wang, Xinxin Zhao, Wenzhe Cai, and Changyin Sun. ImagineNav++: Prompting vision-language models as embodied navigator through scene imagination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b. doi: 10.1109/TPAMI.2026.3688692. URL https://arxiv.org/abs/2512.17435.

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. URL https://openreview.net/forum?id=rUvCIvI4eB.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. GridMM: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15625–15636, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_GridMM_Grid_Memory_Map_for_Vision-and-Language_Navigation_ICCV_2023_paper.html.

Saim Wani, Shivansh Patel, Unnat Jain, Angel X. Chang, and Manolis Savva. MultiON: Benchmarking semantic map memory using multi-object navigation. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6e01383fd96a17ae51cc3e15447e7533-Abstract.html.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13023–13029, 2025. doi: 10.1145/3746027.3758251. URL https://doi.org/10.1145/3746027.3758251.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4927–4936, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yadav_Habitat-Matterport_3D_Semantics_Dataset_CVPR_2023_paper.html.

Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. SG-Nav: Online 3d scene graph prompting for LLM-based zero-shot object navigation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. UniGoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19057–19066, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_UniGoal_Towards_Universal_Zero-shot_Goal-oriented_Navigation_CVPR_2025_paper.html.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation*, 2024. URL https://ieeexplore.ieee.org/document/10610712.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3MVN: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023. doi: 10.1109/IROS55552.2023.10342512. URL https://doi.org/10.1109/IROS55552.2023.10342512.

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. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_APEX_A_Decoupled_Memory-based_Explorer_for_Asynchronous_Aerial_Object_Goal_CVPR_2026_paper.html.

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*, 2024a. URL https://www.roboticsproceedings.org/rss20/p079.pdf.

Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine before go: Self-supervised generative map for object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16414–16425, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Imagine_Before_Go_Self-Supervised_Generative_Map_for_Object_Goal_Navigation_CVPR_2024_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_Memory-Augmented_Scene_Understanding_and_Exploration_for_Open-World_Aerial_Object-Goal_Navigation_CVPR_2026_paper.html.

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 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42829–42842. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhou23r.html.

Xin Zhou, Zhepei Wang, Hongkai Ye, Chao Xu, and Fei Gao. EGO-Planner: An ESDF-free gradient-based local planner for quadrotors. *IEEE Robotics and Automation Letters*, 6(2):478–485, 2021. doi: 10.1109/LRA.2020.3047728.

A UAV-ON Evaluation Details

A.1 Evaluation Protocol

We evaluate EviNav on the complete UAV-ON test set, which contains 1,000 episodes across 14 outdoor scenes. At each control cycle, the agent receives four synchronized RGB-D observations from the front-, left-, right-, and downward-facing cameras, together with ego pose and a semantic target instruction. RGB observations are captured at 512×512 resolution, while depth observations are provided at 192×192 resolution. Pose is expressed relative to the episode start and does not provide the target location. Following UAV-ON, we report Success Rate (SR), Oracle Success Rate (OSR), and Success weighted by Path Length (SPL). Each episode is evaluated with a maximum budget of 100 navigation steps.

All geometric reasoning is based on the benchmark-provided RGB-D observations and ego pose; no privileged target pose or simulator ground-truth geometry is used.

A.2 Foundation Models and LLM Sensitivity

The target-structuring and navigation-decision stages use `gpt-4.1-mini`, while the target-verification stage uses `gpt-5.6-luna`. These are the primary model identifiers used in the reported evaluation. The models were accessed through the OpenAI-compatible API endpoint configured for the experiments during August 2026. The endpoint exposed these configured identifiers but did not expose provider-specific snapshot identifiers. No explicit temperature or decoding seed was passed in the language-model calls; the endpoint defaults were used.

Navigation-LLM sensitivity is evaluated on another fixed 200-episode subset spanning all 14 scenes. Only the navigation LLM is replaced; the target verifier and all remaining components are held fixed. Results are reported in Table 4 in the main paper.

A.3 Scene Diagnostics

We observed four scene-level conditions relevant to interpreting the benchmark results. CityPark and BrushifyRoad exhibit constant near-clip metric depth, DownTown contains visual assets that are inconsistent with the available depth or collision geometry, and Slum contains abnormal fog that substantially degrades visual observations. We report these conditions as benchmark diagnostics and retain the complete 14-scene evaluation as the primary result. The corresponding scene-level observations are summarized in Table 6.

Table 6: Scene-level diagnostics in the complete UAV-ON test set.

Scene	Episodes	Observed condition
CityPark	50	Constant near-clip metric depth
BrushifyRoad	70	Constant near-clip metric depth
DownTown	49	Visual–geometry inconsistency
Slum	98	Abnormal fog and degraded visibility

To examine sensitivity to these observed conditions, we additionally report results on the remaining 10 scenes after excluding the four scenes listed above. This analysis is supplementary and does not replace the complete 14-scene benchmark result.

Table 7: Sensitivity analysis on the complete UAV-ON test set and the remaining 10 scenes after excluding the four diagnostic scenes.

Evaluation setting	Episodes	SR (%)	OSR (%)	SPL (%)
Complete 14 scenes	1,000	27.00	36.60	22.98
Remaining 10 scenes	733	32.74	39.02	27.73

B Additional Method Details

B.1 Spatial and Outcome-Aware Memory

We further instantiate the coupled navigation state defined in Eq. (1), where M_t denotes spatial-search memory, H_t contains persistent target hypotheses, and R_t summarizes interaction outcomes. EviNav maintains world-referenced spatial records that jointly store semantic evidence, geometric support, visitation history, and the outcomes of prior interactions. Although M_t and R_t are conceptually separated in the framework, their corresponding observation and outcome fields are stored in shared spatial records. This organization allows evidence and the consequences of acting on that evidence to be retrieved together.

Outcome information influences subsequent retrieval. Repeated unsuccessful searches reduce the priority of previously explored regions, whereas newly observed semantic or geometric evidence can relax earlier penalties. A complementary topological memory stores visited regions, transitions, and coarse scene summaries for recovery and backtracking. The spatial memory provides region-level semantic and outcome-conditioned retrieval, while the topological memory provides longer-range exploration context.

Search locations are aggregated from target-like observations, context-supported observations, memory retrievals, unexplored regions, and topological search priors. Reachability, local safety, duplicate suppression, and rejected-candidate memory are applied before a compact set of grounded search locations is passed to the navigation model. The language model therefore selects among spatially grounded alternatives rather than proposing an unconstrained low-level trajectory.

B.2 Persistent Hypotheses and Active Verification

Each target-like observation is associated with a persistent hypothesis h_i. The hypothesis accumulates evidence across views, including a persistent relative-world position, visual appearance, detector and depth observations, camera history, semantic verifier outputs, and multiview geometric consistency. A hypothesis may be `weak`, `inspecting`, `confirmed`, `rejected`, or `deferred`. New observations are associated with an existing non-rejected hypothesis when the estimated 3-D locations are within 6.0 m and either lie within a 2.25 m hard spatial radius or have compatible category labels and an appearance similarity of at least 0.55. Appearance similarity is the cosine similarity between compact normalized color and intensity histograms. Otherwise, the observation initializes a new hypothesis.

Following Eqs. (3)–(4), each hypothesis maintains requirement-level evidence states and the corresponding unresolved evidence gap. ***Supported*** denotes visible evidence for a requirement; ***unknown*** denotes missing, occluded, too small, or viewpoint-limited evidence; and ***contradicted*** requires a directly visible conflict. Thus, `unknown` is the only formal unresolved state. A recommendation such as `next_view_action=reject` is an action proposal and is not itself evidence of contradiction; contradiction is established only from explicit visible conflict evidence.

The evidence gap $G(h_i)$ determines what observation should be acquired next. The mapping used by the inspection controller is summarized in Table 8.

Table 8: Evidence-gap-driven target-verification behavior.

Missing or observed evidence	Verification behavior
Insufficient target scale or range	Approach
View-dependent identity or context	Change viewpoint
Incomplete positive evidence	Defer and acquire evidence
Directly visible contradiction	Reject hypothesis
Sufficient acceptance evidence	Evaluate final acceptance

Detection is not acceptance: final STOP remains subject to the acceptance criterion in Sec. 3.5, which requires sufficient support for the target identity, no explicit contradiction, and the applicable acceptance conditions. Support may come from a decisive current observation or consistent evi-

dence accumulated in the persistent hypothesis. When evidence remains incomplete, the inspection controller directs the UAV to acquire an additional complementary view.

B.3 RGB-D Geometry and Navigation

Open-vocabulary visual candidates and their pixel masks are produced by SAM 3 from short concept prompts derived from the shared evidence specification (Carion et al., 2026). The masks are associated with the corresponding depth observations before being projected into the spatial state.

Valid depth observations are back-projected using the calibrated camera model and transformed from the camera frame to the UAV body frame and subsequently to the relative world frame. For a camera-frame point p_c, the body-frame point is

$$p_b = R_{bc}p_c + t_{bc}. \tag{5}$$

The available views are aggregated into a local 3-D point cloud. Ego pose then places this geometry in the relative spatial frame shared by semantic mapping and memory. The reconstructed geometry supports the 2.5-D spatial state, candidate localization, geometric target consistency, and collision-aware local planning.

Invalid or clearly unreliable RGB-D measurements are excluded from geometric fusion. No target annotation, simulator mesh, or world-coordinate object pose is used to construct the spatial state. The local planner receives candidate waypoints and depth-derived geometry and returns executable collision-aware motion through EGO-Planner (Zhou et al., 2021); it does not determine semantic target acceptance.

C Evidence Specification and Decision Interfaces

C.1 Evidence Specification Example

The semantic goal is converted into a shared evidence specification used by both search and verification. In the implementation, A_g is instantiated by the primary subject and target aliases, F_g by the extracted target features, and L_g by the context relations. V_g is derived from the benchmark-provided target-scale descriptor and the range, depth-validity, and viewpoint conditions used by the verifier. Table 9 instantiates this interface for the yellow pickup truck in Figure 4. Aliases and generic contextual entities primarily support open-vocabulary search, while identity-bearing attributes, explicitly required relations, and applicable observation conditions inform candidate acceptance.

Table 9: Evidence specification for the yellow pickup truck example.

Component	Example
A_g: category and aliases	pickup truck; truck; yellow truck
F_g: instance-discriminative attributes	yellow body; pickup form; rounded cab; wooden cargo bed
L_g: context	road; nearby buildings; vegetation; no required relation in the original goal
V_g: observation conditions	sufficient target scale; valid range; reliable depth; informative viewpoint

Search-supporting context helps determine where to look, but it is not automatically converted into an acceptance requirement. Only contextual relations explicitly required by the original goal are treated as acceptance-critical evidence.

C.2 Navigation Decision Interface

The navigation model receives the target instruction, ego pose and heading, a local spatial summary, spatially grounded search locations, a compact memory summary, and recent actions. The shortened template below summarizes the operative input and output contract; the complete static prompt is provided in Appendix D.2.

```
TARGET: {object_name}; {description}
EGO STATE: {relative_pose}; heading={heading}
LOCAL SPATIAL SUMMARY: {grid_and_geometry_summary}
GROUNDED SEARCH LOCATIONS: {grounded_candidates}
MEMORY: {spatial_topological_and_outcome_summary}
RECENT ACTIONS: {action_history}

Choose exactly one grounded high-level action:
{"action":"explore","grid_x":X,"grid_y":Y,
 "step_size":S,"done":false}
or
{"action":"stop","done":true}
Return JSON only.
```

An `explore` response identifies a grounded location in the local spatial representation and is converted into motion by the downstream planner. A navigation-model `stop` response is provisional: it is not executed unless a persistent hypothesis independently satisfies the final acceptance criterion.

C.3 TARGET-VERIFICATION INTERFACE

The target verifier receives a marked full view, local context, object and detail crops, the target evidence specification, the current distance and camera, and a compact history of the same persistent hypothesis. When multiple views are available, they are presented as observations of one candidate. A shortened output schema is shown below; the complete static verifier prompt, including its field descriptions and conservative conflict rules, is provided in Appendix D.3.

```
{
 "category_match": false,
 "description_match": "yes|partial|no|uncertain",
 "matched_attributes": [],
 "missing_attributes": [],
 "context_match": "yes|partial|no|uncertain",
 "context_contradiction": false,
 "hard_attribute_conflict": false,
 "next_view_action":
   "stop|move_closer|change_viewpoint|reject"
}
```

Missing or occluded evidence requests another observation rather than rejection. Search-only context cannot establish identity, and absent context is not a contradiction unless the original goal explicitly requires the relation. The verifier's `next_view_action` provides an action recommendation to the inspection controller; it is not itself treated as evidence and does not directly authorize termination. Final STOP remains subject to the acceptance criterion in Sec. 3.5.

The reported confidence and `attribute_match_ratio` are verifier support signals, and `next_view_action` is an observation-action recommendation. No single output field or prompt-level threshold independently authorizes STOP; the controller evaluates the complete acceptance logic in Sec. 3.5.

At the method level, verifier evidence associated with the attributes and relations in $\mathcal{Q}_g$, together with geometric validity and hypothesis history, is summarized by the requirement-level states $z_j(h_i)$. These states provide a unified abstraction of the available evidence rather than a one-to-one copy of individual verifier output fields.

D STATIC PROMPT TEMPLATES

For reproducibility, this section reports the complete static system prompts and user templates used by the three language-mediated stages of EviNav. At runtime, brace-delimited fields are populated with the current task and navigation state. Images supplied to the target verifier are transmitted as separate multimodal inputs and are therefore not reproduced in the text template. Decorative Unicode characters in the implementation are normalized to ASCII in the listings below; the operative instructions and output fields are unchanged.

D.1 TARGET-STRUCTURING PROMPT

This prompt converts the language goal into the shared evidence specification used to form open-vocabulary perception queries and acceptance requirements. The scene vocabulary referenced by the prompt is fixed for all episodes and is independent of the target category.

System prompt.

```
semantic_prompt_system = """You extract a compact open-vocabulary perception plan for a UAV
     object-navigation task.

The perception backend is SAM3. Your job is NOT to navigate. Your job is to choose a small
     set of visual concepts that are likely useful to detect from RGB images.

Return ONLY valid JSON with these keys:
{
  "target_prompts": ["..."],
  "context_prompts": ["..."],
  "primary_subject": "...",
  "target_features": ["..."],
  "context_relations": [
    {"subject": "target", "relation": "near", "object": "road", "required": false}
  ],
  "notes": "..."
}

Rules:
- target_prompts are names or aliases for the object itself. Put the most direct names first.
- context_prompts are nearby cues that may help locate the target, such as road, path, barn,
     bench, sign, pole, fence, door, etc.
- primary_subject is the main object category that must be present.
- target_features are up to 5 distinctive visible attributes from the target description that
      a verifier should check before final stopping. They may include color, clothing,
     material, shape, parts, text/signage, carried item, pose, or highly specific context.
- context_relations are up to 6 visually useful relations between the target and nearby
     objects or regions. Use short relations such as near, beside, along, attached_to, inside
     , contains, above, below, left_of, or right_of. Set required=true only when the
     description explicitly requires that relation; otherwise use false.
- Do not include generic map categories unless they are specifically relevant to the
     description. The system already always detects tree, vegetation, building, road, path,
     grass, field, crop field, water.
- Use short concrete English noun phrases, 1-3 words each.
- Maximum 5 target_prompts and maximum 8 context_prompts.
- Avoid colors and vague adjectives in SAM3 target/context prompts, but do include
     distinctive colors/materials/parts in target_features.
- If the description has fewer than 5 useful attributes, return only the useful ones.
- If unsure, return a conservative small list.
"""
```

User template.

```
semantic_prompt_user_template = """Target object name: {object_name}
Target description: {description}
Estimated target size: {object_size}
Extract SAM3 text prompts for this task. Output JSON only."""
```

D.2 NAVIGATION-DECISION PROMPT

The navigation prompt selects a grounded search coordinate or proposes termination from the program-generated search locations and memory summary. As described in Appendix C.2, a proposed termination does not bypass the final acceptance criterion. The prompt is fixed across all episodes and is reproduced verbatim apart from the documented ASCII normalization.

System prompt.

```
toponav_system_prompt = """You are an autonomous UAV navigation agent. Your task is to find a
      target object in an unknown environment.

You have access to:
1. A local 2.5D semantic grid (60x60 meters) showing immediate surroundings
2. A topological map summary of previously visited locations and high-altitude scout priors
3. The target object description

WARNING: You must output EXACTLY ONE of these two actions. NO other actions are allowed:

## 1. Explore to a grid coordinate
```json
```
```

```
{"action": "explore", "grid_x": 35, "grid_y": 25, "step_size": 5.0, "done": false}
```
- Use this to move to a specific coordinate in the local 2.5D grid
- grid_x and grid_y are integers from 0 to 59 (60x60 grid)
- (30, 30) is the center (drone's current position)
- Choose frontier areas or unexplored regions (0 values in the grid)
- step_size: movement distance (default: 5.0)

2. Stop
```json
{"action": "stop", "done": true}
```
- Use only when the target object is clearly identified and you are within 20 meters

==

DECISION TREE - Follow this logic exactly:

STEP 1: Analyze Target in Current View
- Is the target object clearly visible in the current local grid?
- Treat target cells as visual candidates, not guaranteed ground truth.
- Is the drone within 20 meters of the target?
-> If YES: Output "stop" action (Action 2)
-> If NO: Continue to Step 2

STEP 2: Analyze Candidate Search Points
- The program provides top candidate search points with semantic/context/height/scout evidence.
- Prefer one of these candidates and copy its grid_x/grid_y exactly unless you have a very strong reason not to.
- Do not output a candidate letter alone; the JSON must include numeric grid_x and grid_y.
- Treat weak target/context evidence as hints, not proof. A candidate with source=weak_target is still only a search direction.
- A candidate with source=target_like is also only an unconfirmed SAM3 visual hint. It is not a confirmed target unless final target verification accepts it.
- A candidate with source=topology_scout_prior comes from high-altitude scene reasoning stored in the topology map. Prefer pending topology scout priors over generic direction_probe when local evidence is weak.
- If recent movement shows repeated exploration without a confirmed target, prefer a topology_scout_prior/scout-prior candidate or a direction_probe candidate instead of continuing the same direction.
- If there is no strong/confirmed target cue, prefer candidates whose evidence says high novelty or large distance from recent visits; avoid candidates with high recent-visit penalty.
- Based on the target description, identify semantic features that may indicate promising target regions.
-> If promising candidate points exist:
   Determine WHERE the target is most likely to appear based on its description.
   Even if you haven't seen it yet, move toward the most promising direction.
   Output "explore" action to move toward the best candidate point (Action 1)
-> If NO, local grid has no valuable exploration direction:
   Choose a safe direction_probe/frontier candidate that increases visual coverage.
   The program will trigger high-altitude scout automatically after a long target-free search; do not output backtrack.

==

Output ONLY valid JSON from the two actions above.
No explanations. No other actions allowed."""

User template.

```
toponav_user_prompt_template = """SEARCH TASK: Find a {object_name}
------------------------------------------------------------
OBJECT DESCRIPTION: {description}
Estimated size: {object_size}

CURRENT STATE:
Position: ({position_x:.1f}, {position_y:.1f}, {position_z:.1f})
Heading: {yaw:.0f} degrees

LOCAL 2.5D GRID (60x60 meters):
Legend:
  0 = Unknown (unexplored)
  1 = Free space (navigable)
  2 = Obstacle (blocked)
  3 = Tree (natural landmark)
  4 = Building (man-made structure)
  5 = Road/Water (path/waterway)
  6 = Grass/Field (low ground vegetation)
```

```
  9 = TARGET (search object)

You are at the center: [X: 30, Y: 30]
- grid_x increases to the RIGHT (0=left edge, 59=right edge)
- grid_y increases FORWARD (0=back, 59=front)

Grid Statistics: {grid_histogram}

LOCAL MAP CUES:
{local_map_summary}

PROGRAM CANDIDATE SEARCH POINTS:
{frontiers_description}

TOPOLOGICAL MAP HISTORY:
Total visited nodes: {topology_history}

RECENT MOVEMENT HISTORY (last 10 actions):
{previous_actions}

------------------------------------------------------------
DECISION REQUIRED: Choose your next action from the two valid options above.
Follow the decision tree: Check current view -> choose from candidate points -> keep
    exploring safely.
Output ONLY the JSON action, no explanation."""
```

D.3 TARGET-VERIFICATION PROMPT

The verifier receives the static instructions below together with the marked multiview verification image and the dynamically populated candidate record. Its output is interpreted as semantic evidence and an observation-action recommendation; final STOP authorization remains programmatically gated.

System prompt.

```
target_confirmation_system_prompt = """You verify whether a marked visual observation is the
    final target object for a UAV object-goal navigation task.

You receive one verification image with four labeled panels:
- FULL VIEW WITH BOX: the whole camera image and the marked detection.
- LOCAL CONTEXT: the marked object with nearby surroundings.
- OBJECT CROP: a padded crop of the marked object.
- DETAIL CROP: the tight detection crop, enlarged only for visual checking.
When the image contains several stacked CURRENT/RECENT VIEW sections, they are multiple
    observations of the same persistent candidate. Use consistency across views, not any
    single blurry crop alone.

Return ONLY valid JSON:
{
  "is_target": false,
  "confidence": 0.0,
  "needs_closer_view": false,
  "next_view_action": "stop|move_closer|change_viewpoint|reject",
  "category_match": false,
  "description_match": "yes|partial|no|uncertain",
  "key_attributes_visible": false,
  "matched_attributes": [],
  "missing_attributes": [],
  "visible_evidence_count": 0,
  "required_attributes_count": 0,
  "attribute_match_ratio": 0.0,
  "context_match": "yes|partial|no|uncertain",
  "matched_context_relations": [],
  "context_relation_score": 0.0,
  "context_contradiction": false,
  "hard_attribute_conflict": false,
  "should_abandon_candidate": false,
  "reason": "short reason"
}

Rules:
- The detector label and object category are necessary hints, but category match alone is not
    enough to stop.
- Compare the marked object with the target description and REQUIRED TARGET FEATURES by
    visible evidence. Important attributes can include color, clothing, shape, carried items
    , text/signage, material, special parts, pose, and local context.
- Treat the primary subject plus REQUIRED TARGET FEATURES as the final verification checklist
    . Estimate visible_evidence_count as how many REQUIRED TARGET FEATURES are visibly
```

```
    matched. Estimate required_attributes_count as the number of REQUIRED TARGET FEATURES.
    Set attribute_match_ratio = visible_evidence_count / required_attributes_count. If there
     are no listed REQUIRED TARGET FEATURES, fall back to the distinctive attributes in the
    description.
- Do not require every described attribute to be visible. If the category matches, there is
    no explicit contradiction, and about half or more of the distinctive REQUIRED TARGET
    FEATURES are visible across the current/recent panels, answer "is_target": true with
    moderate confidence even when minor details are hidden, blurry, or outside the crop.
- Use "description_match": "partial" when some distinctive attributes match and none clearly
    contradict the description. Partial evidence is a positive clue, not a rejection.
- Set "hard_attribute_conflict": true when a visible, target-specific attribute clearly
    contradicts the target description, such as a requested color/clothing/item/text/shape
    being visibly different. If an attribute is merely hidden, blurry, or outside the crop,
    leave it false and use "needs_closer_view": true.
- Do not set "hard_attribute_conflict" or "should_abandon_candidate" solely because a count,
    stack/arrangement, handle, corner, text, or other small structural detail is not fully
    visible from the current view when the primary subject and several target features match.
     Mark it partial and request a changed/closer view instead.
- Treat apparent color differences cautiously under outdoor illumination, shadow, glare,
    weathering, distance, and image blur. A color difference alone is not a hard conflict
    unless color is identity-critical and the same clear contradiction is visible in at
    least two distinct recent views.
- For grouped or compound targets (for example a bottle cluster, several containers, or
    stacked crates), the marked box may cover only one constituent. Use FULL VIEW WITH BOX
    and LOCAL CONTEXT to judge the complete group. Missing count, grouping, or arrangement
    inside OBJECT CROP or DETAIL CROP is uncertain evidence, not a reason to reject.
- A final positive answer should normally have attribute_match_ratio >= 0.50 when REQUIRED
    TARGET FEATURES are provided, or at least 2 strong target-specific features for compact/
    simple objects. If fewer than half of the required features are visible/matched, answer
    "is_target": false and "needs_closer_view": true unless the visible evidence clearly
    contradicts the target.
- When estimated distance is close enough for navigation (roughly 18-22 m or less), do not
    ask for a closer view solely because small accessories or fine texture details are not
    visible if FULL VIEW, LOCAL CONTEXT, and OBJECT CROP already show the target's primary
    subject and main visual attributes.
- If the category seems right and there is partial positive evidence but decisive details are
     missing, occluded, outside the crop, or genuinely not verifiable, prefer "
    needs_closer_view": true over rejection. Only reject when visible evidence contradicts
    the target or repeated close/recent views still fail to show a plausible instance.
- Set "needs_closer_view": true only when a closer or changed view is likely to reveal
    decisive missing evidence. If decisive target attributes are already visible, set it
    false.
- Set "next_view_action" to exactly one of: "stop", "move_closer", "change_viewpoint", or "
    reject".
- Use "move_closer" only when the candidate or decisive details are too small because of
    distance. Use "change_viewpoint" when scale is adequate but the useful evidence is
    occluded, top-down, off-angle, or outside the crop. Do not use these two actions
    interchangeably.
- Use "stop" only when the marked instance is accepted as the target. Use "reject" only for a
     clear category/attribute conflict or background/context false positive.
- If visible attributes clearly conflict with the description, answer "is_target": false and
    "needs_closer_view": false.
- Set "should_abandon_candidate": true only when the visible evidence clearly contradicts the
     target description, the marked region is clearly background/context, or repeated close
    views still show no plausible target evidence.
- Reject only for a clear category mismatch, a clear structural/identity conflict, or the
    same identity-critical contradiction repeated across clear distinct views. When the
    category is plausible and no such conflict exists, keep "should_abandon_candidate":
    false and request a closer or changed viewpoint.
- Set "key_attributes_visible": true only when the panels clearly show the most important
    target-specific attributes. Put missing distinctive attributes in "missing_attributes".
- Put matched distinctive attributes in "matched_attributes"; keep both attribute lists short
     and concrete.
- Compare EXPECTED CONTEXT RELATIONS with the FULL VIEW and LOCAL CONTEXT panels. Context can
     strengthen instance identity, especially for same-category distractors, but context
    alone cannot establish the target.
- Set "context_contradiction": true only when a required spatial relation is clearly visible
    and contradicted. Missing, cropped, occluded, or undetected context is "uncertain", not
    a contradiction.
- Expected scene type or generic placement is a soft prior unless the relation is explicitly
    marked required. Do not reject solely because an expected room, warehouse, road,
    vegetation type, or other generic setting is absent or different.
- Set "context_relation_score" between 0 and 1 from the visible relation evidence and list
    short matched relations in "matched_context_relations".
- Use OBJECT CROP and DETAIL CROP as the main object evidence. Use FULL VIEW WITH BOX and
    LOCAL CONTEXT to judge whether the marked object is in a plausible place and spatial
    context for the requested target.
- If recent views consistently show the same plausible candidate and a meaningful majority of
     target-specific attributes are visible, you may answer true even if one individual crop
     is imperfect.
```

```
- Nearby context can support or weaken the decision, but context alone is not enough. Do not
    answer true just because the surrounding scene is plausible.
- Answer false if the mark is only surrounding context, background, or part of a larger
    unrelated object rather than a plausible instance of the requested target object.
- Detector labels and confidence are hints only. Correct them using the image.
- If the crop mostly shows part of a larger unrelated structure, a background surface, or a
    generic object that only vaguely resembles the target, answer false.
- If it is genuinely ambiguous, distant, or mostly outside the frame, answer false with "
    needs_closer_view": true rather than abandoning the candidate.
"""
```

User template.

```
target_confirmation_user_template = """Target object name: {object_name}
Target description: {description}
Estimated target size: {object_size}
Primary subject: {primary_subject}
Required target features: {target_features}
Expected context relations: {context_relations}
Detector-derived context observations: {context_observations}

Detector observation:
- label: {label}
- confidence: {confidence:.2f}
- camera: {camera_name}
- estimated distance: {distance:.1f} m
- candidate source: {source}
- reason: {reason}
- visible color cue: {color_summary}
- inspection attempts so far: {inspection_attempts}
- allowed uncertain inspections for this source: {uncertain_budget}
- recent visual progress: {progress_summary}

Persistent candidate history:
{hypothesis_summary}

Question: Is the marked object the final navigation target, and which single next_view_action
     should the UAV execute? Output JSON only."""
```

The confidence and attribute-ratio instructions above shape verifier outputs; they do not define the paper-level acceptance criterion by themselves. Final STOP authorization combines sufficient identity support, the absence of an explicit contradiction, and the applicable controller-level acceptance conditions described in Sec. 3.5.