HIERARCHICAL PERCEPTION–DECISION TEST-TIME ADAPTATION FOR UAV VISION–LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unmanned aerial vehicle vision–language navigation (UAV-VLN) has attracted increasing attention, requiring an embodied agent to follow natural-language instructions by sequentially reasoning over visual observations in open 3D environments and executing navigation actions. Despite recent progress, more drastic viewpoint and scene variations in unseen environments can induce substantial train–test distribution shifts, manifested as cross-modal misalignment and feature drift at the perception stage, as well as biased action prediction at the decision stage. To address these issues, we propose Hierarchical Perception–Decision Test-Time Adaptation for UAV-VLN. At the perception stage, we propose a Memory-Augmented Cross-modal Perception Adaptation (MACPA) module. By constructing an environment semantic memory bank and introducing adaptive tokens, MACPA generates adaptive semantic guidance prompts to mitigate semantic drift. At the decision stage, we propose an Entropy-Gated Historical Decision Calibration (EHDC) module, which aggregates high-confidence historical action distributions based on the similarity between the current state and cached features, alleviating decision drift and enhancing the consistency and stability of predictions. Extensive experiments on the UAV-VLN benchmarks, OpenFly and OpenUAV, demonstrate that our method achieves better navigation performance and more stable decision behavior than state-of-the-art approaches in unseen environments.

1 INTRODUCTION

Unmanned Aerial Vehicle Vision-and-Language Navigation (UAV-VLN) Tian et al. (2025); Sautenkov et al. (2025); Xiao et al. (2025); Lykov et al. (2025); Serpiva et al. (2025) requires an agent to interpret natural-language instructions, continuously perceive open 3D environments, and sequentially predict navigation actions until reaching the target. By integrating cross-modal perception with sequential decision-making, UAV-VLN has broad applications in search and rescue Grogan et al. (2018); Lyu et al. (2023), infrastructure inspection Yao et al. (2019); Ham et al. (2016), environmental monitoring Hardin et al. (2019); Hu et al. (2021), and aerial reconnaissance Rashid et al. (2023); Puri (2005). As illustrated in Fig. 1, UAV-VLN can be viewed as two consecutive stages: a perception stage that aligns the language instruction with the current visual observation to form a cross-modal representation, and a decision stage that predicts the next action based on this representation. However, train–test distribution shifts in unseen environments can affect both stages: they disrupt language–vision alignment at the perception level and weaken the reliability of state–action mappings at the decision level.

At the perception stage, compared with ground navigation Song et al. (2025); Gao et al. (2024a); Yao et al. (2025); Qi et al. (2020), UAV-VLN operates in larger-scale and more structurally complex three-dimensional environments. In unseen scenes, broader motion ranges and larger variations in viewing distance and scene coverage make visual representations more susceptible to distribution shifts, thereby exacerbating the mismatch between static language instructions and dynamic visual observations. This mismatch degrades the accuracy and consistency of language–vision alignment. At the decision level, variations in spatial layouts and traversable paths make offline-learned state–

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.

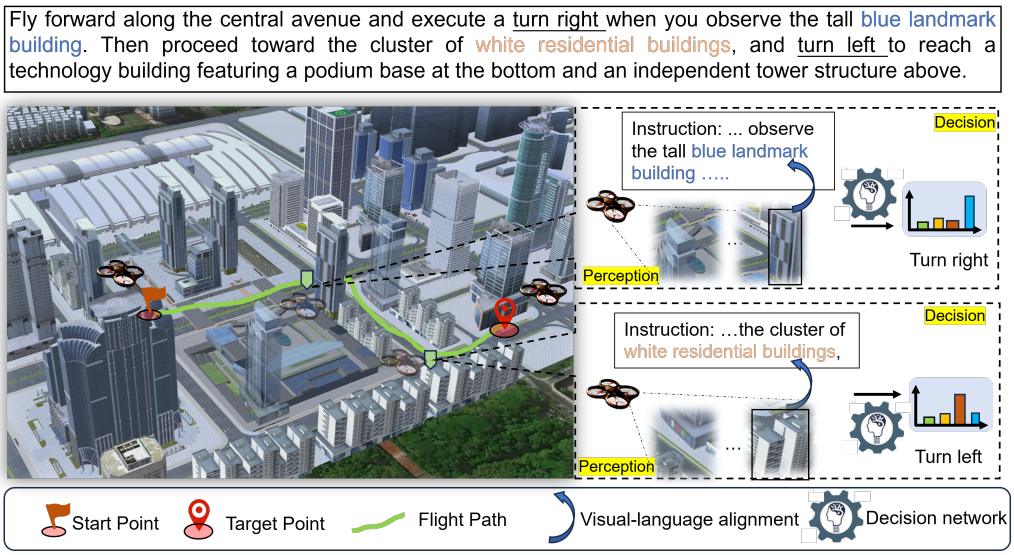


Figure 1: UAV vision–language navigation pipeline. This process is commonly decomposed into two stages: (i) a perception stage that aligns and fuses instruction semantics with the current observation, and (ii) a decision stage that predicts and executes actions based on the fused representation.

action mappings less reliable, while accumulated perception errors can further increase action uncertainty and destabilize navigation trajectories.

These challenges motivate test-time adaptation at both the perception and decision levels. However, most existing UAV-VLN methods Liu et al. (2023); Gao et al. (2025) follow a static inference paradigm after training and lack mechanisms for adapting to test-time distribution changes. Existing TTA approaches Zhang et al. (2025); Lo & Patel (2024); Zhao et al. (2023) often rely on gradient-based optimization, such as entropy minimization, to update model parameters online. Although effective for mitigating distribution shifts, repeated optimization introduces additional computational overhead and can be sensitive to noisy test-time signals, which is undesirable for sequential UAV navigation that requires stable and efficient online inference.

To address these limitations, we propose **Hierarchical Perception–Decision Test-Time Adaptation (HPD)**, a gradient-free framework that performs coordinated adaptation at both the perception and decision stages without updating backbone parameters. At the perception level, we introduce **Memory-Augmented Cross-modal Perception Adaptation (MACPA)**, which maintains an online semantic memory and uses adaptive tokens to retrieve and aggregate environment-relevant semantics. This provides environment-conditioned semantic guidance for improving language–vision alignment under distribution shifts. At the decision level, we introduce **Entropy-Gated Historical Decision Calibration (EHDC)**, which selectively retrieves reliable historical decision information when prediction uncertainty increases and uses it to calibrate the current action distribution. In this way, MACPA improves perception robustness, while EHDC further stabilizes decision-making under residual uncertainty.

Our main contributions are summarized as follows:

- We propose HPD, a gradient-free hierarchical test-time adaptation framework for UAV-VLN that jointly addresses perception-level distribution shifts and decision-level uncertainty without updating backbone parameters.
- We introduce MACPA for environment-conditioned cross-modal adaptation and EHDC for uncertainty-aware historical decision calibration, enabling complementary adaptation across perception and decision stages.
- Extensive experiments on OpenFly and OpenUAV demonstrate consistent improvements in navigation performance and generalization, while maintaining low test-time computational overhead.

2 Related Work

2.1 Unmanned Aerial Vehicle Vision–Language Navigation

Unmanned Aerial Vehicle Vision–Language Navigation (UAV-VLN) Liu et al. (2023); Lin et al. (2025); Zhao et al. (2025) targets instruction-following navigation for UAVs in outdoor environments, requiring continuous egocentric perception and vision–language reasoning to localize goals and plan trajectories. In recent years, datasets such as Aerial-VLN Liu et al. (2023), OpenUAV Lin et al. (2025), CityNav Lee et al. (2025), and OpenFly Gao et al. (2025) have been introduced, providing more realistic and challenging city-scale evaluation environments for UAV-VLN. Building on these benchmarks, prior methods improve UAV-VLN from multiple angles, including structured spatial representations (e.g., grid-based view selection Zhao et al. (2025), cognitive maps Xu et al. (2025), topological graphs Gao et al. (2024b), and hierarchical scene graphs Wang et al. (2024)) to reduce search complexity, hierarchical planning or coarse-to-fine reasoning to decompose long-horizon tasks into sub-goals Jia et al. (2025), and dual-altitude collaboration to separate global reasoning from fine-grained navigation across spatial scales Wu et al. (2025). Recent advances further incorporate multimodal foundation models (VLMs/MLLMs) via zero-shot decision-making Gao et al. (2024b), map-enhanced fine-tuning Galstyan et al., or reinforcement learning Lin et al. (2025) to enhance long-range planning and open-world generalization. Despite recent advances, most UAV-VLN methods still follow a static inference paradigm with frozen parameters and lack the ability to adapt online to test-time distribution changes, which may lead to cross-modal semantic drift and unstable decision-making. To address this issue, we introduce a test-time adaptation framework that dynamically regulates perception and decision processes without updating backbone parameters, thereby improving robustness and generalization in complex aerial environments.

2.2 Test-Time Adaptation

Test-time adaptation (TTA) Wang et al. (2021); Zhang et al. (2023); Gao et al. (2024a); Kim et al. (2025) exploits unlabeled test-time statistics to mitigate train–test distribution shifts commonly encountered in real deployments arising from illumination changes, sensor noise, and environmental variations. Existing methods can be broadly grouped into two categories: (i) parameter-updating approachesWang et al. (2021); Niu et al. (2023); Lee et al. (2024), that adapt model parameters at test time, typically using unsupervised objectives to adjust normalization layers, and (ii) parameter-free approachesZhang et al. (2023); Zanella et al. (2025); Karmanov et al. (2024); Huang et al. (2025) that keep the backbone frozen and refine predictions via auxiliary mechanisms such as clustering, statistical correction, or cache-based memory. The former is more flexible but relies on online gradient updates, which may lead to instability and cumulative drift, whereas the latter is more controllable and easier to deploy but depends on the design of external structures and can under-adapt in complex scenarios. Recent studies Gao et al. (2024a); Kim et al. (2025); Ko et al. (2026); Hu et al. (2026) have introduced TTA into VLN, exploring online parameter updates, feedback-driven optimization, uncertainty-aware adaptation, and historical knowledge reuse to improve generalization to unseen environments. Compared with indoor VLN, UAV-VLN operates in larger-scale and more complex 3D aerial environments, where unseen visual and spatial variations can impair cross-modal perception and increase decision uncertainty, while repeated gradient-based adaptation incurs additional latency and computation. To address this, HPD performs gradient-free hierarchical adaptation, using MACPA for perception-level semantic adaptation and EHDC for uncertainty-aware decision calibration, thereby improving environmental adaptability while enhancing navigation efficiency.

3 Method

3.1 Task setting

Given a natural-language instruction X, the agent must execute a sequence of actions by reasoning over its visual observations. At each time step t, the UAV receives the current visual observation I_t, which is encoded by an image encoder into visual features V_t. A learnable policy network takes the instruction X, the current visual features V_t, and a short history of past visual features H_t as input, and outputs the next action a_t, where H_t denotes the historical visual context. The simulator executes the predicted actions iteratively until the agent outputs the `[stop]` action or reaches the

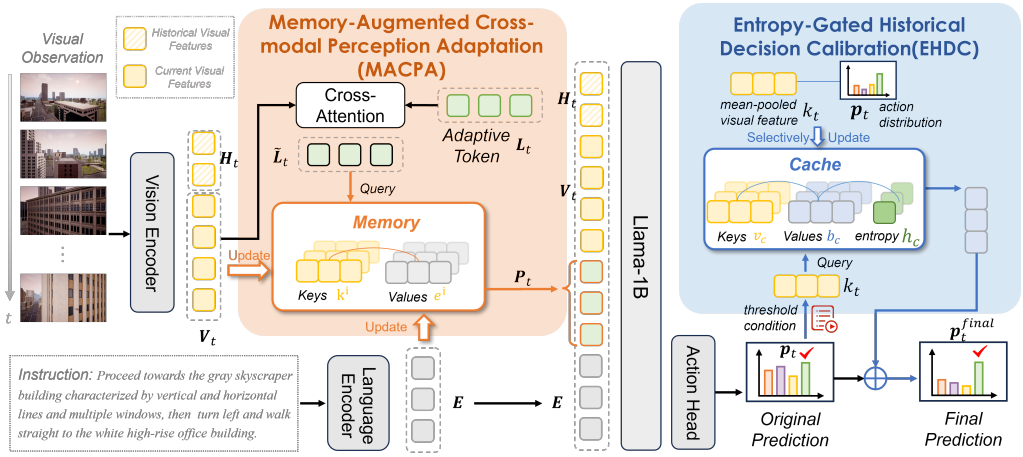


Figure 2: The proposed HPD framework consists of two modules, MACPA and EHDC. At the perception stage, MACPA maintains an online environment semantic memory bank and introduces adaptive tokens to retrieve and softly aggregate environment-relevant semantics, generating adaptive semantic guidance prompts that mitigate semantic drift. At the decision stage, EHDC performs entropy-gated calibration by aggregating high-confidence historical action distributions according to the similarity between the current state and cached prototypes, alleviating decision drift and improving the consistency and stability of action predictions.

maximum number of steps. Navigation is considered successful if the final UAV position lies within 20 m of the target location; otherwise, the episode is regarded as a failure.

3.2 Memory-Augmented Cross-modal Perception Adaptation

During testing in UAV-VLN, the instruction remains fixed within an episode, whereas visual observations evolve continuously due to viewpoint changes, illumination variations, and sensor noise. This non-stationarity weakens language–vision alignment and induces semantic drift, which in turn increases prediction uncertainty and exacerbates error accumulation in closed-loop navigation. To this end, we propose **Memory-Augmented Cross-modal Perception Adaptation (MACPA)**, which injects an online, environment-conditioned semantic prior into perception through a controlled retrieve–aggregate–smooth process.

Given an instruction X, a frozen language encoder encodes it into instruction token embeddings $E \in \mathbb{R}^{q\times d}$. At time step t, given the observation image I_t, a frozen vision encoder extracts visual tokens $V_t \in \mathbb{R}^{m\times d}$. MACPA maintains an online memory bank

$$M_t = \{(k_i, e_i)\}_{i=1}^{|M_t|}, \qquad |M_t| \le K. \tag{1}$$

Each memory entry is represented as (k_i, e_i), where $k_i \in \mathbb{R}^d$ denotes the visual-state prototype used as the key, and $e_i \in \mathbb{R}^d$ denotes the instruction-conditioned semantic embedding used as the value. For the current state, we obtain k_t by pooling the visual tokens V_t into a compact state-level representation. To construct e_t, we use V_t to attend to the instruction tokens E and aggregate the attended features. In this way, e_t encodes instruction semantics grounded in the current visual observation, serving as a semantic anchor for more reliable memory retrieval and cross-modal alignment under noisy observations or test-time distribution shifts.

During inference, MACPA maintains adaptive tokens $L_t \in \mathbb{R}^{N\times d}$ as a persistent state linking the static instruction with dynamic visual observations. At the beginning of each episode, L_0 is initialized as a set of lightweight token vectors and is subsequently updated online throughout navigation. At each step, L_t interacts with the current visual tokens V_t through cross-attention to produce $\widetilde{L}_t$, where L_t preserves accumulated semantic state and $\widetilde{L}_t$ serves as a transient query for retrieving environment-relevant semantics from the online memory. Concretely, for each query token $\tilde{L}_t^{(j)}$, we select its top-5 nearest memory keys under cosine similarity and compute $P_t^{(j)}$ by a temperature-

scaled softmax-weighted aggregation of the corresponding values:

$$P_t^{(j)} = \sum_{i \in \text{Top-5}(\tilde{L}_t^{(j)})} \frac{\exp\big(\cos(\tilde{L}_t^{(j)}, k_i)/\tau\big)}{\sum_{\ell \in \text{Top-5}(\tilde{L}_t^{(j)})} \exp\big(\cos(\tilde{L}_t^{(j)}, k_\ell)/\tau\big)} e_i. \tag{2}$$

Here, τ is a temperature parameter, set to $\tau = 1$ by default, and $P_t \in \mathbb{R}^{N \times d}$ denotes the environment-adaptive semantic prompt tokens. These prompts summarize instruction-relevant semantics observed in similar environmental states, providing contextually grounded representations to stabilize perception under distribution shifts.

To smoothly incorporate the retrieved semantics, we update the adaptive tokens with an exponential moving average, $L_{t+1} = \lambda L_t + (1 - \lambda)P_t$, where λ controls the contribution of the accumulated state and the current semantic prompt. This update preserves temporal continuity while gradually incorporating newly retrieved semantics. The resulting prompt P_t, together with the instruction tokens E, navigation history H_t, and current visual tokens V_t, is then fed into the frozen multimodal backbone f_θ to obtain the perception representation h_t for downstream action prediction.

Importantly, MACPA performs adaptation without test-time gradient updates and without modifying any backbone parameters. All adaptation is realized through online memory retrieval and token updates, making it lightweight and stable for UAV deployment. At the beginning of test-time evaluation, the memory is initialized once with K source-domain entries and is continuously retained and updated across subsequent trajectories. After each incoming pair (k_t, e_t) is obtained, we retrieve a compact set of highly relevant memory entries by selecting the ρ keys with the largest cosine similarities:

$$\mathcal{I}_t = \underset{\substack{\mathcal{S} \subseteq \{1,\dots,|M_t|\} \\ |\mathcal{S}|=\rho}}{\arg\max} \sum_{i \in \mathcal{S}} \frac{k_t^\top k_i}{\|k_t\|_2 \|k_i\|_2}. \tag{3}$$

where ρ denotes the retrieval budget. We set $\rho = 5$ in all experiments. We keep the retrieved keys unchanged and perform element-wise momentum updates on the corresponding semantic values using the current representation e_t:

$$\forall i \in \mathcal{I}_t : \quad e_i \leftarrow \mu e_i + (1 - \mu)e_t, \qquad \mu \in [0, 1). \tag{4}$$

This fixed-capacity online refinement preserves stable retrieval geometry while continuously injecting environment-conditioned semantics into the activated memory values, thereby stabilizing cross-modal perception under test-time distribution shifts.

3.3 Entropy-Gated Historical Decision Calibration

After the perception stage, the UAV encodes the current environment and aligns visual observations with the language instruction through cross-modal fusion, yielding a navigation state representation h_t. The agent then enters the decision stage, where the policy predicts the next action based on this representation. Specifically, at time step t, the frozen policy head maps the step representation h_t through a lightweight MLP to produce action scores, which are then normalized into an action distribution p_t over the available control primitives.

However, in unseen environments, beyond perception-level semantic shift and cross-modal alignment errors, the decision stage also suffers from pronounced distribution shift and instability: spatial layouts and traversable-path patterns can differ markedly from the training data, making offline-learned state–action mappings difficult to transfer. Moreover, semantic ambiguity and feature drift from perception may accumulate over time, further enlarging the uncertainty of the action distribution and ultimately resulting in policy jitter and unstable trajectories. To this end, we propose **Entropy-Gated Historical Decision Calibration (EHDC)**. EHDC maintains a lightweight historical decision cache and performs entropy-triggered retrieval and calibration only when the current prediction becomes uncertain, thereby stabilizing sequential decision-making.

At time step t, the frozen policy outputs an action distribution $p_t \in \mathbb{R}^C$ over the action space. We use the mean-pooled visual feature k_t from the frozen perception encoder as the retrieval key. EHDC maintains a class-wise key–value cache

$$\mathcal{D} = \{(v_c, b_c, h_c)\}_{c=1}^{C}, \qquad (v_{c_t}, b_{c_t}, h_{c_t}) \leftarrow (k_t, p_t, H(p_t)), \quad \text{if } \mathcal{D}_{c_t} = \emptyset \text{ or } H(p_t) < h_{c_t}. \tag{5}$$

Here, each action class keeps at most one high-confidence prototype, where $v_c \in \mathbb{R}^d$ denotes the cached visual-state prototype, $b_c \in \mathbb{R}^C$ is the cached pseudo-label distribution, and h_c stores the associated entropy as a confidence indicator. This design yields constant memory per class and negligible overhead.

Rather than directly adding the two distributions, We performs an uncertainty-aware convex fusion between the current prediction and the historical prior:

$$\alpha_t = \mathrm{clip}\left(\frac{H(p_t) - \gamma}{\log C - \gamma}, 0, 1\right), \qquad p_t^{\mathrm{final}} = (1 - \alpha_t)p_t + \alpha_t \tilde{p}_t, \tag{6}$$

where C denotes the number of discrete actions and $\mathrm{clip}(x, 0, 1) = \min(1, \max(0, x))$. As the prediction uncertainty increases, α_t assigns a larger weight to the historical prior, allowing EHDC to provide stronger calibration when the current policy is less confident. such that EHDC intervenes more strongly only when the policy is unreliable. The final action is selected as $\hat{a}_t = \arg\max_c p_t^{\mathrm{final}}(c)$. Overall, EHDC suppresses action oscillation and mitigates error accumulation in long-horizon UAV navigation while keeping the backbone entirely frozen and the online computation lightweight.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our method on two large-scale UAV vision–language navigation benchmarks, OpenFly Gao et al. (2025) and OpenUAV Wang et al. (2024). OpenFly is built on high-fidelity simulators (e.g., Unreal Engine, GTA V, Google Earth, and 3D Gaussian Splatting) and provides over 100K trajectories. Due to partial scene-release restrictions, we use the public subset with 1,392 Test-Seen and 404 Test-Unseen trajectories. OpenUAV spans 22 urban/rural/natural scenarios with 89 target categories and 12,149 human-collected trajectories (50–400 m), and is split into Train, Test-Seen, Test-Unseen-Map, and Test-Unseen-Object to assess generalization to unseen environments and objects. Following standard protocols, we report SR, OSR, SPL, and NE. SR is the success rate of reaching the goal; OSR measures whether the agent passes sufficiently close to the goal along its trajectory; SPL accounts for both success and path efficiency; and NE is the final Euclidean distance to the target.

4.2 Implementation Details

We adopt OpenFly-Agent as the baseline model. OpenFly-Agent follows the OpenVLA architecture, which employs a fused vision backbone combining SigLIP [9] and DINOv2 [25] encoders, and uses LLaMA-7B as the multimodal decision-making model for vision–language interaction. Considering the requirements of lightweight deployment and inference efficiency, we replace the original 7B backbone with a more compact LLaMA-1B model, and introduce a multilayer perceptron (MLP) as the prediction head for action inference. For the proposed Hierarchical Perception–Decision Test-Time Adaptation method, both the MACPA and EHDC modules are incorporated. In MACPA, the memory size K is empirically set to 1000, the number of tokens N is 5, and λ is set to 0.6. In EHDC, the γ is set to 0.3. These settings achieve a favorable balance between adaptation capacity and computational efficiency. All experiments are conducted on a single NVIDIA RTX 5090 GPU.

4.3 Quantitative Results

We compare HPD with state-of-the-art methods on the OpenFly dataset. As shown in Table 1, when applied to the 7B OpenFly-Agent, HPD improves SR and SPL by 4.3 and 11.3 percentage points on Test-Seen, and by 7.6 and 8.3 percentage points on Test-Unseen, respectively, demonstrating consistent improvements in navigation performance. Although OSR decreases slightly, this behavior reflects a more conservative and goal-directed stopping strategy, which trades broader goal-region coverage for more precise task completion. HPD also yields consistent gains on the lightweight 1B model. On Test-Seen, SR and SPL increase by 6.1 and 6.3 percentage points, respectively, while on Test-Unseen they improve by 5.8 and 3.6 percentage points. Moreover, the 1B+HPD model achieves the best NE on Test-Unseen and attains SR and SPL comparable to those of the 7B OpenFly-Agent,

Table 1: **Quantitative comparison on the OpenFly dataset.** NE is measured in meters (m), while all other metrics are reported in percentage (%). Random means randomly selecting one action to execute until the stop action is chosen. **bold** indicates better performance than the baseline.

Method	test-seen				Test-Unseen			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Random	242	0.7	0.8	0	301	0.1	0.1	0
Seq2SeqKrantz et al. (2020)	291	0.1	3.2	0.6	291	0.1	0.9	0.6
CMAKrantz et al. (2020)	144	5.7	30.0	5.1	268	0.7	6.6	0.7
AerialVLNLiu et al. (2023)	130	6.6	40.8	7.0	227	2.3	19.8	2.2
NavidZhang et al. (2024)	142	9.9	24.3	3.9	231	4.2	27.0	3.5
NaVilaCheng et al. (2024)	140	14.2	41.7	6.7	226	4.5	25.1	2.9
OpenFly-Agent-7BGao et al. (2025)	93	34.3	64.3	24.9	154	22.6	56.2	19.1
+HPD(Ours)	**71**	**38.6**	57.5	**36.2**	**77**	**30.2**	50.7	**27.4**
OpenFly-Agent-1BGao et al. (2025)	81	31.3	58.6	28.9	78	23.1	46.1	21.5
+HPD(Ours)	**66**	**37.4**	56.1	**35.2**	**64**	**28.9**	**47.6**	**25.1**

indicating that the observed gains mainly stem from the proposed perception–decision adaptation mechanism rather than simply from increasing backbone scale. Overall, HPD improves navigation performance and generalization while maintaining a lightweight model configuration.

Table 2: **Quantitative comparison on the OpenUAV dataset.** NE is measured in meters (m), while all other metrics are reported in percentage (%). S and US represent the *Test-seen* and the *Test-unseen*, respectively. Regarding task difficulty, Easy refers to trajectories shorter than 250 meters, while Hard refers to those exceeding 250 meters. The best results are **bold**, and the second-best are underlined.

Method	Test Set	Full				Easy				Hard			
		NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑
Random	S	222.20	0.14	0.21	0.07	142.07	0.26	0.39	0.13	320.12	0.00	0.00	0.00
Fixed	S	188.61	2.27	8.16	1.40	121.36	3.48	11.48	2.14	270.69	0.79	4.09	0.49
CMA Krantz et al. (2020)	S	140.93	4.89	11.56	4.41	83.58	7.35	17.81	6.53	210.91	1.89	3.94	1.83
OpenUAV Wang et al. (2024)	S	106.28	16.10	44.26	14.30	68.78	18.84	47.61	16.39	152.04	12.76	40.16	11.76
NavFoM Zhang et al. (2026)	S	93.05	29.17	49.24	25.03	58.98	32.91	53.16	27.87	143.83	23.58	43.40	20.80
HPD (Ours)	S	**89.31**	**31.06**	**52.46**	**26.84**	**58.36**	**33.64**	**57.03**	**28.72**	**115.45**	**27.10**	**50.39**	**22.78**
Random	US	225.64	0.06	0.06	0.06	164.66	0.19	0.19	0.19	280.58	0.00	0.00	0.00
Fixed	US	193.30	1.76	5.36	1.09	140.33	3.19	8.08	1.88	245.96	0.85	3.08	0.55
CMA Krantz et al. (2020)	US	147.27	4.98	12.41	4.74	102.54	8.03	17.52	7.52	191.30	2.76	7.53	2.71
OpenUAV Wang et al. (2024)	US	130.60	11.41	31.13	10.45	96.27	12.47	33.31	11.29	167.49	10.62	28.91	9.81
NavFoM Zhang et al. (2026)	US	118.34	15.63	30.46	14.21	89.77	16.98	32.22	15.35	155.69	14.35	27.79	13.16
HPD (Ours)	US	**114.71**	**19.53**	**36.36**	**18.44**	**84.62**	**21.09**	**41.73**	**19.47**	**147.55**	**18.43**	**32.18**	**16.23**

Table 2 reports the results on OpenUAV under the Test-Seen and Test-Unseen settings. HPD achieves the best performance across all metrics on the Full, Easy, and Hard splits. Compared with the OpenUAV baseline, HPD improves SR/SPL by 14.96/12.54 percentage points and reduces NE by 16.97 m on Test-Seen Full; on Test-Unseen Full, SR/SPL increase by 8.12/7.99 points and NE decreases by 15.89 m. Compared with the stronger NavFoM baseline, HPD further improves SR/SPL by 1.89/1.81 points on Test-Seen Full and 3.90/4.23 points on Test-Unseen Full, demonstrating the consistent effectiveness of the proposed hierarchical perception–decision adaptation in both seen and unseen environments.

4.4 Ablation Experiments

Comparison with Different TTA Strategies. To ensure a fair comparison, all methods are evaluated using the same 1B OpenFly-Agent backbone, identical Test-Unseen trajectories, and the same trajectory order. We additionally report GPU memory usage and average per-step inference latency to assess test-time efficiency. As shown in Table 3, HPD is compared with representative TTA methods, including SAR, FSTTA, ZeroSiam, TDA, and LTPO, with implementation details provided in the supplementary material. While these methods generally improve over the Base model, they often introduce additional optimization, memory, or inference overhead. In contrast, HPD reduces NE from 78 to 64, improves SR from 23.14% to 28.85%, and increases SPL from 21.46% to 25.13%,

Table 3: Comparison of different TTA methods on the OpenFly Test-Unseen split in terms of navigation performance and inference efficiency. GPU denotes the evaluation-time memory usage in MB, and Time denotes the average per-step inference latency in ms.

Method	Test-Unseen				
	NE ↓	SR ↑	SPL ↑	GPU ↓	Time ↓
Base	78	23.14	21.46	5322	**298**
SARNiu et al. (2023)	71	26.43	23.52	5910	396
TDAKarmanov et al. (2024)	69	26.56	22.71	**5334**	322
FSTTAGao et al. (2024a)	70	26.95	23.87	5790	368
ZeroSiamChen et al. (2026)	74	25.73	22.86	5824	343
LTPOYe et al. (2026)	69	27.17	24.40	5879	540
HPD (Ours)	**64**	**28.85**	**25.13**	5390	319

with only 68 MB additional GPU memory and 21 ms extra latency. Compared with LTPO, HPD further improves SR and SPL by 1.68 and 0.73 percentage points, respectively, while reducing latency by 221 ms. These results demonstrate that HPD achieves a favorable balance between navigation performance and computational efficiency, making it well suited for real-time UAV-VLN.

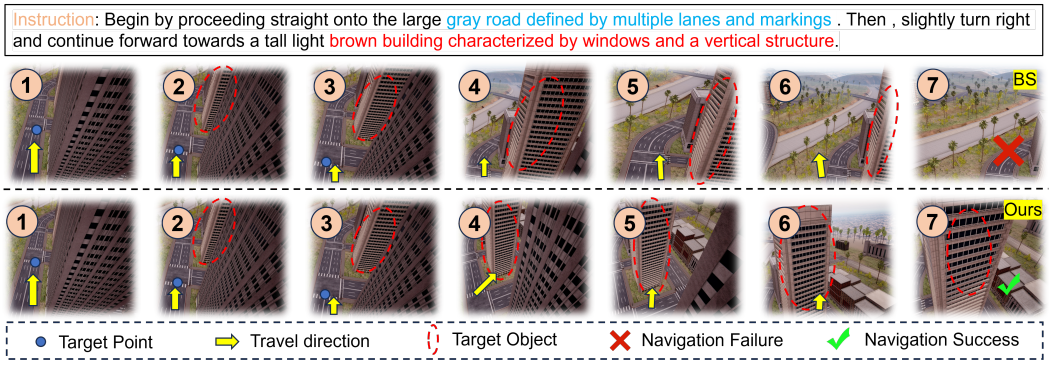
Table 4: Ablation study of the MACPA and EHDC modules under the OpenFly Test-Unseen split. The best result for each metric is highlighted in **bold**.

Method			Test-Unseen			
Baseline	MACPA	EHDC	NE↓	SR↑	OSR↑	SPL↑
✓			78	23.14	46.13	21.46
✓	✓		73	25.54	**47.79**	23.68
✓		✓	69	26.11	46.67	24.65
✓	✓	✓	**64**	**28.85**	47.62	**25.13**

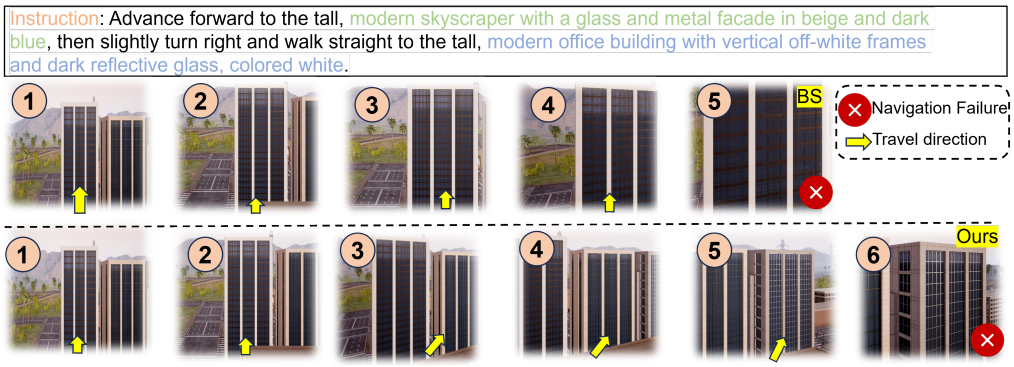
Impact of MACPA and EHDC on Experimental Performance. We further evaluate MACPA and EHDC on the more challenging Test-Unseen split, where the agent must generalize to previously unseen environments. As shown in Table 4, adding MACPA alone improves SR from 23.14% to 25.54%, OSR from 46.13% to 47.79%, and SPL from 21.46% to 23.68%, while reducing NE from 78 to 73. This indicates that memory-augmented cross-modal perception adaptation helps obtain more transferable semantic representations under domain-shifted visual observations. EHDC alone further improves SR to 26.11% and SPL to 24.65%, with a larger NE reduction from 78 to 69, suggesting that entropy-gated historical decision calibration effectively suppresses unstable action predictions in unfamiliar environments. When combined, MACPA and EHDC achieve the best NE, SR, and SPL of 64, 28.85%, and 25.13%, respectively. Compared with using either module alone, their combination yields further improvements, indicating complementary effects across the perception and decision stages. MACPA first improves the robustness of the navigation representation, while EHDC subsequently calibrates residual uncertainty in the resulting action predictions. This sequential coupling enables the two modules to address different sources of test-time degradation within a unified inference pipeline.

4.5 Qualitative Analysis

Visualization Results. To validate the effectiveness of our method, we provide a qualitative comparison with the baseline in Fig. 3a. After reaching the gray road, the baseline fails to follow the instructed right turn and continues forward, deviating from the target direction. In contrast, our method leverages MACPA to generate environment-conditioned semantic prompts, reducing language–vision misalignment under changing observations. Consequently, it better grounds the instruction to the scene, identifies the brown building as the key landmark, performs the correct right turn, and successfully reaches the destination. Fig. 3b presents a challenging failure case for both



(a) Visualization comparison between HPD and the baseline, illustrating a representative case in which the baseline fails to complete the navigation task, while the proposed method successfully reaches the target destination.



(b) Visualization comparison between HPD and the baseline in scenarios where both methods fail to complete the navigation task.

Figure 3: Qualitative navigation examples comparing HPD with the baseline on successful and failed episodes.

models. The baseline directly moves toward the first target due to insufficient instruction grounding and only completes a partial subtask. Our method correctly turns left after reaching the first target, showing improved perception alignment, but later fails to stop at the proper position and moves toward an incorrect target. This failure may be caused by EHDC: imperfectly matched retrieval priors can bias the action distribution in ambiguous regions, leading to overshooting and incorrect termination. This suggests that fixed gating may be suboptimal in complex environments, motivating future work on adaptive triggering and reliability-aware retrieval.

5 Conclusion

We propose HPD, a hierarchical perception–decision test-time adaptation framework for UAV-VLN that operates without updating backbone parameters. At the perception stage, MACPA maintains an online semantic memory and employs adaptive tokens to retrieve and aggregate environment-relevant information, thereby improving the robustness of language–vision alignment. At the decision stage, EHDC performs entropy-gated historical calibration to refine highly uncertain action predictions and enhance temporal decision consistency. Extensive experiments on OpenFly and OpenUAV demonstrate consistent improvements in navigation performance and decision stability. However, as environment scale and complexity increase, semantic memory management and retrieval may become more challenging, motivating more selective and reliability-aware memory update and retrieval strategies. In addition, the fixed entropy threshold in EHDC may not always be optimal. Future work will investigate adaptive gating based on temporal entropy statistics, confidence margins, or quantile-based uncertainty estimates, with the goal of further improving robustness under severe distribution shifts while preserving real-time inference efficiency.

Statement on the Use of AI Assistants

An AI assistant was used to inspect the implementation, organize the manuscript, and improve exposition. Numerical claims were checked against source code, training metadata, complete evaluation logs, or cited primary papers. The authors remain responsible for the scientific claims, citations, and final text.

References

Guohao Chen, Shuaicheng Niu, Deyu Chen, Jiahao Yang, Zitian Zhang, Mingkui Tan, Pengcheng Wu, and Zhiqi Shen. Zerosiam: An efficient asymmetry for test-time entropy optimization without collapse. In *The Fourteenth International Conference on Learning Representations*, 2026.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

Tigran Galstyan, Hakob Tamazyan, and Narek Nurijanyan. Teaching visual language models to navigate using maps. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*.

Junyu Gao, Xuan Yao, and Changsheng Xu. Fast-slow test-time adaptation for online vision-and-language navigation. 2024a.

Yunpeng Gao, Zhigang Wang, Pengfei Han, Linglin Jing, Dong Wang, and Bin Zhao. Exploring spatial representation to enhance llm reasoning in aerial vision-language navigation. *arXiv preprint arXiv:2410.08500*, 2024b.

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, Zhigang Wang, Bin Zhao, and Xuelong Li. Openfly: A comprehensive platform for aerial vision-language navigation. *CoRR*, abs/2502.18041, 2025.

Sean Grogan, Robert Pellerin, and Michel Gamache. The use of unmanned aerial vehicles and drones in search and rescue operations–a survey. *Proceedings of the PROLOG*, pp. 1–13, 2018.

Youngjib Ham, Kevin K Han, Jacob J Lin, and Mani Golparvar-Fard. Visual monitoring of civil infrastructure systems via camera-equipped unmanned aerial vehicles (uavs): a review of related works. *Visualization in Engineering*, 4(1):1, 2016.

Perry J Hardin, Vijay Lulla, Ryan R Jensen, and John R Jensen. Small unmanned aerial systems (suas) for environmental remote sensing: Challenges and opportunities revisited. *GIScience & Remote Sensing*, 56(2):309–322, 2019.

Yiqiang Hu, Chuanxun Yang, Ji Yang, Yanyan Li, Wenlong Jing, and Sijing Shu. Review on unmanned aerial vehicle remote sensing and its application in coastal ecological environment monitoring. In *IOP Conference Series: Earth and Environmental Science*, volume 821, pp. 012018. IOP Publishing, 2021.

Zixuan Hu, Xuantuo Huang, Yancheng Li, Yichun Hu, Shengyong Xu, and Ling-Yu Duan. Turning adaptation into assets: Cross-domain bridging for online vision-language navigation. *arXiv preprint arXiv:2605.23257*, 2026.

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9772–9781, 2025.

Huitong Jia, Hongyuan He, Bing Zhang, et al. Exploratory uav autonomous visual navigation with spatio-temporal cognition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Sungjune Kim, Gyeongrok Oh, Heeju Ko, Daehyun Ji, Dongwook Lee, Byung-Jun Lee, Sujin Jang, and Sangpil Kim. Test-time adaptation for online vision-language navigation with feedback-based reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025.

Heeju Ko, Sung June Kim, Gyeongrok Oh, Jeongyoon Yoon, Honglak Lee, Sujin Jang, Seungryong Kim, and Sangpil Kim. Active test-time vision-language navigation. *Advances in Neural Information Processing Systems*, 38:44756–44775, 2026.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow*, 2020.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*, 2024.

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.

Peican Lin, Gan Sun, Chenxi Liu, Fazeng Li, Weihong Ren, and Yang Cong. Openvln: Open-world aerial vision-language navigation. *arXiv preprint arXiv:2511.06182*, 2025.

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. 15384–15394, 2023.

Shao-Yuan Lo and Vishal M Patel. Adaptive batch normalization networks for adversarial robustness. In *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6. IEEE, 2024.

Artem Lykov, Valerii Serpiva, Muhammad Haris Khan, Oleg Sautenkov, Artyom Myshlyaev, Grik Tadevosyan, Yasheerah Yaqoot, and Dzmitry Tsetserukou. Cognitivedrone: A vla model and evaluation benchmark for real-time cognitive task solving and reasoning in uavs. *arXiv preprint arXiv:2503.01378*, 2025.

Mingyang Lyu, Yibo Zhao, Chao Huang, and Hailong Huang. Unmanned aerial vehicles for search and rescue: A survey. *Remote Sensing*, 15(13):3266, 2023.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023.

Anuj Puri. A survey of unmanned aerial vehicles (uav) for traffic surveillance. *Department of computer science and engineering, University of South Florida*, pp. 1–29, 2005.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Adib Bin Rashid, Ashfakul Karim Kausik, Ahamed al Hassan Sunny, and Mehedy Hassan Bappy. Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International journal of intelligent systems*, 2023(1):8676366, 2023.

Oleg Sautenkov, Yasheerah Yaqoot, Artem Lykov, Muhammad Ahsan Mustafa, Grik Tadevosyan, Aibek Akhmetkazy, Miguel Altamirano Cabrera, Mikhail Martynov, Sausar Karaf, and Dzmitry Tsetserukou. Uav-vla: Vision-language-action system for large scale aerial mission generation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 1588–1592. IEEE, 2025.

Valerii Serpiva, Artem Lykov, Artyom Myshlyaev, Muhammad Haris Khan, Ali Alridha Abdulkarim, Oleg Sautenkov, and Dzmitry Tsetserukou. Racevla: Vla-based racing drone navigation with human-like behaviour. *arXiv preprint arXiv:2503.02572*, 2025.

Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12078–12088, 2025.

Yonglin Tian, Fei Lin, Yiduo Li, Tengchao Zhang, Qiyao Zhang, Xuan Fu, Jun Huang, Xingyuan Dai, Yutong Wang, Chunwei Tian, et al. Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility. *Information Fusion*, 122:103158, 2025.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

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. *arXiv preprint arXiv:2410.07087*, 2024.

Ruipu Wu, Yige Zhang, Jinyu Chen, Linjiang Huang, Shifeng Zhang, Xu Zhou, Liang Wang, and Si Liu. Aeroduo: Aerial duo for uav-based vision and language navigation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2576–2585, 2025.

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.

Haotian Xu, Yue Hu, Chen Gao, Zhengqiu Zhu, Yong Zhao, Yong Li, and Quanjun Yin. Geonav: Empowering mllms with explicit geospatial reasoning abilities for language-goal aerial navigation. *arXiv preprint arXiv:2504.09587*, 2025.

Huang Yao, Rongjun Qin, and Xiaoyu Chen. Unmanned aerial vehicle for remote sensing applications—a review. *Remote sensing*, 11(12):1443, 2019.

Xuan Yao, Junyu Gao, and Changsheng Xu. Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5536–5546, 2025.

Wengao Ye, Yan Liang, and Lianlei Shan. Thinking on the fly: Test-time reasoning enhancement via latent thought policy optimization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r1WEQzkCQv.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

Chenyu Zhang, Kunlun Xu, Zichen Liu, Yuxin Peng, and Jiahuan Zhou. Scap: Transductive test-time adaptation via supportive clique-based attribute prompting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30032–30041, 2025.

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. *Robotics: Science and Systems*, 2024.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 127293–127322, 2026.

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International conference on machine learning*, pp. 41647–41676. PMLR, 2023.

Ganlong Zhao, Guanbin Li, Jia Pan, and Yizhou Yu. Aerial vision-and-language navigation with grid-based view selection and map construction. *arXiv preprint arXiv:2503.11091*, 2025.

Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2305.18010*, 2023.

6 Implementation Details

We adopt OpenFly-Agent as the baseline model. OpenFly-Agent follows the OpenVLA architecture, using a fused visual backbone composed of SigLIP and DINOv2 encoders and LLaMA-7B for vision–language interaction and decision making. To improve deployment efficiency, we replace the original 7B model with a more compact LLaMA-1B model and employ a multilayer perceptron (MLP) as the action prediction head. For the proposed HPD framework, both MACPA and EHDC are enabled. In MACPA, the memory size is set to $K = 1000$, the number of adaptive tokens to $N = 5$, and λ to 0.6. At the beginning of test-time evaluation, we perform trajectory-balanced sampling over the training split and select K source-domain samples to initialize the semantic memory, preventing the initial memory from being dominated by a small number of trajectories. The memory is then continuously retained and updated during testing. In EHDC, the entropy-gating threshold γ is set to 0.3, and its historical decision state is likewise maintained throughout test-time evaluation. All experiments are conducted on a single NVIDIA RTX 5090 GPU.

We also provide detailed implementation settings for the compared TTA methods to facilitate reproducibility. To ensure a fair comparison, all methods are evaluated using the same OpenFly-Agent-1B model on the OpenFly Test-Unseen split, with identical test trajectories, trajectory order, and evaluation protocol. SAR updates only the LLM RMSNorm parameters and performs step-wise adaptation using reliable entropy minimization with SAM, with the entropy filtering threshold set to $0.4 \log C$. FSTTA updates only the LLM normalization parameters and minimizes action entropy through a combination of step-wise fast adaptation and trajectory-level slow refinement, using a gradient-history window of 5, a maximum PCA rank of 4, a suppression ratio of 0.7, and one slow refinement every 20 trajectory snapshots. ZeroSiam updates a lightweight predictor together with the LLM normalization parameters and jointly employs entropy minimization and prediction-consistency regularization, with the consistency weight set to 0.5. LTPO freezes the backbone and optimizes only four latent vectors, performing two optimization rounds per navigation step with four candidates sampled in each round and using normalized negative action entropy as the reward. For TDA, the per-class capacities for positive and negative samples are set to 3 and 2, respectively, with corresponding weights of 4.0 and 0.2.

7 Supplementary Experiments

Table 5: Ablation studies of MACPA on the OpenFly Test-Unseen split. (a) Effect of the number of adaptive tokens N. (b) Effect of the memory size K. The best result for each metric is highlighted in **bold**.

(a) Number of adaptive tokens N

N	NE↓	SR↑	OSR↑	SPL↑
0	78	23.14	46.13	21.46
2	69	25.75	44.87	22.84
3	67	25.96	45.63	23.36
4	**63**	26.92	**47.76**	24.93
5	64	**28.85**	47.62	**25.13**
6	66	27.14	45.53	24.12
7	70	26.44	46.13	24.56

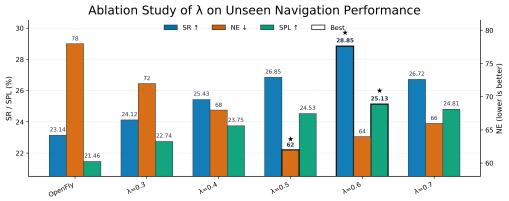
(b) Memory size K

K	NE↓	SR↑	OSR↑	SPL↑
0	78	23.14	46.13	21.46
500	69	26.41	45.02	23.34
1000	**64**	**28.85**	47.62	**25.13**
1500	67	27.21	46.14	24.18
2000	68	26.72	**47.84**	24.67
3000	67	25.32	44.19	22.87

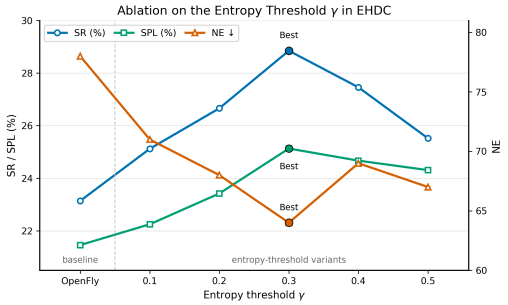
Impact of adaptive Token Number N in MACPA. We investigate the effect of the number of adaptive tokens on the OpenFly Test-Unseen split. As shown in Table 5 (a), introducing adaptive

tokens generally improves NE, SR, and SPL over the baseline, indicating that adaptive tokens help the model capture environment-conditioned semantic cues under distribution shift. When increasing N from 2 to 5, SR and SPL show an overall upward trend, with the best SR of 28.85% and SPL of 25.13% achieved at $N = 5$. Meanwhile, $N = 4$ obtains the lowest NE of 63 and the highest OSR of 47.76%, suggesting that a slightly smaller token set can better preserve broad goal-region coverage. However, further increasing N to 6 or 7 leads to degraded performance, which may be caused by redundant or noisy prompts that weaken representation selectivity. Overall, these results suggest that MACPA benefits from a compact yet expressive set of adaptive tokens, and we adopt $N = 5$ for its best overall SR/SPL performance.

Impact of Memory Size K in MACPA. We further study the effect of memory bank capacity K on Test-Unseen generalization. As reported in Table 5 (b), increasing K from 0 to 1000 substantially improves SR from 23.14% to 28.85% and SPL from 21.46% to 25.13%, while reducing NE from 78 to 64. These results indicate that a properly sized memory bank provides useful historical and environment-conditioned cues for retrieving transferable semantic knowledge in unseen scenes. However, when K exceeds 1000, the main navigation metrics generally decline, suggesting that overly large memory banks may introduce redundant or less relevant entries, which can dilute retrieval precision and bring noisy context. Although K=2000 achieves the highest OSR, its SR and SPL are lower than those of K=1000, indicating that broader goal-region coverage does not necessarily lead to more accurate or efficient task completion. Therefore, a moderate memory capacity achieves the best balance between contextual coverage and retrieval precision.



(a) Impact of the MACPA update magnitude parameter λ.

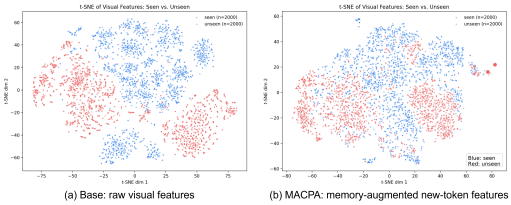


(b) Ablation on the entropy threshold γ in EHDC.

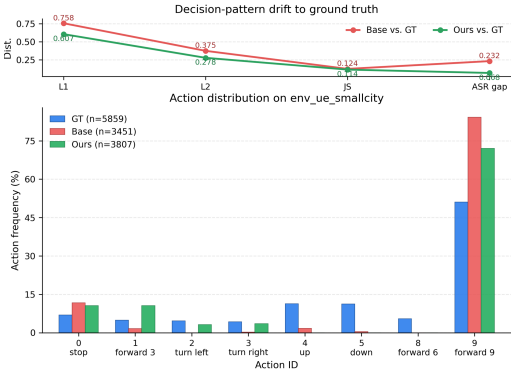
Figure 4: Sensitivity analysis of key hyperparameters on the OpenFly Test-Unseen split. (a) Impact of the MACPA update magnitude parameter λ for adaptive tokens. (b) Impact of the entropy threshold γ in EHDC.

Analysis of the Update Coefficient λ in MACPA on Test-Unseen. We analyze the effect of the EMA update coefficient λ in MACPA, which controls the balance between retaining historical adaptive tokens and incorporating retrieved environment-conditioned semantic cues. As shown in Figure 4 (a), λ=0.6 achieves the best SR of 28.85% and SPL of 25.13%, while maintaining a competitive NE of 64. Although λ=0.5 obtains the lowest NE, λ=0.6 provides the best overall navigation performance in terms of task success and path efficiency. This indicates that a moderate update coefficient enables the adaptive tokens to effectively accumulate useful historical semantics while still absorbing relevant cues from the current unseen environment. When λ is too small, the update relies excessively on the newly retrieved cues, making the tokens sensitive to noisy or mismatched memory entries. In contrast, when λ is too large, the update becomes overly conservative and may fail to adapt sufficiently to the current visual context. Overall, λ should balance historical stability and environmental adaptability, and we set λ=0.6 in our experiments.

Analysis of the entropy threshold γ in EHDC. We evaluate the effect of the entropy threshold γ, which controls when EHDC activates historical decision calibration. As shown in Fig.4 (b), performance improves as γ increases from 0.1 to 0.3, with the best SR and SPL achieved at $\gamma = 0.3$ (28.85% SR and 25.13% SPL), while NE is also reduced to 64 m. This indicates that triggering EHDC only under sufficiently uncertain predictions can effectively suppress unstable action outputs and correct decision drift. When γ is too small, EHDC is activated too frequently, causing historical



(a) Base visual features vs. MACPA representations



(b) Action distributions of Ground Truth, Base, and HPD

Figure 6: Additional analysis on the OpenFly Test-Unseen split. (a) t-SNE comparison between Base visual features and MACPA representations. (b) Comparison of action distributions among Ground Truth, Base, and HPD in unseen environments.

decisions to interfere with confident current predictions and introducing unnecessary calibration noise. Conversely, when γ is too large, EHDC is triggered less often, weakening its ability to correct uncertain decisions. Overall, the results validate entropy-guided gating as an effective mechanism for balancing current-state prediction and historical decision calibration.

Sensitivity to Test Sample Ordering. Since the memory and cache in our method are continuously updated across test samples, their internal states may be influenced by previously processed samples. Consequently, different test sample orders may alter the accumulation of historical information and further affect subsequent predictions. To examine whether the method is sensitive to such order dependence, we use five different random seeds to generate different test sample orders while keeping the test set, model parameters, hyperparameters, and all other evaluation settings unchanged. As shown in Fig. 5, the performance remains stable across different sample orders. Over the five runs, NE, SR, OSR, and SPL are 63.80 ± 0.84 m, 28.65 ± 0.22, 47.52 ± 0.32, and 25.03 ± 0.07, respectively. The maximum differences in SR, OSR, and SPL are only 0.50, 0.75, and 0.17 percentage points, while NE varies by only 2 m. Overall, different test sample orders introduce only minor performance variations, with the overall impact remaining within an acceptable range. This indicates that the method is largely insensitive to test sample ordering and exhibits stable evaluation performance.



Figure 5: Sensitivity to test sample ordering on OpenFly Test-Unseen. Orders 1–5 are generated using five random seeds. NE is in meters; SR, OSR, and SPL are percentages.

8 Qualitative Analysis

Can MACPA effectively mitigate semantic drift? As shown in Fig. 6 (a), we visualize the feature distributions of seen and unseen samples on the OpenFly dataset to examine the representation discrepancy across environments. For the baseline model, the raw visual features from seen and unseen environments exhibit a noticeable separation in the t-SNE projection, with unseen samples forming several isolated clusters. This pattern is consistent with the presence of a seen–unseen representation shift in the baseline feature space. We further visualize the memory-augmented adaptive-token representations generated by MACPA and fed into the downstream multimodal model. Compared with

the baseline raw visual features, the MACPA representations exhibit greater visual overlap between seen and unseen samples in the t-SNE projection. Notably, MACPA does not modify the frozen visual encoder, but instead reshapes the decision-input representations through adaptive token construction and memory retrieval. The resulting visualization suggests improved alignment between seen and unseen representations at the decision-input level. Together with the quantitative improvements in Table 4, these observations provide complementary qualitative evidence that MACPA may alleviate perception-level representation discrepancy under test-time distribution shifts.

Can EHDC effectively alleviate instability in action prediction? To analyze decision drift in unseen environments, we compare the action distributions of ground-truth trajectories, the baseline model, and our method on the OpenFly Test-Unseen split. L1/L2 distances measure action-frequency deviation, JS divergence quantifies distributional mismatch, and the ASR gap evaluates differences in action-transition dynamics. As shown in Fig. 6, (b) the baseline model exhibits action homogenization by overusing the long-distance `forward 9` action, whose proportion reaches 84.2%, much higher than 51.1% in ground-truth trajectories. In contrast, our method reduces `forward 9` to 72.1% and increases short-distance forward and turning actions, yielding a more balanced action pattern closer to the ground truth. Quantitatively, it reduces L1 from 0.758 to 0.607, L2 from 0.375 to 0.278, JS divergence from 0.124 to 0.114, and ASR gap from 0.232 to 0.068. These results show that EHDC-enhanced decision calibration mitigates decision rigidity and stabilizes action prediction under distribution shift.