THE WORLD OUTSIDE, THE MOTION WITHIN: HOLISTIC SPATIAL INTELLIGENCE FOR UAVS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial intelligence for unmanned aerial vehicles (UAVs) relies on the tight coupling of external environmental perception and internal ego-motion awareness. However, existing models struggle to preserve spatial knowledge across viewpoint changes and to disentangle ego-motion from object motion while capturing how the UAV's motion shapes future observations. To address these challenges, we introduce *EnvSelf-UAV*, a unified framework that couples external spatial perception with internal ego-motion understanding. Specifically, a Dynamic Environment Spatial Memory (DESM) module maintains spatially grounded memory by updating regional appearance and geometry across video frames. An Ego-Motion Dynamics Modeling (EMDM) module infers UAV motion from shared geometric changes across matched regions and uses a latent world model to predict future environmental representations conditioned on this motion. A Question-Guided Memory Injection (QGMI) mechanism then selectively integrates environmental, motion, and predictive evidence into a language decoder for spatial reasoning. Extensive evaluations across three key benchmarks demonstrate that EnvSelf-UAV achieves superior performance, while downstream navigation tasks further validate the effectiveness of the proposed architecture.[1]

1 INTRODUCTION

Spatial intelligence is essential for unmanned aerial vehicles (UAVs) to operate autonomously in open 3D environments, with applications ranging from search and rescue to infrastructure inspection and navigation (Zhang et al., 2026c). During flight, UAVs continuously change their position, altitude, and orientation, causing substantial variations in both visual observations and the spatial relationships between the vehicle and surrounding objects (Zhou et al., 2026). As illustrated in Figure 1, these two aspects are inherently coupled: scene observations provide cues for estimating ego-motion, while ego-motion determines how the scene is subsequently perceived. Connecting the world outside with the motion within is thus fundamental to holistic spatial intelligence for UAVs (Zou et al., 2026; Zhang et al., 2026b).

Existing research provides several foundations toward this goal. General vision–language models align visual representations with language reasoning, enabling semantic understanding and question answering over images and videos (Li et al., 2023; Liu et al., 2023a; Lin et al., 2024). Building on this capability, spatially grounded models introduce region-level features, explicit coordinates, and quantitative spatial supervision to strengthen object localization and geometric reasoning (You et al., 2024; Chen et al., 2024; Guan et al., 2026). In aerial settings, AerialVLN studies language-guided navigation from UAV observations (Liu et al., 2023b), while more recent benchmarks extend evaluation to spatial relationships, cross-view reasoning, and motion understanding (Zhang et al., 2026c;b). SIS-Bench further evaluates both spatial cognition and self-awareness, with its accompanying SIS-Motion model incorporating optical flow into visual representations to improve these capabilities (Zou et al., 2026). Collectively, these advances highlight an emerging need to move beyond isolated scene understanding or motion estimation toward a unified framework that jointly models spatial perception and ego-motion dynamics.

Realizing such a unified framework, however, requires addressing two key challenges: **(1) Limited Spatial Cognition.** Image-based representations are inherently tied to the current viewpoint, while

[1]Upon acceptance, we will publicly release all code and model weights.

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.

temporal aggregation in video models does not explicitly preserve and update object-centric information under continuous viewpoint changes. As a result, environmental knowledge can become fragmented over time, limiting the persistent spatial understanding required for reliable reasoning. **(2) Weak Ego-motion Modeling.** UAV ego-motion and independently moving objects jointly alter visual observations, making it difficult to isolate the changes induced by the UAV itself. Beyond recognizing observed motion patterns, a unified framework must capture how ego-motion evolves and shapes future observations. Existing approaches often provide limited modeling of this temporal relationship, making it difficult to use inferred motion to anticipate subsequent observations.

To address these challenges, we propose *EnvSelf-UAV*, a unified framework that couples external spatial perception with internal ego-motion understanding. First, *Dynamic Environment Spatial Memory* (DESM) constructs spatially grounded memory cells from visual observations and continuously updates their appearance and geometry across frames, thereby preserving persistent scene evidence under changing viewpoints. Building on this memory, *Ego-Motion Dynamics Modeling* (EMDM) establishes region correspondences across adjacent frames and estimates UAV motion from their shared geometric transformations. A latent world model (LWM) further predicts future scene representations conditioned on the inferred ego-motion, explicitly connecting motion dynamics with environmental evolution. Finally, *Question-Guided Memory Injection* (QGMI) selectively retrieves and integrates relevant evidence from spatial memory, estimated motion, and predicted future states into the language decoder for answer generation. Extensive experiments on three representative benchmarks demonstrate consistent improvements in spatial cognition, ego-motion understanding, and navigation decision-making, while further supporting the framework's zero-shot transfer across tasks.

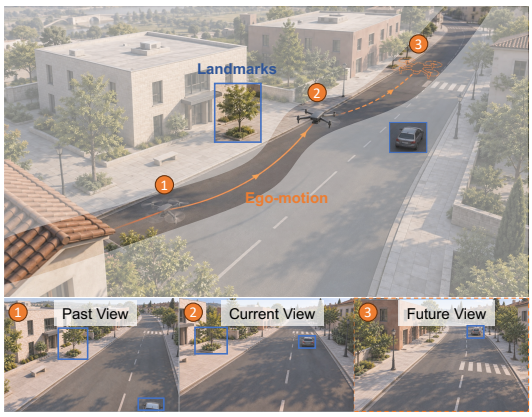

Figure 1: UAV spatial intelligence requires jointly understanding the external environment and the UAV's own motion.

Our contributions are fourfold:

- We propose EnvSelf-UAV, a framework for holistic UAV spatial intelligence that couples external scene perception, internal ego-motion understanding, and their temporal interaction through a shared spatial memory.
- We introduce DESM to carry spatially grounded memory across video frames and update object appearance and geometry, supporting persistent environment perception under changing viewpoints.
- We develop EMDM to infer UAV motion from common geometric changes and predict future environmental representations with a latent world model, connecting ego-motion understanding with environmental dynamics.
- Extensive experiments demonstrate the superiority of the proposed framework, while ablation studies validate the effectiveness of its components and support the underlying design.

2 Related Work

UAV-centric Spatial Intelligence. Research on UAV spatial intelligence spans navigation and object grounding, alongside broader investigations into spatial perception and motion understanding. AerialVLN extends language-guided navigation to UAVs in city-scale environments, where flight altitude and complex spatial relations introduce additional challenges (Liu et al., 2023b). AerialVG emphasizes positional relations to distinguish visually similar objects in aerial imagery (Liu et al., 2025), while memory-augmented navigation (Zhou et al., 2026) maintains persistent 3D representations for scene understanding and exploration. Recent benchmarks broaden evaluation beyond individual tasks: SpatialSky-Bench (Zhang et al., 2026c) examines spatial understanding for navigation,

SpatialUAV covers perception, collaboration across viewpoints, and motion reasoning (Zhang et al., 2026b), and SIS-Bench evaluates both spatial cognition and self-awareness (Zou et al., 2026). Its accompanying SIS-Motion model combines optical flow with visual features to improve both capabilities. Our framework uses DESM to retain and update spatial evidence across frames, connecting persistent environmental memory with ego-motion inference and future scene prediction to support UAV spatial reasoning.

Latent World Models for UAV Dynamics. Latent world models learn compact states and temporal dynamics for prediction and control, typically conditioning transitions on actions (Ha & Schmidhuber, 2018). PlaNet combines deterministic and stochastic dynamics for online planning in latent space (Hafner et al., 2019), while Dreamer learns behaviors by propagating value gradients through imagined trajectories (Hafner et al., 2020). V-JEPA learns video representations by predicting masked features without pixel reconstruction, capturing motion and appearance information without explicitly modeling the observer's motion (Bardes et al., 2024). Related geometric approaches learn depth and camera motion through view synthesis (Zhou et al., 2017), while VLM-3R incorporates reconstruction features into vision–language reasoning (Fan et al., 2026). For UAV visual reasoning, our EMDM infers ego-motion from common geometric changes across matched regions and conditions its LWM on this motion to predict future environmental representations.

3 METHOD

As illustrated in Figure 2, given visual observations and a textual question q, DESM first constructs and progressively updates spatial memory cells that encode regional appearance and geometric information. EMDM then establishes region-level correspondences across adjacent frames and estimates UAV ego-motion from their shared geometric transformations. The estimated motion further conditions LWM to predict future environmental representations. Finally, QGMI retrieves question-relevant evidence from the environmental, motion, and future memories and injects the aggregated information into the language decoder for spatial reasoning.

3.1 DYNAMIC ENVIRONMENT SPATIAL MEMORY

DESM maintains an environmental memory that jointly encodes regional appearance and spatial information. We first learn spatially grounded memory representations from individual images and then transfer the learned parameters to the video stage, where temporal updates preserve, refine, and integrate environmental evidence across frames.

Spatial Memory Learning from Images. Let $\mathcal{X} = (I_1, \ldots, I_F)$ denote a video of F ordered frames, with $F = 1$ for a single image. The visual encoder extracts features from frame I_t, which are then projected into visual tokens $\mathbf{V}_t \in \mathbb{R}^{N_t \times d}$, where N_t is the number of visual tokens and d is the feature dimension. We introduce K memory cells, each initialized from a learnable content vector $\mathbf{m}_k \in \mathbb{R}^d$ and a normalized box prior $\mathbf{b}_k \in [0,1]^4$, for $k = 1, \ldots, K$. The content vector provides an initial feature representation, while the box prior supplies a spatial reference in the image plane. A geometry encoder $\phi : [0,1]^4 \to \mathbb{R}^d$ maps each box prior to an embedding using Fourier features and a learned projection. We combine the content and geometry through a feature gate:

$$\mathbf{s}_k = \mathrm{LN}(\mathbf{m}_k + \sigma(\mathbf{W}_g[\mathbf{m}_k; \phi(\mathbf{b}_k)]) \odot \phi(\mathbf{b}_k)), \tag{1}$$

where $\mathbf{W}_g \in \mathbb{R}^{d \times 2d}$ is a learned projection, σ is the sigmoid function, $[\cdot;\cdot]$ denotes feature concatenation, $\odot$ denotes elementwise multiplication and LN denotes layer normalization. Stacking the transposed cell vectors and box priors gives $\mathbf{S}_{\text{init}} \in \mathbb{R}^{K \times d}$ and $\mathbf{B}_{\text{init}} \in [0,1]^{K \times 4}$, respectively. These priors and the content vectors are learned during image pretraining and shared across inputs.

We next jointly encode the memory cells $\mathbf{S}_{\text{init}}$ with image tokens $\mathbf{V}_1$ and question q, allowing the cells to collect regional visual evidence. The hidden states at the memory positions are then refined by a low-rank mixer (Sun et al., 2025), which exchanges information among the K cells through a smaller set of latent vectors. The resulting image-conditioned cells form $\mathbf{S}_{\text{img}} \in \mathbb{R}^{K \times d}$. Spatial supervision associates these cells with region boxes, descriptive phrases, and spatial relations, encouraging their features to capture both appearance and geometry. Details of the low-rank mixer and spatial training objectives are provided in Appendix A.1.

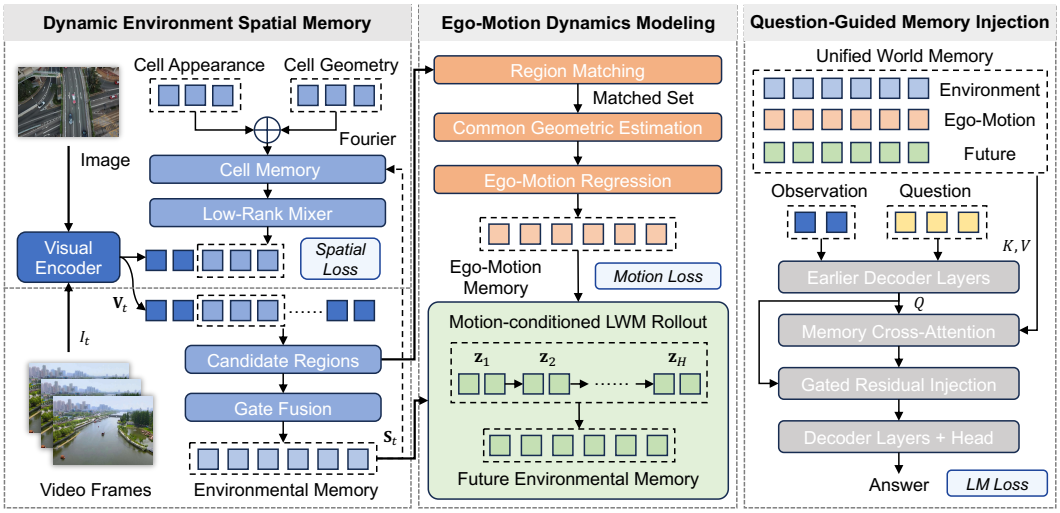

Figure 2: Overview of EnvSelf-UAV, comprising three components: DESM, EMDM, and QGMI.

Dynamic Memory Updates. We transfer the image-pretrained backbone and spatial memory parameters to video adaptation and introduce temporal gates to regulate memory updates. At each frame $t \in \{1, \ldots, F\}$, the previous memory queries the current visual tokens, and the low-rank mixer combines existing information with the retrieved evidence:

$$\bar{\mathbf{S}}_t = \mathrm{Mix}(\mathbf{S}_{t-1} + \mathrm{Attn}(\mathbf{S}_{t-1}, \mathbf{V}_t, \mathbf{V}_t)), \tag{2}$$

where each video starts from $\mathbf{S}_0 = \mathbf{S}_{\text{init}}$ and $\mathbf{B}_0 = \mathbf{B}_{\text{init}}$, constructed from the learned spatial memory parameters. Attn takes queries, keys, and values in that order, while Mix exchanges information among cells and preserves the $K \times d$ shape. An appearance gate compares the previous and candidate states to control how much new information each cell incorporates:

$$\boldsymbol{\alpha}_t = \sigma\big(f_u([\mathbf{S}_{t-1}; \bar{\mathbf{S}}_t; \bar{\mathbf{S}}_t - \mathbf{S}_{t-1}; \bar{\mathbf{S}}_t \odot \mathbf{S}_{t-1}])\big), \tag{3}$$

where $f_u : \mathbb{R}^{4d} \to \mathbb{R}$ operates independently on each cell, producing $\boldsymbol{\alpha}_t \in (0,1)^{K\times 1}$. A separate learned geometry gate $\boldsymbol{\beta}_t \in (0,1)^{K\times 1}$ regulates box updates. The memory evolves as follows:

$$\begin{aligned} \widetilde{\mathbf{S}}_t &= (1 - \boldsymbol{\alpha}_t) \odot \mathbf{S}_{t-1} + \boldsymbol{\alpha}_t \odot \bar{\mathbf{S}}_t, \\ \mathbf{B}_t &= (1 - \boldsymbol{\beta}_t) \odot \mathbf{B}_{t-1} + \boldsymbol{\beta}_t \odot h_b(\bar{\mathbf{S}}_t), \\ \mathbf{S}_t &= \mathcal{G}(\widetilde{\mathbf{S}}_t, \mathbf{B}_t), \end{aligned} \tag{4}$$

where $h_b : \mathbb{R}^d \to [0,1]^4$ predicts a normalized region box for each candidate cell. The appearance and geometry gates are broadcast over the d feature channels and four box coordinates, respectively. The geometry writer $\mathcal{G}$ embeds the updated boxes $\mathbf{B}_t$ using ϕ and fuses these embeddings with the intermediate appearance state $\widetilde{\mathbf{S}}_t$, yielding $\mathbf{S}_t \in \mathbb{R}^{K\times d}$. The resulting sequence $\{\mathbf{S}_t\}_{t=1}^F$ provides environmental context for motion inference in EMDM, while $\mathbf{S}_F$ supplies accumulated environmental evidence to QGMI.

3.2 Ego-Motion Dynamics Modeling

EMDM relates temporal changes in environmental memory to the UAV's ego-motion. It first establishes region-level correspondences across adjacent frames and then aggregates their geometric changes to estimate a shared scene transformation. The resulting motion representation is subsequently used by LWM to predict future environmental dynamics.

Environment-aware Motion Reasoning. We reuse the DESM candidate features $\bar{\mathbf{S}}_t$ and box head to obtain $\widetilde{\mathbf{B}}_t = h_b(\bar{\mathbf{S}}_t)$. From row j of $\widetilde{\mathbf{B}}_t$, we extract a center $\mathbf{p}_{t,j} \in \mathbb{R}^2$ and a descriptor $\mathbf{g}_{t,j} \in \mathbb{R}^4$ containing its center and scale information. Feature matching between $\bar{\mathbf{S}}_{t-1}$ and $\bar{\mathbf{S}}_t$ yields correspondences $\mathcal{M}_t$, where (i,j) links candidate regions in frames $t-1$ and t. For $t = 2, \ldots, F$, we

estimate the common geometric change from the matched centers:

$$\widehat{\mathbf{T}}_t = \underset{\mathbf{T}\in\mathbb{R}^{2\times 3}}{\operatorname{argmin}} \sum_{(i,j)\in\mathcal{M}_t} \ell_\kappa\big(\|\mathbf{p}_{t,j} - \mathbf{T}[\mathbf{p}_{t-1,i}; 1]\|_2\big)\,, \tag{5}$$

where $\mathbf{T} \in \mathbb{R}^{2\times 3}$ is the affine transformation matrix to be estimated, and ℓ_κ is a robust penalty that limits inconsistent correspondences. Using $\widehat{\mathbf{T}}_t$ as a common motion reference, we assign each correspondence the weight $w_{t,ij} = \chi_{t,ij}\exp\big(-\|\mathbf{p}_{t,j} - \widehat{\mathbf{T}}_t[\mathbf{p}_{t-1,i}; 1]\|_2^2/(2\tau^2)\big)$, where $\chi_{t,ij} \in (0, 1]$ is the cosine similarity of the retained feature match and $\tau > 0$ controls tolerance to geometric deviations. This weighting favors reliable matches whose observed centers agree with the locations predicted by the common transformation. We aggregate their encoded geometric changes to obtain the motion state:

$$\mathbf{a}_t = f_m\left(\frac{\sum_{(i,j)\in\mathcal{M}_t} w_{t,ij}\psi([\mathbf{g}_{t-1,i}; \mathbf{g}_{t,j} - \mathbf{g}_{t-1,i}])}{\sum_{(i,j)\in\mathcal{M}_t} w_{t,ij} + \epsilon}\right), \tag{6}$$

where $\epsilon > 0$ stabilizes normalization, and ψ and f_m are learned nonlinear maps producing $\mathbf{a}_t \in \mathbb{R}^{d_m}$. Encoding each change together with its spatial origin allows the aggregation to represent translation, rotation, and expansion patterns. Further details, including matching and motion supervision, are provided in Appendix A.2.

Latent World Model. The LWM predicts how the environment may evolve under the inferred motion. We construct its environmental input as $\mathbf{e}_t = [\mathrm{LN}(\mathrm{SetPool}(\mathbf{S}_t)); \mathrm{LN}(\mathbf{v}_t)] \in \mathbb{R}^{2d}$, where $\mathrm{SetPool}$ summarizes the regional memory independently of cell order, and $\mathbf{v}_t = N_t^{-1}\sum_{j=1}^{N_t}(\mathbf{V}_t)_{j,:}^\top$ averages the current visual tokens. Given observations up to frame t, with $2 \le t \le F$, the LWM initializes its world and motion states from $\mathbf{e}_t$ and $\mathbf{a}_t$ and predicts $H \ge 1$ future steps:

$$\begin{aligned}
\mathbf{z}_0 &= \mathbf{P}_z\mathbf{e}_t, \qquad \widehat{\mathbf{a}}_0 = \mathbf{P}_a\mathbf{a}_t,\\
\mathbf{z}_h &= \mathbf{z}_{h-1} + f_z([\mathbf{z}_{h-1}; \widehat{\mathbf{a}}_{h-1}; \mathbf{z}_{h-1}\odot\widehat{\mathbf{a}}_{h-1}])\,,\\
\widehat{\mathbf{a}}_h &= \widehat{\mathbf{a}}_{h-1} + f_a(\widehat{\mathbf{a}}_{h-1}),\\
\widehat{\mathbf{e}}_{t+h} &= \mathbf{e}_t + \mathbf{P}_e(\mathbf{z}_h - \mathbf{z}_0), \qquad h = 1, \dots, H.
\end{aligned} \tag{7}$$

where $\mathbf{P}_z$ and $\mathbf{P}_a$ are learned projections for state initialization, and f_z and f_a predict residual updates to the world and motion states, respectively. The learned projection $\mathbf{P}_e \in \mathbb{R}^{2d\times d_m}$ maps the accumulated world-state change $\mathbf{z}_h - \mathbf{z}_0$ into an environmental feature change. This change is added to $\mathbf{e}_t$ to produce $\widehat{\mathbf{e}}_{t+h}$, anchoring each forecast to the observed environment.

During training, a prefix ending at $2 \le t < F$ provides the predictor inputs, and subsequent frames provide targets without entering the rollout. The prediction objective and supervision conditions are detailed in Appendix A.2. At inference, we set $t = F$ and provide QGMI with the observed motion sequence $\{\mathbf{a}_t\}_{t=2}^F$ and predicted environmental representations $\{\widehat{\mathbf{e}}_{F+h}\}_{h=1}^H$.

3.3 Question-Guided Memory Injection

QGMI unifies the outputs of DESM and EMDM into a shared memory space and retrieves evidence conditioned on the textual question q and the current decoding context.

Unified World Memory. We use $\mathbf{U}_{\mathrm{env}} = \mathbf{S}_F \in \mathbb{R}^{K\times d}$ as the environmental memory. The learned projections $\mathbf{P}_{\mathrm{mot}} \in \mathbb{R}^{d\times d_m}$ and $\mathbf{P}_{\mathrm{fut}} \in \mathbb{R}^{d\times 2d}$ map the observed motion states and predicted environmental representations to the decoder dimension d, respectively. Stacking the projected vectors as token rows yields the motion memory $\mathbf{U}_{\mathrm{mot}} = \big[\mathbf{P}_{\mathrm{mot}}\mathbf{a}_2, \dots, \mathbf{P}_{\mathrm{mot}}\mathbf{a}_F\big]^\top \in \mathbb{R}^{(F-1)\times d}$ and future memory $\mathbf{U}_{\mathrm{fut}} = \big[\mathbf{P}_{\mathrm{fut}}\widehat{\mathbf{e}}_{F+1}, \dots, \mathbf{P}_{\mathrm{fut}}\widehat{\mathbf{e}}_{F+H}\big]^\top \in \mathbb{R}^{H\times d}$. Each row corresponds to one observed motion transition or future prediction step, with temporal order preserved within each memory.

To distinguish observed environmental evidence, inferred motion, and predicted environmental evidence, we add a learned source embedding to each memory before concatenation:

$$\mathbf{W} = \mathrm{cat}_{\mathrm{tok}}(\mathbf{U}_{\mathrm{env}} + \boldsymbol{\tau}_{\mathrm{env}}, \mathbf{U}_{\mathrm{mot}} + \boldsymbol{\tau}_{\mathrm{mot}}, \mathbf{U}_{\mathrm{fut}} + \boldsymbol{\tau}_{\mathrm{fut}})\,, \tag{8}$$

where $\mathrm{cat}_{\mathrm{tok}}$ concatenates matrices along the token dimension. The embeddings $\boldsymbol{\tau}_{\mathrm{env}}$, $\boldsymbol{\tau}_{\mathrm{mot}}$, and $\boldsymbol{\tau}_{\mathrm{fut}}$ are learned d-dimensional vectors identifying different memory sources.

Table 1: Performance on SIS-Bench across spatial cognition and self-awareness tasks. For each column, the best and second-best results are highlighted in bold and underlined, respectively.

Model	Overall	Spatial Cognition								Self-Awareness				
		Obj Exist	*Obj Attr*	*Rel Dir*	*Land Order*	*Land Recall*	*Pos Rel*	*Spat Consist*	*ST Consist*	*Act Recog*	*Act Seq*	*Act Recall*	*Act Pred*	*Path Plan*
Random	25.0	24.7	24.9	24.9	25.2	25.2	25.1	25.0	25.2	24.7	25.3	25.0	24.9	25.2
Human Performance	91.7	95.9	96.6	84.4	97.0	93.8	94.6	88.3	68.8	92.4	96.6	94.1	86.9	83.6
Proprietary Models														
Gemini-3-Flash	71.6	97.2	80.6	75.0	94.4	<u>84.0</u>	**89.7**	71.8	<u>57.7</u>	66.5	**84.1**	42.4	61.6	53.7
Kimi-2.5	71.0	97.0	78.0	<u>80.5</u>	**98.0**	81.9	87.3	**76.4**	53.1	50.9	77.8	53.0	65.4	57.0
Qwen3.5-Plus	70.1	97.6	76.2	78.5	<u>97.7</u>	83.3	87.7	<u>75.9</u>	54.8	52.8	<u>82.5</u>	42.4	**69.2**	<u>58.5</u>
GPT-5.4	70.0	<u>98.4</u>	74.9	67.0	94.8	78.6	86.9	64.1	57.3	57.3	77.8	49.9	65.8	**58.8**
Open-source Models														
SIS-Motion	69.1	96.5	71.1	66.0	93.8	68.2	73.0	60.0	39.8	<u>88.1</u>	81.3	55.9	44.1	23.5
Qwen3-VL-8B-Instruct	64.6	97.2	82.2	73.5	95.8	74.9	82.9	64.6	51.0	55.1	60.3	32.2	57.0	48.9
Qwen3-VL-8B-Thinking	64.7	95.5	76.0	72.5	95.1	78.6	79.8	64.6	55.6	53.4	62.9	33.6	64.6	46.3
Qwen3-VL-4B-Instruct	62.5	96.3	80.4	71.0	93.8	70.7	82.1	73.3	47.3	53.4	61.3	29.1	57.0	36.4
Qwen3.5-4B	61.1	93.5	76.7	66.5	90.5	67.5	79.0	68.2	46.5	46.5	60.6	38.8	51.3	36.3
InternVL3.5-8B	61.1	96.1	76.5	70.5	86.3	72.0	73.8	55.4	47.3	56.1	52.4	31.8	55.1	41.9
Qwen3-VL-30B-A3B-Instruct	61.0	97.2	78.6	79.0	95.1	76.5	85.3	67.7	47.7	39.7	44.8	28.4	59.7	48.9
GLM-4.1V-9B-Thinking	60.4	92.1	77.5	72.0	93.5	64.1	75.4	63.6	49.8	48.1	55.9	34.8	48.7	43.4
MiMo-VL-7B-RL	59.2	96.3	74.2	73.5	80.1	75.2	81.3	72.3	47.3	44.8	43.2	29.7	51.7	40.8
EnvSelf-UAV-4B (Ours)	<u>76.9</u>	98.0	<u>85.8</u>	78.0	95.1	81.5	86.1	68.7	53.1	<u>88.1</u>	80.3	<u>60.5</u>	64.4	43.7
EnvSelf-UAV-9B (Ours)	**79.9**	**98.5**	**88.3**	**82.0**	95.3	**84.8**	<u>88.0</u>	72.9	**61.5**	**89.6**	81.8	**65.5**	<u>67.9</u>	48.4

Memory Retrieval and Injection. Let $\mathbf{h}_n \in \mathbb{R}^d$ be the intermediate decoder state used to predict output token y_n. Conditioned on the observations, question q, and preceding output tokens $y_{<n}$, this state queries the unified memory through multihead cross-attention:

$$\begin{aligned} \mathbf{c}_n &= \mathrm{Read}(\mathbf{h}_n, \mathbf{W}), \\ \widetilde{\mathbf{h}}_n &= \mathbf{h}_n + g_n \mathbf{W}_O \mathbf{c}_n, \end{aligned} \tag{9}$$

where the operator Read uses $\mathbf{h}_n$ as the query and $\mathbf{W}$ to construct keys and values, returning the retrieved evidence $\mathbf{c}_n \in \mathbb{R}^d$. A learned scalar gate $g_n \in (0, 1)$, computed from $\mathbf{h}_n$ and $\mathbf{c}_n$, controls the contribution of this evidence. The learned projection $\mathbf{W}_O \in \mathbb{R}^{d \times d}$ maps it into a decoder update.

The updated state $\widetilde{\mathbf{h}}_n$ passes through the remaining decoder layers to predict y_n. QGMI is trained through the language modeling loss. Retrieval details, the gating formulation, and the joint training objective are provided in Appendix A.3.

4 Experiments

4.1 Experimental Details

We evaluated our framework on three representative benchmarks, SIS-Bench (Zou et al., 2026), SpatialUAV (Zhang et al., 2026b), and OpenUAV-QA (Wang et al., 2025b). SIS-Bench contains 4,856 multiple-choice questions from 1,646 UAV videos across 13 tasks, assessing spatial cognition and self-awareness at the levels of perception, memory, and reasoning. SpatialUAV comprises 4,331 instances across 14 tasks involving single images, multiple viewpoints, and videos, with a focus on spatial grounding, reasoning across aerial and ground views, and motion understanding in low-altitude environments. OpenUAV-QA contains 3,895 multiple-choice test instances derived from simulated UAV flight trajectories, assessing navigation decision-making and action sequence reasoning from egocentric videos. Further evaluation protocols, baseline descriptions, and implementation details are provided in Appendix B.

4.2 Main Results

Q1: Does the unified framework improve overall UAV spatial intelligence? Table 1 shows that EnvSelf-UAV improves both spatial cognition and self-awareness. The 4B variant achieves 76.9% overall accuracy, exceeding its backbone Qwen3.5-4B by 15.8%. Landmark recall improves by 14.0%, while action recognition and prediction increase by 41.6% and 13.1%, respectively. These gains cover environmental memory, understanding of observed motion, and anticipation of future

Table 2: Zero-shot performance on SpatialUAV. For each column, the best and second-best values are highlighted in bold and underlined, respectively.

Method	Avg.	Semantic		Spatial		Aerial–Aerial					Aerial–Ground				Motion
		Reg Recog	*Ano Det*	*Dir Recog*	*Dist Comp*	*Collab Recog*	*Shared Assoc*	*Obj Match*	*Cam Trans*	*Occl Remov*	*Shared Assoc*	*Collab Recog*	*View Trans*	*Path Plan*	*Global Motion*
Human Level	89.0	99.3	100.0	95.4	97.6	99.5	92.9	85.6	40.2	98.3	92.7	97.8	98.2	96.4	89.1
Random Choice	14.4	8.0	17.7	7.5	9.2	26.9	0.8	1.4	0.9	25.4	1.0	22.6	5.3	14.6	24.1
Proprietary Models															
GPT-5.4	**56.7**	56.7	64.7	50.6	53.6	**92.6**	**32.8**	**28.8**	**8.2**	**49.2**	**44.2**	79.8	54.0	45.8	**66.6**
Claude-Opus-4-7	51.3	71.3	64.7	39.1	60.2	85.7	26.2	22.2	5.4	22.2	0.0	**80.4**	0.0	51.4	51.9
Gemini-3.1-Flash	45.9	72.1	**76.5**	53.6	64.4	46.9	31.6	6.1	7.2	36.5	41.2	65.2	60.5	**54.9**	37.6
Spatial-specific Models															
VST-7B-SFT	29.7	47.1	8.8	9.9	47.5	15.5	2.4	3.8	7.7	46.0	2.1	23.4	10.5	20.8	54.4
SpaceEra++	28.2	43.6	6.1	8.7	44.3	41.7	6.2	6.3	3.3	44.4	1.6	43.1	13.8	19.7	31.9
SpaceEra	27.3	42.8	6.4	8.1	43.6	40.8	5.8	6.7	3.0	43.7	1.3	42.2	12.9	20.2	30.1
SpatialVLM	25.7	30.8	5.9	8.4	32.6	26.7	5.1	9.0	8.0	20.6	8.4	24.6	38.2	4.2	46.3
Cambrian-S-7B	24.1	11.4	11.8	12.8	40.2	24.3	1.1	1.4	0.0	23.8	1.7	26.4	10.5	19.4	42.7
SenseNova-SI-1.3-8B	21.0	50.0	5.9	6.4	36.8	26.0	6.6	8.5	1.0	41.3	0.9	31.0	7.9	0.0	16.6
Open-source Models															
Qwen3.6-27B	49.5	66.2	**76.5**	**60.4**	**69.7**	61.9	24.7	9.0	7.4	33.3	36.5	58.2	47.4	**54.9**	50.5
Qwen3.5-9B	42.5	70.0	64.7	51.0	66.3	37.9	11.1	7.1	6.2	33.3	21.7	37.0	64.5	36.1	48.4
Qwen3.6-35B-A3B	37.3	54.6	17.7	51.9	65.5	19.5	16.8	10.4	0.9	30.2	30.1	34.8	50.0	28.5	47.7
Qwen3.5-4B	36.0	62.3	35.3	19.1	61.3	32.9	12.6	8.0	6.9	33.3	27.3	33.2	53.9	34.7	45.3
InternVL3.5-8B	30.8	44.6	11.8	13.9	50.6	26.2	6.2	6.1	7.4	20.6	9.9	35.4	22.4	6.9	49.1
InternVL3.5-14B	30.4	39.0	5.9	22.6	49.4	21.9	4.6	5.2	4.7	22.2	7.5	18.6	18.4	34.0	55.3
EnvSelf-UAV-4B (Ours)	49.8	77.8	47.0	40.6	66.8	61.8	14.8	10.3	2.6	33.3	26.2	55.1	73.7	52.1	58.6
EnvSelf-UAV-9B (Ours)	53.9	**82.9**	66.0	60.1	68.3	66.3	13.6	9.6	1.8	32.5	21.7	57.5	**80.7**	52.8	61.3

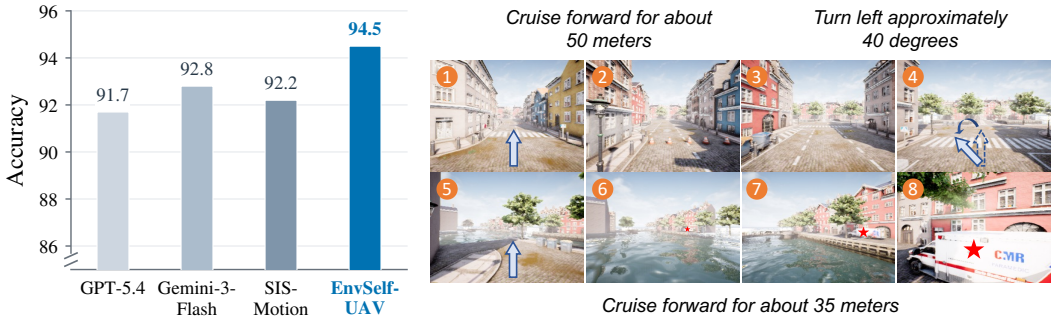

Figure 3: Navigation performance on OpenUAV-QA. Accuracy comparison of different methods (Left). A qualitative example of the navigation decision generated by our method (Right).

actions. Besides, EnvSelf-UAV achieves the highest overall accuracy among evaluated models of comparable size and also outperforms larger open-source baselines while using fewer parameters.

Q2: Can the framework transfer to other tasks in a zero-shot setting? Table 2 reports the zero-shot results on SpatialUAV, where EnvSelf-UAV-9B achieves an average score of 53.9, the highest among the evaluated non-proprietary models. Compared with Qwen3.5-9B, aerial–aerial collaboration recognition improves by 28.4%, view translation by 16.2%, and global motion understanding by 12.9%. These improvements span collaboration across viewpoints and motion description, supporting transfer across different task formulations and input configurations. However, camera transformation remains weak, indicating that transfer to precise geometric reasoning is still limited.

Q3: Does the framework remain effective across backbone scales? Scaling EnvSelf-UAV from 4B to 9B improves SIS-Bench accuracy by 3.0% and the SpatialUAV average by 4.1% (Tables 1 and 2). On SpatialUAV, the gains over the corresponding Qwen3.5 backbones are 13.8% at 4B and 11.4% at 9B. The framework therefore provides substantial benefits at both evaluated scales, while larger capacity further improves overall performance. The gains from increasing backbone size vary across tasks.

Q4: Can the framework improve downstream UAV navigation? Figure 3 presents quantitative and qualitative results on OpenUAV-QA. In the left panel, EnvSelf-UAV achieves the highest accuracy of 94.5%, exceeding SIS-Motion by 2.3% and outperforming both evaluated proprietary models. The right panel illustrates a navigation decision that connects successive observations to an ordered action sequence: *cruising forward*, *turning left*, and *cruising forward toward the tar-*

Table 3: Ablation study with Qwen3.5-4B. Modules are progressively added to the SFT baseline to assess their incremental contributions.

Variant	SIS-Bench					SpatialUAV
	Overall	Landmark Recall	Positional Relationship	Action Recall	Path Planning	Avg.
Qwen3.5-4B	61.1	67.5	79.0	38.8	36.3	36.0
Qwen3.5-4B (SFT)	66.9	72.0	81.0	42.2	35.2	44.1
+ DESM (Image only)	71.6	77.6	84.2	48.8	29.6	47.2
+ DESM	74.7	**81.9**	85.9	56.0	30.4	47.4
+ EMDM (Ego-motion only)	75.3	81.5	86.0	58.3	35.7	48.5
+ EMDM	76.6	80.8	**86.5**	**60.7**	42.1	48.9
+ QGMI (Ours)	**76.9**	81.5	86.1	60.5	**43.7**	**49.8**

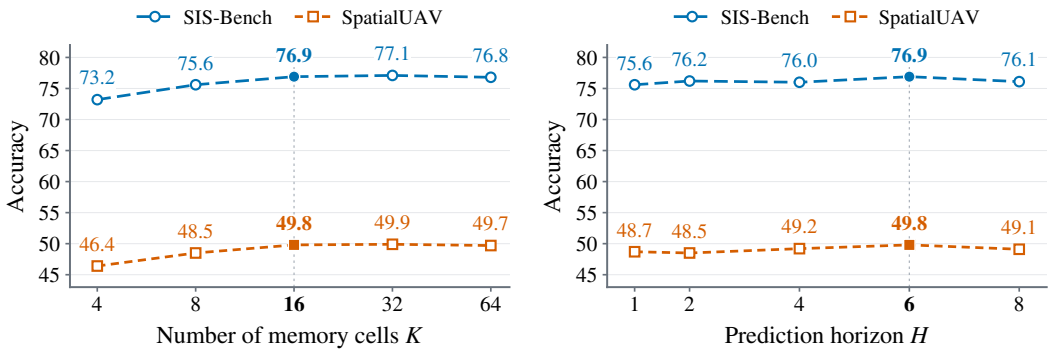

Figure 4: Parameter sensitivity to environmental memory capacity K (Left) and future prediction horizon H (Right).

get. The sequence combines forward movement with heading adjustment based on changing visual observations, illustrating how environmental and motion understanding can support multi-step navigation toward a spatial goal. Together, these comparisons support the framework's applicability to downstream UAV navigation decision-making.

4.3 Ablation Study

Table 3 shows consistent aggregate improvements on both benchmarks as components are progressively added. **(1) SFT provides effective task adaptation.** Fine-tuning Qwen3.5-4B on SIS-Motion-54K raises SIS-Bench overall accuracy from 61.1% to 66.9% and SpatialUAV average from 36.0% to 44.1%, providing a stronger task-adapted baseline for subsequent component analysis. **(2) DESM strengthens spatial cognition.** Compared with SFT, full DESM improves landmark recall and positional relationship accuracy by 9.9% and 4.9%, respectively. Extending DESM from image-only training to temporal adaptation further increases overall accuracy by 3.1%, supporting the value of maintaining environmental evidence across frames. **(3) Ego-motion inference enhances self-awareness.** Adding the ego-motion component of EMDM to DESM improves action recall by 2.3% and path planning by 5.3%, indicating that explicit motion modeling complements environmental memory. **(4) LWM further benefits prediction and planning.** Completing EMDM with the LWM raises action recall by another 2.4% and path planning by 6.4%. The larger planning gain suggests that predicted environmental representations provide useful evidence for reasoning about future actions. **(5) QGMI improves evidence integration.** Adding QGMI increases SIS-Bench overall accuracy by 0.3%, path planning by 1.6%, and the SpatialUAV average by 0.9%. These incremental gains support selectively incorporating complementary memories into decoding.

4.4 Parameter Analysis

We analyzed the number of environmental memory cells K and the prediction horizon H, which control spatial memory capacity and the extent of future environmental prediction, respectively. Figure 4 presents the corresponding curves on SIS-Bench and SpatialUAV. **(1) Memory capacity.** The left panel of Figure 4 shows that both curves rise steadily as K increases from 4 to 16. Increasing



Figure 5: Qualitative comparisons on SIS-Bench. Underlined options indicate reference answers.

K to 32 provides only marginal gains, followed by a slight decline at $K = 64$. This trend suggests diminishing benefits from additional memory cells. We therefore use $K = 16$, which provides performance close to the peak on both benchmarks while keeping the memory compact. **(2) Prediction horizon.** The right panel of Figure 4 shows an overall improvement as H increases toward 6, with small intermediate fluctuations. Both curves peak at $H = 6$ and decline at $H = 8$, indicating that a longer prediction horizon does not consistently improve performance. One possible explanation is that prediction uncertainty accumulates during longer rollouts, offsetting the benefit of additional future context. We therefore selected $H = 6$ as the reference horizon.

4.5 Qualitative Analysis

Figure 5 presents two examples covering action sequence recognition and spatiotemporal consistency. In the **first** example, EnvSelf-UAV identifies the correct sequence (B), including the final forward motion while descending. Gemini-3-Flash and SIS-Motion instead predict hovering at the final stage, while GPT-5.4 reverses the initial lateral direction. This example illustrates successful recognition of compound UAV actions across consecutive clips. The **second** example requires reconstructing a flight path from temporally shuffled clips with overlapping content and identifying a landmark passed along the route. All compared models answer incorrectly. EnvSelf-UAV selects the gray low-rise office building (A), whereas the reference answer is the industrial factory complex with red roofs (D). This error is consistent with confusing a building visible in the scene with one passed along the flight path. It highlights a remaining challenge in relating landmark observations to the UAV's trajectory, beyond recognizing their visual appearance.

5 Conclusion

We presented *EnvSelf-UAV*, a unified framework that connects external spatial perception with internal motion understanding for holistic UAV spatial intelligence. DESM maintains spatially grounded environmental memory across changing viewpoints, while EMDM infers ego-motion from common geometric changes across matched regions and predicts future environmental representations through a latent world model. QGMI selectively integrates these complementary sources of evidence into language decoding. Experiments show improved spatial cognition and self-awareness on SIS-Bench, zero-shot transfer on SpatialUAV, and navigation decision-making on OpenUAV-QA. These findings support jointly modeling the environment and ego-motion for UAV spatial reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, mathematical formula verification, and statistical aggregation of benchmark scores. We have not used generative AI tools for other tasks requiring disclosure, and the remaining disclosure categories are not applicable to this work. All AI-assisted content was reviewed and verified by the authors, including the accuracy of reported results and the correctness of the manuscript. We take full responsibility for the accuracy, originality, and final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A present the model architecture, mathematical formulations, and training objectives. Section 4 and Appendix B provide details on the datasets, data preparation, evaluation protocols, model configurations, training hyperparameters, and inference settings. Upon acceptance, we will publicly release our source code to further facilitate reproducibility and support future research.

REFERENCES

Anthropic. Claude opus 4.7 system card. Technical report, Anthropic, 2026. URL https://anthropic.com/claude-opus-4-7-system-card.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/pdf/2404.08471.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cai_Scaling_Spatial_Intelligence_with_Multimodal_Foundation_Models_CVPR_2026_paper.html.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

Google DeepMind. Gemini 3 flash model card. Technical report, Google DeepMind, 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fan_VLM-3R_Vision-Language_Models_Augmented_with_Instruction-Aligned_3D_Reconstruction_CVPR_2026_paper.html.

Weili Guan, Haoyu Zhang, Meng Liu, Qianlong Xiang, Yaowei Wang, and Liqiang Nie. Spaceera++: A unified framework towards 3d spatial reasoning in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2026. URL https://doi.org/10.1109/TPAMI.2026.3692302.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024. URL https://aclanthology.org/2024.emnlp-main.342/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. Aerialvg: A challenging benchmark for aerial visual grounding by exploring positional relations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5177–5187. IEEE, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_AerialVG_A_Challenging_Benchmark_for_Aerial_Visual_Grounding_by_Exploring_ICCV_2025_paper.html.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15338–15348. IEEE, 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/html/Liu_AerialVLN_Vision-and-Language_Navigation_for_UAVs_ICCV_2023_paper.html.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. URL https://arxiv.org/pdf/2601.03267.

Mengyang Sun, Yihao Wang, Tao Feng, Dan Zhang, Yifan Zhu, and Jie Tang. A stronger mixture of low-rank experts for fine-tuning foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 57712–57727. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25s.html.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026. URL https://arxiv.org/abs/2602.02276.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a. URL https://arxiv.org/abs/2508.18265.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Hongsheng Li, Yue Liao, Si Liu, et al. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. In *International Conference on Learning Representations*, volume 2025, pp. 7292–7310, 2025b. URL https://openreview.net/forum?id=rUvCIvI4eB.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. In *European Conference on Computer Vision*, pp. 192–211. Springer, 2026a. URL https://arxiv.org/abs/2511.05491.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 2026b. URL https://arxiv.org/abs/2511.04670.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pp. 57153–57180, 2024. URL https://openreview.net/forum?id=2msbbX3ydD.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025. URL https://arxiv.org/abs/2506.03569.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data. *Advances in Neural Information Processing Systems*, 38:103202–103229, 2026a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9541101fbc2f24bce5f2462b95db88c4-Abstract-Conference.html.

Haoyu Zhang, Meng Liu, Qianlong Xiang, Kun Wang, Yaowei Wang, and Liqiang Nie. Spatialuav: Benchmarking spatial intelligence for low-altitude uav perception, collaboration, and motion. *arXiv preprint arXiv:2606.27876*, 2026b. URL https://arxiv.org/abs/2606.27876.

Lingfeng Zhang, Yuchen Zhang, Hongsheng Li, Haoxiang Fu, Yingbo Tang, Hangjun Ye, Long Chen, Xiaojun Liang, Xiaoshuai Hao, and Wenbo Ding. Is your vlm sky-ready? a comprehensive spatial intelligence benchmark for uav navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25904–25913, 2026c. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Is_your_VLM_Sky-Ready_A_Comprehensive_Spatial_Intelligence_Benchmark_for_CVPR_2026_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.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6612–6619. IEEE, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Unsupervised_Learning_of_CVPR_2017_paper.html.

Zhishan Zou, Guoyan Sun, Zhiwei Wei, Jiancheng Pan, Yujie Li, Mugen Peng, and Wenjia Xu. Self in space: Benchmarking self-awareness and spatial cognition in uav embodied intelligence. *arXiv preprint arXiv:2607.12477*, 2026. URL https://arxiv.org/abs/2607.12477.

A Additional Method Details

A.1 Dynamic Environment Spatial Memory

Memory Construction and Mixing. The content vectors and box priors in Eq. (1) are shared learnable parameters, whereas the resulting image and video memories depend on the observations. During image pretraining, the memory-position hidden states are refined by the same low-rank mixer used in Eq. (2). For an input $\mathbf{M} \in \mathbb{R}^{K\times d}$ and $R < K$ learnable latent vectors $\mathbf{L} \in \mathbb{R}^{R\times d}$, the mixer is as follows:

$$\begin{aligned} \mathbf{Z} &= \mathrm{Attn}_{\mathrm{agg}}(\mathbf{L}, \mathbf{M}, \mathbf{M}), \\ \mathrm{Mix}(\mathbf{M}) &= \mathbf{M} + \mathrm{Attn}_{\mathrm{dist}}(\mathbf{M}, \mathbf{Z}, \mathbf{Z}), \end{aligned} \tag{10}$$

where two attention operators $\mathrm{Attn}_{\mathrm{agg}}$ and $\mathrm{Attn}_{\mathrm{dist}}$ have separate learned projections. The intermediate $\mathbf{Z} \in \mathbb{R}^{R\times d}$ allows cells to exchange information through a compact bottleneck while preserving the $K \times d$ output shape.

Spatial Supervision. Soft assignments associate image-conditioned cells with annotated regions. The spatial objective combines box, phrase, and relation supervision:

$$\mathcal{L}_{\mathrm{spa}} = \lambda_b \mathcal{L}_{\mathrm{box}} + \lambda_p \mathcal{L}_{\mathrm{phrase}} + \lambda_r \mathcal{L}_{\mathrm{rel}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{box}}$ supervises the shared box head h_b, $\mathcal{L}_{\mathrm{phrase}}$ aligns cell features with the assigned region descriptions, and $\mathcal{L}_{\mathrm{rel}}$ supervises spatial relations between assigned region pairs. Box priors, predicted boxes, and target boxes use the same normalized coordinate convention. Image pretraining minimizes $\mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{spa}}\mathcal{L}_{\mathrm{spa}}$, where the language loss is defined in Eq. (17).

Transfer to Video. Every video starts from the learned $\mathbf{S}_{\mathrm{init}}$ and $\mathbf{B}_{\mathrm{init}}$, and memory is not carried between videos. At each frame, Eq. (2) produces candidate features $\bar{\mathbf{S}}_t$. The appearance and geometry gates in Eq. (4) separately control feature and box updates, after which $\mathcal{G}$ fuses the updated appearance with box embeddings from ϕ. Thus, $\mathbf{S}_t$ retains accumulated environmental evidence. EMDM uses the candidate features $\bar{\mathbf{S}}_t$ and candidate boxes $h_b(\bar{\mathbf{S}}_t)$ for adjacent-frame matching and geometric changes, while QGMI uses the accumulated memory $\mathbf{S}_F$.

A.2 Ego-Motion Dynamics Modeling

Region Correspondence and Motion Aggregation. For each candidate box $h_b(\bar{\mathbf{S}}_t)_{j,:}$, let $\mathbf{p}_{t,j} \in \mathbb{R}^2$ be its center and $\mathbf{u}_{t,j} \in \mathbb{R}^2_{>0}$ its width–height vector. In normalized corner coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$, these are $\mathbf{p}_{t,j} = ((x_{\min} + x_{\max})/2, (y_{\min} + y_{\max})/2)^\top$ and $\mathbf{u}_{t,j} = (x_{\max} - x_{\min}, y_{\max} - y_{\min})^\top$. The geometric descriptor is $\mathbf{g}_{t,j} = [\mathbf{p}_{t,j}; \log \mathbf{u}_{t,j}] \in \mathbb{R}^4$, with an element-wise logarithm. Descriptor differences therefore encode center displacement and log scale ratios. Invalid or degenerate boxes and near-zero feature vectors are excluded before taking logarithms or computing cosine similarities. We retain mutual nearest neighbors whose cosine similarity exceeds a positive threshold η and exceeds the next-best alternative by a margin Δ in both directions. These matches form $\mathcal{M}_t$.

The robust penalty in Eq. (5) is $\ell_\kappa(r) = \min(r^2, \kappa^2)$. A consensus-based affine fit identifies a common transformation supported by spatially distributed correspondences. The consistency weights in Eq. (6) then reduce the contribution of matches that disagree with this transformation. The encoder $\psi : \mathbb{R}^8 \to \mathbb{R}^{d_m}$ receives each region's previous descriptor and descriptor change, and f_m refines their weighted aggregate. The affine transformation summarizes common image-plane motion. The learned state $\mathbf{a}_t$ is a motion representation rather than a directly recovered metric camera pose.

Fallback and Motion Supervision. If the correspondences do not support a valid affine fit, or their total consistency weight is too small, motion is inferred from the global frame summaries:

$$\mathbf{a}_t = \phi_g([\mathbf{v}_{t-1}; \mathbf{v}_t - \mathbf{v}_{t-1}]), \qquad \phi_g : \mathbb{R}^{2d} \to \mathbb{R}^{d_m}. \tag{12}$$

This fallback is also sampled during training so that it receives supervision. Matching, affine fitting, and consistency weights are computed without gradients. The selected geometric descriptors remain differentiable through h_b.

An auxiliary classifier uses a question-conditioned temporal summary $\bar{\mathbf{a}} = \sum_{t=2}^{F} \pi_t(q)\mathbf{a}_t$, where the nonnegative attention weights sum to one. Its objective is:

$$\mathbf{p}^m = \sigma(\mathbf{W}_m \bar{\mathbf{a}}), \qquad \mathcal{L}_{\text{mot}} = \text{BCE}(\mathbf{p}^m, \mathbf{y}^m), \tag{13}$$

where $\mathbf{W}_m \in \mathbb{R}^{C \times d_m}$ predicts C motion classes, $\mathbf{y}^m \in \{0,1\}^C$ contains labels extracted from training reference text, and $\text{BCE}_{\mathbf{w}}$ is binary cross-entropy. This loss is used only when valid labels are available.

Future Environmental Prediction. The LWM operates on the environmental descriptor $\mathbf{e}_t \in \mathbb{R}^{2d}$ and motion state $\mathbf{a}_t$. Its world and motion states both have dimension d_m, making the element-wise interaction in Eq. (7) well-defined. Each predicted descriptor is anchored to $\mathbf{e}_t$ through a learned feature change.

For training, an observed prefix ending at $2 \le t < F$ provides $H_t = \min(H, F - t)$ future targets. Define $\widehat{\boldsymbol{\delta}}_{t,h} = \widehat{\mathbf{e}}_{t+h} - \mathbf{e}_t$ and $\boldsymbol{\delta}_{t,h} = \text{sg}(\mathbf{e}_{t+h} - \mathbf{e}_t)$, where sg stops gradients. We supervise the predicted changes as follows:

$$\mathcal{L}_{\text{LWM}} = \frac{\sum_{h=1}^{H_t} \rho^{h-1} \,\text{SmoothL1}\left(\widehat{\boldsymbol{\delta}}_{t,h}, \boldsymbol{\delta}_{t,h}\right)}{\sum_{h=1}^{H_t} \rho^{h-1}}, \tag{14}$$

where SmoothL1 averages over the $2d$ coordinates and $\rho \in (0,1]$ discounts more distant targets. Future descriptors are obtained from subsequent observations for supervision only. The predictor's visual features, memory, and motion state use only the prefix. This loss is omitted when no eligible prefix exists.

A.3 Question-Guided Memory Injection and Joint Training

Memory Reading. For videos with $F \ge 2$, the memory in Eq. (8) contains $n_{\text{env}} = K$ environmental tokens, $n_{\text{mot}} = F - 1$ observed motion tokens, and $n_{\text{fut}} = H$ predicted environmental tokens. Source embeddings distinguish these three types of evidence. Each attention head has width $d_a = d/J$, where J is the number of heads and divides d. Let $\mathbf{q}_n^{(\ell)}$ be the projected decoder query, and $\mathbf{k}_j^{(\ell)}$ and $\boldsymbol{\nu}_j^{(\ell)}$ the projected key and value of memory token j. Head ℓ retrieves as follows:

$$\begin{aligned} \omega_{nj}^{(\ell)} &= \text{softmax}_j\left(\frac{(\mathbf{q}_n^{(\ell)})^\top \mathbf{k}_j^{(\ell)}}{\sqrt{d_a}} + b_{r(j)} - \log n_{r(j)}\right), \\ \mathbf{c}_n^{(\ell)} &= \sum_j \omega_{nj}^{(\ell)} \boldsymbol{\nu}_j^{(\ell)}, \qquad \mathbf{c}_n = [\mathbf{c}_n^{(1)}; \ldots; \mathbf{c}_n^{(J)}], \end{aligned} \tag{15}$$

where $r(j)$ identifies the source of token j, and b_r is a learned source bias shared across heads. The $-\log n_r$ term compensates for differing source lengths when token logits are equal. A NULL entry, with a learned key, zero value, and source count one, absorbs attention without contributing to the retrieved value. The softmax and sum include the real memory tokens and NULL, with padding masked. The query inherits the question and preceding answer context from the decoder state $\mathbf{h}_n$.

Gated Injection and Causal Decoding. The scalar gate in Eq. (9) is:

$$g_n = \sigma(f_{\text{inj}}([\mathbf{h}_n; \mathbf{c}_n; \mathbf{h}_n \odot \mathbf{c}_n])), \qquad f_{\text{inj}} : \mathbb{R}^{3d} \to \mathbb{R}. \tag{16}$$

Memory is constructed from the observations and question, excluding target answer tokens, and its keys and values are reused during decoding. Under teacher forcing, the state predicting y_n is taken from the preceding decoder position. For y_1, this is the final input-prefix position. The injected state then passes through the remaining decoder layers.

Training Objectives. For reference text $y = (y_1, \ldots, y_T)$, answer generation uses the autoregressive loss as follows:

$$\mathcal{L}_{\mathrm{LM}} = -\sum_{n=1}^{T} \log p(y_n \mid y_{<n}, \mathcal{X}, q). \tag{17}$$

After spatial image pretraining, video adaptation jointly trains temporal memory, motion inference, future prediction, and memory injection as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \mathbb{I}_{\mathrm{mot}} \lambda_{\mathrm{mot}} \mathcal{L}_{\mathrm{mot}} + \mathbb{I}_{\mathrm{pred}} \lambda_{\mathrm{world}} \mathcal{L}_{\mathrm{LWM}}, \tag{18}$$

where $\mathbb{I}_{\mathrm{mot}}$ indicates available motion labels, and $\mathbb{I}_{\mathrm{pred}}$ indicates an eligible prefix with future targets. Unavailable auxiliary terms are omitted. QGMI is optimized through $\mathcal{L}_{\mathrm{LM}}$, while the auxiliary objectives provide explicit supervision for motion and environmental prediction.

B Additional Experimental Details

B.1 Evaluation Benchmarks

SIS-Bench. SIS-Bench (Zou et al., 2026) evaluates UAV spatial cognition and self-awareness through 4,856 multiple-choice questions from 1,646 videos. Its 13 tasks cover perception, memory, and reasoning. The eight spatial cognition tasks are object existence, object attribute, relative direction, landmark order, landmark recall, positional relationship, spatial consistency, and spatiotemporal consistency. The five self-awareness tasks are action recognition, action sequence, action recall, action prediction, and path planning. Together, these tasks assess whether a model can recognize the environment, retain spatial evidence across observations, and understand the UAV's own motion. We extracted one of the four option labels from each response and compute exact-match accuracy, reporting both task-level results and overall accuracy across all evaluation examples in Table 1.

SpatialUAV. SpatialUAV (Zhang et al., 2026b) contains 4,331 instances across 14 tasks, with inputs ranging from single images and multiple viewpoints to videos. It covers five groups: semantic discrimination (region recognition and anomaly detection), spatial relations (direction recognition and distance comparison), aerial–aerial collaboration (collaboration recognition, shared association, object matching, camera transformation, and occlusion removal), aerial–ground collaboration (shared association, collaboration recognition, view translation, and path planning), and motion understanding (global motion). We evaluated in the zero-shot setting and retain the benchmark's task-specific input formats, answer formats, and evaluators. Table 2 reports task scores and the benchmark's aggregate score on a 0–100 scale.

OpenUAV-QA. OpenUAV-QA contains 3,895 multiple-choice test instances derived from simulated UAV flight trajectories in OpenUAV (Wang et al., 2025b). Given egocentric video observations, a model answers questions about navigation decisions and action sequences. The benchmark complements the spatial and motion evaluations above by testing whether observations can support reasoning about ordered flight maneuvers and navigation toward a goal. We evaluated the selected answer against the reference option and report accuracy in Figure 3.

B.2 Comparison Baselines

Proprietary Models. The proprietary model comparisons include GPT-5.4 (Singh et al., 2025), Gemini-3-Flash (DeepMind, 2025), Gemini-3.1-Flash, Claude-Opus-4-7 (Anthropic, 2026), Kimi-2.5 (Team et al., 2026), and Qwen3.5-Plus[2], as listed in the main result tables. These models provide references for the spatial and motion reasoning capabilities of general-purpose commercial multimodal systems. The evaluated variants differ across benchmarks. Their corresponding results are reported in Tables 1 and 2.

Open-source Models. The open-source comparisons include Qwen3-VL (Bai et al., 2025), Qwen3.5, Qwen3.6[3], InternVL3.5 (Wang et al., 2025a), GLM-4.1V (Hong et al., 2025), and MiMo-VL (Yue et al., 2025) variants. They cover different backbone scales and instruction-tuned or

[2]Official Qwen3.5 family release: `https://qwen.ai/blog?id=qwen3.5`.

[3]Official Qwen3.6 model collection: `https://huggingface.co/collections/Qwen/qwen36`.

reasoning-oriented configurations. Qwen3.5-4B and Qwen3.5-9B are also the corresponding backbones of EnvSelf-UAV, allowing comparison at both evaluated scales. SIS-Motion is included as an open-source UAV baseline that combines optical flow with visual features for spatial cognition and self-awareness. The Qwen3.5-4B SFT baseline and progressive component comparisons in Table 3 further assess image-based spatial memory, temporal updates, ego-motion inference, future prediction, and QGMI.

Spatial-specific Models. The spatial model comparisons comprise VST-7B-SFT (Yang et al., 2026a), SpaceEra (Zhang et al., 2026a), SpaceEra++ (Guan et al., 2026), SpatialVLM, Cambrian-S-7B (Yang et al., 2026b), and SenseNova-SI-1.3-8B (Cai et al., 2026). These models provide references for transferring specialized spatial capabilities to aerial perception, collaboration across viewpoints, and motion understanding. Their results are reported separately in Table 2.

Table 4: Sample-weighted category accuracy (%) on SIS-Bench. Row order and overall scores follow Table 1. Bold and underlined values mark the best and second-best model results, excluding human and random references.

Method	Overall	Spatial Cognition	Self-Awareness
Random	25.0	25.0	25.0
Human Performance	91.7	91.5	91.9
Proprietary Models			
Gemini-3-Flash	71.6	<u>83.7</u>	58.6
Kimi-2.5	71.0	83.5	57.6
Qwen3.5-Plus	70.1	83.5	55.7
GPT-5.4	70.0	80.6	58.6
Open-source Models			
SIS-Motion	69.1	74.2	63.7
Qwen3-VL-8B-Instruct	64.6	80.5	47.4
Qwen3-VL-8B-Thinking	64.7	79.9	48.3
Qwen3-VL-4B-Instruct	62.5	79.1	44.5
Qwen3.5-4B	61.1	75.9	45.1
InternVL3.5-8B	61.1	75.6	45.5
Qwen3-VL-30B-A3B-Instruct	61.0	80.8	39.8
GLM-4.1V-9B-Thinking	60.4	75.6	44.1
MiMo-VL-7B-RL	59.2	77.3	39.7
EnvSelf-UAV-4B (Ours)	<u>76.9</u>	83.5	<u>69.7</u>
EnvSelf-UAV-9B (Ours)	**79.9**	**86.2**	**73.0**

B.3 Implementation Details

Data and Training. Training proceeds in two stages. Image pretraining uses 37,768 instruction examples converted from the AerialVG training split (Liu et al., 2025). The region boxes, referring phrases, and positional relations provide the spatial supervision described in Appendix A.1. Video adaptation uses SIS-Motion-54K (Zou et al., 2026) to train temporal memory updates, motion inference, future prediction, and answer generation. We adapted the backbone using LoRA with rank 32, scaling parameter 64, and dropout 0.05. Each stage runs for one epoch with a per-device batch size of 1, eight gradient accumulation steps, a learning rate of 1×10^{-5} with cosine decay, weight decay 0.01, and random seed 42. Training uses bfloat16 precision, gradient checkpointing, and DeepSpeed ZeRO-2. The maximum sequence lengths are 4,096 tokens for image pretraining and 2,048 tokens for video adaptation. Video training samples 4–16 frames at 2 fps, while evaluation uses 8–32 frames per video. At inference, we used greedy decoding with at most 128 newly generated tokens and scored responses using the corresponding benchmark protocol.

Model Parameters. We instantiated EnvSelf-UAV with Qwen3.5-4B and Qwen3.5-9B, and used the 4B backbone for the component ablations. The feature dimension d follows the chosen backbone. DESM uses $K = 16$ environmental memory cells and a low-rank mixer with $R = 4$ latent vectors

and four attention heads. The box, phrase, and relation loss weights in Eq. (11) are $\lambda_b = 1.0$, $\lambda_p = 0.1$, and $\lambda_r = 0.1$, respectively. The EMDM motion representation and LWM latent states have dimension $d_m = 256$. The reference prediction horizon is $H = 6$, consistent with Section 4.4. The auxiliary weights in Eq. (18) are $\lambda_{\text{mot}} = 0.05$ and $\lambda_{\text{world}} = 0.03$. QGMI uses four attention heads and is inserted at approximately 67% of the decoder depth. For a video of $F \geq 2$ observed frames, its memory contains K environmental tokens, $F - 1$ observed motion tokens, and H future tokens, following Eq. (8). The injection residual uses a 200-step warm-up followed by a 300-step linear ramp.

Table 5: Sample-weighted category scores on SpatialUAV. Row order and overall scores follow Table 2. Bold and underlined values mark the best and second-best model results, excluding human and random references.

Method	Overall	Semantic	Spatial	Aerial–Aerial	Aerial–Ground	Motion
Human Level	89.0	99.3	96.2	74.4	97.2	89.1
Random Choice	14.4	8.3	8.1	11.1	17.7	24.1
Proprietary Models						
GPT-5.4	**56.7**	56.9	51.7	**44.1**	**68.1**	**66.6**
Claude-Opus-4-7	51.3	71.1	46.8	<u>37.8</u>	60.6	51.9
Gemini-3.1-Flash	45.9	72.2	57.5	23.5	<u>60.9</u>	37.6
Spatial-specific Models						
VST-7B-SFT	29.7	46.0	23.6	11.3	19.9	54.4
SpaceEra++	28.2	42.5	21.7	19.2	32.5	31.9
SpaceEra	27.3	41.8	21.0	18.8	31.9	30.1
SpatialVLM	25.7	30.1	17.2	15.0	20.8	46.3
Cambrian-S-7B	24.1	11.4	22.8	9.8	21.5	42.7
SenseNova-SI-1.3-8B	21.0	48.7	17.5	13.2	20.6	16.6
Open-source Models						
Qwen3.6-27B	49.5	66.5	**63.8**	28.6	54.8	50.5
Qwen3.5-9B	42.5	69.8	56.6	18.9	38.2	48.4
Qwen3.6-35B-A3B	37.3	53.6	56.9	11.4	34.7	47.7
Qwen3.5-4B	36.0	61.5	34.5	17.7	35.0	45.3
InternVL3.5-8B	30.8	43.7	27.3	14.2	26.8	49.1
InternVL3.5-14B	30.4	38.1	32.4	11.5	20.5	55.3
EnvSelf-UAV-4B (Ours)	49.8	<u>76.9</u>	50.2	26.4	54.0	58.6
EnvSelf-UAV-9B (Ours)	<u>53.9</u>	**82.4**	<u>63.1</u>	27.4	55.9	<u>61.3</u>

C Additional Experimental Results

C.1 Performance across Task Categories

Results on SIS-Bench. Table 4 shows improvements in both categories. EnvSelf-UAV-4B reaches 83.5% in spatial cognition and 69.7% in self-awareness, improving on Qwen3.5-4B by 7.6 and 24.6 percentage points, respectively. Its spatial cognition score is close to Gemini-3-Flash's 83.7%, while its self-awareness score exceeds the strongest baseline, SIS-Motion at 63.7%. The larger gain in self-awareness is consistent with the framework's emphasis on motion understanding alongside environmental memory.

Scaling to EnvSelf-UAV-9B raises spatial cognition to 86.2% and self-awareness to 73.0%, improving on the 4B variant by 2.7 and 3.3 percentage points. It achieves the highest model score in both categories, above the strongest baseline scores of 83.7% and 63.7%. The remaining gaps to human performance are 5.3 and 18.9 points, respectively, highlighting the greater remaining challenge in understanding and reasoning about the UAV's own motion.

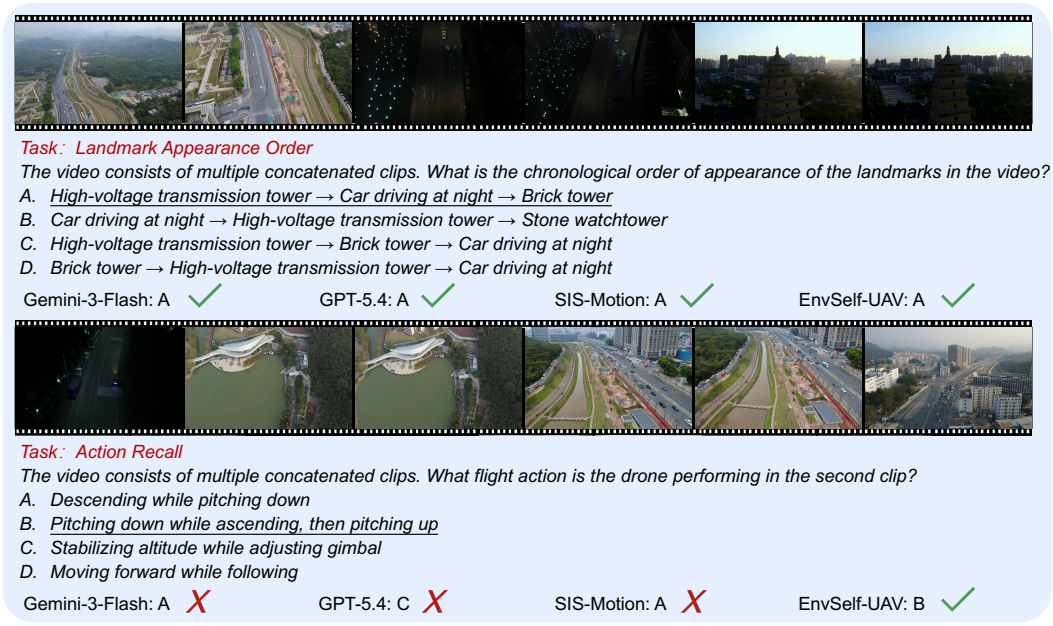

Figure 6: Qualitative comparisons on SIS-Bench. Underlined options indicate reference answers.

Results on SpatialUAV. Table 5 shows that both EnvSelf-UAV variants outperform their corresponding backbones across all five categories. EnvSelf-UAV-4B gains 19.0 points in aerial–ground collaboration and 13.3 points in motion understanding over Qwen3.5-4B. EnvSelf-UAV-9B leads non-proprietary models in semantic understanding, aerial–ground collaboration, and motion understanding, with its motion score increasing from the backbone's 48.4 to 61.3. These results support the effectiveness of combining persistent environmental memory with ego-motion modeling. However, its aerial–aerial score of 27.4 still trails GPT-5.4's 44.1, highlighting the remaining limitations in reasoning across aerial viewpoints, particularly precise camera transformation.

C.2 More Qualitative Analysis

More Cases on SIS-Bench. Figure 6 presents examples of landmark ordering and action recall. In the landmark ordering case, all four models correctly select A, recovering the sequence from a transmission tower to a car at night and then a brick tower. The action recall case requires distinguishing changes in altitude from changes in camera orientation within a specified clip. EnvSelf-UAV correctly selects B, identifying pitching down while ascending, followed by pitching up. Gemini-3-Flash and SIS-Motion instead select A, confusing ascent with descent, while GPT-5.4 selects C, describing stable altitude with gimbal adjustment. This example illustrates more accurate recall of compound motion beyond recognizing the visible scene.

More Cases on SpatialUAV. Figure 7 shows two examples of improved reasoning across viewpoints and a motion understanding failure. For camera transformation, EnvSelf-UAV predicts a heading offset of 95.0° and a translation distance of 67.0 m, closer to the reference values of 100.0° and 77.5 m than GPT-5.4's estimates of -164.0° and 44.0 m. The remaining distance error nevertheless indicates imperfect geometric estimation. In aerial–ground shared association, EnvSelf-UAV recovers both reference pairs, `1-0` and `6-1`, while GPT-5.4 incorrectly replaces `1-0` with `1-2`. These cases illustrate the ability to estimate relative viewpoints and associate regions across aerial and ground observations.

The motion understanding example reveals a limitation. Both models describe forward flight, whereas the reference indicates rightward and backward motion. EnvSelf-UAV additionally describes descent and tighter framing, which are unsupported by the reference description. Despite recognizing the farmers and the surrounding field, the model fails to recover the correct motion di-



Task: Camera Transformation
What are the signed heading offset angle from image 1 to image 2 and the signed relative translation distance of image 1 with respect to the target pose in image 2?
GT: *100.0 degrees, 77.5 meters*
GT-5.4: *-164.0 degrees, 44.0 meters*
EnvSelf-UAV: *95.0 degrees, 67.0 meters*



Task: Shared Association
List all shared region pairs between the aerial image and the ground image.
GT: *6-1, 1-0*
GT-5.4: *1-2, 6-1*
EnvSelf-UAV: *1-0, 6-1*



Task: Motion Understanding
This video is captured from a drone-mounted camera. Please describe, in temporal order, how the drone moves, how the camera angle changes, and how the main object shifts in the frame, ending with the final viewing focus.
GT: *The drone flies to the right and backward while keeping two individuals working in a muddy field in the center of the frame, with no significant tracking or movement observed.*
GT-5.4: *The drone moves slowly forward and slightly to the left over a flooded paddy field, keeping an oblique downward camera angle. As the view progresses, the camera gently pans left, bringing more of the green crop field and houses into view. The main worker in blue gradually shifts from the central area toward the lower-left side of the frame, while another worker near the field ridge becomes more visible on the right. In the final frames, the viewing focus settles on the two farmers working in the muddy field, especially the right-side worker near the ridge and the surrounding sowing area.*
EnvSelf-UAV: *The video begins with a high-angle, wide shot of two farmers working in a flooded rice paddy. The drone is positioned at a distance, capturing the entire scene, including the workers, the adjacent cornfield, and the surrounding rural landscape. As the video progresses, the drone initiates a slow and steady forward movement, flying directly over the rice paddy. This forward motion causes the camera to gradually descend, bringing the subjects closer into view. Consequently, the framing tightens, and the background elements like the cornfield and distant hills begin to recede and become less prominent.*

Figure 7: Qualitative comparisons on SpatialUAV. Underlined text indicates reference answers.

rection. This case highlights the remaining gap between scene recognition and reliable ego-motion description in open-ended responses.

More Cases on OpenUAV-QA. Figure 8 shows two path-planning examples with different outcomes. In the first example, EnvSelf-UAV correctly recovers the complete navigation sequence, including cruising forward for about 9 m, ascending 31 m, turning left by approximately 34°, cruising forward for about 171 m, and descending to land near the target. SIS-Motion and Gemini-3-Flash also select the correct route, whereas GPT-5.4 chooses a similar trajectory but underestimates the second forward segment as 50 m. This example illustrates the ability of EnvSelf-UAV to combine motion direction, travel distance, and altitude changes into a coherent multi-step plan.

The second example reveals a more challenging failure case. All evaluated models produce incorrect navigation plans, and the provided reference answer is itself inconsistent with the visual trajectory. The predictions mainly differ in the turning direction and subsequent motion, indicating difficulty in recovering the correct maneuver sequence from changing viewpoints. Since the reference is also erroneous, this case reflects both the remaining challenge of fine-grained trajectory reasoning and the sensitivity of path-planning evaluation to annotation quality.

D Limitations and Future Work

Despite the overall improvements of EnvSelf-UAV, several limitations remain. **(1) Precise geometric reasoning.** Accurately estimating camera transformations and relating landmarks to flight trajectories remain challenging in some cases, particularly under substantial viewpoint changes. Future work could incorporate more explicit geometric constraints to improve geometric reasoning accuracy. **(2) Fixed memory capacity and prediction horizon.** The current framework uses a fixed environmental memory capacity K and prediction horizon H, without adapting them to scene com-



Task: Path Planning

How should the drone navigate to reach the blue and white car on rocky terrain with dirt tracks and scattered rocks?

A. *Take off, ascend 31 meters, turn right 34 degrees, cruise forward 171 meters, descend 27 meters, and land near the target*

B. *Cruise forward 9 meters, take off, ascend 31 meters, turn left 34 degrees, cruise forward 50 meters, descend 27 meters, and land near the target*

C. *Cruise forward for about 9 meters. Take off and ascend approximately 31 meters. Turn left approximately 34 degrees. Cruise forward for about 171 meters. Descend approximately 27 meters and land near the target*

D. *Cruise forward 9 meters, take off, ascend 80 meters, turn left 34 degrees, cruise forward 171 meters, descend 27 meters, and land near the target*

Gemini-3-Flash: C ✓ GPT-5.4: B ✗ SIS-Motion: C ✓ EnvSelf-UAV: C ✓



Task: Path Planning

What maneuvers must the drone carry out to approach the woman in a red wheelchair near red-canopied buildings?

A. *Cruise forward for about 82 meters. Turn right approximately 78 degrees and fly about 4 meters. Cruise forward for about 104 meters*

B. *Cruise forward for about 82 meters. Turn left approximately 78 degrees and fly about 4 meters. Cruise forward for about 104 meters. Ascend 20 meters before landing*

C. *Cruise forward for about 30 meters. Turn left approximately 78 degrees and fly about 4 meters. Cruise forward for about 104 meters*

D. *Cruise forward for about 82 meters. Turn left approximately 78 degrees and fly about 4 meters. Cruise forward for about 104 meters*

Gemini-3-Flash: B ✗ GPT-5.4: B ✗ SIS-Motion: A ✗ EnvSelf-UAV: A ✗

Figure 8: Qualitative comparisons on OpenUAV-QA. Underlined text indicates reference answers.

plexity or motion dynamics. Exploring adaptive memory allocation and prediction horizons could better balance reasoning performance and computational cost across different flight conditions.