AeroVLA: A Vision-Language-Action Model for UAV Navigation via Minimalist End-to-End Control

Peng Xu[1,2], Zhengnan Deng[1,2], Jiayan Deng[1,2], Zonghua Gu[3], and Shaohua Wan[1,2]*

[1] University of Electronic Science and Technology of China, Chengdu, China
[2] Shenzhen Institute for Advanced Study, UESTC, Shenzhen, China
{202512281039, 202522280719, 202522280717}@std.uestc.edu.cn, shaohua.wan@uestc.edu.cn
[3] Department of Computer Science, Hofstra University, Hempstead, USA
zonghua.gu@hofstra.edu

Abstract. Vision-Language Navigation (VLN) for Unmanned Aerial Vehicles (UAVs) demands complex visual interpretation and continuous control in dynamic 3D environments. Existing hierarchical approaches rely on dense oracle guidance or auxiliary object detectors, creating semantic gaps and limiting genuine autonomy. We propose AeroVLA, a minimalist end-to-end Vision-Language-Action framework mapping raw visual observations and fuzzy linguistic instructions directly to continuous physical control signals. First, we introduce a streamlined dual-view perception strategy that reduces visual redundancy while preserving essential cues for forward navigation and precise grounding, which additionally facilitates future simulation-to-reality transfer. To reclaim genuine autonomy, we deploy a fuzzy directional prompting mechanism derived solely from onboard sensors, completely eliminating the dependency on dense oracle guidance. Ultimately, we formulate a unified control space that integrates continuous 3-Degree-of-Freedom (3-DoF) kinematic commands with an intrinsic landing signal, freeing the agent from external object detectors for precision landing. Extensive experiments on the TravelUAV benchmark demonstrate that AeroVLA achieves state-of-the-art performance in seen environments. Furthermore, it exhibits superior generalization in unseen scenarios by achieving nearly three times the success rate of leading baselines, validating that a minimalist, autonomy-centric paradigm captures more robust visual-motor representations than complex modular systems. Code is available at: https://github.com/XuPeng23/AeroVLA

Keywords: Vision-Language Navigation · Vision-Language-Action · UAV Control · End-to-End Learning

* Corresponding author.

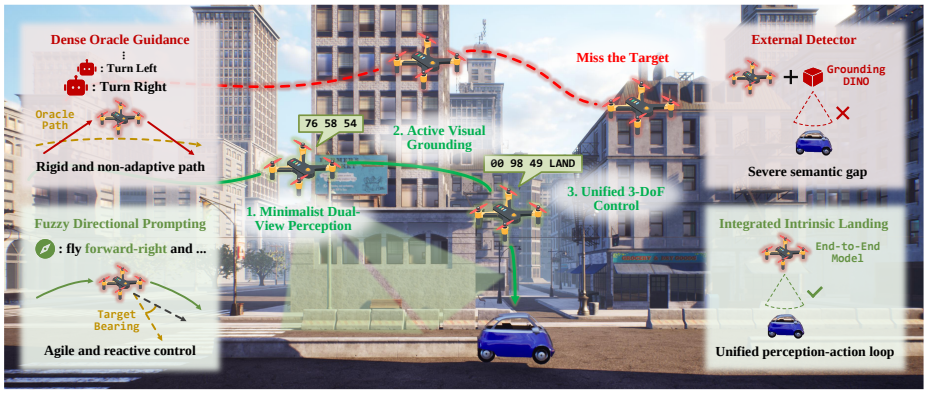

Fig. 1: Comparison of UAV navigation paradigms. While existing modular methods (red) rely on oracle guidance and external detectors, AeroVLA (green) achieves agile autonomous navigation and precise landing via a unified end-to-end policy driven by fuzzy onboard hints and intrinsic stopping.

1 Introduction

Vision-Language Navigation (VLN) has seen remarkable progress in ground-based agents. However, extending VLN to Unmanned Aerial Vehicles (UAVs) introduces unique challenges in 3D open-world environments. This transition marks a paradigm shift in aerial robotics [13, 26, 33], evolving from automated flight to agentic UAVs capable of autonomous perception and reasoning. The field has rapidly advanced alongside the development of specialized benchmarks, ranging from object goal navigation [38] to high-level spatial reasoning [8, 42] and realistic flight simulation [9, 36]. While these benchmarks rigorously evaluate cognitive scene understanding, mastering continuous 6-Degree-of-Freedom (6-DoF) physical execution remains arduous.

Unlike ground robots constrained to a 2D plane, UAVs navigate a full 6-DoF state space, managing 3D position and orientation while interpreting visual cues from actively changing ego-centric viewpoints. This introduces complex continuous control challenges under gravitational and inertial constraints. The goal of UAV-VLN is to enable drones to navigate to a natural-language-described target autonomously. This capability is critical for applications ranging from search and rescue to remote inspection, where GPS signals may be unreliable or target coordinates are unknown.

Despite this potential, existing UAV-VLN approaches often suffer from a reliance on what we term "double crutches", which limits their autonomy in the wild. A primary issue is the dependency on oracle guidance. State-of-the-art benchmarks such as TravelUAV [36] and recent methods [14, 41] typically inject dense, ground-truth-derived directional hints (e.g., "Turn Right") directly into the input prompts. This practice effectively degrades the navigation agent into a passive instruction follower, bypassing the core challenge of active spatial

reasoning and path planning. Furthermore, these modular pipelines frequently necessitate an external object detector, such as Grounding DINO [19], to trigger the landing phase due to a lack of fine-grained visual grounding. This dependency creates a disjointed perception-control loop where the policy learns how to move but relies on an external black box to decide when to stop, thereby reducing system robustness when detectors fail in open-world scenarios.

As illustrated in Figure 1, to address the aforementioned double crutches, we present **AeroVLA**, a minimalist end-to-end Vision-Language-Action framework. Unlike existing modular approaches (red trajectory) that suffer from rigid paths and semantic gaps due to dense oracle guidance and external detectors, AeroVLA (green trajectory) establishes a unified perception-action loop. By replacing exact oracles with fuzzy onboard directional hints, our agent is driven to perform active visual grounding, enabling agile and robust spatial reasoning. Concurrently, by mapping raw observations directly to continuous physical signals, AeroVLA unlocks an integrated intrinsic landing, seamlessly unifying long-distance cruising and precision stopping within a single policy.

In summary, our main contributions are three-fold:

- **Minimalist Dual-View Perception:** We fuse front and down views into a streamlined visual interface aligned with consumer UAV hardware. This design discards multi-camera redundancies while preserving essential geometric and semantic cues for forward navigation and precise target grounding.
- **Fuzzy Directional Prompting:** We eliminate the reliance on step-by-step oracle guidance by introducing fuzzy directional prompts. Derived solely from onboard IMU estimations, this formulation accommodates real-world localization uncertainty and forces the agent to learn robust, active spatial reasoning rather than passive instruction following.
- **High-DoF Control via Numerical Tokenization:** We tokenize a continuous 3-DoF action space compatible with standard UAV control APIs, effectively leveraging the pre-trained numerical reasoning of LLMs. This unlocks an end-to-end intrinsic stopping policy, unifying cruising and precision landing without external object detectors.

2 Related Work

2.1 Benchmarks and Datasets for Aerial Agents

High-quality simulators drive aerial autonomy. AerialVLN [20] and AVDN [7] pioneered instruction-following and dialog tasks, while UAV-ON [38] introduced open-world object navigation. Recent focus expanded to cognitive capabilities: SpatialSky-Bench [42] evaluates spatial reasoning, and UAVBench [8] assesses mission planning. However, these benchmarks primarily evaluate high-level perception or abstract logic. While OpenFly [9] offers diverse trajectories via discrete macro-actions, TravelUAV [36] focuses on 3D continuous navigation precision. We adopt TravelUAV to rigorously evaluate our agent's continuous maneuvering capabilities.

2.2 Vision-Language Navigation for UAVs

As a critical domain of embodied intelligence, the UAV-VLN task requires the integration of multimodal perception, spatial reasoning, and motion planning in complex 3D environments [39]. Unlike prior surveys that categorize methods by algorithmic foundations, we classify existing approaches based on their control granularity to explicitly highlight end-to-end execution challenges.

High-Level Planning and Waypoint Prediction. Most works, including early baselines like CMA [1] and AVDN [7], formulate navigation as a planning problem, predicting spatial waypoints for low-level controllers (e.g., PID). These include modular systems (CityNavAgent [43], SkyVLN [17], AeroDuo [37]) generating long-horizon or collaborative plans, and mission-generation frameworks (UAV-VLA [27]). Learning-based policies (TravelUAV [36], OpenVLN [18], LongFly [14], NavFoM [41], FlightGPT [3]) regress waypoints, targets, or trajectories via fine-tuned foundation models and spatiotemporal contexts, while ANWM [44] evaluates candidate trajectories using a generative world model. While effective for task decomposition, decoupling perception from control discards fine-grained visual cues critical for aerodynamic stability, introduces severe inference latency, and forces reliance on semantically unaware low-level controllers.

Training-Free and Reasoning-Centric Approaches. A parallel trend exploits frozen foundation models [23,31] to bypass policy training. For instance, SPF [11] prompts them to predict 2D waypoints for heuristic 3D unprojection, which often violates physical constraints due to absent depth perception. Meanwhile, TypeFly [4] generates code primitives for task planning. Despite their strong zero-shot capabilities, the sequential token generation of these massive LLMs creates inference latency fundamentally incompatible with real-time aerial agility.

End-to-End Continuous Control. This paradigm maps visual observations directly to continuous control signals (e.g., velocity, thrust). RaceVLA [28] and CognitiveDrone [22] pioneer this direction by adapting fine-tuned VLA models for flight. Nevertheless, these works are primarily confined to structured environments like racing tracks, simplifying navigation into discrete gate-selection tasks. While CognitiveDrone adds an auxiliary VLM, this decoupled, lower-frequency reasoning remains separated from the high-frequency control loop. Lacking intrinsic and unified spatial reasoning, these models struggle with the semantic grounding required for unstructured open-world exploration.

To overcome the dichotomy between decoupled hierarchical planning and structurally confined continuous control, AeroVLA introduces a minimalist VLA paradigm. By unifying fuzzy directional prompting with an intrinsic landing policy, we achieve robust, open-world reasoning and precise maneuvering without relying on external detectors or heavy reasoning chains.

2.3 Vision-Language-Action Models: From Ground to Air

The paradigm of Vision-Language-Action (VLA) models has revolutionized robotic control by unifying perception and action within a single transformer. Generalist

models like RT-2 [45] and OpenVLA [15] demonstrate remarkable generalization by fine-tuning VLMs to output robot actions directly as language tokens. While end-to-end learning has been successfully applied to ground navigation [16,29,32] to bypass traditional explicit map construction [2], existing VLA successes remain predominantly focused on quasi-static manipulation tasks or 2D ground rovers.

Extending the VLA paradigm to aerial platforms introduces fundamental challenges absent in ground-based settings. First, UAVs operate in a fully dynamic 6-DoF state space subject to continuous gravitational and inertial forces, where pitch and roll dynamics are tightly coupled with translational motion. Second, unlike reversible manipulation actions, flight control requires managing tight kinematic constraints where failure is often terminal (e.g., collision). AeroVLA bridges this gap by adapting the VLA architecture specifically for aerial embodiment. We demonstrate that a direct token-prediction approach, when grounded in large-scale expert demonstrations, effectively masters the continuous 3-DoF control required for high-stakes, open-world aerial navigation.

3 Method

3.1 Overview

We present **AeroVLA**, an end-to-end Vision-Language-Action framework designed to endow Unmanned Aerial Vehicles (UAVs) with autonomous navigation capabilities in open-world environments. As illustrated in Figure 2, our architecture is built upon the OpenVLA-7B [15] backbone, which unifies a hybrid visual encoder with a Llama 2 [34] language model using the Transformer architecture [35]. To adapt this generalist model for aerial navigation in dynamic 6-DoF state space, we introduce three key innovations: (1) a streamlined **Dual-View Perception** interface that balances informational gain with computational efficiency; (2) a **Fuzzy Directional Prompting** mechanism eliminating reliance on oracle assistance; and (3) a **Numerical Action Tokenization** strategy that leverages the pre-trained numerical reasoning of LLMs for precise control.

3.2 Minimalist Dual-View Perception

While recent UAV-VLN benchmarks typically simulate multi-camera arrays (e.g., five distinct views), processing such high-dimensional inputs increases computational overhead and inference latency, hindering real-time deployment. Since redundant views offer marginal utility for forward-facing tasks, we propose a minimalist dual-view strategy aligned with consumer UAV hardware. By processing only the front and down views, the front provides critical cues for obstacle avoidance and target identification, while the down view ensures precise ground alignment and landing maneuvers.

Formally, given *front* and *down* images $I_{\text{front}}, I_{\text{down}} \in \mathbb{R}^{H \times W \times 3}$ from AirSim [30], we perform vertical concatenation to form a composite observation

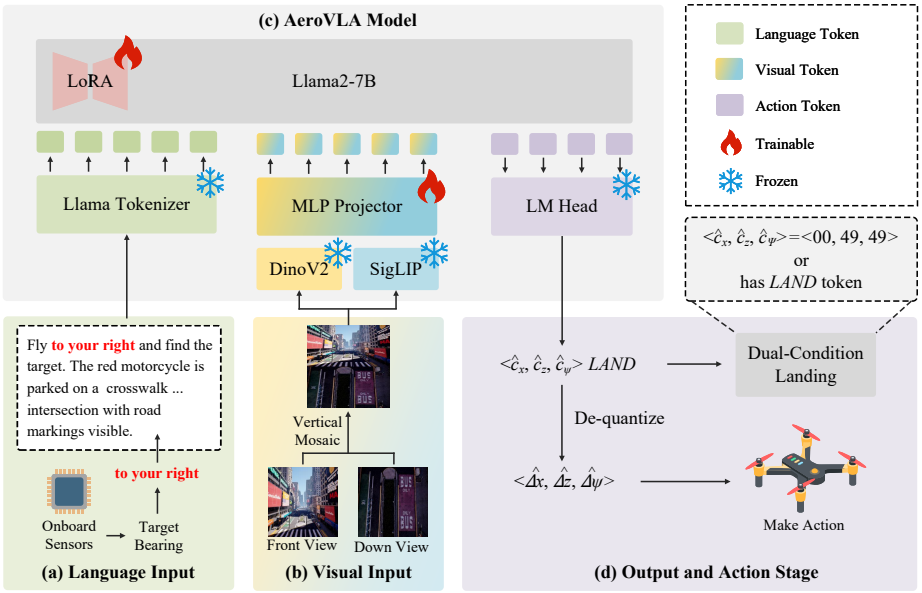

Fig. 2: The architecture of AeroVLA. The framework processes multimodal inputs to generate continuous control signals end-to-end. **(a) Language Input** constructs prompts with fuzzy directional hints derived from the IMU, eliminating oracle reliance. **(b) Visual Input** fuses front and down views via a vertical mosaic. **(c) The AeroVLA Model** utilizes a Llama-2 backbone with LoRA to autoregressively predict numerical tokens. **(d) Output and Action Stage** decodes tokens into spatial offsets for velocity control, or triggers the dual-condition landing.

$I_{\text{comp}} = [I_{\text{front}}; I_{\text{down}}] \in \mathbb{R}^{2H \times W \times 3}$. This composite is resized to the required 224×224 resolution and processed by a hybrid ViT [6] encoder combining SigLIP [40] for language-aligned semantics via CLIP-style [25] objectives, and DINOv2 [24] for robust, fine-grained spatial representations. We fully fine-tune the visual projector to map this input into the LLM embedding space, preserving crucial cues like object categories and terrain textures. Furthermore, AirSim's default 90° Field-of-View (FOV) naturally aligns the views at their horizontal boundary, establishing physical and semantic continuity. Significantly, the 224×224 dimension inherently aligns the stitching seam with the ViT's non-overlapping 14×14 patch grids, preventing intra-patch corruption and enabling the self-attention mechanism to cleanly resolve cross-view spatial relationships.

3.3 Fuzzy Directional Prompting

Existing UAV-VLN methods (e.g., TravelUAV [36]) often rely on dense, step-by-step oracle guidance from optimal trajectories, degrading the agent into a passive instruction follower. AeroVLA removes this dependency by prompting the agent with only fuzzy directional hints derived from onboard sensors (IMU/GPS).

Table 1: Fuzzy directional hint mapping from onboard IMU-derived relative angle.

Angle Range $\lvert\theta\rvert$	Fuzzy Hint ($\theta > 0$ / $\theta < 0$)
$0^\circ \leq \lvert\theta\rvert \leq 15^\circ$	"straight ahead"
$15^\circ < \lvert\theta\rvert \leq 60^\circ$	"forward-right" / "forward-left"
$60^\circ < \lvert\theta\rvert \leq 120^\circ$	"to your right" / "to your left"
$120^\circ < \lvert\theta\rvert \leq 180^\circ$	"to your right rear" / "to your left rear"

Specifically, we define a mapping function $\mathcal{M}$ that discretizes the relative bearing θ of the target into coarse-grained semantic buckets, as detailed in Table 1. Stripping away these dense oracles subjects our agent to a more rigorous learning objective. By replacing exact trajectory alignments with coarse directional priors, we introduce necessary spatial ambiguity. This formulation aligns with imperfect real-world localization and provides training tolerance. Lacking precise angular resolution, the model is forced to rely on active visual grounding, enhancing robustness against sensor noise and environmental ambiguity.

As illustrated in Figure 3, the prompt is structured as a natural sequence commencing with the visual placeholder, followed by the directional hint and the detailed target description. This format integrates multimodal context into a cohesive narrative for the LLM. Distinctively, AeroVLA executes a fully reactive policy based strictly on the current observation and immediate hint. By intentionally bypassing complex spatiotemporal memory buffers, this minimalist design drastically reduces inference latency and prevents the accumulation of cascading errors from stale states. Ultimately, prioritizing this instantaneous agility establishes a highly robust foundation for reactive visual-motor control in dynamic environments.

Geometry-Consistent Supervision. Training an autonomous policy on fuzzy prompts requires strictly resolving mathematical ambiguities in expert demonstrations. Modular baselines utilize dense oracle guidance, providing sufficient local context to justify detour maneuvers. Conversely, AeroVLA relies solely on coarse hints. Pairing a lateral target hint with a straight-flight label in an obstacle-free environment introduces causal confusion, which typically stems from delayed human pilot reactions. To resolve this ambiguity while preserving critical collision avoidance skills, we propose a geometry-consistent filtering strategy using lateral depth maps d_{lat}. Formally, we evaluate frames exhibiting a significant lateral target bearing ($|\theta| > 60^\circ$) coupled with a near-zero expert yaw rate ($\omega_\psi \approx 0$). For these ambiguous frames, we inspect the corresponding side-view depth: if the lateral space is clear ($\min(d_{\text{lat}}) > 20\text{m}$), we discard the sample. However, if a nearby obstacle is detected ($\min(d_{\text{lat}}) \leq 20\text{m}$), we retain the straight flight label as a valid, essential evasion maneuver. This geometric curation removes approximately 4% of the total training frames, guaranteeing that AeroVLA learns to navigate structural constraints, such as urban intersections, without being penalized by contradictory supervision.

AeroVLA Prompt Formulation

<image>
Fly straight ahead and find the target. The blue and white car is situated on a rocky terrain, surrounded by expansive, arid landscape with scattered rocks and distant mountains in the background under a clear sky. This rugged area appears barren with no visible vegetation, and the car is aligned among the rocks, partially camouflaged by the terrain.
Action: 00 49 49 LAND

Fig. 3: AeroVLA prompt formulation. The structured prompt comprises four components: (i) an `<image>` token for visual input, (ii) a fuzzy directional hint (red), (iii) a detailed target description, and (iv) the corresponding numerical control actions (blue).

3.4 High-DoF Control via Numerical Tokenization

Task-Oriented Action Space and Unified Landing. We define a continuous 3-DoF action space $\mathcal{A} = \langle \Delta x, \Delta z, \Delta\psi \rangle$ to decouple altitude, forward progression, and heading adjustments. This formulation delegates low-level stabilization (e.g., roll/pitch dynamics) to the flight controller, aligning with high-level motion primitives exposed by commercial UAV APIs (e.g., DJI SDK, PX4) to facilitate real-world deployment. Crucially, unlike modular baselines that rely on external object detectors (e.g., Grounding DINO [19]) to trigger flight termination, AeroVLA learns an intrinsic stopping policy. During training, we explicitly align terminal frames with a zero-displacement label vector $\langle 0, 0, 0 \rangle$ and the standard text token `LAND`. Consequently, landing is executed end-to-end via a robust dual-condition check: either the generation of the `LAND` token or the prediction of near-zero spatial offsets. This unifies navigation and landing into a single behavior cloning objective, enabling the agent to autonomously trigger flight termination upon visual convergence.

Standard Numerical Tokenization. Previous VLA approaches like RT-2 [45] typically introduce special action tokens (e.g., `<act_0>` to `<act_255>`). In data-constrained UAV domains, training these new embeddings from scratch creates a severe cold-start problem, forcing the model to re-learn basic ordinal relationships. While predicting actions via native tokens to circumvent this issue has been explored in foundation models like VLA-0 [10], our core contribution lies in its domain-specific integration for UAVs. Instead, AeroVLA maps continuous action dimensions, discretized into $N = 99$ bins, directly to existing numerical tokens within the LLM vocabulary. This elegantly leverages the understanding of magnitude and order inherent in the pre-trained model, yielding faster convergence and smoother control trajectories. At each timestep t, the model autoregressively predicts three integer tokens:

$$\langle \hat{c}_x, \hat{c}_z, \hat{c}_\psi \rangle = \mathrm{LLM}(E_{\mathrm{vis}}, E_{\mathrm{prompt}}) \tag{1}$$

where $\hat{c}_k \in \{0, 1, \ldots, 98\}$. These categorical indices are deterministically dequantized into continuous physical commands $\langle \hat{\Delta}x, \hat{\Delta}z, \hat{\Delta}\psi \rangle$. Specifically, $\hat{\Delta}x \in$

$[0, 5]$ and $\hat{\Delta z} \in [-5, 5]$ represent the forward and vertical displacements in meters, while $\hat{\Delta \psi} \in [-1.1, 1.1]$ denotes the yaw change in radians. These action boundaries are strictly derived from the statistical distribution of the expert dataset. To ensure robust physical execution, we map these spatial offsets to continuous velocity commands. Rather than relying on default position-control APIs that often induce erratic accelerations, we enforce a constant cruise speed (1.0 m/s) to dynamically calculate the corresponding flight duration. The UAV is then driven via the `moveByVelocityAsync` interface offered within the AirSim environment. This explicit velocity-duration mapping prevents abrupt kinematic transitions and minimizes camera motion blur, ensuring stable, high-quality visual observations for subsequent autoregressive inference steps. Finally, while our standard cross-entropy objective successfully preserves the pre-trained LLM distribution for smooth 3-DoF control, explicitly modeling ordinal relationships via soft labels [5] remains a promising direction for future refinement.

3.5 Training Objective

We formulate the learning process as a Behavior Cloning (BC) problem. Given a dataset of expert demonstrations $\mathcal{D} = \{(I_t, P, a_t^*)\}$, where I_t represents the visual observation, P the structured language prompt, and a_t^* the corresponding ground-truth expert action, we optimize the model to minimize the negative log-likelihood of the expert action tokens. The autoregressive training objective is defined as:

$$\mathcal{L} = -\mathbb{E}_{(I_t, P, a_t^*) \sim \mathcal{D}} \left[\sum_{k \in \{x, z, \psi\}} \log p(a_{t,k}^* \mid I_t, P, a_{t,<k}^*) \right] \tag{2}$$

where $a_{t,k}^*$ denotes the discrete token for the k-th dimension of the expert action. This straightforward frame-level supervision perfectly aligns with our reactive navigation policy.

4 Experiments

4.1 Dataset and Evaluation Metrics

Dataset. We evaluate AeroVLA on the TravelUAV benchmark [36], specifically the *UAV-Need-Help* task, which contains ∼12k human-piloted trajectories. Departing from the original oracle-assisted setting, we evaluate purely autonomous navigation guided exclusively by fuzzy directional prompts. We strictly adhere to the official data splits, training on 7,922 trajectories and evaluating across three distinct test sets comprising 1,418 trajectories for the Seen split, 629 for Unseen Object, and 958 for Unseen Map.

Evaluation Metrics. Following standard protocols [36], we report Navigation Error (NE) measuring the Euclidean distance to the target, Success Rate (SR) indicating the percentage of episodes ending within 20 meters of the target, Oracle Success Rate (OSR), and Success weighted by Path Length (SPL) to balance success with trajectory efficiency.

Table 2: Comparison on the Test Seen Set. SR, OSR, and SPL are reported in percentage (%). Bold and underline denote the best and second-best model results.

Method	Full				Easy				Hard			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Human	14.15	94.51	94.51	77.84	11.68	95.44	95.44	76.19	17.16	93.37	93.37	79.85
Random Action	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 Action	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 [1]	135.73	8.37	18.72	7.90	84.89	11.48	24.52	10.68	197.77	4.57	11.65	4.51
TravelUAV-DA [36]	98.66	17.45	48.87	15.76	66.40	20.26	51.23	18.10	138.04	14.02	45.98	12.90
NavFoM [41]	93.05	29.17	49.24	25.03	58.98	32.91	53.16	27.87	143.83	23.58	43.40	20.80
LongFly [14]	**60.02**	<u>36.39</u>	**65.87**	<u>31.07</u>	**38.10**	<u>38.52</u>	**71.90**	<u>31.24</u>	**85.20**	<u>33.94</u>	**58.94**	<u>30.88</u>
Ours	<u>65.88</u>	**47.96**	<u>57.69</u>	**38.54**	<u>43.76</u>	**49.30**	<u>61.30</u>	**37.14**	<u>93.16</u>	**46.30**	<u>53.23</u>	**40.26**

4.2 Implementation Details

Model and Training Setup. We instantiate AeroVLA using the OpenVLA-7B backbone [15], training on 420k frames from 7,922 expert trajectories. We apply LoRA [12] ($r = 64$, $\alpha = 128$, dropout 0.05) to the language backbone, yielding ∼2.98% trainable parameters. Both visual encoders remain frozen, while the projector is fully fine-tuned. Optimization uses AdamW [21] (weight decay 0.03, gradient clip 1.0) with a cosine scheduler (peak LR 2×10^{-4}, 5% warmup). Training runs on 4× RTX 4090 (24GB) GPUs with a global batch size of 64 for 5 epochs (∼35 hours) in BF16 precision.

Simulation Configuration. To strictly align with the decoupled action space in Sec. 3.4, we configure AirSim [30] to operate in `MaxDegreeOfFreedom` mode rather than the standard `ForwardOnly` mode. This enables the agent to execute independent yaw rotation ($\Delta\psi$) alongside a forward velocity vector, providing the holonomic agility essential for reactive obstacle avoidance.

Computational Efficiency. Evaluated on an RTX 4090, AeroVLA requires 17GB VRAM and 0.38s total latency, outperforming the 20GB and 0.63s of TravelUAV. While our dual-vision VLA model takes 0.35s versus their 0.26s backbone and prediction head, completely eliminating their 0.37s Assist and Grounding DINO modules drastically reduces system latency. With fuzzy directional hints adding merely 0.03s, our framework offers a vastly faster and more compact architecture for real-time navigation.

4.3 Baselines

To evaluate our pure end-to-end paradigm, we compare AeroVLA against three baseline categories. Unlike our single autoregressive stream, these methods inherently rely on specialized decoder heads or external auxiliary modules. All baseline results are directly sourced from their original publications [14, 36] using identical data splits.

Heuristic Baselines. Random Action and Fixed Action serve as lower bounds to quantify task difficulty.

Table 3: Comparison on the Test Unseen Object Set. SR, OSR, and SPL are reported in percentage (%). Bold and underline denote the best and second-best results, respectively.

Method	Full				Easy				Hard			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Random Action	260.14	0.16	0.16	0.16	174.10	0.48	0.48	0.48	302.96	0.00	0.00	0.00
Fixed Action	212.84	3.66	9.54	2.16	151.66	6.70	13.88	3.72	243.29	2.14	7.38	1.38
CMA [1]	155.79	9.06	16.06	8.68	102.92	14.83	22.49	13.90	182.09	6.19	12.86	6.08
TravelUAV [36]	118.11	22.42	46.90	20.51	86.12	24.40	49.28	22.03	134.03	21.43	45.71	19.75
NavFoM [41]	108.04	29.83	47.99	27.20	70.51	32.54	50.72	29.54	133.01	28.03	46.18	25.64
LongFly [14]	<u>66.74</u>	<u>43.87</u>	<u>64.56</u>	<u>38.39</u>	<u>54.84</u>	<u>38.01</u>	<u>56.84</u>	<u>31.36</u>	**57.07**	<u>50.25</u>	**74.16**	<u>45.27</u>
Ours	**61.45**	**56.60**	**64.86**	**46.61**	**45.72**	**56.94**	**64.11**	**43.76**	<u>69.27</u>	**56.43**	<u>65.24</u>	**48.03**

Hybrid VLM Approaches. We evaluate CMA [1], a traditional recurrent neural network baseline. Notably, we include TravelUAV [36], which combines a language model feature extractor with hierarchical decoders for trajectory regression, while relying on an external Grounding DINO [19] for target detection.

Recent Generalist Models. NavFoM [41] is a foundation model trained across various embodiments using a specialized trajectory planning head, and LongFly [14] leverages explicit spatiotemporal encoding modules.

4.4 Quantitative Results

Tables 2, 3, and 4 summarize our evaluation across the Seen, Unseen Object, and Unseen Map splits, demonstrating the distinct advantages of our fully autonomous paradigm over assistant-reliant baselines.

Performance on Seen Environments. As shown in Table 2, AeroVLA achieves a new state-of-the-art **47.96% SR** and **38.54% SPL**, surpassing the strongest baseline (LongFly) by substantial margins (+11.57% SR and +7.47% SPL). The performance gap becomes even more pronounced on the *Hard* split involving long-horizon flights, where our SR advantage widens to +12.36%. While LongFly achieves a higher OSR by incorporating ground-truth directional hints, its precipitous drop from OSR (65.87%) to SR (36.39%) exposes a critical failure in the final termination phase. In contrast, AeroVLA demonstrates superior OSR-to-SR conversion efficiency. By unifying navigation and landing with an intrinsic stopping mechanism, our agent autonomously executes precise terminal maneuvers without relying on external oracle triggers.

Generalization to Unseen Objects. Table 3 evaluates robustness against novel target categories, where AeroVLA maintains superiority with **56.60% SR** overall. While baselines achieve high OSR on the *Hard* split via oracle guidance, their reliance on explicit object detectors severely limits their ability to recognize and stop at out-of-distribution targets. In contrast, our method directly grounds novel visual concepts to control actions. This confirms that our minimalist framework leverages the open-vocabulary representations inherent in

Table 4: Comparison on the Test Unseen Map Set. SR, OSR, and SPL are reported in percentage (%). Bold and underline denote the best and second-best results, respectively.

Method	Full				Easy				Hard			
	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
Random Action	202.98	0.00	0.00	0.00	158.46	0.00	0.00	0.00	265.88	0.00	0.00	0.00
Fixed Action	180.47	0.52	2.61	0.39	132.89	0.89	4.28	0.67	247.72	0.00	0.25	0.00
CMA [1]	141.68	2.30	10.02	2.16	102.29	3.57	14.26	3.33	197.35	0.50	4.03	0.50
TravelUAV [36]	138.80	4.18	20.77	3.84	102.94	4.63	22.82	4.24	189.46	3.53	17.88	3.28
NavFoM [41]	125.10	6.30	18.95	5.68	102.41	6.77	20.07	6.04	170.58	5.36	15.71	4.97
LongFly [14]	<u>108.32</u>	<u>11.27</u>	<u>30.27</u>	<u>9.32</u>	<u>78.56</u>	<u>12.96</u>	<u>34.31</u>	<u>10.32</u>	<u>148.10</u>	<u>9.02</u>	<u>24.88</u>	<u>7.98</u>
Ours	**67.42**	**37.58**	**52.92**	**28.22**	**44.99**	**41.89**	**58.47**	**29.72**	**99.11**	**31.49**	**45.09**	**26.11**

the LLM to identify unseen targets, rather than merely overfitting to training categories.

Adaptability to Unseen Maps. The Unseen Map Test Set (Table 4) provides the strongest evidence of our generalization capability. In entirely novel environments, LongFly suffers a drastic degradation to 11.27% SR due to its heavy reliance on historical context and map-specific priors. In contrast, AeroVLA shows remarkable zero-shot adaptability, achieving **37.58% SR** and **28.22% SPL**. Both metrics are approximately three times those of the SOTA baseline. This indicates that AeroVLA acquires a fundamental visual servoing capability that seamlessly transfers to novel geometries. By strictly relying on instantaneous observations rather than accumulated spatial memory, our reactive approach exhibits superior robustness against severe environmental shifts.

4.5 Qualitative Analysis

Figure 4 visualizes representative trajectories highlighting two distinct navigation behaviors, demonstrating robust decision-making capabilities of the agent.

Precision Maneuvering in Clutter. In unstructured settings like extensive forests (Figure 4, Top), AeroVLA leverages its decoupled 3-DoF action space for fine-grained control. The agent horizontally aligns with the target while maintaining altitude above vegetation, followed by a precise vertical descent to safely land.

Active Error Correction against Distractors. Under visual ambiguity (Figure 4, Bottom), the agent approaches a distractor object, briefly hovers for inspection, and upon recognizing the mismatch, autonomously climbs to resume the search. This active perception loop proves a capacity for self-correction that strictly exceeds simple trajectory regression.

4.6 Ablation Study

We conduct ablation experiments to validate the individual contributions of our core architectural designs, with quantitative results summarized in Table 5.

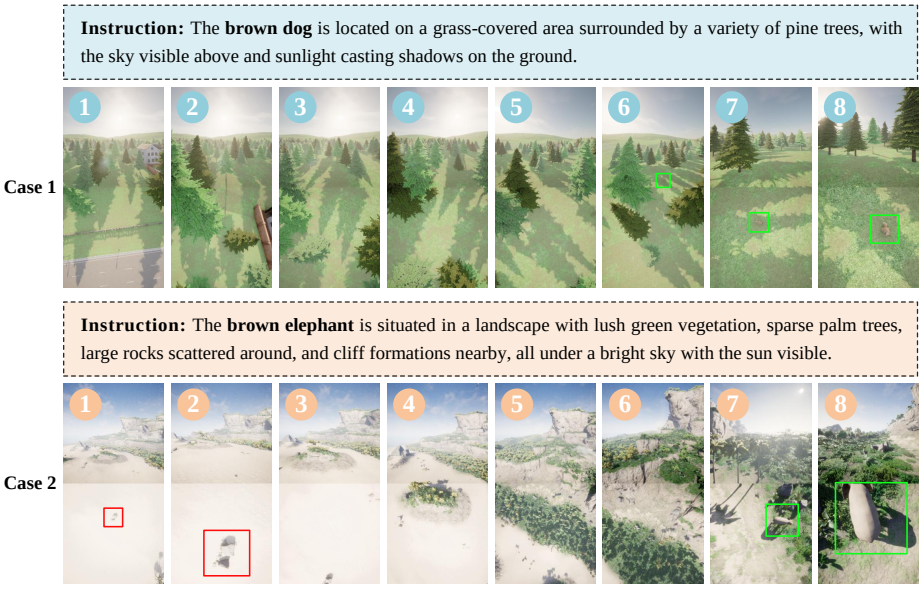

Fig. 4: Qualitative visualization of our proposed AeroVLA. We display the vertical mosaic inputs (Front/Down) at key timesteps. The agent demonstrates *precision maneuvering in clutter* (Top) and *active error correction against distractors* (Bottom), validating the robustness of the end-to-end policy.

Robustness to Raw Demonstrations. To verify that our performance gains stem from the architecture rather than dataset curation, we trained a variant on raw, uncleaned demonstrations. While noisy label conflicts predictably cause a slight degradation (e.g., 5.43% SR drop on unseen maps), this raw variant still achieves 32.15% SR, preserving a nearly threefold advantage over LongFly. This confirms the inherent robustness of the proposed architecture, proving that geometry-consistent filtering primarily serves to resolve mathematical ambiguities rather than artificially boosting baseline performance.

Efficacy of Minimalist Perception. A five-view variant incorporating redundant side and rear cameras severely degrades performance, plummeting the unseen map SR from 37.58% to 21.71%. This substantial decline corroborates our hypothesis that excessive visual inputs distract the agent and induce overfitting to background clutter, empirically validating our streamlined dual-view design.

Advantage of Numerical Tokenization. A variant utilizing custom action tokens instead of standard numerical digits experiences significant SR and SPL drops across all splits. This performance collapse validates our analysis in Sec. 3.4: training novel embeddings from scratch introduces a severe cold-start problem, whereas adopting a pure numerical format leverages the pre-trained magnitude awareness of the LLM to ensure reliable, fine-grained control.

Table 5: Ablation on data curation, perception views, and action tokenization. Metrics are in percentage (%).

Training Data / Method	Seen		Unseen Object		Unseen Map	
	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
Baseline LongFly [14]	36.39	31.07	43.87	38.39	11.27	9.32
Ours - custom tokens	39.84	31.73	38.79	32.77	26.51	19.98
Ours - 5-view input	41.54	32.69	51.51	42.33	21.71	13.46
Ours w/o filtering	40.90	32.59	51.99	42.67	32.15	22.67
Ours	**47.96**	**38.54**	**56.60**	**46.61**	**37.58**	**28.22**

5 Limitations and Future Work

While highly effective, the minimalist design of AeroVLA presents specific avenues for future refinement. First, by prioritizing instantaneous reactive control over explicit historical memory, the agent occasionally faces challenges with global backtracking in highly repetitive urban structures (e.g., the *NewYorkCity* map). Second, bounded by the nature of behavior cloning, the policy acts conservatively in extreme out-of-distribution scenarios. For instance, when targets are severely occluded by dense canopies in the unseen *ModularPark* map, the agent defaults to safe hovering rather than executing complex multi-stage exploration. To address these trade-offs, future work will integrate lightweight memory mechanisms for global reasoning and explore reinforcement learning fine-tuning to enable active exploration beyond static expert demonstrations.

To validate deployment feasibility, hardware-in-the-loop simulation using an INT8-quantized Qwen3-VL-2B (retaining 7B-level performance) on an NVIDIA Orin NX achieves a viable ∼520 ms inference latency. Furthermore, offline evaluations on real-world dual-view imagery demonstrate accurate zero-shot action predictions, leaving dynamic sim-to-real gaps (e.g., severe motion blur) for future physical flights to address.

6 Conclusion

This paper presents AeroVLA, fundamentally rethinking the prevailing modular methodologies in UAV vision-language navigation. Our work introduces a novel perspective to the field by establishing a minimalist end-to-end Vision-Language-Action paradigm. By navigating solely via onboard fuzzy hints and unifying cruising with precise termination into a single autonomous policy, we completely decouple the agent from dense oracle guidance and explicit object detectors. Extensive evaluations prove that discarding redundant spatial memory and complex auxiliary modules inherently fosters exceptional zero-shot generalization across unseen targets and novel map geometries. We anticipate that this pure end-to-end philosophy will serve as a robust foundation for future natively intelligent aerial agents operating in unconstrained open-world environments.

Broader Impact and Ethics Statement

This research advances autonomous UAV navigation using public simulation benchmarks without human subjects. We advocate for responsible deployment that strictly adheres to local aviation regulations and safety standards.

Acknowledgments

This work was supported in part by Shenzhen Science and Technology Program JCYJ20220818103200002, JCYJ20240813114223031, and the Natural Science Foundation of Guangdong Province under Grant No. 2024A1515011155, 2026A1515010179.

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018) 4, 10, 11, 12
2. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE T-RO **32**(6), 1309–1332 (2017) 5
3. Cai, H., Dong, J., Tan, J., Deng, J., Li, S., Gao, Z., Wang, H., Su, Z., Sumalee, A., Zhong, R.: FlightGPT: Towards generalizable and interpretable UAV vision-and-language navigation with vision-language models. In: EMNLP. pp. 6659–6676 (2025) 4
4. Chen, G., Yu, X., Ling, N., Zhong, L.: TypeFly: Low-latency drone planning with large language models. IEEE TMC **24**(09), 9068–9079 (2025) 4
5. Diaz, R., Marathe, A.: Soft labels for ordinal regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4738–4747 (2019) 9
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 6
7. Fan, Y., Chen, W., Jiang, T., Zhou, C., Zhang, Y., Wang, X.: Aerial vision-and-dialog navigation. In: ACL Findings. pp. 3043–3061 (2023) 3, 4
8. Ferrag, M.A., Lakas, A., Debbah, M.: UAVBench: An open benchmark dataset for autonomous and agentic AI UAV systems via LLM-generated flight scenarios. arXiv preprint arXiv:2511.11252 (2025) 2, 3
9. Gao, Y., Li, C., You, Z., Liu, J., Zhen, L., Chen, P., Chen, Q., Tang, Z., Wang, L., Yang, P., Tang, Y., Tang, Y., Liang, S., Zhu, S., Xiong, Z., Su, Y., Ye, X., Li, J., Ding, Y., Wang, D., Wang, Z., Zhao, B., Li, X.: OpenFly: A comprehensive platform for Aerial vision-language navigation. In: ICLR (2026) 2, 3
10. Goyal, A., Hadfield, H., Yang, X., Blukis, V., Ramos, F.: Vla-0: Building state-of-the-art vlas with zero modification. arXiv preprint arXiv:2510.13054 (2025) 8

11. Hu, C.Y., Lin, Y.S., Lee, Y., Su, C.H., Lee, J.Y., Tsai, S.R., Lin, C.Y., Chen, K.W., Ke, T.W., Liu, Y.L.: See, point, fly: A learning-free VLM framework for universal unmanned aerial navigation. In: CoRL. pp. 4697–4708 (2025) 4
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) 10
13. Javaid, S., Fahim, H., He, B., Saeed, N.: Large language models for UAVs: Current state and pathways to the future. IEEE OJVT **5**, 1166–1192 (2024) 2
14. Jiang, W., Wang, L., Huang, K., Fan, W., Liu, J., Liu, S., Duan, H., Xu, B., Ji, X.: LongFly: Long-horizon UAV vision-and-language navigation with spatiotemporal context integration. arXiv preprint arXiv:2512.22010 (2025) 2, 4, 10, 11, 12, 14
15. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: OpenVLA: An open-source vision-language-action model. In: CoRL. pp. 2679–2713 (2025) 5, 10
16. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: EMNLP. pp. 4392–4412 (2020) 5
17. Li, T., Huai, T., Li, Z., Gao, Y., Li, H., Zheng, X.: SkyVLN: Vision-and-language navigation and NMPC control for UAVs in urban environments. In: IROS (2025) 4
18. Lin, P., Sun, G., Liu, C., Li, F., Ren, W., Cong, Y.: OpenVLN: Open-world aerial vision-language navigation. arXiv preprint arXiv:2511.06182 (2025) 4
19. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55 (2024) 3, 8, 11
20. Liu, S., Zhang, H., Qi, Y., Wang, P., Zhang, Y., Wu, Q.: AerialVLN: Vision-and-language navigation for UAVs. In: ICCV. pp. 15384–15394 (2023) 3
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 10
22. Lykov, A., Serpiva, V., Khan, M.H., Sautenkov, O., Myshlyaev, A., Tadevosyan, G., Yaqoot, Y., Tsetserukou, D.: CognitiveDrone: A VLA model and evaluation benchmark for real-time cognitive task solving and reasoning in UAVs. arXiv preprint arXiv:2503.01378 (2025) 4
23. OpenAI, Achiam, J., Adler, S., Agarwal, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 4
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024) 6
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 6
26. Sapkota, R., Roumeliotis, K.I., Karkee, M.: UAVs meet agentic AI: A multidomain survey of autonomous aerial intelligence and agentic UAVs. arXiv preprint arXiv:2506.08045 (2025) 2
27. Sautenkov, O., Yaqoot, Y., Lykov, A., Mustafa, M.A., Tadevosyan, G., Akhmetkazy, A., Cabrera, M.A., Martynov, M., Karaf, S., Tsetserukou, D.: UAV-VLA: Vision-language-action system for large scale aerial mission generation. In: HRI. pp. 1588–1592 (2025) 4

28. Serpiva, V., Lykov, A., Myshlyaev, A., Khan, M.H., Abdulkarim, A.A., Sautenkov, O., Tsetserukou, D.: RaceVLA: VLA-based racing drone navigation with human-like behaviour. arXiv preprint arXiv:2503.02572 (2025) 4
29. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: ViNT: A foundation model for visual navigation. In: CoRL (2023) 5
30. Shah, S., Dey, D., Lovett, C., Kapoor, A.: AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In: FSR. pp. 621–635 (2018) 5, 10
31. Team, G., Anil, R., Borgeaud, S., Wu, Y., et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 4
32. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: DriveVLM: The convergence of autonomous driving and large vision-language models. In: CoRL (2024) 5
33. Tian, Y., Lin, F., Li, Y., Zhang, T., Zhang, Q., Fu, X., Huang, J., Dai, X., Wang, Y., Tian, C., Li, B., Lv, Y., Kovács, L., Wang, F.Y.: UAVs meet LLMs: Overviews and perspectives towards agentic low-altitude mobility. Inf. Fusion **122**, 103158 (2025) 2
34. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 5
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 5
36. Wang, X., Yang, D., Wang, Z., Kwan, H., Chen, J., Wu, W., Li, H., Liao, Y., Liu, S.: Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In: ICLR (2025) 2, 3, 4, 6, 9, 10, 11, 12
37. Wu, R., Zhang, Y., Chen, J., Huang, L., Zhang, S., Zhou, X., Wang, L., Liu, S.: AeroDuo: Aerial duo for UAV-based vision and language navigation. In: ACM MM (2025) 4
38. Xiao, J., Sun, Y., Shao, Y., Gan, B., Liu, R., Wu, Y., Guan, W., Deng, X.: UAV-ON: A benchmark for open-world object goal navigation with aerial agents. In: ACM MM. pp. 13023–13029 (2025) 2, 3
39. Yao, F., Liu, Y., Zhang, W., Zhu, Z., Li, C., Liu, N., Hu, P., Yue, Y., Wei, K., He, X., Zhao, X., Wei, Z., Xu, H., Wang, Z., Shao, G., Yang, L., Zhao, D., Yang, Y.: AeroVerse-Review: Comprehensive survey on aerial embodied vision-and-language navigation. Innov. Inform. **1**(1), 100015 (2025) 4
40. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023) 6
41. Zhang, J., Li, A., Qi, Y., Li, M., Liu, J., Wang, S., Liu, H., Zhou, G., Wu, Y., Li, X., Fan, Y., Li, W., Chen, Z., Gao, F., Wu, Q., Zhang, Z., Wang, H.: Embodied navigation foundation model. In: ICLR (2026) 2, 4, 10, 11, 12
42. Zhang, L., Zhang, Y., Li, H., Fu, H., Tang, Y., Ye, H., Chen, L., Liang, X., Hao, X., Ding, W.: Is your VLM sky-ready? a comprehensive spatial intelligence benchmark for UAV navigation. arXiv preprint arXiv:2511.13269 (2025) 2, 3

43. Zhang, W., Gao, C., Yu, S., Peng, R., Zhao, B., Zhang, Q., Cui, J., Chen, X., Li, Y.: CityNavAgent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. In: ACL. pp. 31292–31309 (2025) 4
44. Zhang, W., Tang, P., Zeng, X., Man, F., Yu, S., Dai, Z., Zhao, B., Chen, H., Shang, Y., Wu, W., Gao, C., Chen, X., Wang, X., Li, Y., Zhu, W.: Aerial world model for long-horizon visual generation and navigation in 3D space. arXiv preprint arXiv:2512.21887 (2025) 4
45. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.W.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023) 5, 8