AIR-VLA+: DECOUPLING MOVEMENT AND MANIPULATION VIA CASCADED DUAL-ACTION DECODERS WITH ASYMMETRIC MOE FOR AERIAL ROBOTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial manipulation systems have long suffered from representation coupling in end-to-end control, as platform-level Unmanned Aerial Vehicle (UAV) movement and end-effector-level arm manipulation differ substantially in action scale, dynamics, and control objectives. In this paper, we propose AIR-VLA+, a flow matching action generation architecture specifically designed for aerial manipulation, featuring cascaded dual-action decoders and an asymmetric feature-level Mixture of Experts (MoE). We construct cascaded manipulation and movement decoders, allowing the UAV to unidirectionally observe the manipulator's intent during movement to achieve workflow coordination, while isolating the impact of UAV movement information backpropagation on arm manipulation stability. Addressing the characteristic that UAV movement is highly dependent on high-level semantics and responsible for task state transitions in aerial manipulation, we design an input feature enhancement module for the UAV movement decoder. This module introduces an implicit visual grasp projector to perceive the interaction state between the gripper and the object, and injects compressed global semantic features. Within the UAV movement decoder, we deploy an implicit MoE architecture, enabling different movement experts to spontaneously exhibit capacity inclinations for various task stages during training. Through dense soft blending computation on the feature manifold, the UAV movement is endowed with stronger task-stage adaptability. Experiments on the standardized AIR-VLA benchmark demonstrate that our method comprehensively surpasses all baselines with an overall average score of 48.0. The overall task completion score improves by 80.2% compared to the single-head $\pi_{0.5}$ policy, effectively mitigating the heterogeneous coordinated control conflicts of composite robots.

1 INTRODUCTION

Empowering physical entities with multimodal large models is a core pathway toward Embodied Artificial Intelligence. Driven by the surging demand for complex, long-horizon interactions, Vision-Language-Action (VLA) models have emerged as a prominent paradigm. Particularly for continuous trajectory generation, flow matching has demonstrated strong convergence and trajectory generation capabilities when modeling high-dimensional action spaces. The rise of this end-to-end architecture enables composite robots, such as aerial manipulation systems, to directly map natural language instructions into high-dimensional, closed-loop continuous action sequences.

However, applying existing VLA models directly to aerial manipulation systems often encounters highly heterogeneous action space requirements. The Unmanned Aerial Vehicle (UAV) focuses on macroscopic spatial navigation (where the action output consists of relative positional changes), whereas the manipulator focuses on microscopic, local 3D high-precision alignment (where the action output consists of joint angles and gripper states). Traditional models force these two distinct streams into joint optimization within a single shared latent space, which can introduce representation interference and coordination conflicts. At the physical execution level, such interference can manifest as UAV drifting, jittering, and uncoordinated transitions across task stages.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

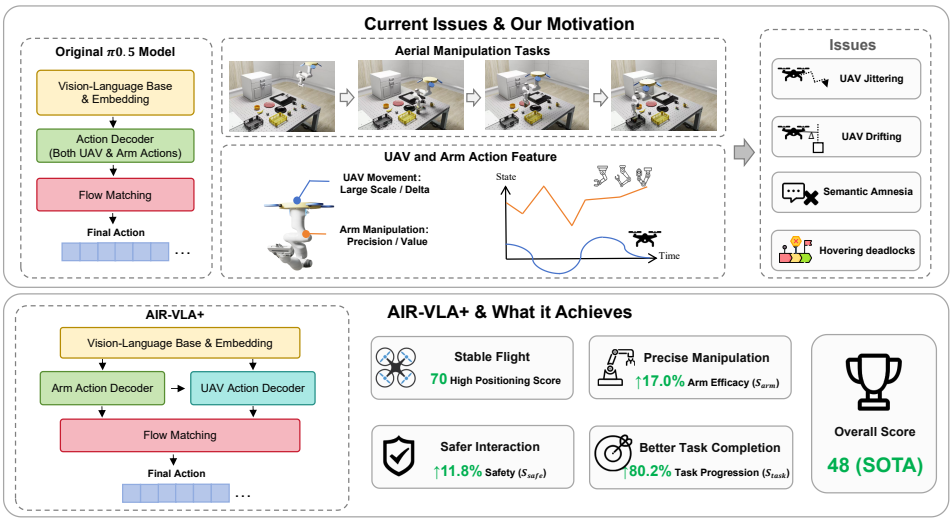


Figure 1: **Overview of the motivation and achievement of AIR-VLA+.**

In aerial manipulation tasks, the UAV must execute large-scale movements after identifying the target object, hover steadily while the manipulator interacts with the object, and proceed to the next target only after confirming a successful grasp. Conversely, the manipulator strictly focuses on precisely grasping the target. Due to this divergence in task objectives, the UAV heavily depends on continuous semantic understanding to clarify target objects across different workflow stages, deciding whether to hover or navigate based on the manipulator's interaction state. Importantly, this functional asymmetry does not imply that the UAV and manipulator are physically decoupled. UAV motion continuously affects manipulator reachability, visual observations, and physical interaction, while both action branches remain conditioned on the coupled system's shared multimodal observations.

Recent mobile-manipulation VLAs have explored separating base and arm action generation. EchoVLA employs independent, factorized diffusion policies for the mobile base and manipulator Lin et al. (2026), while InCoM introduces decoupled flow decoders with bidirectional cross-attention for whole-body coordination Liu et al. (2026). These studies demonstrate the effectiveness of factorized action generation, yet they also raise a critical question regarding how heterogeneous action streams should be coordinated. Unlike standard mobile manipulation, aerial manipulation exhibits a characteristic asymmetric task dependency: the UAV frequently needs to observe the manipulator's interaction intent to determine whether to approach, hover, or transition to the next target. In contrast, directly propagating UAV movement optimization gradients through the explicit cross-stream interaction pathway may disturb fine-grained arm control learning. This asymmetric dependency motivates us to investigate a dedicated arm-to-UAV coordination mechanism for aerial manipulation.

To address these challenges, we propose **AIR-VLA+**, a novel architecture that achieves representation-specialized yet coordinated action generation for UAV movement and arm manipulation. We achieve action-level decoupling via a cascaded dual-action decoder structure. For the UAV movement decoder, we design an input feature enhancement module that introduces a miniature implicit perception network to extract gripper-object interaction-state representations, alongside a compression module to inject global semantic information. Furthermore, we employ a unidirectional cascade channel from the arm manipulation decoder to the UAV movement decoder. Regulated by a stop-gradient operation, this channel allows the UAV to observe manipulation intent while preventing UAV-side gradients from propagating through this explicit cascade pathway into the manipulation projection. Finally, we adopt a Feature-Level Mixture of Experts (MoE) architecture within the UAV movement decoder. Through dense soft blending on the feature manifold, this MoE enhances the UAV's adaptability during task-stage transitions and provides stage-adaptive movement representations.

The core contributions of this paper are summarized as follows:

- We propose **AIR-VLA+**, an asymmetric coordination architecture featuring cascaded dual-action decoders tailored to the heterogeneous nature of aerial manipulation. Unlike existing independent or bidirectionally coordinated base-arm decoders, AIR-VLA+ introduces a unidirectional arm-to-UAV cascade with gradient isolation, enabling the UAV to condition its movement on manipulation intent while blocking UAV-side gradients along the explicit cascade pathway.
- We design an input feature enhancement module (extracting global semantics and implicit visual grasp representations) alongside an asymmetric Feature-Level MoE for the UAV movement decoder. Together, these components provide robust, stage-adaptive movement representations and facilitate efficient closed-loop coordination.
- Comprehensive evaluations on the standardized AIR-VLA benchmark demonstrate that our method significantly outperforms the evaluated baselines in UAV positioning, manipulation efficacy, safety, and task progression.

2 Related Works

Aerial Manipulation Systems. UAVs have seen widespread application across various domains; however, constrained by their mechanical design, traditional single-flight platforms struggle with complex interaction tasks such as pick-and-place operations. Consequently, aerial manipulator systems have emerged as a highly promising research direction. In terms of mechanical structure, researchers have enhanced manipulation flexibility through bionic and cooperative designs, including lightweight bionic robotic limbs to reduce payload Liu et al. (2024), continuum manipulators for compliant target interaction Peng et al. (2025), and dual-arm structures for multi-step collaborative tasks Ghorbani et al. (2023). Concurrently, advanced control methods have significantly improved manipulation accuracy and stability. These include refined anti-disturbance controllers Liu et al. (2024), adaptive prescribed performance control Liang et al. (2023), hierarchical control strategies Deshmukh et al. (2025), and image-based visual impedance control Xu et al. (2023). Furthermore, techniques such as coupled motion planning and whole-body integrated control have optimized overall system dynamics Wang et al. (2025). Despite these advances Zhang et al. (2025); Carvajal et al. (2024); Deng et al. (2025), existing modular systems primarily focus on optimizing physical control metrics and often lack holistic semantic understanding of dynamic environments, motivating the integration of multimodal intelligence.

Vision-Language-Action Models. VLA models provide a unified end-to-end architecture capable of mapping multimodal inputs directly to physical robotic actions. Compared to traditional modular pipelines, VLAs provide stronger semantic understanding and scene generalization capabilities. Since 2022, foundation models ranging from RT-1 Brohan et al. (2023) and OpenVLA Kim et al. (2024) to the π series Black et al. (2026; 2025); Intelligence et al. (2025; 2026) have rapidly evolved from short-horizon interactions toward increasingly generalizable long-horizon tasks. For composite mobile manipulators, recent works have begun explicitly modeling the distinct roles of base movement and arm manipulation. EchoVLA Lin et al. (2026) utilizes synergistic memories to guide separate per-part diffusion policies, while InCoM Liu et al. (2026) introduces decoupled flow matching decoders with bidirectional cross-attention for whole-body coordination. While these approaches demonstrate the value of separating heterogeneous action spaces, they primarily address ground-based mobile manipulation. AIR-VLA+ instead targets the asymmetric task dependency observed in aerial manipulation: manipulation intent is transmitted to the UAV movement decoder, while UAV-side gradients are blocked along the explicit cascade pathway.

VLA for Aerial Systems. Driven by powerful multimodal fusion, VLA architectures have significantly advanced UAV intelligence. Recent applications span coordinated navigation Liu et al. (2025), obstacle avoidance Sun et al. (2026), highly dynamic visual tracking Zhang et al. (2026); Xu et al. (2026), and cognitive trajectory planning Lykov et al. (2025). For resource-constrained onboard platforms, lightweight models utilizing 3D Gaussian Splatting and geometric safety corrections have achieved efficient closed-loop navigation Wu et al. (2025). However, VLA research specifically tailored for aerial manipulation remains nascent and has largely focused on transferring or adapting foundation models to aerial platforms. Sun et al. Sun et al. (icml 2026) proposed the AIR-VLA benchmark, verifying the transferability of ground-based VLAs while revealing perfor-

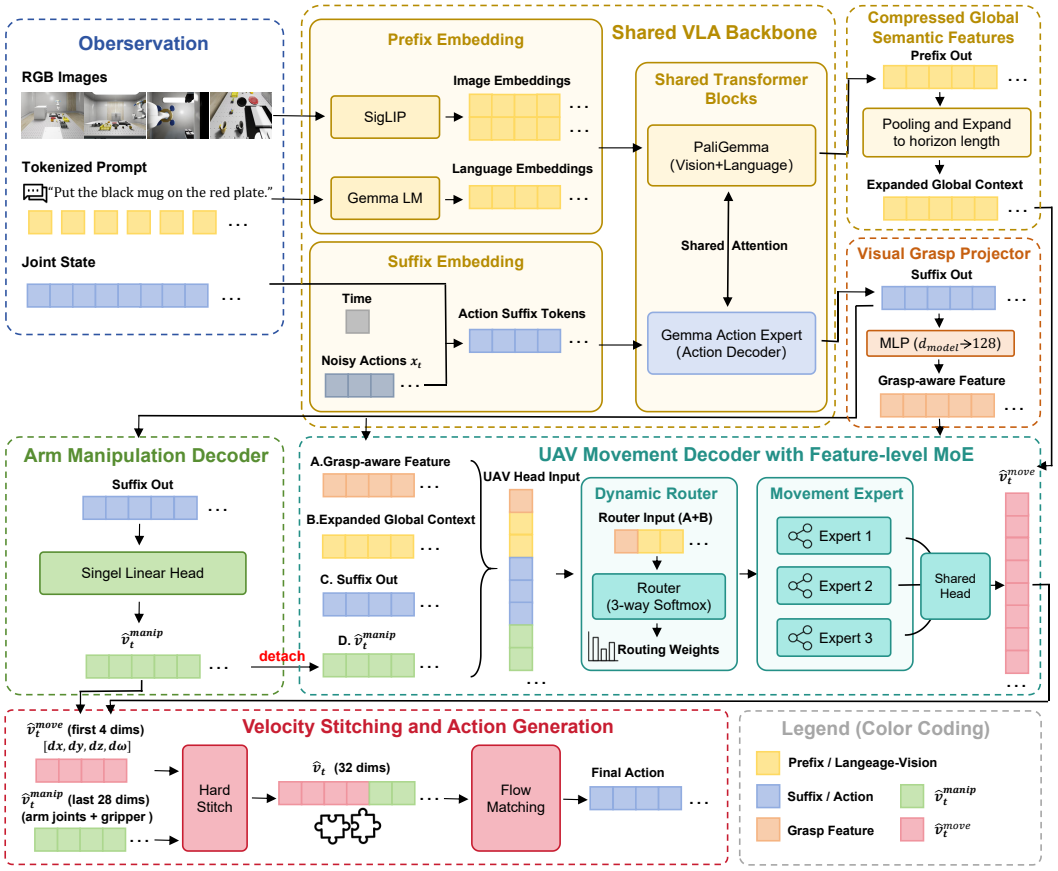


Figure 2: **Overview of the AIR-VLA+ architecture.** The system comprises a shared multimodal backbone, cascaded dual-action decoders, feature enhancement, unidirectional Arm→Move coordination, and an asymmetric feature-level MoE.

mance degradation associated with floating-base dynamics. Tucker et al. Tucker et al. (2026) explored fixed-base VLA transfer by injecting payload-aware physical constraints during flow matching sampling, emphasizing the importance of underactuated dynamics. Building upon these findings, this paper proposes AIR-VLA+ to explicitly address heterogeneous action coordination between the aerial platform and manipulator. By introducing asymmetric feature-level dynamic routing and a unidirectional arm-to-UAV cascade, our method reduces action representation interference while preserving system-level coordination.

3 Methodology

This section details the proposed Asymmetric Feature-Level MoE Flow Matching architecture. Aerial manipulation systems face severe challenges in long-horizon interaction tasks. To mitigate representation coupling between heterogeneous action spaces, we propose the AIR-VLA+ action generation architecture featuring cascaded dual-action decoders, enabling representation-specialized yet coordinated generation of UAV movement and arm manipulation actions (Figure 2).

3.1 Preliminaries: Flow Matching for Continuous Action Generation

Flow matching provides an efficient framework for training continuous-time generative models (e.g., Ordinary Differential Equations (ODEs)), demonstrating strong convergence and trajectory generation capabilities for high-dimensional robot action spaces.

We model the composite robot's physical action space as $x \in \mathbb{R}^{4+d_{\text{arm}}}$. The first 4 dimensions represent the UAV's movement control (dx, dy, dz, dw), while the remaining d_{arm} dimensions denote the manipulator's joint and gripper actions. Given a true action x_0 from the expert distribution $q(x_0)$ and noise $x_1 \sim \mathcal{N}(0, \mathbf{I})$, the conditional target vector field follows a straight interpolation trajectory:

$$x_t = t \cdot \epsilon + (1 - t) \cdot x_0, \qquad t \in [0, 1]. \tag{1}$$

The objective is to fit the derivative field $v_\theta(x_t, t, c)$ by minimizing the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|v_\theta(x_t, t, c) - (\epsilon - x_0)\right\|_2^2\right], \tag{2}$$

where c is the conditional input.

However, jointly predicting UAV and manipulator actions through a single shared output projection can introduce optimization interference between the two heterogeneous action subspaces. The UAV focuses on macroscopic, large-scale movement, whereas the manipulator focuses on microscopic, local 3D high-precision alignment. Their substantially different action distributions and control objectives may compete within the same action-generation representation, manifesting as UAV drifting, jittering, or unstable coordination during task-stage transitions.

3.2 Cascaded Dual-Action Decoders

To alleviate the bottleneck of joint action prediction, we construct cascaded manipulation and movement action decoders on top of a shared Vision-Language-Action backbone. The shared multimodal backbone first processes the visual observations, language instructions, noisy actions, and flow-matching timestep to extract foundational action representations $\mathbf{f}_{\text{suffix}}$.

The arm manipulation decoder directly projects these shared representations into the manipulation vector field:

$$\hat{v}_{\text{manip}} = P_{\text{manip}}\left(\mathbf{f}_{\text{suffix}}\right), \tag{3}$$

where P_{manip} denotes the manipulation action projection.

To achieve workflow coordination, we employ a unidirectional arm-to-UAV cascade channel from the manipulation decoder to the UAV movement decoder. Specifically, the predicted manipulation intent is provided as an additional condition for UAV movement generation. A stop-gradient operation is applied before transmitting this intent:

$$\tilde{v}_{\text{manip}} = \text{sg}\left(\hat{v}_{\text{manip}}\right), \tag{4}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation.

This asymmetric interface allows the UAV movement branch to observe the manipulator's current action intent, which provides important information for determining whether to approach, hover, or transition to the next task stage. Meanwhile, the stop-gradient operation prevents the UAV movement objective from backpropagating through this explicit manipulation-intent pathway and directly perturbing the manipulation projection decoder. Both branches still share the multimodal backbone and therefore remain conditioned on the coupled physical and perceptual state of the complete aerial manipulation system.

3.3 UAV Movement Decoder with Input Feature Enhancement

Addressing the characteristic that UAV movement strongly depends on task-level semantics and manages state transitions in aerial manipulation, we design an input feature enhancement module for the UAV movement decoder.

Compressed Global Semantic Features. We employ Mean Pooling as an information bottleneck to compress the sequence output of the multimodal prefix tokens, including images and language instructions, into a global intent vector. This feature is temporally broadcast across the action sequence horizon to form compressed global semantic features C_{global}, maintaining the UAV's continuous understanding of task instructions:

$$C_{\text{global}} = \text{Broadcast}\left(\text{MeanPool}\left(\mathbf{f}_{\text{prefix}}\right)\right). \tag{5}$$

Implicit Visual Grasp Projector. Aerial grasping requires definitive feedback on the interaction state between the manipulator and the target object. We introduce a miniature implicit perception network, implemented as a two-layer multi-layer perceptron (MLP), to extract interaction-state representations from the shared multimodal action features $\mathbf{f}_{\text{suffix}}$. Since these action features are conditioned on the visual-language prefix through the shared backbone, they contain observation-dependent information related to the current gripper-object interaction state. The visual grasp feature is obtained as

$$\mathbf{f}_{\text{grasp}} = P_{\text{grasp}}\left(\mathbf{f}_{\text{suffix}}\right). \tag{6}$$

Ultimately, the enhanced input for the UAV movement decoder, $\mathbf{f}_{\text{move_input}}$, is constructed by concatenating the shared foundational action features, compressed global semantic features, visual grasp features, and detached manipulation intent:

$$\mathbf{f}_{\text{move_input}} = \text{Concat}\left([\mathbf{f}_{\text{suffix}}, C_{\text{global}}, \mathbf{f}_{\text{grasp}}, \text{sg}\left(\hat{v}_{\text{manip}}\right)]\right). \tag{7}$$

3.4 Asymmetric Feature-Level MoE

Within the UAV movement decoder, we deploy an implicit feature-level MoE architecture to process $\mathbf{f}_{\text{move_input}}$. The routing network (Router) acts as a decision bottleneck, computing activation weights for the experts based on the global semantics and grasp-state features:

$$\mathbf{w} = \text{Softmax}\left(\text{MLP}_{\text{router}}\left([C_{\text{global}}, \mathbf{f}_{\text{grasp}}]\right)\right), \tag{8}$$

where w_i denotes the activation weight of the i-th movement expert.

The parallel movement experts $E_1, \ldots, E_K$ independently transform the enhanced UAV features into expert-specific latent representations:

$$\mathbf{h}_i = E_i\left(\mathbf{f}_{\text{move_input}}\right), \qquad i = 1, \ldots, K. \tag{9}$$

Instead of directly blending independently predicted action vector fields, we perform dense soft blending on the intermediate feature manifold:

$$\mathbf{h}_{\text{move}} = \sum_{i=1}^{K} w_i \cdot \mathbf{h}_i. \tag{10}$$

A shared movement projection subsequently maps the blended feature into the UAV movement vector field:

$$\hat{v}_{\text{move}} = P_{\text{move}}\left(\mathbf{h}_{\text{move}}\right). \tag{11}$$

This feature-level computation enables different movement experts to develop capacity inclinations toward different task stages while maintaining a continuous shared action projection. Through dense soft blending on the feature manifold, the UAV movement decoder gains stronger task-stage adaptability without directly switching or averaging independently generated physical actions.

3.5 Action Space Composition and Flow Matching Integration

After obtaining the UAV movement vector field $\hat{v}_{\text{move}}$ and the manipulation vector field $\hat{v}_{\text{manip}}$, we compose them according to their predefined physical action subspaces before the ODE solver. Specifically, the first four dimensions are provided by the UAV movement branch, while the remaining dimensions are provided by the manipulation branch:

$$\hat{v}_t^{(j)} = \begin{cases} \hat{v}_{\text{move}}^{(j)}, & j \in \{1, 2, 3, 4\}, \\ \hat{v}_{\text{manip}}^{(j)}, & j > 4. \end{cases} \tag{12}$$

Equivalently, the complete vector field can be written as

$$\hat{v}_t = \left[\hat{v}_{\text{move}}^{1:4}; \hat{v}_{\text{manip}}^{5:D}\right]. \tag{13}$$

The unified flow-matching objective is then applied to the composed vector field:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\left\|\hat{v}_t - (\epsilon - x_0)\right\|_2^2\right]. \tag{14}$$

The dimension-wise action composition ensures that the UAV and manipulation projection branches are directly supervised only within their corresponding action subspaces. The shared multimodal backbone remains jointly optimized by the complete flow-matching objective, while the stop-gradient operation specifically isolates the explicit manipulation-intent cascade from UAV-side gradient propagation.

During inference, the composed global derivative field $\hat{v}_t$ guides the flow matching ODE solver. Starting from pure Gaussian noise $x_1 \sim \mathcal{N}(0, \mathbf{I})$, the solver integrates along the vector field over the continuous time horizon $t \in [1, 0]$ to iteratively denoise the action trajectory. This process ultimately yields the coordinated continuous action sequence $x_0 \in \mathbb{R}^{4+d_{\text{arm}}}$, which is directly deployed for closed-loop aerial manipulation.

4 EXPERIMENTS

To systematically evaluate the effectiveness of the proposed architecture in Aerial Manipulation Systems, this section aims to answer the following core research questions (RQs):

- **RQ1 (System Performance):** Compared to action-level blending or hard stitching, can the asymmetric feature-level MoE effectively mitigate heterogeneous action interference and achieve superior performance in complex tasks?
- **RQ2 (Module Contribution and Coordination Direction):** What are the independent contributions of components such as the input feature enhancement module and the unidirectional arm-to-UAV cascade channel? Furthermore, among independent, bidirectional, and unidirectional interaction structures, which information-flow direction is more suitable for coordinating UAV movement and arm manipulation?
- **RQ3 (Interpretability):** Do the implicit expert networks spontaneously emerge cognitive specialization and state estimation capabilities during end-to-end optimization? Detailed experimental results and analyses are deferred to the Appendix.

4.1 EXPERIMENTAL SETUP AND EVALUATION METRICS

This study utilizes the **AIR-VLA** (Sun et al., icml 2026) benchmark, which is specifically designed for aerial manipulation. All experiments are conducted within Isaac Sim, controlling a quadrotor UAV equipped with a 7-DoF manipulator, constituting a 12-DoF control space, and adhering to standard multi-view perception configurations.

This benchmark encompasses four core task suites: Base Manipulation, Object & Spatial, Semantic Understanding, and Long-Horizon. It provides a comprehensive evaluation of aerial manipulation performance across four dimensions: UAV positioning, arm manipulation efficacy, safety, and overall task completion.

4.2 MAIN RESULTS: ARCHITECTURE COMPARISON (RQ1)

We compare Feature-MoE with traditional architectures (ACT, Diffusion Policy) and flow matching large models (π_0-FAST, π_0, $\pi_{0.5}$). To verify internal architectural gains, we use $\pi_{0.5}$ as the base and compare the single head, Action-Level Soft MoE, Action-Level Hard MoE, and our method (Table 1).

Quantitative Analysis. Traditional policy models lack cross-entity pre-training and perform extremely poorly. Flow matching bases such as π_0, despite having generalization priors, exhibit obvious bottlenecks when jointly modeling heterogeneous UAV movement and arm manipulation actions. The **Single Head** ($\pi_{0.5}$) baseline is constrained by the shared action-generation representation between macroscopic UAV movement and microscopic arm manipulation, which may introduce optimization interference between the two action subspaces.

Conventional schemes attempting to decouple heterogeneous actions directly at the action level all perform poorly. **Action-Level Soft MoE** directly blends independently predicted derivative fields,

Table 1: **Detailed performance evaluation of models across four task suites and overall average.** The table displays normalized scores for sub-metrics and weighted total scores for each model under different task types. Weights are set as $w_{\text{pos}} = 0.25$, $w_{\text{arm}} = 0.25$, $w_{\text{safe}} = 0.10$, and $w_{\text{task}} = 0.40$. Abbreviations: **Pos**: Base Positioning Accuracy (S_{pos}); **Arm**: Manipulator Efficacy (S_{arm}); **Safe**: Environmental Safety (S_{safe}); **Task**: Task Progression (S_{task}); **Tot**: Weighted Total Score (S_{total}). Best results are highlighted in bold.

Model	Base Manipulation					Object & Spatial					Semantic Understanding					Long-Horizon					Overall Average				
	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**
ACT	0.0	20.0	**100.0**	0.0	15.0	0.0	16.0	**100.0**	0.0	14.0	0.0	16.0	**100.0**	0.0	14.0	0.0	10.0	**100.0**	0.0	12.5	0.0	16.5	**100.0**	0.0	13.9
Diffusion Policy	0.0	20.0	**100.0**	0.0	15.0	0.0	16.0	**100.0**	0.0	14.0	0.0	16.0	**100.0**	0.0	14.0	0.0	10.0	**100.0**	0.0	12.5	0.0	16.5	**100.0**	0.0	13.9
π_0-FAST	27.5	23.9	98.1	0.0	22.7	5.1	18.9	98.3	0.0	15.8	6.5	16.9	99.3	0.0	15.8	41.0	32.6	99.4	0.0	28.3	15.0	20.8	98.5	0.0	18.8
π_0	56.6	48.0	87.5	8.0	38.1	58.1	38.2	75.4	2.0	32.4	55.7	36.1	85.1	2.0	32.3	**65.1**	38.5	85.9	2.0	32.3	58.8	42.0	83.5	3.5	34.5
Single Head ($\pi_{0.5}$)	78.8	49.9	87.2	11.0	45.3	**70.9**	44.1	81.2	13.5	42.3	**65.4**	44.7	82.7	**11.0**	40.2	62.4	40.8	84.8	7.0	37.1	**71.3**	45.9	84.7	10.6	42.0
Action-Level Soft MoE	54.9	42.3	41.4	0.5	28.6	64.1	38.9	47.2	6.5	33.1	62.5	39.5	46.6	5.5	32.3	52.1	42.4	29.5	9.5	30.4	52.4	40.8	41.2	5.5	31.1
Action-Level Hard MoE	0.0	20.0	**100.0**	0.0	15.0	0.4	19.8	**100.0**	0.0	15.1	0.0	19.2	**100.0**	0.0	14.8	1.2	17.4	99.5	0.3	14.7	0.4	19.1	99.9	0.1	14.9
Ours (Feature-MoE)	**87.0**	**68.8**	99.2	**33.0**	**62.1**	68.3	**49.4**	87.2	**15.6**	**44.4**	65.0	**45.1**	96.4	10.0	**41.2**	59.1	**60.0**	96.1	**17.5**	**44.1**	70.0	**53.7**	94.7	**19.1**	**48.0**

Table 2: **Detailed ablation study.** We evaluate compressed global semantics, visual grasp perception, coordination direction, stop-gradient information transmission, and multi-expert routing. Variants (d)–(g) compare independent, bidirectional, and two unidirectional coordination structures.

Model Variants	Base Manipulation					Object & Spatial					Semantic Understanding					Long-Horizon					Overall Average				
	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**	Pos	Arm	Safe	Task	**Tot**
(a) Original $\pi_{0.5}$ Base	78.8	49.9	87.2	11.0	45.3	**70.9**	44.1	81.2	13.5	42.3	**65.4**	44.7	82.7	11.0	40.2	62.4	40.8	84.8	7.0	37.1	**71.3**	45.9	84.7	10.6	42.0
(b) Multi-Head (Move+Manip)	85.0	61.6	99.6	17.0	53.4	27.9	25.8	93.7	4.0	24.4	25.0	22.9	89.9	5.0	23.0	**63.7**	51.6	92.7	20.0	46.1	50.4	40.5	94.0	11.5	36.7
(c) (b) + Global Semantics	35.3	0.0	**99.9**	0.0	10.1	13.4	13.0	**96.6**	0.0	16.3	12.9	13.4	95.7	0.0	16.1	53.9	41.8	94.4	5.0	35.4	20.2	17.1	**96.7**	1.25	19.5
(d) (b) + Semantics + Grasp (Independent)	79.1	55.6	84.1	12.0	47.9	30.5	29.1	91.6	2.0	24.9	32.3	32.0	86.1	7.0	27.9	71.4	51.5	95.9	**20.5**	**48.5**	53.3	42.1	91.9	10.3	37.2
(e) (b) + Semantics + Grasp (Bidirectional)	74.8	59.5	97.9	13.0	48.6	55.5	43.2	92.5	12.0	38.7	55.8	42.4	93.8	9.0	37.5	49.8	46.1	93.7	9.0	36.9	59.0	47.8	94.5	10.8	40.4
(f) (b) + Semantics + Grasp (Move→Arm)	80.7	60.1	97.5	14.0	50.6	53.3	39.5	90.5	11.0	36.7	59.8	40.0	89.3	8.0	34.6	60.1	48.8	93.3	9.5	40.4	61.0	47.1	92.7	10.6	40.5
(g) (b) + Semantics + Grasp (Arm→Move)	85.9	62.3	98.1	18.0	54.1	67.6	**50.5**	89.1	**20.0**	**46.4**	63.0	44.9	86.3	14.0	**41.2**	55.2	44.0	**98.0**	10.0	38.6	67.9	50.4	92.9	15.5	45.1
(h) Full Feature-MoE	**87.0**	**68.8**	99.2	**33.0**	**62.1**	68.3	49.4	87.2	15.6	**44.4**	65.0	**45.1**	**96.4**	10.0	**41.2**	59.1	**60.0**	96.1	17.5	**44.1**	70.0	**53.7**	94.7	**19.1**	**48.0**

resulting in substantially degraded trajectory consistency and safety in our experiments. **Action-Level Hard MoE** selects a single action expert during routing, introducing abrupt changes between expert-specific vector fields and leading to severe performance degradation.

In contrast, **Ours (Feature-MoE)**, via dense soft blending on the feature manifold and physical isolation of UAV and manipulator action subspaces at the output, comprehensively surpasses all evaluated baselines with an overall score of **48.0**. Compared to the single-head $\pi_{0.5}$ model, its task progression score improves from 10.6 to 19.1, corresponding to an 80.2% relative improvement. It also achieves substantial improvements in manipulation efficacy and safety, effectively mitigating heterogeneous coordinated control conflicts.

4.3 Ablation Study (RQ2)

To verify the contribution of each component and investigate the information-flow direction between UAV movement and arm manipulation, we conduct a bottom-up ablation study (Table 2).

Module Contribution Analysis. Compared with the original base (a), the dual-action decoder (b) separates movement and manipulation generation but lacks sufficient semantic and interaction information. Adding global semantics alone (c) is insufficient for reliable coordination, while introducing the visual grasp representation (d) restores local interaction-state information and improves overall performance.

Variants (d)–(g) further isolate the effect of coordination direction. Independent decoding (d) obtains an overall score of 37.2, while bidirectional interaction (e) improves it to 40.4. The reverse **Move→Arm** design (f) achieves a similar score of 40.5. In contrast, the proposed **Arm→Move** coordination (g) reaches 45.1, outperforming independent, bidirectional, and reverse unidirectional coordination by 7.9, 4.7, and 4.6 points, respectively. This result supports our asymmetric design: UAV movement benefits from manipulation intent for determining approach, hover, and task-stage transitions, whereas feeding movement information into the fine-grained arm stream provides limited benefit.

Table 3: **Sensitivity analysis on the number of latent experts (N).** The table reports the performance of Feature-MoE variants with different expert capacities. Values in parentheses indicate the performance gap ($\downarrow \Delta$) compared to the optimal $N = 3$ setting.

Capacity	S_{pos}	S_{arm}	S_{safe}	S_{task}	S_{total}
$N = 3$ (Ours)	**70.0**	**53.7**	**94.7**	**19.1**	**48.0**
$N = 2$ Experts	64.8 ($\downarrow$ 5.2)	45.8 ($\downarrow$ 7.9)	94.6 ($\downarrow$ 0.1)	7.0 ($\downarrow$ 12.1)	39.9 ($\downarrow$ 8.1)
$N = 4$ Experts	67.1 ($\downarrow$ 2.9)	48.5 ($\downarrow$ 5.2)	94.7	11.5 ($\downarrow$ 7.6)	43.0 ($\downarrow$ 5.0)

Finally, the complete **Feature-MoE (h)** further improves the overall score from 45.1 to **48.0** by introducing stage-adaptive movement representations. These results show that asymmetric Arm→Move coordination and feature-level expert routing provide complementary gains.

Sensitivity to the Number of Experts. The ablation study on the number of UAV movement experts (N) (Table 3) demonstrates that at $N = 2$, residual representation coupling persists, failing to smoothly partition transition states and resulting in degradation of core metrics. Conversely, at $N = 4$, the expanded parameter space does not yield the anticipated gains; certain experts maintain persistently low activation weights and lack distinct features that adapt independently to task phases. This corroborates that aerial manipulation tasks inherently converge into three latent spaces, establishing $N = 3$ as the optimal number of experts.

5 Conclusion

To address representation interference and coordinated control conflicts in aerial manipulation, we propose AIR-VLA+, a flow-matching architecture utilizing cascaded dual-action decoders and an asymmetric feature-level MoE. An input feature enhancement module introduces an implicit visual grasp projector for gripper-object interaction perception and injects global semantics to maintain UAV task awareness. A stop-gradient unidirectional arm-to-UAV cascade channel enables the UAV movement decoder to observe manipulator intent for coordination, while preventing UAV-side optimization from propagating through this explicit manipulation-intent pathway. Additionally, the UAV decoder's MoE employs dense soft blending on the feature manifold before the final movement projection, enhancing task-stage adaptability while preserving specialized manipulation generation. AIR-VLA benchmark experiments validate that our approach significantly improves UAV positioning, manipulation efficacy, safety, and task progression over the evaluated baselines, demonstrating the effectiveness of asymmetric coordination and stage-adaptive movement generation for heterogeneous aerial manipulation systems.

AI Use Statement

Generative AI tools were used to assist with language polishing and improving the readability of the manuscript. They were also used to assist with implementing and debugging research code based on methods and designs specified by the authors. All AI-assisted code was manually reviewed, tested, and validated by the authors. The authors reviewed all AI-assisted manuscript edits and take full responsibility for the final content, experimental results, and conclusions of this work.

References

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=vlhoswksBO.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023. URL https://arxiv.org/abs/2212.06817.

Christian P. Carvajal, Gabriela M. Andaluz, Víctor H. Andaluz, Flavio Roberti, Guillermo Palacios-Navarro, and Ricardo Carelli. Multitask control of aerial manipulator robots with dynamic compensation based on numerical methods. *Robotics and Autonomous Systems*, 173:104614, 2024. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2023.104614. URL https://www.sciencedirect.com/science/article/pii/S0921889023002531.

Weiliang Deng, Hongming Chen, Biyu Ye, Haoran Chen, Ziliang Li, and Ximin Lyu. Whole-body integrated motion planning for aerial manipulators. *IEEE Transactions on Robotics*, 41:6661–6679, 2025. doi: 10.1109/TRO.2025.3626619.

Shlok Deshmukh, Javier Alonso-Mora, and Sihao Sun. Global end-effector pose control of an underactuated aerial manipulator via reinforcement learning, 2025. URL https://arxiv.org/abs/2512.21085.

Shahab Ghorbani, Zahra Samadikhoshkho, and Farrokh Janabi-Sharifi. Dual-arm aerial continuum manipulation systems: modeling, pre-grasp planning, and control. *Nonlinear Dynamics*, 111(8):7339–7355, Apr 2023. ISSN 1573-269X. doi: 10.1007/s11071-022-08212-w. URL https://doi.org/10.1007/s11071-022-08212-w.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a vla that learns from experience, 2025. URL https://arxiv.org/abs/2511.14759.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, Vedant Choudhary, Foster Collins, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Maitrayee Dhaka, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachlan Groom, Haroun Habeeb, Hunter Hancock, Karol Hausman, Gashon Hussein, Victor Hwang, Brian Ichter, Connor Jacobsen, Szymon Jakubczak, Rowan Jen, Tim Jones, Gregg Kammerer, Ben Katz, Liyiming Ke, Mairbek Khadikov, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Brendon LeCount, Sergey Levine, Xinyu Li, Adrian Li-Bell, Vladislav Lialin, Zhonglin Liang, Wallace Lim, Yao Lu, Enyu Luo, Vishnu Mano, Nandan Marwaha, Aikys Mongush, Liam Murphy, Suraj Nair, Tyler Patterson, Karl Pertsch, Allen Z. Ren, Gavin Schelske, Charvi Sharma, Baifeng Shi, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Jiaming Tang, Jimmy Tanner, Shalom Tekeste, Marcel

Torne, Kyle Vedder, Quan Vuong, Anna Walling, Haohuan Wang, Jason Wang, XuDong Wang, Chris Whalen, Samuel Whitmore, Blake Williams, Charles Xu, Sukwon Yoo, Lili Yu, Wuming Zhang, Zhuoyang Zhang, and Ury Zhilinsky. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities, 2026. URL https://arxiv.org/abs/2604.15483.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Jiacheng Liang, Yanjie Chen, Yangning Wu, Zhiqiang Miao, Hui Zhang, and Yaonan Wang. Adaptive prescribed performance control of unmanned aerial manipulator with disturbances. *IEEE Transactions on Automation Science and Engineering*, 20(3):1804–1814, 2023. doi: 10.1109/TASE.2022.3186315.

Min Lin, Xiwen Liang, Bingqian Lin, Jingzhi Liu, Zijian Jiao, Kehan Li, Ziang Yan, Yu Sun, Weijia Liufu, Yuhan Ma, Jiarui Hu, Yuecheng Liu, Shen Zhao, Yuzheng Zhuang, and Xiaodan Liang. Echovla: Robotic vision-language-action model with synergistic declarative memory for mobile manipulation, 2026. URL https://arxiv.org/abs/2511.18112.

Jiahao Liu, Cui Wenbo, Zhongpu Xia, Yongliang Wang, Haoran Li, and Dongbin Zhao. Incom: Intent-driven perception and structured coordination for mobile manipulation, 2026. URL https://arxiv.org/abs/2602.23024.

Qianyuan Liu, Yuhang Liu, Zeshuai Chen, Kexin Guo, Xiang Yu, Youmin Zhang, and Lei Guo. A compact aerial manipulator: Design and control for dexterous operations. *Journal of Intelligent & Robotic Systems*, 110(2):66, Apr 2024. ISSN 1573-0409. doi: 10.1007/s10846-024-02090-7. URL https://doi.org/10.1007/s10846-024-02090-7.

Xu Liu, Yu Liu, Hanshuo Qiu, Yang Qirong, and Zhouhui Lian. Indooruav: Benchmarking vision-language uav navigation in continuous indoor environments, 2025. URL https://arxiv.org/abs/2512.19024.

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

Rui Peng, Yu Wang, Minghao Lu, and Peng Lu. A dexterous and compliant aerial continuum manipulator for cluttered and constrained environments. *Nature Communications*, 16(1):889, Jan 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55157-2. URL https://doi.org/10.1038/s41467-024-55157-2.

Jianli Sun, Bin Tian, Qiyao Zhang, Chengxiang Li, Zihan Song, Zhiyong Cui, Yisheng Lv, and Yonglin Tian. Air-vla: Vision-language-action systems for aerial manipulation, icml 2026. URL https://arxiv.org/abs/2601.21602.

Xiaolou Sun, Wufei Si, Wenhui Ni, Yuntian Li, Dongming Wu, Fei Xie, Runwei Guan, He-Yang Xu, Henghui Ding, Yuan Wu, Yutao Yue, Yongming Huang, and Hui Xiong. Autofly: Vision-language-action model for uav autonomous navigation in the wild, 2026. URL https://arxiv.org/abs/2602.09657.

Johnathan Tucker, Denis Liu, Aiden Swann, Allen Ren, Javier Yu, Jiankai Sun, Brandon Kim, Lachlain McGranahan, Quan Vuong, and Mac Schwager. π, but make it fly: Physics-guided transfer of vla models to aerial manipulation, 2026. URL https://arxiv.org/abs/2603.25038.

Meng Wang, Shangke Lyu, Qianyuan Liu, Ziqi Yang, Kexin Guo, and Xiang Yu. Precise end-effector control for an aerial manipulator under composite disturbances: Theory and experiments. *IEEE Transactions on Automation Science and Engineering*, 22:4006–4021, 2025. doi: 10.1109/TASE.2024.3406754.

Yuze Wu, Mo Zhu, Xingxing Li, Yuheng Du, Yuxin Fan, Wenjun Li, Zhichao Han, Xin Zhou, and Fei Gao. Vla-an: An efficient and onboard vision-language-action framework for aerial navigation in complex environments, 2025. URL https://arxiv.org/abs/2512.15258.

Mengxin Xu, An Hu, and Hesheng Wang. Image-based visual impedance force control for contact aerial manipulation. *IEEE Transactions on Automation Science and Engineering*, 20(1):518–527, 2023. doi: 10.1109/TASE.2022.3162207.

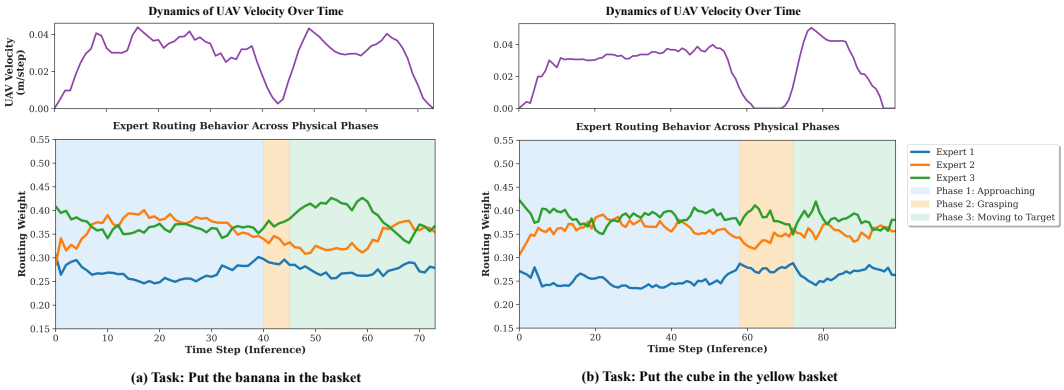
Peng Xu, Zhengnan Deng, Jiayan Deng, Zonghua Gu, and Shaohua Wan. Aerialvla: A vision-language-action model for uav navigation via minimalist end-to-end control, 2026. URL https://arxiv.org/abs/2603.14363.

Qiyao Zhang, Shuhua Zheng, Jianli Sun, Chengxiang Li, Xianke Wu, Zihan Song, Zhiyong Cui, Yisheng Lv, and Yonglin Tian. Uav-track vla: Embodied aerial tracking via vision-language-action models, 2026. URL https://arxiv.org/abs/2604.02241.

Zhaopeng Zhang, Hai Yu, Yi Chai, Zhichao Yang, Xiao Liang, Yongchun Fang, and Jianda Han. An end-effector-oriented coupled motion planning method for aerial manipulators in constrained environments. *IEEE/ASME Transactions on Mechatronics*, 30(6):6027–6037, 2025. doi: 10.1109/TMECH.2025.3550562.

A INTERPRETABILITY: EMERGENCE OF EXPERTS (RQ3)

This section investigates whether implicit expert networks spontaneously emerge physical cognition and functional division of labor under flow matching end-to-end optimization.



(a) Task: Put the banana in the basket

(b) Task: Put the cube in the yellow basket

Figure 3: **Routing Dynamics in Long-Horizon Task.** Within a single task cycle, the three implicit experts demonstrate a highly logical temporal handover mechanism.

Routing Dynamics. As illustrated in Fig. 3, we track the evolution of expert weights throughout a single task cycle. Starting from an initial uniform distribution (0.333), the weights exhibit a significant phase transition after training: during the *Approaching* stage, Expert 2's weight demonstrates an upward trend alongside UAV activation; during the *Grasping* stage, Expert 1 reaches a local peak; whereas in the *Moving to target* stage, Expert 3 leaps significantly and dominates.

This trend intuitively reflects the Router's success in executing dynamic control based on task state transitions. In summary, Expert 1 exhibits a capacity inclination for the *grasping* stage, Expert 2 for the *approaching* stage, and Expert 3 for the *moving to target* stage.

Expert Isolation and Hard-Routing Analysis. To verify the essence of physical decoupling, we conducted forced routing experiments by locking the routing weights to one-hot vectors (Table 4). We observed varying degrees of performance degradation when only a single expert was activated. Activating only Expert 1 results in the minimal overall performance drop, yet the UAV frequently exhibits sluggishness when moving toward the target object. Activating only Expert 2 or Expert 3 leads to significant performance degradation, specifically manifesting as the UAV maintaining excessively high velocities, frequently overshooting the target object, and failing to hover at the appropriate position. This phenomenon quantitatively corroborates that a single network struggles to balance the diverse dynamic requirements across all stages of an aerial manipulation task.

Table 4: **Performance degradation under expert isolation (Hard-Routing).** The table shows the absolute scores when routing weights are forcibly locked to a single expert. Values in parentheses indicate the performance drop (↓ Δ) or gain (↑ Δ) compared to the dynamic Feature-MoE baseline.

Routing Condition	S_{pos}	S_{arm}	S_{safe}	S_{task}	S_{total}
Dynamic (Ours)	**70.0**	**53.7**	94.7	**19.1**	**48.0**
Expert 1 Only ([1,0,0])	69.7 (↓ 0.3)	53.2 (↓ 0.5)	93.6 (↓ 1.1)	18.8 (↓ 0.3)	47.6 (↓ 0.4)
Expert 2 Only ([0,1,0])	67.2 (↓ 2.8)	48.6 (↓ 5.1)	94.9 (↑ 0.2)	10.8 (↓ 8.3)	42.7 (↓ 5.3)
Expert 3 Only ([0,0,1])	67.8 (↓ 2.2)	50.1 (↓ 3.6)	**95.8** (↑ 1.2)	14.8 (↓ 4.3)	45.0 (↓ 3.0)

Latent Manifold Analysis. To explore the latent manifold nature of the perceptor, we extracted hidden representations in each task suite and conducted t-SNE projections and Linear Probing tests (Fig. 4). In simple **Base Manipulation**, features are highly linearly separable (93.24% accuracy).

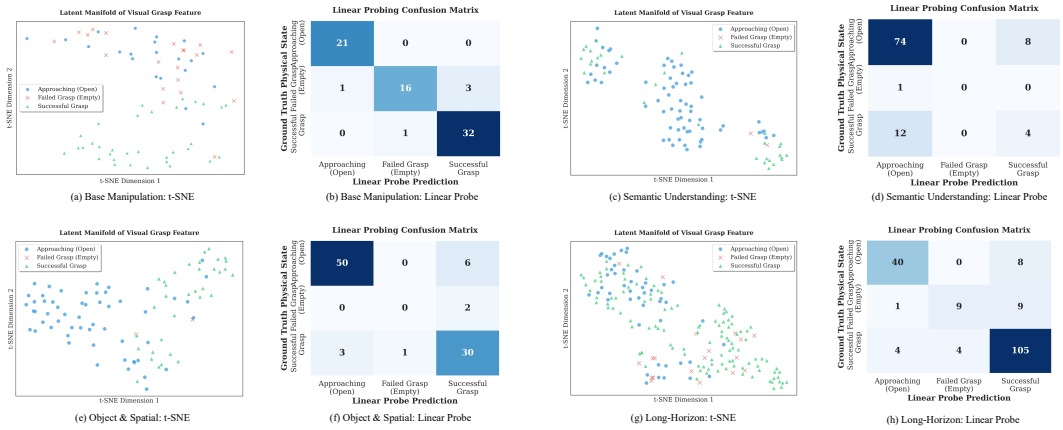


Figure 4: **Cross-Task Qualitative and Quantitative Analysis of Latent Manifold.** The figure displays the t-SNE dimensionality reduction projections of 128-dimensional features and the cross-validation confusion matrices of pure Linear Probing across four tasks. Even facing severe reduction artifacts and visual overlap, the linear probe maintains extremely high physical state separation accuracy.

Adversarial samples representing "gripper closed but not touching object" are accurately expelled from the "successful grasp" cluster by high-dimensional boundaries, confirming the model evades shortcut traps of merely fitting motion coordinates.

Progressing to **Object**, **Long-Horizon**, and **Semantic** tasks, while the native manifold geometry becomes highly complex due to environmental noise and the 2D projections visually overlap, its linear classification accuracy in the native space remains high at 78%–87%. This ability to maintain a robust hyperplane under extremely dense semantic noise indicates the network does not mechanically memorize visual appearances but successfully extracts interaction semantics into the decision input of the UAV movement decoder.

B Implementation and Training Details

Architecture Specifications. Our proposed AIR-VLA+ builds upon the Physical Intelligence $\pi_{0.5}$ foundation model as the primary vision-language backbone. To construct the *input feature enhancement module*, we apply mean pooling over the sequence output of the prefix tokens (images and text) to derive a highly condensed global context representation, which is then temporally broadcast across the action sequence horizon.

The *implicit visual grasp projector* is instantiated as an MLP consisting of two linear layers with a ReLU activation, squeezing the latent dimension from the backbone width (256 dimensions) down to 128 dimensions to enforce physical-state discriminability.

For the *unidirectional information transmission cascade channel*, the fused input for the UAV movement decoder concatenates the visual features, the broadcast global context, the visual grasp features, and the intended action from the arm manipulation decoder. Crucially, a *stop-gradient operation* (`.detach()`) is explicitly applied to the manipulator's intended action tensor before concatenation, ensuring strict isolation of backpropagation dynamics.

The *feature-level MoE* routing network utilizes a two-layer MLP with a hidden dimension of 128, followed by a Softmax function to distribute weights across $N = 3$ linear movement experts. Finally, *physical hard isolation* is strictly enforced on the action dimension indices, mapping the first four dimensions strictly to the UAV movement decoder and the remaining dimensions to the arm manipulation decoder.

Two-Stage Training Paradigm. To prevent catastrophic forgetting of the foundational physical priors and stabilize the MoE routing, we devised a two-stage training pipeline:

- **Stage 1 (Visual Cascade Adaptation):** The network is initialized from the base $\pi_{0.5}$ weights. The entire framework, including the newly added visual extractor and asymmetric dual decoders, is trained for 30,000 steps. We employ the AdamW optimizer with a peak learning rate of 5×10^{-5}, warming up over 10,000 steps, and applying a cosine decay schedule.
- **Stage 2 (Feature MoE Fine-Tuning):** Initialized from the Stage 1 checkpoint, we freeze the vast majority of the foundational multimodal backbone to protect the representation manifold. Gradients are selectively unblocked only for the MoE routing network, the $N = 3$ movement experts, the visual grasp projector, and the manipulation projection decoder. This stage trains for an additional 10,000 steps with a reduced peak learning rate of 2×10^{-5}, with a warmup of 2,000 steps and decay to 1×10^{-5}.

Hardware, Inference, and Evaluation Setup. All training stages were conducted using Distributed Data Parallel (DDP) across two NVIDIA A100 GPUs, maintaining a global effective batch size of 32. Gradients were clipped at a maximum norm of 1.0 to ensure stability.

During inference, continuous trajectory generation is solved using an Euler ODE solver, configured with 10 Neural Function Evaluations (NFE) uniformly spaced across the integration time horizon $t \in [0, 1]$.

Finally, to ensure statistical reliability during evaluation, the reported performance metrics for each individual task are calculated as the average over 20 independent rollout trials.