AIR-VLA: Vision-Language-Action Systems for Aerial Manipulation

Jianli Sun [1 2] **Bin Tian** [1] **Qiyao Zhang** [3] **Chengxiang Li** [4] **Zihan Song** [5] **Zhiyong Cui** [6] **Yisheng Lv** [1]
Yonglin Tian [1 *]

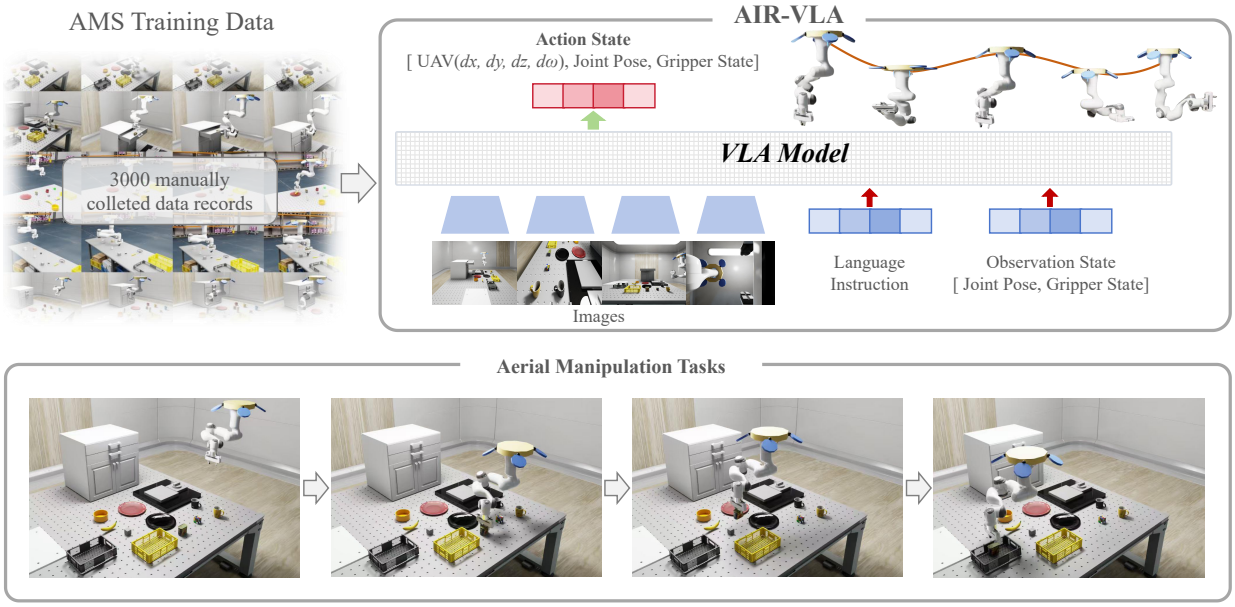


Figure 1. **Overview of AIR-VLA model framework and dataset of aerial manipulation tasks.**

Abstract

While Vision-Language-Action (VLA) models have achieved remarkable success in ground-based embodied intelligence, their application to Aerial Manipulation Systems (AMS) remains a largely unexplored frontier. The inherent characteristics of AMS, including floating-base dynamics, strong coupling between the UAV and the manipulator, and the multi-step, long-horizon nature of operational tasks, pose severe challenges to existing VLA paradigms designed for static or 2D mobile bases. To bridge this gap, we propose **AIR-VLA**, the first VLA benchmark specifically tailored for aerial manipulation. We construct a physics-based simulation environment and release a high-quality multimodal dataset comprising 3000 manually teleoperated demonstrations, covering base manipulation, object & spatial understanding, semantic reasoning, and long-horizon planning. Leveraging this platform, we systematically evaluate mainstream VLA models and state-of-the-art VLM models. Our experiments not only validate the feasibility of transferring VLA paradigms to aerial systems but also, through multi-dimensional metrics tailored to aerial tasks, reveal the capabilities and boundaries of current models regarding UAV mobility, manipulator control, and high-level planning. **AIR-VLA** estab-

[1]The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. [2]University of Chinese Academy of Sciences, Beijing 100049, China. [3]School of Automation, Beijing Institute of Technology, Beijing 100081, China. [4]School of Information and Intelligent Engineering, University of Sanya, Sanya 572000, Hainan Province, China. [5]School of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, Hunan Province, China. [6]State Key Lab of Intelligent Transportation Systems, School of Transportation Science and Engineering, Beihang University, Beijing, China.. Correspondence to: Yonglin Tian <yonglin.tian@ia.ac.cn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

lishes a standardized testbed and data foundation for future research in general-purpose aerial robotics. The resource of AIR-VLA will be available at https://github.com/SpencerSon2001/AIR-VLA.

1. Introduction

With breakthrough advancements in Large Language Models (LLMs) and Vision-Language Models (VLMs), Embodied AI is undergoing a paradigm shift from specialized skill learning to general-purpose intelligence. Recently, Vision-Language-Action (VLA) models, represented by RT-1 (Brohan et al., 2023), OpenVLA (Kim et al., 2025), and π_0 (Black et al., 2026), have demonstrated exceptional capability in handling open-world tasks driven by natural language instructions. However, existing VLA research is predominantly confined to Ground Mobile Manipulators, where the operational space is restricted to 2D planar navigation and limited working heights. This significantly constrains the potential of embodied intelligence in exploring broader three-dimensional physical spaces.

As a pivotal yet challenging branch of embodied intelligence, Aerial Manipulation Systems (AMS) integrate the high mobility of multi-rotor UAVs with the dexterous manipulation capabilities of robotic arms. By transcending the physical limitations of terrain and altitude, AMS demonstrate irreplaceable value in complex scenarios such as logistics transportation, disaster rescue, infrastructure inspection, and high-altitude operations. However, extending VLA models to aerial platforms introduces unique physical and control challenges. AMS must contend not only with the instability of a floating base and the kinematic coupling between the UAV and the manipulator, but also with the demands of 3D spatial reasoning within an expansive workspace and the inherently long-horizon, multi-step nature of operational tasks. Consequently, this requires that models possess not only general semantic understanding, but also the mastery of complex coordination strategies, ranging from 3D omnidirectional navigation to fine-grained end-effector manipulation, along with high-level planning capabilities for complex missions.

Despite its promising prospects, research on VLA for aerial manipulation currently faces a severe "data scarcity" and an "absence of benchmarks". Existing embodied intelligence benchmarks are predominantly designed for tabletop manipulation(Liu et al., 2023) or ground navigation(Yenamandra et al., 2023). Consequently, they fail to effectively evaluate omnidirectional mobility in 3D space, manipulation robustness under floating-base conditions, and long-horizon temporal reasoning capabilities conditioned on language. Furthermore, current research applying VLA to the UAV domain remains focused on navigation(Mehboob et al., 2026), task generation(Sautenkov et al., 2025), or single-DoF VLA manipulation tightly coupled with path planning(Mehboob et al., 2026), leaving the application of VLA to full-spectrum aerial manipulation tasks largely unexplored.

However, with the maturation of AMS hardware and control theory, combined with the rapid advancement of VLMs in robotics, applying the scene generalization capabilities of VLAs and the high-level planning abilities of VLMs to aerial manipulation tasks has demonstrated immense potential. Therefore, there is an urgent need in the academic community for a dataset and evaluation framework specifically focused on aerial manipulation.

To address these challenges, we propose **AIR-VLA**, the first VLA training and evaluation benchmark designed specifically for Aerial Manipulation Systems. We construct a simulation environment integrated with a physics engine based on NVIDIA Isaac Sim, simulating the collaborative operation of UAVs and manipulators under complex dynamic conditions. On this basis, we construct and release a multi-modal dataset covering tasks ranging from basic primitive actions to complex 3D dynamic interactions requiring long-horizon planning. Spanning diverse simulated scenarios, covering indoor rooms, factories, and outdoor environments, the dataset comprises **3000 high-quality manually teleoperated aerial manipulation episodes**. **AIR-VLA** aims to provide a comprehensive performance profile of aerial embodied intelligence across multiple dimensions, including basic motion control, fine-grained visual and spatial understanding, semantic instruction following, and long-horizon task planning.

The main contributions of this paper are summarized as follows:

- **Pioneering Aerial Manipulation VLA Benchmark:** We propose the first VLA benchmark testbed specifically designed for AMS, filling the evaluation gap in the domain of 3D aerial manipulation and providing a standardized research tool for the community. Tailored to the unique characteristics of aerial operations, we design a multi-suite dataset rich in sensory information (RGB, depth, proprioception) and diverse language instructions, providing high-quality data support for training aerial VLA manipulation policies.
- **Tailored Evaluation Metrics:** Addressing the unique characteristics of aerial manipulation, specifically the tight coordination between the UAV and manipulator and the long-horizon, multi-step nature of tasks, we move beyond relying solely on task success rates. We introduce multi-dimensional evaluation metrics designed to assess the aerial manipulation performance of VLAs and the long-horizon planning capabilities of

VLMs. These metrics, which cover dimensions such as UAV localization, navigation, and spatial perception, allow a deep exploration of the performance boundaries of the models.

- **Comprehensive Baseline Analysis:** We conduct extensive benchmarking and in-depth analysis of mainstream VLA and VLM models on **AIR-VLA**. By quantifying the performance of current mainstream VLA models on aerial tasks and the high-level planning capabilities of VLMs, we reveal critical challenges in the transfer process from ground to aerial platforms and provide guidance for future research directions.

2. Related Works

2.1. Aerial Manipulation Systems

AMS extend UAV capabilities from passive reconnaissance to active physical interaction, though the coupling between the floating base and manipulator introduces severe nonlinear dynamics. In mechanical design, research has prioritized bio-inspired and compliant architectures to enhance adaptability and stability, such as owl-inspired limbs (Liu et al., 2024) and continuum elephant-trunk manipulators (Peng et al., 2025). Dual-arm systems have also been explored to execute complex multi-step tasks infeasible for single-arm platforms (Ghorbani et al., 2023). On the control frontier, strategies address external disturbances and uncertainties through refined anti-disturbance techniques (Liu et al., 2024), adaptive prescribed performance control (Liang et al., 2023), and Reinforcement Learning (RL)-driven hierarchical policies (Deshmukh et al., 2025). Additionally, visual impedance control (Xu et al., 2023) and whole-body planning (Wang et al., 2025a) are employed to improve interaction precision. Despite these advances, traditional AMS frameworks generally lack semantic grounding. Operating under closed-world assumptions prevents them from generalizing to unstructured environments or interpreting abstract instructions (Zhang et al., 2025; Carvajal et al., 2024; Deng et al., 2025), limiting their utility as intelligent assistants.

2.2. Vision-Language-Action Models

VLA models shift Embodied AI from modular pipelines to end-to-end architectures, unifying perception, language, and action into a single transformer backbone. Foundational models like RT-1 (Brohan et al., 2023) and OpenVLA (Kim et al., 2025) demonstrate that large-scale training enables open-vocabulary task execution. The field is rapidly evolving toward high-frequency control and long-horizon capabilities. The π series, including π_0 (Black et al., 2026) and $\pi_{0.5}$ (Black et al., 2025), introduces flow-matching for smooth action generation. Notably, $\pi_{0.6}$ (Physical Intelligence et al., 2025) exhibits robust autonomous execution of long-horizon tasks, indicating that VLAs are maturing into reliable systems for complex physical interaction.

2.3. VLA for Aerial Robotics

Recognizing VLA potential, recent works have integrated these models into aerial platforms, focusing primarily on navigation and trajectory generation. Benchmarks like *IndoorUAV* (Liu et al., 2026) leverage task decomposition and π_0 for 3D language grounding, while *RaceVLA* (Serpiva et al., 2025) enables high-speed agility via video-conditioned control. To address onboard constraints, frameworks utilizing 3D Gaussian Splatting (Wu et al., 2025), fine-grained instruction mapping (Wang et al., 2026), and cognitive reasoning fusion (Mehboob et al., 2026) have been proposed. However, current applications remain predominantly confined to trajectory planning or high-level mission generation. Approaches like UAV-VLA (Sautenkov et al., 2025) and DroneVLA (Mehboob et al., 2026) largely limit the VLA's role to mission planning or simple 1-DoF gripping. These methods fail to leverage the UAV's potential for complex physical interaction in unstructured environments. The lack of a mature framework for full-spectrum aerial manipulation represents a critical research gap.

2.4. Benchmarks for Robot Manipulation

High-quality benchmarks drive VLA research but currently focus on ground-based platforms. For static manipulation, *LIBERO* (Liu et al., 2023) and *RoboTwin 2.0* (Chen et al., 2025) evaluate lifelong learning and dual-arm coordination, respectively. For mobile manipulation, *OVMM* (Yenamandra et al., 2023) and *Mobile ALOHA* (Fu et al., 2024) address open-vocabulary tasks and low-cost imitation learning. While comprehensive for ground robots, these benchmarks are ill-suited for AMS. Specifically, they lack the robotic platforms, corresponding kinematic designs, and aerial manipulation task data specific to Aerial Manipulation Systems. This absence restricts the development of aerial VLA models. To bridge this gap, we propose *AIR-VLA*, the first benchmark specifically designed to extend VLA intelligence to the challenging realm of aerial mobile manipulation.

3. AIR-VLA

To comprehensively evaluate the generalization capability and policy robustness of VLA models in the aerial manipulation domain, AIR-VLA is systematically designed across four dimensions: task taxonomy, evaluation framework, simulation environment, and dataset construction. This design aims to provide standardized metrics while simulating authentic physical characteristics.

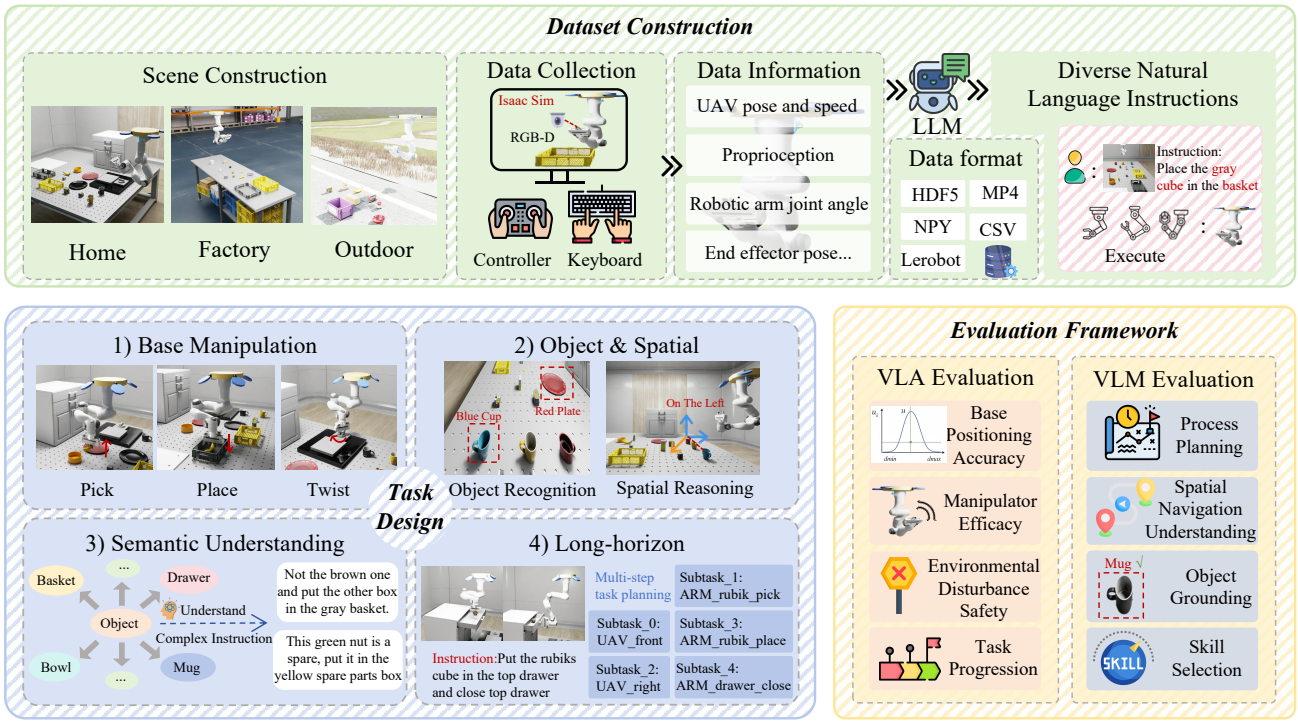


Figure 2. **Overview of the AIR-VLA benchmark.** AIR-VLA serves as a full-stack Vision-Language-Action platform tailored for aerial manipulation systems. It integrates a simulation-based teleoperation data acquisition pipeline, an online simulation environment, and diverse multimodal datasets. Furthermore, it provides a comprehensive benchmark for evaluating mainstream VLA and VLM models across varied aerial manipulation tasks.

3.1. Task Design Taxonomy

AIR-VLA comprises four core task suites targeting manipulation precision, semantic reasoning, visual perception, and planning. Unlike ground-based benchmarks, it emphasizes deep coordination between 3D UAV positioning and manipulator actions, covering a spectrum from atomic actions to complex interactions. Tailored to AMS characteristics, the dataset features an average task length of approximately 475 time steps, significantly exceeding traditional benchmarks to reflect temporal complexity.The specific objectives of the four suites are:

- **Base Manipulation:** Evaluates low-level motion control and UAV-arm coordination with minimal distractors. Tasks prioritize optimal 3D hovering and precise manipulation, requiring the model to handle complex contact dynamics.
- **Object & Spatial:** Assesses fine-grained understanding of physical attributes (color, shape, material) and spatial relationships. This suite imposes stringent requirements on cross-modal alignment and 3D geometric reasoning.
- **Semantic Understanding:** Tests robustness to unstructured language instructions. Tasks feature diverse styles and implicit intents, requiring the model to infer underlying goals from complex natural language.
- **Long-Horizon:** Evaluates multi-step reasoning with temporal dependencies. The AMS must alternate between wide-range maneuvers and local fine-grained operations. Any coordination failure in the chain results in total task failure, representing the most challenging scenario.

3.2. Evaluation Framework

Compared to traditional ground robot tasks, aerial mobile manipulation introduces unique challenges such as dynamic coupling of the floating base, volumetric workspaces, and temporal complexity of long-horizon tasks. To comprehensively evaluate the limit performance of embodied intelligence models in this high-dimensional domain, we constructed a two-layer evaluation framework covering perception, planning, and control. For VLA models, we conduct closed-loop evaluations based on an online simulation environment, focusing on UAV-Arm Coordination capabilities

under real-time inference. Specifically, we quantify the feasibility of UAV spatial navigation, the manipulation precision of the end-effector, and dynamic safety during long-horizon task execution. For long-horizon tasks, considering the critical role of logical reasoning, we propose a multi-dimensional evaluation metric set for the high-level task planning capabilities of VLMs. This assesses the embodied reasoning ability of VLMs from dimensions including sub-task process construction, 3D spatial perception, fine-grained object grounding, and atomic skill selection. **For detailed definitions and mathematical formulations of these metrics, please refer to Appendix A.3.**

3.3. Simulation Environment

Simulator. Built on NVIDIA Isaac Sim, AIR-VLA leverages the Omniverse platform to bridge the Sim-to-Real gap. We utilize the PhysX 5 engine and ray-tracing to ensure realistic simulation from fluid dynamic disturbances to lighting reflections. Furthermore, the platform's USD pipeline and GPU parallel acceleration enable flexible scene configuration and efficient data collection.

Robot System. To fully test the adaptability of VLA models to high-dimensional state spaces, we constructed an AMS consisting of a quadrotor UAV equipped with a 7-DoF Franka Panda manipulator. The UAV control space is defined as position changes in Cartesian coordinates and Yaw rotation; the manipulator control space includes 7 joint poses and the gripper state. This configuration constitutes a high-dimensional control problem characterized by high redundancy and coupling.

3.4. Dataset Construction

Data collection utilizes the Isaac Sim-based environment, emphasizing diversity and physical realism.

Data Collection Strategy. To capture atypical aerial coordination patterns, we employ **Human Teleoperation** rather than scripted generation. Expert demonstrators utilize gamepad interfaces to generate task-oriented trajectories, covering multiple collaborative solutions for the UAV and manipulator within unified tasks.

Multimodal Perception System. Tailored to aerial perspectives, the sensor configuration comprises: (1) a UAV front-down RGB-D camera for global bird's-eye views, (2) a manipulator wrist RGB-D camera for local fine-grained observation, and (3) a traditional third-person perspective camera.

Action and State Space. The dataset records full proprioceptive information, including the UAV's 4D pose and derivatives (velocity/angular velocity), manipulator joint poses, and end-effector pose. Standardized data interfaces ensure compatibility with the input layers of diverse VLA models.

Scene and Instruction Diversity. To enhance generalization, we constructed diverse environments (residential, industrial, outdoor) with varying lighting conditions. Furthermore, LLMs are leveraged to generate natural language instructions featuring complex structures and implicit intents, ensuring dense coverage of the semantic space.

4. Experiments

4.1. VLA Experiments

We evaluated a series of advanced VLA models to assess their performance in aerial manipulation tasks. While VLA models have demonstrated superior efficacy in fixed-base and ground mobile manipulation tasks, we apply them to AMS for the first time. Our experiments are structured around the following research questions:

Q1: Can VLA models maintain superior task completion performance in high-DoF (12-DoF in experiments) robotic system manipulation tasks?

Q2: Can VLA models cope with external disturbances in AMS and complete tasks under random base jitter?

Q3: What is the dependency of VLA models on visual inputs in AMS tasks, and which viewing angles significantly enhance task completion?

Q4: How do VLA models perform in UAV movement and manipulator operation respectively in AMS tasks, and how is their coordination?

Q5: Safety constraints are critical in robotic tasks. Do VLA models introduce more severe safety issues in AMS tasks?

4.1.1. Experimental Setup

Baseline Models. To establish a representative benchmark, we evaluate six diverse models: π_0 (Black et al., 2026) and $\pi_{0.5}$ (Black et al., 2025), Flow Matching-based foundation models pre-trained on cross-embodiment data, represent large-scale transfer capabilities to the aerial domain. π_0**-FAST**(Pertsch et al., 2025) serves as an efficiency-optimized variant, utilizing step distillation to enable the high-frequency inference crucial for agile aerial maneuvers. **ACT** (Zhao et al., 2023) combines CVAEs and Transformers to predict "Action Chunks," effectively modeling long-term dependencies. **Diffusion Policy** (Chi et al., 2025) utilizes conditional diffusion for iterative denoising, robustly handling multi-modal distributions essential for floating-base stability.

Dataset & Training. The training dataset is derived from our human-teleoperated simulation data, fully reflecting the diversity of collaborative actions between UAV and manipulator. We adopted a differentiated fine-tuning strategy based on task difficulty: 30 trajectories were used for fine-tuning simple single-step tasks with less physical interaction, while 50 trajectories were used for multi-step long-horizon tasks or tasks with rich physical interactions.

Robustness Evaluation. To simulate real flight environments (Sim-to-Real Gap), we introduced dynamic disturbance test scenarios. During testing, Gaussian noise was injected into the UAV's state space to simulate the effects of wind gusts or sensor drift on flight stability. Additionally, we designed a "No Fixed Third-Person View" test, retaining only onboard and wrist cameras, to investigate the model's dependency on global perspectives.

4.1.2. Main Results and Analysis

In this section, we report the quantitative performance of baseline models on the four task suites of AIR-VLA. Table 1 presents the comprehensive evaluation results under standard disturbance-free conditions.

Overall Performance. Experimental results indicate that large-scale pre-trained models, represented by $\pi_{0.5}$ and π_0, demonstrate significant advantages in the AIR-VLA evaluation, outperforming traditional imitation learning baselines such as ACT and Diffusion Policy across all metrics. However, in the complex domain of aerial manipulation tasks, $\boldsymbol{\pi_0}$-FAST exhibits a marked performance decline compared to π_0, suggesting that the performance degradation induced by distillation and step reduction is significantly amplified in aerial scenarios.

Overall, the results validate the feasibility of transferring VLA paradigms to Aerial Manipulation Systems. Furthermore, they demonstrate that pre-training on massive cross-embodiment data (including mobile robotics) enables models to acquire generalizable mechanical manipulation priors. Notably, even under few-shot fine-tuning settings with only 30–50 demonstrations, foundation models like $\pi_{0.5}$ rapidly adapt to aerial manipulation paradigms unseen during pre-training, achieving the highest performance among all evaluated models. Nevertheless, there remains substantial room for improvement in overall task completion rates. Compared to low-DoF ground-based platforms, the performance of existing VLA models on high-DoF aerial platforms remains suboptimal.

π_0 achieves its peak success rate in Base Manipulation tasks characterized by minimal environmental interference; however, its performance exhibits a significant decaying trend with the introduction of intense visual distractors, complex semantic instructions, and long-horizon planning requirements. Specifically, within the Object & Spatial suite, dense visual distractors lead to a marked increase in target misidentification rates. Notably, in spatial understanding tasks, the models exhibit Spatial Grounding Failure: although the correct object category is identified, the agent manipulates an identical object at an incorrect location due to an inability to accurately parse relative positional instructions. In comparison, $\pi_{0.5}$ demonstrates superior robustness against object interference relative to π_0. In the Semantic Understanding suite, where visual scene complexity is controlled to be comparable to that of the Object & Spatial suite, we observe a slight performance decline in both $\pi_{0.5}$ and π_0, suggesting that complex semantic instructions impose a distinct penalty on model performance. In the Long-Horizon suite, we observe that successful executions are predominantly confined to the initial sub-task, with performance on the subsequent sub-task exhibiting notable degradation compared to the first.

Regarding collaborative control, VLA models demonstrate significantly better performance in UAV motion control than in manipulator operation. This is partly attributable to the fact that UAV movement in aerial manipulation is relatively coarse-grained, whereas manipulator operation necessitates extremely high precision. Existing models have successfully acquired UAV navigation strategies to guide the system to feasible operating positions; however, executing fine-grained manipulation at these positions remains a critical technical bottleneck. Furthermore, safety constraints cannot be overlooked. Due to the inherent characteristics of the floating base, collisions and unreasonable physical interactions cause significantly more severe disturbances to the system than in ground-based robotics. Our evaluation revealed destructive interactions with the environment in certain episodes, indicating that addressing safety constraints in Aerial Manipulation Systems remains a pivotal direction for future research.

Robustness Analysis. We selected the best-performing model, $\pi_{0.5}$, for UAV dynamic disturbance and third-person lacking tests (results in Table 2). Under UAV disturbance, the decline in performance metrics for $\pi_{0.5}$ is relatively small, indicating that VLA models possess a degree of implicit compensation capability. However, in the absence of fixed third-person view input, model performance decays significantly. This reveals the current models' dependency on stable global perspectives; the lack of a global view severely limits the system's executable space and exploration capabilities. Introducing active perception strategies based on onboard UAV cameras may be an effective solution to this problem.

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 $w_{pos} = 0.25, w_{arm} = 0.25, w_{safe} = 0.10, w_{task} = 0.40$. 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}).

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	2.00	**32.3**	**65.1**	38.5	85.9	2.0	**32.3**	58.8	42.0	83.5	3.50	**34.5**
$\pi_{0.5}$	**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**

Table 2. **Robustness evaluation of $\pi_{0.5}$ under disturbance and perception-deprived conditions.** The table shows absolute scores for each metric, with values in parentheses indicating the performance drop ($\downarrow \Delta$) compared to the Standard condition.

Condition	S_{pos}	S_{arm}	S_{safe}	S_{task}	S_{total}
Standard (None)	**71.3**	**45.9**	**84.7**	**10.6**	**42.0**
UAV Disturb.	68.2 (↓3.1)	45.6 (↓0.3)	83.2 (↓1.5)	10.5 (↓0.1)	41.0 (↓1.0)
View Lacking	58.9 (↓12.4)	35.6 (↓10.3)	85.6 (↑0.9)	5.1 (↓5.5)	34.5 (↓7.5)

4.2. VLM Experiments

Leveraging the capability of VLMs to integrate visual and linguistic instructions for high-level planning, we conduct a comprehensive evaluation of several mainstream VLMs. Specifically, we assess their planning performance across the dimensions of Process Planning, Spatial Navigation, Object Grounding, and Skill Selection, aiming to verify their potential for aerial manipulation tasks.

4.2.1. Experimental Setup

Baseline Models. To establish a representative benchmark, we analyze six diverse open-source VLMs. **Molmo-7B-D-0924** (Deitke et al., 2024) employs a data-centric strategy via the PixMo dataset for efficient alignment. The Qwen series, comprising **Qwen3-VL-8B-Instruct** (Bai et al., 2025a) and **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025b), excels in adaptability; the former leverages dynamic resolution, while the latter utilizes M-RoPE to unify image and video understanding. **GLM-4V-9B** (GLM Team et al., 2024) focuses on deep vision-language alignment for high-resolution features. **InternVL3_5-8B** (Wang et al., 2025b) adopts a "strong vision encoder + strong language decoder" design, targeting geometric perception. Finally, **llava-onevision-qwen2-7b-ov-hf** (Li et al., 2024) balances static and temporal analysis through a unified mixed-data paradigm.

Assessment Protocol. In the VLM evaluation phase, we adopted a non-interactive evaluation process. We input the initial four-view images of the task scene and description instructions to the VLM, providing skill library descriptions, output format requirements, and few-shot examples as System Prompts. The VLM is required to generate high-level task planning sequences conforming to specifications, enabling standardized capability assessment.

4.2.2. Results and Analysis

Quantitative results demonstrate that **Qwen3-VL-8B-Instruct** and **Qwen2.5-VL-7B-Instruct** establish a substantial performance advantage in high-level planning for aerial manipulation tasks. Notably, Qwen3-VL achieves state-of-the-art (SOTA) performance across all baseline models in four core dimensions: Process Planning, Spatial Navigation, Object Grounding, and Skill Selection, highlighting its exceptional potential as a multimodal "brain."

Regarding metric distribution, the capability profiles across models exhibit high consistency. While demonstrating strong performance in semantic understanding and skill matching, all models reveal a significant deficiency in **Spatial Navigation**. In-depth analysis indicates that this lack of 3D spatial awareness is the primary bottleneck limiting the end-to-end planning Success Rate.

In terms of instruction robustness, the performance gap between explicit, structured instructions and implicit, complex instructions is negligible. This suggests that current VLMs possess powerful semantic generalization capabilities, effectively aligning unstructured linguistic inputs with specific operational intents. Regarding task scenarios, VLMs demonstrate excellent temporal stability; their performance in Long-Horizon tasks does not exhibit significant deterioration, standing in sharp contrast to the **performance decay** often observed in traditional VLA models over long sequences. However, in scenarios introducing visual Distractor Interference, most models show a slight decline in object grounding metrics, reflecting the challenge complex backgrounds pose to fine-grained visual perception.

In summary, VLMs hold immense potential for high-level planning in aerial manipulation, particularly in mitigating the long-horizon reasoning limitations of VLA models. However, overcoming the critical bottleneck of **spatial ori-**

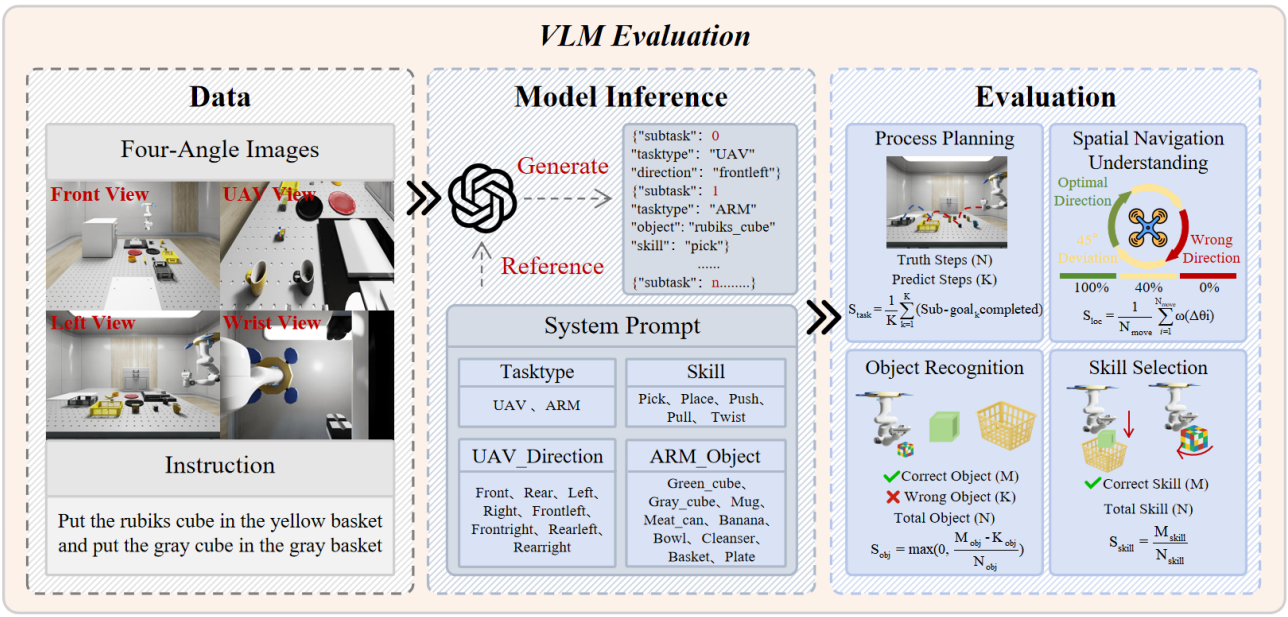


Figure 3. **Evaluation pipeline for VLM high-level planning capabilities in aerial manipulation tasks.**

Table 3. **Detailed evaluation of VLM task planning capabilities.** The table displays normalized sub-metric scores and planning success rates (Succ, %) for each model across different task scenarios and instruction types. Models are sorted in ascending order by their Overall Average Total score. Abbreviations: **Exp.**: Explicit; **Imp.**: Implicit. Weights: $w_{plan} = 0.25, w_{loc} = 0.25, w_{obj} = 0.25, w_{skill} = 0.25$.

Model	Instr.	Base Manipulation						Distractor Interference						Long-Horizon						Overall Average					
		Plan	Loc	Obj	Skill	**Tot**	Succ	Plan	Loc	Obj	Skill	**Tot**	Succ	Plan	Loc	Obj	Skill	**Tot**	Succ	Plan	Loc	Obj	Skill	**Tot**	Succ
LLaVA-OV	Exp.	10.0	38.0	0.0	80.0	**32.0**	0.0	12.0	38.0	0.0	75.0	**31.3**	0.0	72.0	29.3	22.5	62.5	**46.6**	0.0	31.8	35.4	9.6	73.2	**37.5**	0.0
	Imp.	6.0	38.0	0.0	85.0	**32.3**	0.0	22.0	43.0	10.0	70.0	**36.3**	0.0	68.7	26.0	25.0	66.7	**46.6**	0.0						
Molmo-7B	Exp.	33.0	42.0	0.0	70.0	**36.3**	0.0	38.0	47.0	0.0	50.0	**33.8**	0.0	72.0	37.8	13.3	77.5	**50.2**	0.0	47.6	40.4	4.9	61.2	**38.5**	0.0
	Imp.	27.0	37.0	0.0	55.0	**29.8**	0.0	43.0	49.0	0.0	40.0	**33.0**	0.0	72.4	29.5	15.8	75.0	**48.2**	0.0						
InternVL3.5	Exp.	23.0	0.0	30.0	100.0	**38.3**	0.0	32.5	8.0	20.0	80.0	**35.1**	0.0	37.7	3.5	100	100	**60.3**	0.0	31.8	3.2	56.1	89.4	**45.2**	0.0
	Imp.	30.0	0.0	80.0	100	**52.5**	0.0	32.0	8.0	30.0	80.0	**37.5**	0.0	35.6	0.0	76.7	76.7	**47.2**	0.0						
GLM-4V	Exp.	76.5	7.00	65.0	90.0	**59.6**	10.0	81.0	14.0	20.0	85.0	**50.0**	0.0	76.1	26.7	36.7	75.0	**53.6**	10.0	79.8	16.8	46.2	84.4	**56.8**	6.7
	Imp.	80.0	11.0	60.0	100.0	**62.8**	10.0	79.0	13.0	40.0	80.0	**53.0**	0.0	86.0	29.3	55.8	76.7	**62.0**	10.0						
Qwen2.5-VL	Exp.	94.0	37.0	75.0	100	**76.5**	10.0	92.0	46.0	50.0	90.0	**69.5**	0.0	76.2	27.2	60.9	75.0	**59.8**	20.0	92.2	38.7	75.1	88.9	**73.7**	13.3
	Imp.	100.0	37.0	100.0	100.0	**84.3**	20.0	98.0	44.0	95.0	80.0	**79.3**	10.0	93.0	41.2	70.0	88.3	**73.1**	20.0						
Qwen3-VL	Exp.	93.0	36.0	100.0	100.0	**82.3**	20.0	100.0	49.0	90.0	100	**84.8**	20.0	98.7	43.5	90.0	90.0	**80.5**	30.0	96.9	43.5	93.3	96.7	**82.4**	23.3
	Imp.	93.0	38.0	100.0	100.0	**82.8**	30.0	100.0	44.0	90.0	100.0	**83.5**	10.0	96.7	46.0	90.0	90.0	**80.7**	30.0						

entation understanding remains a pivotal challenge for achieving highly reliable aerial agents in the future.

5. Limitations

While AIR-VLA provides a comprehensive benchmark for aerial manipulation, several limitations remain to be addressed in future work. First, the current framework does not adequately consider the dynamic coupling between the robotic arm and the UAV under payload conditions. Manipulating objects with varying masses shifts the system's center of mass and introduces complex inertial disturbances, which poses significant challenges to flight stability and control precision that are not fully modeled in our existing evaluation. Second, the current models exhibit a lack of generalization capability in large-scale workspace operations.

Although the policies perform well in relatively confined and structured scenarios, their spatial reasoning and navigation capabilities struggle to generalize effectively across expansive, unconstrained 3D environments. Future research should focus on integrating payload-aware dynamic compensation and enhancing the spatial generalization of VLA models for large-scale aerial tasks.

6. Conclusion

This paper presents **AIR-VLA**, the first full-stack VLA benchmark for Aerial Manipulation Systems. Through the construction of a simulation environment, a physics-aware dataset, and a multi-dimensional evaluation framework, we systematically quantify the capability boundaries of current mainstream VLA models in 3D aerial manipulation tasks. Our findings reveal that while transferring pre-trained VLA models to aerial platforms is feasible, existing models still face severe challenges in handling floating-base dynamic coupling, 3D long-horizon planning, and disturbance rejection. AIR-VLA not only provides a standardized evaluation tool for the community but also highlights a critical future direction: tightly integrating explicit multi-body dynamics priors with the generalized reasoning capabilities of large models. We hope this benchmark will inspire further exploration into general-purpose aerial embodied intelligence, propelling robotics from the 2D ground into the vast 3D space.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (52441202, 62303460), the Key Research and Development Program of Shaanxi Province (2024CY2-GJHX-49) and the Science and Technology Development Fund of Macau SAR (No. 0093/2023/RIA2 and 0145/2023/RIA3) ..

Impact Statement

This paper presents work whose goal is to advance the field of aerial robotics and embodied AI. The primary societal implication of this research is the potential development of more autonomous and versatile unmanned aerial vehicles (UAVs), which could benefit applications such as search and rescue, logistics, and infrastructure inspection. However, integrating VLMs into physical control systems introduces challenges related to safety, reliability, and the potential for unintended physical actions due to model hallucinations or grounding failures. By systematically evaluating these failure modes, our work aims to promote the safe and robust deployment of such systems. We do not feel there are other specific negative societal consequences that must be highlighted here.

References

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-VL technical report, 2025a.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-VL technical report, 2025b.

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Galliker, M. Y., Ghosh, D., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A. Z., Shi, L. X., Smith, L., Springenberg, J. T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., and Zhilinsky, U. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. In *9th Annual Conference on Robot Learning*, 2025.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L. X., Tanner, J., Vuong, Q., Walling, A., Wang, H., and Zhilinsky, U. π_0: A Vision-Language-Action Flow Model for General Robot Control, 2026.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems XIX*, 2023.

Carvajal, C. P., Andaluz, G. M., Andaluz, V. H., Roberti, F., Palacios-Navarro, G., and Carelli, R. 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.

Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Li, Z., Liang, Q., Lin, X., Ge, Y., Gu, Z., Deng, W., Guo, Y., Nian, T., Xie, X., Chen, Q., Su, K., Xu, T., Liu, G., Hu, M., ang Gao, H., Wang, K., Liang, Z., Qin, Y., Yang, X., Luo, P., and Mu, Y. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation, 2025.

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11): 1684–1704, 2025.

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.-H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N. A., Hajishirzi, H., Girshick, R., Farhadi, A., and Kembhavi, A. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models, 2024.

Deng, W., Chen, H., Ye, B., Chen, H., Li, Z., and Lyu, X. Whole-body integrated motion planning for aerial manipulators. *IEEE Transactions on Robotics*, 41:6661–6679, 2025. doi: 10.1109/TRO.2025.3626619.

Deshmukh, S., Alonso-Mora, J., and Sun, S. Global end-effector pose control of an underactuated aerial manipulator via reinforcement learning, 2025.

Fu, Z., Zhao, T. Z., and Finn, C. Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *8th Annual Conference on Robot Learning*, 2024.

Ghorbani, S., Samadikhoshkho, Z., and Janabi-Sharifi, F. 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.

GLM Team, Zeng, A., Xu, B., Wang, B., et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools, 2024.

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., and Finn, C. OpenVLA: An open-source vision-language-action model. In Agrawal, P., Kroemer, O., and Burgard, W. (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 06–09 Nov 2025.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., and Li, C. LLaVA-OneVision: Easy visual task transfer, 2024.

Liang, J., Chen, Y., Wu, Y., Miao, Z., Zhang, H., and Wang, Y. 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.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791. Curran Associates, Inc., 2023.

Liu, Q., Liu, Y., Chen, Z., Guo, K., Yu, X., Zhang, Y., and Guo, L. 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.

Liu, X., Liu, Y., Qiu, H., Qirong, Y., and Lian, Z. IndoorUAV: Benchmarking vision-language UAV navigation in continuous indoor environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23864–23872, 2026.

Mehboob, F., James, M. W., Habel, A. A., Sam, J., Altamirano Cabrera, M., and Tsetserukou, D. DroneVLA: VLA-based aerial manipulation. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1135–1139, 2026.

Peng, R., Wang, Y., Lu, M., and Lu, P. 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.

Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., and Levine, S. FAST: Efficient action tokenization for vision-language-action models, 2025.

Physical Intelligence, Amin, A., Aniceto, R., et al. $\pi^*_{0.6}$: A VLA That Learns From Experience, 2025.

Sautenkov, O., Yaqoot, Y., Lykov, A., Mustafa, M. A., Tadevosyan, G., Akhmetkazy, A., Cabrera, M. A., Martynov, M., Karaf, S., and Tsetserukou, D. UAV-VLA: Vision-language-action system for large scale aerial mission generation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 1588–1592, 2025. doi: 10.1109/HRI61500.2025.10974117.

Serpiva, V., Lykov, A., Myshlyaev, A., Khan, M. H., Abdulkarim, A. A., Sautenkov, O., and Tsetserukou, D. RaceVLA: VLA-based racing drone navigation with human-like behaviour, 2025.

Wang, M., Lyu, S., Liu, Q., Yang, Z., Guo, K., and Yu, X. Precise end-effector control for an aerial manipulator under composite disturbances: Theory and experiments. *IEEE Transactions on Automation Science and Engineering*, 22:4006–4021, 2025a. doi: 10.1109/TASE.2024.3406754.

Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., and Luo, G. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b.

Wang, X., Yang, D., Liao, Y., Zheng, W., Dai, B., Li, H., Liu, S., et al. UAV-Flow Colosseo: A real-world benchmark for flying-on-a-word UAV imitation learning. *Advances in Neural Information Processing Systems*, 38, 2026.

Wu, Y., Zhu, M., Li, X., Du, Y., Fan, Y., Li, W., Han, Z., Zhou, X., and Gao, F. VLA-AN: An efficient and onboard vision-language-action framework for aerial navigation in complex environments, 2025.

Xu, M., Hu, A., and Wang, H. 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.

Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A. S., Khanna, M., Gervet, T., Yang, T.-Y., Jain, V., Clegg, A., Turner, J. M., et al. HomeRobot: Open-vocabulary mobile manipulation. In *Conference on Robot Learning*, pp. 1975–2011. PMLR, 2023.

Zhang, Z., Yu, H., Chai, Y., Yang, Z., Liang, X., Fang, Y., and Han, J. 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.

Zhao, T., Kumar, V., Levine, S., and Finn, C. Learning fine-grained bimanual manipulation with low-cost hardware. *Robotics: Science and Systems XIX*, 2023.

A. Benchmark Implementation Details

A.1. Task Description and Taxonomy

Task Overview. AIR-VLA aims to establish a comprehensive evaluation platform covering the full skill spectrum of aerial mobile manipulation. We have constructed a diverse family of tasks comprising 60 atomic and composite tasks, logically categorized into four core suites: Base Manipulation, Object & Spatial, Semantic Understanding, and Long-Horizon. The complete task list is detailed in Table 5. These tasks range from omnidirectional pick-and-place in 3D space and precision insertion in confined spaces, to articulated object interaction involving drawers and cabinets, and twist-and-rotate operations requiring torque control. Furthermore, the benchmark integrates hybrid navigation-manipulation tasks, requiring agents to seamlessly switch between extensive scene movement and local fine-grained manipulation. Beyond basic motor skills, this benchmark emphasizes deep evaluation of advanced embodied intelligence across multiple dimensions, including voxel-level spatial perception, dynamic pose compensation of the floating base, cross-modal semantic generalization, and long-horizon temporal logical planning.

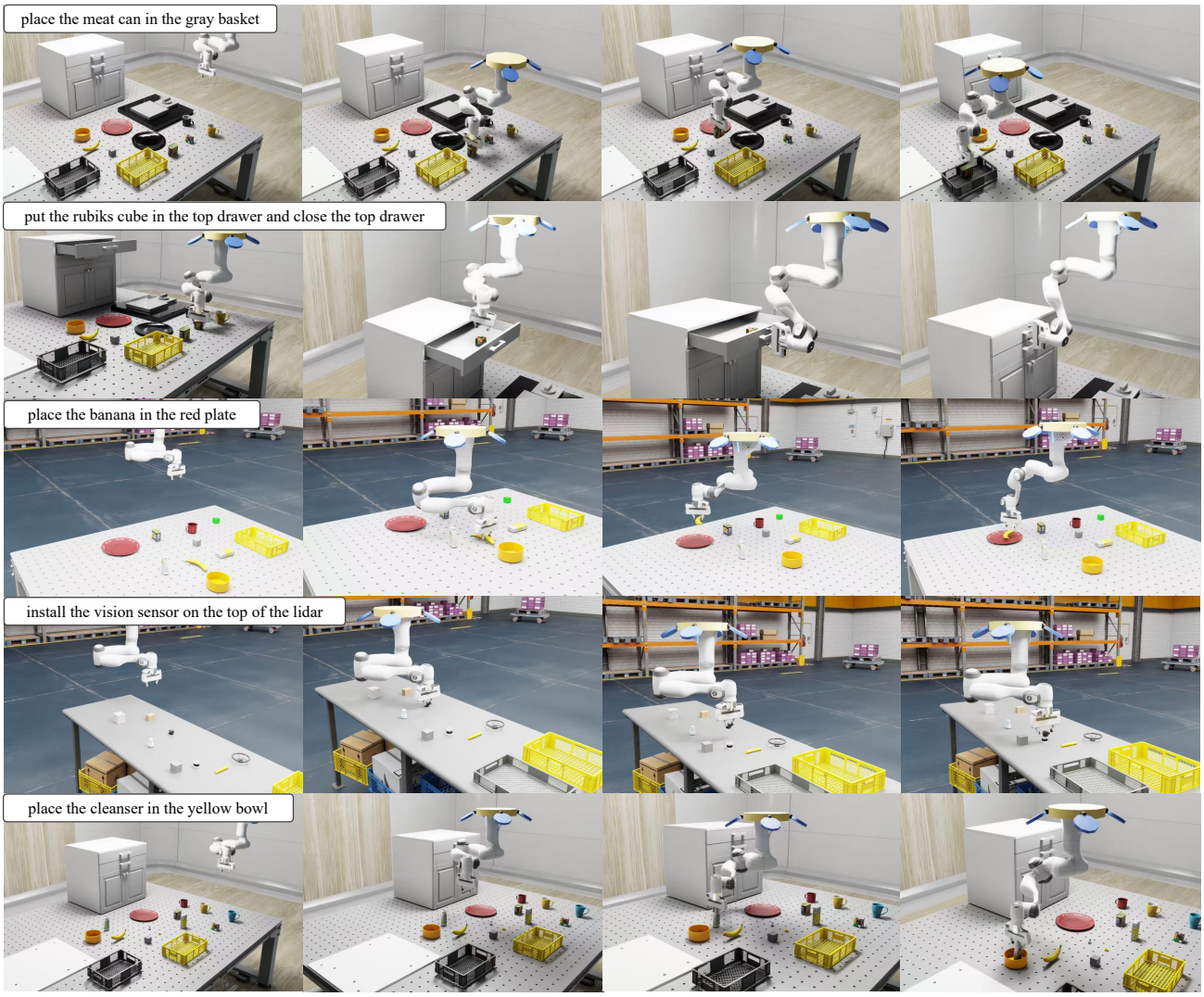


Figure 4. Task examples in AIR-VLA dataset.

Coupled Dynamics and Long-Horizon Reasoning. Unlike traditional benchmarks that often treat navigation and manipulation as decoupled phases or focus solely on 2D planar movement, the core feature of AIR-VLA is the emphasis

on spatio-temporal coupling between the UAV base and the manipulator. In our context, "reasoning" is endowed with a dual meaning: first, kinematic feasibility reasoning, where the agent must real-time assess whether the target is within the manipulator's reachable workspace under the floating state of the UAV with drift and jitter; second, long-horizon temporal planning, where early interaction actions (e.g., opening a drawer and maintaining a specific opening) must create necessary preconditions for subsequent geometric constraints (e.g., placing an object inside). Our dataset statistics indicate that the average episode length is approximately 475 time steps, with the Long-Horizon suite exceeding 600 time steps (at 20Hz control frequency), significantly surpassing traditional tabletop manipulation datasets. This long-horizon characteristic requires policies not only to maintain logical consistency over extended periods but also to continuously output control signals to stabilize the base and resist environmental disturbances while executing fine-grained sub-tasks.

A.2. Observation Space

To support perception and decision-making in complex environments, AIR-VLA constructs a multimodal observation space rich in semantic information. In terms of visual perception, the system synchronously captures RGB-D image streams from four perspectives: a front-down chassis camera provides a first-person flight view for global path planning and target approach; an eye-in-hand camera mounted on the manipulator end-effector resolves occlusion issues and supports fine manipulation; while global overhead and side-view cameras provide third-person god views to assist state estimation and training supervision. Regarding proprioception, the system records full state vectors at high frequency, including the UAV's position, orientation quaternion, linear velocity, and angular velocity in the world frame, the manipulator's 7 joint poses and velocities, as well as the end-effector's 6-DoF pose and gripper state. All sensory data is strictly timestamp-synchronized, providing a precise data alignment foundation for multimodal learning.



Figure 5. Observation Space. This figure illustrates instances of observations from various camera perspectives.

A.3. Evaluation Metrics

A.3.1. VLA Evaluation Metrics

Traditional binary metrics fail to capture continuous state changes in aerial manipulation. To evaluate UAV movement, coordination, and safety at a fine granularity, we propose a four-dimensional system. The total score S_{total} is a weighted sum of metrics $S_i \in [0, 100]$, where the maximum possible score is 100:

$$S_{total} = w_{pos}S_{pos} + w_{arm}S_{arm} + w_{safe}S_{safe} + w_{task}S_{task} \tag{1}$$

Base Positioning Accuracy (S_{pos}) This metric assesses if the UAV reaches a Feasible Workspace (collision-free Inverse Kinematics (IK) solution). We model positioning quality relative to optimal distance μ using a Gaussian decay function, scored hierarchically by interaction state on a 0-100 scale:

$$s_{pos}^{(i)} = \begin{cases} 100, & \text{if operation succeeds} \\ 100 \cdot \mathcal{G}(d_i; \lambda_{act}), & \text{if interaction attempted} \\ 100 \cdot \mathcal{G}(d_i; \lambda_{hov}), & \text{if hovering only} \end{cases} \tag{2}$$

where $\mathcal{G}(d; \lambda) = \lambda \cdot \exp\left(-\frac{(d-\mu)^2}{2\sigma^2}\right) \cdot \mathbb{I}(d_{min} \leq d \leq d_{max})$. $\lambda_{act} > \lambda_{hov}$ rewards active attempts.

Manipulator Efficacy (S_{arm}) To quantify resilience against floating-base jitter, we combine orientation consistency $S_{ori} \in \{0, 1\}$ (Euler angle alignment) and interaction precision $s_{int}^{(i)}$ to form a score out of 100:

$$S_{arm} = 100 \cdot \left(\beta S_{ori} + \frac{1-\beta}{N} \sum_{i=1}^{N} s_{int}^{(i)}\right) \tag{3}$$

where interaction precision uses a Gaussian kernel on end-effector distance d_{ee}:

$$s_{int}^{(i)} = \exp\left(-\frac{d_{ee}^2}{2\sigma_{arm}^2}\right) \cdot \mathbb{I}(d_{ee} \leq d_{thresh}) \tag{4}$$

Environmental Disturbance Safety (S_{safe}) Collision risk is measured via the total displacement Δ_{total} of non-target objects. The score decays from 100 (static environment) to 0 (exceeding limit Δ_{limit}):

$$S_{safe} = 100 \cdot \exp\left(-\frac{\Delta_{total}^2}{2\sigma_{safe}^2}\right) \cdot \mathbb{I}(\Delta_{total} < \Delta_{limit}) \tag{5}$$

Task Progression (S_{task}) For Long-Horizon Tasks with K sub-goals, we measure the percentage of completion to reveal planning decay:

$$S_{task} = \frac{100}{K} \sum_{k=1}^{K} \mathbb{I}(\text{Sub-goal}_k \text{ completed}) \tag{6}$$

A.3.2. VLM Task Planning Evaluation Metrics

We evaluate high-level planning across process logic, spatial understanding, object grounding, and skill selection using scores scaled to [0, 100]. The total weighted score is:

$$S_{total} = w_{plan} S_{plan} + w_{loc} S_{loc} + w_{obj} S_{obj} + w_{skill} S_{skill} \tag{7}$$

Process Planning Capability (S_{plan}) Evaluates action sequence correctness via **Type Accuracy** (agent match M) and **Length Consistency** (step count K vs. ground truth N), scaled to 100:

$$S_{plan} = 100 \cdot \left[\alpha \cdot \frac{M}{N} + (1-\alpha) \cdot \max\left(0, 1 - \frac{|K-N|}{N}\right)\right] \tag{8}$$

Spatial Navigation Understanding (S_{loc}) Measures UAV movement logic. For N_{move} steps, we score the deviation $\Delta\theta_i$ between predicted and optimal directions using a discrete point system:

$$\omega(\Delta\theta_i) = \begin{cases} 100, & \text{if } \Delta\theta_i = 0^\circ \text{ (Correct)} \\ 40, & \text{if } \Delta\theta_i = 45^\circ \text{ (Neighboring)} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$S_{loc} = \frac{1}{N_{move}} \sum_{i=1}^{N_{move}} \omega(\Delta\theta_i) \tag{10}$$

Object Grounding Capability (S_{obj}) Penalizes hallucinations during manipulation. With M_{obj} correct and K_{obj} incorrect predictions out of N_{obj} total objects, the score is calculated out of 100:

$$S_{obj} = 100 \cdot \max\left(0, \frac{M_{obj} - K_{obj}}{N_{obj}}\right) \tag{11}$$

Skill Selection Capability (S_{skill}) The coverage percentage of correctly selected atomic skills M_{skill} out of N_{skill} total skills:

$$S_{skill} = 100 \cdot \frac{M_{skill}}{N_{skill}} \tag{12}$$

B. Dataset Construction

B.1. Scene Diversity

To minimize the Sim-to-Real gap and provide diverse visual inputs, we built a simulation environment based on NVIDIA Isaac Sim. We designed six variant environments covering three core space types: household, industrial, and outdoor settings.

B.2. Assets Library

Our asset library leverages the Universal Scene Description (USD) ecosystem of Isaac Sim and deeply integrates high-quality assets from ShapeNet and Google Scanned Objects. The library contains objects of various categories, covering a wide range of interactive objects from rigid bodies to articulated mechanisms. To ensure the realism of contact dynamics, all objects undergo rigorous physical property calibration, including convex decomposition optimization of collision meshes, inertia tensor calculation for mass distribution, and fine-tuning of surface friction coefficients.

Table 4. Statistics of interactive assets in the simulation environment.

ASSET CATEGORY	QUANTITY
STOVE	1
MUG	4
CABINET	1
TOY	2
SNACK	3
INGREDIENT	1
CONDIMENT	2
BOWL	3
BOX	5
BOX CONTAINER	4
TRAY	3
HOUSEHOLD	2
INDUSTRIAL PART	7
TABLE	3

B.3. Data Collection Pipeline

To capture natural and diverse action distributions and avoid the behavioral rigidity of rule-based scripts, we employed a "Human-in-the-Loop" simulation teleoperation data collection scheme. Operators synergistically control the UAV's flight and the manipulator's fine movements via gamepad or keyboard. Successful task segments are retained and processed into formats including raw Isaac Sim data, Robomimic-compatible HDF5 format, and LeRobot standardized format adapted for the Hugging Face open-source ecosystem, satisfying the requirements of different training frameworks.

B.4. VLM Evaluation Setup

To quantitatively evaluate the task planning capabilities of VLMs, we defined a standardized task description format based on JSON. This format structurally articulates the ideal execution flow of long-horizon tasks, detailing sub-task decomposition

logic, execution agents (UAV or ARM), target objects, and required skills. This structured data serves as Ground Truth for automatically evaluating the correctness of VLM output planning logic.

```
[
  {
    "subtask": 0,
    "tasktype": "UAV",
    "direction": "frontleft"
  },
  {
    "subtask": 1,
    "tasktype": "ARM",
    "object": "rubiks_cube",
    "skill": "pick"
  }
]
```

B.5. Instruction Generation and Augmentation

To comprehensively evaluate the generalization capabilities of models in natural language understanding, we developed an instruction augmentation engine leveraging LLMs. For each specific task, the LLM is prompted to generate novel instructions that retain complete action semantics while employing diverse linguistic variations, including object rephrasing, implicit action inference, and environmental descriptions. This approach ensures the inclusion of unseen yet semantically equivalent instructions within the training set, enabling effective assessment of the model's semantic robustness against unstructured linguistic inputs.The specific prompt utilized for instruction generation is presented below:

Listing 1. Prompt for instruction generation.

```
The current action description is a bit one-dimensional. I need to make some modifications
    : 1. Replace the description of the object, 2. Replace the description of the action,
    3. Add additional semantic scene information, 4. It can be expressed in a more
    conversational way.
```

C. Experimental Implementation

To establish robust baselines and evaluate the transferability of VLA models to the aerial manipulation domain, we selected five representative architectures for full-parameter fine-tuning: $\boldsymbol{\pi_0}$, $\boldsymbol{\pi_{0.5}}$, $\boldsymbol{\pi_0}$-FAST, **ACT** (Action Chunking Transformer), and **Diffusion Policy**.

Training Configuration. All models accept RGB images from four perspectives (resized to 224×224) alongside 7-DoF manipulator joint positions and gripper states as inputs. The models predict a hybrid action vector comprising 4-DoF relative UAV displacement commands, 7-DoF manipulator joint position targets, and the gripper state. We adopted the officially recommended hyperparameters for all evaluated models. To ensure a fair comparison regarding training volume, the training duration was standardized across all baselines; specifically, the training schedule was aligned based on the number of effective epochs, where one epoch is strictly defined as a single complete pass over the entire training dataset. All experiments were executed in parallel on two NVIDIA A100 GPUs (80GB VRAM). We utilized the official open-source implementations for each model while maintaining default architectural hyperparameters. This setup was chosen to verify the potential of general-purpose VLA architectures to adapt to high-dimensional aerial manipulation tasks without domain-specific modifications.

Evaluation Protocol. During the evaluation phase, we maintained consistency with the training data regarding the formatting of images, language instructions, observation states, and action states, adopting a control frequency of 20Hz. For the inference strategy, specifically for π_0, $\pi_{0.5}$ and π_0-FAST, we executed the first 5 actions of the predicted action chunk at each inference step. For Diffusion Policy and ACT, we directly executed the predicted actions. For each specific task, we conducted 10 independent evaluation trials and reported the average metrics across these trials as the final result.

D. Full Task List

Table 5. Detailed Task List of AIR-VLA (Summary Version)

Task Name	Suite	Description
Standard Pick & Place	Base Manip.	Basic task: Pick gray cube, place in yellow basket.
Place on Flat Surface	Base Manip.	Precision placement: Pick cube, place stably on plate.
Color Distractor (Green)	Base Manip.	Color discrimination: Pick green cube amidst gray distractors, place in basket.
Container Geometry (Bowl)	Base Manip.	Container adaptation: Place green cube into small-aperture bowl.
Deformable Object (Banana)	Base Manip.	Irregular object: Pick banana and place in basket, handling non-rigid grasping.
Irregular on Plate	Base Manip.	Placement stability: Place banana stably on plate.
Tool Manipulation (Mug)	Base Manip.	Handled object: Pick red mug and place on plate.
Precise Insertion (Mug)	Base Manip.	Fine manipulation: Insert mug accurately into bowl, avoiding collision.
Cylinder Manipulation	Base Manip.	Pose constraint: Pick meat can and place upright on plate.
Tall Object (Cleanser)	Base Manip.	Center of mass control: Pick tall cleanser bottle and place in deep basket.
Variant: Banana in Box	Base Manip.	Variant: Place cube in purple storage box.
Variant: Teddy Bear	Base Manip.	Complex model: Pick teddy bear and place in purple storage box.
Variant: Meat Can in Box	Base Manip.	Variant: Place meat can in purple storage box.
Variant: Chocolate Box	Base Manip.	Box-shaped object: Pick chocolate box and place in purple storage box.
Variant: Wood Block	Base Manip.	Textured object: Pick wood block and place in purple storage box.
Variant: Tomato Can	Base Manip.	Variant: Place chocolate box in red bowl.
Variant: Teddy in Bowl	Base Manip.	Size adaptation: Place large teddy bear in bowl.
Variant: Banana in Bowl	Base Manip.	Variant: Adapt banana curvature to bowl shape.
Stacking: Can on Wood	Base Manip.	Simple stacking: Place meat can on top of wood block.
Stacking: Box on Wood	Base Manip.	Simple stacking: Place chocolate box on top of wood block.
Target Color (Yellow)	Object & Spatial	Multi-container grounding: Identify yellow basket among multiple and place cube.
Target Color (Gray)	Object & Spatial	Low-contrast grounding: Identify gray basket similar to ground color and place.
Target Category (Banana)	Object & Spatial	Category & location: Identify banana and place in specified gray basket.
Pick Left Cup (Red)	Object & Spatial	Spatial & visual matching: Pick the left red cup and place on red plate.
Visual Combination (Blue-Red)	Object & Spatial	Color combination: Place blue cup on red plate, testing instruction following.
Cross-Category Transfer	Object & Spatial	Cross-category move: Place meat can into gray basket.
Slim Object in Bowl	Object & Spatial	Slim object grounding: Place cleanser into yellow bowl.
Spatial Ref. (Middle Drawer)	Object & Spatial	Vertical spatial grounding: Identify middle handle, pull drawer > 0.3m.
Small Object (Bolt)	Object & Spatial	Tiny object grounding: Identify tiny bolt and place in bowl.
Spatial Ref. (Top Drawer)	Object & Spatial	Dynamic height adaptation: Ascend to identify top handle, pull drawer > 0.3m.
Pick Left Box (White)	Object & Spatial	Spatial constraint: Pick the white box on the left and place in the left basket.
Variant: Green Nut	Object & Spatial	Small part: Pick green nut and place in yellow basket.
Assembly (Camera on Lidar)	Object & Spatial	Fine assembly: Place camera module on top of LiDAR.
Non-Prehensile (Push Cube)	Object & Spatial	Non-prehensile: Push cube next to gray basket.
Clear Desk (Push Box)	Object & Spatial	Clearing action: Push brown box off the table.
Synonym: Container	Semantic Und.	Hypernym: "Container" $\rightarrow$ Basket. Action: Cube $\rightarrow$ Yellow Basket
Synonym: Block & Bin	Semantic Und.	Double synonym: "Block" $\rightarrow$ Cube, "Bin" $\rightarrow$ Basket.
Verb: Drop	Semantic Und.	Verb nuance: "Drop" implies lower precision release. Action: Banana $\rightarrow$ Gray Basket
Synonym: Cup & Dish	Semantic Und.	Noun substitution: "Cup" $\rightarrow$ Mug, "Dish" $\rightarrow$ Plate.
Preposition: Onto	Semantic Und.	Preposition: "Onto" emphasizes surface support. Action: Blue Mug $\rightarrow$ Red Plate

Continued on next page...

Table 5 – Continued

Task Name	Suite	Description
Material: Tin	Semantic Und.	Material reference: "Tin" → Metal Can. Action: Meat Can → Gray Basket
Function: Detergent	Semantic Und.	Function reference: "Detergent" → Cleanser. Action: Cleanser → Yellow Bowl
Synonym: Center	Semantic Und.	Position semantic: "Center" → Middle Drawer. Pull action.
Attribute: Metal	Semantic Und.	Attribute reference: "Metal" → Bolt. Place in bowl.
Superlative: Highest	Semantic Und.	Extreme semantic: "Highest" → Top Drawer. Pull action.
Negation Constraint	Semantic Und.	Negation: "Not the brown one...". Exclude brown, pick white box.
Contextual: Spare Part	Semantic Und.	Contextual: "Spare part" refers to green nut, place in parts box.
Jargon: Vision Sensor	Semantic Und.	Jargon: "Vision sensor" → Camera. Place on LiDAR.
Preposition: By / Don't	Semantic Und.	Position & Negation: "By" + "Don't put in". Push action.
Abstract Intent: Clear	Semantic Und.	Abstract intent: "Clear the desk" → Push object off table.
Multi-Object Sorting	Long-Horizon	Multi-step sorting: 1. Rubiks → Yellow; 2. Cube → Gray.
Batch Collection	Long-Horizon	Batch collection: 1. Rubiks → Yellow; 2. Cube → Yellow.
Color-Matched Sorting	Long-Horizon	Color matching: 1. Black Mug → Black Plate; 2. Yellow Mug → Red Plate.
Cross-Color Sorting	Long-Horizon	Spatial crossing: 1. Black Mug → Left(Red); 2. Yellow Mug → Right(Black).
Category Sorting	Long-Horizon	Category sorting: 1. Food(Can) → Right; 2. Fruit(Banana) → Right.
Stacking & Nesting	Long-Horizon	Nesting dependency: 1. Move Bowl → Plate; 2. Place Cube → Inside Bowl.
Tower Stacking	Long-Horizon	Stacking balance: 1. Place Rubiks → Basket; 2. Place Cube → On Rubiks.
Hybrid Interaction (Stove)	Long-Horizon	Hybrid skill: 1. Rotate Switch $> 75^{\circ}$; 2. Place Bowl → Stove.
Drawer Interaction	Long-Horizon	Mechanism interaction: 1. Pull Top Drawer; 2. Place Banana → Inside.
Store & Close	Long-Horizon	Full storage: 1. Place Rubiks → Drawer; 2. Close Drawer.