Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

Flying Hand: End-Effector-Centric Framework for Versatile Aerial Manipulation Teleoperation and Policy Learning

Guanqi He*†, Xiaofeng Guo*†, Luyi Tang†, Yuanhang Zhang†, Mohammadreza Mousaei†, Jiahe Xu†,
Junyi Geng‡, Sebastian Scherer† and Guanya Shi†
*Equal contribution, Alphabetical order
†Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA
Email: {guanqihe, xguo2, luyit, yuanhanz, mmousaei, jiahex, basti, guanyas}@andrew.cmu.edu
‡Department of Aerospace Engineering, Pennsylvania State University, University Park, PA 16802 USA
Email: jgeng@psu.edu

Abstract—**Aerial manipulation has recently attracted increasing interest from both industry and academia. Previous approaches have demonstrated success in various specific tasks. However, their hardware design and control frameworks are often tightly coupled with task specifications, limiting the development of cross-task and cross-platform algorithms. Inspired by the success of robot learning in tabletop manipulation, we propose a unified aerial manipulation framework with an end-effector-centric interface that decouples high-level platform-agnostic decision-making from task-agnostic low-level control. Our framework consists of a fully-actuated hexarotor with a 4-DoF robotic arm, an end-effector-centric whole-body model predictive controller, and a high-level policy. The high-precision end-effector controller enables efficient and intuitive aerial teleoperation for versatile tasks and facilitates the development of imitation learning policies. Real-world experiments show that the proposed framework significantly improves end-effector tracking accuracy and can handle multiple aerial teleoperation and imitation learning tasks, including writing, peg-in-hole, pick and place, changing light bulbs, etc. We believe the proposed framework provides one way to standardize and unify aerial manipulation into the general manipulation community and to advance the field. Project website: https://lecar-lab.github.io/flying_hand/.**

I. Introduction

Uncrewed Aerial Manipulators (UAMs), which target complex tasks at high altitudes [45], hold significant potential to reduce human labor in many elevated operations, such as changing light bulbs on tall towers, inspecting aircraft wings or turbine blades, and painting bridges, which are not only costly but also pose substantial risks to human safety. Previous works have demonstrated the ability to achieve different specific complex aerial manipulation tasks, including drawing calligraphy [20], grasping [36], perching [30], drilling [13], etc. However, most previous works have been tailored to specific tasks, developing unique platforms and algorithms accordingly, lacking the ability to handle different types of tasks. In real-world scenarios, manipulation tasks can be complex and typically consist of multiple sub-tasks. For example, changing a light bulb can involve several motion primitives, including interaction, grasping, insertion, and rotation. This raises a requirement for a more general-purpose, versatile aerial manipulation system, which essentially requires a **versatile aerial manipulation framework** to handle multiple tasks.

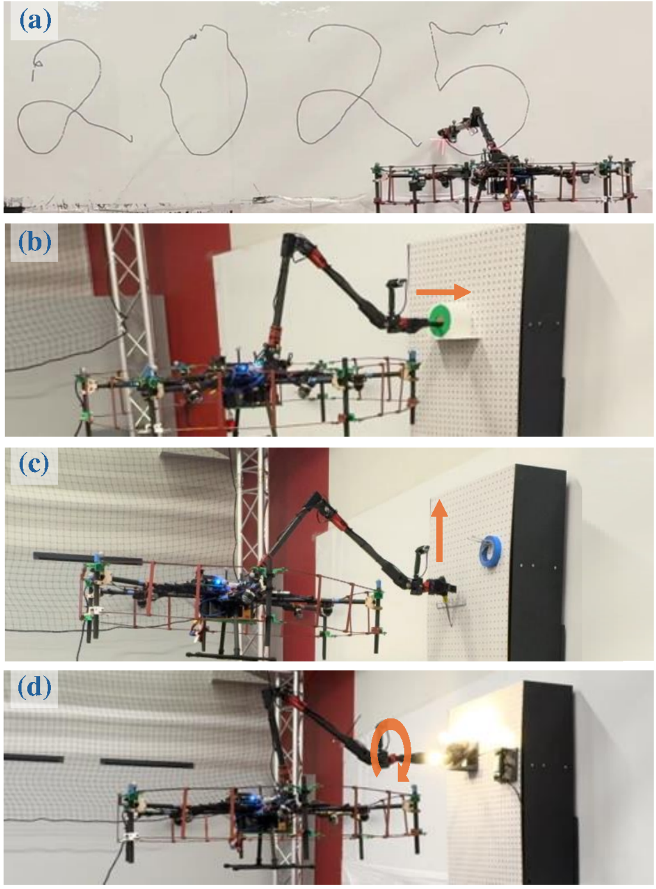

Fig. 1. The proposed framework and system can accomplish multiple typical aerial manipulation tasks precisely and robustly, such as (a) writing ”2025”, (b) peg-in-hole, (c) pick-and-place, and (d) changing light bulbs.

In robotic manipulation [35], the end-effector-centric (ee-centric) approach is widely used. It defines tasks and policies [31, 32] in Cartesian space instead of the specific robotic arm configuration space. By effectively decoupling high-level policies from low-level control, it enables the development of embodiment-agnostic policies [47], [10], [43] and policy-agnostic low-level controllers [34], enhancing framework versatility, cross-embodiment adaptability, and algorithm reuse capability [54]. Although the end-effector-centric paradigm has shown the advantage of versatility in the manipulation field, applying it to aerial manipulation systems presents significant challenges due to the UAM's floating-base dynamics and the coupling effects between the UAV and the manipulator.

In this work, inspired by the success of ee-centric interfaces in classic manipulation, we reformulate aerial manipulation from the **manipulation** perspective by proposing a versatile aerial manipulation framework with the ee-centric interface to address various aerial manipulation tasks. The framework consists of a versatile aerial manipulation platform capable of executing multiple tasks, a policy-agnostic controller that precisely tracks the target end-effector state, and an ee-centric policy module responsible for generating target end-effector states. Specifically, we develop a fully-actuated hexarotor with a 4 DoF robotic arm, providing a sufficiently large workspace and wrench space for diverse tasks. We then develop an ee-centric whole-body Model Predictive Controller (**ee-centric MPC**) that precisely tracks the target end-effector state, even in the presence of model uncertainties. Moreover, to bring human cognitive skills into policy development and benefit from the ee-centric interface, we develop an ee-centric teleoperation interface and an imitation-learning-based framework to acquire autonomous policy learned from human demonstration. To the best of our knowledge, this is the first imitation learning-based framework for aerial manipulation. Real-world experiments show that the proposed framework achieves high-precision end-effector tracking and enables a wide range of aerial manipulation tasks, including aerial writing, peg-in-hole, pick and place, light bulb replacement, etc., as shown in Fig. 1. It also demonstrates how this modular and standardized ee-centric framework effectively decouples the high-level policy from the low-level controller, which enables seamless integration of existing standard high-level policy modules from the broader manipulation community, such as teleoperation and imitation learning, into the field of aerial manipulation. We believe the proposed framework provides a step toward standardizing and unifying aerial manipulation into the broader manipulation community, advancing the field toward greater versatility and generalization.

In summary, our contributions are:

1) We reformulated the aerial manipulation problem within the unified manipulation paradigm, consisting of a UAM system, a controller encapsulated by the ee-centric interface, and a high-level policy.

2) We developed an end-effector-centric whole-body MPC for aerial manipulation that precisely tracks the target end-effector state while maintaining robustness against disturbances through L1 adaptation.

3) We developed an ee-centric teleoperation system and an imitation-learning-based autonomous system that learns from human teleoperation demonstration.

4) Rich real-world experiments demonstrated the versatility of the proposed framework, the effectiveness of the user-friendly teleoperation interface, and the potential to incorporate learning-based policies and other manipulation policies.

II. Related Works

A. Aerial Manipulation

There have been many research efforts exploring aerial manipulation for various kinds of tasks [39]. Based on the motion primitives they require, common aerial manipulation tasks can be categorized into: 1) Aerial Interaction, which requires maintaining contact with external objects, for tasks such as inspection [3] [4] [19], aerial writing [33] [49] [22] or pushing a target [6]. Researchers mostly developed a point-contact arm, such as a rigid rod, and proposed the hybrid motion-force control framework, although achieving high-precision tracking performance, struggled to handle scenarios requiring grasping; 2) Aerial Grasping [37], where previous works mainly focused on designing different custom end-effectors, such as claw [41] or soft gripper [16]. Some work also showed amazing results, achieving high-speed grasping, or grasping moving objects [50], but sacrificed payload capacity or precision due to specialized hardware designs. 3) Aerial Insertion. Typical work includes [42] where they proposed a specific hole searching policy for bolt screwing tasks, and [52] where they achieved mm-level peg-in-hole task; 4) Manipulate articulated objects such as doors [44], or valves [7]. In general, although different works have shown success on different specific tasks, the specific system design and algorithm development make the same hardware and algorithm hard to deploy to different tasks, reducing its potential for practical long-horizon versatile aerial manipulation tasks. In our work, we target these four types of aerial manipulation tasks, developing a versatile framework to handle all of them.

B. Mobile Manipulation Framework and EE-Centric Interface

Combining a whole-body tracking controller with a high-level policy has shown promise in mobile manipulation systems such as humanoids. [25] proposed a framework that consists of a robust humanoid whole-body controller with a high-level policy, either an autonomous agent like GPT-4o or an imitation learning policy learned from teleoperation. [17] developed a system that consists of a transformer-based low-level control and imitation learning for humanoids. [26] discussed the interface between high-level policy and low-level controller, and proposed a unified controller for humanoid whole-body control supporting versatile interfaces. Beyond the explicit whole-body motion, some work extends the interface to latent variables and extends the whole-body control to a vision-motor manipulation policy. [57] proposed a hierarchical framework that consists of the understanding module, a pre-trained large visual-language model running in low-frequency, and the execution module, a visual-based action policy running in high-frequency. Gemini robotics [46], Helix [15], and Isaac GR00T N1 [2] also adopted similar structures.

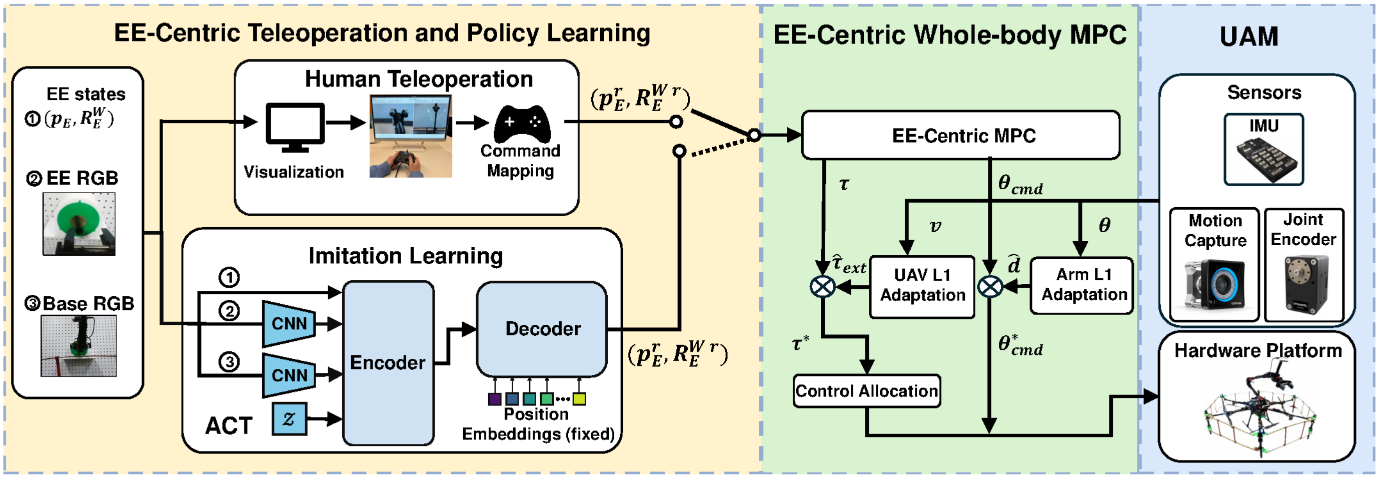

Fig. 2. The proposed end-effector-centric aerial manipulation framework includes the UAM platform, the ee-centric whole-body MPC, and the high-level policy including an ee-centric teleoperation interface, and an imitation learning-based framework using Action Chunk with Transformer (ACT) [59]. The high-level policy, either the human teleoperation or learned autonomous policy, sends the target end-effector state to ee-centric MPC, which then generates motor commands for the UAM platform to execute.

Defining tasks or commands using ee-centric approaches is widely adopted in general manipulation fields, as it is more intuitive and can be cross-embodiment. For example, Båberg et al. [8] developed a teleoperation interface to enable full control of the end-effector pose. The Universal Manipulation Interface [10] [21] demonstrates a data-collection and policy-learning framework that allows direct skill transfer from in-the-wild human demonstrations to multiple robot embodiments. Their system employs a hand-held gripper and carefully designed hardware-agnostic policies, showcasing the potential for ee-centric solutions in multi-platform scenarios. Similarly, other mobile manipulation strategies, such as N2M2 [28] [27] and HarmonicMM [55], reduced the operator's burden by extracting feasible base motions from end-effector trajectories. However, these works generally remain limited to ground robots. Although several aerial manipulation studies have adopted end-effector-centric methods [20] [5], they primarily focused on developing controllers to track specified end-effector trajectories without a systematic framework that tackles various tasks comprehensively.

In our work, we propose to adopt the two-layer hierarchical system for aerial manipulation and propose a unified framework with the ee-centric interface for versatile aerial manipulation tasks.

C. Teleportation and Imitation Learning

Developing a robust and practical autonomous aerial manipulation policy is extremely challenging due to complex real-world environments and high precision and safety requirements. Moreover, policies are typically designed to handle specific tasks and lack the generality to handle unexpected conditions. Therefore, teleoperation, which takes human effort into the loop for policy development, attracts researchers' interest as a practical solution. For example, [1] developed the UAM with a fully actuated UAV with a 0 DoF arm and controlled the end-effector directly by teleoperation, but their method is highly coupled with the specific UAM design, and the system struggles with versatile tasks due to the workspace limitation. In most structured UAM teleoperation works, users control each DoF separately [56] or explicitly switch to different modes during different phases [11]. Both of these increase the human teleoperator's burden and require the teleoperator to have a rich understanding of the specific hardware system. Moreover, even if incorporated with human input, the previous works still have not shown vast versatility in different tasks and scenarios.

Recently, imitation learning (IL) has demonstrated significant potential for autonomous policy learning due to its high data efficiency, straightforward framework, and outstanding performance. Recent progress in both systems, such as ALOHA [59] and mobile ALOHA [18], and algorithms, such as ACT [59] and diffusion policy [9], have significantly facilitated the success of long-horizon, contact-rich, complex manipulation tasks. However, there is no precedent to incorporate such IL-based policy into aerial manipulation fields due to the lack of a mature demonstration collection system, such as a well-developed teleoperation system, and the lack of a proven framework. In this work, we develop an intuitive teleoperation system using the ee-centric framework. It also helps to collect human demonstration data, enabling us to develop an imitation learning-based policy for autonomous aerial manipulation.

III. System Overview

Our aerial manipulation system is designed to enable precise and versatile operations. The system incorporates an end-effector-centric (ee-centric) interface to decouple high-level decision-making from low-level control, increasing the framework's versatility. As shown in Fig. 2, our system consists of an aerial manipulator platform, an ee-centric whole-body MPC, and an ee-centric policy module. The hardware platform consists of a fully-actuated hexarotor and a 4 DoF robotic arm. The platform has a large enough workspace and wrench space for different tasks. A motion capture system and onboard IMUs are used for drone state estimation. The joint encoders are used to get arm joint angles, and the end-effector states are then calculated based on forward kinematics. The ee-centric whole-body MPC reads the end-effector state target from high-level policy and generates the reference trajectory and reference control for both the UAV and the robotic arm. An L1 online adaptation control term is designed to further improve the tracking performance. The UAV control commands are then sent to the control allocation to generate the motor commands for the UAV to execute. At the most high-level, the ee-centric policy module gets current observations and generates the target end-effector states online without the need to consider the specific platform jointly. We developed two high-level policy modules. The first is the ee-centric teleoperation interface that allows human users to directly control the end-effector pose. The second is an imitation learning framework, where we adopt Action Chunking with Transformers (ACT) [60], to learn autonomous policies from human teleoperation demonstrations. The following sections will introduce the developed modules accordingly.

IV. Hardware Design

A. Fully-Actuated UAV

The foundation of the system is a fully-actuated hexarotor UAV capable of independently generating six-dimensional forces and torques, based on our previous works [19] [20] [22]. This capability allows precise control of position and orientation, which is essential for executing complex aerial manipulation tasks. The robust design ensures stability in dynamic environments while ensuring high-precision end-effector tracking. We use Tarot680 as the drone base, 6 KDE 4215XF motors with 12-inch 2-blade propellers as our driving force, LiPo batteries for on-board power supply, an Intel NUC for on-board computation, and a customized PX4 autopilot for low-level flight control and information processing.

B. Manipulator

The UAV integrates a 4-DOF robotic manipulator optimized for versatile and precise task execution. The arm features three pitch joints and one roll joint, driven by Dynamixel XM540 and XM430 servos. Its configuration allows high-precision operations. The system achieves whole-body manipulation capabilities by combining the UAV's actuation with the manipulator, enhancing task execution across diverse scenarios. The manipulator includes a modular end-effector, allowing interchangeable tools for specific tasks. For instance, a two-finger gripper with replaceable tips enables precision handling, while a circular gripper is ideal for changing light bulbs.

C. Perception

In this work, we mainly focus on the indoor environment, and we include a brief discussion about outdoor environment applications in Appendix B. We use both the motion capture system and the PX4 onboard IMU for drone state estimation. The motion capture system provides the UAV's position and can be replaced with other localization methods, such as SLAM[58]. The manipulator arm joint angles are estimated by the joint encoders, and the end-effector states are online calculated based on forward kinematics.

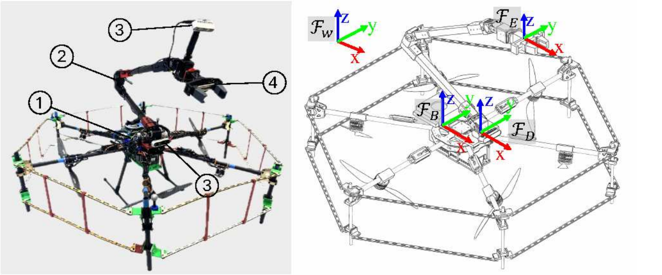

Fig. 3. UAM hardware system design, illustrating the key components: (1) fully-actuated hexarotor as the base structure, (2) 4 Dof manipulator, (3) Intel RealSense cameras for vision-based perception and feedback, and (4) end-effector gripper for object interaction. The frame notations in the right diagram represent the coordinate axes associated with the system.

To further improve drone perception for teleoperation and autonomous policy development, we equip the aerial manipulator with two RealSense RGBD cameras. One camera is mounted on the UAV base to capture a broad view of the entire workspace, while the other is positioned near the end-effector to deliver detailed close-up views of the target area. This dual-camera setup ensures teleoperators and vision-based policies can maintain precise control and situational awareness during complex manipulation tasks.

V. System Modeling

A. Frames and Notation

The frames depicted in the Fig 3 are defined as follows: $\mathcal{F}_W$ is an inertial world frame with its z-axis opposite to the gravity vector, ensuring $\hat{z}_W$ points upward. $\mathcal{F}_B$ is rigidly attached to the UAV's body at its center of gravity, with axes (x_B, y_B, z_B) aligned with the UAV body frame. $\mathcal{F}_D$ is the manipulator base frame, centered at the attachment point of the manipulator on the UAV. The transformation from $\mathcal{F}_B$ to $\mathcal{F}_D$ is defined by a constant translation $\mathbf{p}_D \in \mathbb{R}^3$ and a fixed orientation $\boldsymbol{R}_B^D \in SO(3)$. $\mathcal{F}_E$ is the end-effector frame with axes (x_E, y_E, z_E), where x_E is aligned with the roll axis of the 4th joint of the manipulator, while y_E remains horizontal. Other symbols in the paper are listed in Table I for the convenience of the following discussion.

B. Fully-Actuated UAV Dynamics

A fully-actuated UAV is adopted as the base of the aerial manipulator, which can generate six-dimensional force and torque independently. Let the generalized position of the UAV be represented as $\boldsymbol{q} = [\boldsymbol{p}, \boldsymbol{R}_B^W]$, where $\boldsymbol{p} \in \mathbb{R}^3$ represents the position of the UAV in the global coordinate frame, and $\boldsymbol{R}_B^W \in SO(3)$ represents its orientation. The generalized velocity is denoted as $\boldsymbol{v} = [\dot{\boldsymbol{p}}, \boldsymbol{\omega}]$, where $\boldsymbol{\omega} \in \mathbb{R}^3$ is the angular velocity.

TABLE I
NOTATION OVERVIEW

Symbol	Dimension	Description
m	$\mathbb{R}$	Mass
$\boldsymbol{J}$	$\mathbb{R}^{3\times 3}$	Inertia tensor
$\boldsymbol{p}$, $\boldsymbol{R}_B^W$	$\mathbb{R}^3$, $SO(3)$	UAV position expressed in $\mathcal{F}_W$ and orientation between $\mathcal{F}_B$ and $\mathcal{F}_W$
$\boldsymbol{v}$	$\mathbb{R}^3$	Generalized velocity, expressed in $\mathcal{F}_W$
$\boldsymbol{\tau}$	$\mathbb{R}^6$	Control wrench
$\boldsymbol{\tau}_{\text{ext}}$	$\mathbb{R}^6$	External disturbance wrench
$\boldsymbol{g}_W$	$\mathbb{R}^6$	Gravity vector in $\mathcal{F}_W$
$\boldsymbol{p}_E$, $\boldsymbol{R}_E^W$	$\mathbb{R}^3$, $SO(3)$	End-effector position expressed in $\mathcal{F}_W$ and orientation between $\mathcal{F}_E$ and $\mathcal{F}_W$
$\boldsymbol{\theta}$	$\mathbb{R}^4$	Current joint angles
$\boldsymbol{\theta}_{\text{cmd}}$	$\mathbb{R}^4$	Commanded joint angles
$\boldsymbol{d}$	$\mathbb{R}^4$	Joint servo disturbance
$\boldsymbol{\zeta}$	$\mathbb{R}^{3\times 4}$	DH parameter with the i^{th} joint component $[\theta_i, l_i, a_i, \alpha_i]^T$
$\boldsymbol{\beta}$	$\mathbb{R}$	Joint motor delay constant

The UAV dynamics can be formulated using Newton-Euler equations for rigid body motion as follows:

$$\boldsymbol{M}\dot{\boldsymbol{v}} + \boldsymbol{C}\boldsymbol{v} + \boldsymbol{g} = \boldsymbol{\tau} + \boldsymbol{\tau}_{\text{cxt}} \tag{1}$$

with inertia matrix $\boldsymbol{M} \in \mathbb{R}^{6\times 6}$, centrifugal and Coriolis term $\boldsymbol{C} \in \mathbb{R}^{6\times 6}$, gravity wrench $\boldsymbol{g} \in \mathbb{R}^6$, control wrench $\boldsymbol{\tau} \in \mathbb{R}^6$ from the UAM actuators, and unknown external wrench $\boldsymbol{\tau}_{ext} \in \mathbb{R}^6$ from model mismatch and manipulator interaction.

Specifically, we have

$$\boldsymbol{M} = \text{diag}\left(\begin{bmatrix} m\boldsymbol{I}_{3\times 3} & \boldsymbol{J} \end{bmatrix}\right), \tag{2}$$

$$\boldsymbol{C} = \text{diag}\left(\begin{bmatrix} m[\boldsymbol{\omega}]_\times & -\boldsymbol{J}[\boldsymbol{\omega}]_\times \end{bmatrix}\right), \tag{3}$$

$$\boldsymbol{g} = m \ \text{diag}\left(\begin{bmatrix} \boldsymbol{R}_B^W & 0_{3\times 3} \end{bmatrix}\right) \boldsymbol{g}_W. \tag{4}$$

where m is the vehicle mass, $\boldsymbol{J}$ is the moment of inertia at the vehicle center of mass in the body frame, $\boldsymbol{g}_W = [0, 0, g, 0, 0, 0]^\top$ is the gravitational acceleration in $\mathcal{F}_W$, and $[*]_\times$ is the skew-symmetric matrix associated with vector $*$.

C. Manipulator Kinematics

In this work, the manipulator employs servos as joint actuators, which cannot accurately control joint torques directly. Therefore, only the kinematics of the manipulator is considered in the system modeling. The interaction between the manipulator and the fully-actuated UAV is treated as a disturbance and is compensated in real-time using L1 adaptive control.

We use the standard Denavit-Hartenberg (DH) convention [12] to model the forward kinematics of our 4-DoF robotic arm. Under the DH formulation, the adjacent frame transformation $\boldsymbol{T}_i^{i-1}$ is characterized by four parameters θ_i, d_i, a_i, α_i, where the first one is the joint angle and the last three are pre-identified and fixed during robot movement. Define DH parameter $\zeta_i = [\theta_i, l_i, a_i, \alpha_i]^\top \in \mathbb{R}^4$. The frame transformation from end-effector frame to arm base body frame can be written as

TABLE II
SYSTEM IDENTIFICATION RESULTS

Mass Matrix $\boldsymbol{M}$	$\text{diag}(0.105, 0.121, 0.101, 0.025, 0.011, 0.013)$
Motor Delay β	$(0.66,\ 0.68,\ 0.81,\ 0.85)$
Joint 1 DH Param $\boldsymbol{\zeta}_1$	$d_1 = 0.0$, $a_1 = 0.363$, $\alpha_1 = 0.10$
Joint 2 DH Param $\boldsymbol{\zeta}_2$	$d_2 = 0.050$, $a_2 = 0.441$, $\alpha_2 = -0.10$
Joint 3 DH Param $\boldsymbol{\zeta}_3$	$d_3 = 0.0$, $a_3 = 0.007$, $\alpha_3 = -1.578$
Joint 4 DH Param $\boldsymbol{\zeta}_4$	$d_4 = 0.076$, $a_4 = 0.200$, $\alpha_4 = 0.0$

$$\boldsymbol{T}_E^D(\boldsymbol{\theta}; \boldsymbol{\zeta}) = \prod_{i=1}^{4} \boldsymbol{T}_i^{i-1}(\theta_i; \zeta_i) \tag{5}$$

Then the transformation from world frame to end-effector frame can be computed as follows:

$$\boldsymbol{T}_E^W = \begin{bmatrix} \boldsymbol{R}_E^W & \boldsymbol{p}_E^W \\ 0_{1\times 3} & 1 \end{bmatrix} = \boldsymbol{T}_B^W \cdot \boldsymbol{T}_D^B \cdot \boldsymbol{T}_E^D \tag{6}$$

where $\boldsymbol{T}_B^W$ can be obtained from UAV odometry and $\boldsymbol{T}_D^B$ is a fixed transformation between the manipulator base and the UAV.

The accurate DH parameters are obtained through system identification. We collect motion data of the manipulator using a motion capture system and compute the DH parameters via least squares regression. The detailed parameter values are presented in Table II.

D. Manipulator Motor Delay

The servo motor dynamics are approximated as first-order systems to account for command-to-state delay. For the 4-DoF manipulator, the relationship between commanded joint angles $\boldsymbol{\theta}_{\text{cmd}} \in \mathbb{R}^4$ and actual joint angles $\boldsymbol{\theta} \in \mathbb{R}^4$ is governed by:

$$\text{diag}(\boldsymbol{\beta})\dot{\boldsymbol{\theta}} + \boldsymbol{\theta} = \boldsymbol{\theta}_{\text{cmd}} + \boldsymbol{d} \tag{7}$$

where $\boldsymbol{\beta} \in \mathbb{R}$ are the joint-specific time delay constants, and $\boldsymbol{d} \in \mathbb{R}^4$ is the unknown disturbance in servo control. This formulation captures the transient response characteristics of each actuator. The motor delay coefficients $\boldsymbol{\beta}$ are identified alongside the DH parameters using least squares regression, with the results presented in Table II.

VI. END-EFFECTOR-CENTRIC WHOLE-BODY CONTROL WITH ONLINE ADAPTATION

Given the over-actuated nature of our system and the users' primary focus on the end-effector motion, we employ model predictive control to regulate the end-effector trajectory. This approach enables whole-body coordination between the manipulator and the fully-actuated UAV, ensuring precise and efficient end-effector motion control. As discussed in previous sections, the complex interaction between the manipulator and the UAV is treated as the disturbance in the modeling stage, which introduces uncertainty in the nominal model used in the whole-body MPC. To mitigate the disturbances and model uncertainties, we integrate the L1 adaptive controller, ensuring robust disturbance compensation and accurate tracking performance. The diagram of the control algorithm is illustrated in Fig. 2.

A. End-Effector-Centric Model Predictive Controller

In the following, we describe the whole-body MPC framework used to optimize the end-effector reference trajectory.

For the MPC formulation, we define the following state and control variables:

$$\boldsymbol{x} := \begin{bmatrix}\boldsymbol{p}_E & \boldsymbol{R}_E^W & \boldsymbol{v} & \boldsymbol{\theta}\end{bmatrix} \quad \boldsymbol{u} := \begin{bmatrix}\boldsymbol{\tau} & \boldsymbol{\theta}_{\text{cmd}}\end{bmatrix} \tag{8}$$

We use the following error functions for the position of the end-effector, the UAM velocity, the wrench control input, and the manipulator joint angle, respectively:

$$\boldsymbol{e}_p = \boldsymbol{p}_E - \boldsymbol{p}_E^r \tag{9a}$$

$$\boldsymbol{e}_R = \frac{1}{2}\left(\boldsymbol{R}_E^{W^r\top}\boldsymbol{R}_E^W - \boldsymbol{R}_E^{W^\top}\boldsymbol{R}_E^{W^r}\right)^\vee \tag{9b}$$

$$\boldsymbol{e}_v = \boldsymbol{v} - \boldsymbol{v}^r \tag{9c}$$

$$\boldsymbol{e}_\theta = \boldsymbol{\theta} - \boldsymbol{\theta}^r \tag{9d}$$

$$\boldsymbol{e}_u = \boldsymbol{u} - \boldsymbol{u}^r \tag{9e}$$

where $(*)^\vee$ is the vee-operator that extracts a vector from a skew-symmetric matrix $*$ and $(\cdot)^r$ represents the reference state values. The reference signals $\boldsymbol{p}_E^r$ and $\boldsymbol{R}_E^{W^r}$ are provided by a high-level teleoperation command or derived from an imitation learning policy. Manipulator default joint angle $\boldsymbol{\theta}^r$ is pre-selected, and we define the reference control $\boldsymbol{u}^r = [0_6, \hat{\boldsymbol{\theta}}]$, where $\hat{\boldsymbol{\theta}}$ is the current joint states.

The MPC formulation minimizes a cost function over a finite time horizon H while subject to system dynamics and constraints:

$$\boldsymbol{u}_{\text{opt}} = \arg\min_{\boldsymbol{u}} \left\{ L_e(\boldsymbol{x}_H, \boldsymbol{x}_H^r) + \sum_{i=1}^{H} L_r(\boldsymbol{x}_n, \boldsymbol{x}_n^r, \boldsymbol{u}_n) \right\} \tag{10a}$$

$$\text{s.t.} \quad \boldsymbol{x}_{n+1} = \boldsymbol{f}_{dyn}(\boldsymbol{x}_n, \boldsymbol{\tau}_n) \tag{10b}$$

$$\boldsymbol{x}_0 = \hat{\boldsymbol{x}}, \quad \boldsymbol{x}_n \in \mathcal{X} \tag{10c}$$

$$\boldsymbol{u}_{\text{lb}} \leq \boldsymbol{u} \leq \boldsymbol{u}_{\text{ub}} \tag{10d}$$

Eq. (10a) defines the optimization objective, where H represents the discrete prediction horizon. The stage and terminal costs, L_r and L_e, are quadratic functions of the tracking errors, given by $\boldsymbol{e}_i^\top \boldsymbol{Q}_i \boldsymbol{e}_i$, where $\boldsymbol{e}_i \in \{\boldsymbol{e}_p, \boldsymbol{e}_R, \boldsymbol{e}_v, \boldsymbol{e}_\theta, \boldsymbol{e}_u\}$. The gain matrices $\boldsymbol{Q}_i$ are positive definite and tuned experimentally to balance precision and robustness.

Eq. (10b) enforces the discrete-time system dynamics, incorporating the fully actuated UAV dynamics Eq. (1), manipulator kinematics Eq. (6), and joint servo dynamics Eq. (7). The continuous system dynamics are discretized using a fourth-order Runge-Kutta (RK4) integration scheme to maintain numerical stability and accuracy. Disturbances $\boldsymbol{\tau}_{\text{ext}}$ and $\boldsymbol{d}$ are ignored in the MPC formulation and solving process, but will be handled in the following Section VI-B via online L1 adaptation.

Eq. (10c) introduces state constraints, where $\hat{\boldsymbol{x}}$ represents the latest state estimate. The feasible state space $\mathcal{X}$ is defined by: 1) Self-collision avoidance constraints: ensuring that the manipulator does not collide with the UAV structure. 2) Environment collision constraints: preventing the UAV contact with external obstacles. 3) Safety constraints: including velocity limits and joint angle restrictions to ensure safe operation.

Eq. (10d) imposes actuation limits on the aerial manipulator, where $\boldsymbol{u}_{\text{lb}}$ and $\boldsymbol{u}_{\text{ub}}$ define the lower and upper bounds of the control inputs.

The end-effector centric whole-body MPC formulation is a general framework that can adapt to various vehicle types and can extend to systems with multiple end-effectors. The control inputs, UAV dynamics, and arm kinematics can be tailored to specific vehicle and manipulator configurations, ensuring flexibility across different aerial manipulation systems.

B. L1 Online Adaptation

As discussed in previous sections, the complex interaction between the manipulator and the UAV is treated as the disturbance, which introduces uncertainty in the nominal model used in the whole-body MPC. Such model uncertainties are typically bounded, and prior knowledge about their characteristics is usually available [40, 23]. To mitigate the disturbances and model uncertainties, we integrate the L1 adaptive controller, ensuring robust disturbance compensation and accurate tracking performance.

We adopt the L1 adaptive controller from [53, 29] in both the fully-actuated UAV motion control and the manipulator joint angle tracking control, to compensate the disturbance $\boldsymbol{\tau}_{\text{ext}}$ in Eq. (1) and $\boldsymbol{d}$ in Eq. (7).

The adaptation law is designed by

$$\boldsymbol{M}\dot{\hat{\boldsymbol{v}}} + \boldsymbol{C}\hat{\boldsymbol{v}} + \boldsymbol{g} = \boldsymbol{\tau} + \hat{\boldsymbol{\tau}}_{\text{ext}} + \boldsymbol{A}_v(\hat{\boldsymbol{v}} - \boldsymbol{v}) \tag{11a}$$

$$\hat{\boldsymbol{\tau}}'_{\text{ext}} = -(e^{\boldsymbol{A}_v dt} - \boldsymbol{I}_{6\times 6})^{-1}\boldsymbol{A}_v e^{\boldsymbol{A}_v dt}(\hat{\boldsymbol{v}} - \boldsymbol{v}) \tag{11b}$$

$$\hat{\boldsymbol{\tau}}_{\text{ext}} \leftarrow \text{low pass filter}(\hat{\boldsymbol{\tau}}_{\text{ext}}, \hat{\boldsymbol{\tau}}'_{\text{ext}}) \tag{11c}$$

where $\hat{\boldsymbol{v}} \in \mathbf{R}^6$ denotes the estimated UAV velocity, $\boldsymbol{A}_v$ is a Hurwitz matrix, dt is the discretization step length, and $\hat{\boldsymbol{\tau}}_{\text{ext}} \in \mathbb{R}^6$ encapsulates the unknown wrench disturbances. Here Eq. (11a) is a velocity estimator and Eq. (11b) and Eq. (11c) update and filter the disturbance $\hat{\boldsymbol{\tau}}_{\text{ext}}$.

Thus, the total UAV wrench control command $\boldsymbol{\tau}^*$ is computed as

$$\boldsymbol{\tau}^* = \boldsymbol{\tau}_{\text{mpc}} + \hat{\boldsymbol{\tau}}_{\text{ext}} \tag{12}$$

Similarly, the L1 adaptive controller for the manipulator joint angles is formulated to compensate for dynamic disturbances and model uncertainties:

$$\text{diag}(\boldsymbol{\beta})\dot{\hat{\boldsymbol{\theta}}} + \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\text{cmd}} + \hat{\boldsymbol{d}} + \boldsymbol{A}_d(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{13a}$$

$$\hat{\boldsymbol{d}}' = -(e^{\boldsymbol{A}_d dt} - \boldsymbol{I}_{4\times 4})^{-1}\boldsymbol{A}_d e^{\boldsymbol{A}_d dt}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{13b}$$

$$\hat{\boldsymbol{d}} \leftarrow \text{low pass filter}(\hat{\boldsymbol{d}}, \hat{\boldsymbol{d}}'). \tag{13c}$$

where $\hat{\boldsymbol{\theta}}$ is the estimated joint state, and the disturbance term $\hat{\boldsymbol{d}} \in \mathbb{R}^4$. Thus, the final joint control command $\boldsymbol{\theta}^*$, incorporating the adaptive disturbance compensation, is given by:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_{\text{cmd}} + \hat{\boldsymbol{d}}. \tag{14}$$

VII. EE-Centric Teleoperation and Policy Learning

As we mentioned, our framework enables the decoupling between the high-level policy and low-level controller, with the ee-centric interface serving as the sole connection between them. This allows the policy to be embodiment-agnostic, eliminating the need to consider low-level tracking control. In this section, we introduce two aerial manipulation systems we developed based on this framework: the ee-centric aerial teleoperation system and the imitation-learning-based autonomous aerial manipulation system.

A. EE-Centric Aerial Teleoperation

We developed an aerial manipulation teleoperation system with the ee-centric interface, allowing the operator to focus solely on controlling the target end-effector pose, as if they have complete control of a freely moving hand in 3D space.

Robotic teleoperation requires bidirectional communication between the user and the robot. For the user-to-robot command, we developed a gamepad program so that the user can control the end-effector position ${\boldsymbol{p}_E}^r$ and orientation ${\boldsymbol{R}_E^W}^r$ with buttons and joysticks.

For robot-to-user communication, different from tabletop and mobile manipulation settings, users in aerial manipulation settings often lack direct visual access to the workspace, making it necessary to rely on onboard perception systems. In our work, we address this limitation by providing real-time visualization of RGB images captured from cameras mounted on both the end-effector and the base of the UAM. These images are displayed on monitors for continuous user observation. Further enhancing teleoperation efficacy, we have found it crucial to also visualize the user's inputs directly. To this end, we render the commanded target end-effector pose trajectories in real-time within 3D world frame plots. This dual approach of visual feedback not only improves spatial awareness but also significantly enhances user performance in teleoperation tasks.

B. EE-Centric Policy Learning

To establish the autonomous aerial manipulation policy for versatile tasks, we develop an ee-centric policy learning framework based on imitation learning. Specifically, we adopt Action Chunk with Transformer (ACT) as the network structure [59]. ACT utilizes a Conditional Variational Autoencoder (CVAE) where the encoder compresses action sequences and joint observations into a latent style variable. The transformer-based decoder generates action sequences from the latent variable (only during training and set to be the mean of the prior during testing), current joint observations, and encoded image features. The action chunking mitigates compounding errors and enhances the model's ability by predicting multiple future actions at once.

In this work, the ACT policy as well as policy observation and action are defined as follows:

$$\boldsymbol{a}_{t:t+K} = \pi_\varphi(\boldsymbol{o}_t) \tag{15}$$

$$\boldsymbol{o}_t = \left\{I_E,\ I_B,\ \boldsymbol{p}_E,\ \boldsymbol{R}_E^W\right\}_t \tag{16}$$

$$\boldsymbol{a}_t = \left\{\boldsymbol{p}_E^r,\ {\boldsymbol{R}_E^W}^r\right\}_t \tag{17}$$

where π denotes the ACT policy and φ is the network parameter. I_B, I_E are RGB images from the base camera and the end-effector camera, each with 640×480 resolution. K is the chunking size. $\boldsymbol{p}_E$, $\boldsymbol{R}_E^W$, $\boldsymbol{p}_E^r$, ${\boldsymbol{R}_E^W}^r$ denote the current and target UAM position and orientation, respectively. We use ResNet-18 as the backbone to encode the RGB images before inputting them into the transformer encoder. The flowchart of the algorithm implementation is illustrated in Fig. 2.

VIII. Experiments

To validate the effectiveness of our proposed framework, we conduct a series of experiments focusing on end-effector trajectory tracking, aerial teleoperation, and policy learning for autonomous aerial manipulation[1]. We first assess whether the proposed whole-body MPC with L1 adaptation approach enhances trajectory tracking accuracy. Additionally, we evaluate how the ee-centric interface facilitates intuitive and precise teleoperation, reducing operator burden and improving task execution in a series of teleoperation tasks. Finally, we investigate whether the high-quality teleoperation demonstrations can be leveraged to train imitation learning-based policies for autonomous aerial manipulation in both simulation and real-world environments.

A. Experimental Setup

1) Trajectory Tracking Task Setup: To show the effectiveness of our proposed method in end-effector trajectory tracking tasks, we perform a comparison between our control methods against two baseline approaches:

- **w.o. MPC**: This baseline replaces the ee-centric MPC with the Direct Force Feedback Control(DFFC) method from [38], which directly controls the end-effector acceleration based on the current reference pose. However, it lacks a prediction horizon to account for future trajectories.
- **w.o. L1**: This baseline excludes the L1 adaptive component, leaving disturbances from UAV and manipulator interactions and modeling uncertainties uncompensated during control execution.

We conduct experiments with three types of reference trajectories for the end-effector, each lasting 60 seconds. The setpoint trajectory requires the aerial manipulator to keep the end-effector hovering at a fixed position $\boldsymbol{p}_E = [0.0, 0.0, 1.3]$. The ellipse trajectory requires tracking a sinusoidal trajectory defined as $\boldsymbol{p}_E = [0.5\sin(0.3t), 0.0, 1.4 + 0.2\sin(0.3t + 0.75)]$. The figure-8 trajectory requires tracking a trajectory $\boldsymbol{p}_E =$

[1] Please check out our project page for more visualization videos: https://lecar-lab.github.io/flying_hand/.

TABLE III
CONTROLLER PARAMETERS

Horizon Length T	2.5 s
Horizon Steps N	100
State Cost $\boldsymbol{Q}_p$	$\mathrm{diag}(12, 12, 12)$
Rotation Cost $\boldsymbol{Q}_R$	$\mathrm{diag}(10, 10, 10)$
Velocity Cost $\boldsymbol{Q}_v$	$\mathrm{diag}(0.1, 0.1, 0.1)$
Joint Angle Cost $\boldsymbol{Q}_\theta$	$\mathrm{diag}(0.1, 0.1, 0.1)$
Control Cost $\boldsymbol{Q}_u$	$\mathrm{diag}(0.03, 0.03, 0.03, 0.1, 0.1, 0.1)$

$[0.1 + 0.6\sin(0.3t), 0.0, 1.35 + 0.25\sin(0.6t)]$. The maximum velocity in the reference trajectory is about 0.2 m/s. The pitch, row, and yaw attitude of the end-effector are all fixed at zero during the tracking. Root Mean Square Error (RMSE) is used as the tracking performance evaluation criterion. Each trajectory is repeated three times to compute the mean and standard deviation.

2) Aerial Manipulation Task Setup: We conducted a series of experiments to evaluate the capabilities and applications of our aerial manipulation system. We select different typical tasks from each category we discussed in Sec. II-A, including:

- **Aerial Writing**: Drawing a target shape (the digit "2025") on a vertical wall, with an overall size of approximately 3m×0.8m. This task required precise specification and tracking of the end-effector pose trajectory while maintaining stable contact with the surface.
- **Aerial Peg-in-Hole**: Inserting a 20 mm diameter pole into a 50 mm diameter hole positioned around 150 cm above the ground.
- **Rotate Valve**: Manipulating the articulated valve by grasping its handle and rotating it along a 20 cm diameter circle, emulating industrial valve manipulation.
- **Aerial Pick and Place**: Grasping and placing various objects with different shapes and sizes, including the screwdriver, pen, tape, and glue bottle.
- **Unmount Light Bulb**: Grasping a mounted light bulb and unscrewing it from the socket.
- **Mount Light Bulb**: A long horizon task that requires a sequence of motion, including inserting a light bulb into a socket, screwing it in, and subsequently turning it on by pressing the button.

For different tasks, different end-effectors are adopted, including the parallel jaw gripper for the pick and place and peg-in-hole task, a passive elastic claw for grasping the light bulb, and a bucket-shaped gripper for rotating the valve.

B. Implementation Details

The optimal control problem in the ee-centric MPC is implemented using ACADOS [51] with a 25ms discretisation step and a 2.5s constant prediction horizon, running in 100 Hz, and other controller parameters are listed in Table III. The control output is executed in a receding horizon style, where at each iteration, only the first control input $\boldsymbol{u}_0$ is applied to the system.

Since both the L1 adaptive controller and the MPC controller require accurate system modeling $\boldsymbol{f}_{dyn}$ to achieve effective control performance, we perform system identification to estimate uncertain parameters shown in Table II. We excite the system with two types of motions. First, arm motion-only trajectories are executed while keeping the UAV stationary to calibrate the DH parameters $\boldsymbol{\zeta}$ and joint servo delay $\boldsymbol{\beta}$. These trajectories ensure that the manipulator kinematic parameters and joint motor dynamics accurately reflect the actual manipulator motion response. Second, UAV free-flight trajectories are conducted to identify the drone dynamics described in Eq. (1).

C. Experiment I: Control Performance

Table IV shows the comparison of our approach against w.o. MPC and w.o. L1 baselines in three types of reference end-effector trajectories. The results show that our proposed method achieves the lowest tracking error, with approximately 1 cm in hover and 4 cm during motion. In contrast, the baseline w.o. L1 exhibits 1.3 cm and 6.5 cm, respectively, while the baseline w.o. MPC performs the worst, with 2 cm in hover and 8 cm in motion. As shown in Fig. 4, compared with our method (blue), the baseline w.o. L1 (green) exhibits overshoot in the X and Z axes and bias in Y, indicating that the L1 controller effectively mitigates both transient and steady-state errors caused by model uncertainties. The baseline w.o. MPC (orange) suffers from significant motion lag, as DFFC fails to account for trajectory feedforward. Fig. 5 depicts the error distribution for all trajectories using the proposed methods and two baselines, showing that our method achieves the smallest and most centered error, whereas the L1 baseline exhibits steady-state errors and the MPC baseline displays a broader error spread due to dynamic lag. These results confirm the effectiveness of our proposed control scheme.

To further analyze the contribution of the L1 adaptive control, specifically its effectiveness in handling model mismatch and interaction disturbance, we visualized the base external wrench τ_{ext} measured from motion capture by numerical differentiation and the estimated wrench $\hat{\tau}_{ext}$ from L1 adaptation. The interaction force is within 5 N and the torque within 1.4 Nm. Fig. 8 shows disturbances along the base x (red), z (blue) and θ_{pitch} (green), respectively. The disturbances and model uncertainties primarily arise from arm motions, inaccurate thrust gain, and hovering thrust bias. Shaded areas indicate obvious interaction torque in base pitch resulting from arm motion. The results demonstrate that the L1 adaptive controller accurately compensates for base disturbances, effectively mitigating the impact of model mismatch.

We also investigate the contribution of arm flexibility to end-effector tracking performance by increasing the arm control cost 5 times. Fig. 6a *Slow-Arm* results show a 35% larger end-effector tracking error when arm flexibility is restricted, with a more oscillating end-effector trajectory (Fig. 6c) compared to the MPC with flexible arm (Fig. 6b). This aligns with the intuition that higher arm flexibility improves end-effector

TABLE IV
END-EFFECTOR TRAJECTORY TRACKING PERFORMANCE

RMSE (cm)	Setpoint	Ellipse	Figure-8
Our Method	**1.00 ± 0.11**	**3.98 ± 0.41**	**4.62 ± 0.58**
w.o. L1 Adaptation	1.33 ± 0.16	6.67 ± 0.42	6.28 ± 0.50
w.o. MPC	2.07 ± 0.19	8.52 ± 0.79	7.35 ± 0.51

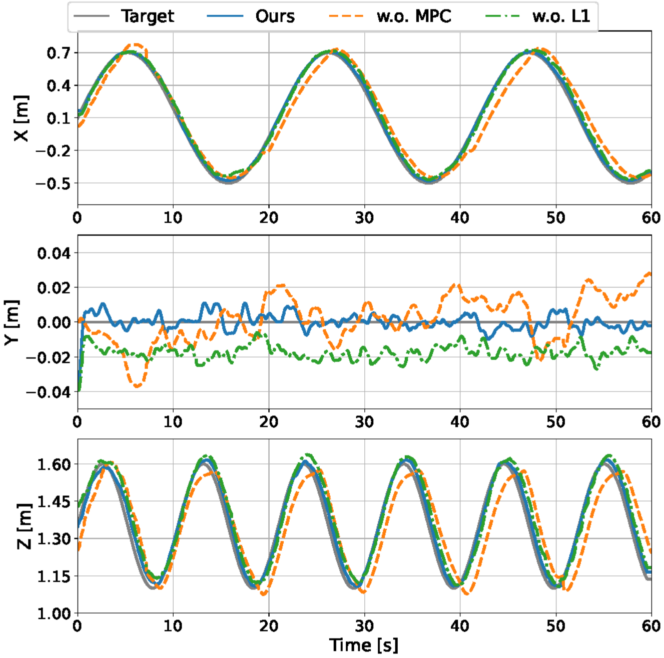

Fig. 4. End-effector tracking performance of aerial manipulator in Ellipse trajectory. Tracking results indicate that the w.o. MPC baseline exhibits significant tracking lag, while the w.o. L1 baseline suffers from static tracking errors due to model mismatches.

tracking performance, as the arm typically responds faster than the drone base.

Despite the high tracking performance of the proposed methods, Fig. 7 reveals that tracking error increases at lower altitudes (around 1m), likely due to unmodeled ground and wall effect disturbances. Additionally, oscillations in the end-effector trajectory are observed across all methods. We notice servo backlash (around 0.5° dead zone), which results in a 2 cm control dead zone in the end-effector task space, limiting tracking precision during fast UAV maneuvers. Further improvements can be achieved through more accurate system modeling and higher-precision hardware to enhance tracking accuracy.

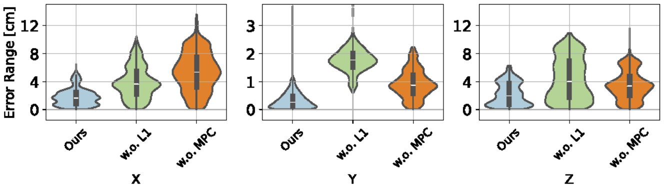

Fig. 5. End-effector tracking error distribution for three types of trajectories using our methods and two baselines.

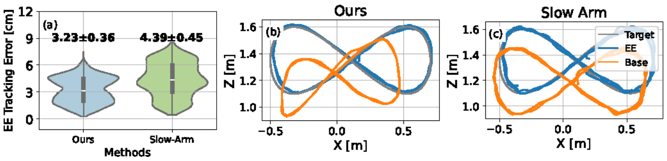

Fig. 6. Arm flexibility ablation study for MPC controller.

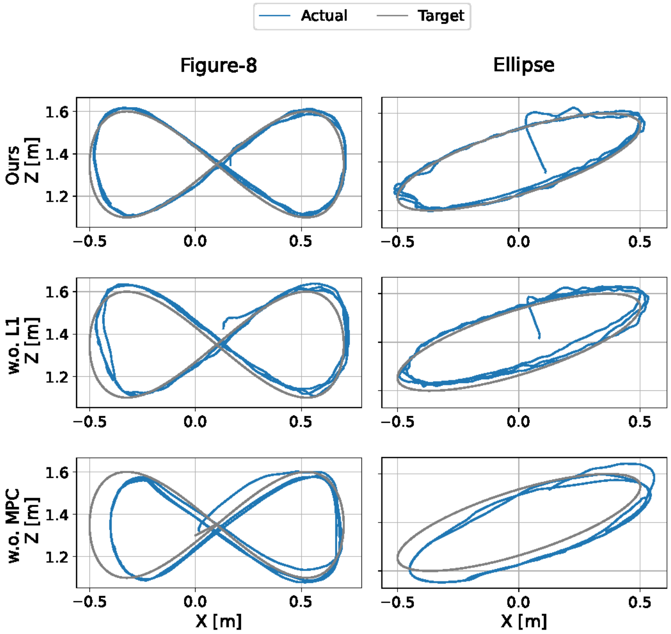

Fig. 7. Comparison of Figure-8 and Ellipse trajectory tracking performance across three methods. Our approach achieves the lowest tracking error in dynamic trajectory tracking tasks.

D. Experiment II: Aerial Teleoperation

First, we demonstrate the effectiveness and versatility of the proposed teleoperation system by targeting aerial writing, rotating the valve, aerial pick and place, unmount, and mount light bulb tasks. As shown in Fig. 10 and Fig. 11, human teleoperators can easily achieve all aerial manipulation tasks with little learning and operation cost. One key success factor we attribute is the ee-centric interface, which reduces human effort and improves the quality of teleoperation data for future policy learning.

In a subsequent experiment, we evaluate the benefits of directly controlling the end-effector's pose using our framework against controlling each degree of freedom (DoF) for UAVs and robotic arms [56] in a simulated peg-in-hole task. The teleoperation command trajectories are illustrated in Fig. 9a. Direct control of the end-effector allowed operators to issue more fluid command trajectories, significantly enhancing the precision of end-effector movements and decreasing the time required to complete the task.

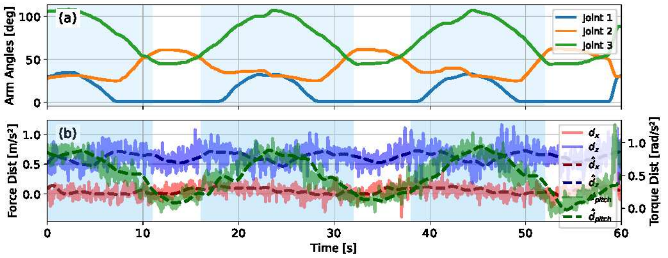

Fig. 8. (a). Arm joint angles (b). Disturbance τ_{ext} and L1 disturbance estimation $\hat{\tau}_{ext}$ of x, z, and θ_{pitch}.

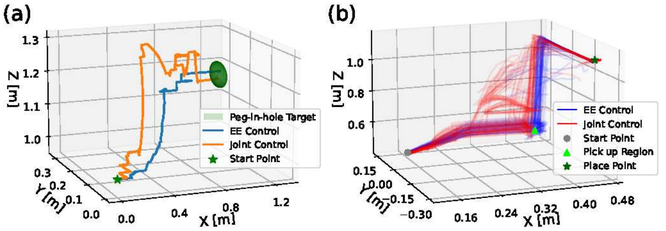

Fig. 9. (a). End-effector command trajectory using the ee-centric teleoperation interface and full-DoF teleoperation interface, in a simulated peg-in-hole task. (b). End-effector command trajectory of the learned autonomous policy during 50 test trials using the ee-centric interface and using the full-DoF interface, in a simulated pick and place task.

TABLE V
IMITATION LEARNING SIMULATION SUCCESS RATE

	Rotate Valve	Pick & Place	Peg in Hole	Open & Retrieve
Joint Space	50/50	38/50	9/50	8/50
EE	50/50	**48/50**	**23/50**	**17/50**

E. Experiment III: Learning from Demonstration

1) Simulation Experiments: We first demonstrate our learning from demonstration framework in Mujoco [48] simulator with four tasks: (i) *peg-in-hole*, (ii) *rotate valve*, (iii) *pick and place* and (iv) *open and retrieve*, as shown in Fig. 12 (a). To collect demonstrations, we use a scripted policy for each task. Every episode of the scripted policy lasts about 12 seconds for each task. We collect 50 episodes for each task. Note that with our ee-centric interface, we do not consider any joint configuration when collecting demonstrations, which allows us to efficiently collect smooth demonstrations without tediously adjusting each joint position to complete the task. Our ACT policy for each task in the simulation is trained with the action chunk size of 100 and limited 5000 epochs. More details on implementation are included in Appendix A. After training, we choose the policy with the least validation loss to perform 50 evaluation trials. The evaluation result is shown in Table V.

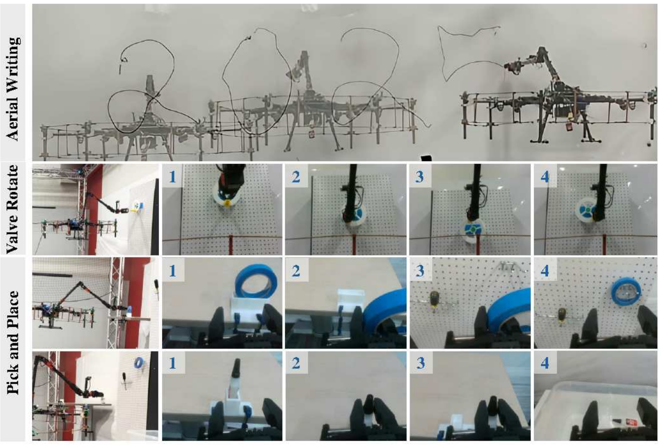

Fig. 10. Aerial Teleoperation Manipulation Tasks. We target 1) Aerial Writing: UAM with a marker pen writes ’2025’ on a whiteboard. 2) Rotate Valve: UAM grasps the handle and rotates the valve with one loop. 3) Pick-and-Place: UAM grasps an object and relocates it to a designated area.

To show the advantage of learning from an ee-centric demonstration compared to a joint space demonstration, we use the same demonstration trajectory but change the observation and action to be in UAM configuration space, i.e., the UAV position and orientation, and each joint angle of the robotic arm. After that, we train a joint space ACT policy with the same training setting as the ee-centric ACT policy, except that the end-effector pose in the observation and action space is replaced by the drone base pose and full manipulator joint angles. Their success rate comparison is summarized in Table V. It shows that with limited 5000 training epochs, ee-centric policy outperforms the joint space policy in challenging tasks including *pick and place*, *peg in hole*, and *open and retrieve*. As illustrated in Fig. 9b, the ee-centric policy targets the pickup region and the place point more precisely, while the joint space policy is more prone to generate wrong targets.

Overall, the experimental results reveal several complementary insights:

- **Geometric Precision Advantage**: Our ee-centric policy achieves 2.5× higher success rate in geometrically sensitive *peg in hole* task, directly benefiting from task-space supervision that eliminates the accumulated end-effector error from the joint space.
- **Multi-Skill Composition**: In the *open and retrieve* task, our ee-centric policy achieves 2× higher success rate than the joint space policy, which demonstrates its inherent advantages in multi-skill decomposition and execution.

2) Real-world Experiments: We adopt the aerial peg-in-hole task to demonstrate our capability to derive an autonomous policy from human demonstrations for aerial manipulation in the real world. The task configurations are illustrated in Fig. 13 and Fig. 14.

We collected 25 episodes of demonstration data via human teleoperation, varying the hole’s horizontal position, with each episode taking approximately 2 minutes, culminating in a total of around 50 minutes of operational data and about 2 hours of wall-clock time. The data is downsampled to 10 Hz, and the action chunk size is empirically set to 100 during the

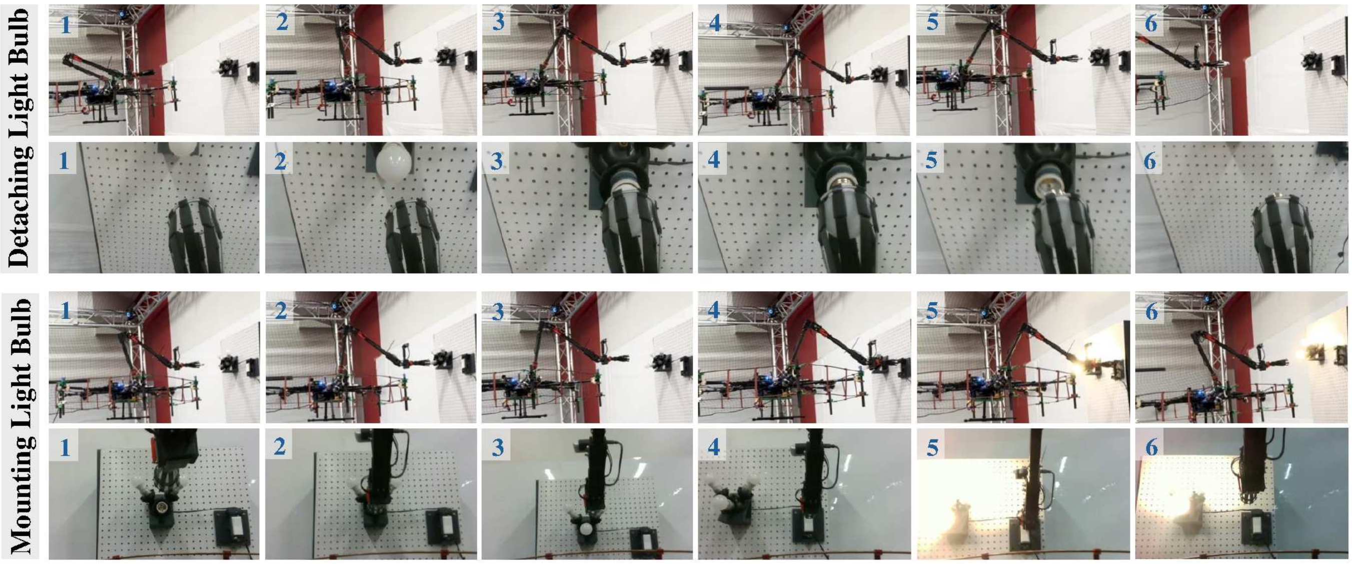

Fig. 11. Long horizon aerial teleoperation light bulb changing task. UAM grasps the light bulb and unscrews it during the first flight. And it inserts, screws a new light bulb, and presses the button to turn on the light during the second flight.

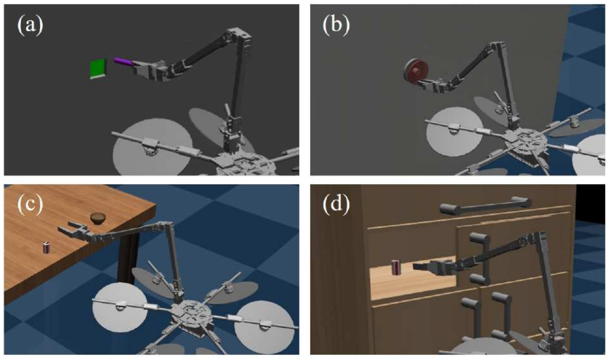

Fig. 12. Task setup in Mujoco simulation, including (a) Peg-in-Hole; (b) Rotate the Valve; (c) Pick and Place; and (d) a long horizon Open and Retrieve task.

training process. After training through 100,000 epochs, the policy with the least validation loss is selected. We tested with random unseen horizontal hole positions and the learned policy successfully completed 4 out of 5 real-world peg-in-hole tests, i.e., 80% successful rate. The UAM pushed the peg forward to the edge of the hole and didn't insert it inside successfully. These results underline the potential of learning-based approaches in aerial manipulation under our developed framework, while also highlighting the need to develop robust recovery policies, especially for aerial manipulation.

F. Discussion

The experiments demonstrate our framework in end-effector trajectory tracking, aerial teleoperation, and policy learning for autonomous aerial manipulation. The precise end-effector control framework demonstrated superior end-effector tracking accuracy with minimal error. The high-precision control enables efficient, user-friendly aerial teleoperation, allowing human operators to perform multiple complex tasks, which also helps high-quality demonstration data collection. Leveraging the ee-centric framework, advanced high-level policies such as imitation learning can be easily incorporated into aerial manipulation, which further expands the development of this field.

IX. Limitations

Although we have demonstrated the proposed framework through various real-world experiments, there are still several limitations due to time constraints and methodological limitations. First, all of our experiments were conducted indoors within a motion capture system, where we achieved millimeter-level state estimation of the UAV and end-effector states using forward kinematics. This setup limits its practical application in real-life scenarios. Second, the current safety constraints are predefined. Incorporating onboard perception to detect obstacles and generate safety constraints in real-time will be our next step, as various studies have demonstrated the feasibility of UAV collision-free flight. Third, the current performance is limited by the robotic arm actuators, which have relatively large backlash and would generate unavoidable vibrations. Finally, although the proposed framework, which decouples different modules, demonstrates the potential for cross-platform compatibility and integration with general manipulation fields, more real-world experiments are planned to be conducted to further validate its effectiveness.

X. Conclusion

This work presents a unified end-effector-centric aerial manipulation framework for versatile aerial manipulation tasks. Our system includes a versatile hardware platform that consists of a fully actuated UAV and a 4-DOF manipulator, an

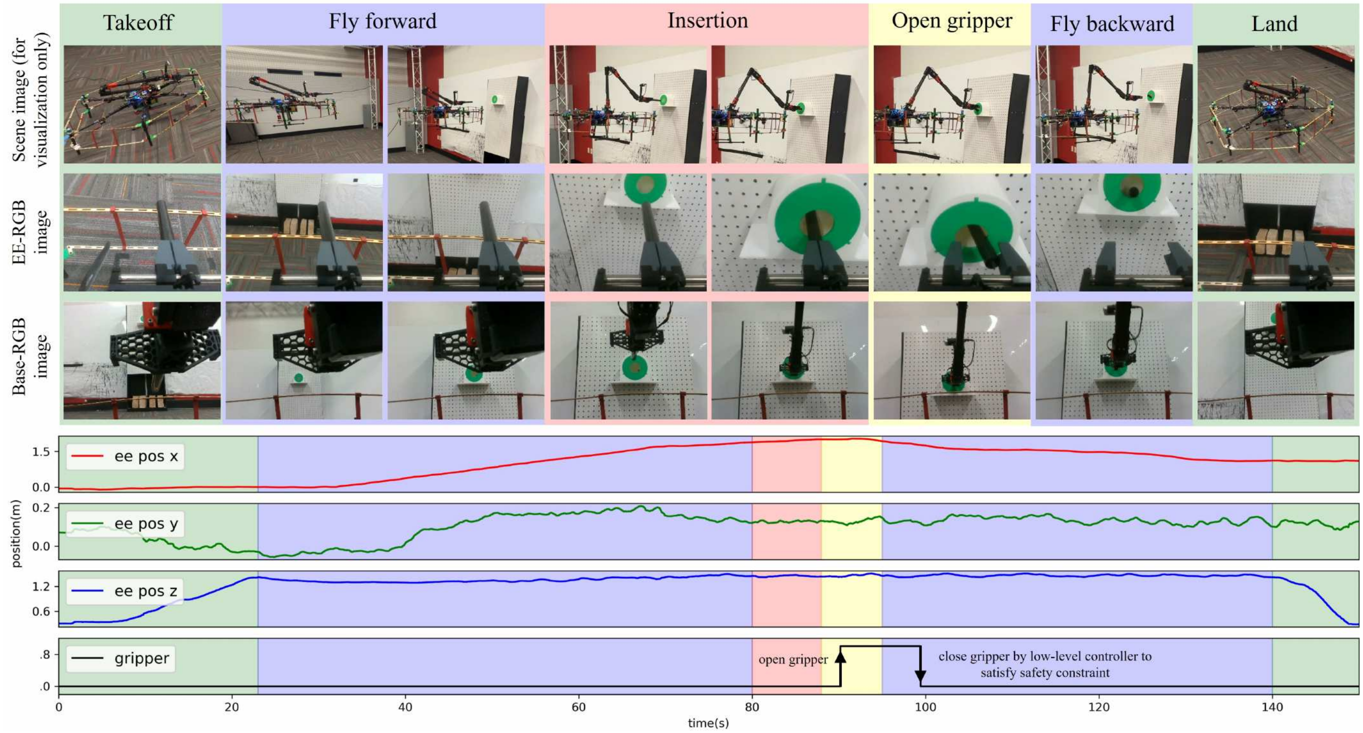

Fig. 13. Autonomous aerial manipulation peg-in-hole policy experiment. The UAM inserts the peg precisely, highlighting both the accuracy of the learned policy and the low-level controller.

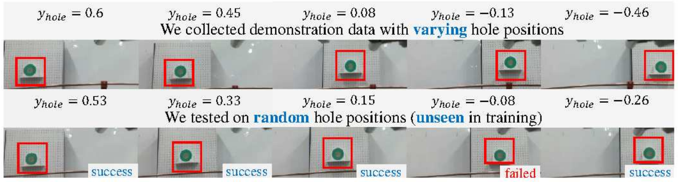

Fig. 14. Task scenario randomization for peg-in-hole data collection and test.

ee-centric whole-body MPC to ensure precise end-effector tracking, and an ee-centric high-level policy, where we developed both an intuitive teleoperation system and an imitation learning-based autonomous system. Through extensive real-world experiments, we demonstrated the system's versatility across various aerial manipulation tasks, including writing, peg-in-hole, pick-and-place, valve rotating, and light bulb replacement. More importantly, we demonstrate how this modular and standardized ee-centric framework effectively decouples the high-level policy from the low-level controller, which enables seamless integration of existing standard high-level policy modules from the broader manipulation community, such as teleoperation and imitation learning, into the field of aerial manipulation. The proposed framework achieved high precision, adaptability, and robust performance, making it a significant step toward standardizing aerial manipulation within the broader manipulation field. Future work will extend the framework's applicability to outdoor environments, incorporate onboard perception for obstacle avoidance, and further improve the end-effector tracking performance.

Acknowledgments

This research is partially sponsored by the CMU Manufacturing Futures Institute (MFI). Any opinions, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the MFI.

References

[1] Mike Allenspach, Nicholas Lawrance, Marco Tognon, and Roland Siegwart. Towards 6dof bilateral teleoperation of an omnidirectional aerial vehicle for aerial physical interaction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9302–9308. IEEE, 2022.

[2] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[3] Karen Bodie, Maximilian Brunner, Michael Pantic, Stefan Walser, Patrick Pfändler, Ueli Angst, Roland Siegwart, and Juan Nieto. An omnidirectional aerial manipulation platform for contact-based inspection. *arXiv preprint arXiv:1905.03502*, 2019.

[4] Karen Bodie, Maximilian Brunner, Michael Pantic, Stefan Walser, Patrick Pfändler, Ueli Angst, Roland Siegwart, and Juan Nieto. Active interaction force control for contact-based inspection with a fully actuated aerial

vehicle. *IEEE Transactions on Robotics*, 37(3):709–722, 2020.

[5] Karen Bodie, Marco Tognon, and Roland Siegwart. Dynamic end effector tracking with an omnidirectional parallel aerial manipulator. *IEEE Robotics and Automation Letters*, 6(4):8165–8172, 2021.

[6] Maximilian Brunner, Livio Giacomini, Roland Siegwart, and Marco Tognon. Energy tank-based policies for robust aerial physical interaction with moving objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2054–2060. IEEE, 2022.

[7] Maximilian Brunner, Giuseppe Rizzi, Matthias Studiger, Roland Siegwart, and Marco Tognon. A planning-and-control framework for aerial manipulation of articulated objects. *IEEE Robotics and Automation Letters*, 7(4): 10689–10696, 2022.

[8] Fredrik Båberg, Yuquan Wang, Sergio Caccamo, and Petter Ögren. Adaptive object centered teleoperation control of a mobile manipulator. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 455–461, 2016. doi: 10.1109/ICRA.2016.7487166.

[9] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[10] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots, 2024. URL https://arxiv.org/abs/2402.10329.

[11] Andre Coelho, Yuri Sarkisov, Xuwei Wu, Hrishik Mishra, Harsimran Singh, Alexander Dietrich, Antonio Franchi, Konstantin Kondak, and Christian Ott. Whole-body teleoperation and shared control of redundant robots with applications to aerial manipulation. *Journal of Intelligent & Robotic Systems*, 102:1–22, 2021.

[12] Jacques Denavit and Richard S Hartenberg. A kinematic notation for lower-pair mechanisms based on matrices. 1955.

[13] Caiwu Ding, Lu Lu, Cong Wang, and Caiwen Ding. Design, sensing, and control of a novel uav platform for aerial drilling and screwing. *IEEE Robotics and Automation Letters*, 6(2):3176–3183, 2021.

[14] Yanming Feng, Jinling Wang, et al. Gps rtk performance characteristics and analysis. *Positioning*, 1(13), 2008.

[15] Figure AI. Helix: A vision-language-action model for generalist humanoid control. https://www.figure.ai/news/helix, February 2025. Accessed: 2025-04-23.

[16] Joshua Fishman, Samuel Ubellacker, Nathan Hughes, and Luca Carlone. Dynamic grasping with a" soft" drone: From theory to practice. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4214–4221. IEEE, 2021.

[17] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.

[18] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation, 2024. URL https://arxiv.org/abs/2401.02117.

[19] Xiaofeng Guo, Guanqi He, Mohammadreza Mousaei, Junyi Geng, Guanya Shi, and Sebastian Scherer. Aerial interaction with tactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1576–1582. IEEE, 2024.

[20] Xiaofeng Guo, Guanqi He, Jiahe Xu, Mohammadreza Mousaei, Junyi Geng, Sebastian Scherer, and Guanya Shi. Flying calligrapher: Contact-aware motion and force planning and control for aerial manipulation. *IEEE Robotics and Automation Letters*, 2024.

[21] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers, 2024. URL https://arxiv.org/abs/2407.10353.

[22] Guanqi He, Yash Jangir, Junyi Geng, Mohammadreza Mousaei, Dongwei Bai, and Sebastian Scherer. Image-based visual servo control for aerial manipulation using a fully-actuated uav. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5042–5049. IEEE, 2023.

[23] Guanqi He, Yogita Choudhary, and Guanya Shi. Self-supervised meta-learning for all-layer dnn-based adaptive control with stability guarantees. *arXiv preprint arXiv:2410.07575*, 2024.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[26] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. *arXiv preprint arXiv:2410.21229*, 2024.

[27] Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Learning kinematic feasibility for mobile manipulation through deep reinforcement learning. *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[28] Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. $\mathrm{N}^2\mathrm{m}^2$: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments. *IEEE Transactions on Robotics*, 2023. doi: 10.1109/TRO.2023.3284346.

[29] Kevin Huang, Rwik Rana, Alexander Spitzer, Guanya Shi, and Byron Boots. Datt: Deep adaptive trajectory tracking for quadrotor control. In *Conference on Robot*

Learning, pages 326–340. PMLR, 2023.

[30] Jialin Ji, Tiankai Yang, Chao Xu, and Fei Gao. Real-time trajectory planning for aerial perching. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10516–10522. IEEE, 2022.

[31] Oussama Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal on Robotics and Automation*, 3(1):43–53, 1987.

[32] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of machine learning research*, 22(30):1–82, 2021.

[33] Christian Lanegger, Marco Ruggia, Marco Tognon, Lionel Ott, and Roland Siegwart. Aerial layouting: Design and control of a compliant and actuated end-effector for precise in-flight marking on ceilings. *Proceedings of Robotics: Science and System XVIII*, page p073, 2022.

[34] Frank L Lewis, Darren M Dawson, and Chaouki T Abdallah. *Robot manipulator control: theory and practice*. CRC Press, 2003.

[35] Matthew T Mason. Toward robotic manipulation. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):1–28, 2018.

[36] Daniel Mellinger, Quentin Lindsey, Michael Shomin, and Vijay Kumar. Design, modeling, estimation and control for aerial grasping and manipulation. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2668–2673. IEEE, 2011.

[37] Jiawei Meng, Joao Buzzatto, Yuanchang Liu, and Minas Liarokapis. On aerial robots with grasping and perching capabilities: A comprehensive review. *Frontiers in Robotics and AI*, 8:739173, 2022.

[38] Gabriele Nava, Quentin Sablé, Marco Tognon, Daniele Pucci, and Antonio Franchi. Direct force feedback control and online multi-task optimization for aerial manipulators. *IEEE Robotics and Automation Letters*, 5(2):331–338, 2020. doi: 10.1109/LRA.2019.2958473.

[39] Anibal Ollero, Marco Tognon, Alejandro Suarez, Dongjun Lee, and Antonio Franchi. Past, present, and future of aerial robotic manipulators. *IEEE Transactions on Robotics*, 38(1):626–645, 2021.

[40] Michael O'Connell, Guanya Shi, Xichen Shi, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural-fly enables rapid learning for agile flight in strong winds. *Science Robotics*, 7(66): eabm6597, 2022.

[41] William RT Roderick, Mark R Cutkosky, and David Lentink. Bird-inspired dynamic grasping and perching in arboreal environments. *Science Robotics*, 6(61): eabj7562, 2021.

[42] Micha Schuster, David Bernstein, Paul Reck, Salua Hamaza, and Michael Beitelschmidt. Automated aerial screwing with a fully actuated aerial manipulator. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3340–3347, 2022. doi: 10.1109/IROS47612.2022.9981979.

[43] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3):4978–4985, 2020.

[44] Yao Su, Jiarui Li, Ziyuan Jiao, Meng Wang, Chi Chu, Hang Li, Yixin Zhu, and Hangxin Liu. Sequential manipulation planning for over-actuated unmanned aerial manipulators. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6905–6911. IEEE, 2023.

[45] Alejandro Suarez, Victor M Vega, Manuel Fernandez, Guillermo Heredia, and Anibal Ollero. Benchmarks for aerial manipulation. *IEEE Robotics and Automation Letters*, 5(2):2650–2657, 2020.

[46] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[47] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[48] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

[49] Dimos Tzoumanikas, Felix Graule, Qingyue Yan, Dhruv Shah, Marija Popovic, and Stefan Leutenegger. Aerial manipulation using hybrid force and position nmpc applied to aerial writing. *arXiv preprint arXiv:2006.02116*, 2020.

[50] Samuel Ubellacker, Aaron Ray, James M. Bern, Jared Strader, and Luca Carlone. High-speed aerial grasping using a soft drone with onboard perception. *npj Robotics*, 2(1):5, 2024. ISSN 2731-4278. doi: 10.1038/s44182-024-00012-1. URL https://doi.org/10.1038/s44182-024-00012-1.

[51] Robin Verschueren, Gianluca Frison, Dimitris Kouzoupis, Jonathan Frey, Niels van Duijkeren, Andrea Zanelli, Branimir Novoselnik, Thivaharan Albin, Rien Quirynen, and Moritz Diehl. acados – a modular open-source framework for fast embedded optimal control. *Mathematical Programming Computation*, 2021.

[52] Meng Wang, Zeshuai Chen, Kexin Guo, Xiang Yu, Youmin Zhang, Lei Guo, and Wei Wang. Millimeter-level pick and peg-in-hole task achieved by aerial manipulator. *IEEE Transactions on Robotics*, 2023.

[53] Zhuohuan Wu, Sheng Cheng, Kasey A. Ackerman, Aditya Gahlawat, Arun Lakshmanan, Pan Zhao, and Naira Hovakimyan. L1adaptive augmentation for geo-

metric tracking control of quadrotors. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1329–1336, 2022. doi: 10.1109/ICRA46639.2022.9811946.

[54] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024.

[55] Ruihan Yang, Yejin Kim, Rose Hendrix, Aniruddha Kembhavi, Xiaolong Wang, and Kiana Ehsani. Harmonic mobile manipulation, 2024. URL https://arxiv.org/abs/2312.06639.

[56] Grigoriy A Yashin, Daria Trinitatova, Ruslan T Agishev, Roman Ibrahimov, and Dzmitry Tsetserukou. Aerovr: Virtual reality-based teleoperation with tactile feedback for aerial manipulation. In *2019 19th International Conference on Advanced Robotics (ICAR)*, pages 767–772. IEEE, 2019.

[57] Jianke Zhang, Yanjiang Guo, Xiaoyu Chen, Yen-Jen Wang, Yucheng Hu, Chengming Shi, and Jianyu Chen. Hirt: Enhancing robotic control with hierarchical robot transformers. *arXiv preprint arXiv:2410.05273*, 2024.

[58] Shibo Zhao, Hengrui Zhang, Peng Wang, Lucas Nogueira, and Sebastian Scherer. Super odometry: Imu-centric lidar-visual-inertial estimator for challenging environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8729–8736. IEEE, 2021.

[59] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[60] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.

Appendix

A. Policy Learning Implementation Details

For each task in the simulation, we randomize the task setup during each trial in training and test to evaluate the robustness of the policy, as shown in Table VI. The hyperparameters of the ACT are shown in Table VII. Those hyperparameters are used for both simulation and real-world experiments.

TABLE VI
Task Randomization Settings

Task	Randomized item	Randomization range (m)
peg-in-hole	hole	$p_y \sim \mathcal{U}(-0.3, 0.3)$
rotate valve	UAM	$p_y \sim \mathcal{U}(-0.15, 0.15)$ $p_z \sim \mathcal{U}(0.85, 1.15)$
pick and place	object	$p_y \sim \mathcal{U}(0.05, 0.12)$ $p_z \sim \mathcal{U}(0.5, 0.6)$
open and retrieve	object	$p_y \sim \mathcal{U}(0.2, 0.4)$ $p_z \sim \mathcal{U}(1.16, 1.26)$

TABLE VII
Hyperparameters of ACT

learning rate	1e-5
batch size	16
feedforward dimension	3200
chunk size	100
vision backbone	pretrained ResNet18 [24]

B. Policy Learning with Less Accurate State Estimation

In this work, we use a motion capture system to get the state estimation of the drone itself and further get the state estimation of the end-effector using proprioception information by forward kinematics. However, in more realistic scenarios like the outdoor environment, the high-precision motion capture system may not be applicable, and it further challenges the policy learning. To simulate this, we further evaluate the simulated peg-in-hole task with 1 cm state estimation noise, an achievable precision with an RTK GPS in the real world [14]. It still achieved a 42% successful rate with a limited 5k training epochs, proving the potential to deploy our system in outdoor environments without a motion capture system.