



# JuggleRL: Mastering Ball Juggling with a Quadrotor via Deep Reinforcement Learning

Shilong Ji<sup>1</sup>, Yinuo Chen<sup>2</sup>, Chuqi Wang<sup>2</sup>, Jiayu Chen<sup>2</sup>, Ruize Zhang<sup>1</sup>, Feng Gao<sup>3</sup>, Wenhao Tang<sup>1</sup>,  
Shu'ang Yu<sup>2</sup>, Sirui Xiang<sup>2</sup>, Xinlei Chen<sup>1</sup>, Chao Yu<sup>1</sup>, Yu Wang<sup>2</sup>

**Abstract**—Aerial robots interacting with objects must perform precise, contact-rich maneuvers under uncertainty. In this paper, we study the problem of aerial ball juggling using a quadrotor equipped with a racket, a task that demands accurate timing, stable control, and continuous adaptation. We propose JuggleRL, the first reinforcement learning-based system for aerial juggling. It learns closed-loop policies in large-scale simulation using systematic calibration of quadrotor and ball dynamics to reduce the sim-to-real gap. The training incorporates reward shaping to encourage racket-centered hits and sustained juggling, as well as domain randomization over ball position and coefficient of restitution to enhance robustness and transferability. The learned policy outputs mid-level commands executed by a low-level controller and is deployed zero-shot on real hardware, where an enhanced perception module with a lightweight communication protocol reduces delays in high-frequency state estimation and ensures real-time control. Experiments show that JuggleRL achieves an average of 311 hits over 10 consecutive trials in the real world, with a maximum of 462 hits observed, far exceeding a model-based baseline that reaches at most 14 hits with an average of 3.1. Moreover, the policy generalizes to unseen conditions, successfully juggling a lighter 5 g ball with an average of 145.9 hits. This work demonstrates that reinforcement learning can empower aerial robots with robust and stable control in dynamic interaction tasks. The project page is at <https://jugglerl.github.io/>.

## I. INTRODUCTION

Despite significant advances in navigation [1], motion planning [2] and stable control [3], dynamic interaction with the environment remains one of the central challenges in robotics. Tasks such as pushing, striking, or juggling are inherently reactive, contact-rich, and must be performed under tight temporal constraints, often in the presence of model uncertainty, sensing noise, and environmental variability. Aerial robots face these challenges even more acutely, as they must not only stabilize themselves in flight but also coordinate precise interactions with fast-moving objects. In this paper, we study the problem of aerial ball juggling using a quadrotor equipped with a racket. Despite its simplicity, this task demands accurate timing, robust contact handling,


and rapid adaptation, making it a representative benchmark for real-time interaction under uncertainty.

Traditional control approaches [4], [5], [6] address this problem by predicting ball trajectories, calculating contact points, and planning quadrotor motions to satisfy these constraints. However, such methods are brittle in practice, due to several fundamental limitations:

- **Nonlinear ball dynamics:** The ball's motion model is highly nonlinear and lacks a closed-form solution. Existing methods rely on numerical integration, which inevitably trades off between accuracy and speed. Such computational limitations introduce timing errors that become especially critical under low-bounce conditions, often driving the system into physically infeasible states.
- **Error accumulation:** Trajectory planning must be performed in advance based on predicted contact points. Because subsequent execution follows an open-loop strategy without incorporating new observations, any prediction error cannot be corrected and instead accumulates over time, ultimately leading to a severe degradation in task success rates.
- **Hard trajectory constraints:** Trajectory optimization requires predefined terminal states and hitting times. These hard constraints shrink the feasible solution space and often lead to infeasibility under real-world uncertainties.

Reinforcement learning (RL), in contrast, offers a compelling alternative paradigm. Rather than explicitly computing optimal trajectories or solving inverse dynamics during execution, RL acquires sensorimotor policies through interaction, thereby enabling closed-loop feedback, online error correction, and adaptive responses to novel disturbances.

In this paper, we present **JuggleRL**, the first reinforcement learning-based system for aerial ball juggling. The framework is developed in three stages. First, we perform system identification (SysID) to calibrate the dynamics of both the quadrotor and the ball, narrowing the gap between the physical platform and its digital counterpart. Second, we construct a large-scale GPU-parallel training environment based on Isaac Sim [7], supporting contact-rich interactions. Within this environment, policies are trained using Proximal Policy Optimization (PPO) [8]. The training incorporates reward shaping to encourage racket-centered impacts and sustained juggling, as well as domain randomization over ball position, coefficient of restitution, and other uncertain parameters to improve robustness and sim-to-real transfer. The policy

 Corresponding Authors. {yuchao, yu-wang}@tsinghua.edu.cn

<sup>1</sup> Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China.

<sup>2</sup> Electronic Engineering Department, Tsinghua University, Beijing 10083, China.

<sup>3</sup> Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100083, China.

This research was supported by National Natural Science Foundation of China (No.62406159, 62325405), Postdoctoral Fellowship Program of CPSF under Grant Number (GZC20240830, 2024M761676), China Postdoctoral Science Special Foundation 2024T170496.

outputs mid-level collective thrust and body rates (CTBR) commands, which are executed by a low-level PID controller on the quadrotor. Finally, the trained policy is deployed zero-shot on real hardware, where an enhanced perception module, incorporating a lightweight communication protocol, reduces latency in high-frequency state estimation, ensuring real-time control.

Extensive experiments validate the effectiveness of JuggleRL. In simulation, we compared JuggleRL with a traditional model-based predictive planner under varying ball drop heights. JuggleRL consistently maintained juggling performance close to the maximum episode length across all heights, whereas the baseline failed whenever the drop height fell below 4m. In real-world experiments, our policy achieved up to 462 consecutive hits, with an average of 311 hits over 10 trials, while the baseline reached at most 14 hits and averaged only 3.1 hits. Moreover, the learned policy generalizes to previously unseen ball weights, successfully juggling a lighter 5g ball with an average of 145.9 hits over 10 consecutive trials, demonstrating robust performance and adaptability to novel physical conditions.

Our contributions can be summarized as follows:

- We present the first model-free RL-based aerial ball-juggling system, JuggleRL, with zero-shot sim-to-real deployment.
- JuggleRL integrates a series of techniques, including system identification, soft-constraint reward shaping, domain randomization, and a lightweight communication protocol in the perception module to ensure robust real-time control on real hardware.
- JuggleRL achieves up to 462 hits with an average of 311 in real experiments under varying ball drop heights, far exceeding the baseline, which reaches at most 14 hits with an average of 3.1.
- JuggleRL could generalize to unseen ball weights, successfully juggling a lighter 5g ball with an average of 145.9 hits over 10 consecutive trials.

## II. RELATED WORK

### A. Motion Planning

Motion planning and control for autonomous quadrotors have traditionally followed two major paradigms: optimization-based methods and reinforcement learning (RL). Optimization-based methods decompose the problem into path planning and trajectory optimization, producing smooth and executable trajectories. This line of research has evolved from the seminal Minimum Snap trajectory generation [9] to advanced formulations such as Minimum Jerk planning [10], polynomial-based navigation in dense environments [11], and gradient-based approaches that avoid explicit signed-distance fields [2], with further extensions to decentralized swarm flight [12] and navigation [13].

In contrast, RL provides an end-to-end alternative, directly mapping sensory inputs to control commands. Policies are typically trained in high-fidelity simulators [14] and executed with high inference efficiency in the real world.

RL has enabled quadrotors to achieve impressive capabilities, including effective scheduling [15], stable hovering under disturbances [16], minimum-time navigation in cluttered environments [17], superhuman-level drone racing [18], and aggressive flight maneuvers [19], [20].

Despite their differences, both paradigms have largely concentrated on free-flight control, where the quadrotor operates without physical interaction with external objects. In this paper, we investigate a fundamentally different problem: aerial ball juggling. Unlike free-flight tasks, juggling requires sustained dynamic interaction between the quadrotor and a ball in motion. The vehicle must execute precise spatio-temporal strikes—arriving at specific positions with appropriate velocities and orientations—while continuously adapting to the ball’s real-time dynamics. These requirements—precise timing, rapid response, and robustness to sensing and communication delays—make aerial ball juggling far more demanding than conventional flight control, where traditional model-based methods achieve at most 14 consecutive hits [4], [6]. By contrast, our RL-based system succeeds in achieving up to 462 consecutive hits, demonstrating that reinforcement learning can enable robust, adaptive, and interactive quadrotor control.

### B. Aerial Ball-juggling

Previous work has explored quadrotor-based ball juggling primarily through prediction–planning frameworks. Müller et al. [4] proposed a seminal system that predicts ball states with a Kalman filter, models impact dynamics with a restitution-based rigid body model, and solves an open-loop interception trajectory for the quadrotor. Ritz et al. [5] improved the planning module by optimizing trajectories with a minimum-acceleration criterion, enabling smoother and more flyable paths for cooperative juggling with two quadrotors. Dong et al. [6] extended the framework by introducing minimum-jerk reference trajectories [21] and robust feedback control, addressing under-actuation and enhancing efficiency. Yu et al. [22] further refined the pipeline by employing an extended Kalman filter (EKF), cubic spline interpolation for trajectory fitting, and a two-stage joint optimization method for contact-aware planning.

Despite these advances, prediction–planning approaches face fundamental limitations. The ball’s motion exhibits strong nonlinearities (e.g., aerodynamic drag) that are difficult to model accurately. Optimization requires manually defined heuristics due to incomplete or uncertain state information. Pre-computed open-loop trajectories inevitably accumulate prediction errors. Collectively, these challenges constrain the robustness and success rate of existing juggling systems.

In contrast, our proposed RL-based system directly optimizes the final objective—the number of successful hits—without relying on predefined intermediate states. Rather than predicting long-term ball trajectories, our method only filters the instantaneous ball state as input. The learned policy updates quadrotor actions in real time, implicitly achieving adaptive motion planning and reducing reliance on

handcrafted models. This end-to-end approach enables stable, robust juggling under diverse and uncertain conditions.

### III. PRELIMINARY

#### A. Problem Formulation

We model the aerial ball-juggling task as a Markov Decision Process, characterized by  $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ , which defines the state space  $\mathcal{S}$ , action space  $\mathcal{A}$ , and observation space  $\mathcal{O}$ , the transition probability  $\mathcal{P}$ , reward function  $\mathcal{R}$ , and discount factor  $\gamma$ . The objective is to learn a stochastic policy  $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ , parameterized by  $\theta$ , that maps the current observation  $\mathbf{o}_t$  to a distribution over actions  $\mathbf{a}_t$ , with the aim of achieving *sustained ball juggling*. The parameters  $\theta$  are optimized to maximize the accumulative reward  $J(\theta) = \mathbb{E}_{\pi_\theta} [\sum_t \gamma^t R(\mathbf{s}_t, \mathbf{a}_t)]$ .

#### B. Quadrotor Dynamics

A quadrotor is modelled as a 6 degree-of-freedom rigid body of mass  $m$  and inertia matrix  $\mathbf{I} = [\mathbf{I}_x, \mathbf{I}_y, \mathbf{I}_z]$ . Its dynamics are modelled by:

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{\mathbf{p}}_{\mathcal{W}} \\ \dot{\mathbf{q}} \\ \dot{\mathbf{v}}_{\mathcal{W}} \\ \dot{\omega}_{\mathcal{B}} \\ \dot{\Omega} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_{\mathcal{W}} \\ \mathbf{q} \otimes [0, \omega_{\mathcal{B}}/2]^T \\ \frac{1}{m} \mathbf{q} \cdot \mathbf{f}_{\text{prop}} \cdot \bar{\mathbf{q}} + \mathbf{g}_{\mathcal{W}} \\ \mathbf{I}^{-1} (\tau_{\text{prop}} - \omega_{\mathcal{B}} \times (\mathbf{I} \omega_{\mathcal{B}})) \\ T_m (\Omega_{\text{cmd}} - \Omega) \end{bmatrix}, \quad (1)$$

where state  $\mathbf{x}$  consists of position  $\mathbf{p}$ , quaternion  $\mathbf{q}$ , linear velocity  $\mathbf{v}$ , angular velocity  $\omega$  and rotation speed of rotors  $\Omega$ . Subscripts  $\mathcal{W}$  and  $\mathcal{B}$  represent the world and body frame. Frame  $\mathcal{B}$  is located at the center of gravity. Notation  $\otimes$  indicates the multiplication of two quaternions.  $\bar{\mathbf{q}}$  denotes the quaternion's conjugate.  $\mathbf{g}_{\mathcal{W}} = [0, 0, -9.81m/s^2]^T$  denotes gravity.  $\mathbf{f}_{\text{prop}}$  and  $\tau_{\text{prop}}$  are the collective force and the torque produced by the propellers, defined as:

$$\mathbf{f}_{\text{prop}} = \sum_j \mathbf{f}_j, \tau_{\text{prop}} = \sum_j \tau_j + \mathbf{r}_{p,j} \times \mathbf{f}_j, \quad (2)$$

where  $\mathbf{r}_{p,j}$  is the location of propeller  $j$  in the body frame,  $\mathbf{f}_j, \tau_j$  the force and torque of propeller  $j$ . The dynamic of  $\Omega$  is modelled as a first-order system with a time constant  $T_m$ . We adopt a widely used model from prior work for the calculation of force and torque. Each motor's thrust and torque are modelled as proportional to the square of its rotation speed [18], determined by thrust coefficient  $k_f$  and drag coefficient  $k_m$ :

$$\mathbf{f}_j = [0, 0, k_f \Omega_j^2]^T, \tau_j = [0, 0, k_m \Omega_j^2]^T. \quad (3)$$

#### C. Task Setup

We study the problem of *aerial ball juggling*, where a quadrotor equipped with a badminton racket must repeatedly strike a ball to keep it aloft. The task objective is to maximize the number of consecutive successful hits before failure.

At each time step, the agent observes both the quadrotor state and the ball state, obtained from simulation or onboard estimation. A hit is considered successful only if two spatial constraints are satisfied: the contact height lies within a

prescribed interval, and the apex of the subsequent trajectory also falls within a valid range, namely

$$z_{\text{min}}^{\text{hit}} \leq z_{\text{hit}} \leq z_{\text{max}}^{\text{hit}}, \quad (4a)$$

$$z_{\text{min}}^{\text{apex}} \leq z_{\text{apex}} \leq z_{\text{max}}^{\text{apex}}. \quad (4b)$$

These constraints ensure that each strike produces a controllable rebound and that the ball remains within the workspace. The episode terminates once the ball touches the ground or leaves the workspace.

This setup imposes tight spatio-temporal requirements: the quadrotor must position itself precisely to satisfy the contact window (see Equ. (4a)) while ensuring the next apex remains feasible (see Equ. (4b)). Even small delays or misalignments quickly lead to failure, reflecting the difficulty of executing accurate strikes with an underactuated platform subject to strict timing constraints.

## IV. JUGGLERL SYSTEM

### A. Overview

The overview of the JuggleRL system is shown in Fig. 1. The framework begins with system identification (SysID) to calibrate the dynamics of both the quadrotor and the ball, ensuring consistency between the physical platform and its digital counterpart. We then construct a large-scale training environment, where policies are trained using PPO [8]. To improve robustness and facilitate sim-to-real transfer, reward shaping and domain randomization are incorporated into training. Finally, for real-world deployment, we design an enhanced perception module equipped with a lightweight communication protocol, which reduces latency in high-frequency state estimation and enables robust real-time control on physical quadrotors.

### B. System Identification

We precisely measure the key dynamic parameters of our quadrotor through SysID to construct a high-fidelity simulation model. As summarized in Tab. I, these parameters include the quadrotor's mass  $m$ , the inertia  $\mathbf{I} = [\mathbf{I}_x, \mathbf{I}_y, \mathbf{I}_z]$ , the motor thrust coefficient  $k_f$ , as well as the mass of the ball  $m_{\text{ball}}$  and the coefficient of restitution between the racket surface and the ball  $R_s$ . The mass  $m$  and inertia  $\mathbf{I}$  of the quadrotor, along with the mass of the ball  $m_{\text{ball}}$ , are measured directly using physical methods. The motor thrust coefficient  $k_f$  is estimated by first calculating the maximum thrust of a single motor from the hover throttle and the quadrotor's mass, and then normalizing the thrust by the square of the maximum motor angular velocity. To determine the coefficient of restitution  $R_s$  between the racket and the ball, we hover the quadrotor and racket stationary in the air and measure the rebound behavior of the ball dropped freely onto the racket at different contact points. A region is defined as the "sweet spot" if its coefficient of restitution remains above 0.75 throughout the area; otherwise, it is regarded as non-sweet. As shown in Fig. 2, the central circular region of the racket, with a radius of 5.5 cm, can be considered the "sweet spot," exhibiting an average coefficient of restitution of 0.82, while

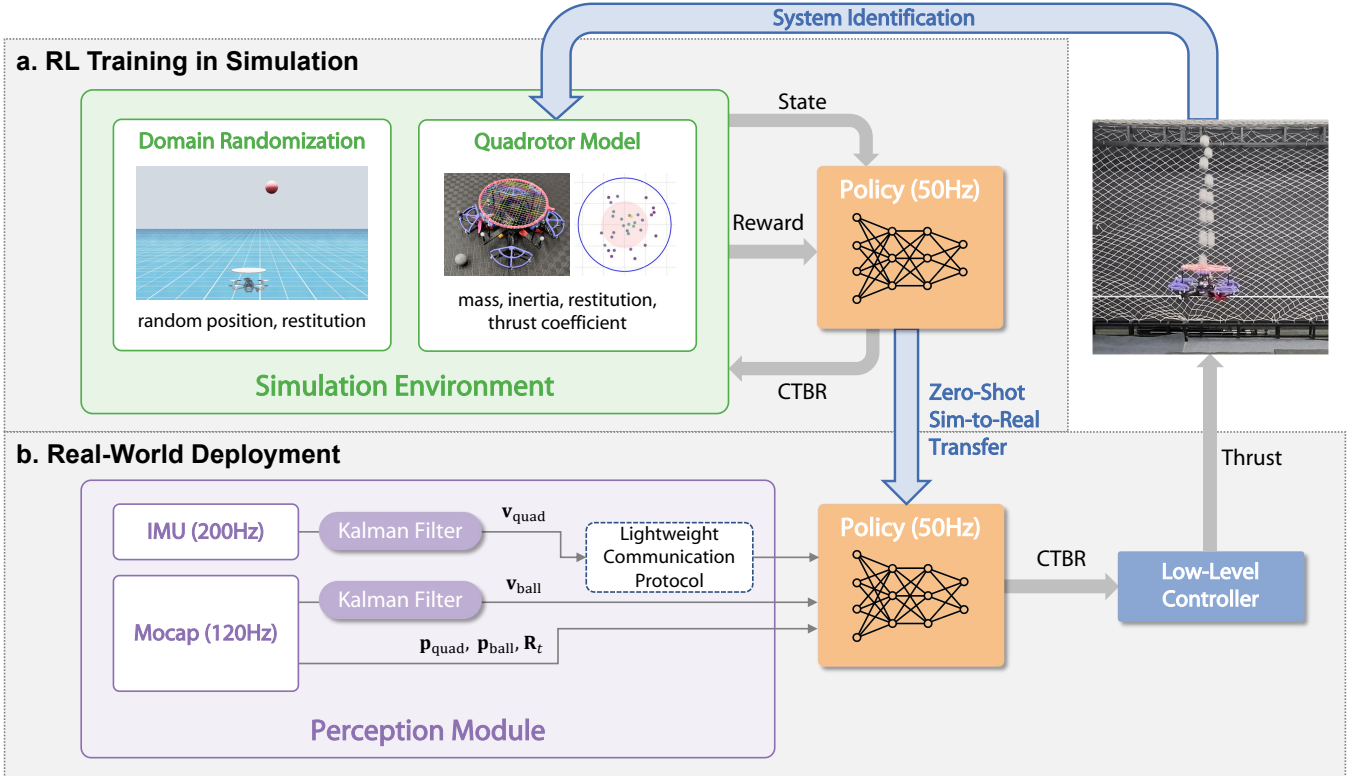


Fig. 1: Overview of the JuggleRL system. The system combines SysID-based dynamics calibration, large-scale GPU-parallel training in Isaac Sim with domain randomization to learn robust juggling policies. The learned policy outputs mid-level CTBR commands executed by a low-level PID controller, and is deployed zero-shot on real hardware with an enhanced perception module using a lightweight communication protocol to minimize latency in high-frequency state estimation.

TABLE I: Parameters of the quadrotor and the ball

Parameter	Value
$m$ (kg)	1.090
$I_x$ ( $kg \cdot m^2$ )	$[5.29 \times 10^{-3}, 0, 0]$
$I_y$ ( $kg \cdot m^2$ )	$[0, 5.62 \times 10^{-3}, 0]$
$I_z$ ( $kg \cdot m^2$ )	$[0, 0, 7.80 \times 10^{-3}]$
$k_f$ ( $kg \cdot m$ )	$5.39 \times 10^{-6}$
$m_{ball}$ (kg)	0.0472
$R_s$	0.82 (sweet), 0.64 (outer)

the outer regions have a lower average of approximately 0.64.

### C. Reinforcement Learning with Generalization Enhancement

We use PPO [8] as the training algorithm and provide a detailed design of the observation space, action space, reward function, and domain randomization below.

**Observation Space.** The quadrotor's input consists of two components: a state space observable by both the actor and critic networks, and privileged information accessible only to the critic. The 24-dimensional state space includes the quadrotor's absolute root state  $\mathbf{o}_{quad,t} = [\mathbf{p}_{quad,t}, \mathbf{R}_t, \mathbf{v}_{quad,t}] \in \mathbb{R}^{15}$ , where  $\mathbf{p}_{quad,t} \in \mathbb{R}^3$ ,  $\mathbf{v}_{quad,t} \in \mathbb{R}^3$  and  $\mathbf{R}_t \in \mathbb{R}^9$  denote the position, linear velocity and flattened rotation matrix of the quadrotor at time step  $t$ , respectively. In addition, the ball's observation vector is  $\mathbf{o}_{ball,t} = [\mathbf{p}_{ball,t}, \mathbf{v}_{ball,t}, \Delta \mathbf{p}] \in \mathbb{R}^9$ ,

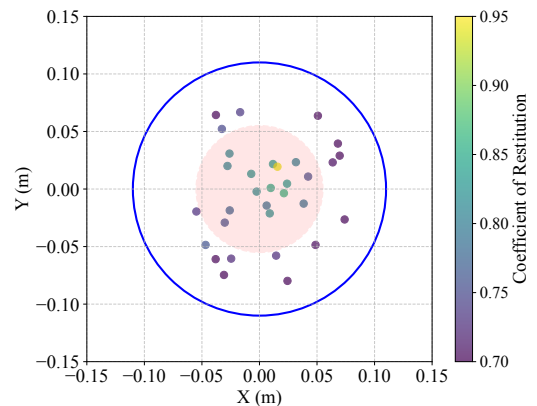


Fig. 2: Visualization of the restitution coefficient between the racket and the ball. The central "sweet spot" exhibits a higher restitution coefficient (average 0.82), while the periphery is less elastic (average 0.64), motivating our domain randomization strategy.

where  $\Delta \mathbf{p} \in \mathbb{R}^3$  denotes the relative position between the ball and the quadrotor. The privileged information for the critic is the current timestep  $t$  which helps the critic network capture temporal dynamics for more accurate value estimation.

**Action Space.** The action space is 4-dimensional, representing the collective thrust and body rates (CTBR). The action

at time step  $t$  is defined as  $\mathbf{a}_t = [c, \omega_x, \omega_y, \omega_z] \in \mathbb{R}^4$ , where  $c$  denotes the collective thrust controlling the quadrotor’s vertical acceleration, and  $(\omega_x, \omega_y, \omega_z)$  are the body rates around the roll, pitch, and yaw axes, respectively.

**Reward Function.** Our reward function combines sparse task rewards with several shaping terms. The core rewards are a hitting reward  $r^{\text{hit}}$  for successful strikes within a specified altitude window (Equ. (4a)) and an apex reward  $r^{\text{apex}}$  for reaching the desired apex height (Equ. (4b)). Shaping rewards include a dense positional term  $r^{\text{rpos}}$ , a center-contact reward  $r^{\text{contact}}$ , a horizontal position penalty  $r^{\text{xy}}$ , a smoothness reward  $r^{\text{smooth}}$ , and a spin reward  $r^{\text{spin}}$ .

The reward at time  $t$  is defined as:

$$r_t^{\text{hit}} = 50 \mathbb{1}_{\text{hit}}, \quad (5a)$$

$$r_t^{\text{apex}} = 50 \mathbb{1}_{\text{apex}}, \quad (5b)$$

$$r_t^{\text{rpos}} = \frac{1}{1 + \max(d_{xy}, 0.2)}, \quad (5c)$$

$$r_t^{\text{contact}} = 10 \exp(-5 d_{\text{axis}}) \mathbb{1}_{\text{hit}}, \quad (5d)$$

$$r_t^{\text{xy}} = -1.0 d_{\text{quad}}, \quad (5e)$$

$$r_t^{\text{smooth}} = 2.0 \exp(-\varepsilon_t), \quad (5f)$$

$$r_t^{\text{spin}} = -10 |\psi_t|, \quad (5g)$$

where  $d_{xy}$  is the horizontal distance between the quadrotor and the ball,  $d_{\text{axis}}$  is the distance of the contact point from the racket centerline,  $d_{\text{quad}}$  is the horizontal deviation of the quadrotor’s position from the origin,  $\psi_t$  is the yaw angle of the quadrotor, and  $\varepsilon_t$  is the action error defined as:

$$\varepsilon_t = \|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2, \quad (6)$$

where  $\mathbf{a}_t$  is the current action and  $\mathbf{a}_{t-1}$  is the previous action. The indicator functions  $\mathbb{1}_{\text{hit}}$  and  $\mathbb{1}_{\text{apex}}$  are triggered by the conditions in Equ. (4a)–(4b).

The overall reward is the sum of all components,

$$r_t = r_t^{\text{hit}} + r_t^{\text{apex}} + r_t^{\text{rpos}} + r_t^{\text{contact}} + r_t^{\text{xy}} + r_t^{\text{smooth}} + r_t^{\text{spin}}. \quad (7)$$

**Domain Randomization.** Tab. II summarizes our domain randomization settings including coefficient of restitution  $R_s$ , initial position of the quadrotor  $X_{\text{quad}0}$ ,  $Y_{\text{quad}0}$ ,  $Z_{\text{quad}0}$ , and initial position of the ball  $X_{\text{ball}0}$ ,  $Y_{\text{ball}0}$ ,  $Z_{\text{ball}0}$ . The height randomization aims to train a policy capable of stabilizing the ball from various initial heights, while the horizontal position randomization is employed to facilitate the policy in learning fast and precise actions to intercept balls with horizontal velocity. Simulating the racket’s central “sweet pot”, we randomize the coefficient of restitution around the average value 0.82 in the “sweet pot” and choose the range of [0.75, 0.90].

#### D. Real-world Deployment

For the real-world experiment, the reinforcement learning policy is zero-shot deployed onboard an NVIDIA Jetson Orin NX carried by the quadrotor platform. The policy takes both the quadrotor and ball states as input. The quadrotor poses are obtained from a motion capture system at 120 Hz, while velocities are estimated on the onboard computer by fusing

TABLE II: Range of domain randomization parameters for the quadrotor and ball.

Parameters	Range
$R_s$	[0.75, 0.90]
$Z_{\text{ball}0}$	[1.5, 2.0]
$Y_{\text{quad}0}$	[0.9, 1.1]
$X_{\text{ball}0}, X_{\text{quad}0}$	[-0.07, 0.07]
$Y_{\text{ball}0}, Y_{\text{quad}0}$	[-0.07, 0.07]

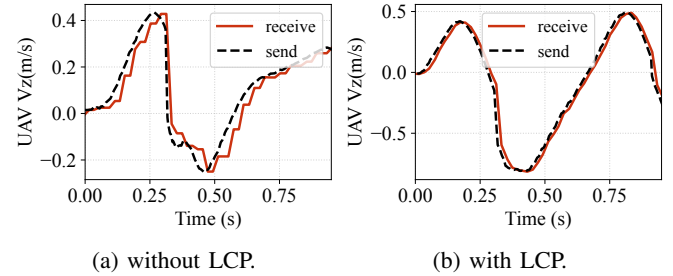


Fig. 3: Latency comparison with the lightweight communication protocol (LCP). Using the original communication scheme, the velocity signal exhibits step-like latency even at a 200 Hz publish rate. In contrast, LCP enables smooth, real-time transmission of velocity.

mocap and IMU measurements at 200 Hz via a Kalman filter. The ball is tracked as a reflective marker in the Mocap system, and its velocity is recovered through Kalman filtering of position measurements.

Because these high-frequency state estimates, especially the 200 Hz quadrotor velocities, incur heavy communication overhead and serialization delays, we design a lightweight communication protocol (LCP) that transmits only the essential motion states, i.e., the quadrotor’s pose and twist. This eliminates latency accumulation and ensures that the policy receives state updates in real time. Running at 50 Hz, the policy then outputs CTBR, which are executed by the onboard low-level flight controller.

## V. EXPERIMENTS

### A. Experiment Setting

We employ OmniDrones [14], a high-parallel UAV simulator, to construct the ball juggling task. The setup consists of a quadrotor (mass 1090 g) juggling a ball (mass 47.2 g, radius 2 cm) falling from above, using a racket with a 5.5 cm radius. Each episode is capped at 500 steps. Training is conducted with 4096 parallel environments running at 50 Hz, producing 2 billion simulation frames in under 3 hours on a desktop workstation (Intel i7-13700K CPU, RTX 4090 GPU).

### B. Evaluation Metric

We adopt the *number of consecutive hits* as the primary evaluation metric for juggling performance. In simulation, where conditions are idealized, the metric is defined as the maximum number of hits achieved within a fixed one-minute episode. Reported results are averaged over 100 episodes for

statistical reliability. In real-world experiments, the quadrotor continues juggling until failure, and we report the maximum number of hits over 10 consecutive trials as an indicator of robustness.

### C. Baseline

We adopt the open-source model-based juggling baseline from Müller et al. [4], referred to as the model-based predictive planner (MBPP). MBPP predicts the ball trajectory with a Kalman filter, generates an open-loop interception trajectory, and executes it through a trajectory tracking controller.

### D. Main Results

**Simulation Results.** We first evaluate the performance of JuggleRL against MBPP in simulation. Fig. 4 compares different initial ball release heights. MBPP requires a fixed target apex height (set to 3 m in our experiments) in order to compute a feasible trajectory, whereas JuggleRL operates within a target apex height range of [2.85, 3.2] m without such constraints. MBPP succeeds only when the ball is released between 4.0 and 4.5 m, and fails completely at lower heights, primarily because it relies on sufficient reaction time for the prediction–planning pipeline. In contrast, JuggleRL consistently maintains near-maximum episode performance (over 58 hits) across all tested heights, demonstrating its reactive capability and robustness.

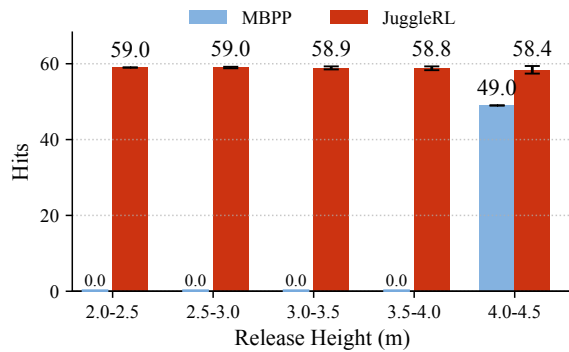


Fig. 4: Performance comparison between JuggleRL and MBPP in simulation. JuggleRL demonstrates robust performance across all tested ball release heights, while MBPP fails at lower release heights.

**Real-world Results.** We further evaluated JuggleRL in real-world experiments under two settings: fixed height, where the ball was released from 1.65 m, and varying height, where the release height was randomized between 1.5 m and 1.95 m. The target apex height range is [1.6, 1.85] m. Tab. III summarizes the results over 10 consecutive trials: under the fixed-height setting, JuggleRL achieved an average of 311 consecutive hits, with a maximum of 414 and a minimum of 165. Under the varying-height setting, it achieved an average of 403.6 hits, with a maximum of 462 and a minimum of 283. The numerical differences between the two settings are

TABLE III: Real-world juggling performance under fixed and varying release heights

Height	Hits of 10 consecutive trials	Avg
Fixed	301, 290, 325, 410, 401, 165, 217, <b>414</b> , 232, 355	311
Varying	454, 400, 460, 432, 460, 307, 355, 423, <b>462</b> , 283	403.6

attributed to experimental noise, and both clearly demonstrate JuggleRL’s stable control performance and robust policy.

Tab. IV compares JuggleRL with MBPP, where the MBPP results are taken from the original paper. It should be noted that the experimental setups differ: MBPP used higher toss heights to allow sufficient reaction time, whereas JuggleRL was evaluated without such constraints. JuggleRL significantly outperforms traditional methods, which achieved at most 14 hits as reported in both [4] and [6]. Videos of the juggling experiments are provided in the supplementary material.

TABLE IV: Comparison of real-world hits between JuggleRL and MBPP

Method	Average	Maximum
MBPP [4]	3.1	14.0
JuggleRL-Fixed	311.0	414.0
JuggleRL-Varying	<b>403.6</b>	<b>462.0</b>

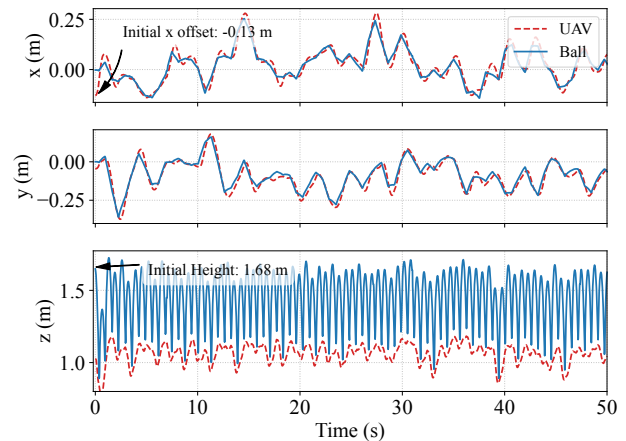


Fig. 5: Real-world juggling trajectory of a trial with a 0.13 m horizontal offset, ball releasing from 1.68 m.

**Behavior Analysis.** In this part, we provide a detailed analysis of a single juggling trial. Fig. 5 shows the  $xyz$  trajectories of the quadrotor and the ball in a trial where the ball was released from a height of 1.68 m. During hundreds of consecutive hits, the ball’s apex height remains consistently within the target range of 1.60–1.85 m. Whenever a strike drives the ball outside this range, the quadrotor exhibits immediate dynamic adjustments that gradually restore stability. Moreover, the average horizontal displacement of the ball from the origin is only 0.129 m. Considering the unavoidable deflections caused by the racket’s uneven surface [4], this highlights the strong closed-loop feedback regulation capability of JuggleRL.

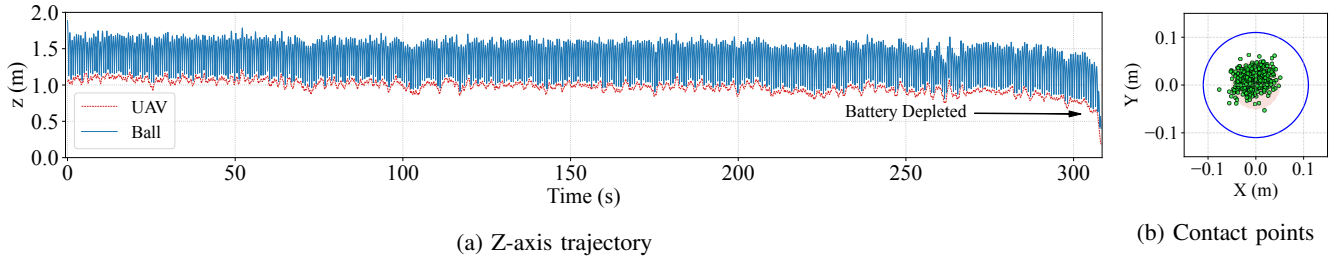


Fig. 6: Statistics of the longest trial (462 hits). (a) The gradual decrease in quadrotor height due to battery depletion. (b) Spatial distribution of contact points on the racket, with the “sweet spot” highlighted.

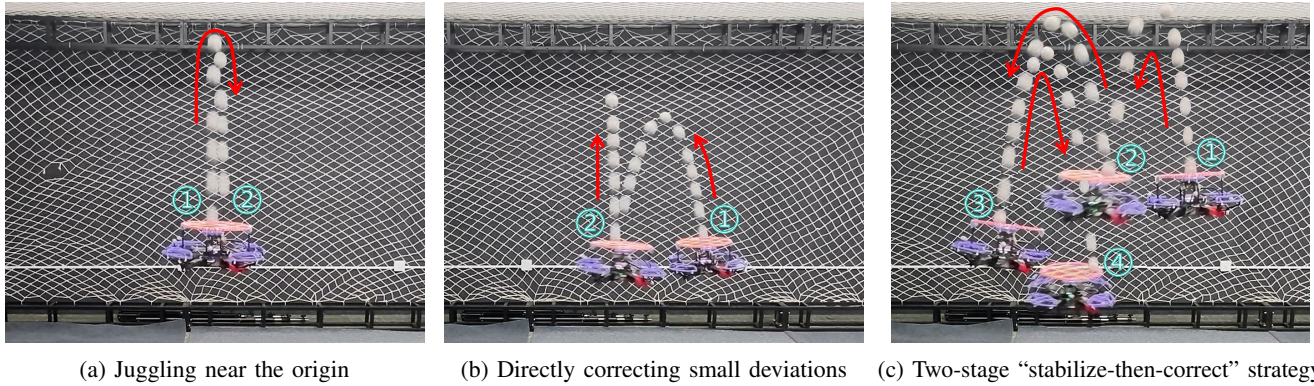


Fig. 7: Time-lapse visualization of JuggleRL’s adaptive strategies. (a) maintaining stable juggling near the origin; (b) directly correcting small deviations; and (c) adopting a two-stage “stabilize-then-correct” strategy when facing larger disturbances.

Fig. 6 presents the  $z$ -axis trajectory from the longest trial of 462 consecutive hits. In the latter part of this trial, the quadrotor gradually descends due to battery depletion, and the experiment ultimately terminates when the battery is exhausted. The spatial distribution of contact points reveals that the policy predominantly strikes within the racket’s “sweet spot,” which reduces rebound unpredictability and further demonstrates the stability and reliability of JuggleRL.

Fig. 7 illustrates the typical adaptive behaviors learned by the policy under different disturbance conditions: (a) maintaining stable juggling near the origin; (b) directly correcting small deviations; and (c) adopting a two-stage “stabilize-then-correct” strategy when facing larger disturbances. These behaviors suggest that JuggleRL does not rely on single-step optimal control. Instead, it achieves stability through gradual adjustments—a flexible control style particularly well suited for dynamic interaction tasks such as aerial juggling.

### E. Ablation Study

We conduct ablation studies on the key components of JuggleRL through real-world experiments with a fixed release height. Tab. V summarizes the results.

**Reward shaping.** To better understand the contribution of each reward component, we conduct ablation studies by removing them individually. Without the horizontal penalty (w/o  $r^{xy}$ ), the quadrotor gradually drifts away from the origin and cannot maintain a centered position. Excluding the yaw penalty (w/o  $r^{yaw}$ ) causes the quadrotor to spin aggressively, breaking flight stability. When the relative

position reward (w/o  $r^{pos}$ ) is omitted, the quadrotor fails to follow the ball smoothly, instead showing jittery and erratic movements. Removing the contact reward (w/o  $r^{contact}$ ) reduces the precision of strikes, leading to frequent off-center hits and more unpredictable rebounds. Finally, without the smoothness reward (w/o  $r^{smooth}$ ), the quadrotor issues unstable commands, resulting in visibly shaky flight. These results collectively demonstrate that each component plays a critical role in achieving stable and robust juggling behavior.

**Domain randomization.** Ablation studies on domain randomization highlight its critical role in robust juggling performance. When horizontal position randomization is removed (w/o H-Pos DR), the quadrotor struggles to compensate for misalignments, achieving only 0.5 hits on average. Omitting restitution randomization (w/o Rest. DR) causes the agent to strike too aggressively, lowering the average hit count to 6.8. Without height randomization (w/o Hgt. DR), the system cannot effectively adapt to variations in ball release elevation, resulting in unstable juggling and an average of 20.3 hits. These results underscore that each randomization component is essential for enabling reliable real-world performance.

**LCP.** Removing the lightweight communication protocol (w/o LCP) reduces the average number of consecutive hits from about 311 to 200.2, while the standard deviation increases noticeably, indicating that communication delays amplify policy decision fluctuations and significantly reduce juggling stability. LCP mitigates high-frequency state estimation latency, improving observation quality and thereby

ensuring more stable real-time control.

TABLE V: Ablation study of key design components.

Variants	Mean	Std Dev
JuggleRL	<b>311.0</b>	86.7
w/o $r^{xy}$	fail	/
w/o $r^{yaw}$	fail	/
w/o $r^{pos}$	58.7	52.2
w/o $r^{contact}$	55.3	56.0
w/o $r^{smooth}$	21.3	19.6
w/o H-Pos DR	0.5	0.6
w/o Rest. DR	6.8	5.2
w/o Hgt. DR	20.3	18.6
w/o LCP	200.2	186.8

TABLE VI: Real-world performance of JuggleRL on a lighter 5g ball. The policy successfully adapts to the new ball weight with an average of 145.9 hits across 10 consecutive trials.

Trial	1	2	3	4	5	6	7	8	9	10	Average
Hits	162	23	252	494	79	92	219	77	19	42	145.9

### F. Generalization

We further evaluate zero-shot transfer to an out-of-distribution setting: juggling a lighter 5g ball. As shown in Tab. VI, the policy achieves an average of 145.9 hits over 10 consecutive trials, with results ranging from 19 to 494. The high variance arises from the amplified sensitivity of the lighter ball to small control errors, yet the upper bound demonstrates that the learned strategy generalizes beyond the training distribution.

## VI. CONCLUSION AND FUTURE WORK

We propose JuggleRL, a quadrotor-based ball juggling system designed for dynamic physical interaction. To enable robust sim-to-real transfer, the system integrates (i) high-fidelity simulation through system identification, (ii) carefully designed reward shaping and domain randomization, and (iii) a lightweight communication protocol that mitigates observation latency. With this design, JuggleRL achieves up to 462 consecutive hits in real-world experiments, dramatically outperforming a model-based baseline that manages at most 14 hits. In addition, the system demonstrates out-of-distribution generalization, successfully juggling a lighter 5 g ball with an average of 145.9 hits, despite its higher sensitivity to control errors. Future work will extend the framework to multi-agent aerial volleyball.

### REFERENCES

[1] B. Zhao, J. Fang, Z. Dai, Z. Wang, J. Zha, W. Zhang, C. Gao, Y. Wang, J. Cui, X. Chen, *et al.*, "Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 32 400–32 423.

[2] X. Zhou, Z. Wang, H. Ye, C. Xu, and F. Gao, "Ego-planner: An esdf-free gradient-based local planner for quadrotors," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 478–485, 2020.

[3] H. Wang, J. Xu, X. Luo, X. Chen, T. Zhang, R. Duan, Y. Liu, and X. Chen, "Ultra-high-frequency harmony: mmwave radar and event camera orchestrate accurate drone landing," in *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, 2025, pp. 15–29.

[4] M. Müller, S. Lupashin, and R. D'Andrea, "Quadcopter ball juggling," in *2011 IEEE/RSJ international conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 5113–5120.

[5] R. Ritz, M. W. Müller, M. Hehn, and R. D'Andrea, "Cooperative quadcopter ball throwing and catching," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 4972–4978.

[6] W. Dong, G.-Y. Gu, Y. Ding, X. Zhu, and H. Ding, "Ball juggling with an under-actuated flying robot," in *2015 IEEE/RSJ IROS*, 2015, pp. 68–73.

[7] NVIDIA Corporation, "Nvidia isaac sim," Available: <https://developer.nvidia.com/isaac-sim>, accessed: Sep. 15, 2025.

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[9] D. Mellinger and V. Kumar, "Minimum snap trajectory generation and control for quadrotors," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 2520–2525.

[10] M. W. Mueller, M. Hehn, and R. D'Andrea, "A computationally efficient motion primitive for quadcopter trajectory generation," *IEEE transactions on robotics*, vol. 31, no. 6, pp. 1294–1310, 2015.

[11] C. Richter, A. Bry, and N. Roy, "Polynomial trajectory planning for aggressive quadrotor flight in dense indoor environments," in *Robotics Research: The 16th International Symposium ISRR*. Springer, 2016, pp. 649–666.

[12] X. Zhou, J. Zhu, H. Zhou, C. Xu, and F. Gao, "Ego-swarm: A fully autonomous and decentralized quadrotor swarm system in cluttered environments," in *2021 IEEE ICRA*, 2021, pp. 4101–4107.

[13] W. Zhang, C. Gao, S. Yu, R. Peng, B. Zhao, Q. Zhang, J. Cui, X. Chen, and Y. Li, "Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 31 292–31 309.

[14] B. Xu, F. Gao, C. Yu, R. Zhang, Y. Wu, and Y. Wang, "Omnidrones: An efficient and flexible platform for reinforcement learning in drone control," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2838–2844, 2024.

[15] Y. Xu, J. Zha, J. Ren, X. Jiang, H. Zhang, and X. Chen, "Scalable multi-agent reinforcement learning for effective uav scheduling in multi-hop emergency networks," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 2028–2033.

[16] J. Hwangbo, I. Sa, R. Siegwart, and M. Hutter, "Control of a quadrotor with reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2096–2103, 2017.

[17] R. Penicka, Y. Song, E. Kaufmann, and D. Scaramuzza, "Learning minimum-time flight in cluttered environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7209–7216, 2022.

[18] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, "Champion-level drone racing using deep reinforcement learning," *Nature*, vol. 620, no. 7976, pp. 982–987, 2023.

[19] J. Chen, C. Yu, Y. Xie, F. Gao, Y. Chen, S. Yu, W. Tang, S. Ji, M. Mu, Y. Wu, *et al.*, "What matters in learning a zero-shot sim-to-real rl policy for quadrotor control? a comprehensive study," *IEEE Robotics and Automation Letters*, 2025.

[20] Q. Sun, J. Fang, W. X. Zheng, and Y. Tang, "Aggressive quadrotor flight using curiosity-driven reinforcement learning," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 12, pp. 13 838–13 848, 2022.

[21] M. W. Mueller, M. Hehn, and R. D'Andrea, "A computationally efficient algorithm for state-to-state quadcopter trajectory generation and feasibility verification," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 3480–3486.

[22] H. Yu, J. Tu, P. Wang, Z. Zheng, K. Zhang, G. Lu, F. Gao, and J. Wang, "Bat planner: Aggressive flying ball player," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5307–5314, 2023.