Whole-Body Coordination for Dynamic Object Grasping with Legged Manipulators

Qiwei Liang[1], Boyang Cai[2], Rongyi He[2], Hui Li[2], Tao Teng[3], Haihan Duan[1], Changxin Huang[2], Runhao Zeng[1,*]

[1]Shenzhen MSU-BIT University
[2]Shenzhen University
[3]The Chinese University of Hong Kong
zengrh@smbu.edu.cn

Abstract

Quadrupedal robots with manipulators offer strong mobility and adaptability for grasping in unstructured, dynamic environments through coordinated whole-body control. However, existing research has predominantly focused on static-object grasping, neglecting the challenges posed by dynamic targets and thus limiting applicability in dynamic scenarios such as logistics sorting and human–robot collaboration. To address this, we introduce DQ-Bench, a new benchmark that systematically evaluates dynamic grasping across varying object motions, velocities, heights, object types, and terrain complexities, along with comprehensive evaluation metrics. Building upon this benchmark, we propose DQ-Net, a compact teacher–student framework designed to infer grasp configurations from limited perceptual cues. During training, the teacher network leverages privileged information to holistically model both the static geometric properties and dynamic motion characteristics of the target, and integrates a grasp fusion module to deliver robust guidance for motion planning. Concurrently, we design a lightweight student network that performs dual-viewpoint temporal modeling using only the target mask, depth map, and proprioceptive state, enabling closed-loop action outputs without reliance on privileged data. Extensive experiments on DQ-Bench demonstrate that DQ-Net achieves robust dynamic objects grasping across multiple task settings, substantially outperforming baseline methods in both success rate and responsiveness.

Introduction

Quadruped robots have emerged as a promising platform for mobile manipulation due to their superior mobility and terrain adaptability (Yang et al. 2022; Youm et al. 2023; Long et al. 2024; Sun et al. 2024; Mei et al. 2024; Zeng et al. 2024). With the integration of robotic arms, these systems are capable of performing manipulation tasks in complex environments (Bharadhwaj et al. 2024; Yokoyama et al. 2023; Qiu et al. 2024b; Sleiman et al. 2021; Ma et al. 2022; Mittal et al. 2022), demonstrating significant potential in applications such as search and rescue, logistics, and human-robot collaboration. By coordinating multi-joint movements of the limbs and arm, quadrupeds can realize whole-body control, thereby enhancing both their dynamic responsiveness and operational workspace (Pan et al. 2024; Portela et al. 2024a; Fu, Cheng, and Pathak 2022; Hao et al. 2021).

While prior research has made substantial progress in manipulation with quadruped systems, most efforts focus on grasping static objects (Liu et al. 2024; Wang et al. 2025; Zhang et al. 2024; Qiu et al. 2024a; Ha et al. 2024), assuming the target remains stationary. This simplification overlooks the critical challenges posed by dynamic object manipulation, where objects are in continuous motion—common in real-world scenarios such as conveyor-based logistics, interactive mobile targets, and human handovers (Xie et al. 2025; Fang et al. 2023). These dynamic settings demand rapid perception and agile control, making stable and efficient grasping under motion- one of the core challenges in quadruped robot research (Huang, Yu, and Jain 2023; Akinola et al. 2021).

To study this problem, we introduce the first benchmark tailored for whole-body **d**ynamic grasping with **q**uadruped robots: **DQ-Bench**. This benchmark provides a reproducible and comprehensive evaluation framework to assess algorithmic generalization and decision efficiency under dynamic and challenging conditions. DQ-Bench supports physically realistic target motion, non-planar terrain, and multi-level task difficulties. It includes diverse common daily objects partitioned into seen and unseen sets to rigorously test perception robustness and policy transferability. Evaluation metrics include Grasp Success Rate (GSR), One-Shot Success Rate (OSSR), and Timesteps to Completion (TSC), jointly measuring grasping effectiveness, decision quality, and responsiveness.

Upon the benchmark, we seek to find a solution to the core problem of dynamic grasping. Unlike static manipulation, the key challenge here lies in maintaining whole-body stability while precisely controlling the arm's end-effector in the presence of continuously changing relative motion. Even minor grasping pose deviations may lead to failure due to relative velocity between the robot and the target. Thus, accurately and efficiently predicting grasp poses at each time step from input signals is crucial yet highly challenging. A straightforward approach is to employ high-performance grasping networks to decode object geometry and predict optimal grasp poses (Ma and Huang 2023; Qin et al. 2023; Chen et al. 2023; Cai et al. 2022; Dai et al. 2022; Wen et al. 2022). However, this is problematic in dynamic tasks: (i) these networks often

*Corresponding author: zengrh@smbu.edu.cn

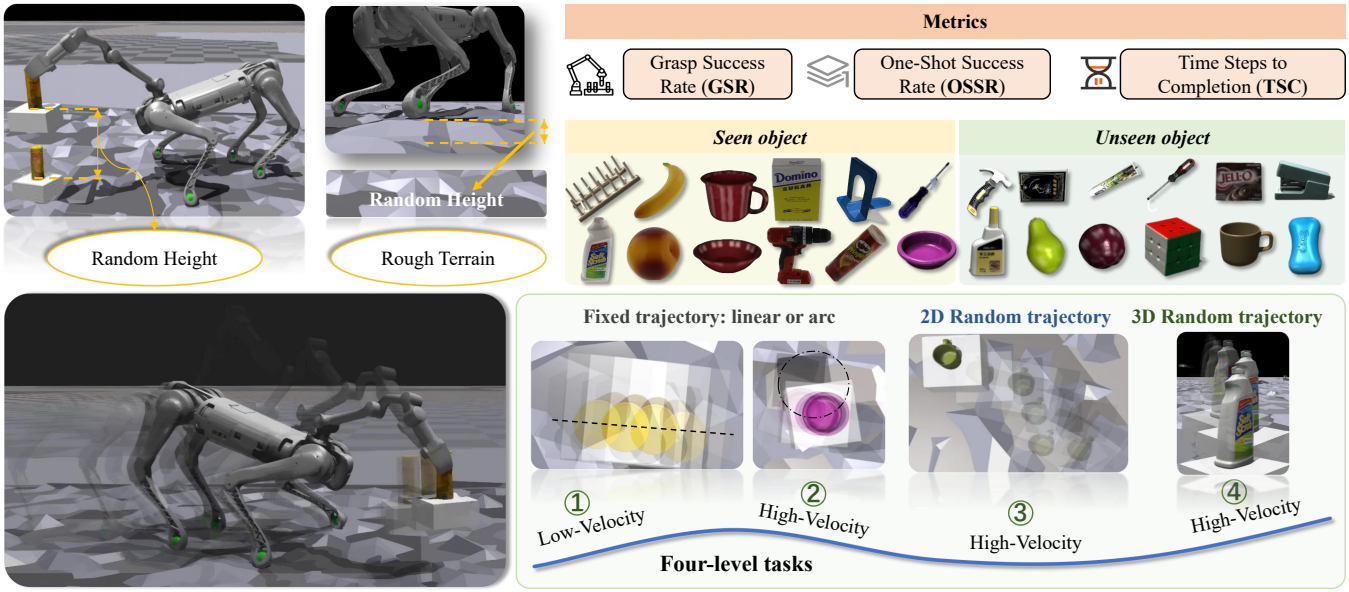

Figure 1: Overview of DQ-Bench: Our proposed benchmark provides a standardized and reproducible platform for evaluating dynamic object grasping with whole-body quadruped control. It systematically incorporates diverse grasping objects, multi-level task designs, and rigorous evaluation metrics to assess the adaptability, generalization, and efficiency of grasping strategies in dynamic scenarios.

rely on high-quality visual input, while quadrupeds typically observe targets from long distances with limited resolution; (ii) invoking such networks at every timestep incurs high computational costs, hindering real-time performance and training efficiency; (iii) conventional grasping networks output static grasp poses, lacking temporal consistency needed for continuous motion planning.

To overcome these limitations, we propose a framework for **d**ynamic object grasping with **q**uadruped robots, called **DQ-Net**. At the core of this framework is the **Grasp Fusion Module (GFM)**, designed to improve grasp quality and temporal consistency under motion. The GFM maintains a memory bank of multi-reference grasp poses, pre-generated in simulation based on object geometry and pose, and transforms these into a unified local coordinate system. At each timestep, the module constructs a query by using the object's 6D pose and encoded point cloud, then the query will be matched against the memory bank via attention mechanisms to produce a refined and robust grasp pose for control. This fused grasp pose, together with the object's motion state and robot proprioception, is fed into the policy head and trained via reinforcement learning to enable whole-body dynamic grasping. However, this pipeline relies on privileged information (object pose and velocity), which is challenging to obtain in practice due to the expense and measurement uncertainties of specialized sensors.

To bridge this gap, we further design a lightweight student policy network trained via knowledge distillation. It leverages only onboard sensory input, including dual-perspective visual observations (from the base and end-effector) and proprioceptive data. These inputs are separately encoded and fused using a dual-stream Transformer to capture both global semantic information and fine-grained object dynamics. The fused high-level features are decoded by the policy head to enable closed-loop dynamic grasping control. Extensive experiments on DQ-Bench demonstrate that DQ-Net significantly outperforms existing baselines across multiple dynamic tasks, achieving superior grasp success rates and faster response times. **Our contributions are summarized as follows:**

- We construct **DQ-Bench**, the first benchmark for dynamic object grasping with quadruped robots, supporting realistic dynamics, diverse objects, multi-level task difficulty, and comprehensive evaluation across perception and control.
- We propose **DQ-Net**, a framework combining a grasp memory-based fusion module and a lightweight student network relying solely on dual-perspective vision and proprioception for stable and efficient whole-body dynamic grasping.
- We conduct extensive evaluations across challenging dynamic tasks, where DQ-Net consistently outperforms prior methods in terms of grasp success and policy responsiveness.

Related Work

Whole-Body Control in Quadrupedal Manipulation

Recent work has advanced unified whole-body control for quadrupedal robots integrating manipulation (Portela et al. 2024b). Fu et al.(Fu, Cheng, and Pathak 2022) propose Deep Whole-Body Control, a reinforcement learning policy coordinating both arm and leg joints for agile behaviors like picking and button-pressing. Liu et al.(Liu et al. 2024) extend this with a vision-guided hierarchical policy that maps RGB-D

inputs to whole-body trajectories, showing success on real hardware. However, these methods assume static targets and execute single-shot grasps without considering object motion. More recent approaches like QuadWBG (Wang et al. 2025) and GAMMA (Zhang et al. 2024) integrate grasp-aware planning into quadruped systems but still assume fixed objects during inference. Overall, existing whole-body control strategies excel under static assumptions but lack support for continuous target motion and dynamic tracking.

Benchmarks for Robotic Grasping

Large-scale datasets like GraspNet-1Billion (Fang et al. 2020) and TARGO (Xia et al. 2024) have enabled robust learning and evaluation of grasping models, yet both focus on fixed-base arms interacting with static objects. Efforts like DGBench (Burgess-Limerick et al. 2022) and GAP-RL (Xie et al. 2025) target dynamic grasping by introducing moving objects and reactive grasp policies, but are still limited to tabletop settings with fixed manipulators and constrained motion. General benchmarks like ManiSkill (Mu et al. 2021) and BEHAVIOR (Li et al. 2022) emphasize task diversity and mobile manipulation, yet lack support for dynamic grasping and do not involve locomotion-manipulation coupling. In contrast, our DQ-Bench introduces a physically realistic evaluation suite that combines non-planar terrain, 6-DoF target motion, and whole-body quadruped grasping, filling a critical gap in existing benchmarks.

DQ-Bench: Dynamic Grasping Benchmark for Quadruped Robots

Despite significant progress in static grasping tasks with quadruped robots, real-world applications often involve continuously moving targets that require real-time perception and grasping decisions. Currently, there is a lack of a standardized evaluation platform to systematically assess quadruped performance in dynamic grasping scenarios. To address this gap, we propose the first benchmark—**DQ-Bench**—for dynamic object grasping with full-body quadruped control. This benchmark provides a comprehensive and reproducible platform to evaluate the generalization and decision-making efficiency of various algorithms in complex dynamic environments. Our design systematically considers key factors from four dimensions—environment modeling, object selection, task difficulty levels, and evaluation metrics—ensuring high-fidelity modeling and rigorous assessment of dynamic grasping tasks. The overview of DQ-Bench is shown in Figure 1.

Environment Setup

To ensure physical realism and support parallel training, we build the evaluation environment on the physics-based simulation platform *Isaac Gym* (Makoviychuk et al. 2021). Objects are placed on a movable floating platform that follows predefined or randomized trajectories in 3D space. Compared to directly assigning random motion to objects, this method better reflects real-world object dynamics under external forces, avoiding unnatural motion that could distort grasping strategies. Additionally, we introduce uneven terrain to test the robot's adaptability and whole-body coordination across varied ground conditions.

Objects for Grasping

We select representative objects from the YCB dataset (Calli et al. 2015) that vary in shape, size, and weight to ensure task diversity and challenge. To test generalization, objects are split into two groups: *seen* (used in both training and testing) and *unseen* (used only during testing). This setup introduces real-world unpredictability and demands greater robustness and generalization from the grasping strategies.

Multi-Level Task Design

To comprehensively evaluate the quadruped's capability in dynamic grasping, we design four progressively challenging task levels based on object speed, trajectory complexity, terrain conditions, and spatial degrees of freedom:

- **Level 1:** Object moves at low speed ($0 \sim 15\,\text{cm/s}$) along fixed trajectories (linear or arc).
- **Level 2:** Object moves at high speed ($15 \sim 30\,\text{cm/s}$) along fixed trajectories.
- **Level 3:** Object moves along random trajectories at high speed ($0 \sim 30\,\text{cm/s}$).
- **Level 4:** Object moves freely along the z-axis, resulting in random 3D trajectories.

These four levels span from 2D to 3D motion, flat to complex terrains, and low to high speeds. At each episode reset, the object's initial height is randomly sampled within $[0.2, 0.7]$ meters, and the terrain is randomly varied within a flat range of 0–0.1 meters.

Evaluation Metrics

To quantitatively assess grasping performance across task levels, we introduce three core metrics:

- **Grasp Success Rate (GSR):** The proportion of successfully completed grasps, indicating the overall effectiveness of the strategy.
- **One-Shot Success Rate (OSSR):** The percentage of successful grasps completed in a single attempt without any re-adjustment, reflecting decisiveness.
- **Time Steps to Completion (TSC):** The number of time steps taken from task initiation to successful grasp, representing grasping efficiency.

These metrics jointly evaluate grasping strategies from the perspectives of success rate, decision quality, and response efficiency, providing insights into behavioral differences and performance bottlenecks in dynamic grasping scenarios.

Proposed Method

Overview

We aim to develop a lightweight control system for quadruped robots that achieves whole-body coordination and dynamic grasping using only accessible perceptual inputs, such as depth maps and segmentation masks. The control framework

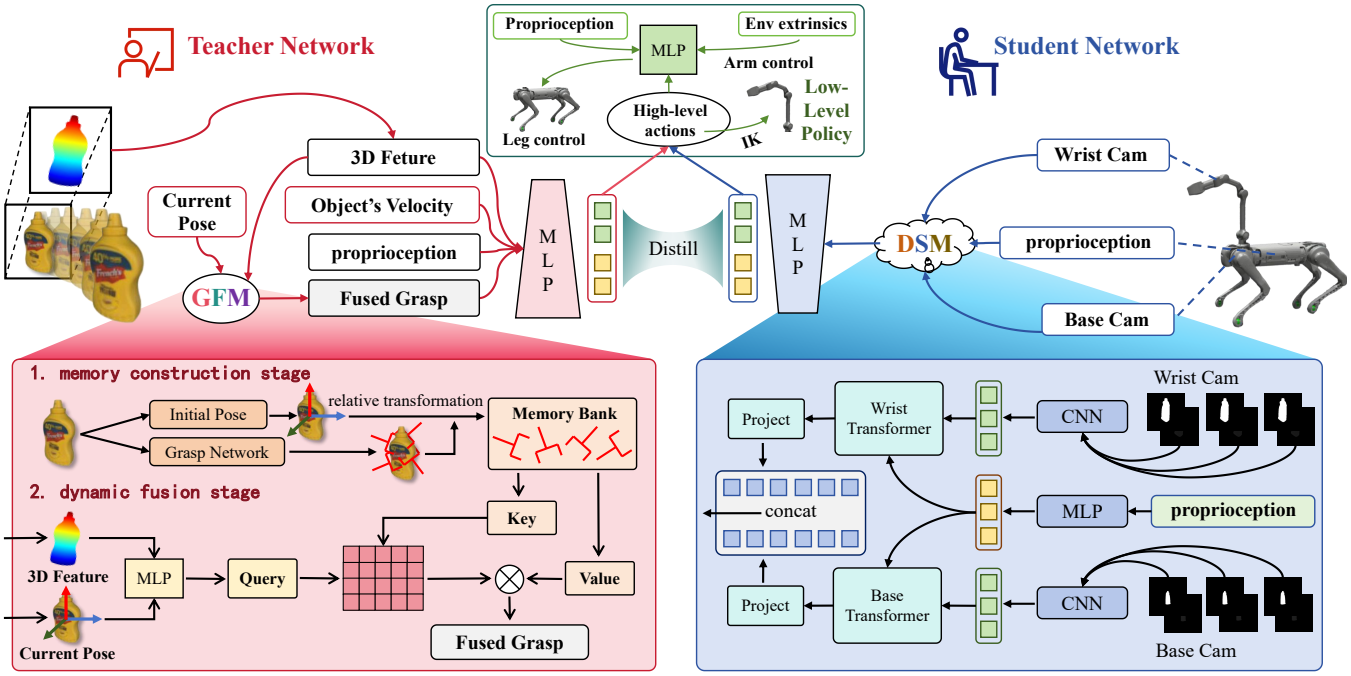

Figure 2: Overview of DQ-Net: We prepose a teacher-student framework for quadruped robots to perform dynamic object grasping. The teacher network takes point cloud features, object motion, robot proprioception, and a grasp representation from a Grasp Fusion Module (GFM), which builds a grasp memory and fuses object features with current pose via attention. The student network uses a dual-stream architecture to encode three-frame sequences of target masks and depth maps from wrist and base cameras. Features from both streams are fused with proprioception, processed by Transformers, and decoded into high-level actions. These high-level actions actions are further mapped to low-level control for end-to-end loco-manipulation.

comprises: **1)** a low-level policy (LLP) trained via reinforcement learning to produce full-body control signals from target velocity and end-effector poses; and **2)** a high-level policy (HLP) that maps perceptual observations to high-level commands for the LLP.

However, training a vision-only policy via reinforcement learning is challenging due to the task's complexity and the limited informativeness of raw visual inputs. To address this, we adopt a teacher-student distillation framework (Fig. 2). The teacher is trained in simulation with privileged inputs (e.g., object pose and velocity) to learn a high-quality policy, while the student learns to imitate the teacher using only depth and segmentation maps.

While the teacher benefits from rich inputs, dynamic grasping still requires accurate and timely grasp pose estimation due to object movement. Traditional grasp pose estimation methods often incur high computational cost, suffer from latency, and rely on static visual inputs, which limits their effectiveness in long-range prediction and grasp pose refinement during approach.

In this paper, we introduce the Grasp Fusion Module (GFM) that dynamically selects grasp poses in real time. GFM maintains a memory bank of multi-directional grasp candidates and predicts the optimal grasp pose $\mathbf{g}_t$ using attention mechanisms, conditioned on point cloud features $\mathbf{f}_p$ and object pose $\mathbf{x}_t$. Then, $\mathbf{g}_t$ is fed into the policy network together with the object pose $\mathbf{x}_t$ and velocity $\mathbf{v}_t$, point cloud features $\mathbf{f}_t$, and robot proprioception $\mathbf{b}_t$ to generate the high-level action:

$$\mathbf{a}_t = \pi\left(\mathbf{f}_p, \mathbf{x}_t, \mathbf{v}_t, \mathbf{g}_t, \mathbf{b}_t\right). \tag{1}$$

Recognizing the limitations of obtaining complete point clouds or accurate pose estimates in deployments, we design a student policy that relies only on depth and segmentation images from wrist and base-mounted cameras. A Transformer-based visual encoder captures temporal dynamics from multi-frame inputs $(\mathcal{I}_t^{\text{arm}}, \mathcal{I}_t^{\text{base}})$, and fuses them with proprioception $\mathbf{b}_t$ to generate the final action:

$$\mathbf{a}_t^{\text{stu}} = \pi^{\text{stu}}\left(\mathcal{I}_t^{\text{arm}}, \mathcal{I}_t^{\text{base}}, \mathbf{b}_t\right). \tag{2}$$

Lastly, the student is trained to imitate the teacher by minimizing the mean squared error between their action outputs:

$$\mathcal{L}_{\text{KD}} = \frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{a}_t^{\text{stu}} - \mathbf{a}_t^{\text{tea}}\right\|_2^2. \tag{3}$$

This enables robust and real-time control in perception-limited scenarios, ensuring reliable dynamic grasping with minimal sensing requirements.

High-Level Teacher Policy Network with Grasp Fusion Module

In the simulation environment, the teacher network utilizes privileged information of the target object, including ground-truth pose $\mathbf{x}_t$ and velocity $\mathbf{v}_t$, to learn robust grasping and motion control policies. To tackle dynamic object grasping, we

propose the Grasp Fusion Module (GFM) before the policy network, which has two stages: grasp memory construction and dynamic fusion.

In the **grasp memory construction stage**, under a fixed camera pose, a pretrained grasp pose prediction network (e.g., Contact-GraspNet (Sundermeyer et al. 2021)) generates N grasp candidates $\{\mathbf{G}_i \in \mathbb{SE}(6)\}_{i=1}^{N}$ in a single forward pass. The top-K candidates with highest scores are stored in a memory bank as relative transformations $\tilde{\mathbf{G}}_i$ in the object's local frame, reducing online computation.

In the **dynamic fusion stage**, the object point cloud is encoded by pretrained PointNet (Qi et al. 2017) into a global feature vector $\mathbf{f}_\text{p}$. At each time step t, the current object pose $\mathbf{x}_t$ and $\mathbf{f}_\text{p}$ form a query vector via an MLP:

$$\mathbf{q}_t = \text{MLP}_q\left(\mathbf{f}_\text{p};\ \mathbf{x}_t\right). \tag{4}$$

Relative transformations $\{\tilde{\mathbf{G}}_i\}$ are converted to world-frame grasps $\{\mathbf{G}_i^t\}$ using $\mathbf{x}_t$. Each $\mathbf{G}_i^t$ is flattened and mapped to key-value pairs:

$$\mathbf{k}_i^t = \text{MLP}_k(\text{vec}(\mathbf{G}_i^t)), \quad \mathbf{v}_i^t = \text{MLP}_v(\text{vec}(\mathbf{G}_i^t)). \tag{5}$$

An attention mechanism weights the values to produce the optimal grasp representation:

$$\alpha_{i,t} = \frac{\exp(\mathbf{q}_t^\top \mathbf{k}_i^t)}{\sum_{j=1}^{K} \exp(\mathbf{q}_t^\top \mathbf{k}_j^t)}, \quad \mathbf{g}_t = \sum_{i=1}^{K} \alpha_{i,t} \mathbf{v}_i^t. \tag{6}$$

Finally, the fused grasp $\mathbf{g}_t$, proprioception $\mathbf{b}_t$, point cloud feature $\mathbf{f}_\text{p}$, object pose $\mathbf{x}_t$, and velocity $\mathbf{v}_t$ are input to an MLP to generate high-level actions.

High-Level Student Policy Network: Temporal Modeling with Dual-View Fusion

We design a lightweight student policy network that takes only camera inputs. To ensure accurate control in dynamic grasping, we propose a Dual-Stream visual Modeling (DSM) structure, which fuses observations from a wrist-mounted and a base-mounted camera. This allows the policy to capture both local details and global spatial layouts.

Each view provides three consecutive frames, each containing a target mask $\mathbf{m}_t$ (from a pretrained Track-SAM (Cheng et al. 2023)) and a depth map $\mathbf{d}_t$. Compared to RGB, mask maps reduce the sim-to-real domain gap, while depth maps encode the 3D geometry between the camera and the target. The inputs are denoted as $\mathcal{I}_\text{arm}$ and $\mathcal{I}_\text{base}$.

Both input streams are passed through a shared CNN encoder CNN_θ to extract feature sequences $\mathbf{F}^a$ and $\mathbf{F}^b$. Meanwhile, the robot's proprioceptive state is encoded by an MLP into an embedding $\mathbf{e}_b$, which is concatenated with each visual feature, forming fused tokens $\tilde{\mathbf{f}}_t^a$ and $\tilde{\mathbf{f}}_t^b$.

Temporal encodings are added, and the sequences are processed by two separate Transformer encoders (Trans_a, Trans_b) to model time-aware attention from both views. The resulting outputs are projected via a linear layer and concatenated:

$$\mathbf{z} = [\text{Proj}(\text{Trans}_a(\tilde{\mathbf{f}}^a));\ \text{Proj}(\text{Trans}_b(\tilde{\mathbf{f}}^b))],$$

which is then fed into an action head to generate $\hat{\mathbf{a}}_t$.

Low-Level Policy For Locomotion

The high-level policy outputs the end-effector position increment $\Delta\hat{\mathbf{p}} \in \mathbb{R}^3$, orientation increment $\Delta\hat{\mathbf{r}} \in \mathbb{R}^3$, as well as base linear velocity v_lin and yaw rate ω_yaw, which need to be further converted into executable low-level control signals.

The low-level controller first integrates these increments into target end-effector states $(\hat{\mathbf{p}}, \hat{\mathbf{r}})$, and combines them with base motion commands to form a low-level command vector:

$$\mathbf{u}_t = [\hat{\mathbf{p}}, \hat{\mathbf{r}}, v_\text{lin}, \omega_\text{yaw}] \in \mathbb{R}^8. \tag{7}$$

This command vector, together with the robot's proprioceptive state $\mathbf{b}_t$ and the terrain embedding $\mathbf{z}_t$, is fed into a multi-layer perceptron π_low to output target joint angles for the legs:

$$\mathbf{q}_t^* = \pi_\text{low}(\mathbf{b}_t, \mathbf{u}_t, \mathbf{z}_t). \tag{8}$$

The quadruped executes actions through PD control. For the manipulator, inverse kinematics is used to compute joint angles $\boldsymbol{\theta}_\text{arm}^*$ from the target end-effector state $(\hat{\mathbf{p}}_\text{arm}, \hat{\mathbf{r}}_\text{arm})$. The final action output is:

$$\mathbf{a}_{low} = [\mathbf{q}_t^*, \boldsymbol{\theta}_\text{arm}^*]. \tag{9}$$

Training Details

To efficiently train hierarchical control policies, we use a staged training approach. First, the low-level policy is trained with PPO (Schulman et al. 2017) and then frozen. Next, a high-level teacher policy is trained via reinforcement learning using a modified reward based on static grasping. This includes a yaw-angle penalty: when the absolute yaw angle $|\psi|$ exceeds $60°$, a quadratic penalty is applied; if it exceeds $70°$, the episode terminates early to improve sampling efficiency. After freezing the teacher, a student policy is trained with DAgger (Ross, Gordon, and Bagnell 2011) by minimizing the MSE between student actions and teacher labels.

Experiment

Environments

We adopt **DQ-Bench** as the unified evaluation platform. Each method is executed for 5,000 steps per task level to ensure statistically stable results. All methods are trained under the Level 4 setting and are evaluated on Levels 1 through 4. To better simulate real-world conditions, the visual observations are delayed by four frames to mimic perception latency.

Compared Methods

VBC (Liu et al. 2024). A strong baseline for quadruped grasping static object via whole-body control.

VBC-D. A modified version of VBC with a reward function and angle constraints adapted for dynamic object grasping, aligned with our settings for fair comparison.

DQ-Net w/o GFM. A variant of DQ-Net where the GFM is removed from the teacher policy.

DQ-Net w/o vel. DQ-Net where object velocity is excluded from the teacher policy.

Method	Level 1			Level 2			Level 3			Level 4		
	GSR-T	GSR-S	OSSR	GSR-T	GSR-S	OSSR	GSR-T	GSR-S	OSSR	GSR-T	GSR-S	OSSR
VBC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VBC-D	57.5	28.6	27.4	57.4	29.0	27.6	48.7	19.1	17.8	42.7	16.0	15.2
DQ-Net w/o GFM	65.2	49.8	44.7	65.2	49.7	44.8	60.0	36.0	31.3	55.0	33.3	28.9
DQ-Net w/o vel	78.5	47.4	46.9	78.6	46.9	46.4	74.0	38.4	37.6	70.3	34.8	34.0
DQ-Net	**80.8**	**55.8**	**53.2**	**80.8**	**55.6**	**53.2**	**77.9**	**44.8**	**41.8**	**74.3**	**41.0**	**38.5**

Table 1: Grasp Success Rate (GSR) and One-Shot Success Rate (OSSR) (%) across four levels of increasing difficulty. Here, GSR-T and GSR-S denote the teacher and student strategies, respectively. DQ-Net consistently outperforms all baselines.

Method	L1	L2	L3	L4	Params (M)
CNN-Based	51.61	51.44	40.13	36.14	8.43
Ours	**55.82**	**55.63**	**44.87**	**41.04**	**5.37**

Table 2: Grasp Success Rate (GSR) and model size comparison between student policies. Our Transformer-based method outperforms the CNN-based baseline across all difficulty levels with fewer parameters.

Implementation Details

All training and evaluation are performed on a single NVIDIA RTX 4090 GPU using the Unitree B1 quadruped robot with a Unitree Z1 robot arm. The low-level control policy is shared across all methods to ensure consistent control capability. The high-level teacher policy is trained in 6,000 parallel environments with a rollout length of 24, totaling 80,000 timesteps. The high-level student policy is trained in 200 parallel environments for the same number of timesteps. We set the number of top-ranked grasp candidates in the GFM module to $K = 30$ throughout all experiments. More details are put in the supplementary material.

Experimental Results

Grasping Performance under Varying Motion Complexity. We evaluate all methods across four difficulty levels using Grasp Success Rate (GSR). As shown in Table 1, the DQ-Net student policy consistently achieves the highest GSR, outperforming VBC-D by 25% at Level 4. Ablation variants slightly degrade performance but still surpass VBC-D at all levels. VBC-D, despite dynamic adaptation, suffers from limited motion modeling, while static VBC almost completely fails. Overall, DQ-Net remains robust under increasing motion complexity, with both GFM and velocity input contributing to effective dynamic grasp prediction.

Success Rate on One-Time Grasping. Unlike static object grasping with multiple adjustment opportunities, dynamic grasping often permits only one attempt due to object motion. Thus, the one-shot grasp success rate (OSSR) is a key metric for evaluating legged robots in dynamic scenarios. To assess each method's decision efficiency, we compare their OSSR across five scenarios in Table 1. DQ-Net consistently outperforms all baselines, demonstrating superior real-time decision-making. Even with a single-attempt constraint, it achieves strong performance across all difficulty levels.

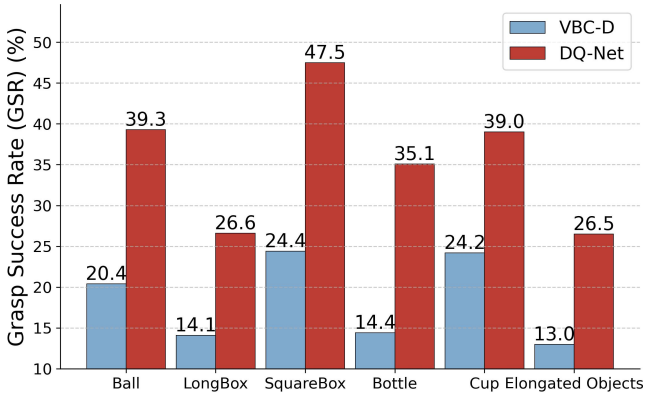

Figure 3: Grasp success rates on unseen objects under Level 4 difficulty.

Method	Level 1	Level 2	Level 3	Level 4
VBC-D	43.71	43.86	41.20	41.56
DQ-Net w/o GFM	36.20	35.35	36.91	35.43
DQ-Net w/o vel	37.01	37.13	35.22	35.10
DQ-Net	**35.45**	**35.24**	**35.27**	**34.32**

Table 3: Average timesteps to successful completion (TSC) across four difficulty levels. Lower values indicate faster and more efficient grasping.

Timesteps Comparisons. To assess execution efficiency in dynamic grasping, we measure the average time steps per grasp (TSC). As shown in Table 3, DQ-Net and its ablated variants require significantly fewer steps than VBC-D. The three DQ-Net variants show similar TSC across all levels, indicating that both GFM and velocity input enhance object dynamics understanding and action speed. Without these cues, the policy becomes more cautious and less efficient.

Generalization to Unseen Objects. We further test on unseen YCB objects (ball, long box, square box, bottle, cup, elongated objects) held out from training, all under Level 4. As shown in Figure 3, DQ-Net consistently surpasses VBC-D in grasp success rate across all categories, demonstrating strong generalization under challenging dynamics. Full per-object results are provided in the appendix.

Our approach vs. CNN-based policy for student strat-

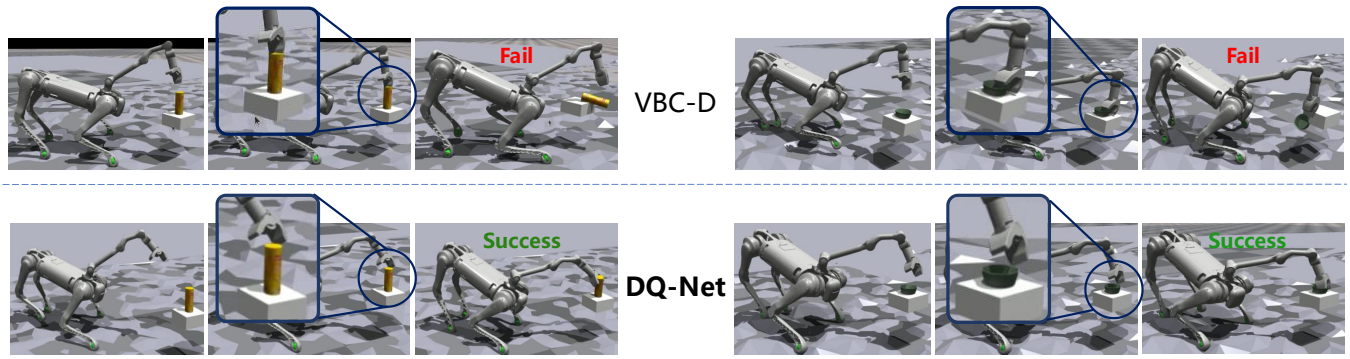

Figure 4: End-effector poses before contact in dynamic scenes. Both methods approach the object, but only DQ-Net achieves precise alignment, enabled by grasp priors from GFM. VBC-D suffers from misalignment, often leading to failure.

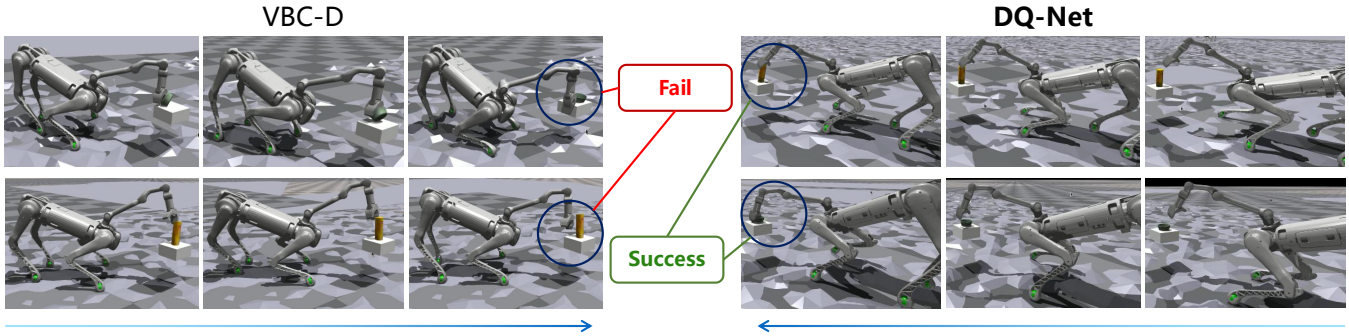

Figure 5: Comparison of motion anticipation. DQ-Net predicts future object trajectories and adjusts accordingly, while VBC-D reacts with delay. GFM and velocity cues enable DQ-Net's predictive and adaptive grasping.

egy. We design a lightweight Transformer-based student policy that captures temporal dependencies and fine-grained visual cues via a dual-view strategy. Compared to the CNN-based student in VBC (Liu et al. 2024), our model achieves consistently higher grasp success rates (Table 2) with only 5.37M parameters versus 8.43M. The performance gain stems from separate modeling of camera inputs and the temporal modeling strength of the Transformer backbone.

Qualitative Results

To better understand the effectiveness of our method in dynamic object grasping, we visualize and compare the behavioral differences between different policies. In such tasks, both the grasping pose of the end-effector and its spatial alignment with the object's motion trajectory are crucial to grasp success. We present a qualitative comparison between VBC-D and DQ-Net in representative scenarios.

Better Grasping Pose. Figure 4 illustrates the end-effector poses just before contact. While both methods bring the arm near the object, only DQ-Net aligns the gripper precisely with the target, whereas VBC-D suffers from noticeable misalignment, often leading to grasp failure. We attribute this to the GFM, which enables DQ-Net to dynamically integrate grasp priors from memory to predict a more suitable pose.

More Predictive Pose Estimation. Even with a theoretically correct static grasp pose, successful grasping may still fail in dynamic scenarios if the robotic arm cannot properly adapt to the object's continuous motion. As shown in Figure 5, VBC-D exhibits delayed reactions and lacks motion anticipation, whereas DQ-Net predicts the object's future trajectory and proactively adjusts its gait and body posture. This predictive behavior fundamentally stems from DQ-Net's enhanced ability to model object motion dynamics in real time, effectively leveraging velocity cues and the GFM's memory retrieval mechanism to generate grasp poses that are optimally aligned with anticipated future object positions.

Conclusion

We introduced DQ-Bench, the first benchmark for dynamic object grasping with quadruped robots, featuring realistic target motion, challenging terrains, multi-level difficulty, and diverse objects. Building on this, we proposed DQ-Net, a unified whole-body grasping framework that combines a memory-driven Grasp Fusion Module with a lightweight student policy using dual-view vision and proprioception to fuse spatiotemporal information from both robot and environment. Experiments show that DQ-Net consistently surpasses strong baselines across difficulty levels and object categories, indicating robust generalization and responsiveness. In future work, we will deploy DQ-Net on real quadruped platforms, and extend it to deformable or articulated objects, multi-object settings, and collaborative manipulation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62202311 and 62403325; the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011512; the Key Scientific Research Project of the Department of Education of Guangdong Province under Grant 2024ZDZX3012; the Shenzhen Science and Technology Foundation under Grant JCYJ20250604173210013; the Shenzhen Natural Science Foundation (Stable Support Plan Program) under Grant 20231122104038002; and the Key Field Projects of Ordinary Universities in Guangdong Province under Grant 2025ZDZX3050.

References

Akinola, I.; Xu, J.; Song, S.; and Allen, P. K. 2021. Dynamic grasping with reachability and motion awareness. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9422–9429.

Bharadhwaj, H.; Mottaghi, R.; Gupta, A.; and Tulsiani, S. 2024. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, 306–324.

Burgess-Limerick, B.; Lehnert, C.; Leitner, J.; and Corke, P. 2022. DGBench: An Open-Source, Reproducible Benchmark for Dynamic Grasping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022*, 3218–3224.

Cai, J.; Su, J.; Zhou, Z.; Cheng, H.; Chen, Q.; and Wang, M. Y. 2022. Volumetric-based Contact Point Detection for 7-DoF Grasping. In *Conference on Robot Learning (CoRL)*.

Calli, B.; Walsman, A.; Singh, A.; Srinivasa, S.; Abbeel, P.; and Dollar, A. M. 2015. Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set. *IEEE Robotics and Automation Magazine*, 22(3): 36–52.

Chen, Y.; Lin, Y.; Xu, R.; and Vela, P. A. 2023. Keypointgraspnet: Keypoint-based 6-dof grasp generation from the monocular rgb-d input. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 7988–7995. IEEE.

Cheng, Y.; Li, L.; Xu, Y.; Li, X.; Yang, Z.; Wang, W.; and Yang, Y. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558*.

Dai, Q.; Zhang, J.; Li, Q.; Wu, T.; Dong, H.; Liu, Z.; Tan, P.; and Wang, H. 2022. Domain Randomization-Enhanced Depth Simulation and Restoration for Perceiving and Grasping Specular and Transparent Objects. In *European Conference on Computer Vision (ECCV)*.

Fang, H.; Wang, C.; Gou, M.; and Lu, C. 2020. GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 11441–11450.

Fang, H.-S.; Wang, C.; Fang, H.; Gou, M.; Liu, J.; Yan, H.; Liu, W.; Xie, Y.; and Lu, C. 2023. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5): 3929–3945.

Fu, Z.; Cheng, X.; and Pathak, D. 2022. Deep Whole-Body Control: Learning a Unified Policy for Manipulation and Locomotion. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, 138–149.

Ha, H.; Gao, Y.; Fu, Z.; Tan, J.; and Song, S. 2024. UMI on Legs: Making Manipulation Policies Mobile with Manipulation-Centric Whole-body Controllers. In *Proceedings of the 2024 Conference on Robot Learning*.

Hao, J.; Yuan, Y.; Wang, C.; and Wang, Z. 2021. ED2: Environment Dynamics Decomposition World Models for Continuous Control. *arXiv preprint arXiv:2112.02817*.

Huang, B.; Yu, J.; and Jain, S. 2023. EARL: Eye-on-hand reinforcement learner for dynamic grasping with active pose estimation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2963–2970.

Li, C.; Zhang, R.; Wong, J.; Gokmen, C.; Srivastava, S.; Martín-Martín, R.; Wang, C.; Levine, G.; Lingelbach, M.; Sun, J.; Anvari, M.; Hwang, M.; Sharma, M.; Aydin, A.; Bansal, D.; Hunter, S.; Kim, K.; Lou, A.; Matthews, C. R.; Villa-Renteria, I.; Tang, J. H.; Tang, C.; Xia, F.; Savarese, S.; Gweon, H.; Liu, C. K.; Wu, J.; and Fei-Fei, L. 2022. BEHAVIOR-1K: A Benchmark for Embodied AI with 1, 000 Everyday Activities and Realistic Simulation. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, 80–93.

Liu, M.; Chen, Z.; Cheng, X.; Ji, Y.; Qiu, R.; Yang, R.; and Wang, X. 2024. Visual Whole-Body Control for Legged Loco-Manipulation. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, 234–257.

Long, J.; Wang, Z.; Li, Q.; Cao, L.; Gao, J.; and Pang, J. 2024. Hybrid Internal Model: Learning Agile Legged Locomotion with Simulated Robot Response. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Ma, H.; and Huang, D. 2023. Towards scale balanced 6-dof grasp detection in cluttered scenes. In *Conference on robot learning*, 2004–2013.

Ma, Y.; Farshidian, F.; Miki, T.; Lee, J.; and Hutter, M. 2022. Combining learning-based locomotion policy with model-based manipulation for legged mobile manipulators. *IEEE Robotics and Automation Letters*, 7(2): 2377–2384.

Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A.; et al. 2021. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*.

Mei, Y.; Wang, Y.; Zheng, S.; and Jin, Q. 2024. Quadrupedgpt: Towards a versatile quadruped agent in open-ended worlds. *arXiv preprint arXiv:2406.16578*.

Mittal, M.; Hoeller, D.; Farshidian, F.; Hutter, M.; and Garg, A. 2022. Articulated object interaction in unknown scenes with whole-body mobile manipulation. In *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 1647–1654. IEEE.

Mu, T.; Ling, Z.; Xiang, F.; Yang, D.; Li, X.; Tao, S.; Huang, Z.; Jia, Z.; and Su, H. 2021. ManiSkill: Generalizable Manipulation Skill Benchmark with Large-Scale Demonstrations.

In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Pan, G.; Ben, Q.; Yuan, Z.; Jiang, G.; Ji, Y.; Li, S.; Pang, J.; Liu, H.; and Xu, H. 2024. RoboDuet: Whole-body Legged Loco-Manipulation with Cross-Embodiment Deployment. *arXiv preprint arXiv:2403.17367*.

Portela, T.; Cramariuc, A.; Mittal, M.; and Hutter, M. 2024a. Whole-body end-effector pose tracking. *arXiv preprint arXiv:2409.16048*.

Portela, T.; Margolis, G. B.; Ji, Y.; and Agrawal, P. 2024b. Learning force control for legged manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 15366–15372.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qin, R.; Ma, H.; Gao, B.; and Huang, D. 2023. RGB-D grasp detection via depth guided learning with cross-modal attention. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8003–8009.

Qiu, R.-Z.; Hu, Y.; Song, Y.; Yang, G.; Fu, Y.; Ye, J.; Mu, J.; Yang, R.; Atanasov, N.; Scherer, S.; et al. 2024a. Learning generalizable feature fields for mobile manipulation. *arXiv preprint arXiv:2403.07563*.

Qiu, R.-Z.; Song, Y.; Peng, X.; Suryadevara, S. A.; Yang, G.; Liu, M.; Ji, M.; Jia, C.; Yang, R.; Zou, X.; et al. 2024b. WildLMa: Long Horizon Loco-Manipulation in the Wild. *arXiv preprint arXiv:2411.15131*.

Ross, S.; Gordon, G. J.; and Bagnell, D. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, 627–635.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Sleiman, J.-P.; Farshidian, F.; Minniti, M. V.; and Hutter, M. 2021. A unified mpc framework for whole-body dynamic locomotion and manipulation. *IEEE Robotics and Automation Letters*, 6(3): 4688–4695.

Sun, J.; Zhou, L.; Geng, B.; Zhang, Y.; and Li, Y. 2024. Leg state estimation for quadruped robot by using probabilistic model with proprioceptive feedback. *IEEE/ASME transactions on mechatronics*.

Sundermeyer, M.; Mousavian, A.; Triebel, R.; and Fox, D. 2021. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 13438–13444.

Wang, J.; Rajabov, J.; Xu, C.; Zheng, Y.; and Wang, H. 2025. QuadWBG: Generalizable Quadrupedal Whole-Body Grasping. arXiv:2411.06782.

Wen, H.; Yan, J.; Peng, W.; and Sun, Y. 2022. TransGrasp: Grasp Pose Estimation of a Category of Objects by Transferring Grasps from Only One Labeled Instance. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Xia, Y.; Ding, R.; Qin, Z.; Zhan, G.; Zhou, K.; Yang, L.; Dong, H.; and Cremers, D. 2024. TARGO: Benchmarking Target-driven Object Grasping under Occlusions. arXiv:2407.06168.

Xie, P.; Chen, S.; Chen, Q.; Tang, W.; Hu, D.; Dai, Y.; Chen, R.; and Wang, G. 2025. GAP-RL: Grasps as Points for RL Towards Dynamic Object Grasping. *IEEE Robotics Autom. Lett.*, 10(1): 40–47.

Yang, R.; Zhang, M.; Hansen, N.; Xu, H.; and Wang, X. 2022. Learning Vision-Guided Quadrupedal Locomotion End-to-End with Cross-Modal Transformers. In *International Conference on Learning Representations*.

Yokoyama, N.; Clegg, A.; Truong, J.; Undersander, E.; Yang, T.-Y.; Arnaud, S.; Ha, S.; Batra, D.; and Rai, A. 2023. Asc: Adaptive skill coordination for robotic mobile manipulation. *IEEE Robotics and Automation Letters*, 9(1): 779–786.

Youm, D.; Jung, H.; Kim, H.; Hwangbo, J.; Park, H.-W.; and Ha, S. 2023. Imitating and finetuning model predictive control for robust and symmetric quadrupedal locomotion. *IEEE Robotics and Automation Letters*, 8(11): 7799–7806.

Zeng, R.; Zhou, D.; Liang, Q.; Liu, J.; Li, H.; Huang, C.; Li, J.; Hu, X.; and Sun, F. 2024. Video2Reward: Generating Reward Function from Videos for Legged Robot Behavior Learning. *arXiv preprint arXiv:2412.05515*.

Zhang, J.; Gireesh, N.; Wang, J.; Fang, X.; Xu, C.; Chen, W.; Dai, L.; and Wang, H. 2024. GAMMA: Graspability-Aware Mobile MAnipulation Policy Learning based on Online Grasping Pose Fusion. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, 1399–1405.