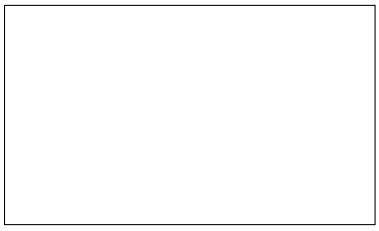
Graphical Abstract

CoPickVLM: A Vision-Language Model Guided Dual-Arm Collaborative System for Occlusion-Aware Tomato Harvesting

Yan Luo,Zhenhua Xiong,Han Ding



Highlights

CoPickVLM: A Vision-Language Model Guided Dual-Arm Collaborative System for Occlusion-Aware Tomato Harvesting

Yan Luo,Zhenhua Xiong,Han Ding

- Vision-language model
- Occlusion-aware recognition
- Agricultural automation

CoPickVLM: A Vision-Language Model Guided Dual-Arm Collaborative System for Occlusion-Aware Tomato Harvesting⋆

Yan Luo[a], Zhenhua Xiong[a,*] and Han Ding[a,b]

[a]State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b]State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Keywords:
Occlusion-aware recognition
Tomato harvesting
Dual-arm robot
Vision-language model
Collaborative manipulation
Agricultural automation

ABSTRACT

This paper presents **CoPickVLM**, an occlusion-aware tomato recognition and dual-arm collaborative harvesting system based on vision-language model (VLM) reasoning, designed to address automated harvesting challenges in greenhouse environments with complex leaf occlusion. The system centers on leveraging VLMs for task-level semantic planning, integrated with occlusion-aware recognition and dual-arm coordination strategies. To address occlusion-induced visual ambiguity, we adopt a vision-language reasoning approach that integrates multi-image prompting and attention-based localization, combined with color-based leaf segmentation for grasp point grounding. The harvesting task is decomposed into three semantic subtasks—occlusion identification, 2D-to-3D spatial localization, and vision-language reasoning for action planning—each mapped to structured motion primitives for execution. A dual-arm control scheme is implemented, assigning the auxiliary arm to gently remove occlusions while the primary arm executes precise cutting to ensure fruit integrity. Experimental results show that the proposed system achieves a tomato harvesting success rate of 86.9% in unoccluded scenes, 78.2% under partial occlusion, and 58.7% under full occlusion, showing significant improvements over conventional methods. These results validate the effectiveness of vision-language reasoning and dual-arm collaboration in enabling robust harvesting in visually cluttered agricultural scenarios.

1. Introduction

The rapid advancement of agricultural automation is reshaping traditional farming practices, driven by global population growth, escalating labor shortages, and increasing demands for efficiency and sustainability. Among the various branches of smart agriculture, robotic harvesting systems have emerged as a critical solution to address labor constraints while enhancing productivity and operational consistency [1, 2]. In particular, fruit and vegetable harvesting—traditionally reliant on intensive manual labor—stands to benefit significantly from robotic interventions, which can reduce labor costs, improve picking precision, and contribute to the modernization of agricultural production [3, 4].

Tomatoes, as one of the world's most widely cultivated and economically significant crops, present a series of unique technical challenges for automation. Manual tomato picking is not only labor-intensive and inefficient but also increasingly unsustainable in the face of dwindling rural labor availability [5, 6]. The harvesting process is complicated by factors such as non-uniform fruit maturity, complex plant structures, and dense foliage, all of which demand high levels of perception, dexterity, and adaptability from automated systems [7, 8]. These challenges underscore the urgent need for intelligent and robust robotic systems capable of operating effectively in unstructured agricultural environments.

A primary obstacle in automated tomato harvesting is the frequent and severe occlusion of fruit by leaves and stems. In such conditions, conventional computer vision techniques often fail to maintain adequate recognition accuracy, leading to high rates of missed detections or false positives [9, 10]. Furthermore, occlusions complicate the physical interaction required for successful picking, increasing the likelihood of mechanical failure or crop damage. As such, enhancing the system's perception and reasoning ability under occlusion has become a central research focus in agricultural robotics [11, 12].

In this context, traditional single-arm robotic systems face inherent limitations. Although structurally simple and cost-effective, single-arm manipulators struggle to concurrently execute complex, multi-step tasks such as navigating occlusions, adjusting viewpoints, and manipulating objects. Their limited workspace and operational flexibility further constrain their applicability in dense crop layouts, where dynamic and adaptive behaviors are often required [13, 14, 15].

To address these shortcomings, dual-arm collaborative systems have garnered increasing attention. By distributing functional roles—such as assigning one arm for auxiliary tasks like occlusion removal, and the other for precision harvesting—dual-arm robots can achieve higher task efficiency, environmental adaptability, and motion flexibility [16, 17]. This bimanual approach allows for more human-like, coordinated interactions with complex agricultural scenes, ultimately improving the success rate

⋆This work was supported by Jiangsu Provincial Science and Technology Department Program Special Funds (Basic Research on Frontier Leading Technologies) Project SBK2023050162.
*Corresponding author
**Principal corresponding author
luoyan99@sjtu.edu.cn (Y. Luo); mexiong@sjtu.edu.cn (Z. Xiong); hding@sjtu.edu.cn (H. Ding)
ORCID(s):

Figure 1: Execution sequence: The right arm removes occluding leaves, allowing the left arm to complete the harvest. This dual-arm coordination is essential under heavy occlusion.

and quality of automated harvesting operations [18, 19].

Simultaneously, the emergence of large language models (LLMs) and multimodal foundational models has introduced a paradigm shift in robotic intelligence. These models exhibit strong capabilities in integrating visual inputs with semantic knowledge, enabling context-aware perception and high-level reasoning across diverse domains [20, 21]. In agriculture, such capabilities can be leveraged to overcome long-standing limitations of traditional algorithms by enabling robots to understand occluded scenes, infer task-relevant actions, and coordinate multi-arm behaviors based on natural language or structured prompts [22, 23, 24].

In this work, we propose an occlusion-aware dual-arm tomato harvesting system guided by a vision-language model-based reasoning framework. The system integrates vision-language understanding for robust tomato detection under occlusion and employs a collaborative manipulation strategy in which one robotic arm lifts obstructing foliage while the other executes the picking action. This synergistic approach combines the cognitive advantages of vision-language model (VLM) with the mechanical versatility of dual-arm robotics to address the core challenges of automated tomato harvesting in complex field environments.

The main contributions of this paper are as follows:

1. We propose **CoPickVLM**, a novel **VLM-guided** dual-arm collaborative tomato harvesting framework that integrates vision-language reasoning with semantic-aware action planning, enabling high-level understanding and execution of complex harvesting tasks under occlusion.
2. We design an **occlusion-aware tomato recognition algorithm** that leverages multimodal fusion of RGB, depth, and semantic cues to accurately perceive ripe targets and their occlusion states in dense agricultural environments.
3. We implement and deploy the proposed system on a **wheeled mobile humanoid robot**, validating its effectiveness through real-world greenhouse experiments under various occlusion scenarios.

The remainder of the paper is organized as follows: Section 2 reviews related work; Section 3 describes the problem formulation and overall system architecture; Section 4 details the CoPickVLM-based reasoning methodology; Section 5 presents experimental validation and analysis; Section 6 concludes the paper and outlines future directions.

2. Related Work

Automated tomato harvesting is a highly integrated task that involves the convergence of multiple advanced technologies to address the challenges posed by complex and unstructured agricultural environments. Specifically, effective tomato harvesting systems require advancements in four key technical domains: **occlusion-aware perception and recognition**, which enables accurate detection of target fruits under dense foliage; **robotic harvesting systems**, which ensure reliable and gentle fruit handling; **dual-arm collaborative manipulation**, which enhances dexterity and adaptability in constrained spaces; and **multimodal vision-language reasoning frameworks**, which provide high-level task understanding and decision-making capabilities. The integration of these technologies is essential for achieving robust, efficient, and intelligent harvesting in real-world greenhouse conditions.

2.1. Occlusion-Aware Perception and Recognition

The task of tomato recognition in greenhouse environments has evolved from traditional vision methods to more robust deep learning approaches [25], driven largely by the challenge of occlusion. Early techniques based on color segmentation and geometric features [26, 27, 28] suffered from poor adaptability to lighting changes and cluttered foliage.

With the rise of CNN-based detectors, models like the improved YOLOv4-tiny model [29], YOLO-tomato [30], and S-YOLO [31] achieved significant improvements in both accuracy and real-time performance. More recent methods have turned to multimodal fusion, integrating RGB, depth, and near-infrared (NIR) data to enhance robustness under severe occlusion [32, 33]. NeRF-based 3D reconstruction [34] further introduced geometric reasoning into occlusion-aware perception, enabling accurate fruit localization and volume estimation even in dense foliage.

Despite progress in detection methods, current systems often lack semantic understanding required for robotic harvesting, such as reasoning about occlusions and manipulation strategies. Vision-language models have shown promise by enabling contextual scene comprehension and zero-shot inference [35, 36]. However, the capability of vision-language models to understand occlusion-rich and complex multi-modal scenes remains limited, particularly in terms of integration with robotic manipulation and motion control [37]. Enhancing this understanding and coordination is essential for advancing autonomous dual-arm harvesting systems.

2.2. Robotic Harvesting Systems

Robotic harvesting systems have witnessed substantial progress across various crop types, including sweet peppers [38, 39], eggplants [40], strawberries [41], cherries [42], tea leaves [43], and small spherical fruits [44]. These systems often integrate six-degree-of-freedom collaborative robotic arms mounted on mobile platforms such as unmanned ground vehicles or rail-guided systems to navigate structured agricultural environments. Advances in mechanical design and control, such as those discussed by Droukas et al. [45] and Mail et al. [46], have enabled greater adaptability to field conditions and diversified crop geometries. Furthermore, recent surveys have outlined a range of end-effector technologies tailored to specific fruit morphologies and detachment mechanisms, highlighting the increasing sophistication of harvesting tools [47].

Tomato harvesting, as a representative case of soft, occlusion-prone crops, has received focused attention in recent studies. Lili et al. [48] developed a tomato harvesting robot specifically designed for greenhouse conditions, incorporating a vision-guided system and a picking arm. Wang et al. [49] proposed an adaptive end-effector pose control scheme to enhance grasping precision, while Xie et al. [50] optimized a bionic compliant gripper to improve fruit protection during harvesting. Similarly, Rong et al. [51] introduced a selective harvesting robot for cherry tomatoes with a redesigned mechanical structure and field-tested performance metrics. These works demonstrate significant progress in both the actuation systems and the design of harvesting grippers, enabling more reliable and efficient fruit detachment in structured environments.

Despite these advancements, several limitations persist. Most current systems rely on a single robotic manipulator for the entire harvesting process, which proves inadequate in scenarios involving occlusions or clustered fruits. Such configurations struggle with leaf obstructions that impair visual perception and grasp planning. Additionally, the efficiency of single-arm systems is inherently constrained by the sequential nature of perception and manipulation tasks, limiting throughput in real-world deployments [45, 47].

2.3. Dual-Arm Collaborative Manipulation

Dual-arm collaborative manipulation has become increasingly important in robotic systems that require coordinated motion planning, stable grasping, and complex object handling. Traditional research has largely focused on model-based approaches, such as hybrid position/force control and coordinated trajectory planning [16, 52]. These methods rely on accurate kinematic and dynamic modeling to achieve synchronized control of both arms. In agricultural applications, dual-arm configurations have demonstrated advantages in handling complex spatial constraints and improving manipulation efficiency. For instance, Li et al. [53] developed a multi-arm system for apple harvesting that integrates perception and task planning. Lammers et al. [54] and Sepúlveda et al. [55] presented dual-arm solutions for apple and aubergine harvesting, respectively, showing improved task stability in structured settings. However, these systems are still limited in dealing with challenges such as heavy occlusion or unstructured foliage, which require more adaptive perception and reasoning.

Recent studies have turned to data-driven methods to enhance the flexibility and adaptability of dual-arm manipulation. Reinforcement learning (RL) and imitation learning have been applied to improve control policies in uncertain and variable environments. Cui et al. [56] proposed a task-adaptive RL framework that allows dual-arm robots to adjust manipulation strategies based on task context. In the context of tomato harvesting, Li et al. [57] introduced a dual-arm motion planning strategy using deep RL to manage navigation and manipulation efficiency in greenhouse settings. While these methods represent significant progress, the problem of dynamic occlusion handling during manipulation remains under-addressed. Other studies have explored cooperative human-robot manipulation [58, 59] and integrated visual feedback for better coordination. A systematic review by Abbas et al. [60] summarized key developments in dual-arm modeling, planning, control, and vision strategies, highlighting ongoing challenges in adapting to unstructured environments.

Looking forward, the integration of foundation models and multimodal reasoning is poised to significantly enhance dual-arm robotic capabilities. Vision-language-action frameworks such as $\pi 0$ [61] and Mobility-VLA [62] enable robots to interpret complex instructions, reason across modalities, and generalize to novel tasks. Ze et al. [63] proposed a generalizable framework based on 3D diffusion policies for coordinated humanoid manipulation. In parallel, efforts such as Mobile Aloha [64] and low-cost teleoperation-based learning [65] aim to democratize bimanual manipulation. In agriculture, Jin and Han [1]

highlighted the increasing demand for intelligent dual-arm systems capable of handling delicate tasks like harvesting. These trends point toward a new generation of dual-arm systems that can integrate perception, planning, and semantic reasoning. However, effectively addressing occlusion-aware manipulation through such systems remains an open and essential challenge.

2.4. Multimodal Reasoning in Robotics

Recent progress in multimodal reasoning frameworks has enabled robots to interpret and act upon visual and linguistic inputs jointly, offering a unified interface between perception, language understanding, and control. Representative models such as RT-2 [66], PaLM-E [67], and ROSGPT_Vision [68] explore the integration of web-scale vision-language knowledge with robotic control pipelines. While these models demonstrate the feasibility of transferring abstract knowledge to physical tasks via language prompts, their deployment in real-world applications remains limited by issues of robustness, interpretability, and data efficiency. Recent studies have proposed incorporating physical constraints [69] and leveraging instruction tuning [70, 71] to improve alignment between high-level semantic instructions and downstream manipulation policies. A systematic review by Firoozi et al. [72] summarizes these advancements, identifying both technical progress and the remaining challenges toward reliable deployment.

In contrast, the application of multimodal models in agricultural robotics is still in its early stages and presents unique challenges. For example, Li et al. [73] developed AgroMind, a benchmark for evaluating multimodal understanding in agricultural scenes, and found that even state-of-the-art models underperform when confronted with complex backgrounds and fine-grained semantics. Smart farming systems incorporating mobile robots have begun to explore such models for navigation and monitoring tasks [74], yet their effectiveness in dynamic, occlusion-rich environments such as fruit orchards remains constrained. Zhu et al. [75] review recent applications of large-scale multimodal models in agriculture and highlight key issues such as limited generalization, domain-specific adaptation, and inconsistent reasoning in unstructured field conditions. On the decision-making side, language models have been utilized for high-level mission planning in agricultural settings [76], and imitation learning has been studied as a practical interface for knowledge transfer and manipulation policy learning in agricultural robots [77].

Looking forward, the integration of multimodal reasoning capabilities into embodied agricultural systems—particularly those involving mobile platforms with dual-arm manipulators—offers a promising direction for enabling autonomous fruit harvesting under occlusion. By combining structured language-driven inference with real-time multimodal perception [78], future systems may better interpret complex scenes and coordinate collaborative manipulation actions in agricultural environments.

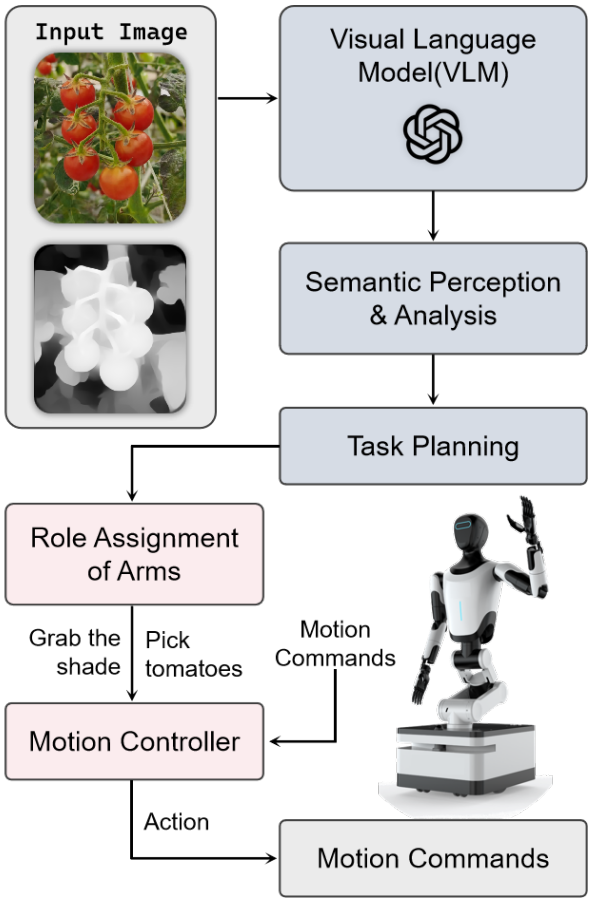


Figure 2: Overview of the proposed semantic-aware dual-arm harvesting system, combining large vision-language models with a mobile humanoid platform.

3. Problem Formulation and System Architecture

3.1. Problem Formulation

In this study, we focus on the automated harvesting of tomatoes in greenhouse environments under complex occlusion conditions using a dual-arm robot system enhanced by large language model (LLM) reasoning. The problem is formulated as a multimodal perception and decision-making task involving three key components: (1) occlusion-aware tomato recognition, (2) dual-arm collaborative action planning, and (3) LLM-based semantic reasoning.

Let the greenhouse environment be represented by an observable multimodal state space $S = \{I_v, D, \mathcal{M}_{sem}\}$, where I_v denotes the RGB visual input, D the corresponding depth map from the RGB-D sensor, and $\mathcal{M}_{sem}$ a semantic segmentation map capturing object-level understanding. The tomato harvesting task is defined as finding an optimal action sequence $\mathcal{A} = \{a_1, a_2, \dots, a_T\}$ over a time horizon T, where each action a_t belongs to the combined action space $\mathcal{A} = \mathcal{A}_L \times \mathcal{A}_R$, with $\mathcal{A}_L$ and $\mathcal{A}_R$ denoting the action spaces of the left and right manipulators respectively.

Each tomato target $x_i \in \mathcal{X}$ is described by a tuple:

$$x_i = (p_i, o_i), \tag{1}$$

where $p_i \in \mathbb{R}^3$ is the estimated 3D position and $o_i \in [0,1]$ is the occlusion ratio computed from depth and leaf overlap analysis. The objective is to maximize the expected harvesting success rate P_{succ} over the set of ripe tomatoes $\mathcal{X}_{\text{ripe}} \subset \mathcal{X}$:

$$\max_{\mathcal{A}} \quad \mathbb{E}[P_{\text{succ}}(\mathcal{A}, \mathcal{X}_{\text{ripe}})]. \tag{2}$$

3.2. System Architecture

As illustrated in Fig. 2, the proposed semantic-aware dual-arm tomato harvesting system integrates perception, reasoning, planning, and control into a unified closed-loop framework. It consists of the following four key modules:

1. **Visual-Language Perception Module:** This module receives RGB and depth images as input from the onboard RGB-D sensor. A large vision-language model (VLM), such as GPT-4o, processes the multimodal inputs to extract semantic understanding of the scene, including fruit locations, occlusion relationships, and contextual object roles (e.g., leaf versus fruit).
2. **Semantic Analysis and Task Planning Module:** The interpreted scene information is passed to a semantic perception and reasoning component, which leverages LLM capabilities to analyze occlusions and generate a task plan. This includes subgoal decomposition (e.g., uncover then harvest), priority determination (e.g., ripeness and occlusion level), and motion intent formulation for each manipulator.
3. **Role Assignment and Motion Command Module:** Based on the semantic plan, this module performs role allocation: one arm is designated to handle occlusions (e.g., lifting leaves), while the other executes harvesting. The module translates high-level subgoals into low-level motion commands and resolves potential arm conflicts, ensuring coordinated actions.
4. **Motion Control Module:** A hierarchical motion controller receives motion commands for both arms, combining impedance control for safe leaf interaction and precision control for fruit grasping. The robot executes the commands and provides feedback, which can trigger re-evaluation by the VLM if occlusion conditions change or task failure occurs.

This architecture enables dynamic understanding of occluded environments, flexible dual-arm coordination, and adaptive decision-making through continuous multimodal reasoning. By combining semantic inference with physically grounded control, the system achieves efficient and gentle tomato harvesting even in cluttered and unpredictable greenhouse conditions.

4. Methodology

4.1. Occlusion-Aware Grasp Point Detection via Vision-Language Reasoning

To enable accurate detection of occlusion-related grasp points under limited supervision, we adopt a vision-language reasoning framework enhanced by a multi-image prompting strategy. As illustrated in Fig. 3, the system constructs a composite visual input by horizontally concatenating multiple reference images—each containing annotated occlusion masks—with a target image lacking annotations. This layout provides the vision-language model (VLM) with both prior examples and contextual cues, allowing it to infer semantically valid grasp locations in the unmarked image through analogy.

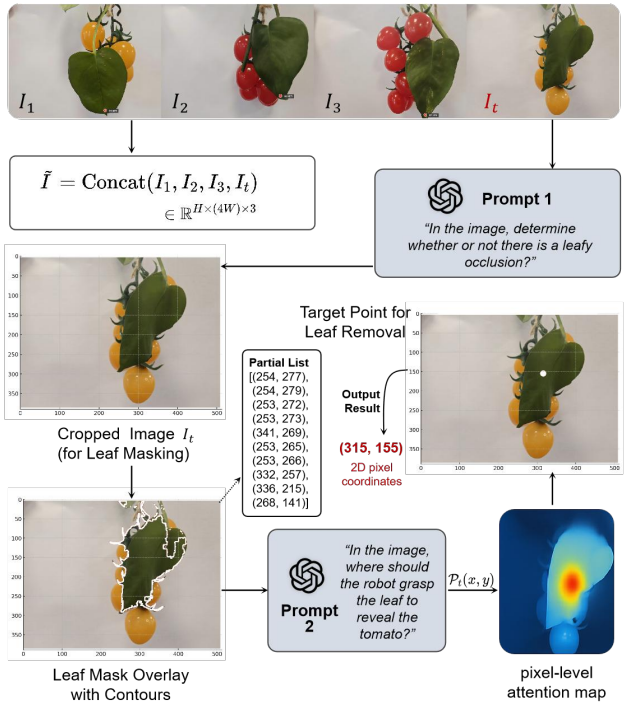


Figure 3: Illustration of the occlusion-aware grasp point detection pipeline using multi-image vision-language reasoning. The input consists of three reference images I_1, I_2, I_3 with annotated occlusions and one unmarked target image I_t, concatenated to form a composite image $\tilde{I}$. The system interacts with a frozen Vision-Language Model (VLM) using two textual prompts: (1) to verify the presence of a leafy occlusion, and (2) to localize the optimal leaf grasp point. The VLM returns a pixel-level attention map $\mathcal{P}_q$, which is constrained by a green region mask extracted from I_t to produce the final grasp point $g_{\text{leaf}} = (315, 155)$ in image space. This point is passed to the motion planning module for occlusion removal.

Formally, let $\{I_1, I_2, I_3\}$ denote three reference images with labeled leaf grasp points, and let I_t denote the target image. These images are concatenated to form a single composite input:

$$\tilde{I} = \text{Concat}(I_1, I_2, I_3, I_t) \in \mathbb{R}^{H\times(4W)\times 3}. \tag{3}$$

The VLM is first queried with a binary classification prompt: *"In the image, determine whether or not there is a leafy occlusion?"*, to verify whether occlusion is present in I_t. If confirmed, a second instruction is issued: *"In the fourth image, where should the robot grasp the leaf to reveal the tomato?"* This prompt, along with the composite image, is

processed by a frozen vision-language model $\mathcal{F}_{\text{VLM}}$, such as GPT-4o-VL or BLIP-2, which outputs a dense attention map $\mathcal{P}_q \in [0,1]^{H\times(4W)}$ indicating the pixel-level relevance to the query across the entire concatenated image.

To isolate the attention corresponding to the target image I_t, we extract the rightmost quarter of $\mathcal{P}_q$, denoted as $\mathcal{P}_t \in [0,1]^{H\times W}$. Each pixel value in $\mathcal{P}_t$ reflects the model's belief that the robot should interact with that specific location in the target image to perform the desired occlusion-removal action. This attention map provides a transparent and interpretable representation for identifying the optimal grasp point.

To further constrain the prediction to physically valid graspable regions, we perform a color-based leaf segmentation on the target image I_t. The image is converted to HSV color space, and a binary mask $\mathcal{M}_{\text{leaf}} \in \{0,1\}^{H\times W}$ is generated by applying thresholding rules in the green hue range (e.g., $H \in [35°, 85°]$, $S > 40, V > 40$). The mask is refined by morphological operations and connected component filtering to isolate the dominant occluding leaf region.

The final grasp point is obtained by selecting the pixel with the highest attention score within the valid leaf region:

$$g_{\text{leaf}} = \arg\max_{(x,y)} \left[\mathcal{P}_t(x,y) \cdot \mathcal{M}_{\text{leaf}}(x,y)\right]. \tag{4}$$

This ensures that the chosen grasp candidate is both semantically aligned with the task instruction and physically grounded on the occluding surface.

This multi-image prompt design enhances the model's ability to generalize across novel occlusion scenarios by enabling semantic analogy over in-context examples, without requiring task-specific retraining. The pixel-level output $\mathcal{P}_t$ serves not only as a robust basis for grasp selection but also as a transparent indicator of the VLM's internal reasoning. The resulting point g_{leaf} is passed to the robotic system for leaf removal execution.

4.2. Vision-Language Guided 3D Localization of Cutting Points

To support robotic tomato harvesting, we design a vision-language-based pipeline for identifying and localizing semantically meaningful cutting points. The system integrates pixel-level reasoning from a vision-language model (VLM) with depth sensing and camera calibration to accurately project 2D predictions into 3D space.

As shown in Fig. 4, the VLM takes as input a natural image of a tomato bunch along with a descriptive prompt, such as:

"Identify the optimal cut point where the stem meets the tomato cluster, ensuring the full bunch is preserved."

Following a similar strategy as in Sec. 4.1, the input to the VLM is structured as a horizontally concatenated sequence of multiple reference images and one target image. The reference images are annotated with ground-truth cut points from previous examples, while the target image remains unmarked. This multi-image layout, combined with the task-specific prompt, enables the VLM to reason by analogy across the input set.

As in the occlusion removal task of Sec. 4.1, the model produces a pixel-level attention map $\mathcal{P}_q \in [0,1]^{H\times(4W)}$ that reflects the spatial relevance of all regions to the given instruction. To isolate the output for the target image, the rightmost quarter of the attention map is extracted as $\mathcal{P}_t \in [0,1]^{H\times W}$. The final 2D coordinate of the cutting point is obtained by selecting the most salient location:

$$(h_x, h_y) = \arg\max_{(x,y)} \mathcal{P}_t(x,y) \tag{5}$$

Unlike in Sec. 4.1, which focuses on grasping an occluding leaf, this task aims to detect a stable and semantically meaningful cut point on the tomato stem for detaching the full cluster. The attention map $\mathcal{P}_t$ serves as a transparent intermediate representation that enables visual verification and debugging of the model's prediction.

To transform the 2D cut point into a 3D position, we retrieve the depth value $z = D(h_x, h_y)$ from a corresponding aligned depth map D, captured using an RGB-D camera. Using the intrinsic parameters of the camera—focal lengths f_x, f_y and optical center (c_x, c_y)—we compute the real-world coordinates (x, y, z) via the pinhole camera model:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} (h_x - c_x)\dfrac{z}{f_x} \\ (h_y - c_y)\dfrac{z}{f_y} \end{bmatrix} \tag{6}$$

The resulting 3D position $p = (x, y, z)^T$ is then passed to the robotic control module for motion planning and execution. This approach bridges semantic perception and physical control by leveraging VLMs to reason about where a harvest cut should be performed and grounding that decision geometrically in the real world.

Quantitative Evaluation. To quantitatively assess the accuracy of the proposed method, we evaluate the 2D and 3D localization errors of the predicted cutting points. Let (h_x, h_y) be the 2D coordinate predicted by the VLM, and (h_x^*, h_y^*) be the manually annotated ground-truth label. The average pixel-level root-mean-square error (RMSE) is computed as:

$$\text{RMSE}_{2D} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(h_x^{(i)} - h_x^{*(i)})^2 + (h_y^{(i)} - h_y^{*(i)})^2\right]} \tag{7}$$

where N denotes the number of evaluated images.

Furthermore, after projecting the 2D points into 3D space, the spatial error is evaluated by comparing the predicted 3D location $p = (x, y, z)$ with the ground-truth 3D point $p^* = (x^*, y^*, z^*)$ using Euclidean distance:

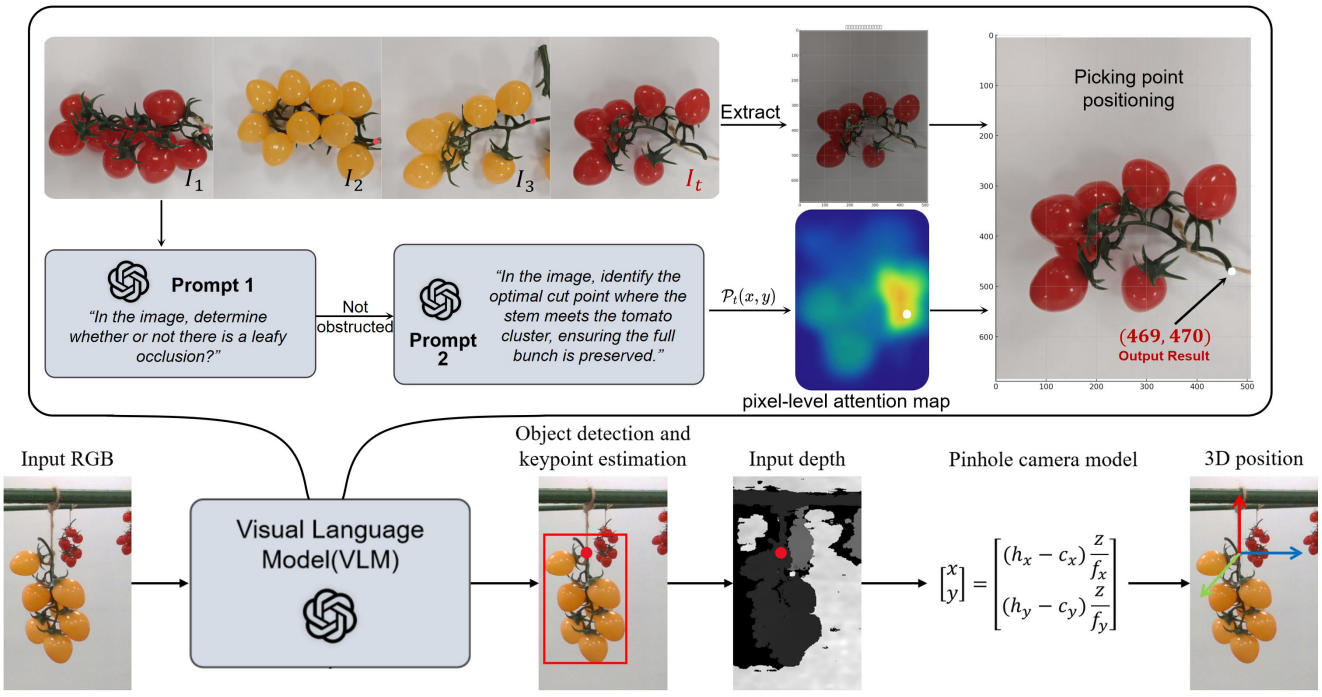


Figure 4: The proposed visual guidance system for object 3D localization using VLM. The system identifies the optimal cutting point in the image space using prompt-based reasoning, and projects it into 3D space through a calibrated pinhole camera model.

$$\text{Error}_{3D} = \frac{1}{N}\sum_{i=1}^{N}\left\|p^{(i)} - p^{*(i)}\right\|_2 \tag{8}$$

These metrics are used in our experiments to benchmark the reliability of the VLM-guided cutting point localization under various conditions. In practice, we observe that the average 2D localization error is within 4–6 pixels across typical resolutions, while the 3D projection error—limited by the accuracy of the RGB-D sensor—is within ± 0.5 cm, indicating that the predicted cut points are both visually and spatially consistent with ground truth.

4.3. Vision-Language Reasoning for Integrated Execution Planning

We propose a unified visual-linguistic framework in which a Vision-Language Model (VLM) serves as the central decision-making engine to coordinate perception, reasoning, task planning, and dual-arm execution for autonomous tomato harvesting. Unlike traditional pipelines that decouple perception and control, our approach enables end-to-end semantic reasoning and action grounding via multimodal prompting and interpretable visual-language outputs.

As illustrated in Fig. 2, the system receives input RGB and depth images (I, D) from an onboard RGB-D sensor. A large Vision-Language Model (e.g., GPT-4o) interprets the scene and sequentially drives decision-making. The pipeline centers around three core functionalities: occlusion detection and reasoning, dual-arm task allocation, and fine-grained motion triggering.

Arm Functionality and Role Assignment. The dual-arm platform consists of two specialized manipulators: the **right arm** is a 6-DoF dexterous hand with five fully actuated fingers, capable of compliant interaction with soft structures such as occluding leaves; the **left arm** is a custom-designed harvesting tool that integrates a scissor-like cutting mechanism and an adaptive gripper, enabling simultaneous severing and support of the tomato bunch to prevent dropping. The `Role Assignment of Arms` module dynamically allocates occlusion removal to the right arm and harvesting to the left arm based on semantic task segmentation.

Motion Control and Feedback Loop. To bridge high-level VLM decisions with low-level robotic control, we adopt a hybrid motion execution paradigm. Key manipulations—such as leaf grasping, lifting, and bunch cutting—are pre-defined as modular motion primitives, each represented in a JSON-based schema containing action type, target coordinates, and actuator parameters. An example command is:

```
{
"arm": "right",
"action": "grasp",
"target": [x, y, z],
"grip_width": 0.03
}
```

Each primitive is invoked by the VLM via prompt-

conditioned decoding, made possible through lightweight finetuning on task-specific examples. This structure enhances both safety and interpretability: actions are verifiable, bounded, and modular. The system performs re-evaluation after each action by re-invoking the VLM with updated (I, D), allowing dynamic adaptation in the presence of changing occlusions or failures.

Execution Logic and Algorithmic Overview. The complete decision-making process is formalized in Algorithm 1. Here, I and D denote the RGB and depth inputs, respectively. Semantic queries Q are issued to the VLM to assess occlusion states and identify target keypoints. If an occlusion is detected, a right-arm command is issued to lift the leaf. This loop continues until no occlusion remains, after which a left-arm harvesting command is executed. The function `ExecuteMotion` maps each semantic action to a pre-defined control primitive triggered via structured JSON commands.

Algorithm 1 Vision-Language Guided Dual-Arm Harvesting with **CoPickVLM**

Require: RGB image $I = I_v$, depth map D
Ensure: Cutting point p_{cut}, motion commands $C_{\text{left}}, C_{\text{right}}$

$Q_1 \leftarrow$ "Is the tomato bunch currently occluded by any leaves?"
`isOccluded` $\leftarrow \text{VLM}(I, D, Q_1)$
while `isOccluded` = True **do**
    $Q_2 \leftarrow$ "Where is the optimal grasp point on the leaf to remove the occlusion?"
    $g_{\text{leaf}} \leftarrow \text{VLM}(I, D, Q_2)$
    `Generate JSON:` $C_{\text{right}} = \text{GraspLeaf}(g_{\text{leaf}})$
    `ExecuteMotion`(C_{right})
    `Update` I, D
    `isOccluded` $\leftarrow \text{VLM}(I, D, Q_1)$
end while
$Q_3 \leftarrow$ "Where is the optimal cutting point on the tomato stem to preserve the bunch?"
$p_{\text{cut}} \leftarrow \text{VLM}(I, D, Q_3)$
`Generate JSON:` $C_{\text{left}} = \text{CutAndGrip}(p_{\text{cut}})$
`ExecuteMotion`(C_{left})
return p_{cut}

Framework Strengths and Generalization. This architecture offers several key advantages. First, it supports interpretable decision-making through transparent prompt-response pairs and visual attention outputs. Second, the modular structure of JSON motion primitives allows rapid extension to new actions or robot platforms. Third, the framework naturally generalizes to richer queries, such as assessing fruit ripeness, estimating yield, or ranking clusters by harvest priority. These capabilities are easily supported by modifying the query prompt, enabling powerful semantic flexibility without retraining.

Overall, this approach exemplifies a vision-language-robotics loop that integrates high-level understanding with grounded motor execution—supporting robust, adaptable, and safe manipulation in complex agricultural environments.

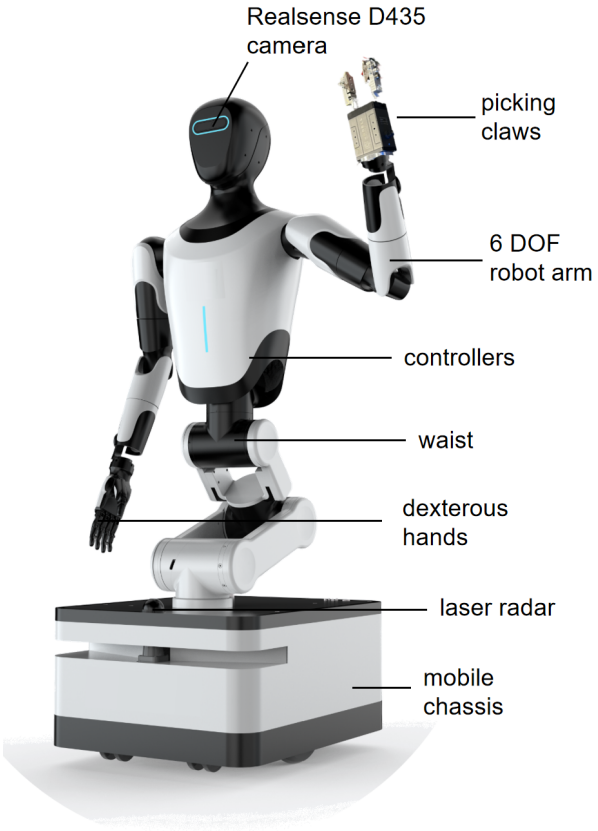


Figure 5: Mobile humanoid robot used in the experiments, featuring dual arms, vision sensors, and a wheeled chassis for autonomous navigation.

5. Experimental Validation

This section presents a comprehensive evaluation of the proposed vision-language guided dual-arm harvesting system. We begin by describing the experimental setup, followed by comparative experiments under occlusion conditions, and conclude with real-world robotic harvesting demonstrations under occlusion scenarios.

5.1. Experimental Setup

The experiments are conducted on a custom-built mobile humanoid robot equipped with dual-arm manipulation capabilities and rich perception modules. As shown in Fig. 5, the platform integrates both high-precision sensing and robust actuation components, enabling complex harvesting behaviors in unstructured environments.

The robot's head is equipped with an Intel RealSense D435 RGB-D camera for acquiring aligned color and depth information, crucial for 3D localization of tomato clusters and occlusions. Mounted on the right arm is a dexterous five-fingered robotic hand capable of delicate manipulation, such as grasping and moving obstructing

leaves. The left arm features a 6-degree-of-freedom (6-DoF) robot manipulator terminated with a custom harvesting end-effector—comprising picking claws designed for synchronized cutting and gripping actions.

To enable robust interaction and coordinated motion planning, the robot's torso houses embedded controllers responsible for low-level motor actuation and feedback. The waist joint allows limited articulation for reorienting the upper body, enhancing reachability in cluttered environments.

The entire platform is built on a laser-radar-guided mobile chassis, providing omnidirectional mobility for in-field navigation. The laser radar supports obstacle detection and SLAM-based localization.

All computational modules, including vision-language reasoning, multi-modal prompt execution, and trajectory generation, run on an onboard industrial-grade computer (Ubuntu 20.04, ROS Noetic). ROS facilitates seamless communication between sensors and actuators via real-time topics and services. The low-level motor control runs at 100 Hz to ensure responsive and stable actuation.

5.2. Comparative Experiments under Occlusion Conditions

To assess the performance of our proposed CoPickVLM system under varying degrees of occlusion, we conduct comparative experiments across three representative categories of tomato clusters: (1) **Non-occluded Cluster (Easy)**, where the stem and fruits are fully visible without any obstruction; (2) **Partially Occluded Cluster (Moderate)**, where leaves partially block the stem but the cut point remains partially visible; and (3) **Fully Occluded Cluster (Hard)**, where the leaf completely obscures the cut point and part of the fruit bunch.

We compare three methods in this study: a heuristic baseline based on color segmentation and rule-based inference, YOLOv8-Pose as a learning-based pose detector trained on 85 annotated images, and our proposed **CoPickVLM**, which leverages GPT-4o-based vision-language reasoning and dual-arm collaboration. Each method is evaluated under real-world greenhouse conditions using five quantitative metrics.

The *mean Average Precision at 0.5 IoU threshold (mAP@0.5)* is used to evaluate the spatial accuracy of the predicted cut-point localization against annotated ground truth. The *F1 score* reflects detection reliability by balancing precision and recall. We also measure the *inference time per frame* in milliseconds to assess performance. Task-level effectiveness is represented by the *success rate*, defined as the percentage of cases in which the system correctly identifies the cut point, executes the cut, and retrieves the tomato bunch intact. Finally, we include a binary *dual-arm reasoning* flag to indicate whether the method supports coordinated planning between the two arms—for instance, assigning one arm to remove occlusions and the other to perform harvesting.

As shown in Table 1, CoPickVLM consistently outperforms the baselines across all occlusion levels. In the **Easy** scenario, all methods achieve reasonable performance due to full visibility, but CoPickVLM still achieves the highest detection accuracy (mAP@0.5 = 0.713) and task success rate (86.9%), demonstrating its strong generalization even in less challenging settings.

Under **Moderate** occlusion, heuristic and YOLO-based methods show substantial performance degradation, particularly in F1 score and success rate, due to their limited capacity to infer partially hidden semantics. CoPickVLM maintains a robust performance (F1 = 0.597, success rate = 78.2%) by leveraging multimodal context and language-informed prompts to reason about occlusion patterns and infer plausible cut-point locations.

The performance gap is most evident in the **Hard** scenario. Here, CoPickVLM achieves a success rate of 58.7%, significantly higher than YOLOv8-Pose (45.7%) and heuristic methods (21.7%). This result underscores the advantage of VLM-based reasoning in highly ambiguous and occluded conditions. By semantically decomposing the scene and coordinating dual-arm behavior—where one arm removes the leaf and the other performs cutting—CoPickVLM enables successful task completion even when the cut point is completely hidden from view. This advantage comes at the cost of increased inference time due to vision-language reasoning with GPT-4o, which may impact real-time deployment in some settings.

Fig. 1 shows a full execution cycle of our proposed CoPickVLM system performing a tomato harvesting task under partial occlusion. The figure illustrates the semantic-guided dual-arm coordination enabled by our framework, with the process unfolding in three distinct stages:

In the first stage, the system identifies the occluding leaf and commands the right manipulator (equipped with a dexterous five-fingered hand) to reach and grasp the obstructing foliage. The grasp point is inferred through VLM-guided reasoning over RGB-D input and in-context prompts.

In the second stage, the right arm gently lifts and moves the occluding leaf away from the tomato bunch, effectively removing the visual obstruction. This enables subsequent perception and cutting point localization by the system.

In the final stage, the left arm—equipped with a hybrid cutting-grasping end-effector—is deployed to execute the harvest. The system localizes the optimal cut point and performs a precise cutting motion while simultaneously supporting the fruit bunch to prevent dropping. The harvested tomatoes are securely held and retracted, completing the task.

This visual rollout demonstrates the system's ability to reason through occlusions, assign roles to each manipulator, and execute a seamless multi-step harvesting procedure.

6. Conclusion

In this work, we present CoPickVLM, a vision-language-guided framework for dual-arm tomato harvesting under occlusion. By integrating semantic reasoning,

Table 1
Performance Comparison of Cut-Point Detection Methods under Varying Occlusion Scenarios

Scenario	Method	mAP@0.5	F1 Score	Time (ms)	Success Rate (%)	Dual-Arm Reasoning
Easy	Heuristic+Seg.	0.472	0.451	5.1	71.7	✗
	YOLOv8-Pose	0.694	0.637	9.3	84.7	✗
	Ours (CoPickVLM)	**0.713**	**0.682**	**320.1**	**86.9**	✓
Moderate	Heuristic+Seg.	0.295	0.252	5.3	47.8	✗
	YOLOv8-Pose	0.482	0.429	9.6	65.2	✗
	Ours (CoPickVLM)	**0.638**	**0.597**	**323.8**	**78.2**	✓
Hard	Heuristic+Seg.	0.102	0.097	5.2	21.7	✗
	YOLOv8-Pose	0.315	0.278	10.1	45.7	✗
	Ours (CoPickVLM)	**0.465**	**0.432**	**326.5**	**58.7**	✓

multimodal perception, and coordinated manipulation, our system enables robust and adaptive execution in visually complex environments. Through experiments, we demonstrate that CoPickVLM outperforms conventional baselines in both accuracy and task success rate, particularly under partial and full occlusions where traditional vision-only methods fail. The proposed architecture not only enables interpretable decision-making through language-based prompts but also supports dynamic task decomposition and dual-arm role assignment. These capabilities are crucial for real-world agricultural settings that demand both precision and flexibility.

Despite the promising results, several limitations remain in the current system. First, the full-body mobility of the humanoid platform, particularly the waist and torso, is not yet leveraged during harvesting. Second, the leaf-removal motion is currently implemented as a predefined primitive. In future work, we aim to incorporate imitation learning to enhance the adaptability and generalization of both perception and motion strategies. By learning from expert demonstrations, the system can dynamically adjust leaf removal and harvesting trajectories, further improving task success rates. We also plan to explore whole-body coordination, enabling more dexterous manipulation and efficient operation in cluttered or large-scale crop environments.

References

[1] Tantan Jin and Xiongzhe Han. Robotic arms in precision agriculture: A comprehensive review of the technologies, applications, challenges, and future prospects. *Computers and Electronics in Agriculture*, 221:108938, 2024.

[2] Shivendra Singh, Ram Vaishnav, Saurabh Gautam, and Somnath Banerjee. Agricultural robotics: A comprehensive review of applications, challenges and future prospects. In *2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, pages 1–8. IEEE, 2024.

[3] Vishnu Rajendran, Bappaditya Debnath, Sariah Mghames, Willow Mandil, Soran Parsa, Simon Parsons, and Amir Ghalamzan-E. Towards autonomous selective harvesting: A review of robot perception, robot design, motion planning and control. *Journal of Field Robotics*, 41(7):2247–2279, 2024.

[4] Mostafa Eissa. Precision agriculture using artificial intelligence and robotics. *Journal of Research in Agriculture and Food Sciences*, 1(2):35–35, 2024.

[5] Marta Benavides, Manuel Cantón-Garbín, Jorge Antonio Sánchez-Molina, and F Rodríguez. Automatic tomato and peduncle location system based on computer vision for use in robotized harvesting. *Applied Sciences*, 10(17):5887, 2020.

[6] Tom Duckett, Simon Pearson, Simon Blackmore, Bruce Grieve, Wen-Hua Chen, Grzegorz Cielniak, Jason Cleaversmith, Jian Dai, Steve Davis, Charles Fox, et al. Agricultural robotics: the future of robotic agriculture. *arXiv preprint arXiv:1806.06762*, 2018.

[7] C Wouter Bac, Eldert J Van Henten, Jochen Hemming, and Yael Edan. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal of field robotics*, 31(6):888–911, 2014.

[8] Hongyu Zhou, Xing Wang, Wesley Au, Hanwen Kang, and Chao Chen. Intelligent robots for fruit harvesting: Recent developments and future challenges. *Precision Agriculture*, 23(5):1856–1907, 2022.

[9] Teng Sun, Wei Zhang, Xuan Gao, Wen Zhang, Nan Li, and Zhonghua Miao. Efficient occlusion avoidance based on active deep sensing for harvesting robots. *Computers and Electronics in Agriculture*, 225:109360, 2024.

[10] JJ Zhuang, SM Luo, CJ Hou, Yu Tang, Yong He, and XY Xue. Detection of orchard citrus fruits using a monocular machine vision-based method for automatic fruit picking applications. *Computers and Electronics in Agriculture*, 152:64–73, 2018.

[11] Gustavo Gil, Daniel Emilio Casagrande, Leonardo Pérez Cortés, and Rodrigo Verschae. Why the low adoption of robotics in the farms? challenges for the establishment of commercial agricultural robots. *Smart Agricultural Technology*, 3:100069, 2023.

[12] Spyros Fountas, Nikos Mylonas, Ioannis Malounas, Efthymios Rodias, Christoph Hellmann Santos, and Erik Pekkeriet. Agricultural robotics for field operations. *Sensors*, 20(9):2672, 2020.

[13] Jiawei Chen, Wei Ma, Hongsen Liao, Junhua Lu, Yuxin Yang, Jianping Qian, and Lijia Xu. Balancing accuracy and efficiency: The status and challenges of agricultural multi-arm harvesting robot research. *Agronomy*, 14(10):2209, 2024.

[14] Sandeep Kumar, Santhakumar Mohan, and Valeria Skitova. Designing and implementing a versatile agricultural robot: A vehicle manipulator system for efficient multitasking in farming operations. *Machines*, 11(8):776, 2023.

[15] Rui Xu and Changying Li. A modular agricultural robotic system (mars) for precision farming: Concept and implementation. *Journal of Field Robotics*, 39(4):387–409, 2022.

[16] Christian Smith, Yiannis Karayiannidis, Lazaros Nalpantidis, Xavi Gratal, Peng Qi, Dimos V Dimarogonas, and Danica Kragic. Dual arm manipulation—a survey. *Robotics and Autonomous systems*, 60(10):1340–1353, 2012.

[17] Zhi He, Li Ma, Yinchu Wang, Yongzhe Wei, Xinting Ding, Kai Li, and Yongjie Cui. Double-arm cooperation and implementing for harvesting kiwifruit. *Agriculture*, 12(11):1763, 2022.

[18] Guoniu Zhu, Xiao Xiao, Changsheng Li, Jin Ma, Godwin Ponraj, AV Prituja, and Hongliang Ren. A bimanual robotic teleoperation architecture with anthropomorphic hybrid grippers for unstructured manipulation tasks. *Applied Sciences*, 10(6):2086, 2020.

[19] Wanteng Ji, Xianhao Huang, Shubo Wang, and Xiongkui He. A comprehensive review of the research of the "eye–brain–hand" harvesting system in smart agriculture. *Agronomy*, 13(9):2237, 2023.

[20] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

[21] Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, et al. Multimodal fusion and vision-language models: A survey for robot vision. *arXiv preprint arXiv:2504.02477*, 2025.

[22] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024.

[23] Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence*, 2024.

[24] Yeseung Kim, Dohyun Kim, Jieun Choi, Jisang Park, Nayoung Oh, and Daehyung Park. A survey on integration of large language models with intelligent robots. *Intelligent Service Robotics*, 17(5):1091–1107, 2024.

[25] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.

[26] Hubert Fonteijn, Manya Afonso, Dick Lensink, Marcel Mooij, Nanne Faber, Arjan Vroegop, Gerrit Polder, and Ron Wehrens. Automatic phenotyping of tomatoes in production greenhouses using robotics and computer vision: from theory to practice. *Agronomy*, 11(8):1599, 2021.

[27] Guohua Gao, Shuangyou Wang, Ciyin Shuai, Zihua Zhang, Shuo Zhang, and Yongbing Feng. Recognition and detection of greenhouse tomatoes in complex environment. *Traitement du Signal*, 39(1), 2022.

[28] Ehud Barnea, Rotem Mairon, and Ohad Ben-Shahar. Colour-agnostic shape-based 3d fruit detection for crop harvesting robots. *Biosystems Engineering*, 146:57–70, 2016.

[29] Philippe Lyonel Touko Mbouembe, Guoxu Liu, Jordane Sikati, Suk Chan Kim, and Jae Ho Kim. An efficient tomato-detection method based on improved yolov4-tiny model in complex environment. *Frontiers in Plant Science*, 14:1150958, 2023.

[30] Guoxu Liu, Joseph Christian Nouaze, Philippe Lyonel Touko Mbouembe, and Jae Ho Kim. Yolo-tomato: A robust algorithm for tomato detection based on yolov3. *Sensors*, 20(7):2145, 2020.

[31] Xiangyang Sun. Enhanced tomato detection in greenhouse environments: A lightweight model based on s-yolo with high accuracy. *Frontiers in Plant Science*, 15:1451018, 2024.

[32] Jiacheng Rong, Hui Zhou, Fan Zhang, Ting Yuan, and Pengbo Wang. Tomato cluster detection and counting using improved yolov5 based on rgb-d fusion. *Computers and Electronics in Agriculture*, 207:107741, 2023.

[33] Wenjun Chen, Yuan Rao, Fengyi Wang, Yu Zhang, Tan Wang, Xiu Jin, Wenhui Hou, Zhaohui Jiang, and Wu Zhang. Mlp-based multimodal tomato detection in complex scenarios: Insights from task-specific analysis of feature fusion architectures. *Computers and Electronics in Agriculture*, 221:108951, 2024.

[34] Hong-Beom Choi, Jae-Kun Park, Soo Hyun Park, and Taek Sung Lee. Nerf-based 3d reconstruction pipeline for acquisition and analysis of tomato crop morphology. *Frontiers in Plant Science*, 15:1439086, 2024.

[35] James Blossom Eleojo. Utilizing vision-language models for detection of leaf-based diseases in tomatoes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29567–29569, 2025.

[36] Jinyang Li, Fengting Zhao, Hongmin Zhao, Guoxiong Zhou, Jiaxin Xu, Mingzhou Gao, Xin Li, Weisi Dai, Honliang Zhou, Yahui Hu, et al. A multi-modal open object detection model for tomato leaf diseases with strong generalization performance using pdc-vld. *Plant Phenomics*, 6:0220, 2024.

[37] Tawseef Ayoub Shaikh, Tabasum Rasool, K Veningston, and Syed Mufassir Yaseen. The role of large language models in agriculture: harvesting the future with llm intelligence. *Progress in Artificial Intelligence*, pages 1–48, 2024.

[38] Chris Lehnert, Chris McCool, Inkyu Sa, and Tristan Perez. Performance improvements of a sweet pepper harvesting robot in protected cropping environments. *Journal of Field Robotics*, 37(7):1197–1223, 2020.

[39] Liesbet van Herck, Polina Kurtser, Lieve Wittemans, and Yael Edan. Crop design for improved robotic harvesting: A case study of sweet pepper harvesting. *Biosystems engineering*, 192:294–308, 2020.

[40] Shigehiko Hayashi, Katsunobu Ganno, Yukitsugu Ishii, and Itsuo Tanaka. Robotic harvesting system for eggplants. *Japan Agricultural Research Quarterly: JARQ*, 36(3):163–168, 2002.

[41] Feng Qingchun, Zheng Wengang, Qiu Quan, Jiang Kai, and Guo Rui. Study on strawberry robotic harvesting system. In *2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, volume 1, pages 320–324. IEEE, 2012.

[42] Qingchun Feng, Wei Zou, Pengfei Fan, Chunfeng Zhang, and Xiu Wang. Design and test of robotic harvesting system for cherry tomato. *International Journal of Agricultural and Biological Engineering*, 11(1):96–100, 2018.

[43] Yatao Li, Shunkai Wu, Leiying He, Junhua Tong, Runmao Zhao, Jiangming Jia, Jianneng Chen, and Chuanyu Wu. Development and field evaluation of a robotic harvesting system for plucking high-quality tea. *Computers and Electronics in Agriculture*, 206:107659, 2023.

[44] Fu Zhang, Zijun Chen, Yafei Wang, Ruofei Bao, Xingguang Chen, Sanling Fu, Mimi Tian, and Yakun Zhang. Research on flexible end-effectors with humanoid grasp function for small spherical fruit picking. *Agriculture*, 13(1):123, 2023.

[45] Leonidas Droukas, Zoe Doulgeri, Nikolaos L Tsakiridis, Dimitra Triantafyllou, Ioannis Kleitsiotis, Ioannis Mariolis, Dimitrios Giakoumis, Dimitrios Tzovaras, Dimitrios Kateris, and Dionysis Bochtis. A survey of robotic harvesting systems and enabling technologies. *Journal of Intelligent & Robotic Systems*, 107(2):21, 2023.

[46] Mohd Fazly Mail, Joe Mari Maja, Michael Marshall, Matthew Cutulle, Gilbert Miller, and Edward Barnes. Agricultural harvesting robot concept design and system components: A review. *AgriEngineering*, 5(2):777–800, 2023.

[47] Eleni Vrochidou, Viktoria Nikoleta Tsakalidou, Ioannis Kalathas, Theodoros Gkrimpizis, Theodore Pachidis, and Vassilis G Kaburlasos. An overview of end effectors in agricultural robotic harvesting systems. *Agriculture*, 12(8):1240, 2022.

[48] Wang Lili, Zhao Bo, Fan Jinwei, Hu Xiaoan, Wei Shu, Li Yashuo, Qiangbing Zhou, and Wei Chongfeng. Development of a tomato harvesting robot used in greenhouse. *International Journal of Agricultural and Biological Engineering*, 10(4):140–149, 2017.

[49] Dong Wang, Yongxiang Dong, Jie Lian, and Dongbing Gu. Adaptive end-effector pose control for tomato harvesting robots. *Journal of Field Robotics*, 40(3):535–551, 2023.

[50] Huaibei Xie, Deyi Kong, and Qiong Wang. Optimization and experimental study of bionic compliant end-effector for robotic cherry tomato harvesting. *Journal of Bionic Engineering*, 19(5):1314–1333, 2022.

[51] Jiacheng Rong, Lin Hu, Hui Zhou, Guanglin Dai, Ting Yuan, and Pengbo Wang. A selective harvesting robot for cherry tomatoes: Design, development, field evaluation analysis. *Journal of Field Robotics*, 41(8):2564–2582, 2024.

[52] Yi Ren, Zhengsheng Chen, Yechao Liu, Yikun Gu, Minghe Jin, and Hong Liu. Adaptive hybrid position/force control of dual-arm cooperative manipulators with uncertain dynamics and closed-chain kine-

matics. *Journal of the Franklin Institute*, 354(17):7767–7793, 2017.

[53] Tao Li, Feng Xie, Zhuoqun Zhao, Hui Zhao, Xin Guo, and Qingchun Feng. A multi-arm robot system for efficient apple harvesting: Perception, task plan and control. *Computers and electronics in agriculture*, 211:107979, 2023.

[54] Kyle Lammers, Kaixiang Zhang, Keyi Zhu, Pengyu Chu, Zhaojian Li, and Renfu Lu. Development and evaluation of a dual-arm robotic apple harvesting system. *Computers and Electronics in Agriculture*, 227:109586, 2024.

[55] Delia SepúLveda, Roemi Fernández, Eduardo Navas, Manuel Armada, and Pablo González-De-Santos. Robotic aubergine harvesting using dual-arm manipulation. *IEEE Access*, 8:121889–121904, 2020.

[56] Yuanzhe Cui, Zhipeng Xu, Lou Zhong, Pengjie Xu, Yichao Shen, and Qirong Tang. A task-adaptive deep reinforcement learning framework for dual-arm robot manipulation. *IEEE Transactions on Automation Science and Engineering*, 22:466–479, 2024.

[57] Yajun Li, Qingchun Feng, Yifan Zhang, Chuanlang Peng, and Chunjiang Zhao. Intermittent stop-move motion planning for dual-arm tomato harvesting robot in greenhouse based on deep reinforcement learning. *Biomimetics*, 9(2):105, 2024.

[58] Xinbo Yu, Shuang Zhang, Liang Sun, Yu Wang, Chengqian Xue, and Bin Li. Cooperative control of dual-arm robots in different human-robot collaborative tasks. *Assembly Automation*, 40(1):95–104, 2020.

[59] Sotiris Makris, Panagiota Tsarouchi, Aleksandros-Stereos Matthaiakis, Athanasios Athanasatos, Xenofon Chatzigeorgiou, Michael Stefos, Konstantinos Giavridis, and Sotiris Aivaliotis. Dual arm robot in cooperation with humans for flexible assembly. *Cirp Annals*, 66(1):13–16, 2017.

[60] Mohamed Abbas, Jyotindra Narayan, and Santosha K Dwivedy. A systematic review on cooperative dual-arm manipulators: modeling, planning, control, and vision strategies. *International Journal of Intelligent Robotics and Applications*, 7(4):683–707, 2023.

[61] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π0: A vision-language-action flow model for general robot control, 2024. *URL https://arxiv. org/abs/2410.24164*.

[62] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, et al. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775*, 2024.

[63] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024.

[64] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[65] 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.

[66] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[67] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023.

[68] Bilel Benjdira, Anis Koubaa, and Anas M Ali. Rosgpt_vision: Commanding robots using only language models' prompts. *arXiv preprint arXiv:2308.11236*, 2023.

[69] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024.

[70] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[71] Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*, 2023.

[72] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025.

[73] Qingmei Li, Yang Zhang, Zurong Mai, Yuhang Chen, Shuohong Lou, Henglian Huang, Jiarui Zhang, Zhiwei Zhang, Yibin Wen, Weijia Li, et al. Can large multimodal models understand agricultural scenes? benchmarking with agromind. *arXiv preprint arXiv:2505.12207*, 2025.

[74] Darío Fernando Yépez-Ponce, José Vicente Salcedo, Paúl D Rosero-Montalvo, and Javier Sanchis. Mobile robotics in smart farming: current trends and applications. *Frontiers in Artificial Intelligence*, 6:1213330, 2023.

[75] Hongyan Zhu, Shuai Qin, Min Su, Chengzhi Lin, Anjie Li, and Junfeng Gao. Harnessing large vision and language models in agriculture: A review. *arXiv preprint arXiv:2407.19679*, 2024.

[76] Marcos Abel Zuzuárregui and Stefano Carpin. Leveraging llms for mission planning in precision agriculture. *arXiv preprint arXiv:2506.10093*, 2025.

[77] Siavash Mahmoudi, Amirreza Davar, Pouya Sohrabipour, Ramesh Bahadur Bist, Yang Tao, and Dongyi Wang. Leveraging imitation learning in agricultural robotics: a comprehensive survey and comparative analysis. *Frontiers in Robotics and AI*, 11:1441312, 2024.

[78] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.