Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Gemini Robotics in Agriculture: Applications, Models, Fruit Industry, and Future Perspectives

mustafa mhamed

mustafamhamed2099@gmail.com

China Agricultural University https://orcid.org/0000-0002-3106-669X

Zhao Zhang

China Agricultural University

Short Report

Keywords: Gemini Robotics, agricultural robotics, embodied AI, VLMs, VLA, fruit harvesting, autonomous navigation, precision agriculture

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11695753/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Gemini Robotics in Agriculture: Applications, Models, Fruit Industry, and Future Perspectives

Mustafa Mhamed[a,b,c], Zhao Zhang[a,b,c]

[a] *College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China*

[b] *Key Laboratory of Smart Agriculture System Integration, Ministry of Education, China Agricultural University, Beijing 100083, China*

[c] *Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China*

Abstract

Agricultural robotics is progressing from task-specific automation toward embodied systems that integrate multimodal perception, language, reasoning, and physical action. This chapter reviews Gemini Robotics and closely related vision-language model (VLM) and vision-language-action (VLA) developments in the context of agriculture, with particular emphasis on fruit production. It first presents a conceptual agricultural architecture linking sensing, cognition, decision and control, actuation, connectivity, power and safety. It then examines emerging evidence from language-guided agricultural navigation, spatial-memory systems, occlusion-aware harvest planning, VLA-based strawberry harvesting, obstacle separation and text-prompted multi-fruit picking. The reviewed studies show growing capability in navigation, semantic reasoning and closed-loop manipulation, but also expose important limitations in agricultural validation, cross-crop generalization, action datasets, real-time computing, safety and commercial readiness. For fruit systems, the principal technical challenges remain occlusion, maturity and accessibility assessment, three-dimensional localization, delicate detachment and adaptation across cultivars, canopies and seasons. The chapter therefore proposes a future pathway centered on multimodal agricultural foundation models, edge VLA inference, soft and tactile manipulation, continual learning, natural-language human-robot collaboration and coordinated UAV-UGV-manipulator systems. The Gemini Robotics provides a promising foundation for moving agricultural robots from isolated perception-action pipelines toward general-purpose embodied intelligence, although extensive field validation is required before broad deployment.

Keywords: Gemini Robotics; agricultural robotics; embodied AI; VLMs; VLA; fruit harvesting; autonomous navigation; precision agriculture.

1. Introduction

Agriculture is changing from conventional mechanisation to intelligent and autonomous production systems based on AI, computer vision, multimodal sensing, and robotics [1-3]. Agricultural robots are increasingly being developed for crop monitoring, precision spraying, weed control, pruning and harvesting, phenotyping, and even post-harvest processing [4-6], but most of the existing systems are task-specific and will be based on predefined perception pipelines, specialized datasets and control methods programmed to optimize them [7-9]; the gap in generalizing from one to another crops, environments and new situations and technologies is still a major roadblock to fully autonomous agriculture.

Recent advances in VLMs, VLA models and embodied AI enable us to move from task-specific robots to general intelligent robots [10]. Unlike conventional robotic pipelines, VLA models connect visual perception and natural language understanding with physical actions [11]. Robots can interpret instructions, reason about their environments, plan tasks, and interact with objects in a more natural manner, which is very important for agriculture, where robots need to work with variable light, extreme occlusion, irregular plant structures, moving vegetation, uneven terrain, and very high biological variability [12].

Gemini Robotics, developed by Google DeepMind, is one of the most significant developments in this direction; based on Gemini's multimodal model, Gemini Robotics integrates vision, language, reasoning and robot actions with a focus on generality, interactivity and dexterity [13]. The embodied reasoning counterpart also supports spatial reasoning, object localization, grasp prediction, trajectory reasoning, and three-dimensional scene interpretation [14]. Gemini Robotics 1.5 has now leaped higher in embodied reasoning, multi-step task execution, and motion transfer among robotic systems. Combine these developments to create robots that are capable of moving beyond perception-action pipelines to perception reasoning planning action adaptation [16].

Such capabilities could be transformational in agriculture, particularly the fruit industry; fruit production is laborious for processing: fruit detection, counting, thinning, bagging, maturity assessment, pruning, harvesting, grading, sorting, and packaging. Autonomous fruit harvesting is particularly challenging to accomplish as fruit are often covered by leaves and branches, are often clustered and distributed in complex three-dimensional canopies [17-20]. New agricultural systems such as AgriVLN in agriculture and VLA-based harvesting systems are emerging in agriculture to support language-guided navigation and generalizable perception–action models in agricultural robotics [21-23]. Gemini Robotics might provide something more mature like this model by allowing a robot to understand what is happening around it, understand crop conditions, take appropriate actions, safely manipulate fruits, and respond to environmental situations.

However, Gemini Robotics has not been fully tested in the real world of agriculture; agriculture presents challenges that are far more challenging than in controlled robotics, such as sunlight variation, wind, rain, dust, seasonal domain change, fragile biological targets, limited connectivity, and strict safety requirements. Computational cost, inference latency, energy consumption, hallucination, explainability, cybersecurity, human-robot interaction, economic feasibility, and physical-action safety are also significant barriers. Most importantly, we still have a significant gap between the general-purpose capabilities of Gemini Robotics and how they are used in agriculture.

This chapter contributes a focused bridge between general-purpose Gemini Robotics and the operational requirements of agriculture. First, it organizes Gemini-enabled agricultural robotics as a reusable perception-reasoning-decision-action architecture rather than a single task-specific machine. Second, it synthesizes emerging VLM/VLA evidence across agricultural navigation, spatial memory, harvest planning and physical fruit manipulation, explicitly separating demonstrated capabilities from prospective applications. Third, it identifies fruit-specific bottlenecks—including occlusion, maturity and reachability assessment, fragile detachment, cultivar/canopy variation and limited action data—and maps each bottleneck to a concrete research direction. Finally, it proposes a staged pathway toward multi-task, adaptive and collaborative fruit robots through multimodal foundation models, edge inference, tactile/soft manipulation, continual learning and UAV-UGV-manipulator coordination.

The remainder of this chapter is organized as follows: Section 2 presents the Gemini Robotics conceptual model and proposed agricultural reference architecture. Section 3 reviews Gemini Robotics and VLM/VLA progress in agricultural domains. Section 4 discusses AI evolution and the prospective outlook for fruit agriculture. Section 5 synthesizes limitations, challenges and future research directions, and Section 6 concludes the chapter.

2. Genmie Robotic Systems: Conceptual Model and Architecture

Genmie is presented here as a modular, layered robotic platform and is designed for a wide range of physical tasks [24-26]; Genmie is thus best understood as a reference architecture: a common concept, a common hardware/software architecture, and a family of interchangeable subsystems that can be adapted for a specific application in agriculture field [27-29]. Figure 1 summarizes the logic underlying each Genmie setup, regardless of which domain it will be deployed to.

During all the stages of this loop, the system gathers sensor data regarding itself and its environment (Sense), combines that data with a new understanding of the task and environment (Interpret & Plan), executes the motion or task commands (Act), and retains the data to improve its internal model for the next iteration (Learn & Adapt). This closed-loop structure allows Genmie systems to be flexible and is the foundation for the conceptual model and architecture that will be explored in the next step.

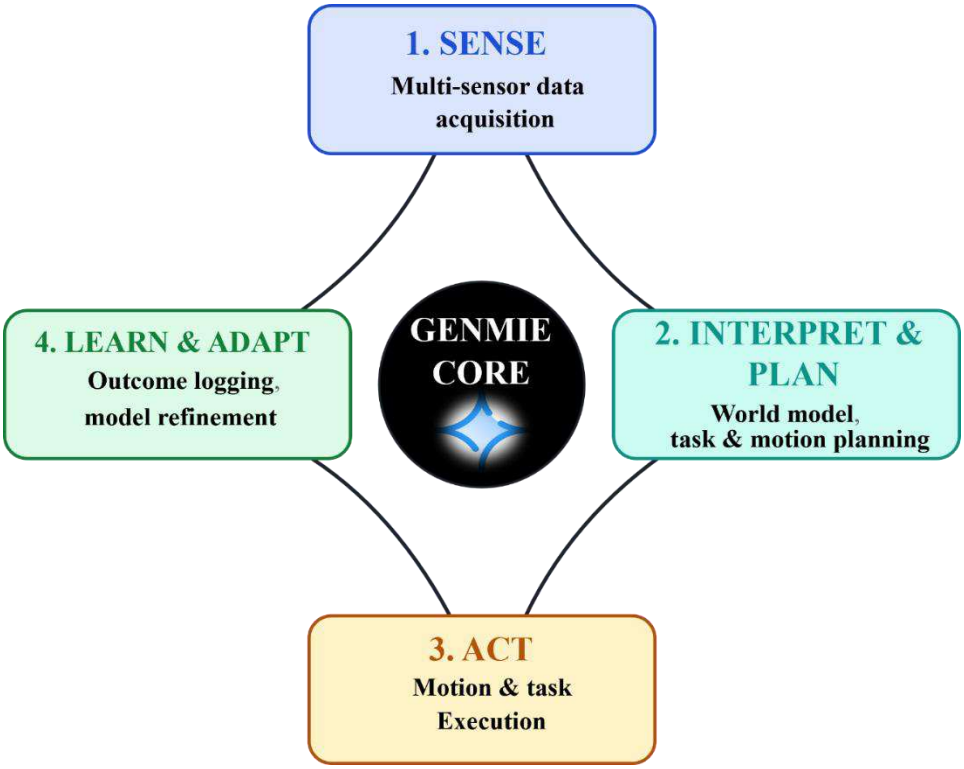


Fig. 1. Overall operational workflow of the Genmie Robotic System: a continuous Sense, interpret & Plan, act, Learn cycle coordinated by the Genmie core.

2.1 Conceptual Model and Architecture

The Genmie model categorizes robotic functions into four vertically aligned layers that each has a certain level of abstraction, and an overall layer that connects all four (Figure 2).

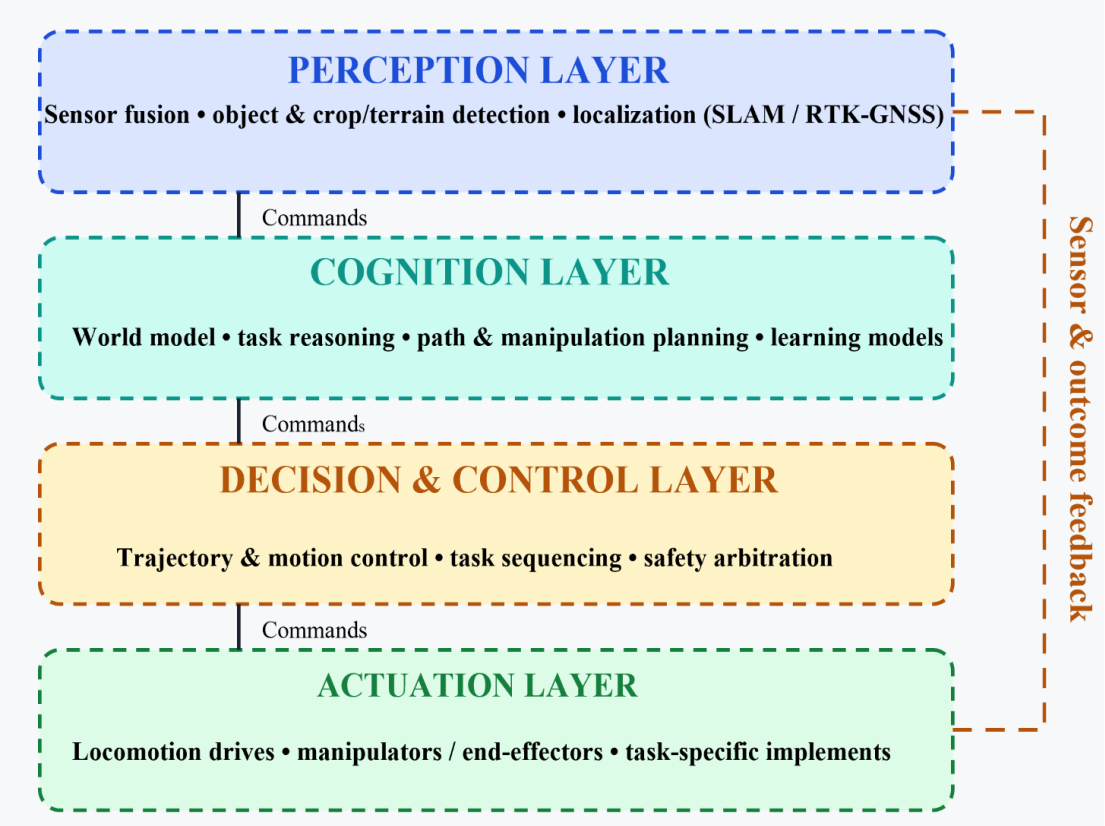


Fig. 2. The conceptual model of the Genmie Robotic System: four functional layers connected by a continuous perception-to-action feedback path, with a cross-cutting safety and communication layer

The four layers are as follows:

- Perception Layer - the sensor fusion, object and terrain/crop detection and self-localization (SLAM, RTK-GNSS) [30, 31].
- Cognition Layer - the world model inside, task reasoning and path and manipulation plan, and the learned models.
- Decision & Control Layer - turns the plan into a path/task and decision and decision-making and safety constraints.
- Actuation Layer - the physical hardware: locomotion drives, manipulators/end-effectors, and task-specific ones.

A feedback path brings the sensors and the results of the tasks back to the Perception Layer, completing the loop of Figure 1. The cross-cutting safety monitoring and fleet communication layer runs in addition to all four layers: It hosts the emergency-stop logic, continuous health monitoring, and telemetry and fleet coordination channel for the Genmie units working together. This model is implemented in hardware and software, as shown in Figure 2. The system is composed of a sensing suite, an edge compute core, and an actuation group, and is complemented by a power management unit, a safety controller, and a connectivity module to connect the platform to cloud-based fleet management.

In the edge compute core, the perception module takes raw sensor data to create a coherent and localized view of the world; the planning and cognition module uses it to fine-tune the world model and design action and motion plans; and the motion/task controller translates the action into detailed direction/action from the sensors and actuators. The safety controller works autonomously from this pipeline to stop it, and is a sentinel to the system, no matter if it is updated or running. The power management unit gives energy to the sensor and actuator parts and the connectivity module connects the platform to a cloud-based fleet management platform to see the fleet in real time, upgrade the system from the sky and share the work with other robots.

2.2 Comprehensive Structure of the Genmie Robotic System

The Genmie platform is organized into seven collaborative subsystems that are managed by a central Genmie Core Controller that orchestrates and synchronizes tasks on the platform.

Each component communicates its status and directives with the central controller to control the whole platform's internal model and how the tasks are executed within each subsystem. In addition to this hub-and-spoke coordination, neighboring subsystems are also able to interact directly (e.g., sensing suite, mobility and chassis are able to communicate without waiting to send messages to the central controller) for terrain-aware navigation. This framework is independent of the specific areas of the application: seven subsystems and coordination are present in all Genmie implementations, including the agricultural setup.

2.3 Cross-Domain Applications of Gemini Robotics

Gemini Robotics extends conventional task-specific robotics to general embodied intelligence via multimodal perception, language understanding, reasoning, planning, and physical action [15]. These capabilities provide a common foundation for transferring robotic intelligence across different embodiments and operational environments (Figure 3).

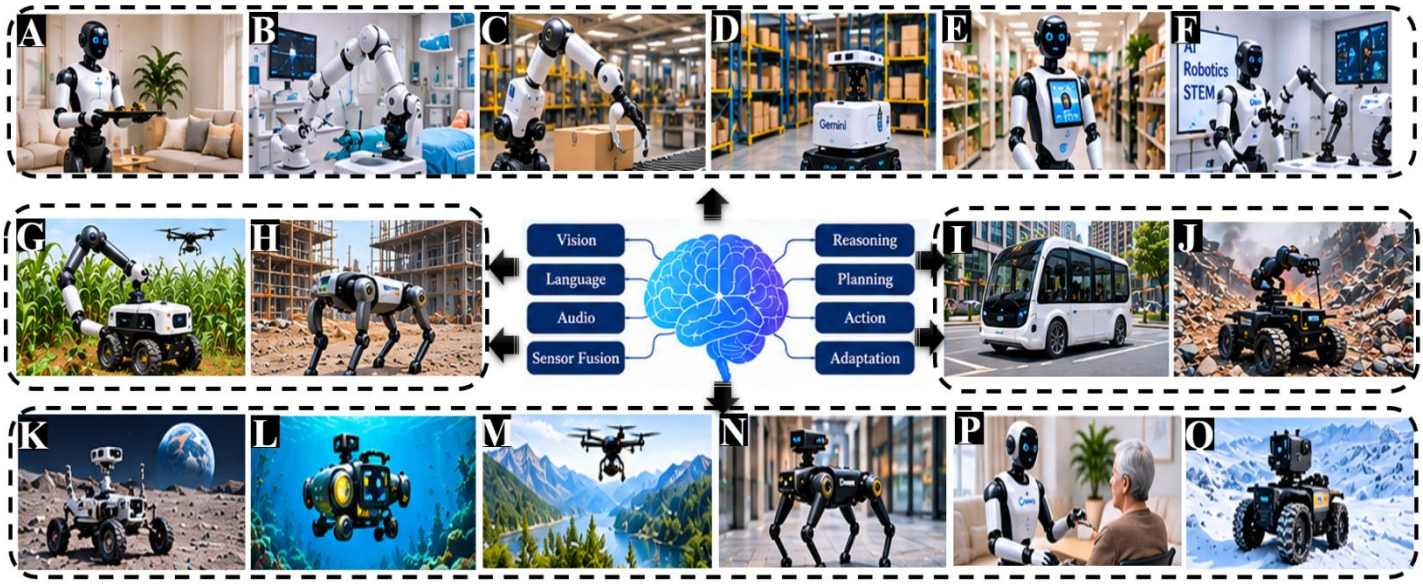


Fig. 3. Cross-domain applications of Gemini Robotics and embodied AI. The central multimodal framework integrates vision, language, audio, and sensor fusion with reasoning, planning, action, and adaptation to support diverse robotic applications: (A) home and service robotics, (B) healthcare and medical robotics, (C) industrial and manufacturing robotics, (D) logistics and warehousing, (E) retail and hospitality, (F) education and research, (G) agriculture and farming, (H) construction and infrastructure, (I) transportation and mobility, (J) disaster response and rescue, (K) space exploration, (L) marine and underwater robotics, (M) environmental monitoring, (N) security and surveillance, (P) elderly-care and assistive robotics, and (Q) extreme-environment robotics. The framework illustrates the potential transfer of general-purpose embodied intelligence across different robot embodiments and operational environments.

In home and service robotics (Figure 3A), Gemini-like systems can support natural language interaction, object recognition, manipulation, and household assistance. In healthcare and elderly care (Figure 3B), embodied AI can support rehabilitation, monitoring, assistive manipulation, and daily-living support; however, the safety of these methods remains to be verified. Industrial manufacturing and logistics (Figure 3C, D) are especially promising domains for adaptive assembly, inspection, material handling, picking, packing, navigation, and human-robot collaboration.

The same embodied framework (Figure 3E, F) can be applied to retail and hospitality through interactive service robots and education and research through language-guided demonstrations, robot learning, and human-robot interaction. In agriculture (Figure 3G), VLM/VLA-enabled robots could combine crop perception, reasoning, navigation, spraying, and fruit harvesting, but robust outdoor validation and agricultural action datasets remain important research needs .

More challenging areas are construction, transportation, disaster response, security, and environmental

monitoring (Figure 3H-J), where adaptive perception and spatial reasoning could support inspection, autonomous navigation, anomaly detection, search and rescue, and operation in dynamic or hazardous environments. Similarly, marine, space, and other extreme environments (Figure 3K-Q) are prospective applications where autonomous reasoning, navigation, manipulation, and adaptation could reduce dependence on direct human intervention.

All these domains show Gemini Robotics should not be seen as a collection of domain-specific robots. Rather, it offers a potential general embodied-AI foundation that links perception, understanding, reasoning, planning, action, adaptation across a range of robotic platforms and applications.

Gemini Robotics has also grown from a research-oriented vision–language–action to a broader embodied-AI ecosystem where foundation models are used in collaboration with different robotic embodiments and technology partners. Google DeepMind has demonstrated the capabilities of Gemini Robotics in ALOHA, Franka arms, Apptronik Apollo humanoids, Agile Robots platforms, Boston Dynamics robots and so on, showing that the whole family of robots can be used for reasoning, manipulation, body control, and task adaptation (https://deepmind.google/models/gemini-robotics/). In parallel, Google DeepMind Robotics Accelerator is also available to support Apptronik, Agile Robots, Intrinsic, Teleo, FieldAI, Physical Intelligence, Tangible AI, Bubble Robotics and Dexterity, all of which are robotics in the field (humanoids, industrial automation, autonomous machines, general robotics). These advancements are indicative of a growing transition from foundation model work to actual use in the real world but the platform is still not yet a commercial/developer ecosystem in itself and we should not think of Gemini Robotics as a set of fully commercialized Gemini-branded robots (https://deepmind.google/accelerators/robotics/).

2.4 Genmie Robotic Structure in Agriculture

The Genmie framework (Section 2.2) is tailored to agricultural purposes; each of the seven broad subsystems is an agriculture-specific equivalent, and the whole system is built around a transportable chassis for field use (Figure 4). The Genmie Field Platform is situated at the center, with the modules for agriculture-specific sensors, computation/AI, tasks, and connectivity/power.

The Sensing Suite is converted into an agriculture-specific ensemble (RTK-GNSS, multispectral/RGB cameras, soil-moisture, and NDVI sensors); the Compute & AI Core is enhanced with models for crop and weed identification, row-following navigation, and yield/growth predictions; the Manipulation subsystem transforms into a set of interchangeable Task Modules (seeding, precision spraying, mechanical weeding, and harvesting end-effectors); and the Communication and Power subsystems are enhanced with a farm-IoT gateway, integration with a farm management system (FMS), and field-suitable power sources like solar-assisted battery packs [32-36]. The agricultural framework may be characterized as a series of six physical and logical strata, reflecting the overall design of Figure 4; however, each layer is tailored for row-crop field activities. Figure 5 elucidates this stratified configuration.

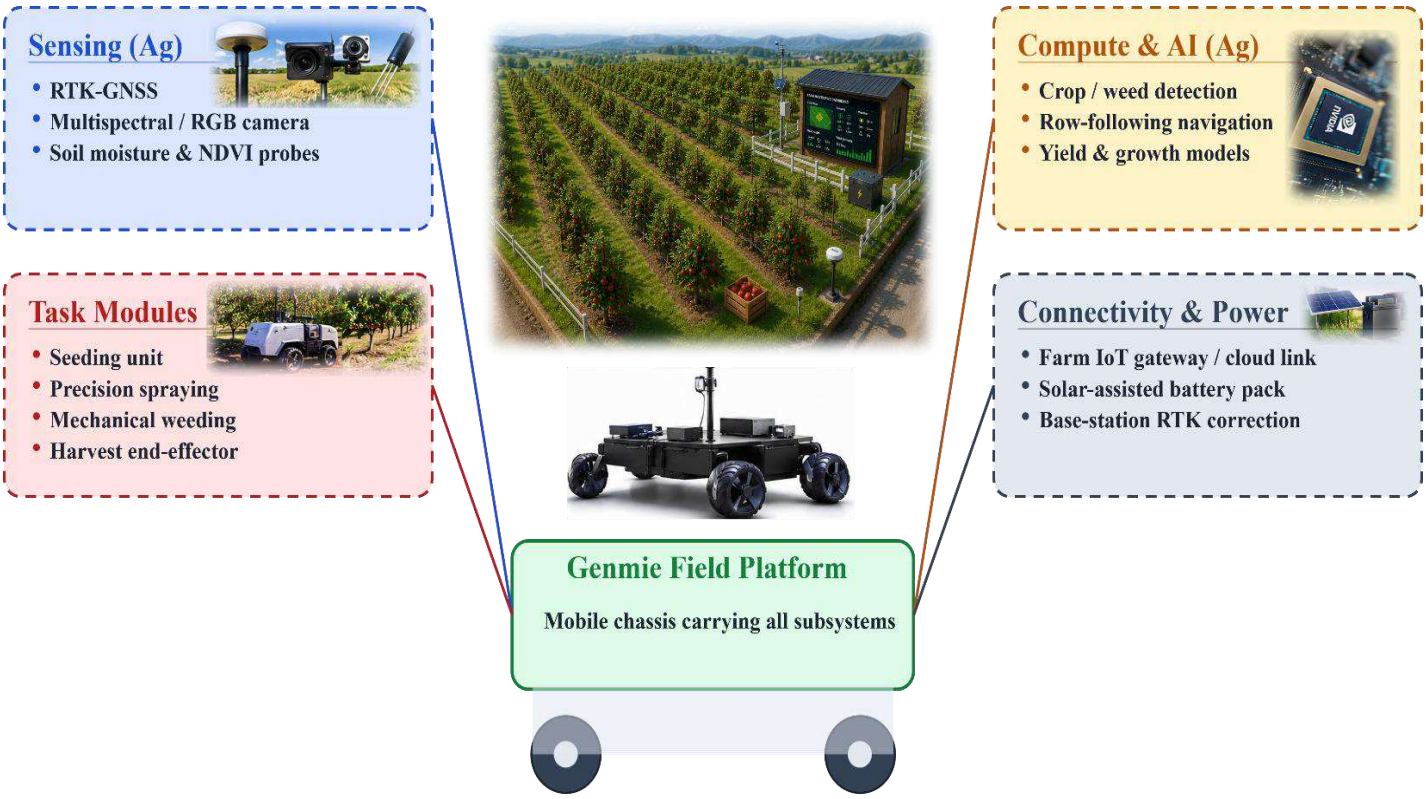


Fig. 4. The Genmie robotic structure in agriculture: the general Genmie subsystems of Figure 5 mapped onto a field-deployed configuration built around the Genmie Field Platform.

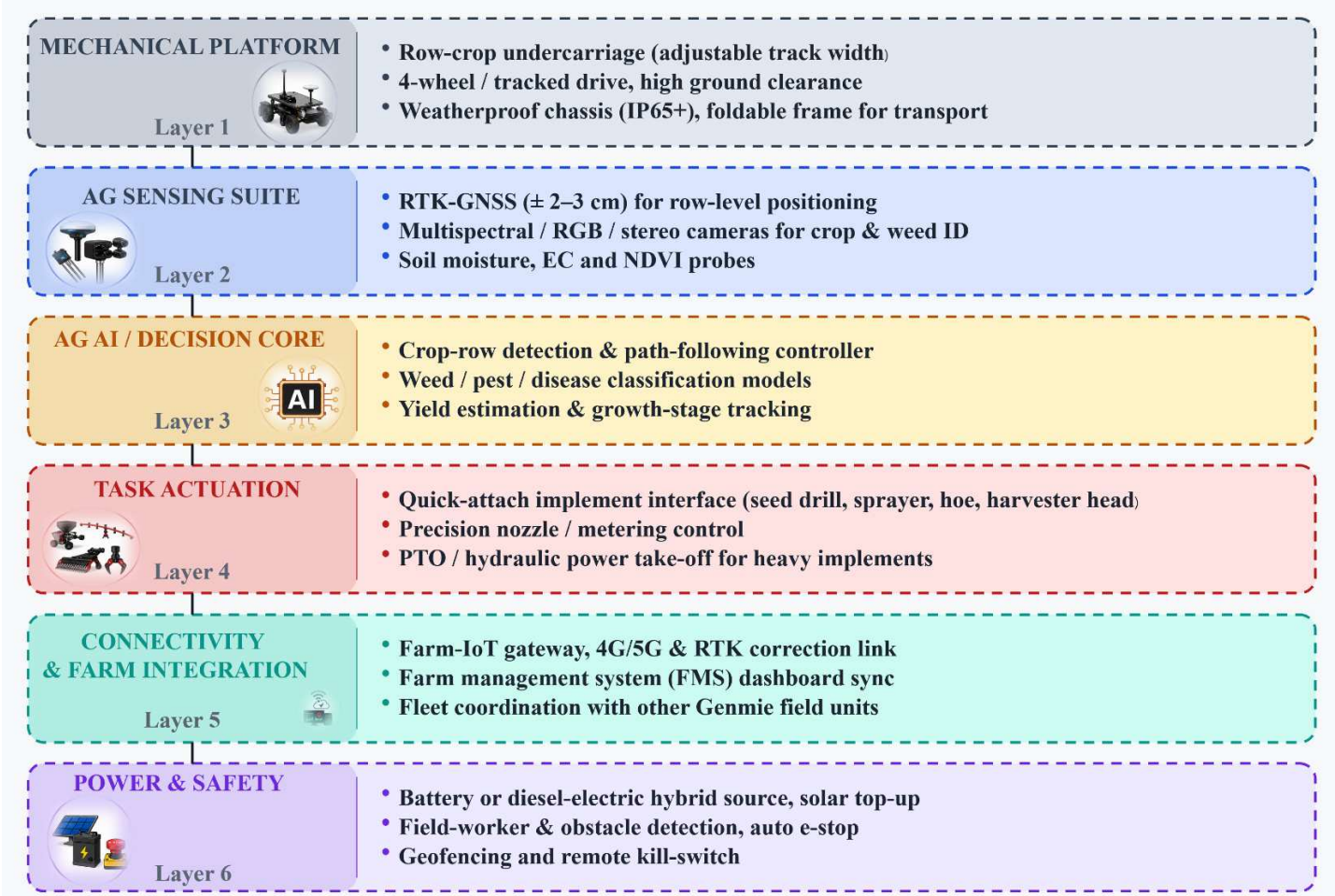


Fig. 5. The detailed structure of the Genmie agricultural configuration: six stacked layers, from the mechanical platform up through power and safety, specialized for row-crop field deployment.

The proposed Gemini Robotics-enabled agricultural architecture is separated into six layers to turn a conventional agricultural machine into an intelligent and perception-driven autonomous system. Layer 1, the mechanical platform, provides the physical foundation for our system via a row-crop-compatible undercarriage (with a wide track width, high ground clearance, four-wheel or tracked locomotion) and weather-resistant and transportable chassis that can be used in various fields. Layer 2, the Agricultural Sensing Suite, integrates multimodal environmental perception by integrating RTK-GNSS for centimetre-level localization, RGB, multispectral and stereo/depth cameras for crop, weed and obstacle perception, and complementary soil-moisture, electrical-conductivity and vegetation-index sensing for crop and field characterization. Layer 3, the agricultural AI and decision core, translates these diverse field observations into actionable intelligence in the form of crop-row identification, autonomous path following, weed, pest and disease identification and growth stage analysis, yield estimation, spatial reasoning and task-level decision-making. Within a Gemini Robotics system, it can further add vision-language–action and embodied reasoning capabilities to turn land and plant knowledge in the natural language of agriculture into context-driven robotic decisions. Layer 4, Task Actuation, transforms this decision into life-like operations via a modular quick-attach interface for seed drills, precision sprayers, mechanical weeders, harvesting heads, etc. and variable-rate nozzle or metering control, electrical, hydraulic, or power-take-off systems (if a higher power action is needed). Layer 5, Connectivity and Farm Integration, connects the autonomous unit to the whole ecosystem of digital farming through farm-IoT gateways, 4G/5G communications, RTK correction services, farm-management information systems, cloud/edge data synchronization and coordination of many field robots. Layer 6, Power and Safety, provides the physical and energy envelope with battery-electric or hybrid power, human and obstacle detection, collision avoidance and automatic emergency stopping, geofencing and remote intervention. Layers 1 and 6 establish the physical, energy and safety envelope, Layers 2–4 are the core perceive–reason–decide–act loop and Layer 5 connects the individual robot to farm-level intelligence and fleet management. This multi-layer modularity provides a reusable architecture rather than a single machine; orchard, greenhouse, vineyard, row crop and fruit production can be designed and built without having to re-specialize sensing, reasoning and actuation modules to make the autonomous agriculture system autonomous from scratch.

3. Gemini Robotics and VLM/VLA Progress in Agricultural Domains

Recent research, as shown in Table 1, indicates a definitive shift from general-purpose embodied AI to agriculture-focused VLM/VLA systems. Gemini Robotics [37] laid the groundwork by amalgamating vision, language, embodied thinking, and robotic manipulation, showcasing robust generalization, interaction, and dexterity, but without specialized assessment in agricultural settings. Gemini Robotics 1.5 [38] enhanced these functionalities via multi-step reasoning and motion transfer across various robotic forms, establishing a crucial technical foundation for versatile agricultural robots, while still missing

direct validation at the crop and field levels (Figure 6A).

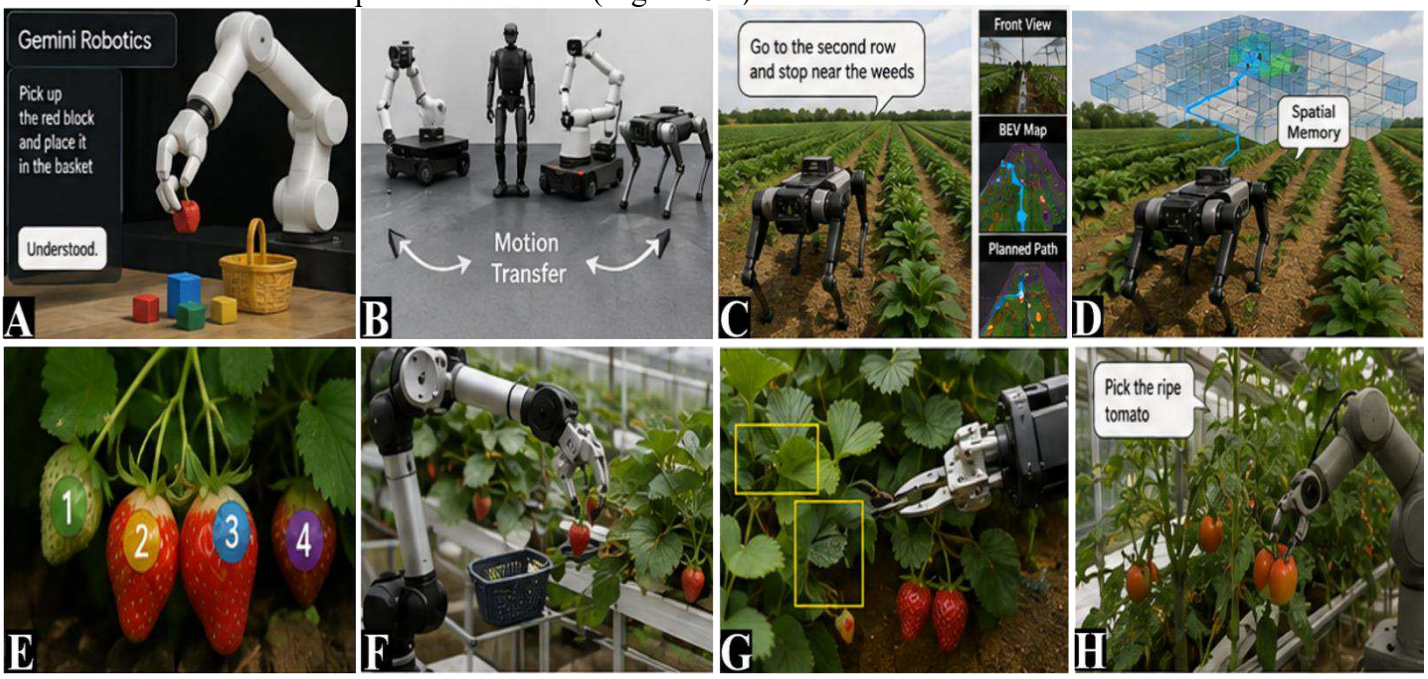


Fig. 6. Gemini Robotics and VLM/VLA developments and their emerging applications in agriculture. (A) embodied manipulation; (B) reasoning and motion transfer; (C) language-guided navigation; (D) spatial-memory navigation; (E) occlusion-aware harvest planning; (F) VLA strawberry harvesting; (G) VLM-guided obstacle removal; and (H) text-prompted multi-fruit harvesting.

Agricultural research has therefore started to convert these ideas into particular practices; AgriVLN [39] implemented language-directed navigation for agricultural robots using the A2A benchmark, including 1,560 episodes across six farming settings, enhancing navigation success from 0.33 to 0.47 via instruction decomposition (Figure 6B). SUM-AgriVLN [40] included spatial memory to enhance environmental comprehension and navigation. Collectively, these investigations illustrate the capability of VLMs to advance agricultural robotics from fixed trajectories to navigation that is informed by language and context, despite the ongoing difficulties posed by intricate directives, spatial reasoning, and physical manipulation (Figure 6C).

The shift toward fruit-centric embodied intelligence is notably seen in contemporary harvesting research. The occlusion-aware strawberry harvesting research [41] used VLM/LLM reasoning to enhance the sequence of harvesting amidst fruit occlusion, illustrating the significance of semantic thinking prior to manipulation (Figure 6D). HarvestFlex [42] enhanced this idea to include closed-loop VLA regulation in actual greenhouse strawberry harvesting. With just 227 demos (3.71 hours of data), the optimal configuration attained a harvesting success rate of 74.0% with a 4.1% fruit-damage rate, offering significant proof that generic VLA rules may be modified for sensitive agricultural practices (Figure 6E). Nonetheless, observability, contact dynamics, and crop-specific generalization continue to pose significant constraints.

Contemporary systems are progressively integrating logic with physical engagement. Active Obstacle

Separation [43] used RGB-D observations, a scene graph, and VLM reasoning to ascertain the necessary movement of leaves or other impediments prior to strawberry harvesting, resulting in an 83.90% success rate in clearing and an 82.8% success rate in end-to-end picking during real-robot experiments (Figure 6F). Segment and Pick Any Fruit [44] further exemplifies the shift toward language-directed and more universally applicable fruit recognition and collection (Figure 6G), whereas the comprehensive agricultural VLM review [45] highlights the significant promise—and ongoing deficiency—of physically instantiated VLM/VLA systems in the agricultural sector (Figure 6H).

Nonetheless, a notable study deficiency persists: existing agricultural VLM/VLA investigations mostly focus on singular activities, specific crops, or regulated settings instead of offering an integrated multi-crop, multi-task artificial intelligence. Future agricultural practices using Gemini Robotics should thus include multimodal sensing, natural language communication, spatial reasoning, autonomous navigation, skilled manipulation, action validation, and ongoing adaptation within a unified robotic system.

Table 1. Gemini Robotics and related VLM/VLA systems for agriculture.

Year	Reference	Task	Crop/Fruit	Application	Model	Data	Key result	Gap / limitation
	[37]	Manipulation	General	Embodied robot control	Gemini Robotics; Robotics-ER	Multimodal robot demonstrations	Strong generalization, interaction and dexterity	No agricultural validation
	[38]	Reasoning & manipulation	General	Multi-step robot tasks	Gemini Robotics 1.5; ER 1.5	Multi-embodiment robot data	Improved reasoning and motion transfer	No field/crop evaluation
2025	[39]	Navigation	Multiple agricultural environments	Language-guided field navigation	AgriVLN + VLM	A2A; 1,560 episodes	Success rate improved **0.33→0.47**	Difficult long instructions
	[40]	Navigation & memory	Multiple crops	Spatially aware navigation	SUM-AgriVLN + VLM	A2A benchmark	Success rate improved **0.47→0.54**	No manipulation
	[41]	Harvest planning	Strawberry	Picking-order optimization	GPT-4V; Qwen-7B + LoRA	Strawberry images + attributes	Up to **91%** occlusion-aware success	Planning only; limited physical control
	[42]	Harvesting	Strawberry	Closed-loop robotic picking	WALL-OSS	227 episodes; 3.71 h	**74%** harvest success; **4.1%** damage	Single crop; 26% failures
2026	[43]	Obstacle removal	Strawberry	Occlusion clearing + picking	VLM + scene graph	RGB-D; real/simulation trials	**82.8%** end-to-end picking	Crop-specific; limited generalization
	[44]	Detection & harvesting	Tomato, strawberry + other crops	Zero-shot fruit picking	VLM + SAM2	Multiple fruit datasets	Tomato accuracy **95.54%**; multi-crop zero-shot perception	Prompt/occlusion sensitivity

4. AI Evolution and Future Outlook in Fruit Agriculture

Figure 7 shows the evolution of artificial intelligence in fruit agriculture from mechanization and rule-based systems to computer vision, deep learning, precision agriculture, VLM/VLA models, and now embodied robotic intelligence. Historically, sensing, machine vision, UAVs, deep learning, and robotic manipulation have gradually moved fruit production systems from manual observation and decision-making to automated perception and decision-making [46-50]. Recent VLM/VLA developments have further evolved that trend by linking visual perception, language understanding, reasoning, planning, and physical robotic actions that provide the technological basis for more general-purpose agricultural robots.

The 2025–2075 path in Figure 7 should not be viewed as a prediction but as evidence of future imagination. With the continuing development of AI, robotics, multimodal sensing, edge computing, and agricultural automation, fruit production could evolve from AI-driven systems to semi-autonomous systems, multi-task harvesting robots, coordinated multi-robot orchards, and eventually a more autonomous and adaptive ecosystem. This would create better harvesting efficiency, resource use efficiency, product quality and traceability, and reduce labor-related losses and production costs [51-53]. But the pace of the transition will depend on the technological reliability, affordability, safety, farmer adoption, infrastructure, regulation, and the ability of future embodied AI systems to adapt to crops, orchards, seasons, and environments.

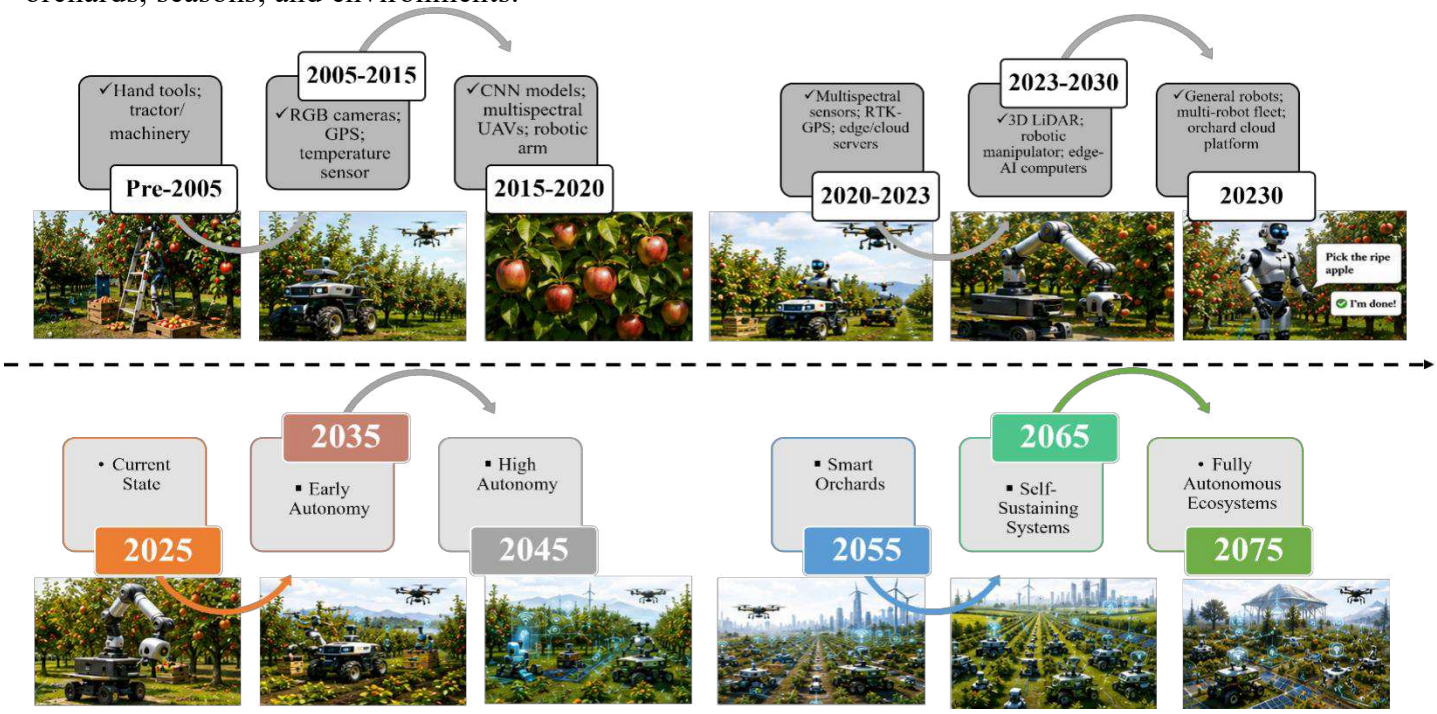


Fig. 7. Historical evolution and prospective 50-year trajectory of AI and embodied robotics in fruit agriculture. The upper panel summarizes the transition from mechanization and early machine vision to deep learning, precision agriculture, VLM/VLA systems, and emerging Gemini Robotics-like embodied intelligence. The lower panel presents an evidence-informed scenario for 2025–2075, illustrating the potential progression from AI-assisted fruit production toward adaptive, multi-robot, and increasingly autonomous orchard ecosystems. Future values and milestones represent prospective scenarios rather than measured predictions.

5. Limitations and Future Research

5.1 Challenges

The robust development of Gemini Robotics and other VLA models, their application in agriculture is still in its early stages (Table 2). Agricultural environments are highly dynamic and less structured than industrial settings. The lighting, rain, dust, wind, uneven terrain, dense vegetation and seasonal changes can significantly affect perception, localization and robotic control. Moreover, contributing to general-purpose embodied models to agriculture requires incorporating RGB, depth, LiDAR, multispectral and environmental data and thus a low latency for real-time physical operations.

Another major challenge is the need for generalization and reliability; agricultural robots are working on different crops, cultivars, farms, seasons and geographic environments and need to be adjusted to them without retraining. A model that can navigate one greenhouse and recognize one crop may not be able to do so when it comes to an unknown orchard or field. More importantly, embodied AI errors can also result in physical errors. If the reasoning is wrong, crop damage, inappropriate spraying, collisions, or human-robot interactions might result. Agricultural Gemini Robotics systems therefore require uncertainty-aware reasoning, collision avoidance, fail-safe control, human intervention and continuous verification of completed actions.

In the fruit industry, there will be various challenges to manipulate these fruits; fruits are often covered by leaves, branches or neighboring fruit and are very distinct in size, shape, color, maturity and orientation. Harvesting, thinning, pruning, bagging and bag removal require accurate three-dimensional localization of biological targets and highly controlled manipulation of them. A robot needs to determine not only where they are, but also whether it is mature, reachable and safe for manipulating that fruit, which approach trajectory should be chosen, and how the mission has gone without damage. Limited agricultural VLA data, computational complexities, and the cost of robotic hardware also limit the commercial adoption.

Future research should move beyond the specialized agricultural robots into general agricultural embodied intelligence. Rather than develop independent systems for navigation, monitoring and spraying, weeding and harvesting in a vacuum, Gemini Robotics-like architecture will have to be able to have a common perception reasoning and action foundation which can adapt to the environment and to the natural language instructions. The development of multi-modal agricultural foundation models based on images, video, depth, sensor measurements, language instructions and robot-action trajectories is particularly key to achieving this.

5.2 Future directions

The fruit industry is now ripe for the development of a multi-task robotic system that can continue to operate in all stages of the production process. Fruit robots might take over fruit detection, maturity assessment, thinning, bagging, bag removal, harvesting, quality inspection, sorting and check its actions

regularly to keep the fruit healthy and adapt to different cultivars and canopy shapes. Tactile sensing, soft grippers, active vision and 3D reasoning could also help to reduce fruit damage and better manipulate the fruit in the case of a serious occlusion.

Future systems should also emphasize edge/on-device intelligence, continual learning and human-robot collaboration. Efficient VLA models running directly on agricultural robots could reduce dependence on cloud connectivity and improve response time. Continual learning would help robots to adapt to new farms, crops and seasons from experience in the field, while natural language interaction would enable farmers to assign tasks, alter operating constraints and intervene when uncertainty is high. Multi-robot cooperation between UAVs, ground vehicles and manipulators could also extend embodied intelligence from individual plants to entire orchards and farms.

In the final analysis, Gemini Robotics' future in agriculture is in moving from "detect and execute" to "observe, understand, reason, decide, act, verify, adapt." This transition would lead to a new breed of safe, logical and generalizable agricultural robots that are capable of understanding the biological environment and human goals and autonomous in handling physical operations.

Table 2. Key challenges and future directions for Gemini Robotics in agriculture and fruit production.

Area	Current challenge	Fruit-specific issue	Future direction
Perception	Variable lighting, weather and occlusion	Fruit–branch–leaf occlusion	Multimodal RGB–depth–spectral perception
Reasoning	Limited agricultural knowledge	Maturity, accessibility and target selection	Agricultural embodied foundation models
Navigation	Uneven and changing environments	Narrow orchard rows and canopy obstacles	Semantic VLA navigation + spatial memory
Manipulation	Unstructured biological targets	Fragile fruits and complex detachment	Soft/tactile manipulation + 3D reasoning
Generalization	Crop/farm/season domain shifts	Different cultivars and canopy systems	Cross-crop and cross-orchard learning
VLA data	Few agricultural action datasets	Limited fruit-manipulation demonstrations	Large-scale agricultural VLA benchmarks
Computing	High latency and energy demand	Real-time picking/bagging requirements	Edge/on-device VLA models
Safety	AI errors become physical actions	Fruit/tree damage and operational risk	Uncertainty-aware, fail-safe control
Human interaction	Limited intuitive supervision	Complex task-specific instructions	Natural-language human–robot collaboration
Multi-task autonomy	Most robots perform one task	Separate systems for bagging/harvesting	General-purpose multi-task fruit robots
Multi-robot systems	Limited coordination	Orchard-scale monitoring/manipulation	UAV–UGV–manipulator collaboration
Adaptation	Models remain relatively static	Seasonal and maturity changes	Continual and lifelong learning
Commercialization	Cost, reliability and maintenance	High cost per harvested fruit	Lightweight, modular and affordable platforms

6. Conclusion

This study underscores the evolution of agricultural robotics from specialized automation to versatile embodied intelligence, particularly focusing on the capabilities of Gemini Robotics and VLM/VLA models in fruit cultivation. Through the amalgamation of multimodal sensing, linguistic comprehension, reasoning, navigation, manipulation, and physical execution, these technologies may facilitate progressively autonomous tasks like fruit surveillance, ripeness evaluation, spraying, bagging, harvesting, and post-harvest management. Nonetheless, contemporary agricultural applications are still in their infancy, facing significant obstacles in field generalization, occlusion, delicate fruit handling, agricultural VLA data, real-time processing, safety, and cost-effectiveness. Future advancements should thus include agricultural foundational models, gentle and tactile handling, edge VLA computing, ongoing learning, and collaborative multi-robot systems. Ultimately, the trajectory of intelligent fruit robotics is anticipated to go beyond basic "detect and execute" mechanisms to robots that can observe, comprehend, reason, make decisions, act, verify, and adapt in intricate academic and agricultural environments.

Funding

This research was supported by Beijing Natural Science Foundation (funding No. IS24089), Project of Fund for Stable Support to Agricultural Sci-Tech Renovation (Project Number: xjnkywdzc-2025003-08-01), and National Modern Agricultural Industry Technology System Guangxi Green Circular Development Innovation Team Project (No. nycytxgxcxtd-20-02).

References

1) LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *nature*, 521(7553), pp.436-444.
2) Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
3) Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T. and Sun, H., 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1), p.3094.
4) Bac, C.W., van Henten, E.J., Hemming, J. and Edan, Y., 2014. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal of Field Robotics*, 31(6), pp.888–911.
5) Silwal, A., Davidson, J.R., Karkee, M., Mo, C., Zhang, Q. and Lewis, K., 2017. Design, integration, and field evaluation of a robotic apple harvester. *Journal of Field Robotics*, 34(6), pp.1140–1159.
6) Rajendran, V., Debnath, B., Mghames, S., Mandil, W., Parsa, S., Parsons, S. and Ghalamzan-E, A., 2023. Towards autonomous selective harvesting: A review of robot perception, robot design, motion planning and control. *arXiv preprint arXiv:2304.09617*.
7) Mhamed, M. and Zhang, Z., 2025. Applications of the Internet of things on agriculture: review and future apple fruit comprehensive automation system. Apple Production Technologies: From Laboratory to Practical Applications: Cutting-Edge Innovative Techniques, pp.45-68.
8) Yu, J., Zhang, Z., Li, Y., Hua, W., Wei, X., Igathinathane, C., Mhamed, M., Zhang, W., Jiao, X., Yang, L. and Zhang, M., 2024. In-field grading and sorting technology of apples: A state-of-the-art review. Computers and Electronics in Agriculture, 226, p.109383.

9) Mhamed, M., Zhang, Z., Yu, J., Li, Y. and Zhang, M., 2024. Advances in apple's automated orchard equipment: A comprehensive research. Computers and Electronics in Agriculture, 221, p.108926.
10) Zhu, Q., Lin, Y., Fu, W., Tang, W., Li, J., Zhang, Y., Li, B., Guo, X., Wang, F., Qi, H. and Sun, C., 2026. Towards clinical-level interpretation of dental panoramic radiography using an instance-guided vision-language model. *Nature biomedical engineering*, pp.1-17.
11) Tangtartharakul, G. and Storrs, K.R., 2026. Visual language models show widespread visual deficits on neuropsychological tests. Nature Machine Intelligence, 8(2), pp.209-219.
12) Li, X., Li, P., Qian, L., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Wang, X., Guo, D. and Kong, T., 2026. What matters in building vision–language–action models for generalist robots. Nature Machine Intelligence, 8(2), pp.158-172.
13) Zhu, H. et al., 2025. Harnessing large vision and language models in agriculture. *Frontiers in Plant Science*.
14) Gibney, E., 2015. DeepMind algorithm beats people at classic video games. *Nature*, 518(7540), pp.465-466.
15) Gemini Robotics Team et al., 2025. Gemini Robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*.
16) Hayes-Roth, B., 1995. An architecture for adaptive intelligent systems. *Artificial intelligence*, 72(1-2), pp.329-365.
17) Zhao, X., Lyu, X., Chen, X. and Li, X., 2026, July. Mde-agrivln: Agricultural vision-and-language navigation with monocular depth estimation. In International Conference on Intelligent Computing (pp. 500-511). Singapore: Springer Nature Singapore.
18) Cuaran, J., Koe, K., Potnis, A., Uppalapati, N.K. and Chowdhary, G., 2026. Visual-language-guided task planning for horticultural robots. Computers and Electronics in Agriculture, 252, p.112058.
19) Edan, Y., Han, S. and Kondo, N., 2009. Automation in agriculture. Springer handbook of automation, pp.1095-1128.
20) Jha, K., Doshi, A., Patel, P. and Shah, M., 2019. A comprehensive review on automation in agriculture using artificial intelligence. Artificial Intelligence in Agriculture, 2, pp.1-12.
21) Abdolmaleki, A. et al., 2025. Gemini Robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*.
22) Zhao, X., Lyu, X. and Li, X., 2025. AgriVLN: Vision-and-language navigation for agricultural robots. *arXiv preprint arXiv:2508.07406*.
23) HarvestFlex authors, 2026. HarvestFlex: Strawberry harvesting via vision-language-action policy adaptation in the wild. *arXiv preprint arXiv:2603.05982*.
24) Siegwart, R., Nourbakhsh, I.R. and Scaramuzza, D., 2011. *Introduction to autonomous mobile robots*. MIT press.
25) Parada, C., 2025. Introducing Gemini Robotics and Gemini Robotics-ER, AI models designed for robots to understand, act and react to the physical world. *Google DeepMind. Retrieved March*, *13*, p.2025.
26) Vougioukas, S.G., 2019. Agricultural robotics. *Annual review of control, robotics, and autonomous systems*, *2*(1), pp.365-392.
27) Comparetti, A., Fagiolini, A., Fountas, S. and Cascio, V., 2025. Field robots for precision agriculture. *AIMS Agriculture and Food*, *10*(4), pp.885-916.
28) Almtireen, N. and Ryalat, M., 2026. Robotics in precision agriculture: Task-, platform-, and evaluation-oriented review. *Robotics*, *15*(4), p.81.
29) Oliveira, L.F., Moreira, A.P. and Silva, M.F., 2021. Advances in agriculture robotics: A state-of-the-art review and challenges ahead. *Robotics*, *10*(2), p.52.
30) Murad, N.Y., Mahmood, T., Forkan, A.R.M., Morshed, A., Jayaraman, P.P. and Siddiqui, M.S., 2023. Weed detection using deep learning: A systematic literature review. *Sensors*, *23*(7), p.3670.
31) Wu, H., Wang, X., Chen, X., Zhang, Y. and Zhang, Y., 2025. Review on key technologies for autonomous navigation in field agricultural machinery. *Agriculture*, *15*(12), p.1297.
32) Navarro, E., Costa, N. and Pereira, A., 2020. A systematic review of IoT solutions for smart farming. *Sensors*, *20*(15), p.4231.

33) Xu, R. and Li, C., 2022. A modular agricultural robotic system (MARS) for precision farming: Concept and implementation. *Journal of Field Robotics*, *39*(4), pp.387-409.
34) Sim, M.Y., Zhang, W.E., Dai, X. and Fang, B., 2025, July. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 24452-24470).
35) Li, Z., Wu, X., Du, H., Nghiem, H. and Shi, G., 2025. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, *1*(1).
36) Zhang, D., Sun, J., Hu, C., Wu, X., Yuan, Z., Zhou, R., Shen, F. and Zhou, Q., 2025. Pure vision language action (vla) models: A comprehensive survey. *arXiv preprint arXiv:2509.19012*.
37) Wilk-Jakubowski, G., Harabin, R. and Ivanov, S. (2022) 'Robotics in crisis management: A review', *Technology in Society*, **68**, 101935.
38) Bac, C.W., van Henten, E.J., Hemming, J. and Edan, Y., 2014. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal of Field Robotics*, 31(6), pp.888–911.
39) Silwal, A., Davidson, J.R., Karkee, M., Mo, C., Zhang, Q. and Lewis, K., 2017. Design, integration, and field evaluation of a robotic apple harvester. *Journal of Field Robotics*, 34(6), pp.1140–1159.
40) Zhao, X., Lyu, X. and Li, X., 2025. SUM-AgriVLN: Spatial understanding memory for agricultural vision-and-language navigation. *arXiv preprint*, arXiv:2510.14357.
41) 2025. Occlusion-aware harvesting sequence planning for strawberries in clusters using large language and vision-language models. *Preprint*.
42) Fan, Z., Su, X., Lu, Y., Fang, Y., Gao, Z., Feng, J. and Hu, Y., 2026. Segment and pick any fruit: Text-prompted robotic harvesting. *Pattern Recognition*, 179, 113836. doi:10.1016/j.patcog.2026.113836.
43) 2026. Active obstacle separation: Vision-language model-driven clearing decisions for robotic harvesting. *Research article*.
44) 2026. Segment and pick any fruit: Text-prompted robotic harvesting. *Research article*.
45) Zhu, H., Qin, S., Su, M., Lin, C., Li, A. and Gao, J., 2025. Harnessing large vision and language models in agriculture: A review. *Frontiers in Plant Science*, **16**, 1579355. doi:10.3389/fpls.2025.1579355.
46) Mhamed, M., Zhang, Z. and Zhang, M., 2026. Global wheat disease identification via AI technologies: An in-depth survey with sustainability and future projections. Computers and Electronics in Agriculture, 251, p.112042.
47) Zhang, X., Liu, J., Zhang, Z., Dong, X., Zhang, P. and Zhang, Z., 2026. Technology Evolution of Natural Rubber Harvesting Mechanization.
48) Mhamed, M., Zhang, Z. and Zhang, M., 2025. Monitoring corn growth with disease identification and yield prediction via advanced intelligent architecture. Journal of Industrial Information Integration, p.100922.
49) Zhang, T., Mhamed, M., Zhang, Q., Yang, L., Xiaohui, Z.H.A.O., Haiyan, G. and Zhang, Z., 2025. Apple varieties, diseases, and distinguishing between fresh and rotten through deep learning approaches. Plos one, 20(5), p.e0322586.
50) Mhamed, M., Zhang, Z., Hua, W., Yang, L., Huang, M., Li, X., Bai, T., Li, H. and Zhang, M., 2025. Apple varieties and growth prediction with time series classification based on deep learning to impact the harvesting decisions. Computers in Industry, 164, p.104191.
51) Nath, S. and Palmer, T., 2026. Can AI models reliably forecast extreme weather events?. Nature, 651(8106), pp.583-584.
52) Soliman, A., 2024. DeepMind AI weather forecaster beats world-class system. Nature, 636(8042), pp.282-283.
53) Chen, C., Jost, M., Outram, M.A., Friendship, D., Chen, J., Wang, A., Periyannan, S., Bartoš, J., Holušová, K., Doležel, J. and Zhang, P., 2023. A pathogen-induced putative NAC transcription factor mediates leaf rust resistance in barley. Nature communications, 14(1), p.5468.