APPLICATION OF LLM + ZERO-SHOT LARGE MODELS FOR FRUIT OBJECT DETECTION
/
应用 LLM + Zero-Shot 大模型实现水果目标检测

Yingdong QIN [1,2,3,*], Haoyu SONG [1,2,3], Jingyi LI [4], Keyao WEN [1]

[1] College of Artificial Intelligence and Low-Altitude Technology, South China Agricultural University, Guangzhou 510642, China;
[2] National Center for International Collaboration Research on Precision Agricultural Aviation Pesticides Spraying Technology (NPAAC), South China Agricultural University, Guangzhou 510642, China;
[3] Center for International Cooperation and Disciplinary Innovation of Precision Agricultural Aviation Applied Technology ('111 Center'), South China Agricultural University, Guangzhou 510642, China;
[4] School of Statistics and Data Science, Guangdong University of Finance & Economics, Guangzhou 510220, China.
*Correspondent author: Qin Yingdong, E-mail: qinyingd@foxmail.com, ORCID: 0009-0008-3403-0554
DOI: https://doi.org/10.35633/inmateh-79-11

Keywords: Fruit Recognition, LLM, Vision Transformer, Zero-Shot, Smart Agriculture

ABSTRACT

Efficient and flexible agricultural image annotation is crucial for intelligent crop monitoring in smart agriculture, yet conventional detection models are limited by fixed class labels and require extensive manual annotations. This study presents a zero-shot annotation framework that integrates OWLv2, Google's second-generation open-vocabulary vision model, with large language models (e.g., GPT-3.5, DeepSeek V1) to enable multilingual, natural language-driven fruit recognition in smart agriculture. A user-friendly interface was developed to support individual or batch image annotation with adjustable sensitivity to meet diverse field requirements. Experimental evaluations demonstrated the framework's strong generalizability and semantic understanding capabilities, allowing recognition of unseen fruit categories and attributes such as ripeness or color. The system significantly reduces annotation time and labor costs, while enhancing accessibility through natural language interaction. To ensure a robust evaluation of generalizability, a cross-domain protocol was employed using a novel dataset from 2025. Results showed that OWLv2 achieved an F1-score of 0.80 and an mAP of 0.8301, significantly outperforming the pre-trained YOLO11 (F1: 0.74, mAP: 0.60) in zero-shot scenarios. OWLv2 exhibited superior flexibility and required no task-specific dataset retraining, although its computational demands remain higher than lightweight models like YOLO11. Notably, while the LLM (DeepSeek) introduced a total one-time API latency of 598.3 ms (called only once for processing multiple images). the actual core computational latency of OWLv2 was only 257.7 ms per image. Despite a total processing time of 891.9 ms (including visualization output), the framework demonstrates superior recall (0.9080) and semantic flexibility without retraining. These results verify the enormous application potential of OWLv2 and similar zero-shot models in agriculture, providing scalable solutions for automated annotation, real-time monitoring, and large-scale data collection.

摘要

在智慧农业领域，传统农业领域图像目标检测模型受限于固定的类别标签，并依赖大量人工标注。本研究提出了一种可以语言交互的零样本（zero-shot）标注框架，将 Google 第二代开放词汇视觉模型 OWLv2 与大语言模型（GPT-3.5、DeepSeek V1）相结合，以实现面向智慧农业的多语言、自然语言驱动的果实识别。为此，我们开发了一个交互界面，支持单张或批量图像标注，并可根据实际田间需求调节检测灵敏度。实验评估结果表明，该框架具备优异的泛化能力与语义理解能力，能够识别未曾在训练集中出现的果品类别及其属性（如成熟度、颜色等）。基于 2025 年果园数据集的跨域评估表明，OWLv2 的 F1 分数（0.80）与 mAP（0.8301）显著优于零样本场景下的预训练 YOLO11（F1: 0.74, mAP: 0.60）。相较于传统模型，OWLv2 展现出更强的灵活性，且无需针对特定任务重新训练数据集；时延分析显示，LLM（DeepSeek）仅引入 598.3 ms 的单次调用延迟（多张图只调一次），OWLv2 的核心计算耗时仅为 257.7 ms/幅。尽管包含图像输出的总处理时间为 891.9 ms，但该框架在无需重训的情况下展现了极高的召回率（0.9080）与语义灵活性。上述结果验证了 OWLv2 及同类零样本模型在农业领域具有为自动化标注、实时监测及大规模数据采的巨大潜力。

INTRODUCTION

In recent years, the emergence of transformer-based architectures has significantly advanced the field of computer vision, with growing implications in agricultural automation and precision farming. One notable breakthrough was Google's Vision Transformer (ViT), which introduced the concept of dividing images into fixed-size patches (e.g., 16×16) and treating them as sequential data, similar to words in natural language processing (*Dosovitskiy et al., 2020*). This approach enabled visual models to benefit from the scalability and learning efficiency of transformer networks, paving the way for a wide range of vision tasks with reduced reliance on task-specific architectures.

Following ViT, transformer-based vision models have evolved rapidly. Notably, OpenAI's CLIP model (*Radford et al., 2021*) demonstrated that joint vision-language pretraining on web-scale data can result in powerful zero-shot capabilities. Similarly, research from institutions such as Google and NVIDIA (*Xu J. et al., 2023*) has explored large-scale multimodal models for open-vocabulary learning, where systems are trained not on fixed class labels but on flexible language prompts. These developments are highly relevant to agriculture, where the diversity of crops, growing conditions, and field scenarios presents a challenge to traditional classification-based systems.

To address the need for more adaptable models, Google introduced the Open-Vocabulary Learner (OWL) framework. The first generation, OWL-ViT (*Minderer et al., 2022*), combined ViT backbones with vision-language pretraining in the style of CLIP, enabling zero-shot object detection based on textual input. However, it still relied heavily on annotated data for downstream adaptation. In response, OWLv2 was released in 2023 (*Minderer et al., 2023*), introducing a self-labelling mechanism called OWL-self. This technique allows the model to automatically generate pseudo-annotations via bounding box proposals, which are then refined through loss computation and further model updates. The concept aligns with the iterative self-supervision loop proposed by Meta AI in its Segment Anything (SA) model (*Kirillov et al., 2023*), which reduces dependence on manual labels by enabling scalable training on unlabeled image datasets.

Compared to chemical methods for fruit detection, image-based methods are more intuitive (*Qin et al., 2023*). *Wei et al. (2023)* verified that imaging techniques enhance the accuracy of conventional sensing methods for food detection. In smart agriculture, open-vocabulary models excel in dynamic environments like orchards or greenhouses, enabling tasks such as fruit recognition, disease monitoring, and crop growth assessment (*Espinoza et al., 2023; Mimma et al., 2022; Zhang et al., 2025*). *Li et al. (2025)* developed an improved YOLOv8n model for strawberry detection, which achieved high precision and lightweight deployment but still relied on labeled fruit datasets and fixed class definitions. *Tang et al. (2025)* proposed an enhanced YOLOv11n model for kiwifruit flower detection under occlusion, demonstrating strong performance in complex orchard environments but requiring task-specific training data. Their work underscores the value of open-vocabulary zero-shot models that support flexible, unannotated fruit and flower recognition. *Tan et al. (2021)* early explored the zero-shot paradigm but were hindered by limited datasets, methods, and computational resources, resulting in unsatisfactory performance.

In this study, the application of OWLv2 to agricultural image segmentation is investigated, with particular emphasis on fruit recognition and annotation under diverse field conditions. OWLv2 was locally deployed using Google's open-source implementation and integrated with the ChatGPT 3.5 API to enhance natural language interaction capabilities. Systems incorporating large language models (LLMs), such as ChatGPT, generally exhibit improved usability and broader user acceptance in practical applications (*Qing et al., 2023*). The proposed system enables users to submit free-form agricultural queries, ranging from crop names to complex descriptions such as "ripe citrus under leaves", and generates corresponding segmentation masks in a zero-shot manner. In addition, a user-friendly interface with configurable parameters was developed to support both single-image and batch annotation workflows.

MATERIALS AND METHODS

Experimental Environment and Hardware

The proposed framework was implemented using Python 3.7 on an Ubuntu 22.04 LTS workstation. To ensure reproducibility, all visual inference tasks were executed on a consistent hardware setup: an Intel Core i7-10700K CPU, 32 GB of DDR4 RAM, and an NVIDIA GeForce RTX 3060 GPU with 12GB of VRAM. The OWLv2 model (google/owlv2-base-patch16) was deployed via the Hugging Face Transformers library.

Logic Architecture and Feature Matching Mechanism

This section describes the deployment of OWLv2 in combination with GPT-3.5 or DeepSeek V1 for open-vocabulary fruit image annotation in an agricultural context. The method involves four key stages: (1) environment setup and model loading, (2) query preprocessing via GPT-3.5 or DeepSeek V1, (3) image inference and prediction using OWLv2, and (4) postprocessing and visualization. The technical schematic of the developed zero-shot annotation framework with language interaction capability is shown in **Fig. 1**.

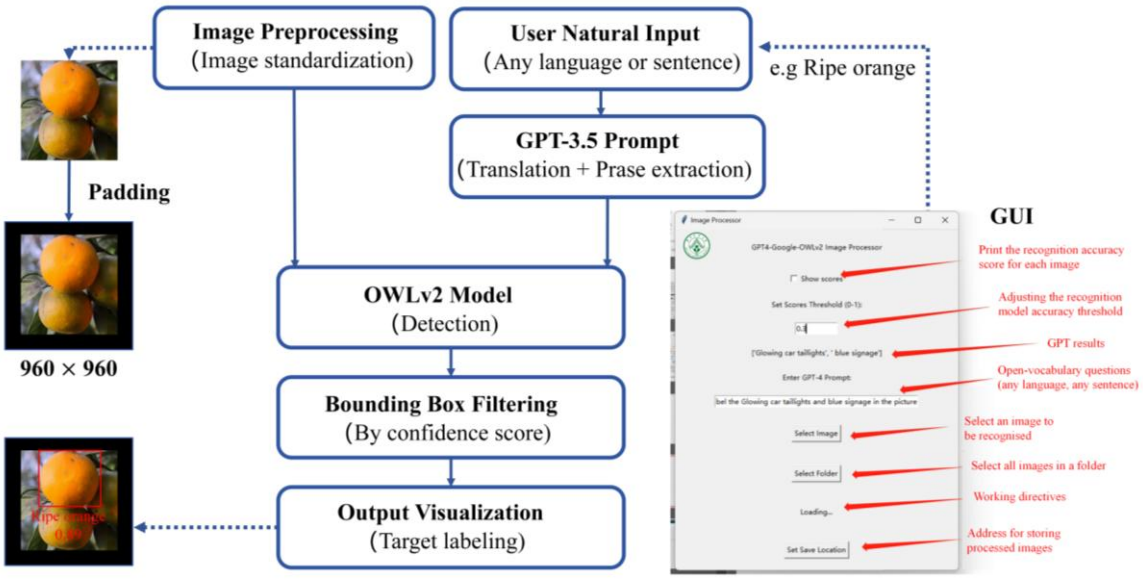

Fig. 1 - Zero-Shot Models for Fruit Object Detection Steps

Since it is based on Transformer architecture, the library is housed in the public repository NielsRogge/Transformers-Tutorials on GitHub*, where the OWLv2 model can be located. Additionally, the SAM model is available for those interested, despite its distinct functionality, exhibiting similar characteristics during application. First and foremost, the installation of the library is crucial.

Subsequently, the workflow involves loading the OWLv2 model, processor, and input image. After completing these steps, a query vocabulary must be defined for object retrieval. In the original OWLv2 framework, queries are required to be provided as comma-separated vocabulary terms in string format, which limits usability and flexibility. To improve user interaction, the proposed system directly integrates the outputs of ChatGPT 3.5 or DeepSeek V1 through a predefined system prompt. First, the user input is translated into English and converted into concise noun phrases representing the target objects. Several examples are included in the prompt design to facilitate few-shot learning by the large language model (LLM). During inference, the user input is first processed by ChatGPT or DeepSeek, and the generated output is subsequently used as the query input for OWLv2. This strategy enables the system to accept arbitrary sentence structures and multilingual inputs while maintaining compatibility with the OWLv2 query format. To ensure deterministic keyword extraction, the LLM API temperature was fixed at 0.0, and the prompt design followed a strict JSON-like few-shot template, as illustrated in **Tab. 1**.

The core of the zero-shot detection lies in the similarity matching between image tokens and text queries. To eliminate mathematical redundancy, the similarity score $S_{i,j}$ is calculated as the normalized dot product (cosine similarity) between the visual and textual embeddings:

$$S_{i,j} = \frac{f_{image}(x_i) \cdot f_{text}(q_j)}{\|f_{image}(x_i)\| \|f_{text}(q_j)\|} \quad (1)$$

where: x_i denotes the i-th visual token, q_j is the j-th text query embedding, f_{image} and f_{text} represent the respective encoders. This score determines the confidence level for each bounding box proposal.

Subsequent steps involve the local deployment of OWLv2, including image standardization through padding to achieve a fixed shape and size (torch size of 960 x 960). After the forward pass, the model's output is obtained. Finally, coordinate annotation and visualization are carried out based on the model's output, accompanied by the labeling of their similarities for later adjustments to the model's sensitivity.

*: NielsRogge/Transformers-Tutorials: https://github.com/NielsRogge/Transformers-Tutorials

Table 1

System prompt used for keyword extraction

Prompt
You are a keyword extraction machine. Your main responsibility is to extract key target words and determine the action objects of the questioner. For instance, when statements like 'I want to find an apple', 'Search for bananas', or 'Where is the watermelon' are given, the targets for the questioner are the apple, bananas, and watermelon, respectively. Only output the object, including object phrases. If there are multiple objects, separate them with commas. If the input is in another language, it is considered to be English. Use English output only. Cannot output the word 'picture'. Regardless of the syntax used by the questioner, your answer can only be an object (a noun or noun phrase with modifying qualifiers, if there are no modifiers only nouns are output). Preferably output modified nouns, e.g., 'girl in a skirt' instead of 'skirt' and 'girl'

Evaluation Protocol and Metrics

To rigorously validate the system's performance beyond a simple demonstration, a cross-domain evaluation protocol was established.

Baseline Model Dataset (YOLO11): To provide a robust benchmark, the YOLO11 model (Released in late 2024) was pre-trained on a specific agricultural dataset consisting of 1,217 RGB images of oranges (1,017 for training and 200 for validation). These images were captured between 2024 and 2025 across various locations in Guangdong, China, covering diverse environmental challenges such as standard lighting, minor occlusions, and fruit clustering.

Test Dataset Definition: A test set of 50 high-resolution images was independently collected in 2025 at a citrus orchard in Sihui, Guangdong. These images include complex variables such as heavy occlusion, variable lighting (noon vs. evening), and fruit clustering.

Quantitative Metrics: The detection performance is quantified using standard computer vision metrics:

1. Precision and Recall: To evaluate accuracy and coverage.
2. F1-Score: The harmonic mean of precision and recall.
3. mAP (mean Average Precision): Calculated at an IoU (Intersection over Union) threshold of 0.5.
4. Inference Latency: Measured in milliseconds (ms) to assess real-time potential.

Ground Truth: Manual annotation was performed by two researchers using the LabelImg tool to provide the benchmark for Comparison with the OWLv2 outputs.

RESULTS

Q&A-annotation Tool

A graphical user interface (GUI) for annotation was developed using Tkinter, a lightweight Python library. The tool supports common image formats, including *.jpg* and *.png*, and allows users to process either individual images or image folders in batch mode. User input is provided through a simple text input box, while the corresponding output information is displayed directly within the interface, as shown in **Fig. 1**. The main functional parameters of the tool are summarized in **Tab. 2**.

Table 2

Q&A-Semantic Annotation Tool Parameters

Performance indicators	parameter
Single Operation Time	1~2 s
Batch Operation Time (30 sheets)	~25 s per sheet
Startup Time	~20 s
Supported File Formats	.jpg, .png, etc.
Interaction Languages	English, Chinese, French, German, Spanish, etc.
Semantic Conversion Tools	ChatGPT, DeepSeek
LLM API Latency	~1 s

The batch processing capability of our program is demonstrated in **Fig. 2**. This function enables rapid and efficient annotation of large image datasets with high accuracy, significantly reducing the time and labor costs associated with manual labeling.

As noted in Google's documentation, such automated annotation avoids errors caused by fatigue during repetitive labeling tasks. The source code of the program is publicly available in the supplementary materials.

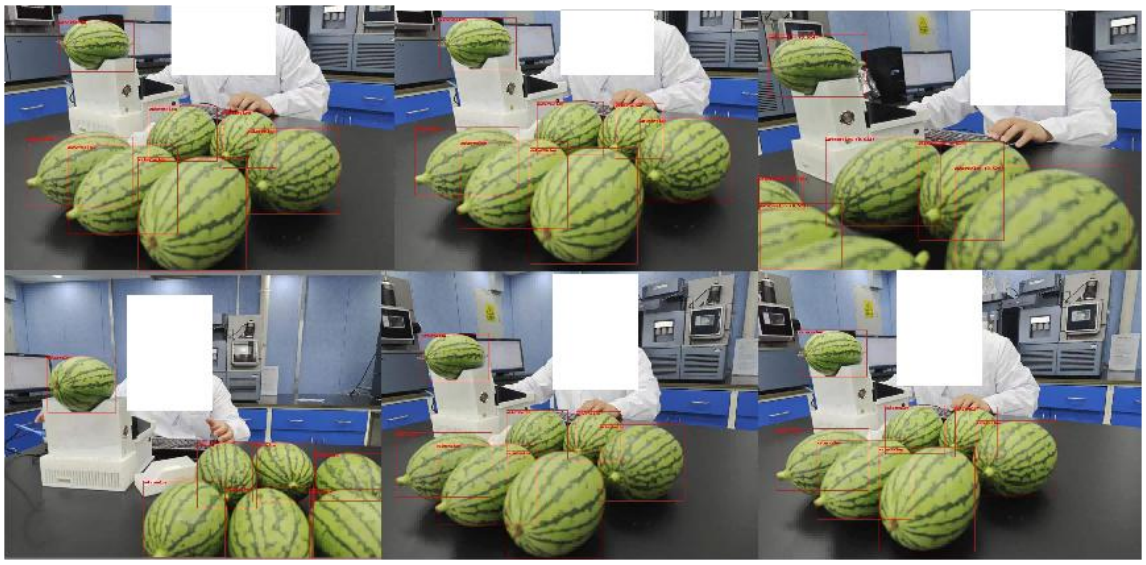

Fig. 2 - Batch Labeling of Watermelon Targets

Understanding of Word Semantics and Generalizability

Compared to traditional object detection models, OWLv2 demonstrates the ability to interpret semantic descriptors, enabling it to analyze attributes such as fruit ripeness without additional fine-tuning. As shown in **Fig. 3**, the model correctly inferred the "ripe" status of Shatangju mandarins produced in Sihui, Guangdong, China. In **Fig. 3b**, the AI-generated maturity scores are visually presented and align closely with human perception, although some misclassifications remain. These can be filtered out using thresholding strategies as demonstrated in **Fig. 3c**.

It is hypothesized that such human-like image-semantic understanding arises from the model's interpretation of not only color features (e.g., the word "yellow" used as a prompt did not strictly correspond to the model's assigned maturity scores) but also more comprehensive features such as orange, red, or green hues, as well as texture information, as indicated in the model's official documentation. Moreover, the proposed open-ended, natural language querying approach facilitates voice-based usage by farmers and enhances accessibility for farmers speaking different languages worldwide. Apart from maturity assessment, citrus images also hold significant potential for semantic applications in pest and disease monitoring (*Zongyin et al., 2025*).

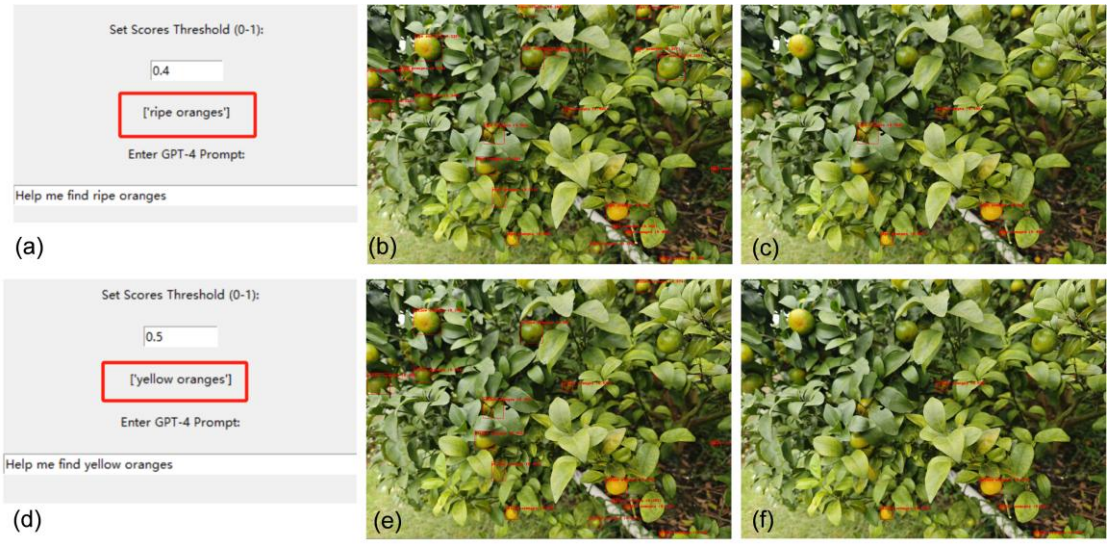

Fig. 3 - Semantic understanding ability.

(a) Questioning and extracting vocabulary (the meaning of ripe fruit); (b) The threshold score is 0.25; (c) The threshold score is 0.4; (d) Questioning and extracting vocabulary (the meaning of fruit color); (e) The threshold score is 0.35; (f) The threshold score is 0.5.

Differences from other model applications

Fig. 4 compares OWLv2 with two widely used publicly available models on Hugging Face, using both network resource images and our own dataset. Targets with a screening score greater than 0.5 were marked for evaluation. OWLv2 was more aggressive in detecting oranges and identified a greater number of fruits overall, though it also produced some misclassifications. It is important to note that these comparison models are not zero-shot models; they require training with annotated datasets. Despite this, OWLv2, trained on a billion-level text-image dataset with strong computational support, demonstrated comparable or even superior performance in certain tasks. While general-purpose models remain effective for simple detection tasks, integrating OWLv2 with APIs such as GPT or DeepSeek further lowers the barrier to its practical application.

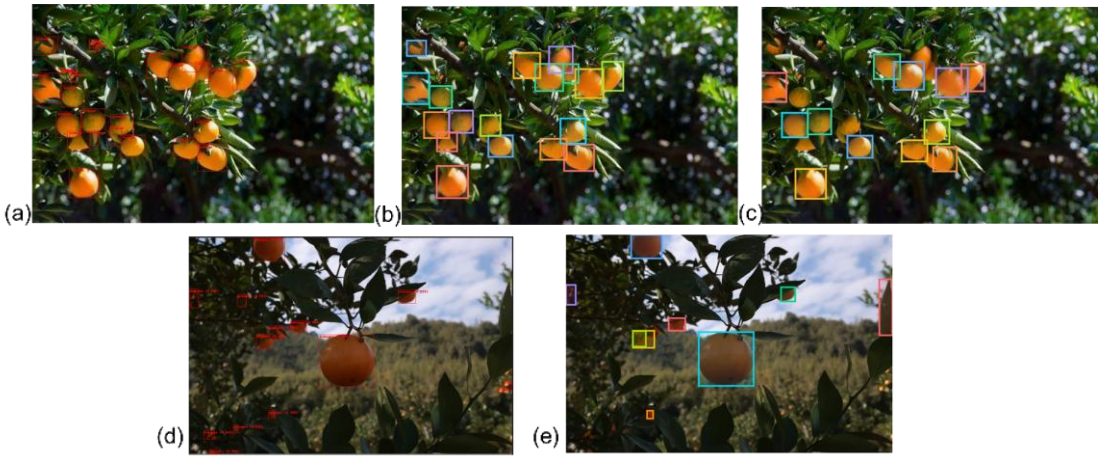

Fig. 4 - Comparison of annotation capabilities across models

(a) Google/owlv2 (network diagram); (b) Facebook/detr-resnet-50 (network diagram); (c) Hustvl/yolos-tiny (network diagram); (d) Google/owlv2 (real-world image); (e) Facebook/detr-resnet-50 (real-world image).

Comparative Evaluation of Zero-shot Foundation Model and YOLO11 for Cross-Domain Object Detection

To benchmark local deployment, zero-shot OWLv2 was compared with a pre-trained YOLO11 model. To prioritize generalization, a cross-domain evaluation was conducted using a novel, independent dataset. This protocol ensures a fair comparison by testing YOLO11 on unseen, heterogeneous data rather than its original training distribution, matching the zero-shot conditions of OWLv2. On its internal validation set, the pre-trained YOLO11 achieved a Precision of 0.9250, a Recall of 0.8333, an F1-score of 0.88, and an mAP of 0.9199. For the comparative test, 50 representative images were selected as the common test set for both the zero-shot OWLv2 and the pre-trained YOLO11 models.

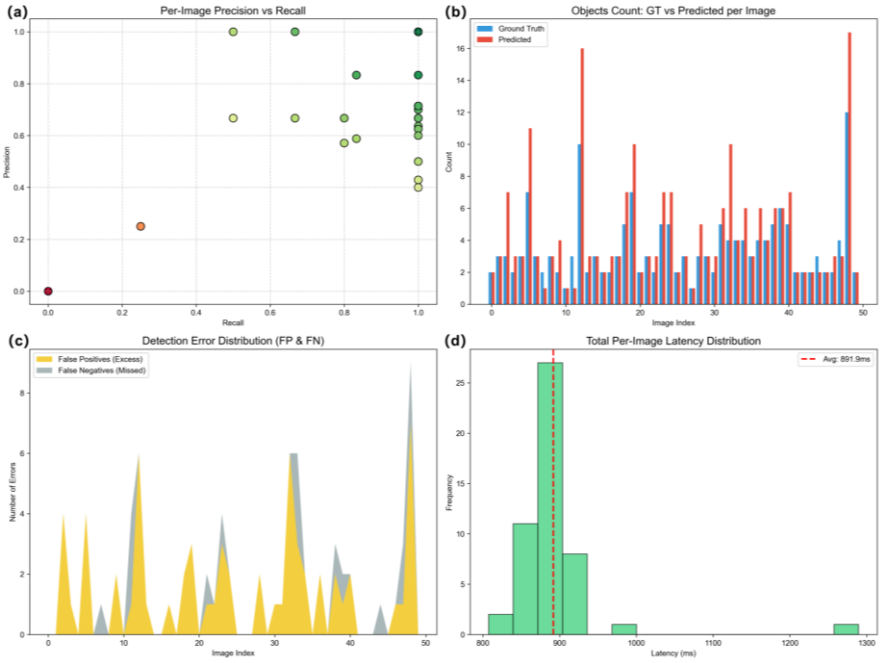

Fig. 5 - Comprehensive performance evaluation of the zero-shot OWLv2 model

(a) PR scatter plot showing per-image precision and recall distribution; (b) Object count consistency between GT and predicted labels; (c) Error distribution identifying fluctuations in FP and FN; (d) Latency distribution with a mean processing time of 891.9 ms.

Table 3

Performance comparison between zero-shot OWLv2 and YOLO11 on the orange dataset

Method	TP	FP	FN	Precision	Recall	F1-Score	mIoU	mAP	Avg. Computation Latency (ms)
OWLv2	158	63	16	0.7149	0.908	0.8000	0.8482	0.8301	257.7
YOLO11	120	27	54	0.8163	0.6897	0.7477	0.8438	0.6055	116.2

The performance of OWLv2 is evaluated from four dimensions **(Fig. 5)**. As illustrated in the Per-Image Precision vs. Recall plot, a significant majority of the data points are clustered in the high-recall and high-precision region (upper right), demonstrating the model's robustness. The Objects Count comparison reveals a high degree of consistency between the Ground Truth (GT) and predicted labels across varying object densities. Regarding error analysis, the Detection Error Distribution suggests that False Positives (FP) constitute the primary source of error compared to False Negatives (FN), indicating a slight tendency towards over-detection in cluttered environments. Furthermore, the Latency Distribution analysis confirms the model's computational efficiency; the total per-image latency follows a near-normal distribution with an average of 891.9 ms. In fact, this is because the program additionally outputs annotated images, while the actual computation latency is only 257.7 ms (**Tab. 3**) and Average DeepSeek latency (Run once before the computational procedure) is only 598.3 ms.

Although the original YOLO validation set achieved a Precision of 0.9250 and an F1-score of 0.88, validation performance typically exceeds that of the test set. Upon switching to the specific test dataset, YOLO11 yielded a Precision of 0.8163 and an F1-score of 0.74, which are considered reasonable within this context. However, owing to a more aggressive detection strategy, OWLv2 significantly outperformed YOLO in terms of Recall (0.9080 vs. 0.6897) while maintaining a respectable Precision of 0.7149. Consequently, OWLv2 achieved a superior F1-score of 0.80. While both models exhibited comparable Mean Intersection over Union (mIoU) values (0.8482 and 0.8438, respectively), OWLv2 demonstrated superior robustness with a substantially higher mAP (0.8301 vs. 0.6055), making it more suitable for detection tasks in complex environments.

Error Analysis of Zero-Shot Foundation Model Annotation via Local Deployment

A comparative review of the OWLv2 outputs against the manually annotated dataset reveals that human annotators often overlook small or partially concealed targets due to visual fatigue or resolution constraints. Consequently, traditional supervised training sets may not fully encompass all ground-truth oranges present in a scene. The zero-shot foundation model consistently identifies minute targets that are only discernible upon magnification, highlighting its significant utility for pre-annotation workflows. This suggests a paradigm shift where foundation models perform automated labeling, leaving humans to act solely as verifiers to mitigate the experiential errors commonly introduced during manual dataset construction.

Detailed manual verification of the detection errors indicates that false positives are primarily driven by extreme texture variations resulting from lighting and shadow fluctuations. Specifically, overexposure (pure white) and deep shadows (black) can lead the model to fragment a single fruit into 2 or 3 separate detections (e.g., as shown on the **Fig. 6** left, where one orange is misidentified as two).

Fig. 6 - Comparison of Zero-Shot OWLv2 and Manual Annotation

(a) Annotations generated by the foundation model under zero-shot conditions;
(b) Standard manual annotations derived from the laboratory's existing dataset.

Discussion

The performance of the zero-shot model demonstrates a "detect-at-all-costs" strategy, prioritizing comprehensive recall over strict precision. This "detect-all" capability is effective not only for close-range operations but also for long-distance path planning by identifying distant fruits in advance. *Jin (2020)* applied visual servo control to agricultural picking robots, emphasizing accurate target positioning as critical for automated harvesting but limited by conventional image detection pipelines. In practice, orange localization was also achieved to assist grasping using traditional object recognition schemes in the previous work, but the factor that most significantly affected operational efficiency was unrelated to fruit coordinate positioning *(Lan et al., 2026).* In this study, switching the detection framework from YOLO11 to OWLv2 only increased the average inference time by approximately 140 ms.

Beyond validating its basic feasibility, the primary advantage of such foundation models lies in their rich semantic information. For instance, the model can theoretically analyze crop growth stages based on prompts like "ripe oranges". However, since the foundation model is pre-encapsulated, the exact semantic threshold for concepts like "maturity" remains a "black box" during API calls. Because both semantic prompts and image features are likely abstracted into high-dimensional vectors, the internal understanding of these concepts can only be calibrated through iterative adjustments of the confidence scores.

In this study, the integration of the Segment Anything Model (SAM) was investigated, as illustrated in **Fig. 7**. SAM offers general image segmentation based on object boundaries, complementing OWLv2's semantic recognition. While SAM does not always outperform OWLv2, their combination could enable instance detection and area measurement for applications such as yield estimation and disease monitoring. Future versions like SAM2 may also support video stream segmentation for real-time use in agricultural machinery or UAVs.

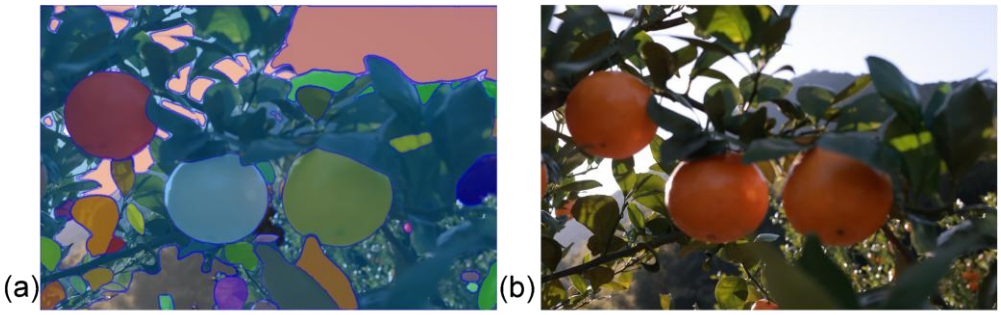

Fig. 7 - Using the SAM model to segment oranges
(a) segmentation image; (b) visible light image.

Moreover, OWLv2 shows certain advantages over models like YOLO, as it requires no annotated datasets or retraining to recognize new classes or descriptive attributes. This makes it suitable for backend decision-making, semantic analysis, and rapid dataset annotation. Its ability to interpret terms like “ripe” or color variations via natural language is valuable in diverse crop scenarios. However, for real-time orchard operations requiring high-speed and lightweight models, YOLO remains more appropriate due to its faster inference and computational efficiency.

CONCLUSIONS

This work presented a zero-shot fruit recognition and annotation tool by integrating OWLv2 with GPT or DeepSeek to support multilingual, natural language queries. The system showed strong generalizability and semantic understanding, allowing recognition of unseen fruit categories and descriptive attributes without retraining. Its ability to interpret complex human language inputs makes it practical for deployment in diverse agricultural environments. Cross-domain experimental results verified that the framework possesses superior generalization capabilities compared to traditional supervised models like YOLO11 when encountering unseen environments. Future work will focus on integrating segmentation models and optimizing inference speed to meet the real-time requirements of smart agriculture.

ACKNOWLEDGEMENT

This study was funded by The Special Fund for Science and Technology Innovation Strategy of Guangdong province (the "Climbing Plan"), [pdjh2025bc042], the University Students' Innovation and Entrepreneurship Training Program [S202510564054S], the ”111 Center” [D18019].

DATA AVAILABILITY

Code and data are available upon request from the corresponding author.

DECLARATIONS OF INTEREST

The authors declare no competing financial interest.

REFERENCES

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010*.11929.

[2] Espinoza, S., Aguilera, C., Rojas, L., & Campos, P. G. (2023). Analysis of fruit images with deep learning: a systematic literature review and future directions. *IEEE Access*, 12, 3837-3859.

[3] Jin, Y. (2020). Recognition technology of agricultural picking robot based on image detection technology. *INMATEH-Agricultural Engineering*, 62(3).191-200.

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. *arXiv preprint arXiv:2304*.02643.

[5] Lan, Y. B., Bian, Z. H., Qin, Y. D., et al. (2026). Current status and future prospects of agricultural drones in multi-scenario application in China. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, x(x), 611-627. doi: 10.11975/j.issn.1002-6819.202510102.

[6] Lan, Y., Song, H., Wang, Y., Liu, X., Li, X., & Li, Y. (2026). Collision-free grasp path planning for citrus-picking manipulator based on ALH-SAC algorithm. Nongye Jixie Xuebao / *Transactions of the Chinese Society of Agricultural Machinery*, 57(5), 82–94.

[7] LI, Z., LI, S., LAN, W., LI, S., CHEN, Y., & LV, P. (2025). Strawberry Fruit Detection Method Based on Improved YOLOv8n. *INMATEH-Agricultural Engineering*, 76(2).697-710.

[8] Minderer, M., Gritsenko, A., & Houlsby, N. (2023). Scaling Open-Vocabulary Object Detection. *arXiv preprint arXiv:2306*.09683.

[9] Mimma, N. E. A., Ahmed, S., Rahman, T., & Khan, R. (2022). Fruits classification and detection application using deep learning. *Scientific Programming*, 2022(1), 4194874.

[10] Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., ... & Shen, Z. Simple open-vocabulary object detection with vision transformers. *arXiv (2022). arXiv preprint arXiv:2205*.06230.

[11] Qin Y, Jia W. Real-time monitoring system for rabbit house environment based on NB-IoT network. *Smart Agric*. (2023) 5:155–65. doi: 10.12133/j. smartag.SA202211008.

[12] Qin, Y., Jia, W., Sun, X., & Lv, H. (2023). Development of electronic nose for detection of micro-mechanical damages in strawberries. *Frontiers in Nutrition*, 10, 1222988.

[13] Qing, J., Deng, X., Lan, Y., & Li, Z. (2023). GPT-aided diagnosis on agricultural image based on a new light YOLOPC. *Computers and electronics in agriculture*, 213, 108168.

[14] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning*. pp. 8748-8763. PMLR.

[15] Tang, Y., Sun, J., Zhang, Y., Wang, C., Zhang, Z., & Shi, F. (2025). Kiwifruit Flower Detection Using an Optimized YOLOv11n Architecture. *INMATEH-Agricultural Engineering*, 77(3). 44-53.

[16] Tan, C., Xu, X., & Shen, F. (2021). A survey of zero-shot detection: Methods and applications. *Cognitive Robotics*, 1, 159–167.

[17] Wei, X., Wang, B., Cao, X., Zhou, H., Wu, Z., & Wang, Z. L. (2023). Dual-sensory fusion self-powered triboelectric taste-sensing system towards effective and low-cost liquid identification. *Nature Food*, 4(8), 721-732.

[18] Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., & De Mello, S. (2023). Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2955-2966.

[19] Zhang, Y., Shao, Y., Tang, C., Liu, Z., Li, Z., Zhai, R., ... & Song, P. (2025). E-CLIP: An Enhanced CLIP-Based Visual Language Model for Fruit Detection and Recognition. *Agriculture*, 15(11), 1173.

[20] Zongyin, C., Cui, L., Yan, X., Han, Y., Yang, W., Yilong, Z., ... & Yubin, L. (2025). Field Evaluation of Different Unmanned Aerial Spraying Systems Applied to Control Panonychus citri in Mountainous Citrus Orchards. *Agriculture*, 15(12), 1283.