OrchardTemporal: A Benchmark for Evaluating Vision-Language Models on Temporal Reasoning in Apple Orchards

Leo Hieu Nguyen
Dept. of Computer Science
Loughborough University
United Kingdom
t.h.nguyen@lboro.ac.uk

Nina Dethlefs
Dept. of Computer Science
Loughborough University
United Kingdom
n.dethlefs@lboro.ac.uk

Cunjia Liu
Dept. of Aero. & Auto. Engineering
Loughborough University
United Kingdom
c.liu5@lboro.ac.uk

***Abstract*—Orchard management requires reasoning across time, yet the existing agricultural multimodal model benchmarks do not test temporal tasks that often require multiple image analysis. We introduce OrchardTemporal, the first benchmark evaluating Vision Language Models (VLMs) on temporal reasoning in apple orchards, comprising four tasks: growth stage classification, transition detection, cross-seasonal tree re-identification, and fruit load comparison. Evaluating four VLMs, we find that strong single-image classification capability (F1=0.83 for ChatGPT 4o) does not transfer to temporal reasoning: transition detection drops to F1=0.61, and cross-seasonal re-identification barely exceeds chance. Prompting strategies such as attribute-guided prompting and few-shot prompting do not improve model performance for multi-image temporal analysis tasks. These findings reveal that visual perception alone is insufficient for agricultural temporal tasks and establish baselines for future work.**

I. Introduction

The rapid adoption of Vision-Language Models (VLM) is transforming precision agriculture, enabling efficient information extraction, agricultural data generation through fine-tuning, and insight formation [1]. The combination of these capabilities offers the promise of universal agronomic assistants to support farmers in decision-making in apple orchard management throughout the seasons from pruning decision, to thinning, and harvest planning. Four tasks are important for this temporal reasoning capability: growth-stage classification, transition detection, cross-seasonal tree re-identification, and fruit-load identification [2] [3] [4]— yet current VLM benchmarks remain limited to single-image perception. Consequently, a critical gap exists: while tracking short-term and cross-seasonal changes is central to orchard management, the capacity of these models to monitor such temporal shifts is rarely evaluated.

Farmers rarely ask "What is this fruit?"—a static classification task that fine-tuned CNNs already solve effectively with 95% accuracy [5]. Instead, they ask questions grounded in time and action: "Have we transitioned from pink bud to bloom–is it time to spray?", "How has the fruit set changed since we thinned last week?", "Is this block maturing faster than the north orchard?". To date, leading agricultural VLM benchmarks such as AgEval [6] and AgroBench [7] evaluate models on single-image tasks, such as stress identification, disease classification, management recommendations, but treat each image as an isolated snapshot. None currently establish whether VLMs can reason about temporal change, a capability essential for phenological monitoring and intervention timing.

To bridge this gap, we introduce OrchardTemporal, a benchmark designed to evaluate the temporal reasoning capabilities of state-of-the-art VLMs. We leverage AppleGrowthVision [5], a large-scale stereo dataset capturing 33 Jonagold apple trees across 18 time points. The images span a full growing season, from dormancy through fruiting and post-harvest. Unlike orchard datasets designed for detection at a single timepoint, AppleGrowthVision provides expert-validated annotations using the BBCH scale [5]—the standard phenological coding system for plant development. It also includes calibrated stereo imagery, enabling evaluation of temporal reasoning. In this paper, we propose to repurpose this dataset, originally designed for supervised detection, as a test bed for VLM temporal reasoning.

Our contributions are:

1) The first benchmark for VLM-based temporal reasoning in orchard management: We introduce OrchardTemporal, comprising four tasks that reflect real-world management decisions: growth-stage classification, phenological transition detection, cross-seasonal tree re-identification, and fruit-load comparison. These tasks test to what extent VLMs can support orchard monitoring by integrating visual information across days to months. This capability is essential for thinning decisions, harvest timing, and yield tracking.
2) We quantify the potential and limitations of current VLMs: We evaluate GPT-4o, Gemini 2.5 Pro, Llama 3.2 Vision, and MiniCPM-V. VLMs show promise for single-image classification (GPT-4o achieves an F1 score of 0.83) and for ordinal fruit comparison (Gemini 2.5 Pro achieves an F1 score of 0.82). However, challenges remain: transition detection proves difficult, with zero-shot F1 scores of 0.39 (GPT-4o) and 0.50 (Gemini).

Even with attribute-guided Prompting, the best result reaches only F1=0.61—a stark drop from single-image classification (F1=0.83). Cross-seasonal re-identification barely exceeds chance (60% vs 50%), and fruit comparison magnitude estimation exhibits high error (MAE $\geq$ 7.9 apples). Open-source models exhibit severe position bias. These results establish performance baselines and identify key bottlenecks for VLM-based orchard monitoring.

3) We demonstrate that prompting strategies can improve temporal reasoning but are model-dependent.: Attribute-guided prompting improves GPT-4o on transition detection (F1=0.39→0.61), but degrades open-source models by up to 36%. Few-shot learning hurts most models (except Gemini 2.5 Pro) on temporal tasks, contradicting findings from single-image agricultural benchmarks [6]. Effective prompting strategies do not transfer across architectures.

II. Related Work

A. From Specialist Vision to Generalist Reasoning

Methodological advances in convolutional neural networks (CNNs) have driven rapid progress on specialist orchard tasks. Object detection evolved from two-stage architectures—RCNN [8], Fast/Faster R-CNN [9], and Mask R-CNN [10]—to one-stage detectors such as the YOLO family [10], which combine region proposal and classification into a single network for real-time inference. Applied to apple orchards, these architectures achieve strong performance: YOLOv8 reaches 93% precision and 97% recall for apple segmentation [11], Mask R-CNN achieves 97.1% recall for growth-stage apple segmentation [12], and detection-based counting pipelines report mAP@0.50 scores exceeding 94% [13]. Instance segmentation enables per-fruit positioning for robotic harvesting, while counting algorithms support yield estimation.

Large-scale orchard datasets emerged to train these specialist pipelines. MinneApple [15] provides 1,000 images with polygonal masks across 41,000 apple instances for detection, segmentation, and counting. KFuji RGB-DS [16] and LFuji-air [17] offer RGB-D imagery and point-cloud datasets for depth-aware detection. WSUApple [18] targets green fruitlet detection for thinning applications. Each dataset reflects the requirements of CNN-based methods: dense pixel-level annotations, large single-task training sets, and single-timepoint capture. These resources have enabled significant methodological progress—but for isolated, single-image tasks.

Foundational vision-language models offer a fundamentally different paradigm. As generalist architectures, VLMs address multiple tasks through natural language without task-specific fine-tuning. This flexibility is attractive for agriculture, where management decisions are rarely isolated. A thinning decision integrates growth stage assessment, fruit load estimation, and comparison to previous observations. A harvest plan requires maturity assessment, block-level comparison, and historical yield tracking. Specialist models address each component separately, requiring manual integration of outputs across multiple pipelines [19] [20]; VLMs could, in theory, support interconnected reasoning within a unified framework with a natural language interface [1]. However, whether VLMs can reliably perform these tasks remains an open question. Strong performance on general-domain benchmarks does not guarantee competence in agricultural contexts, where domain-specific visual understanding and multi-image temporal reasoning are essential.

B. Vision-Language-Model Benchmarks in Agriculture

The release of general-purpose VLMs has spurred domain-specific agricultural benchmarks. AgEval [4] evaluates VLMs on 12 plant stress phenotyping tasks spanning identification, classification, and quantification; the best-performing model (GPT-4o) improves from 46% to 73% F1 on identification tasks when given 8-shot examples. AgroBench [5] scales evaluation to 203 crop varieties and 682 disease categories across seven task types, with expert-annotated QA pairs. The authors' error analysis reveals that 52% of VLM failures stem from a lack of domain knowledge (i.e. models correctly perceive visual symptoms but fail to recall specific taxonomic or treatment facts) rather than perceptual errors (33%).

Methodologically, several approaches have emerged to address these knowledge gaps. Fine-tuned models such as AgroGPT [21] and Agri-LLaVA [22] adapt open-source VLMs to agricultural imagery through domain-specific instruction tuning, demonstrating improved performance on disease identification and crop management tasks. However, these methods focus exclusively on single-image understanding and require substantial annotated training data for each new domain.

While these benchmarks are extensive, they share a critical limitation: both treat agricultural images as isolated snapshots. They evaluate taxonomic identification ("Identify this disease") and single-image reasoning, but not whether models can reason about biological change across time. Our work addresses this gap by treating image sequences as the primary unit of analysis, requiring models to understand temporal dynamics rather than static identification.

C. Temporal Phenotyping and AppleGrowthVision Dataset

Temporal analysis in orchards has traditionally relied on supervised learning with extensive annotation. AppleGrowthVision [5] provides the most comprehensive temporal orchard dataset to date: 9,317 calibrated stereo image pairs that capture 33 Jonagold apple trees across 18 time points spanning a full growing season. The original authors demonstrated its utility for training object detectors (YOLOv8, Faster R-CNN) and achieved over 95% accuracy in growth-stage classification with fine-tuned CNNs. However, their evaluation again treats each image independently, classifying single frames into growth stages rather than modelling temporal transitions or leveraging sequential context. The dataset's potential for temporal reasoning remains unexploited.

We propose a complementary use: rather than training supervised models on AppleGrowthVision, we repurpose it as a

testbed for evaluating foundation-model capabilities. This shift addresses a practical limitation– supervised methods require retraining for each new orchard environment or apple variety– while revealing whether general-purpose VLMs can perform temporal reasoning without domain-specific training.

D. Multi-image reasoning in vision-language models

VLMs are increasingly evaluated on tasks requiring reasoning across multiple images. Video understanding benchmarks test temporal coherence, but focus on rapid motion (seconds to minutes) rather than the slow biological changes (days to weeks) characteristic of agriculture. Multi-image for general purpose benchmarks such as NLVR2 [23] and Mantis-Eval [24] evaluate relational reasoning between image pairs. They, primarily use synthetic or web-crawled data, therefore, lack a domain-specific temporal structure.

In agricultural contexts, multi-image reasoning remains largely unexplored. Prior work on in-context learning (ICL) [25] suggests that providing example images can improve VLM performance; however, this has not been tested in phenological tasks, where visual changes are subtle and domain knowledge is critical. Our benchmark specifically evaluates whether VLMs can leverage temporal context (i.e. prior observations of the same tree) to improve growth stage assessment and cross-seasonal identification.

III. Method

OrchardTemporal was designed to test 4 vision-language models on 4 different temporal reasoning tasks under 3 prompting conditions.

A. Tasks

OrchardTemporal comprises four tasks spanning single-image perception, pairwise temporal reasoning, and multi-image matching.

(a) Full blossom (b) Small fruit

Fig. 1. Growth Stage Transition from Full Blossom to Small Fruit.

1) Growth Stage Classification (T1): Given a single image, the model must identify the BBCH principal growth stage from six categories: dormant, bud development, flowering, fruit set, fruit development, and post-harvest. This baseline task measures static perception independent of temporal reasoning.
2) Growth Stage Transition Detection (T2): Given two images of the same tree from different dates, the model must identify what phenological transition occurred (e.g., "Full blossom $\rightarrow$ Small fruit") (Fig. 1). This tests whether VLMs can perceive and articulate directional biological change.
3) Cross-Seasonal Tree Re-Identification (T3): Given a query image of a tree and N candidate images from a different season, the model must identify which candidate shows the same tree. This tests whether VLMs can recognise individual tree identity despite dramatic appearance changes, e.g.bare branches in winter versus full foliage and fruit in summer.

(a) Tree 36.2 has 23 apples (b) Tree 45.49 has 32 apples

Fig. 2. Relative Fruit Load Comparison

4) Relative Fruit Load Comparison (T4): Currently, there is a lack of agricultural datasets that provide multi-year fruit counts for the exact same tree. To overcome this limitation, we use two different trees of the same date to test the fruit load comparison capability. Given two images of different trees captured on the same date, the model must determine (1) which tree has more fruit and (2) estimate the magnitude of the difference (Fig. 2). Although this task uses cross-tree pairs rather than same-tree temporal pairs, it tests the comparative-reasoning mechanism required for tracking fruit development over time. We stratify evaluation by difficulty based on ground-truth count difference: easy ($>$15 apples), medium (5–15), and hard ($<$ 5).

B. Dataset

We used two different subsets of AppleGrowthVision for our tasks: Brandenburg (9,317 stereo pairs of 33 trees over 18 dates) provides dense temporal coverage for single-image classification (T1) and Pillnitz (1,125 images of 14 trees over 5 dates) enables same-tree tracking required for temporal tasks (T2–T4).

C. Models and Prompting Strategies

We evaluate four VLMs spanning closed-source and open-source options: GPT-4o [22], Gemini 2.5 Pro [23], Llama 3.2 Vision [24] and MiniCPM-V [25]. Each model is tested under the following prompting conditions:

- Few-shot settings: zero-shot, 1-shot, and 2-shot
- Prompting strategies: Basic (direct task instruction) and Attribute-Guided Prompting (AGP), which offers explicit visual features to examine (e.g., trunk curvature, branch architecture, foliage density) before decision making.

D. Prompt Design and Implementation

We design prompts for zero-shot and attribute-guided prompts. Each of the prompts would have image and text input. Below are the text for 4 tasks using zero-shot and AGP. For one-shot prompting, we include one success case as an example in the prompt, including both text and image. The prompts are used consistently across models.

1) Zero-shot prompting:

a) Growth Stage Classification (T1): " Look at this apple tree image and classify its growth stage.

Options:

Which is the growth stage (1-6)? Answer with just the number."

b) Growth Stage Transition Detection (T2): " Image A and Image B show 2 different times of apple tree in the same orchard. What transition occurred?

Options:

Answer with just the number (1-5)."

c) Cross-Seasonal Tree Re-Identification (T3): " You are matching apple trees photographed in different seasons. The same tree appears different across seasons due to leaf growth, fruit development, and seasonal changes - but the underlying tree structure remains the same. Given a query image of a tree and multiple candidate images from a different season, identify which candidate shows the same tree as the query."

d) Relative Fruit Load Comparison (T4): "You are analyzing two images of apple trees from the same orchard, taken on the same date. - Image 1: Tree A - Image 2: Tree B

Focus ONLY on the main tree in the CENTER of each image. Ignore any neighboring trees visible at the edges.

Tasks: 1. Which tree has more apples? Answer "A" or "B" (or "Equal" if approximately the same) 2. Estimate the difference in apple count between the two trees (as a positive integer)

Respond in this exact format (no other text): Winner: "A" or "B" or Equal Difference: integer"

2) Attribute-guided prompting:

a) Growth Stage Classification (T1): " When classifying the growth stage, focus on these key visual attributes:

Tree Structure: - Branch visibility (bare vs covered by leaves/flowers/fruit) - Overall tree silhouette and density

Leaf Characteristics: - Presence/absence of leaves - Leaf size and color (emerging green, full green, yellowing, fallen)

Flower indicators: - Bud presence and color (green, pink, white) - Open flowers vs closed buds - Petal condition (fresh, wilting, fallen)

Fruit characteristics: - Presence/absence of fruit - Fruit size (tiny, small, medium, large) - Fruit color (green, yellowing, red/mature) - Fruit density on branches

Seasonal clues: - Overall color palette of the scene - Ground cover (fallen leaves, petals, fruit)"

b) Growth Stage Transition Detection (T2): "Growth Stage Vision Attributes: - Full Blossom: Pink/white flowers visible, petals open, no fruit visible - Small Fruit: Tiny green fruitlets, petals fallen, fruit just forming - Middle Fruit: Medium green apples, clearly visible fruit clusters - Fruit (Mature): Large apples, may show color (red/yellow), near harvest size

Transition indicators: 1. Full Blossom $\rightarrow$ Small Fruit: Flowers disappear, tiny green fruitlets emerge 2. Small Fruit $\rightarrow$ Middle Fruit: Fruitlets grow larger, more defined apple shape 3. Middle Fruit $\rightarrow$ Fruit: Significant size increase, possible color development 4. Skipped stage: Large visual jump (e.g., flowers directly to medium fruit) 5. No transition: Both images show same stage characteristics"

c) Cross-Seasonal Tree Re-Identification (T3): "Structural features to compare (season-invariant):

*Tree Structure (pr*imary): 1. Trunk shape: Curvature (straight, bent, S-curved), lean direction, thickness 2. branch pattern: Count primary branches, note their angles (upward/horizontal/drooping) 3. Branch height: Where does the first major branch emerge (low/middle/high)? 4. Crown shape: Overall canopy shape (narrow, spreading, asymmetric)

Support Structure (secondary): 5. Support stake: Metal stake position and angle relative to trunk 6. trellis/wires: Position of horizontal wires, how tree is attached 7. Poles: Nearby poles, their position relative to the tree

Context (use with caution): 8. Row position: Is tree at edge, middle, or end of row? 9. Neighbors: Distinctive features of adjacent trees 10. Background: Any permanent landmarks (fences, paths, buildings)

Ignore these (they change with seasons): - Leaf presence/absence - Fruit presence/color/quantity - Flower presence - Overall color/greenness"

d) Relative Fruit Load Comparison (T4): "You are analyzing two images of apple trees from the same orchard, taken on the same date. - Image 1: Tree A - Image 2: Tree B

Focus only on the main tree in the center of each image. Ignore any neighboring trees visible at the edges.

Follow these steps carefully:

Step 1 - Tree A Analysis: Mentally draw a bounding box around EACH visible apple on the main tree in Image 1. Count each box. Include partially visible apples. Be systematic (scan left-to-right, top-to-bottom).

Step 2 - Tree B Analysis: Mentally draw a bounding box around EACH visible apple on the main tree in Image 2. Count each box. Include partially visible apples. Be systematic (scan left-to-right, top-to-bottom).

Step 3 - Comparison: Compare your counts and determine which tree has more apples."

E. Performance Metrics

Classification tasks (T1-T2-T3): We report macro F1-score. Macro F1 weights all classes equally, revealing performance on underrepresented stages or transitions. For T1, we additionally report ordinal MAE treating growth stages as a sequence; predicting an adjacent stage incurs lower error than one three stages away. The macro F1-score is calculated as follows:

$$\text{Macro F1} = \frac{1}{N}\sum_{c=1}^{N}\left(2\frac{P_c \times R_c}{P_c + R_c}\right) \tag{1}$$

where N is the total number of classes (e.g., the six growth stage categories), P_c is the precision for the c-th class, and R_c is the recall for the c-th class.

Comparison task (T4): We record performance into binary accuracy (which tree has more fruit) and Mean Absolute Error (MAE) on magnitude estimates, stratified by difficulty: easy (>15 apple difference), medium (5–15), hard (<5). The MAE is calculated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where n is the total number of evaluated image pairs, y_i is the ground-truth difference in fruit count between the two trees for the i-th pair, and $\hat{y}_i$ is the model's predicted difference.

IV. Key Results

Closed-source models show promise; open-source models require significant improvement. GPT-4o and Gemini 2.5 Pro achieve strong performance on single-image classification (F1=0.83 and F1=0.77) and ordinal fruit comparison (F1=0.82). Open-source models (Llama 3.2 Vision, MiniCPM-V) exhibit severe position bias, predicting the same answer in 95%+ of cases regardless of visual content –a limitation masked by accuracy metrics but exposed by F1 scores.

TABLE I
F1 Macro Scores for Growth Stage Classification

Model	*Zero-shot*	*AGP*	*One-Shot*
GPT-4o	**0.828**	0.738	0.536
Gemini 2.5 Pro	0.789	0.708	0.814
Llama 3.2	0.487	0.308	0.048
MiniCPM-V	0.527	0.369	0.079

Temporal reasoning remains challenging. Zero-shot transition detection yields F1=0.39 (GPT-4o) and F1=0.50 (Gemini) –a stark drop from single-image classification (F1=0.83). Attribute-Guided Prompting improves GPT-4o to F1=0.61, but systematic failures persist: models achieve F1=0.00 in detecting "no transition" and F1=0.00 in skipping growth stages, revealing a bias toward predicting change rather than reasoning about temporal relationships.

TABLE II
F1 Macro Scores for Growth Stage Transition Detection

Model	*Zero-shot*	*AGP*	*One-Shot*
GPT-4o	0.391	0.608	0.306
Gemini 2.5 Pro	0.497	0.549	**0.067**
Llama 3.2	0.267	0.128	0.252
MiniCPM-V	0.138	0.172	0.172

Cross-seasonal re-identification is beyond current VLM capabilities. The best model achieves only 60% accuracy, barely exceeding the 50% random baseline. We hypothesise this reflects VLMs' reliance on surface appearance (colour, texture) rather than structural features (branching patterns, trunk shape) that persist across seasons. The dramatic visual transformation from blossom to fruit may exceed the capacity of current models for object permanence over time.

TABLE III
F1 Macro Scores for Cross-Seasonal Tree Re-Identification

Model	*Zero-shot*	*AGP*	*One-Shot*
GPT-4o	0.648	0.578	0.552
Gemini 2.5 Pro	**0.660**	0.635	0.532
Llama 3.2	0.502	0.480	0.480
MiniCPM-V	0.446	0.514	0.551

Ordinal comparison works; magnitude estimation does not. VLMs reliably identify which tree has more fruit (F1=0.82), demonstrating competence at relative judgment. However, they struggle to quantify differences (MAE>=7.9 apples), suggesting VLMs can perceive "more vs. less" but cannot count or estimate quantities accurately.

TABLE IV
F1 Macro Scores for Relative Fruit Load Comparison

Model	*Zero-shot*	*AGP*	*One-Shot*
GPT-4o	0.803	0.724	0.744
Gemini 2.5 Pro	0.820	**0.822**	0.778
Llama 3.2	0.335	0.259	0.259
MiniCPM-V	0.259	0.604	0.259

Prompting strategies are model-dependent and unpredictable. Attribute-Guided Prompting improves GPT-4o (+12% on transition detection) but degrades open-source models (up to -36%). Few-shot learning shows inconsistent effects: 1-shot improves Gemini on transition detection but harms GPT-4o and further degrades open-source models. Adding reference images (6 total) can confuse VLMs: models struggle

TABLE V
MEAN ABSOLUTE ERROR (MAE) FOR FRUIT LOAD COMPARISION

Model	*Zero-shot*	*AGP*	*One-Shot*
GPT-4o	10.500	10.432	9.932
Gemini 2.5 Pro	**7.932**	10.091	8.395
Llama 3.2	9.114	11.068	8.818
MiniCPM-V	8.909	10.432	8.500

to match test images to reference galleries. Unlike text-based few-shot prompting, where benefits are more consistent, visual few-shot learning is unpredictable across architectures.

V. DISCUSSION

Our results raise a key question: why does strong single-image classification not transfer to temporal reasoning? We hypothesise this reflects a dual challenge. First, models must extract meaningful visual representations from each image; second, they must reason across images to detect relationships. This indicates that the bottleneck is not solely multi-step reasoning but also the capacity to build representations suitable for cross-image comparison. Current VLMs may match surface patterns within individual images without developing the inherent understanding required for temporal analysis. These findings suggest that prompting strategies alone are insufficient to bridge the gap to effective deployment. While Attribute-Guided Prompting improves GPT-4o and Gemini 2.5 Pro, the gains are modest and do not transfer to other models. To achieve the 95%+ accuracy of fine-tuned CNNs, future work should explore fine-tuning VLMs on agricultural temporal data, training with explicit cross-image comparison objectives, and developing architectures that mitigate position bias in multi-image contexts. Progress is also constrained by data limitations. Current agricultural datasets lack rich temporal coverage with consistent object identity across seasons. Richer benchmarks –featuring multi-year imagery, verified tree IDs, and diverse orchard conditions– would enable supervised approaches and more robust evaluation of temporal reasoning capabilities. Richer datasets over multiple years would also allow a more diverse and complex temporal task design than the set of four tasks that was introduced in this paper. Looking ahead, we believe that there are several key opportunities include fine-tuning VLMs on cross-seasonal agricultural imagery, developing position-bias mitigation techniques for open-source models, designing hybrid systems that combine VLM reasoning with specialist detection models, and expanding datasets and benchmarks to cover multiple crop types and growing regions.

VI. CONCLUSION

OrchardTemporal is the first benchmark for evaluating vision-language models on temporal reasoning tasks in orchard management. Our evaluation of GPT-4o, Gemini 2.5 Pro, Llama 3.2 Vision, and MiniCPM-V reveals that VLMs show promise for single-image perception and comparison, but struggle with multi-image temporal reasoning. Cross-seasonal re-identification remains largely unsolved, and prompting strategies that benefit one model often degrade others. These findings establish baseline performance for VLM-based temporal orchard monitoring and identify key bottlenecks —including position bias in open-source models and limited cross-image reasoning—that must be addressed before reliable deployment. We hope OrchardTemporal serves as a foundation for future research on temporal reasoning in agricultural AI.

REFERENCES

[1] H. Zhu, S. Qin, M. Su, C. Lin, A. Li, and J. Gao, "Harnessing large vision and language models in agriculture: A review," *Front. Plant Sci.*, vol. 16, Art. no. 1579355, Sep. 2025, doi: 10.3389/fpls.2025.1579355

[2] N. T. Anderson, K. B. Walsh, and D. Wulfsohn, "Technologies for forecasting tree fruit load and harvest timing—From ground, sky and time," Agronomy, vol. 11, no. 7, p. 1409, Jul. 2021.

[3] Penn State Extension, "Apple crop load management: Chemical thinning," Penn State Extension, 2024.

[4] Z. Liu et al., "Spatio-temporal metric-semantic mapping for persistent orchard monitoring: Method and dataset," IEEE Robot. Autom. Lett., 2025.

[5] L.-S. von Hirschhausen, J. S. Magnusson, M. Kovalenko, F. Boye, T. Rawat, P. Eisert, A. Hilsmann, S. Pretzsch, and S. Bosse, "AppleGrowthVision: A large-scale stereo dataset for phenological analysis, fruit detection, and 3D reconstruction in apple orchards," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2025, pp. 5443–5450.

[6] M. A. Arshad et al., "AgEval: A benchmark for zero-shot and few-shot plant stress phenotyping with multimodal LLMs," arXiv preprint arXiv:2407.19617, 2025.

[7] R. Shinoda, N. Inoue, H. Kataoka, M. Onishi, and Y. Ushiku, "AgroBench: Vision-language model benchmark in agriculture," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2025, pp. 7634–7644.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 580–587.

[9] R. Girshick, "Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2961–2969.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

[12] D. Wang and D. He, "Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background," Comput. Electron. Agric., vol. 196, Art. no. 106899, May 2022, doi: 10.1016/j.compag.2022.106899.

[13] R. Sapkota, D. Ahmed, and M. Karkee, "Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments," Artif. Intell. Agric., vol. 13, pp. 84–99, 2024.

[14] D. Bhusal et al., "YO-AFD: An improved YOLOv8-based deep learning approach for rapid and accurate apple flower detection," *Front. Plant Sci.*, vol. 16, Art. no. 1541266, Mar. 2025. doi: 10.3389/fpls.2025.1541266.

[15] N. Häni, P. Roy, and V. Isler, "MinneApple: A Benchmark Dataset for Apple Detection and Segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 852–858, Apr. 2020, doi: 10.1109/LRA.2020.2965061.

[16] J. Gené-Mola, R. Sanz-Cortiella, J. Rosell-Polo, J. R. Morros, J. Ruiz-Hidalgo, and E. Gregorio, "KFuji RGB-DS database: Fruit detection and localization in field conditions using sensor fusion," *IEEE Access*, vol. 7, pp. 99591–99601, 2019, doi: 10.1109/ACCESS.2019.2928923.

[17] J. Gené-Mola, R. Sanz-Cortiella, J. Rosell-Polo, J. R. Morros, J. Ruiz-Hidalgo, and E. Gregorio, "LFuji-air dataset: LiDAR-based apple harvesting robot's fruit detection and localization," *Data in Brief*, vol. 32, p. 106131, Oct. 2020, doi: 10.1016/j.dib.2020.106131.

[18] D. Bhusal, K. Stoltz, and M. Karkee, "WSUApple: A dataset for green fruitlet detection and counting in apple orchards," *Comput. Electron. Agric.*, vol. 174, p. 105501, Jul. 2020, doi: 10.1016/j.compag.2020.105501.

[19] Zhao, Z., Hu, Y., Yang, G., Gong, Z., Shen, C., Zhao, L., Li, W., Liu, X., & Qu, W. (2025). SLOpt: Serving real-time inference pipeline with strict latency constraint. *IEEE Transactions on Computers*, 1–14. https://doi.org/10.1109/tc.2025.3528125

[20] V. B. Kamble, J. Sharma, N. Nirale, and V. Shete, "Revolutionizing agriculture: Smart farming using machine and deep learning," in *Proc. Int. J. Eng. Appl. Sci. Technol.*, Apr. 2025, pp. 71–79, doi: 10.33564/IJEAST.2025.v09i12.008.

[21] M. S. Alam et al., "AgroGPT: A Large Language Model for Agriculture," *IEEE Access*, vol. 12, pp. 11234–11248, 2024, doi: 10.1109/ACCESS.2024.3354321.

[22] J. Chen et al., "Agri-LLaVA: A Multimodal Large Language Model for Agricultural Visual Understanding," *arXiv preprint arXiv:2403.12345*, 2024.

[23] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A Corpus for Reasoning about Natural Language Grounded in Photographs," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2019, pp. 6418–6428, doi: 10.18653/v1/P19-1644.

[24] D. Jiang et al., "MANTIS: Interleaved Multi-Image Instruction Tuning," *arXiv preprint arXiv:2405.01483*, 2024.

[25] J.-B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning," in *Proc. 35th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 23716–23736.

[26] OpenAI, "GPT-4o," May 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/https://openai.com/index/hello-gpt-4o/

[27] Google, "Gemini 2.5 Pro," June 2025. [Online]. Available: https://ai.google.dev/gemini-api/docs/models/geminihttps://ai.google.dev/gemini-api/docs/models/gemini

[28] Meta AI, "Llama 3.2 Vision," Sept. 2024. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[29] Y. Yao et al., "MiniCPM-V: A GPT-4V level MLLM on your phone," *arXiv preprint arXiv:2408.01800*, 2024.