Beyond Vision: A Multimodal Dataset and Framework for Pest Recognition via Plant Electrophysiological Signals

Lu Wang[1] , **Jiaming Lin**[1] , **Yuting Ye**[1] , **Tao Wei**[1]* , **Chuchu Qin**[2]*
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences
wanglu@szu.edu.cn, 2500101025@mails.szu.edu.cn, 2510103026@mails.szu.edu.cn,
2550101023@mails.szu.edu.cn, qinchuchu@caas.cn

Abstract

Precise pest identification is essential for sustainable agriculture. Current visual recognition systems are brittle in the wild, where performance degrades due to occlusion and variable illumination. In contrast, plant electrophysiological signals serve as a robust, all-weather physiological modality, capable of detecting cryptic feeding behaviors that escape optical sensors. However, this field remains constrained by the scarcity of data and the absence of specialized algorithms. To bridge this gap, we introduce the Herbivory-Induced Plant Bio-signal Multimodal (HIPB-MM) dataset, the first fine-grained dataset comprising 4,023 synchronized plant electrophysiological signal-video pairs recording the feeding processes of three typical pest species. To address the weak and non-stationary nature of these signals, we propose the Herbivory-Induced Physiological Sensing (HIPS) framework. It integrates a Morphological Semantic Decoupling strategy to recover robust slow-wave semantics, and a Generation-State Encoder to model latent physiological states. Complementing this, an auxiliary dual-stream visual branch calibrates signal representations using explicit behavioral and morphological cues. Experiments demonstrate that HIPS establishes a solid benchmark (69.81% accuracy), comprehensively outperforming state-of-the-art baselines. Crucially, this work validates plant electrophysiology as a low-cost, all-weather modality for sustainable crop protection, effectively reducing pesticide dependency and safeguarding ecosystem health.

1 Introduction

In the realm of modern precision agriculture, pest recognition serves as a pivotal cornerstone for sustainable crop protection [Zhou *et al.*, 2024]. Timely identification enables early-stage intervention, significantly mitigating crop yield losses. Prevailing monitoring paradigms predominantly rely on the visual modality [Wadhwa and Malik, 2024]. Pest benchmark datasets ranging from the early IP102 [Wu *et al.*, 2019] to the recent AP162 [Wang *et al.*, 2025] have significantly advanced the application of pest detection models such as the YOLO series [Wang *et al.*, 2024b] and Transformer variants [Fu *et al.*, 2024] in field detection. However, the visual modality faces inherent physical limitations: complex illumination, leaf occlusion, and high concealment of tiny pests severely degrade recognition robustness, coupled with the prohibitive computational overheads of high-dimensional image data preclude all-weather continuous monitoring [de la Parte *et al.*, 2024]. More critically, vision-based monitoring typically exhibits inherent temporal latency, as it often detects infestation only after pest maturation or damage escalation [Wadhwa and Malik, 2024]. This observational lag precludes proactive early warn-

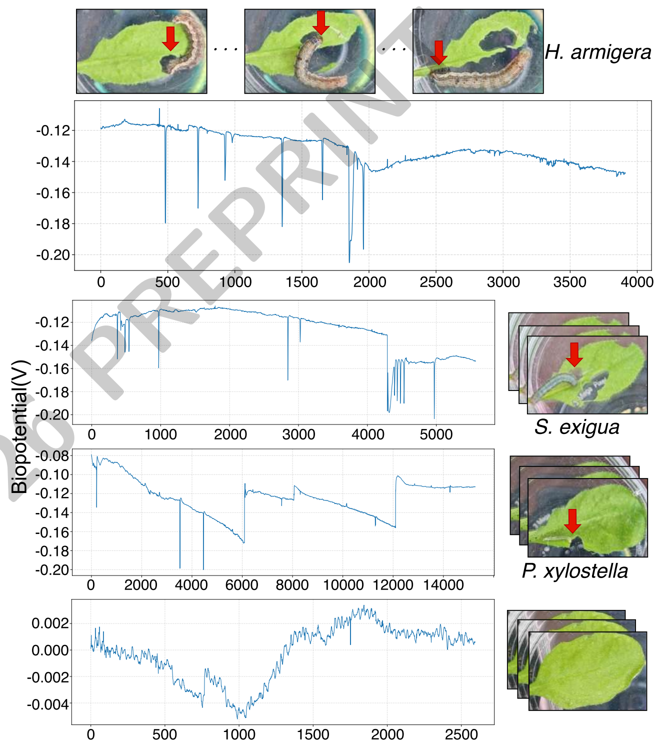

Figure 1: Synchronized Multimodal Samples in the HIPB-MM Dataset. Each voltage drop in the signal corresponds to a discrete pest feeding event captured in the synchronized herbivory imaging.

ing during the larval feeding stage, thus calling for a novel monitoring paradigm.

In light of this, leveraging plant electrophysiological signals—a novel temporal modality characterized by low dimensionality, high sensitivity, and illumination insensitivity—emerges as a promising avenue to surmount the bottlenecks. Recent research indicates that systemic electrical signals, including Action Potentials (APs), Variation Potentials (VPs) and System Potentials (SPs) carry specific physiological response information induced by biotic stress [Li *et al.*, 2025; Farmer *et al.*, 2020]. Unlike occlusion-prone visual cues, the distinct electrophysiological variations within these signals provide unambiguous internal evidence of pest feeding behaviors (Fig. 1). Nevertheless, leveraging these electrophysiological signatures for fine-grained pest identification remains largely unexplored. Effectively harnessing these plant bio-signals is non-trivial. It necessitates overcoming a series of hierarchical obstacles, ranging from biological domain incompatibility and temporal misalignment to a counter-intuitive amplitude-semantic disparity.

The first challenge lies in algorithmic incompatibility. While signal processing is well-established for human physiology (e.g., EEG, ECG), directly transferring these state-of-the-art algorithms [Mohammadi Foumani *et al.*, 2024] to plant electrophysiology faces fundamental challenges. This is primarily because existing time-series modeling methods typically assume that signals are generated by relatively stable or task-consistent mechanisms. For instance, human signals, regulated by the central nervous system and cardiac homeostasis, typically exhibit cyclostationary waveforms. However, during insect herbivory, the generation mechanisms of plant electrophysiological signals evolve non-linearly involving stimulus response, tissue damage, and physiological alteration [Farmer *et al.*, 2020], leading to drastic structural distribution drifts. Such drifts do not originate from observational noise or external environmental changes, making them difficult to mitigate through traditional normalization or domain adaptation. To tackle this challenge, we introduce a Generation-State Encoder (GSE) that explicitly models these latent generation states to condition downstream feature learning.

The second challenge arises from the severe temporal mismatch between the hysteresis of plant physiological responses and the intermittency of insect behavior. Unlike rapid animal bio-signals, plant electrophysiological dynamics operate on a much slower timescale. Coupled with the stochastic nature of insect feeding, this characteristic results in continuous acquisitions containing substantial invalid data. To balance signal validity with real-time monitoring, we adopt an Event-triggered Windowing Strategy, extracting fixed-length segments of 16 seconds exclusively during feeding events. Although this strategy eliminates non-informative intervals, it leaves the intrinsic structural disparity within the bio-signals unresolved.

The third challenge stems from this disparity, manifesting as a "Semantic Inversion" where signal amplitude diverges from semantic significance. Plant bio-signals inherently comprise high-frequency APs and slow-varying VPs/SPs. Specifically, physiological patterns induced by herbivory are pre-

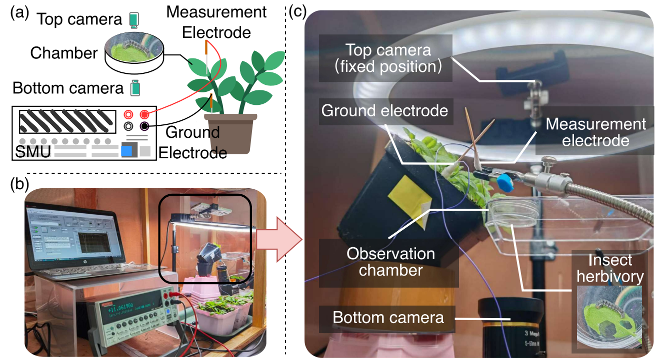

Figure 2: The HIPB-MM Dataset Collection Environment and System Overview. (a) Conceptual schematic of the dual-camera setup and electrode placement; (b) Snapshot of the in-situ experimental environment; (c) Detailed view of the acquisition platform highlighted in (b).

dominantly encoded within the low-amplitude VPs/SPs [Li *et al.*, 2025; Mousavi *et al.*, 2013]. However, these critical cues are easily overshadowed by the co-existing, high-amplitude APs. In this context, the conspicuous APs act as disruptive semantic interference, masking the robust biological features necessary for precise identification. Consequently, directly applying standard normalization causes feature collapse, as sparse AP outliers compress the effective dynamic range of VPs/SPs, severely impairing the model's ability to identify invasive biotic stressors. To mitigate this amplitude-frequency imbalance, we propose a Morphological Semantic Decoupling (MSD) strategy. This strategy is designed to selectively attenuate high-amplitude AP interference while accentuating the feature representation of slow-wave components, thereby recovering the effective dynamic range without compromising critical semantic information.

We summarize our contributions as follows:

- To the best of our knowledge, we release the Herbivory-Induced Plant Bio-signal Multimodal (HIPB-MM) dataset, the first dataset featuring synchronized video and plant electrophysiological signals specifically for insect herbivory, bridging the data gap and establishing a benchmark for future multimodal research.
- We propose the Herbivory-Induced Physiological Sensing (HIPS) framework for plant electrophysiological signals, integrating a MSD strategy to preserve robust slow-wave semantics and GSE to model evolving states.
- Experimental results uncover the evolutionary patterns of herbivory-induced signals in relation to damage severity, providing a crucial empirical foundation for future research on invasive plant stress.

2 Related Work

2.1 Pest Recognition and Benchmarks

Existing pest recognition benchmarks focus on the visual modality. Early works released small-scale datasets, containing fewer than 1,000 samples. Subsequently, larger-scale datasets emerged; notably, IP102 [Wu *et al.*, 2019] comprises 75,222 images across 102 classes, establishing a significant benchmark. Recently, research has pivoted towards fine-grained classification and higher annotation quality [Wang *et al.*, 2025]. These datasets drove algorithmic evolution. Early pest monitoring relied on analyzing environmental factors [Lee and Yun, 2023]. With the advent of deep learning, CNN-based approaches became the dominant paradigm, leveraging transfer learning and data augmentation to mitigate data scarcity [Ayan, 2024]. To address the limitations of standard CNNs in complex scenes, recent studies leverage Vision Transformers (ViTs) for global feature capture [Fu *et al.*, 2024] and optimized YOLO models for real-time, small-object detection [Wang *et al.*, 2024b]. Specialized networks have also been proposed for camouflaged and few-shot tasks [Bi *et al.*, 2021; Wei *et al.*, 2025].

Recent datasets like APTV-99 [Wang *et al.*, 2023] and visual-text fusion methods [Duan *et al.*, 2023] incorporate textual knowledge to assist recognition, yet they remain confined to static imagery. To address this, we propose the HIPB-MM dataset and the HIPS framework, which synchronize dynamic video with plant electrophysiological signals to transcend the limitations of static imagery, establishing the first multimodal benchmark for pest classification.

2.2 Plant Electrophysiological Signals

Recently, deep learning has significantly expanded the boundaries of plant-based sensing, evolving from basic environmental perception to the sophisticated monitoring of plant physiological responses under various biotic and abiotic stresses [Zarbakhsh *et al.*, 2025]. These advancements, coupled with the application of Automated Machine Learning (AutoML) [Aust *et al.*, 2025], have validated the viability of plants as intelligent biosensors. Furthermore, to address the poor generalization caused by pronounced inter-individual variability, researchers have successfully introduced domain adaptation techniques to enhance cross-individual recognition accuracy via feature alignment [Liu *et al.*, 2025a]. However, species-level pest identification via electrophysiological signals remains unexplored. To this end, we propose the HIPB-MM dataset and the HIPS framework, establishing a solid baseline for multimodal pest classification, which is essential for reducing chemical pesticide usage in sustainable agriculture.

3 The HIPB-MM Dataset

3.1 Data Collection

To acquire synchronized leaf electrophysiological time-series data and insect-herbivory video, a dedicated dual-view imaging platform was developed (Fig. 2a). Within this setup, three-week-old *Arabidopsis thaliana* are positioned such that the target leaf extends into a transparent chamber, restricting larval feeding to the designated field of view. The recording electrode contacted the petiole via a KCl/agar salt bridge, while the ground electrode was inserted into the soil to maintain a stable reference potential. All electrophysiological signals were acquired using a Keithley 2401 source-measure

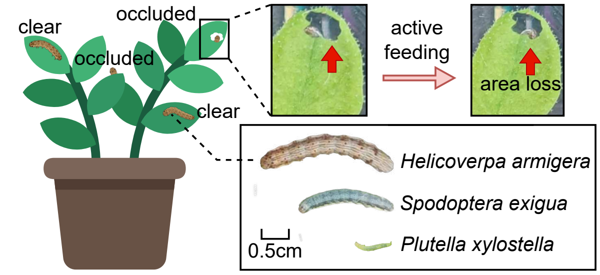

Figure 3: HIPB-MM dataset highlight. The dataset captures various real-world visual challenges including leaf occlusion and hidden feeding. Quantifying herbivory intensity via leaf area loss for dataset construction.

unit. Both electrodes were prepared following a standard protocol [Mousavi *et al.*, 2014].

Video and electrophysiological signal acquisitions were initiated simultaneously, with the trial commencing once the baseline reached a stable state. Subsequently, a second-instar larva was introduced to the target leaf. As the larva engaged in continuous feeding, the resulting tissue damage elicited bioelectrical activity. Each trial spanned 45–85 minutes, yielding synchronized multimodal datasets of dual-view videos and electrophysiological time-series. The larvae selected for this study comprised three common lepidopteran pests: *Helicoverpa armigera*, *Spodoptera exigua*, and *Plutella xylostella* (Fig. 3). The experimental site, data acquisition equipment, and theoretical physiological guidance were provided by our interdisciplinary partners (Fig. 2b,c). This collaboration ensured biological consistency in signal interpretation and dataset fidelity.

Class	*H. armigera*	*S. exigua*	*P. xylostella*	Insect-free
Count	990	977	931	1,125

Table 1: Sample Distribution Across Classes. Insect-free denotes samples in which no insect is present.

3.2 Data Pre-processing

Due to the temporal mismatch between the hysteresis of plant physiological responses and the intermittency of insect behavior, direct utilization of long-duration raw signals introduces substantial redundancy. To construct a high-quality benchmark, we implement a Herbivory-Event-triggered Windowing Strategy for data cleaning and segmentation, strictly adhering to the ”multimodal consistency” principle. Under this protocol, a valid ”Insect Feeding Event” requires simultaneous cross-modal verification: i) visual confirmation through rhythmic larval mandibular movements or explicit leaf area loss; and ii) electrophysiological correlation evidenced by significant non-stationary fluctuations in the plant signal. Applying these criteria, valid feeding intervals were extracted as short-duration segments, while ambiguous or fragmented samples were excluded. Ultimately, we constructed a dataset of 4,023 strictly aligned segments, with the category distribution detailed in Tab. 1.

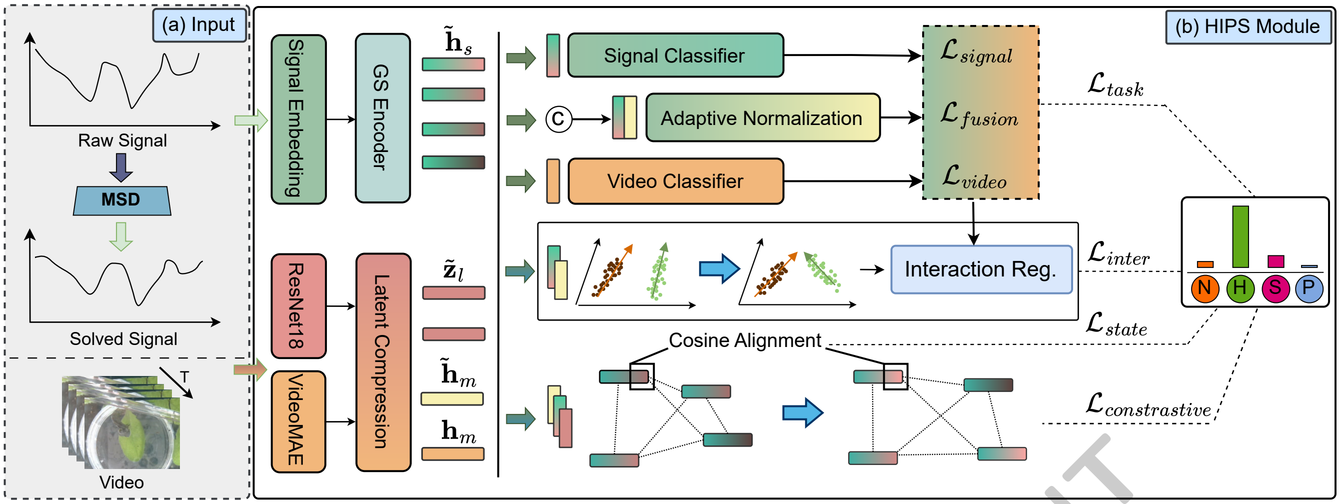

Figure 4: Overview of the HIPS framework. (a) Input Processing: Raw signals are refined via MSD strategy to recover slow-wave semantics and paired with synchronized video. (b) HIPS Module: The Signal Branch employs a GSE to inject latent state information (pink) into signal embeddings (green), while the Visual Branch extracts static and dynamic cues. These multimodal features are subsequently compressed, aligned, and fused to perform robust pest recognition.

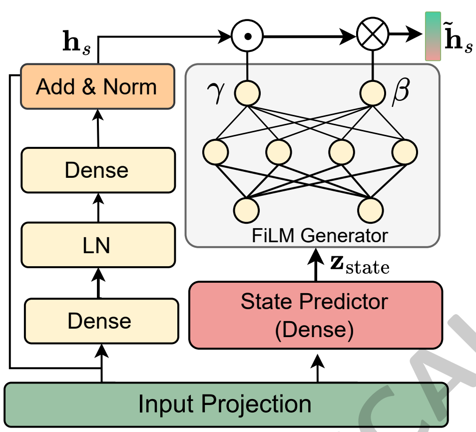

Figure 5: Structure of the Generation-State Encoder. The FiLM generator modulates input $\mathbf{h}_s$ using the latent state $\mathbf{z}_{state}$ to yield $\tilde{\mathbf{h}}_s$.

3.3 Dataset Highlights

The HIPB-MM dataset features synchronously acquired plant electrophysiological signals and dynamic videos, manually screened to isolate valid feeding-induced segments. Unlike traditional static benchmarks, the HIPB-MM dataset retains complex real-world visual challenges—such as occlusions and underside feeding—to ensure robustness without relying on idealized imagery (Fig. 3). The dataset's rare synchronized dual-modality data facilitates the capture of fine-grained behavioral cues, advancing research in plant-insect interactions and signal-based identification. Both aligned feeding segments and raw complete datasets are released to support multi-scale research in multimodal learning and biosignal modeling. By enabling early detection of pest infestations, this resource supports targeted intervention for sustainable agriculture, while extending the benchmark to abiotic and mechanical-stress conditions remains future work.

4 Method

4.1 Framework Overview

We formulate the task as a signal-centric multimodal classification problem, learning a mapping $\mathcal{F}_\theta : (\mathbf{x}_s, \mathbf{x}_v) \rightarrow \mathcal{Y}$ from synchronized electrophysiological signals $\mathbf{x}_s$ and auxiliary video context $\mathbf{x}_v$ to target pest categories $\mathcal{Y}$. The HIPS framework, illustrated in (Fig. 4), is designed as an asymmetric dual-branch architecture that prioritizes the intrinsic dynamics of plant electrophysiological signals while incorporating visual evidence for grounded calibration.

At its core, a primary signal branch is dedicated to decoding the complex biopotential variations; it first recovers stable slow-wave semantics via MSD and subsequently infers latent physiological regimes through a GSE. This process is facilitated by a supplementary visual branch that extracts pest behavioral dynamics and leaf morphological integrity to provide explicit physical priors. By integrating these heterogeneous representations through a conditional modulation mechanism, HIPS enables the robust inverse inference of pest species from non-stationary bioelectrical signatures.

4.2 Morphological Semantic Decoupling (MSD)

To mitigate "Semantic Inversion," MSD employs a robust trend-residual decomposition. We first isolate the physiological trend $\mathbf{v} = \mathcal{G}_\sigma(\mathbf{x}_s)$ and residual $\mathbf{r} = \mathbf{x}_s - \mathbf{v}$. Then, we compute a robust threshold τ via Median Absolute Deviation (MAD) to clip AP outliers and reconstruct the signal $\mathbf{y}$:

$$\tau = \lambda \cdot \text{median}\left(|\mathbf{r} - \text{median}(\mathbf{r})|\right) + \epsilon, \tag{1}$$

$$\mathbf{y} = \mathbf{v} + \alpha \cdot \text{clip}(\mathbf{r}, -\tau, \tau), \tag{2}$$

where λ and α are the robustness factor and reconstruction weight, respectively. This strategy effectively recovers slow-wave semantics for state-aware encoding.

4.3 Generation-State Encoder (GSE)

Drawing on modern multi-scale paradigms, we initialize the signal embedding $\mathbf{h}_s$ by fusing handcrafted multi-scale statistical and spectral features with deep window-based dynamics. Building upon this representation, to mitigate structural distribution drifts arising from progressive tissue damage, the GSE (Fig. 5) models the signal generation mechanism as a latent physiological condition. For a given input $\mathbf{h}_s$, the module infers a state vector $\mathbf{z}_{\text{state}}$ that characterizes the underlying generation regime. Based on this inferred state, we employ Feature-wise Linear Modulation (FiLM) to adaptively calibrate the feature manifold:

$$\tilde{\mathbf{h}}_s = \text{FiLM}(\mathbf{h}_s, \mathbf{z}_{\text{state}}) = (1+\gamma(\mathbf{z}_{\text{state}}))\odot\mathbf{h}_s+\beta(\mathbf{z}_{\text{state}}), \quad (3)$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ denote learnable affine transformations initialized to zero. This mechanism aligns signal distributions across varying stages of herbivory, ensuring that extracted features remain invariant to evolving tissue properties.

4.4 Dual-Stream Visual Context Encoding

The visual branch extracts explicit physical references through two specialized streams. The behavioral motion stream utilizes self-supervised VideoMAE [Tong *et al.*, 2022] to capture ethological patterns—including locomotion, feeding, and quiescence—providing a semantic anchor for transient biopotential bursts. Concurrently, the leaf state stream employs pre-trained ResNet18 [He *et al.*, 2016] to map tissue morphological integrity into structural references that counteract damage-induced drifts. Following latent compression, these high-dimensional cues are projected into compact embeddings $(\tilde{\mathbf{h}}_m, \tilde{\mathbf{z}}_l)$, which collectively guide the cross-modal fusion to resolve the ambiguity of implicit biological states.

4.5 Optimization Objectives

HIPS is optimized via a joint multi-task objective that facilitates robust classification while regularizing the feature manifold and cross-modal consistency.

Multi-branch Supervision. To mitigate class imbalance and background dominance, we apply a focal loss $\mathcal{L}_{task}$ across the signal (s), video (v), and fusion (f) branches:

$$\mathcal{L}_{task} = \sum_{k\in\{s,v,f\}} -\alpha_t(1-p_{t,k})^{\gamma}\log(p_{t,k}). \quad (4)$$

The fusion head $\mathbf{h}_f$ is obtained by modulating the calibrated signal $\tilde{\mathbf{h}}_s$ with the behavioral context $\mathbf{h}_m$ via Adaptive Normalization, which dynamically recalibrates signal features using visual cues.

Geometric Shaping and Distillation. To refine the representation space, a global contrastive loss $\mathcal{L}_{con}$ enforces intra-class compactness and inter-class separability:

$$\mathcal{L}_{con} = \mathbb{E}\left[y d_{ij}^2 + (1-y)\max(0, m-d_{ij})^2\right]. \quad (5)$$

Furthermore, we introduce a state distillation loss $\mathcal{L}_{state}$ to align the signal-derived state $\mathbf{z}_{state}$ with the visually grounded leaf integrity $\tilde{\mathbf{z}}_l$ via cosine distance:

$$\mathcal{L}_{state} = 1 - \frac{\tilde{\mathbf{z}}_l\cdot\mathbf{z}_{state}}{\|\tilde{\mathbf{z}}_l\|_2\|\mathbf{z}_{state}\|_2}. \quad (6)$$

This anchors implicit physiological inferences to explicit physical evidence of tissue damage.

Interaction Constraints. To prevent the signal branch from collapsing into the visual branch, $\mathcal{L}_{inter}$ penalizes feature redundancy and enforces semantic consistency between the video branch P_v and signal branch P_s:

$$\mathcal{L}_{inter} = \|\text{Corr}(\tilde{\mathbf{h}}_m, \tilde{\mathbf{h}}_s)\|_F^2 + \beta D_{KL}(P_v\|P_s). \quad (7)$$

The first term minimizes the Frobenius norm of the cross-correlation matrix, while the second term distills knowledge from the robust fusion head back to the unimodal branch.

Total Objective. The entire architecture is trained end-to-end by minimizing:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1\mathcal{L}_{con} + \lambda_2\mathcal{L}_{state} + \lambda_3\mathcal{L}_{inter}, \quad (8)$$

where $\{\lambda_i\}$ are hyperparameters balancing supervision and regularization.

5 Experiments

5.1 Experimental Setup

Baseline Methods. To comprehensively evaluate the performance of our framework, we compare it against four categories of time-series analysis methods: (1) **Conventional DL**: We select TCN [Bai *et al.*, 2018], ConvLSTM [Shi *et al.*, 2015], ResNet-TS [Wang *et al.*, 2017], and InceptionTime [Ismail Fawaz *et al.*, 2020] as standard deep learning benchmarks. (2) **Non-Attention TS**: We employ TimesNet [Wu *et al.*, 2022], DLinear [Zeng *et al.*, 2023], TimeMixer [Wang *et al.*, 2024a], LightTS [Campos *et al.*, 2023], and DisMS-TS [Liu *et al.*, 2025b] to represent recent CNN or MLP-based architectures. (3) **Attention-based TS**: The comparison covers standard Transformer [Vaswani *et al.*, 2017], FEDformer [Zhou *et al.*, 2022], Informer [Zhou *et al.*, 2021], Non-stationary Transformer [Liu *et al.*, 2022], iTransformer [Liu *et al.*, 2024], and ShapeFormer [Le *et al.*, 2024]. (4) **Foundation / SSM TS**: We further incorporate emerging foundation and state-space models, specifically Peri-midFormer [Wu *et al.*, 2024], MOMENT (evaluated in Frozen) [Goswami *et al.*, 2024], and TSCMamba [Ahamed and Cheng, 2025]. Additionally, we report results for HIPS in both Signal-only and Visually-Guided settings to verify the efficacy of the fusion strategy.

Implementation Details. All experiments were implemented using PyTorch on a single NVIDIA RTX A6000 (48GB) GPU. To ensure reproducibility, we fixed the random seed at 42 and applied stratified sampling to split the HIPB-MM dataset into training and testing sets with an 8:2 ratio. Following the established protocol, raw electrophysiological signals are segmented into 16-second windows without normalization, while video inputs are formatted as 128-frame clips. During training, we utilized the Adam optimizer with a batch size of 16. The initial learning rate was set to 5×10^{-4} and adjusted via a ReduceLROnPlateau scheduler based on validation loss. Evaluation metrics include Accuracy, Macro-F1, Macro-Recall, Macro-AUC, MCC, and Cohen's Kappa.

5.2 Performance Evaluation

Extensive benchmarking on the HIPB-MM dataset against 18 baselines demonstrates the superiority of our framework. The

Category	Model	Ref.	Acc↑	F1$_m$↑	Rec$_m$↑	AUC$_m$↑	MCC↑	Kappa↑
Conventional DL	ConvLSTM	NeurIPS'15	0.6037	0.5563	0.5830	0.8484	0.4964	0.4692
	ResNet-TS	IJCNN'17	0.5975	0.5682	0.5787	0.8053	0.4809	0.4615
	TCN	ArXiv'18	0.6161	0.6016	0.6016	0.8281	0.4927	0.4882
	InceptionTime	DMKD'20	0.6050	0.5500	0.5927	0.8396	0.5023	0.4742
Non-Attention TS	TimesNet	ICLR'23	0.6261	0.6097	0.6111	0.8234	0.5066	0.5010
	DLinear	AAAI'23	0.6050	0.5874	0.5901	0.8048	0.4792	0.4731
	TimeMixer	ICLR'24	0.6224	0.6061	0.6063	0.8289	0.4978	0.4957
	LightTS	SIGMOD'23	0.5988	0.5807	0.5829	0.8106	0.4658	0.4645
	DisMS-TS	ACM MM'25	0.5851	0.5037	0.5622	0.8187	0.4711	0.4436
Attention-based TS	Transformer	NeurIPS'17	0.5925	0.5644	0.5736	0.8091	0.4684	0.4549
	Informer	AAAI'21	0.6236	0.6092	0.6087	0.8190	0.5009	0.4978
	FEDformer	ICML'22	0.6137	0.5960	0.5980	0.8129	0.4892	0.4842
	NSTransformer	NeurIPS'22	0.6099	0.5897	0.5928	0.8110	0.4899	0.4788
	iTransformer	ICLR'24	0.6137	0.5982	0.5980	0.8111	0.4870	0.4844
	ShapeFormer	KDD'24	0.6236	0.6057	0.6076	0.8446	0.4971	0.4970
Foundation / SSM TS	Peri-midFormer	NeurIPS'24	0.6024	0.5817	0.5869	0.7993	0.4726	0.4698
	MOMENT (Frozen)	ICML'24	0.5590	0.4647	0.5493	0.7310	0.5047	0.4150
	TSCMamba	TMLR'25	0.6248	0.6103	0.6103	0.8310	0.5023	0.4996
Ours	HIPS (Signal-Only)	–	0.6634	0.6437	0.6516	0.8742	0.5541	0.5509
	HIPS (Visually-Guided)	–	**0.6981**	**0.6838**	**0.6867**	**0.8878**	**0.5976**	**0.5969**

Table 2: Benchmarking results of representative models and HIPS on the HIPB-MM dataset. The best results are highlighted in bold. We evaluate HIPS under two settings: Signal-Only (uni-modal baseline), Visually-Guided (cross-modal training with signal-only inference).

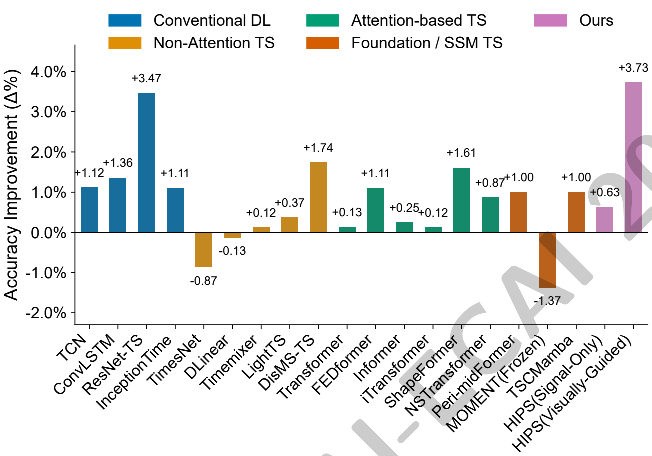

Figure 6: Performance gains of MSD across diverse model architectures. Accuracy improvements (Δ Acc) before and after applying MSD are reported for 18 models, demonstrating consistent performance gains across most architectures.

MSD	GS Enc.	Video	Acc↑	F1$_m$↑	AUC$_m$↑	Kappa↑
×	×	×	0.6435	0.6196	0.8558	0.5247
✓	×	×	0.6496	0.6302	0.8640	0.5327
✓	✓	×	0.6634	0.6437	0.8742	0.5509
✓	×	✓	0.6708	0.6537	0.8725	0.5608
✓	✓	✓	**0.6981**	**0.6838**	**0.8878**	**0.5969**

Table 3: Ablation study on key components of HIPS. Compared with the baseline (1st row), each module contributes positively, and their combination achieves the best performance.

Model	8s		16s		32s	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
HIPS-VG	0.6564	0.6369	**0.6981**	**0.6838**	0.6302	0.6095
HIPS-FF	0.8078	0.7909	0.8248	0.8174	1.0000	1.0000

Table 4: Sensitivity to window length. HIPS-VG and HIPS-FF denote the visually-guided and full-fusion variants; HIPS-VG peaks at 16s, whereas HIPS-FF overfits to visual cues at 32s (100%).

Visually-Guided HIPS achieves state-of-the-art performance with 69.81% accuracy, outperforming the runner-up TimesNet by 7.2%. Notably, despite employing diverse mechanisms, the majority of advanced architectures—including TSCMamba and ShapeFormer—plateau below a 63% accuracy ceiling. This bottleneck indicates that generic sequence modeling struggles to disentangle the stochastic non-stationarity inherent in plant bio-signals. Similarly, foundation models such as MOMENT yield a suboptimal accuracy of 55.9% due to severe domain shifts. Remarkably, our Signal-Only variant achieves a competitive accuracy of 66.34%. This result surpasses all baseline methods and validates the efficacy of MSD and GSE in capturing discriminative semantics independently.

5.3 Ablation Studies

Effect of Structural Components. As shown in Tab. 3, HIPS performance improves incrementally with the integration of proposed modules. Specifically, MSD secures steady gains over the baseline, while GSE further enhances results, indicating their complementary roles in physiological mod-

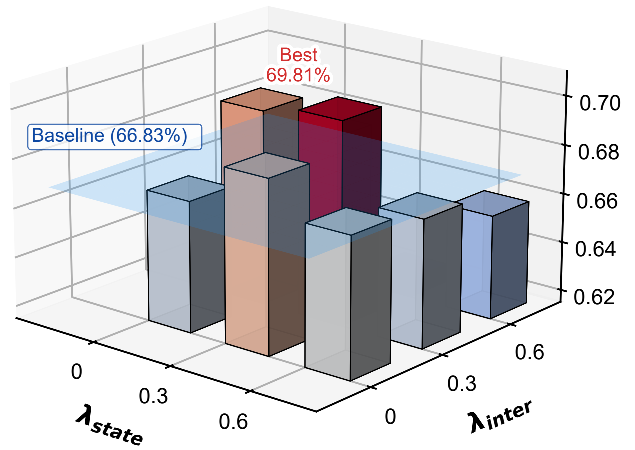

Figure 7: Visualization of accuracy with fixed $\lambda_{con} = 1$. The x and y axes represent λ_{state} and λ_{inter}, respectively. The blue plane marks the baseline accuracy (66.83%), and the highest bar indicates the best performance (69.81%).

Backbone	Strategies	Signal ($\mathbf{x}_s$)		Fusion ($\mathbf{x}_s, \mathbf{x}_v$)	
		Acc↑	F1↑	Acc↑	F1↑
TimesNet	CrossAtt	0.2447	0.1316	0.9714	0.9705
	Concat	0.5217	0.4646	0.9652	0.9645
	KD	0.5068	0.3523	-	-
ShapeFormer	CrossAtt	0.5615	0.4845	0.8671	0.8610
	Concat	0.5193	0.3717	0.8981	0.8938
	KD	0.5677	0.4965	-	-
HIPS	CrossAtt	0.2783	0.2044	0.9988	0.9987
	Concat	0.2484	0.1325	0.9540	0.9523
	KD	0.5242	0.4407	-	-

Table 5: Comparison of fusion strategies. Performance under signal (x_s) and fusion (x_s, x_v) inference highlights that standard fusion strategies suffer from severe modality dominance, collapsing on signal-only inputs irrespective of the backbone architecture.

eling. Notably, Fig. 6 demonstrates that MSD consistently elevates accuracy across the majority of tested architectures, with particularly significant improvements in ResNet-TS and HIPS(Visually-Guided). These results confirm the broad efficacy of MSD in recovering robust physiological semantics, whereas marginal performance decays in a few specific models suggest potential architectural mismatches.

Sensitivity Analysis of Loss Terms. Fig. 7 illustrates the sensitivity of HIPS to loss weights λ_{state} and λ_{inter} with $\lambda_{con} = 1$. The proposed method consistently outperforms the 66.83% baseline, reaching a peak accuracy of 69.81%. Crucially, the removal of either λ_{state} or λ_{inter} results in a marked performance decay compared to the optimal configuration. This trend underscores that both optimization objectives are essential and mutually reinforcing, validating the necessity of our joint training strategy.

Sensitivity to Window Length. To investigate the impact of the observation window on model robustness, we analyzed synchronized signal-video window lengths (Tab. 4).

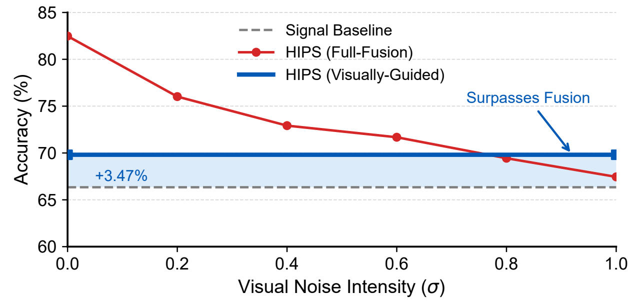

Figure 8: Robustness analysis under visual noise. Our Visually-Guided mode (Blue) maintains complete stability, securing a +3.47% gain over the baseline (Gray) and surpassing the degradation-prone Full-Fusion model (Red) in severe environments.

The Visually-Guided accuracy peaks at 16s, striking the optimal balance between information density and stationarity. In comparison, the Full-Fusion model saturates to 100% at 32s. This anomaly betrays a severe visual bias, where the model bypasses physiological learning for visual shortcuts. This fragility is corroborated in our subsequent robustness analysis, where such visually-dependent models suffer catastrophic collapse under environmental degradation, validating the necessity of our signal-centric design.

5.4 In-depth Analysis of Visual Guidance

Cross-Modal Transfer Efficiency. Table 5 reveals a stark contrast: standard fusion achieves perfect accuracy on joint inputs yet collapses on signal-only inputs. Notably, even the HIPS backbone plummets to 27.83%. Knowledge Distillation offers partial recovery ($\sim$53%) but still trails the 69.81% benchmark set by our Visually-Guided strategy. This confirms that robust physiological learning demands explicit physical alignment rather than generic probability mimicry.

Robustness against Environmental Degradation. We evaluate inference stability by injecting Gaussian noise into visual embeddings (Fig. 8) as an approximation of degraded visual conditions. With increasing noise, the Full-Fusion performance drops significantly, while our Visually-Guided mode remains stable. This failure exposes the high redundancy in visual data; relying on it makes the system fragile when visual quality degrades. This also implies the Full-Fusion model is unreliable, despite the perfect accuracy at 32s (Tab. 4).

6 Conclusion

In this work, we introduce plant electrophysiological signals as a complementary modality to overcome unimodal visual bottlenecks in complex agricultural environments. We present HIPB-MM, the first fine-grained multimodal benchmark coupling plant electrophysiology and video for pest stress recognition, alongside the HIPS framework. By integrating Morphological Semantic Decoupling with a Generation-State Encoder, HIPS synergizes physical priors and latent state dynamics. Experiments show that HIPS consistently outperforms existing methods in fine-grained classification, establishing a new performance benchmark.

Acknowledgments

The research was supported in part by the Key Research and Development Program in Xinjiang Uygur Autonomous Region (2025B04019-001), the China NSFC Grant (No. 62372307, No. U2001207), Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011691), Shenzhen Science and Technology Program (No. RCYX20231211090129039), Shenzhen Science and Technology Foundation (No. JCYJ20230808105906014), Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007), the Project of DEGP (No. 2023KCXTD042), the National Natural Science Foundation of China (No. 22304186). Tao Wei and Chuchu Qin are corresponding authors.

References

[Ahamed and Cheng, 2025] Md Atik Ahamed and Qiang Cheng. Tscmamba: Mamba meets multi-view learning for time series classification. *Information Fusion*, 120:103079, 2025.

[Aust *et al.*, 2025] Till Aust, Eduard Buss, Felix Mohr, and Heiko Hamann. Automated phytosensing: Ozone exposure classification based on plant electrical signals. In *2025 IEEE Symposia on Computational Intelligence for Energy, Transport and Environmental Sustainability (CIETES)*, pages 1–7. IEEE, 2025.

[Ayan, 2024] Enes Ayan. Genetic algorithm-based hyperparameter optimization for convolutional neural networks in the classification of crop pests. *Arabian Journal for Science and Engineering*, 49(3):3079–3093, 2024.

[Bai *et al.*, 2018] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

[Bi *et al.*, 2021] Hongbo Bi, Cong Zhang, Kang Wang, Jinghui Tong, and Feng Zheng. Rethinking camouflaged object detection: Models and datasets. *IEEE transactions on circuits and systems for video technology*, 32(9):5708–5724, 2021.

[Campos *et al.*, 2023] David Campos, Miao Zhang, Bin Yang, Tung Kieu, Chenjuan Guo, and Christian S Jensen. Lightts: Lightweight time series classification with adaptive ensemble distillation. *Proceedings of the ACM on Management of Data*, 1(2):1–27, 2023.

[de la Parte *et al.*, 2024] Mario San Emeterio de la Parte, José-Fernán Martínez-Ortega, Pedro Castillejo, and Néstor Lucas-Martínez. Spatio-temporal semantic data management systems for iot in agriculture 5.0: Challenges and future directions. *Internet of Things*, 25:101030, 2024.

[Duan *et al.*, 2023] Jinli Duan, Haoyu Ding, and Sung Kim. A multimodal approach for advanced pest detection and classification. *arXiv preprint arXiv:2312.10948*, 2023.

[Farmer *et al.*, 2020] Edward E Farmer, Yong-Qiang Gao, Gioia Lenzoni, Jean-Luc Wolfender, and Qian Wu. Wound-and mechanostimulated electrical signals control hormone responses. *New Phytologist*, 227(4):1037–1050, 2020.

[Fu *et al.*, 2024] Xueqian Fu, Qiaoyu Ma, Feifei Yang, Chunyu Zhang, Xiaolong Zhao, Fuhao Chang, and Lingling Han. Crop pest image recognition based on the improved vit method. *Information Processing in Agriculture*, 11(2):249–259, 2024.

[Goswami *et al.*, 2024] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ismail Fawaz *et al.*, 2020] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[Le *et al.*, 2024] Xuan-May Le, Ling Luo, Uwe Aickelin, and Minh-Tuan Tran. Shapeformer: Shapelet transformer for multivariate time series classification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1484–1494, 2024.

[Lee and Yun, 2023] Sangyeon Lee and Choa Mun Yun. A deep learning model for predicting risks of crop pests and diseases from sequential environmental data. *Plant Methods*, 19(1):145, 2023.

[Li *et al.*, 2025] Rui Li, Yongfang Yang, Hao Lou, Weicheng Wang, Jianbin Yan, Daoxin Xie, and Xiaoyi Shan. Electrical and calcium signaling in plant systemic defense: from local wounds to global responses. *New Phytologist*, 2025.

[Liu *et al.*, 2022] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35:9881–9893, 2022.

[Liu *et al.*, 2024] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pages 11116–11140, 2024.

[Liu *et al.*, 2025a] Chenrui Liu, Ji Qi, Xiuxin Xia, Yicheng Wang, Qiuping Wang, Lingfang Sun, and Hong Men. Cross-individual electrophysiological signal recognition in clivia biosensors via domain adaptation. *Applied Soft Computing*, page 113323, 2025.

[Liu *et al.*, 2025b] Zhipeng Liu, Peibo Duan, Binwu Wang, Xuan Tang, Qi Chu, Changsheng Zhang, Yongsheng Huang, and Bin Zhang. Disms-ts: Eliminating redundant multi-scale features for time series classification. In

Proceedings of the 33rd ACM International Conference on Multimedia, pages 10817–10826, 2025.

[Mohammadi Foumani *et al.*, 2024] Navid Mohammadi Foumani, Lynn Miller, Chang Wei Tan, Geoffrey I Webb, Germain Forestier, and Mahsa Salehi. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9):1–45, 2024.

[Mousavi *et al.*, 2013] Seyed AR Mousavi, Adeline Chauvin, François Pascaud, Stephan Kellenberger, and Edward E Farmer. Glutamate receptor-like genes mediate leaf-to-leaf wound signalling. *Nature*, 500(7463):422–426, 2013.

[Mousavi *et al.*, 2014] Seyed AR Mousavi, Chi Tam Nguyen, Edward E Farmer, and Stephan Kellenberger. Measuring surface potential changes on leaves. *nature protocols*, 9(8):1997–2004, 2014.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wadhwa and Malik, 2024] Daisy Wadhwa and Kamal Malik. A generalizable and interpretable model for early warning of pest-induced crop diseases using environmental data. *Computers and Electronics in Agriculture*, 227:109472, 2024.

[Wang *et al.*, 2017] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.

[Wang *et al.*, 2023] Shansong Wang, Qingtian Zeng, Weijian Ni, Cheng Cheng, and Yanxue Wang. Odp-transformer: Interpretation of pest classification results using image caption generation techniques. *Computers and Electronics in Agriculture*, 209:107863, 2023.

[Wang *et al.*, 2024a] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*, 2024.

[Wang *et al.*, 2024b] Zejun Wang, Shihao Zhang, Lijiao Chen, Wendou Wu, Houqiao Wang, Xiaohui Liu, Zongpei Fan, and Baijuan Wang. Microscopic insect pest detection in tea plantations: improved yolov8 model based on deep learning. *Agriculture*, 14(10):1739, 2024.

[Wang *et al.*, 2025] Kaili Wang, Wei Li, Xiao Wu, Jingwen Xu, Zhi-Wen Wang, and Siyu Yang. Ap162: A large-scale dataset for agricultural pest recognition. *Computers and Electronics in Agriculture*, 237:110520, 2025.

[Wei *et al.*, 2025] Linsen Wei, Jingjun Tang, Jinxiu Chen, Carine Pierrette Mukamakuza, Defu Zhang, and Tong Zhang. A lightweight few-shot learning model for crop pest and disease identification. *Artificial Intelligence Review*, 58(10):329, 2025.

[Wu *et al.*, 2019] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8787–8796, 2019.

[Wu *et al.*, 2022] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[Wu *et al.*, 2024] Qiang Wu, Gechang Yao, Zhixi Feng, and Yang Shuyuan. Peri-midformer: Periodic pyramid transformer for time series analysis. *Advances in Neural Information Processing Systems*, 37:13035–13073, 2024.

[Zarbakhsh *et al.*, 2025] Saeedeh Zarbakhsh, Fazilat Fakhrzad, Dragana Rajkovic, Gniewko Niedbała, and Magdalena Piekutowska. Approaches and challenges in machine learning for monitoring agricultural products and predicting plant physiological responses to biotic and abiotic stresses. *Current Plant Biology*, page 100535, 2025.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[Zhou *et al.*, 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pages 27268–27286. PMLR, 2022.

[Zhou *et al.*, 2024] Wentao Zhou, Yashwanth Arcot, Raul F Medina, Julio Bernal, Luis Cisneros-Zevallos, and Mustafa ES Akbulut. Integrated pest management: an update on the sustainability approach to crop protection. *ACS omega*, 9(40):41130–41147, 2024.