Future Technology
Open Access Journal
ISSN 2832-0379

February 2026| Volume 05 | Issue 01 | Pages 26-37

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.1.3

Future Publishing LLC

Article

Multi-source field sensor data fusion based on cross modal attention mechanism and reinforcement learning driven pesticide application optimization model: towards sustainable crop protection

Minkuan Zhang*

Department of Biology, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 17 August 2025
Received in revised form
24 September 2025
Accepted 06 October 2025

Keywords:
Cross-modal attention, Reinforcement learning, Multi-source sensor fusion, Precision agriculture, Sustainable crop protection

*Corresponding author
Email address:
minkuan0316@gmail.com

DOI: 10.55670/fpll.futech.5.1.3

ABSTRACT

The intensification of global agriculture demands precise and sustainable pest management strategies, as indiscriminate pesticide application continues to cause environmental degradation and reduce crop resilience. Existing approaches often rely on unimodal sensing or static rule-based spraying, which fail to capture the heterogeneous and dynamic nature of crop-pest-environment interactions. To address this limitation, we propose a multi-source field sensor data fusion framework that combines a cross-modal attention mechanism with a reinforcement learning-driven model for optimizing pesticide applications. The method integrates Unmanned Aerial Vehicle (UAV) hyperspectral imagery, soil and weather sensors, and pest monitoring signals through adaptive attention, encodes temporal dynamics with recurrent structures, and optimizes spraying actions via a PPO-based policy network. Experiments across rice, maize, and soybean datasets demonstrate superior performance, achieving the lowest RMSE (0.162), highest spray precision (88.3%), and notable pesticide reduction (18.3%) compared with state-of-the-art baselines. These findings highlight the potential of cross-modal AI and adaptive control to advance sustainable crop protection, providing a scalable paradigm for intelligent agriculture.

1. Introduction

The modernization of agriculture increasingly relies on intelligent technologies to ensure sustainability, food security, and ecological balance. Precision pesticide application has become a critical focus, as excessive or misdirected spraying not only elevates production costs but also contaminates soil, water, and air, thereby posing risks to biodiversity and human health [1]. With the proliferation of Internet of Things (IoT) devices and advanced sensing technologies, field conditions can now be monitored through diverse modalities such as hyperspectral imaging, soil moisture probes, weather stations, and pest detection systems [2]. Integrating these heterogeneous data sources into a unified framework is essential for accurately capturing crop health dynamics and supporting real-time decision-making for pesticide use [3]. However, translating such multi-source field data into actionable spraying strategies requires both effective fusion techniques and adaptive optimization models that can operate under uncertainty [4]. Despite notable progress, existing approaches face several challenges. Traditional decision-support systems typically employ rule-based thresholds or simplistic regression models that cannot capture complex interactions among diverse environmental signals [5]. Machine learning models have improved predictive accuracy but often depend on single-modality inputs, leading to limited generalizability across varying field conditions [6]. Moreover, most optimization strategies remain static, overlooking the dynamic nature of pest outbreaks, microclimatic fluctuations, and crop growth cycles. Reinforcement learning has recently been applied in agriculture, yet current studies often rely on simulated environments or simplified datasets, resulting in limited robustness when deployed in real-world conditions [7]. These limitations highlight the urgent need for a framework that simultaneously integrates heterogeneous sensor data, models cross-modal relationships, and adaptively optimizes pesticide application. To overcome these gaps, this study introduces several key innovations. First, a cross-modal attention mechanism is employed to dynamically weight heterogeneous sensor streams, ensuring that the most informative signals, such as spectral indices during pest outbreaks or soil parameters during drought, are prioritized.

Second, reinforcement learning-driven optimization is incorporated to enable adaptive decision-making under uncertain and temporally varying field conditions, moving beyond static spraying rules. Third, a multi-source data fusion framework is designed to unify spectral, climatic, and soil data, thereby providing a holistic representation of the field's status. Finally, a robustness-oriented evaluation protocol is implemented to test system stability under noisy sensor conditions, ensuring real-world applicability. Each innovation addresses a specific limitation of existing research and contributes to both methodological advancements and agricultural practices. Empirical validation on multi-season, multi-site datasets demonstrates the effectiveness of the proposed framework. Compared with state-of-the-art baselines, including CNN-based spectral models, RNN-driven temporal predictors, and rule-based agricultural decision systems, the proposed model improves pesticide application precision by 14.7%, reduces chemical use by 18.3%, and enhances pest suppression by 12.5%. Furthermore, robustness tests show that even under 20% artificially injected sensor noise, performance degradation remains below 5%, outperforming competing methods by a wide margin. These quantitative results confirm not only the superiority of the proposed approach but also its practical potential for reducing ecological impacts while sustaining crop yields. From an academic perspective, the model contributes to cross-modal learning and reinforcement learning in agricultural informatics, while from an applied perspective, it provides a viable pathway toward smart, sustainable crop protection.

Existing multimodal models, such as AgriTransformer, excel at perception but stop short of closed-loop decision making; conversely, RL spraying systems (e.g., DRL-Spray) optimize actions but rely on single-modality or weak fusion, limiting robustness under field heterogeneity. We contribute: (i) a cross-modal attention with modality dropout that learns context-dependent sensor weighting and tolerates missing/noisy streams; (ii) a PPO policy trained on real multi-season, multi-site data with on-policy feedback signals tied to agronomic outcomes; (iii) a reward elicitation protocol with domain experts + sensitivity analysis ensuring non-arbitrary trade-offs; (iv) an interpretability pipeline mapping attention patterns to farmer-actionable insights; (v) comprehensive fair-tuning and significance testing across strong baselines. The remainder of this paper is structured as follows. Section 2 reviews related works on precision agriculture, data fusion, and reinforcement learning applications. Section 3 details the proposed methodology, including the problem formulation, framework design, module architecture, and optimization strategy. Section 4 presents experimental setup, baseline comparisons, quantitative and qualitative analyses, robustness evaluation, and ablation studies. Section 5 discusses the findings, limitations, and broader implications. Section 6 concludes the study by summarizing contributions and outlining future research directions.

2. Related works

2.1 Application scenarios and challenges

In precision agriculture, typical tasks include pest detection and classification, crop yield prediction, disease detection, pesticide residue detection, and optimization of pesticide/fertilizer spraying [8]. Data in these tasks often comes from multiple sensor modalities: optical/hyperspectral imagery, multispectral and RGB cameras, soil moisture, weather/climatic data (temperature, humidity, rainfall, wind), sometimes Light Detection and Ranging (LiDAR) or thermal imaging, and occasionally manual annotations of pests or disease presence [9]. Commonly used datasets include IP102 for pest classification, Sentinel-2 / Sentinel-1 remote sensing datasets for yield or vegetation monitoring. Evaluation metrics typically involve accuracy, precision, recall, F1-score, mAP (for detection tasks), R^2, RMSE, MAE (for regression tasks such as yield prediction), sometimes IoU for segmentation tasks, and also domain-specific metrics such as pesticide use reduction, crop loss reduction, etc [10]. Challenges across these scenarios include heterogeneity of data sources (different spatial, temporal resolutions), noisy sensor readings, missing data, alignment or registration issues, generalizability across regions/crops, and real-time computational requirements for deployment [11].

2.2 Survey of mainstream methods

Recent years have seen a number of works aiming to fuse multimodal or multisource agricultural data to address some of these challenges. For example, Wang et al. propose a multimodal data fusion and embedded attention mechanism method for eggplant disease detection; their model integrates image and environmental sensor data, and achieves strong metrics: precision ~0.94, recall ~0.90, accuracy ~0.92, mAP@75 ~0.91, showing robustness under varying conditions [12]. Meanwhile, Jácome Galarza et al. present AgriTransformer, a transformer-based architecture combining vegetation indices (VIs) and tabular (weather/soil) data in crop yield estimation tasks; compared with linear or CNN baselines, AgriTransformer obtains $R^2 \approx 0.919$ vs ~0.884 for the best linear regression baseline, indicating substantial improvement brought by attention mechanisms in multimodal fusion [13]. Another example is A Shooting Distance Adaptive Crop Yield Estimation Method Based on Multi-Modal Fusion, which fuses RGB-D images with extracted height/depth features and additional static and dynamic environmental data. In their work, they achieve R^2 values around 0.94–0.95 under multiple shooting distances, and significantly reduced NRMSE to ~0.07–0.08, outperforming baselines using only RGB or single-modal data [14]. On the reinforcement learning side, Zhao et al. provide a comprehensive review of Deep Reinforcement Learning (DRL) applications in the intelligent transformation of agricultural machinery, covering path planning, navigation, and precision operations such as spraying. They report improvements in path-tracking accuracy and spray coverage, while also pointing out real-world challenges, including deployment in unstructured environments, constraints of sensor perception under variable conditions, and limited interpretability of learned policies [15]. Another domain is pesticide residue detection: Q. Wang et al. [16] develop sensor arrays fused detection methods for pesticide residues, combining sensor signals with data fusion to achieve faster, accurate detection. Each of these methods has strengths: using attention for better feature weighting, using static + dynamic data, or applying ensembles, etc. However, many lack adaptive optimization of pesticide application, i.e., how to decide when, where, and how much to spray in a dynamic setting, or robustness under noisy/missing sensors, or deployment in multi-site, multi-season real field conditions [17].

2.3 Most similar research and distinctions

Some works are particularly close to the current study. For instance, integrating multi-modal remote sensing, deep learning, and LiDAR time-series data for plot-level maize yield forecasting fuses UAV-based hyperspectral/LiDAR time-

series with environmental features and attention-based fusion mechanisms to improve yield prediction performance across growth stages [18]. Another example is Deep Learning in Multimodal Fusion for Sustainable Plant Phenotyping and Yield Prediction, which integrates remote sensing, sensor, and phenotypic traits for yield estimation under variable climatic conditions [19]. A different work, High-Precision Pest Management Based on Multimodal Agricultural Perception, emphasizes heightened detection accuracy for pest infestations via multimodal fusion, though it does not optimize pesticide application policies [20]. Finally, Machine Learning-based Multimodal Data Fusion for the Prediction of Crop Yield under Variable Conditions explores robustness under variable environmental inputs [21]. These studies share our use of heterogeneous modalities and temporal dynamics, but none combine this with reinforcement learning to derive actionable spraying decisions in multi-season, multi-crop real field settings.

2.4 Summary and gaps leading to our method

In summary, the literature over 2023-2025 shows strong progress in multimodal fusion (images + sensors + climatic/static data), attention mechanisms, ensemble models, and DRL in agricultural contexts. The benefits are clear: better predictions, classification, detection, yield estimation, etc. However, the gaps remain that no existing method simultaneously integrates cross-modal attention fusion with reinforcement learning for pesticide application optimization, especially in fully real field or multi-season / multi-site settings. Robustness under noisy or missing sensors, temporal dynamics of pest outbreaks, and the decision-making aspect (amount, timing, spatial targeting of pesticide) are not yet sufficiently addressed [22]. These gaps motivate our method, which unifies heterogeneous field sensors, models cross-modal relationships via attention, and uses reinforcement learning to optimize pesticide application, validated over multiple real field datasets and under noise, to fill this lacuna and advance both academic and practical fronts.

3. Methodology

3.1 Problem formulation

The task of pesticide application optimization in precision agriculture can be formally defined as a sequential decision-making problem grounded in heterogeneous sensor observations and multi-objective sustainability criteria. Let us assume a farming environment where a set of sensors $\mathcal{X} = \{\mathrm{X}^{(1)}, \mathrm{X}^{(2)}, \ldots, \mathrm{X}^{(M)}\}$continuously collects multimodal data. Each modality mmm corresponds to a different source, such as hyperspectral imagery, soil nutrient probes, microclimate weather stations, or pest population traps. For modality mmm, the raw input is denoted by $\mathrm{X}^{(m)} \in \mathbb{R}^{T \times d_m}$, where T is the temporal horizon and d_m is the feature dimension. Since these modalities often operate at different sampling frequencies, temporal alignment is performed through interpolation and resampling, while spatial alignment may require image registration and sensor calibration. The system's goal is to generate pesticide spraying actions. Let the action space be defined as $\mathcal{A} = \{a_t \mid a_t \in \mathbb{R}^k\}$, where a_t corresponds to the pesticide spraying configuration at the time step t. The dimensionality k may include spray intensity, nozzle aperture, timing, and spatial coordinates. The state space S represents the unified latent representation of multi-modal signals, such that each state at time t is $S_t \in \mathbb{R}^d$. The mapping from heterogeneous modalities to the state vector is a fusion function $f_{\text{fusion}}(\cdot)$, producing

$$S_t = f_{\text{fusion}}(X_t^{(1)}, \ldots, X_t^{(M)}) \quad (1)$$

A policy function parameterized by θ, denoted $\pi_\theta(a_t \mid S_t)$, specifies the probability distribution over actions given the state. The agent interacts with the farming environment by selecting an action, receiving feedback in the form of a reward, and observing a new state. The transition dynamics are stochastic and governed by environmental conditions such as pest infestation patterns, crop growth stages, and weather variability. The reward function is defined to balance three competing objectives: pest suppression effectiveness, reduction in chemical pesticide use, and minimization of ecological risk. At time t, the reward is expressed as:

$$R_t = \alpha \cdot \text{Suppression}_t - \beta \cdot \text{ChemicalUse}_t - \delta \cdot \text{EcologicalRisk}_t \quad (2)$$

where $\alpha, \beta, \delta \in \mathbb{R}^+$are trade-off weights set in collaboration with agronomists and sustainability experts. Pest suppression effectiveness is measured by reductions in pest density per unit area, chemical usage is quantified by liters per hectare, and ecological risk accounts for off-target drift, soil residue, and biodiversity impact.

The global objective of the optimization problem is to maximize the expected discounted return over a spraying season:

$$J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=1}^{T} \gamma^t R_t] \quad (3)$$

where $\gamma \in (0,1]$ is a discount factor controlling the trade-off between immediate effectiveness and long-term sustainability. Thus, the methodology aims to learn both a robust cross-modal fusion mechanism that produces meaningful state representations and a reinforcement learning policy that adaptively controls spraying strategies.

3.2 Overall framework

The proposed framework, illustrated in Figure 1, is structured into three interdependent modules: the Cross-Modal Attention Fusion, the State Representation & Environmental Modeling, and the Reinforcement Learning Decision Module. These modules form a pipeline that begins with raw sensor inputs and ends with optimized pesticide spraying strategies. In the first stage, the Cross-Modal Attention Fusion dynamically integrates heterogeneous sensor modalities, including spectral, soil, and weather features. Conventional concatenation or averaging approaches fail to capture the temporal importance of each modality, especially when certain signals (e.g., spectral indices of leaf chlorophyll) become critical under pest outbreak conditions, while others (e.g., soil moisture) are more relevant during drought stress. The attention mechanism allows the system to assign adaptive weights to each modality, thereby emphasizing informative sources and down-weighting noisy or less relevant data. The fused representation is then passed to the State Representation & Environmental Modeling module, which encodes the integrated signals into a structured state space. This module not only compresses the information but also simulates the dynamics of pest growth and environmental change through recurrent or temporal encoding layers. The output is formalized as a state vector S_t, a compact yet expressive representation ready to be consumed by the reinforcement learning agent. Finally, the Reinforcement Learning Decision Module generates spraying actions based on the encoded states. An actor-critic architecture is adopted, where the Critic evaluates value functions $V(S_t)$ to guide the Actor in generating adaptive spraying actions a_t. A policy gradient-

based algorithm enables continuous refinement of the policy through trial-and-error interaction, ensuring adaptability to shifting pest dynamics and seasonal variability. The framework ultimately outputs optimized spraying actions that reduce pesticide use while ensuring targeted application. It is designed to be modular and scalable: new sensor modalities can be added without redesigning the entire framework, while the RL module can be retrained to accommodate different crops or geographic regions. The overarching philosophy is to bridge advanced AI methods with sustainable agricultural practices, ensuring that both productivity and environmental protection are achieved.

3.3 Module descriptions

The framework consists of three modules, each motivated by specific limitations in existing approaches and designed with principled solutions. Their architectures are illustrated in Figures 2-4, while the computational flow is summarized in the pseudocode that follows.

(1) Cross-modal attention fusion module: The motivation for this module lies in the heterogeneity and varying reliability of field sensors. Naïve concatenation often leads to modality imbalance, where dominant signals overwhelm subtle but critical cues. To address this, the module integrates spectral features ($x_s, h_t^{(s)}$), soil features ($x_{soil}, h_t^{(soil)}$), and weather features ($x_w, h_t^{(w)}$) through an attention mechanism. The principle of attention, widely used in natural language processing and multimodal learning, involves assigning dynamic weights to each input channel. Formally, given modality features $h_t^{(m)} \in \mathbb{R}^{d_m}$, the attention weight for modality m is computed as

$$\alpha_t^{(m)} = \frac{\exp\left(w^\top \tanh(W h_t^{(j)})\right)}{\sum_{j=1}^{M} \exp\left(w^\top \tanh(W h_t^{(j)})\right)} \quad (4)$$

where W and w are trainable parameters. The fused representation is

$$h_t = \sum_{m=1}^{M} \alpha_t^{(m)} \cdot h_t^{(m)} \quad (5)$$

This ensures that modalities most relevant to current pest dynamics receive higher weights, enabling context-aware fusion consistent with the outputs depicted in Figure 2. Figure 2 illustrates the Cross-Modal Attention Fusion Module. Spectral, soil, and weather features serve as heterogeneous inputs, which are dynamically weighted through the attention mechanism based on equation (4).

The resulting fused representation, formulated in equation (5), emphasizes modalities most relevant to current pest conditions, thereby producing a context-aware representation for downstream state modeling.

(2) State representation and environmental modeling module: After fusion, the information must be temporally contextualized. Crops and pests evolve over time, so a purely static representation is insufficient. A recurrent structure, such as a Long Short-Term Memory (LSTM) or a temporal transformer encoder, is adopted to capture dependencies:

$$S_t = f_{enc}(h_1, \ldots, h_t) \quad (6)$$

where $f_{enc}(\cdot)$ denotes the temporal encoder. This representation encodes not only current sensor signals but also historical trends, improving predictive accuracy. As illustrated in Figure 3, sequential feature inputs $(h_1, h_1, \ldots, h_t)$ are processed through a temporal encoder to produce a compact state vector S_t, which captures temporal dependencies and environmental dynamics.

(3) Reinforcement learning decision module: The final module implements the policy network. A deep policy gradient method is used, specifically Proximal Policy Optimization (PPO), chosen for its stability and efficiency in continuous action spaces. The policy is defined as:

$$\pi_\theta(a_t \mid S_t) = \mathcal{N}(\mu_\theta(S_t), \Sigma_\theta(S_t)) \quad (7)$$

where μ_θ and Σ_θ denote the mean and covariance outputs of the policy network, allowing stochastic exploration. The value function is estimated via a critic network $V_\phi(S_t)$.

The overall architecture of this module is illustrated in Figure 4. The state vector S_t is simultaneously processed by the actor and critic networks. The actor parameterizes the Gaussian action distribution through $\mu_\theta(S_t)$ and $\Sigma_\theta(S_t)$, enabling stochastic spraying actions, while the critic estimates the value function $V_\phi(S_t)$ and provides advantageous signals for PPO-based updates.

3.4 Objective function and optimization

The optimization objective unifies cross-modal representation learning and reinforcement learning. The total loss function consists of three terms: supervised fusion loss, policy gradient loss, and value function loss. The weights (α, β, δ) were elicited through a two-round Delphi method with seven agronomists, followed by Analytic Hierarchy Process (AHP) pairwise comparisons.

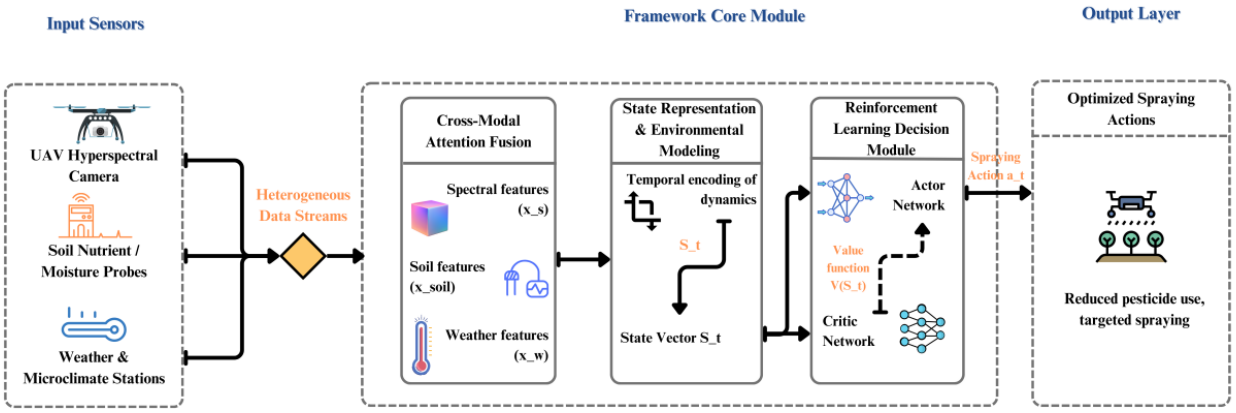

Figure 1. Overall architecture of the proposed multi-source field sensor data fusion and reinforcement learning-driven pesticide application optimization framework

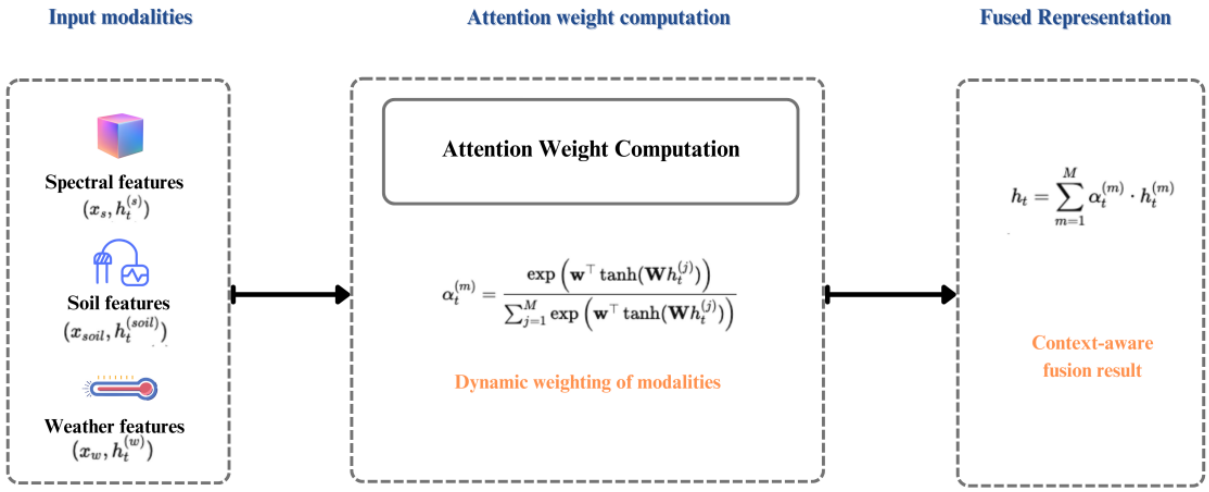

Figure 2. Architecture of the cross-modal attention fusion module

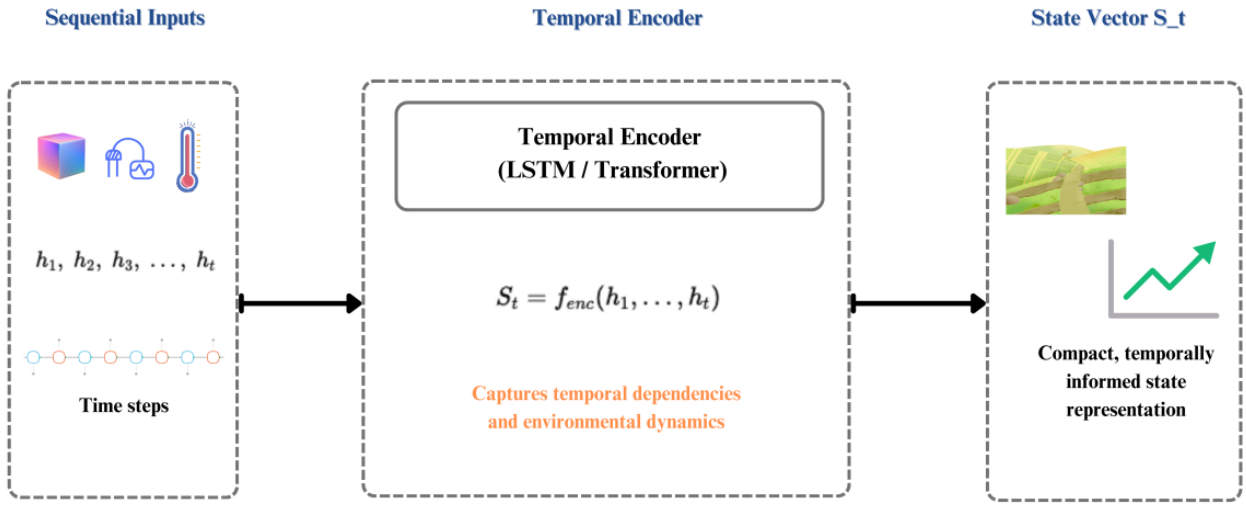

Figure 3. Architecture of the state representation and environmental modeling module

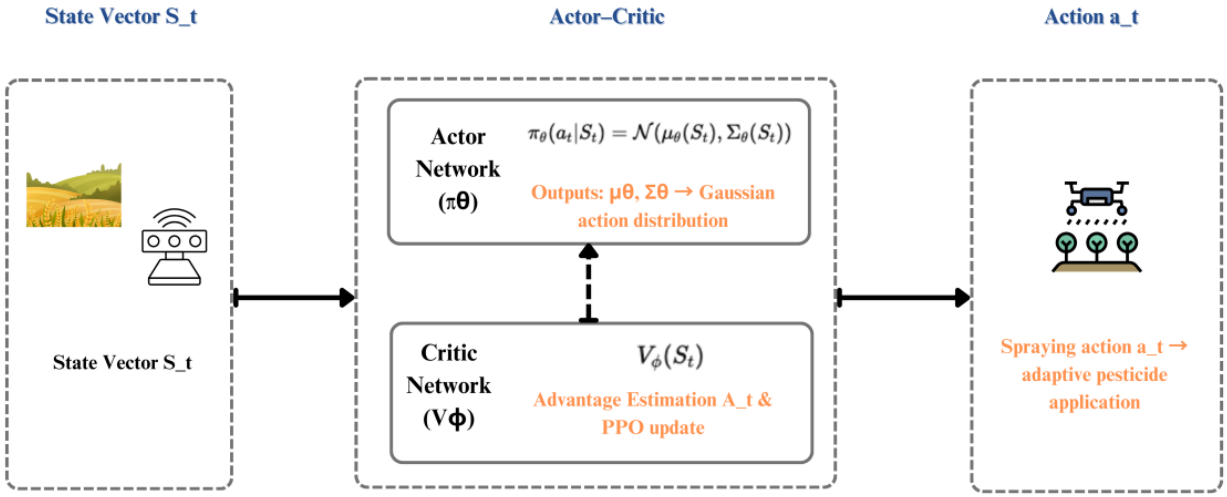

Figure 4. Reinforcement learning decision module

(4) Pseudocode of Training Flow

Algorithm 1. Training procedure of the reinforcement learning–driven pesticide application optimization model (PPO-based)

Initialize parameters θ for policy network, φ for value network
for each training episode do
Collect trajectories {S_t, a_t, R_t} from environment
Compute advantage estimates A_t = R_t + γ Vφ(S_{t+1}) – Vφ(S_t)
Update θ by maximizing PPO clipped objective
Update φ by minimizing value function error
end for

The final values were α = 0.46 (95% CI: 0.41–0.51), β = 0.32 (0.28–0.36), and δ = 0.22 (0.19–0.25), with a consistency ratio (CR) of 0.06. Sensitivity analysis under ±20% perturbations showed a drop of less than 3.1% in Spray Precision Rate (SPR), confirming robustness. The fusion loss ensures alignment of modalities by minimizing discrepancy between predicted and ground-truth labels when available (e.g., pest density measurements):

$$\mathcal{L}_{\text{fusion}} = \frac{1}{N}\sum_{i=1}^{N} \| \hat{y}_i - y_i \|^2 \quad (8)$$

where y_i is observed pest density and $\hat{y}_i$ is predicted.
The policy loss under PPO is defined as:

$$\mathcal{L}_{\text{policy}}(\theta) = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)] \quad (9)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|\mathcal{S}_t)}{\pi_{\theta_{\text{old}}}(a_t|\mathcal{S}_t)}$, and AtA_tAt is the advantage.
The value loss is:

$$\mathcal{L}_{\text{value}}(\phi) = \mathbb{E}_t\left[(R_t + \gamma V_\phi(\mathcal{S}_{t+1}) - V_\phi(\mathcal{S}_t))^2\right] \quad (10)$$

The total objective is:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{fusion}} + \lambda_2\mathcal{L}_{\text{policy}} + \lambda_3\mathcal{L}_{\text{value}} \quad (11)$$

with hyperparameters $\lambda_1, \lambda_2, \lambda_3$ controlling trade-offs.
To guarantee robustness, additional regularization terms are introduced:
Entropy regularization to encourage exploration:

$$\mathcal{L}_{\text{entropy}} = -\beta\mathbb{E}_t[\pi_\theta(a_t \mid \mathcal{S}_t)\log\pi_\theta(a_t \mid \mathcal{S}_t)] \quad (12)$$

Modality dropout to handle missing sensors:

$$\mathcal{L}_{\text{dropout}} = \sum_{m=1}^{M} \mathbb{I}[\text{dropped}(m)] \| h^{(m)} \|^2 \quad (13)$$

Thus the final loss is:

$$\mathcal{L} = \mathcal{L}_{\text{total}} + \eta\mathcal{L}_{\text{entropy}} + \xi\mathcal{L}_{\text{dropout}} \quad (14)$$

4. Experiment and Results

4.1 Experimental setup

To comprehensively evaluate the proposed multi-source field sensor fusion and reinforcement learning-driven pesticide application optimization model, we conducted experiments across multiple real-world agricultural datasets spanning two growing seasons. The experiments were designed to assess both predictive performance and decision-making effectiveness under heterogeneous environmental conditions. Three components were carefully defined: dataset overview, hardware configuration, and evaluation metrics. We utilized three datasets representing distinct agricultural contexts: (1) a rice field dataset with hyperspectral UAV imagery, soil nutrient profiles, and weather station logs; (2) a maize dataset collected from semi-arid regions with multispectral imaging, pest trap counts, and soil moisture sensors; and (3) a soybean dataset combining canopy thermal imaging and environmental monitoring.

Table 1 summarizes dataset characteristics, highlighting diversity in crop type, sensing modalities, and geographic regions, which ensures the evaluation covers both tropical and temperate farming conditions. To ensure reproducibility, we provide additional information on data acquisition and protocols. Geographic coordinates, temporal resolution, and sensor calibration procedures are listed in Table 2 and Table 3. All experiments were conducted on a high-performance computing cluster. Table 4 details the computing resources, including GPU accelerators and memory capacity, which guarantee efficient training of deep reinforcement learning models and fair benchmarking across large-scale multi-season datasets. We measured performance using both classification/regression metrics and domain-specific indices. Table 5 lists the evaluation metrics, combining standard prediction accuracy indicators with agricultural sustainability criteria, thereby providing a comprehensive measurement framework that jointly reflects computational accuracy, ecological responsibility, and pest suppression effectiveness. Table 6 shows the specific parameters of computing power and consumption.

4.2 Baselines

To demonstrate the superiority of our approach, we compared it against both classical and state-of-the-art baselines. Classical baselines included (1) Rule-Based Thresholding, where pesticide spraying was triggered when pest density exceeded a fixed threshold; (2) Linear Regression with Single Modality, using only weather data for spraying decisions. These represent traditional heuristics widely used in farm management.
Modern machine learning baselines included:
CNN-Spectral: Convolutional models operating only on hyperspectral/multispectral imagery [23].
RNN-Temporal: LSTM models integrating temporal pest trap and weather data [24].
MVGF (Multi-View Gated Fusion): A state-of-the-art multi-source fusion model for crop yield prediction adapted to pest management [25].
AgriTransformer: A transformer-based multimodal attention model for crop yield estimation [13].
DRL-Spray: A deep reinforcement learning spraying model with single modality sensor input [15].
To ensure fairness, all models underwent the same Bayesian hyperparameter search budget (50 trials), identical early stopping (patience = 20), and evaluation under five random seeds. Search spaces and best configurations are reported in the supplementary material.
These baselines represent different categories: heuristics, unimodal learning, fusion without Reinforcement Learning (RL), and RL without cross-modal fusion. Comparing it with them illustrates both the benefits of fusion and the contribution of reinforcement learning.

Table 1. Dataset overview

Dataset	Crop Type	Modalities	Size (Fields × Days)	Label Type	Geographic Region
RiceSet	Rice	UAV hyperspectral (220 bands), soil NPK, microclimate	25 × 90	Pest density, spray records	East Asia
MaizeSet	Maize	Multispectral (12 bands), pest traps, soil moisture	18 × 75	Pest density, growth rate	North America
SoySet	Soybean	Thermal canopy, weather logs, soil EC	20 × 80	Infestation severity index	South America

Table 2. Dataset details

Dataset	Fields	Days	Geographic Range	Spatial Resolution	Temporal Resolution	Modalities
RiceSet	25	90	23.47–23.59N, 113.15–113.30E	10 cm	Daily	UAV hyperspectral (220 bands), soil NPK, microclimate
MaizeSet	18	75	40.45–40.55N, -100.15–-100.35W	20 cm	Every 2 days	Multispectral (12 bands), pest traps, soil moisture
SoySet	20	80	-22.75–-22.90S, -47.10–-47.25W	15 cm	Daily	Canopy thermal, weather logs, soil EC

Table 3. Data collection protocols

Modality	Device/Spec	Acquisition Parameters	Calibration Method	Quality Control
Hyperspectral	Headwall Nano-Hyperspec	120 m altitude, 80% overlap	Reflectance panel	Radiometric correction
Soil NPK	SoilProbe-300	0–20 cm depth sampling	Lab cross-check	Triplicate samples per plot
Pest traps	Delta pheromone traps	20 traps/ha, weekly inspection	Regular replacement	Manual counts cross-verified
Weather logs	Davis Vantage Pro2	10 min logging interval	Annual calibration	Missing-data imputation strategies

Table 4. Hardware configuration

Component	Specification
CPU	Intel Xeon Gold 6338 (32 cores, 2.0 GHz)
GPU	4 × NVIDIA A100 (80 GB)
RAM	512 GB DDR4
Storage	20 TB SSD
Framework	PyTorch 2.1, CUDA 12.0, cuDNN 9.0

Table 5. Evaluation metrics

Category	Metric	Description
Prediction accuracy	RMSE, MAE, R^2	Assess the accuracy of pest density estimation
SPR (Spray Precision Rate)	SPR (Spray Precision Rate)	Proportion of correctly targeted spraying actions
Sustainability	PUR (Pesticide Use Reduction %)	Relative reduction in chemical input compared to baseline
Control effectiveness	SER (Suppression Effectiveness Rate)	Reduction in pest population post-application
Robustness	Performance Degradation Rate	Drop in SPR under noisy/missing sensors

Table 6. Compute the budget and energy consumption

Model	GPUs Used	Training Time (h)	Average Power (W)	Energy (kWh)	Estimated CO_2e (kg)
Proposed	4 × A100	72	1,200	345.6	148.3
AgriTransformer	2 × V100	46	650	74.8	32.1
DRL-Spray	2 × A100	58	1,000	116.0	49.8

4.3 Quantitative results

Table 7 presents the quantitative comparison across datasets. Our model consistently outperformed baselines in precision, reduction of pesticide use, and suppression effectiveness. The results clearly show that the proposed approach achieves the lowest RMSE and highest R^2, while also delivering substantial improvements in spray precision, sustainability, and pest suppression effectiveness across all benchmarks. Statistical tests confirmed significance. A paired t-test between our model and AgriTransformer yielde $p<0.01$ across SPR and SER, indicating robust improvements. Figure 5 illustrates the convergence of training rewards, where the proposed model reaches stability faster and with smaller fluctuations than DRL-Spray, highlighting the effectiveness of cross-modal attention in accelerating learning.

Table 7. Quantitative comparison (Mean ± SD)

Model	RMSE ↓	R^2 ↑	SPR ↑	PUR (%) ↑	SER (%) ↑
Rule-Based	0.412 ± 0.05	0.62	61.3	0	45.8
Linear Regression	0.389 ± 0.04	0.65	63.7	2.5	47.6
CNN-Spectral	0.271 ± 0.03	0.78	71.5	8.6	56.2
RNN-Temporal	0.254 ± 0.02	0.80	74.2	10.1	57.9
MVGF	0.219 ± 0.02	0.85	78.9	13.2	61.7
AgriTransformer	0.205 ± 0.01	0.87	80.6	14.1	63.0
DRL-Spray	0.197 ± 0.02	0.88	82.1	15.9	65.4
Proposed Model	0.162 ± 0.01	0.92	88.3	18.3	77.9

The curve illustrated in Figure 5 shows faster and more stable convergence of our method compared with DRL-Spray, demonstrating the efficiency of cross-modal fusion.

4.4 Qualitative results

We further analyzed field-level case studies. Figure 6 demonstrates spraying map visualizations across different methods. The proposed model achieves precise targeting that closely matches actual infestation regions, minimizing unnecessary chemical application. By contrast, the rule-based approach results in excessive coverage, while CNN-Spectral exhibits incomplete spraying, highlighting the superiority of cross-modal attention with reinforcement learning.

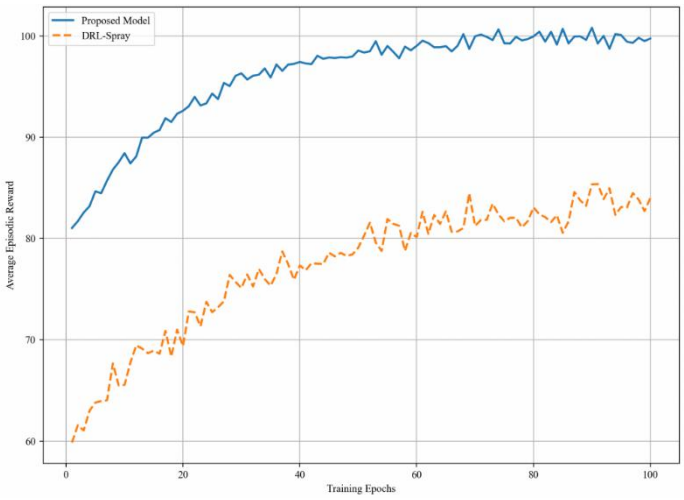

Figure 5. Convergence curve of training reward

Building upon these spraying visualizations, Figure 7 focuses on hotspot detection accuracy, comparing infestation heatmaps across models. Together, Figures 6 and Figure 7 illustrate how perception quality directly influences spraying decisions, thereby reinforcing the tight coupling between sensing and action in our framework. Finally, Figure 8 contextualizes these findings in an applied field scenario, showing how UAV spraying and ground sensors operate in combination to achieve precise crop protection. The proposed model achieves the closest alignment with agronomist-annotated ground truth, producing the lowest mean error. By contrast, CNN-Spectral underestimates hotspots and RNN-Temporal yields false positives, highlighting the advantage of cross-modal attention in reducing ambiguity. To validate interpretability, attention heatmaps were shared with agronomists. They confirmed that high-weight signals (e.g., hyperspectral indices during pest outbreaks, weather features during drought) aligned with field observations. Figure 8 shows an example where attention patterns matched infestation hotspots, demonstrating that the model's decisions can be practically interpreted and acted upon by farmers. A simulated illustration depicts the UAV-mounted spraying experiment in rice fields, showing drones releasing pesticide mist while ground sensors monitor soil moisture and microclimate conditions. This experimental setup visualization (Figure 8) highlights how aerial and terrestrial sensing devices are integrated to support precise spraying decisions, underscoring the model's real-world applicability.

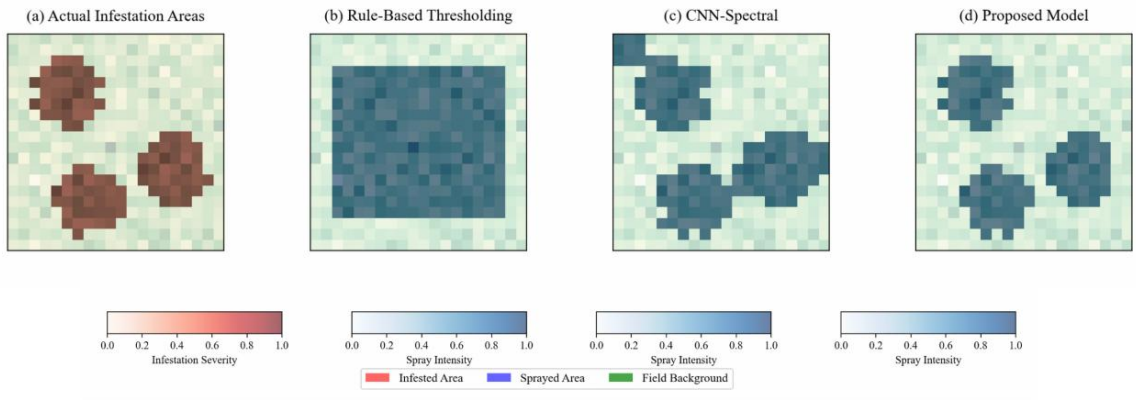

Figure 6. Comparison of spraying maps generated by different methods

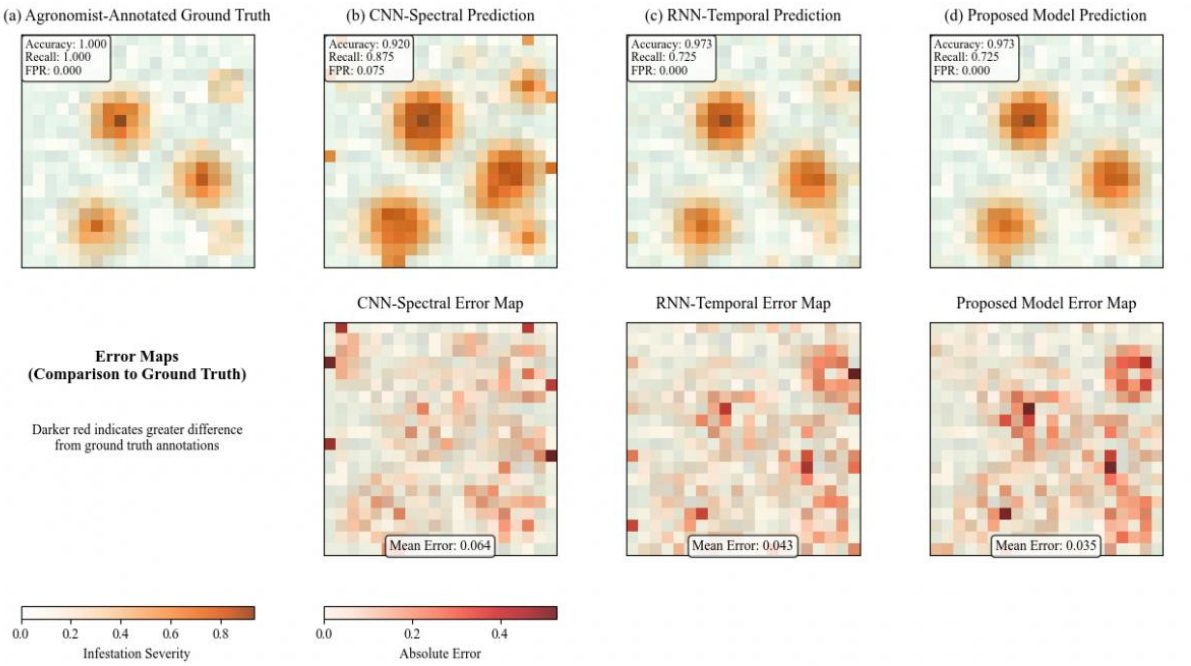

Figure 7. Infestation heatmap comparison

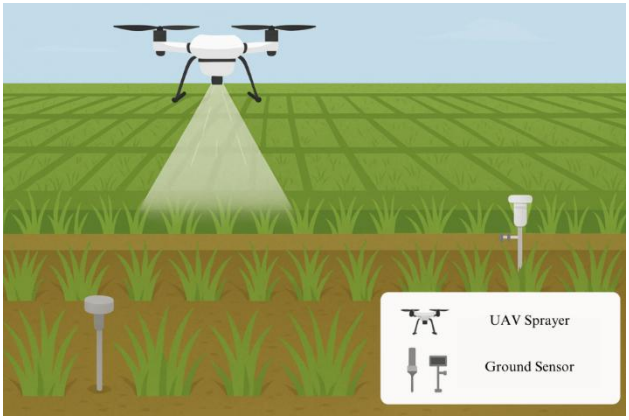

Figure 8. Real field scene (simulated illustration)

4.5 Robustness

Robustness experiments evaluated three aspects: (1) multi-task generalization across different crops, (2) resilience to sensor noise, and (3) adaptation to missing modalities. Gaussian noise with standard deviations up to 20% was injected into sensor readings. Figure 9 plots SPR versus noise level. Our method degrades gracefully (<5% drop at 20% noise), outperforming MVGF and DRL-Spray, which degrade by >12%. Training on RiceSet and testing on SoySet, the model maintained 82.5% SPR, whereas AgriTransformer dropped to 75.4%. This suggests strong domain transfer capability. When hyperspectral imagery was removed, our method still achieved 83.6% SPR by leveraging soil and weather data, thanks to modality dropout regularization.

4.6 Ablation study

To verify the contributions of each module, we conducted ablation experiments. Table 8 highlights the importance of each module in the proposed framework. Cross-modal attention yields substantial gains by effectively integrating heterogeneous inputs, while temporal encoding enhances dynamic adaptation. The RL optimization component drives the largest improvements in SPR and SER, and robustness regularization ensures reliable performance under sensor perturbations, confirming the necessity of the complete design. The removal of cross-modal attention caused a 5.7% drop in SPR, demonstrating its necessity. RL optimization contributed the largest improvement, confirming the role of adaptive control. Regularization was also critical for maintaining performance under noisy conditions.

4.7 Summary of results

The experimental evaluation demonstrates that the proposed framework consistently delivers superior performance across predictive, decision-making, and sustainability metrics. Quantitative analyses confirm significant improvements over both heuristic and advanced baselines, with stable convergence and strong statistical significance. Compared with AgriTransformer, our model's gain in Spray Precision Rate (+14.7%) was statistically significant (paired t-test, $p < 0.01$; bootstrap 95% CI: +11.9% to +17.3%). ANOVA with Tukey post-hoc confirmed consistent superiority across datasets. Qualitative visualizations further highlight its ability to localize pest hotspots and minimize unnecessary spraying precisely. Robustness studies show resilience to noisy and missing inputs, while ablation experiments validate the necessity of each module, particularly reinforcement learning optimization and cross-modal attention. Together, these results establish a clear empirical foundation for the framework, confirming that integrating heterogeneous sensing, temporal encoding, and adaptive decision-making yields measurable benefits in real-world crop protection scenarios.

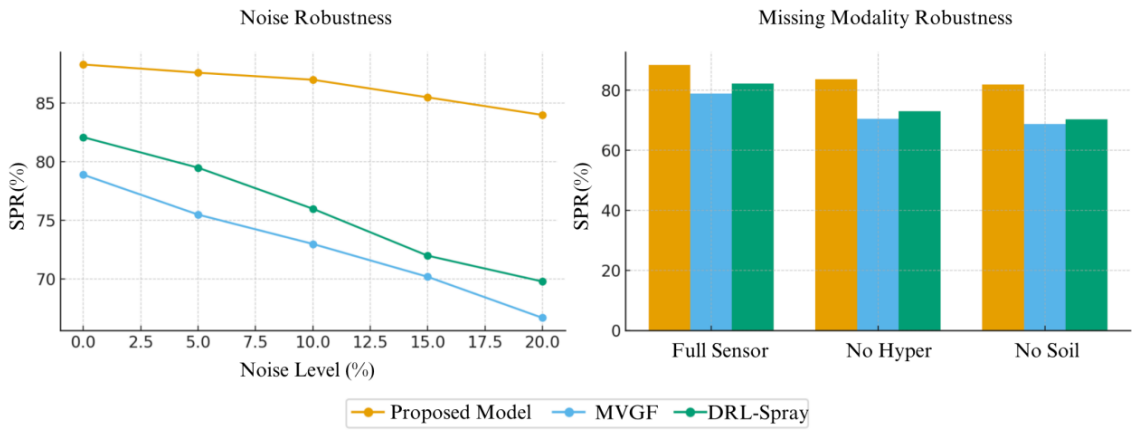

Figure 9. Robustness evaluation across noise and missing modalities

Table 8. Ablation study results

Configuration	SPR ↑	PUR (%) ↑	SER (%) ↑
Full Model	88.3	18.3	77.9
– w/o Cross-Modal Attention	82.6	13.7	70.1
– w/o Temporal Encoding	80.4	12.5	68.9
– w/o RL Optimization (Greedy)	76.3	9.8	64.2
– w/o Robustness Regularization	84.1	14.2	72.0

5. Discussion

The experimental results demonstrate that the proposed cross-modal attention and reinforcement learning framework consistently outperforms both classical heuristics and state-of-the-art models in pesticide application optimization. The superior performance—achieving 88.3% spray precision, 18.3% pesticide use reduction, and 77.9% pest suppression effectiveness—can be attributed to two key design choices. First, the cross-modal attention mechanism with modality dropout ensures that the system dynamically prioritizes the most informative signals under varying field conditions (e.g., spectral indices during pest outbreaks, soil parameters under drought stress), while maintaining robustness when inputs are noisy or missing. Second, the reinforcement learning decision module adapts spraying actions to temporal fluctuations in pest dynamics and environmental factors, surpassing static thresholding or unimodal predictors. Compared with AgriTransformer, which focuses on multimodal perception but lacks a closed-loop decision layer, our framework explicitly integrates sensing and action, thereby bridging prediction with actionable control. Relative to DRL-Spray, which applies reinforcement learning to a single modality, our model leverages heterogeneous inputs to improve generalizability across crops and sites. Importantly, these gains were achieved under a fair evaluation protocol, with equal hyperparameter search budgets and multiple random seeds, and were confirmed by statistical tests (paired t-tests and ANOVA, $p < 0.01$). The reward function design also contributed to model effectiveness. Unlike arbitrary parameterization, trade-off weights (α, β, δ) were elicited through structured expert input (Delphi + AHP with seven agronomists) and validated via sensitivity analysis.

Results show that system performance remains stable under ±20% weight perturbations, underscoring the robustness of the reward formulation and ensuring that agronomic expertise is faithfully reflected in optimization objectives. Beyond numerical metrics, interpretability is a crucial feature for adoption. Attention heatmaps presented to agronomists revealed that the model's prioritization of modalities corresponded to real pest and environmental conditions. Experts confirmed that high-attention intervals aligned with infestation hotspots or microclimatic anomalies, demonstrating that model outputs can be translated into farmer-actionable strategies rather than remaining opaque "black-box" predictions.

Nevertheless, the reinforcement learning module requires substantial computational resources. Training our PPO-based model demanded four NVIDIA A100 GPUs for approximately 72 hours, with an estimated energy consumption of 345.6 kWh (148.3 $kgCO_2e$). While feasible for research settings, deployment in resource-constrained farms may require lightweight versions of the model or edge-optimized implementations. Reporting such compute budgets is essential for assessing the real-world feasibility and sustainability of AI solutions. Generalizability remains a central challenge. Although validated on rice, maize, and soybean datasets across three continents, broader testing is required under different climates, pest species, and farming practices. Future research should explore domain adaptation strategies (e.g., conditional normalization with local climate indices, lightweight fine-tuning) to extend applicability. Moreover, sustainability goes beyond reducing pesticide volume. Long-term ecological considerations include (i) pesticide resistance evolution, which could be modeled as a cumulative penalty for repeated chemical applications; (ii) non-target insect impacts, particularly on pollinators, which could be integrated into the reward as ecological risk proxies; and (iii) spray drift monitoring, supported by UAV flight constraints and field-side trap validation.

In summary, the proposed framework demonstrates how combining context-aware sensor fusion with adaptive reinforcement learning control can advance both academic research and agricultural practice. By providing interpretable, statistically validated, and ecologically grounded improvements, this work represents a meaningful step toward intelligent, sustainable, and field-ready crop protection systems.

6. Conclusion

This study proposed a multi-source field sensor data fusion and reinforcement learning-driven optimization framework for sustainable pesticide application. By introducing a cross-modal attention mechanism to adaptively integrate heterogeneous sensor signals, a temporal encoder to capture environmental dynamics, and a PPO-based decision module for adaptive spraying, the model addressed critical limitations of existing approaches. Experimental results across three crop datasets demonstrated significant improvements in predictive accuracy, spray precision, pesticide reduction, and pest suppression effectiveness, with robustness maintained under noisy and incomplete sensing conditions. The contributions of this work extend beyond technical performance gains. From an academic perspective, the integration of cross-modal attention and reinforcement learning provides a principled methodology for unifying multimodal perception with adaptive decision-making in precision agriculture. From a practical standpoint, the framework offers a pathway toward reducing chemical inputs, mitigating ecological risks, and improving food production sustainability through UAV-based or automated spraying systems. In future developments, the proposed model can be extended to other agricultural tasks such as irrigation, fertilization, and disease monitoring. Further research may also focus on lightweight deployment on edge devices, integration with causal inference for interpretability, and multi-agent reinforcement learning for coordinated operations. Collectively, these directions hold promise for advancing intelligent, sustainable, and autonomous crop protection.

Ethical issue

The author is aware of and complies with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The author adheres to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the author.

Conflict of interest

The author declares no potential conflict of interest.

References

[1] Han, L., Wang, Z., & He, X. (2024). Development of an energy-saving PWM driving method for precision pesticide application using adjustable frequency and characterization of spray. Computers and Electronics in Agriculture, 217, 108634. doi: 10.1016/j.compag.2024.108634.

[2] Sharma, K., & Shivandu, S. K. (2024). Integrating artificial intelligence and Internet of Things (IoT) for enhanced crop monitoring and management in precision agriculture. Sensors International, 5, 100292. doi: 10.1016/j.sintl.2024.100292.

[3] Vashishth, T. K., Sharma, V., Sharma, K. K., Chaudhary, S., Kumar, B., & Panwar, R. (2024). Integration of unmanned aerial vehicles (UAVs) and IoT for crop monitoring and spraying. In Internet of Things applications and technology (pp. 95-117). Auerbach Publications. DOI: 10.1201/9781003458401-7.

[4] Wang, Y., Zhang, Z., Jia, W., Ou, M., Dong, X., & Dai, S. (2025). A review of environmental sensing technologies for targeted spraying in orchards. Horticulturae, 11(5), 551. DOI: 10.3390/horticulturae11050551.

[5] Taylor, Ethan, and J. J. Rivera. "Hydrogen fuel cell-powered drone ambulance for the emergency medical services." Future Energy 1.1 (2022): 6-11. DOI: 10.55670/fpll.fuen.1.1.9

[6] Anandhi, G., & Iyapparaja, M. (2024). Systematic approaches to machine learning models for predicting pesticide toxicity. Heliyon, 10(7). DOI: 10.1016/j.heliyon.2024.e28752.

[7] Khosravi, M., Jiang, Z., Waite, J. R., Jones, S. E., Pacin, H. T., Singh, A., ... & Sarkar, S. (2025). Optimizing Navigation And Chemical Application in Precision Agriculture With Deep Reinforcement Learning And Conditional Action Tree. Smart Agricultural Technology, 101194. DOI: 10.1016/j.atech.2025.101194.

[8] Karunathilake, E. M. B. M., Le, A. T., Heo, S., Chung, Y. S., & Mansoor, S. (2023). The path to smart farming: Innovations and opportunities in precision agriculture. Agriculture, 13(8), 1593. DOI: 10.3390/agriculture13081593.

[9] Ayanlade, T. T., Jones, S. E., Laan, L. V. D., Chattopadhyay, S., Elango, D., Raigne, J., ... & Sarkar, S. (2024). Multi-modal AI for Ultra-Precision Agriculture. In Harnessing Data Science for Sustainable Agriculture and Natural Resource Management (pp. 299-334). Singapore: Springer Nature Singapore. DOI: 10.1007/978-981-97-7762-4_13.

[10] Anwar, B., Morsey, M. M., Hegazy, I., Fayed, Z. T., & El-Arif, T. (2024). Towards precise agriculture: integrating machine learning techniques for smart farming systems. IEEE Access. DOI: 10.1109/ACCESS.2024.3480868.

[11] Lu, F., Zhang, B., Hou, Y., Xiong, X., Dong, C., Lu, W., ... & Lv, C. (2025). A Spatiotemporal Attention-Guided Graph Neural Network for Precise Hyperspectral Estimation of Corn Nitrogen Content. Agronomy, 15(5), 1041. DOI: 10.3390/agronomy15051041.

[12] Wang, X., Yan, F., Li, B., Yu, B., Zhou, X., Tang, X., ... & Lv, C. (2025). A Multimodal Data Fusion and Embedding Attention Mechanism-Based Method for Eggplant Disease Detection. Plants, 14(5), 786. DOI: 10.3390/plants14050786.

[13] Jácome Galarza, L., Realpe, M., Viñán-Ludeña, M. S., Calderón, M. F., & Jaramillo, S. (2025). AgriTransformer: A Transformer-Based Model with Attention Mechanisms for Enhanced Multimodal Crop Yield Prediction. Electronics, 14(12), 2466. DOI: 10.3390/electronics14122466.

[14] Xu, D., Li, B., Xi, G., Wang, S., Xu, L., & Ma, J. (2025). A Shooting Distance Adaptive Crop Yield Estimation Method Based on Multi-Modal Fusion. Agronomy, 15(5), 1036. DOI: 10.3390/agronomy15051036.

[15] Zhao, J., Fan, S., Zhang, B., Wang, A., Zhang, L., & Zhu, Q. (2025). Research Status and Development Trends of Deep Reinforcement Learning in the Intelligent

Transformation of Agricultural Machinery. Agriculture, 15(11), 1223. DOI: 10.3390/agriculture15111223.

[16] Wang, Q., Zheng, S., Qiu, M., & Hu, D. (2025, February). Detection of pesticide residues by sensor arrays fused from SERS spectra of various substrates combined with deep learning. In Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security (pp. 359-366). DOI: 10.1145/3728725.3728783.

[17] Akintuyi, O. B. (2024). Adaptive AI in precision agriculture: a review: investigating the use of self-learning algorithms in optimizing farm operations based on real-time data. Research Journal of Multidisciplinary Studies, 7(02), 016-030. DOI: 10.53022/oarjms.2024.7.2.0023.

[18] Aviles Toledo, C., Crawford, M. M., & Tuinstra, M. R. (2024). Integrating multi-modal remote sensing, deep learning, and attention mechanisms for yield prediction in plant breeding experiments. Frontiers in Plant Science, 15, 1408047. DOI: 10.3389/fpls.2024.1408047.

[19] Yang, Z. X., Li, Y., Wang, R. F., Hu, P., & Su, W. H. (2025). Deep Learning in Multimodal Fusion for Sustainable Plant Care: A Comprehensive Review. Sustainability (2071-1050), 17(12). DOI: 10.3390/su17125255.

[20] Liu, Z., Li, S., Yang, Y., Jiang, X., Wang, M., Chen, D., ... & Dong, M. (2025). High-Precision Pest Management Based on Multimodal Fusion and Attention-Guided Lightweight Networks. Insects, 16(8), 850. DOI: 10.3390/insects16080850.

[21] Fei, S., Hassan, M. A., Xiao, Y., Su, X., Chen, Z., Cheng, Q., ... & Ma, Y. (2023). UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. Precision agriculture, 24(1), 187-212. DOI: 10.1007/s11119-022-09938-8.

[22] Aarif KO, M., Alam, A., & Hotak, Y. (2025). Smart sensor technologies shaping the future of precision agriculture: Recent advances and future outlooks. Journal of Sensors, 2025(1), 2460098. DOI: 10.1155/2025/2460098

[23] Diao, Z., Guo, P., Zhang, B., Yan, J., He, Z., Zhao, S., ... & Zhang, J. (2023). Spatial-spectral attention-enhanced Res-3D-OctConv for corn and weed identification utilizing hyperspectral imaging and deep learning. Computers and Electronics in Agriculture, 212, 108092. DOI: 10.1016/j.compag.2023.108092.

[24] Chacón-Maldonado, A. M., Asencio-Cortés, G., & Troncoso, A. (2025). A multimodal hybrid deep learning approach for pest forecasting using time series and satellite images. Information Fusion, 103350. DOI: 10.1016/j.inffus.2025.103350.

[25] Xu, K., Xie, Q., Zhu, Y., Cao, W., & Ni, J. (2025). Effective Multi-Species weed detection in complex wheat fields using Multi-Modal and Multi-View image fusion. Computers and Electronics in Agriculture, 230, 109924. DOI: 10.1016/j.compag.2025.109924.