

Generative AI-enhanced multimodal deep learning for pest and disease detection in Chinese herbal medicine

An Chang^a, Shalli Rani^b, Byung-Gyu Kim^c, Jianing Gong^a, Zhiyong Qin^a, Keqin Li^d, Jianhui Lv^{a,*}

^a The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, 121000, China

^b Chitkara University, Punjab, 140401, India

^c Sookmyung Women's University, Seoul, 04310, Republic of Korea

^d State University of New York, New Paltz, NY, 12561, USA

HIGHLIGHTS

- A generative AI-enhanced multimodal framework for accurate pest and disease detection is proposed.
- A cross-modal contrast learning mechanism with linguistic disease descriptions is designed.
- The Traditional Chinese Medicine-Disease dataset containing pest and disease manifestations is constructed.

ARTICLE INFO

Keywords:

Chinese herbal medicine
Pest and disease detection
Generative artificial intelligence
Multimodal learning
Few-shot learning
Smart agriculture

ABSTRACT

Chinese herbal medicine cultivation faces mounting pressures from pest infestations and disease outbreaks that compromise both yield quality and therapeutic efficacy. Traditional monitoring approaches struggle with the vast diversity of medicinal plant species and the subtle symptom variations across different growing environments. We present GAI-MDPS, a framework that integrates generative artificial intelligence with multimodal deep learning to enable accurate disease detection from limited training samples. The architecture employs vision transformers initialized with domain-specific disease attention bias priors for extracting global visual features, while leveraging pre-trained language models to generate contextual descriptions that guide visual attention toward diagnostically relevant image regions. Cross-modal contrastive learning aligns linguistic and visual representations, creating synergistic interactions where language understanding enhances visual interpretation. Comprehensive evaluation across PlantVillage and a self-constructed TCM-Disease dataset demonstrates substantial improvements over seven baseline methods, achieving 95.3% and 89.4% accuracy, respectively, with five-fold and ten-fold cross-validation confirming result stability. The generative module produces synthetic disease images with an overall FID of 23.6, and per-severity analysis shows a 9.1 %age point advantage over the best baseline at severity level 1, where early-stage detection delivers the highest practical value for timely intervention. The framework converges 2.1 times faster than existing approaches while maintaining superior generalization across diverse cultivation environments. Ablation studies reveal that multimodal integration reduces attention dispersion by 43.2% and improves lesion localization precision, enabling reliable automated monitoring for smart agriculture platforms serving Chinese herbal medicine production.

1. Introduction

Chinese herbal medicine cultivation occupies a unique position at the intersection of traditional medicine and contemporary

pharmaceutical development. More than 11,000 plant varieties have been recorded for medicinal utility in China alone [1]. Climate variability, intensive cultivation, and shifting pathogen populations are placing increasing pressure on this industry, threatening both crop

* Corresponding author.

E-mail addresses: changa@jzmu.edu.cn (A. Chang), shalli.rani@chitkara.edu.in (S. Rani), bg.kim@sookmyung.ac.kr (B.-G. Kim), gongjn@jzmu.edu.cn (J. Gong), qinzy@stu.jzmu.edu.cn (Z. Qin), lik@newpaltz.edu (K. Li), lvjh@jzmu.edu.cn (J. Lv).

<https://doi.org/10.1016/j.asoc.2026.116377>

Received 8 January 2026; Received in revised form 24 August 2026; Accepted 1 September 2026

Available online 7 September 2026

1568-4946/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

output and the therapeutic quality of harvested material [2]. Medicinal plants require precise control of secondary metabolite production, which determines therapeutic efficacy, whereas conventional agricultural crops are evaluated primarily on yield quantity [3].

The management of pests and diseases associated with Chinese herbal medicine is complicated by the fact that the number of species that are cultivated is absolutely extraordinary, and each of them has a different degree of exposure and a different profile of symptoms [4]. Traditional methods of diagnosis are largely based on the experience of experts over several centuries, but as the geography of cultivation and the appearance of new strains of pathogens are growing faster than the supply of qualified professionals. Moreover, the ban on synthetic pesticide use in certified organic as well as the Good Agricultural Practice (GAP) certified production systems requires timely identification and accurate identification in order to intervene with discriminating biological control plausible solutions [5,6]. Digital agriculture technologies are promising to transform the world, but their utilization has been hampered by factors such as a lack of annotated datasets on rare species of medicine and the insidious nature of early-stage infections [7].

Deep learning systems have transformed the detection of plant diseases in commercial agriculture by approaching expert accuracy on crops such as tomato, potato, and wheat with a well-known history [8]. Nevertheless, there are basic constraints also to the direct application of Chinese herbal medicine: the uneven distribution of available labeled data across species, with most medicinal plants having fewer than 100 annotated disease images imply that most medicinal plants have inadequate training data, and the morphological heterogeneity during different stages of growth makes the extraction of features complex. Generative AI offers a novel approach as it creates realistic training samples using a few samples, which basically increases the coverage of the databases used that belong to underrepresented disease categories [9]. Multimodal learning is also useful in increasing diagnostic potential (the use of textual descriptions in pharmacopoeias and current phytopathological databases) [10]. Muntaqim et al. [11] integrate federated learning with few-shot learning through ensemble voting to enable robust disease classification across heterogeneous, non-IID data distributions without requiring large centralized datasets. Additionally, Muntaqim et al. [12] propose a homomorphic encryption-enhanced federated learning framework that preserves data privacy during collaborative model training for agricultural disease detection.

Robustness to unreliable or partial modality input is treated as a design requirement rather than an afterthought in recent multimodal work outside plant pathology. Federated learning settings, where clients may hold noisy, missing, or heterogeneous modality combinations, have motivated uncertainty-weighted fusion schemes that down-weight unreliable modality streams during aggregation rather than treating every input as equally trustworthy [13] and selective state-space aggregation that carries an explicit uncertainty estimate through sequential multimodal updates [14]. Adversarial noise-imitation training during fusion has likewise been shown to improve resilience to corrupted or dropped input channels [15]. These approaches target sentiment and affect recognition rather than agricultural imagery. Nevertheless, the underlying premise, that a model deployed on incomplete field data should degrade gracefully rather than fail silently, motivates the sensor-dropout and prompt-corruption analyses.

Cross-modal alignment methods typically establish shared embedding spaces by minimizing distributional distances between visual and textual feature vectors, using objectives such as contrastive loss or optimal transport [16,17]. Applications of these methods to plant pathology have been limited by the scarcity of paired image-text data for specialty crops, which makes standard contrastive pretraining insufficient without domain adaptation. A separate line of work has demonstrated that encoding prior biological knowledge directly into model structure, either through auxiliary loss terms that penalize feature distributions inconsistent with known taxonomic relationships or through attention initialization priors reflecting anatomical symptom locations,

can improve sample efficiency when annotation budgets are constrained [18,19]. The third relevant direction concerns methods for detecting subtle lesion characteristics, including color gradients at infection margins, within-lesion texture heterogeneity, and the geometric spread patterns of different pathogen types [20]. Existing systems addressing this problem often require pixel-level segmentation labels that are prohibitively expensive to produce at scale for medicinal plant species. GAI-MDPS addresses all three gaps: spatially initialized disease attention priors encode anatomical domain knowledge, cross-modal contrastive alignment couples BERT-derived disease embeddings with vision transformer patch features, and generative augmentation synthesizes lesion-realistic images that are conditioned on both class prototypes and environmental context.

Our work introduces GAI-MDPS, a comprehensive framework that addresses these multifaceted requirements through the synergistic integration of generative AI, few-shot learning, and multimodal fusion. The framework leverages vision transformers to capture global spatial relationships in disease symptoms, while text encoders process semantic information from disease descriptions and environmental context. A novel cross-modal attention mechanism aligns visual and textual representations, enabling knowledge transfer from data-rich to data-scarce disease categories. Smart agriculture integration provides real-time environmental parameters that modulate diagnostic decisions, accounting for the phenotypic plasticity of symptoms under varying field conditions.

In summary, the main contributions are as follows:

- 1) We propose GAI-MDPS, a generative AI-enhanced multimodal framework that integrates vision transformers with language models for accurate pest and disease detection in Chinese herbal medicine cultivation using limited training samples.
- 2) We design a cross-modal contrast learning mechanism that aligns visual patch representations with linguistic disease descriptions, enabling attention refinement toward diagnostically relevant image regions through vision-language synergy.
- 3) We construct the Traditional Chinese Medicine (TCM)-Disease dataset containing pest and disease manifestations across multiple Chinese herbal medicine species, addressing the critical shortage of specialized labeled data for medicinal plant pathology research.

The rest of the paper is structured as follows. [Section 2](#) describes the materials and methods employed in this study. [Section 3](#) presents experimental results and analysis. [Section 4](#) gives the discussions. Finally, [Section 5](#) concludes this paper.

2. Materials and methods

2.1. Disease dataset construction

The GAI-MDPS framework is a complex of multiple artificial intelligence paradigms to fight the issues based on identifying the disease, which uses the context of the Chinese herbal medicine farming. The architecture represents the spatial decoders of features, semantic comprehension models, data augmentation, generative AI, and a cross-modal fusion system in a single architecture focused on overall disease recognition.

The experimental validation framework applies two supplementary datasets, which are used to cover, in a systematic manner, general principles of plant pathology and, at the same time, the special circumstances of the cultivation of the Chinese herbal medicine. The offered solution, based on two datasets, will allow the proposed approach to give a holistic evaluation of the offered GAI-MDPS framework in various agricultural contexts without becoming uninterested in the question of the identification of medicinal plant diseases.

2.1.1. PlantVillage foundation dataset

It is possible to assume the PlantVillage collection as the starting point for where the transfer learning and evaluation of cross-species generalization can be conducted, since the duo has 54,303 good-quality images, which have 14 examples of crop species, and 38 types of different diseases [21]. Despite the fact that in this dataset the main focus is placed on traditional farming products rather than on medicinal plants, considerable background training on the typical pathological disorders is generated, since it is significant in our frame of vision transformer components. The images in the PlantVillage collection have their images standardized to precheck the images against the baseline of 224×224 -pixels at normalized RGB channel values (0–1) to guarantee that the calculations remain uniform throughout the training pipeline.

The advantage of the utilization of the PlantVillage is that it depicts the overall tendencies in the patterns of diseases, which are not restricted to a specific kind of crop. Fungal, bacterial, viral symptoms, and nutrient deficiencies share the same appearance in the various plant families, and this enables the model to acquire generalized pathological properties with the capacity to specialize in medicinal species. Besides, the overall quality of annotation of the dataset that has been checked by a vast number of reviews by the experts creates a good ground truth for the early convergence of the model.

2.1.2. TCM-Disease dataset development

The TCM-Disease data was made to suit the definite diagnostic requirement of the Chinese herbal medicine development and was hence developed by conducting organized field surveys of interest areas of production in the 2023–2024 growing seasons. The specified specialized collection is a significant move in the history of the medicinal plant pathology documentation that considers not only the traditional Chinese medicine methods of the diagnostic view, but the modern system of phytopathological taxonomy as well.

This information is a list of twelve priority medicinal species, which were selected after extensive consultations with the professionals working in the traditional medicine, agricultural extension, and the representatives of the pharmaceutical industry. These species do not belong to the same plant families, development forms, and farming requirements, which ensures a versatile range of applicability of the involved trained models: The selection includes *Panax ginseng*, an adaptogenic species that has a characteristic disease pressure phenotype given the requirement to grow under special conditions in the shade. Honeysuckle *Lonicera japonica* is particularly highly sensitive to powdery mildew infections that compromise the quality of flowers and concentrations of numerous medicinal substances. Plantation Crops During the extended growing season, there are thorny pests and disease issues that accompany the growth of *Chrysanthemum morifolium*. The primary hazard to the economy is the white rust and aphid destruction. The *astragalus membranaceus*, which is a key ingredient of the classic formula, is also susceptible to root rot pathogens in soils with low drainage capacity, which is characteristic of the areas of large-scale growing.

Other rust diseases may also attack other species like *Glycyrrhiza uralensis*, which causes the plant's foliar symptoms to be indistinguishable from drought stress. The increasing number of insect vectors is straining *Salvia miltiorrhiza* production due to the risk of the transmission of viral diseases through insect vectors. A special strain of pathogens that become accustomed to the cool and damp environment confronts the ecology of *Angelica sinensis* in the high-altitude environment. The ability of multiple pathogens to infect *Rheum palmatum* at the same time has complicated the symptomology of the disease, which complicates the use of separate disease diagnosis procedures.

Coptis chinensis is also present in the data, where cultivation of shade characteristics similar to those of ginseng would offer a good environment for the growth of fungi. *Gastrodia* possess unique disease symptoms, and *Gastrodia* with a single mycoheterotroph growth pattern have their own fungal synergetism. The *Dendrobium nobile*

epiphytic orchid is faced with some specialized pathogens that grow on the trees naturally, as air-grown. Finally, *Fritillaria thunbergii* is a bulbous medicinal plant especially vulnerable to the disease transmitted by soil during the time of dormancy.

2.1.3. Image acquisition and annotation protocols

The field collection campaigns were carried out in eight provinces that are the major production centers of Chinese herbal medicine, such as ginseng, produced in Jilin and Heilongjiang, *chrysanthemum*, produced in Zhejiang and Anhui, licorice, produced in Gansu and Inner Mongolia, and diverse high-altitude medicinal species, produced in Sichuan and Yunnan. This spatial diffusion will ensure the strength of models in various climatic conditions, soils, and also methods of cultivation.

The final TCM-Disease collection is that of 8432, which had annotated images that represented 47 distinct disease-pathogen combinations with economic and significant priorities. A smartphone camera (12–48 megapixels) and professional DSLR cameras were utilized to get images, and the identical capture process was followed; thus, the quality of a diagnosis could be guaranteed. Photography was carried out in the natural lighting conditions that covered a period of 9:00 am to 4:00 pm, so as not to cause interference with the shadows, but still in the real field conditions [22]. A few photographs of each of the diseased specimens were obtained at fixed distances of 10 cm (one can see the disease), 30 cm (organ condition inspection), and 100 cm (a plan of the whole plant inspection).

All images have been highly annotated by qualified plant pathologists who have special expertise in medicinal plant diseases. The annotation protocol had a two-fold classification framework with both modern pathology nomenclature and the customary Chinese medicine syndrome distinction nomenclature noted. The level of disease severity was evaluated using a standardized scale of 0–5, with zero denoting a healthy tissue, 1 signifying less than 5% area affected, 2 depicting a 5–15% infection, 3 indicating 15–35% tissue disease, 4 representing 35–65% tissue disease, and 5 denoting more than 65% tissue disease.

Each image has extensive metadata, such as GPS position with an accuracy of 5 m, a time stamp within seconds, the phenological stage of growth based on the BBCH scale pushed into medicinal use, a practice of cultivation as an organic one, as a GAP-certified participant, or a conventional, and the 7-day weather forecast in the form of temperature range, rainfall, and relative humidity. Parameters in soil samples at the sites of collection include pH, the content of organic matter, the amount of accessible nitrogen, phosphorus, and potassium. This ecological setting is critical in explaining variability in symptom expression and developing resilient frameworks explaining genotype-environment-pathogen interaction.

2.1.4. Dataset partitioning and few-shot task construction

The dataset format is episodic training frameworks on which few-shot learning is based. The PlantVillage collection is further divided into the source domain classes comprising 29 disease classes to train models and initialize them, and nine classes to be used in the target domain to test the potential of cross-species generalization. This division does not entail overlaps in the types of training and evaluation of diseases, but provides a rigid assessment of the model's ability to discover new pathological conditions.

TCM-Disease samples organize into episodic tasks following the N -way K -shot paradigm. Each training episode samples $N = 5$ disease categories, with each category containing $K = 5$ support images for prototype learning and $Q = 15$ query images for evaluation. This configuration balances computational efficiency with sufficient statistical power for gradient estimation. The support set deliberately includes images captured under varying environmental conditions to encourage robust feature learning, while query sets incorporate both typical and edge-case presentations to thoroughly evaluate diagnostic capability.

Test episodes sample disease categories that receive only the K -shot support images during Stage 3 fine-tuning and never appear in the base-class pretraining of Stage 1; we refer to these below as low-shot rather than fully unseen categories, since the text branch's unsupervised masked-language-model pretraining corpus contains descriptive passages for all 47 TCM-Disease combinations, including these categories.

Class imbalance within the TCM-Disease collection was handled through two mechanisms applied jointly. Underrepresented disease categories, defined as those with fewer than 30 annotated images, appeared in training episodes at a frequency inversely proportional to their sample count through a weighted episode sampler. The generative augmentation module additionally produced synthetic images for these categories until each reached a minimum of 150 examples before episodic training began, preventing the model from developing a systematic preference toward more frequently documented diseases.

2.1.5. Quality assurance and preprocessing pipeline

The initial data collected due to motion blur, low resolution, and unspecified symptoms requiring lab confirmation were deleted or found in the process of data quality control, amounting to 3.2 per cent. The remainder of the images were systematically preprocessed in order to normalize the input without affecting diagnostic features. Advanced processing is done on the basis of histogram equalization and is an adaptively controlled contrast adjustment in varying light situations without excessive removal of color detail considered essential to disease detection. Background segmentation using DeepLab v3 + removes non-plant objects when they can assist in modeling focus, but keeps full-context images when diseases whose factors are environmental can be diagnostically useful [23].

Random horizontal flipping (probability=0.5) adds effective data size, random cropping (80–100 per cent of original image size) adds scale invariance, color jittering (maximumdelta=0.2) adds brightness contrast, and saturation adds illumination resistance, and random rotation (maximumdegree=15) is added as an augmentation measure during training. These augmentations are probabilistic during training, and when using validation and test, original images are used to ensure accurate assessment.

The seven environmental sensor channels, namely ambient temperature, relative humidity, photosynthetically active radiation, soil pH, and plant-accessible nitrogen, phosphorus, and potassium, were preprocessed before they entered the model. Each channel was first clipped to the 1st and 99th percentiles of its training-split distribution to limit the influence of faulty readings, then standardized to zero mean and unit variance using statistics computed on the training split alone, so that no validation or test information leaked into the normalization. The standardized values formed the seven-dimensional vectors used by both the generator and the attention modulation pathway. When a channel was missing from the field metadata, it was set to the training mean, which becomes zero after standardization and corresponds to the null-sensor condition. The same fixed training statistics were reused without recomputation at validation and test time.

The complete data model has the full-scale testing of the GAI-MDPS model under a range of disease detection conditions. The specialization of medicinal plants proposed by TCM-Dise, and the versatility of the PlantVillage, also gave some interesting information on finding cross-species disease patterns, and applies to the specific demands of the Chinese herbal medicine development. Such a dual dataset approach allows developing a few-shot system of learning in theory, as well as to apply it practically in the context of agricultural disease management infrastructure.

Fig. 1 presents sample disease manifestations as they have been captured in the TCM-Disease dataset, and reveals the visual complexity that presents itself to automated detection systems. *Cylindrocarpus destructans* induces ginseng root rot characterized by dark brown discoloration that begins in fibrous roots and progresses to the taproot, depleting ginsenoside content through soft tissue degradation. Powdery

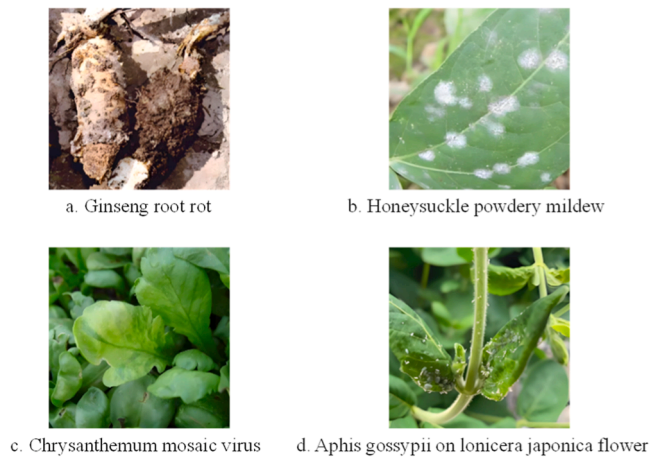


Fig. 1. Representative pest and disease manifestations in Chinese herbal medicine cultivation.

colonies of *Erysiphe lonicerae* on honeysuckle leaves must be distinguished from spray residues and accumulated dust in field settings, as the two appear similar under standard RGB imaging. Chrysanthemum mosaic virus produces yellowing with sharp boundaries between affected and healthy tissue, requiring differentiation from nutritional deficiencies that produce similar but diffuse chlorotic patterns. *Aphis gossypii* feeding on *Lonicera japonica* flower buds causes leaf curl deformation and honeydew secretion that promotes colonization by secondary sooty mold [24].

These visual examples underscore the sophisticated pattern recognition requirements for effective disease detection in Chinese herbal medicine cultivation, where accurate diagnosis demands discrimination among subtle symptom variations across morphologically diverse plant species under variable environmental conditions.

2.2. Multimodal few-shot disease recognition model construction

Although the individual building blocks of GAI-MDPS, including vision transformers, BERT-based text encoding, contrastive cross-modal alignment, prototype-based few-shot classification, and conditional generative augmentation, each have prior instantiations in the broader multimodal learning literature, the technical novelty of the framework rests on three specific design decisions that have not been combined in this form and that are non-trivial relative to existing multimodal few-shot recognition systems. First, the vision transformer's multi-head self-attention is initialized with a learnable disease attention bias matrix $\mathbf{M}_{\text{disease}} \in \mathbb{R}^{197 \times 197}$ that encodes spatial priors derived from phytopathological knowledge about where different pathogen types manifest on plant tissue. Standard vision transformer implementations initialize this bias at zero or learn it purely from data; the structured domain-specific initialization proposed here reduces the number of training episodes required to develop diagnostically focused attention, as evidenced by the 2.1 times faster convergence reported in Section 3.8, and is qualitatively different from prior attention-prior work that derives spatial priors from image statistics rather than biological domain knowledge. Second, the text description pipeline treats TCM syndrome differentiation terminology as a mandatory structural component of every prompt rather than as an optional annotation layer. Prior multimodal plant disease systems that incorporate textual information use botanical symptom language or chemical compound descriptions; the TCM syndrome patterns used here encode centuries of empirical clinical observation about disease progression trajectories, co-occurring symptom clusters, and species-specific pathological responses that have no equivalent in standard phytopathological databases. Third, the conditional GAN generator receives the environmental sensor vector $\mathbf{s} =$

$[T, H, L, pH, N, P, K]^T$ as an explicit conditioning input alongside the class prototype vector $\mathbf{p}_c$ and latent code $\mathbf{z}$, so that synthetic images represent the specific symptom variation induced by real field conditions rather than an average symptom appearance pooled across all cultivation environments. Prior conditional GAN augmentation work in plant disease detection conditions the generator on class labels or prototype vectors alone; the environmental conditioning in GAI-MDPS enables the model to learn how temperature, humidity, soil chemistry, and light intensity modulate symptom expression, which is particularly important for Chinese herbal medicine species because their secondary metabolite profiles cause disease appearances to vary with cultivation conditions in ways that differ substantially from food crops.

The GAI-MDPS framework represents an integration of multiple artificial intelligence paradigms to address the challenges of disease detection in Chinese herbal medicine cultivation. The architecture synthesizes vision transformers for spatial feature extraction, language models for semantic understanding, generative AI for data augmentation, and cross-modal fusion mechanisms for holistic disease recognition.

2.2.1. Overall GAI-MDPS framework architecture

We follow the standard few-shot learning notation established in [25,26] and adopt the N-way K-shot episodic framework. The complete dataset contains paired RGB images, disease labels, text descriptions, and sensor readings. The label space partitions into base classes with abundant labeled images and novel classes with fewer than 20 images per category, where these two sets are disjoint, and their union covers the full label space [27,28].

The complete GAI-MDPS architecture, illustrated in Fig. 2, orchestrates four primary computational pathways that process multimodal inputs through specialized neural architectures before integrating their outputs for final disease classification. The framework design philosophy emphasizes modularity and interpretability, enabling researchers to understand how different information sources contribute to diagnostic decisions.

The vision branch processes disease images through a hierarchical feature extraction pipeline, transforming raw pixel data into semantically meaningful representations that capture both local symptom details and global disease patterns. Simultaneously, the text branch encodes linguistic descriptions from multiple sources, including traditional Chinese medicine texts, modern phytopathological databases, and field observation notes, creating rich semantic embeddings that complement visual features.

The generative augmentation engine uses a modified StyleGAN2 architecture. The style mapping network receives as input the concatenation of a 512-dimensional class prototype vector $\mathbf{p}_c$ (the mean of the support set features for class c), the 7-dimensional environmental sensor vector $\mathbf{s} \in \mathbb{R}^7$, and a 128-dimensional latent code $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ drawn from a unit Gaussian. This 647-dimensional input $[\mathbf{p}_c; \mathbf{s}; \mathbf{z}] \in \mathbb{R}^{647}$ is passed through an 8-layer fully connected mapping network with LeakyReLU activations to produce a 512-dimensional style vector $\mathbf{w} \in \mathbb{R}^{512}$.

The synthesis network generates images at 224×224 resolution through seven progressive upsampling blocks using weight-demodulated convolutions. The discriminator uses a dual-head design: one head computes the Wasserstein distance loss $\mathcal{L}_{WD}$, and a second head predicts whether the generated image is semantically consistent with its associated text description embedding $\mathbf{f}_t$, providing an auxiliary alignment objective during GAN training. The GAN was trained for 50,000 iterations using the Adam optimizer with a learning rate of $\eta = 2 \times 10^{-4}$ and momentum parameters $(\beta_1, \beta_2) = (0.0, 0.99)$, and a gradient penalty coefficient of $\lambda_{gp} = 10$ for the Wasserstein objective. Generated images were accepted into the training pool only when their LPIPS score, computed against the nearest real image $\mathbf{x}_{real}$ in the support set, satisfied $LPIPS(\mathcal{G}(\mathbf{z}, \mathbf{y}, \mathbf{s}), \mathbf{x}_{real}) < 0.35$, preventing low-quality samples from degrading the episodic training signal [29,30]. This component employs a modified StyleGAN architecture adapted for plant pathology, generating synthetic training samples that maintain biological plausibility while increasing effective dataset size. The synthesis process incorporates environmental sensor data $\mathbf{s} = [T, H, L, pH, N, P, K]^T$ to ensure generated images reflect realistic symptom variations under different cultivation conditions. Here T denotes ambient temperature in degrees Celsius, H represents relative humidity as a percentage, L indicates photosynthetically active radiation in micromoles per square meter per second, pH is the soil acidity value, and N, P , and K represent plant-accessible nitrogen, phosphorus, and potassium concentrations in milligrams per kilogram, respectively.

At inference time, the generative augmentation engine plays no role. The model receives a real, unlabeled query image and a fixed support set of K labeled examples per class, each support image paired with a text description generated from the combined template. The query image carries no text input of any kind: at no point does the model have access to a description of the query's own species, symptoms, TCM syndrome pattern, or pathogen identity, since that identity is exactly what is being predicted. Classification proceeds by computing the query's visual embedding, comparing it with the class prototypes derived from the support set's fused visual-text representations using the learned distance d_ϕ in Eq. (22), and assigning the label of the nearest prototype. No ground-truth disease identity for the sample under evaluation is provided to the model at test time; the GAN generator and its conditioning network are likewise not invoked, as previously stated.

The generative AI module operates exclusively during training. It accepts a class-prototype vector computed from the episode support set, concatenated with the environmental sensor vector $\mathbf{s}$, and a latent code $\mathbf{z}$ sampled from a unit Gaussian prior. The conditional generator then produces synthetic disease images whose symptom appearance reflects the environmental conditions encoded in $\mathbf{s}$. These images are inserted into the support set of the corresponding class before each training episode. At inference time, the generator is inactive, and the model classifies query images using only real support examples.

Cross-modal fusion mechanisms align representations from different

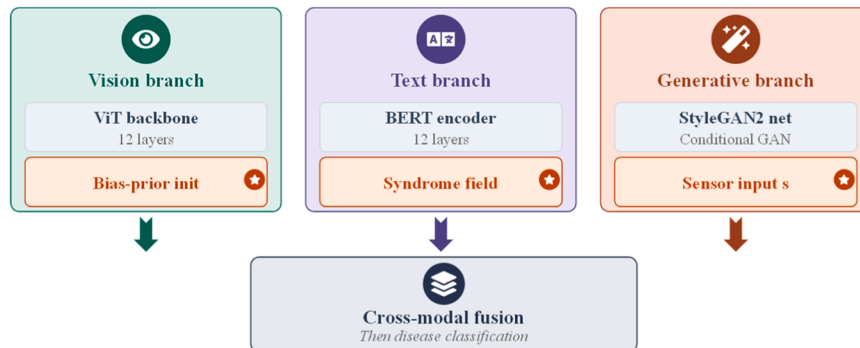


Fig. 2. GAI-MDPS multimodal few-shot learning framework architecture.

modalities through contrastive learning objectives, establishing a shared semantic space where visual and textual features can be directly compared [31,32]. The alignment process employs a temperature-scaled similarity function:

$$\text{sim}(f_v, f_t) = \frac{\exp(\langle \hat{f}_v, \hat{f}_t \rangle / \tau)}{\sum_{i=1}^B \exp(\langle \hat{f}_v, \hat{f}_{ti} \rangle / \tau)} \quad (1)$$

where $\hat{f}_v = f_v / \|f_v\|_2$ and $\hat{f}_t = f_t / \|f_t\|_2$ represent L2-normalized features, $\langle \cdot, \cdot \rangle$ denotes inner product, τ indicates temperature parameter controlling distribution sharpness, and B represents batch size.

The framework incorporates smart agriculture integration through real-time sensor data processing, enabling context-aware disease diagnosis that accounts for environmental factors influencing symptom expression [33]. Sensor readings modulate attention weights in both vision and text branches, allowing the model to adjust its feature extraction focus based on current field conditions.

2.2.2. Few-shot learning (FSL) image branch architecture

The visual encoding pathway employs a vision transformer (ViT-B/16) architecture specifically adapted for few-shot learning constraints in agricultural applications [34,35]. This design choice reflects extensive empirical validation demonstrating the superior performance of transformer architectures for capturing long-range spatial dependencies critical in disease pattern recognition. Fig. 3 illustrates the detailed structure of our FSL visual feature extraction branch, demonstrating how disease images undergo transformation from raw pixels to discriminative feature representations suitable for few-shot classification.

Input images $x \in \mathbb{R}^{224 \times 224 \times 3}$ first undergo patch embedding through non-overlapping spatial division into 14×14 patches of size 16×16 pixels. Each patch p_i transforms into a d -dimensional embedding through learnable linear projection:

$$z_0^{(i)} = p_i W_{\text{embed}} + e_{\text{pos}}^{(i)} + e_{\text{type}} \quad (2)$$

where $W_{\text{embed}} \in \mathbb{R}^{768 \times 768}$ represents the embedding projection matrix, $e_{\text{pos}}^{(i)} \in \mathbb{R}^{768}$ provides learnable positional encoding for spatial location i , and $e_{\text{type}} \in \mathbb{R}^{768}$ indicates whether the token represents an image patch or the special [CLS] token used for global representation.

The transformer encoder consists of $L = 12$ identical layers, each containing multi-head self-attention (MHSA) and position-wise feed-

forward network (FFN) sublayers with residual connections and layer normalization:

$$z'_\ell = \text{LayerNorm}(\text{MHSA}(z_{\ell-1}) + z_{\ell-1}). \quad (3)$$

$$z_\ell = \text{LayerNorm}(\text{FFN}(z'_\ell) + z'_\ell). \quad (4)$$

The multi-head self-attention mechanism computes dot-product attention across $h = 12$ parallel attention heads [36]:

$$\text{MHSA}(Z) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O. \quad (5)$$

where each attention head performs:

$$\text{head}_i = \text{Attention}(Z W_i^Q, Z W_i^K, Z W_i^V). \quad (6)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M_{\text{disease}}\right) V. \quad (7)$$

The learnable disease attention bias matrix $M_{\text{disease}} \in \mathbb{R}^{197 \times 197}$ encodes spatial priors for common disease manifestation patterns. For instance, fungal infections typically initiate at leaf margins, viral symptoms often display systemic distribution, and root diseases manifest in lower plant portions. This architectural innovation enables the model to incorporate domain knowledge while maintaining flexibility for novel disease patterns.

The feed-forward network applies two linear transformations with GELU activation:

$$\text{FFN}(z) = \text{GELU}(z W_1 + b_1) W_2 + b_2. \quad (8)$$

where $W_1 \in \mathbb{R}^{768 \times 3072}$, $W_2 \in \mathbb{R}^{3072 \times 768}$ represent weight matrices, and b_1, b_2 denote bias terms.

Following transformer encoding, the global [CLS] token representation undergoes disease-specific projection through a multi-layer perceptron:

$$f_v = \text{MLP}_{\text{disease}}(z_L^{[\text{CLS}]}) = \text{ReLU}(\text{BN}(z_L^{[\text{CLS}]} W_{\text{proj1}})) W_{\text{proj2}}. \quad (9)$$

where $W_{\text{proj1}} \in \mathbb{R}^{768 \times 1024}$ and $W_{\text{proj2}} \in \mathbb{R}^{1024 \times 512}$ project features to the shared multimodal space, while batch normalization (BN) stabilizes training dynamics.

2.2.3. Text encoding branch

The semantic processing pathway leverages a pre-trained Chinese-BERT-wwm (Whole Word Masking) model fine-tuned on agricultural and traditional medicine corpora [37,38]. This specialized language model captures nuanced relationships between symptom descriptions, pathogen characteristics, and traditional diagnostic terminology crucial for accurate disease identification. Fig. 4 illustrates the architectural design of the text head module, which serves as the critical interface between pre-trained language representations and the disease-specific semantic features required for effective multimodal alignment.

Input text sequences combine multiple information sources formatted as: "[CLS] [Species] [SEP] [Symptoms] [SEP] [TCM Pattern] [SEP] [Environment] [SEP]". This structured input enables the model to differentiate information types while learning cross-segment relationships.

Text descriptions were drawn from three sources: phytopathological monographs documenting Chinese medicinal plant diseases, pharmacopeia entries describing syndrome differentiation patterns for each species, and structured field observation notes recorded by the annotation team during the 2023–2024 data collection campaigns. Two domain experts with dual training in plant pathology and traditional medicine reviewed each description and standardized it to a canonical format before inclusion in the training corpus.

The tokenization process employs WordPiece vocabulary extended with 3847 domain-specific terms, including medicinal plant names, pathogen nomenclature, and traditional medicine terminology. Each

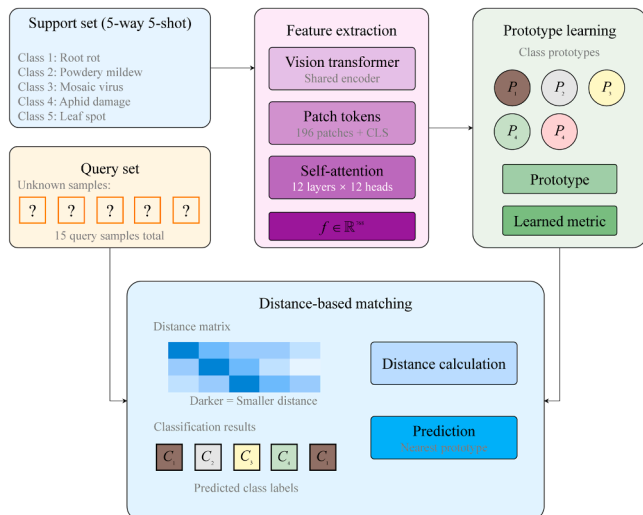


Fig. 3. FSL branch architecture for disease recognition.

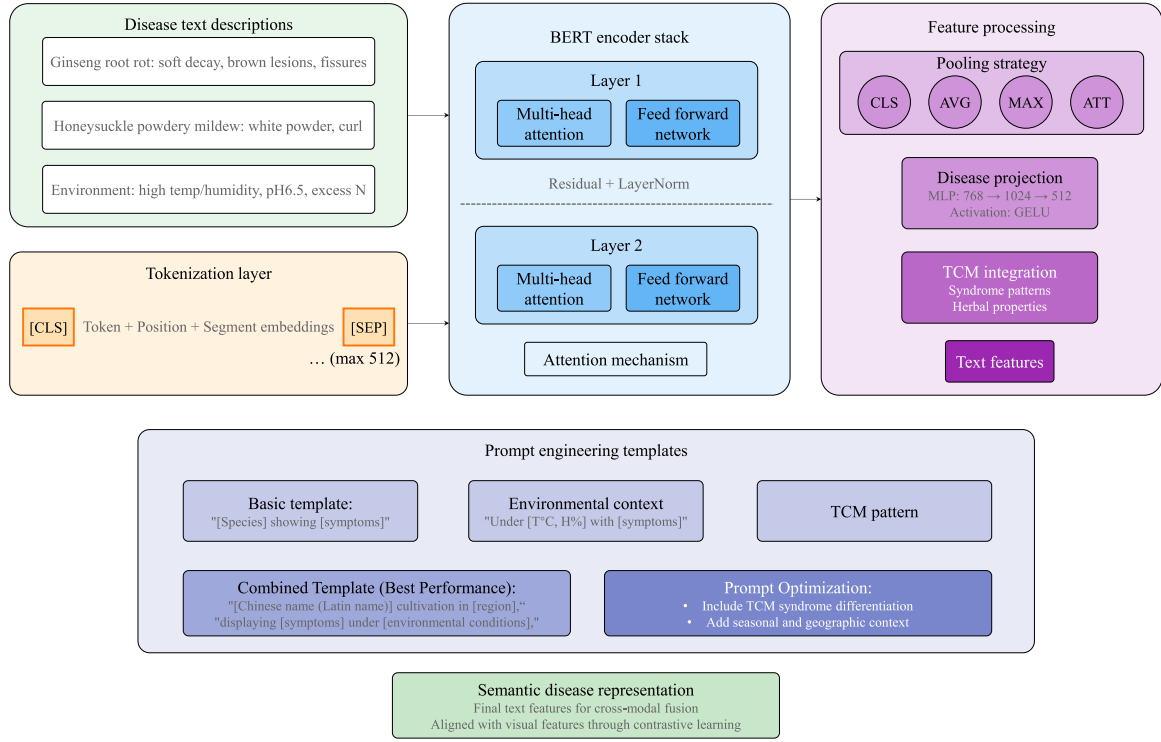


Fig. 4. Text head structure for semantic disease processing.

token t_i maps to a dense embedding vector:

$$h_0^{(i)} = \text{Embed}(t_i) + e_{\text{position}}^{(i)} + e_{\text{segment}}^{(i)}. \quad (10)$$

where embedding dimension equals 768, position embeddings support sequences up to 512 tokens, and segment embeddings distinguish information categories.

The BERT encoder processes token sequences through 12 transformer layers with architecture identical to the original BERT-Base configuration but with weights adapted through domain-specific pre-training on 2.3 million agricultural text documents. Each layer applies:

$$h_\ell = \text{TransformerLayer}(h_{\ell-1}). \quad (11)$$

The domain-specific pre-training corpus of 2.3 million agricultural text documents was assembled from four sources: the Chinese National Knowledge Infrastructure database queried for phytopathology and traditional medicine cultivation publications from 1990 to 2023; the 2020 Chinese Pharmacopoeia, digitized and segmented into disease description passages; plant disease survey reports and agricultural extension bulletins from the Chinese Academy of Agricultural Sciences; and 47,000 structured disease description records created by the annotation team using the Combined prompt template for each of the 47 disease-pathogen combinations in TCM-Disease. The first three sources are subject to licensing restrictions that prevent direct public redistribution. The 47,000 manually created records, which are the most directly relevant component for reproducing the multimodal alignment in GAI-MDPS. Pre-training on the full corpus used a masked language modeling objective for 20 epochs using AdamW with a learning rate of $2e-5$, batch size 64, and a warmup ratio of 0.06.

The final semantic representation aggregates information across all tokens using attention-weighted pooling:

$$\alpha_i = \frac{\exp(w_{\text{att}}^T \tanh(W_{\text{att}} h_L^{(i)}))}{\sum_{j=1}^n \exp(w_{\text{att}}^T \tanh(W_{\text{att}} h_L^{(j)}))}. \quad (12)$$

$$f_{\text{t.pool}} = \sum_{i=1}^n \alpha_i h_L^{(i)}. \quad (13)$$

where $W_{\text{att}} \in \mathbb{R}^{768 \times 768}$ and $w_{\text{att}} \in \mathbb{R}^{768}$ represent learned attention parameters.

The pooled representation undergoes disease-specific transformation:

$$f_t = \text{Dropout}(0.1)(\text{GELU}(\text{LayerNorm}(f_{\text{t.pool}} W_{\text{text1}}))) W_{\text{text2}}. \quad (14)$$

where $W_{\text{text1}} \in \mathbb{R}^{768 \times 1024}$ and $W_{\text{text2}} \in \mathbb{R}^{1024 \times 512}$ project text features to match visual feature dimensionality.

2.2.4. Cross-modal contrast module

The alignment between visual and textual modalities employs a sophisticated contrastive learning framework that establishes correspondence between disease images and their semantic descriptions. This cross-modal alignment enables the model to leverage textual knowledge for improving visual recognition and vice versa.

Given a batch of paired representations $\{(f_v^i, f_t^i)\}_{i=1}^B$, the contrastive objective maximizes similarity between matched pairs while minimizing similarity with non-matching samples:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{2B} \sum_{i=1}^B \left[\log \frac{\exp(\text{sim}(f_v^i, f_t^i))}{\sum_{j=1}^B \exp(\text{sim}(f_v^i, f_t^j))} + \log \frac{\exp(\text{sim}(f_t^i, f_v^i))}{\sum_{j=1}^B \exp(\text{sim}(f_t^i, f_v^j))} \right]. \quad (15)$$

The cross-attention mechanism enables bidirectional information exchange between modalities:

$$f_{v \rightarrow t} = \text{CrossAttention}\left(f_v, f_t, f_t\right) = \text{softmax}\left(\frac{f_v W_Q^v (f_t W_K^t)^T}{\sqrt{d}}\right) f_t W_V^t. \quad (16)$$

$$f_{t \rightarrow v} = \text{CrossAttention}\left(f_t, f_v, f_v\right) = \text{softmax}\left(\frac{f_t W_Q^t (f_v W_K^v)^T}{\sqrt{d}}\right) f_v W_V^v. \quad (17)$$

where projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{512 \times 512}$ learn modality-specific transformations.

The final fused representation combines aligned features using adaptive weighting:

$$f_{\text{fused}} = \alpha \cdot \text{LayerNorm}(f_{v \rightarrow t} + f_v) + (1 - \alpha) \cdot \text{LayerNorm}(f_{t \rightarrow v} + f_t). \quad (18)$$

where the mixing coefficient α is dynamically computed based on input quality:

$$\alpha = \text{sigmoid}(w_a^T [f_v; f_t; f_v \odot f_t; f_v - f_t]). \quad (19)$$

with $w_a \in \mathbb{R}^{2048}$ as learnable parameters and $\odot$ denoting element-wise multiplication.

2.2.5. Loss function construction

The comprehensive training objective integrates multiple loss components that collectively optimize disease recognition performance while maintaining feature quality and preventing overfitting:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{FSL}} + \lambda_2 \mathcal{L}_{\text{contrast}} + \lambda_3 \mathcal{L}_{\text{gen}} + \lambda_4 \mathcal{L}_{\text{consistency}} + \lambda_5 \mathcal{L}_{\text{reg}}. \quad (20)$$

Each term in $\mathcal{L}_{\text{total}}$ addresses a distinct failure mode that the others do not cover on their own. $\mathcal{L}_{\text{FSL}}$ drives the episodic classification objective and is therefore assigned the reference weight $\lambda_1 = 1.0$ against which the remaining terms are scaled. $\mathcal{L}_{\text{contrast}}$ pulls matched image-text pairs together in the shared embedding space; without it, the visual and textual branches are optimized independently, and the cross-attention layers in Eqs. (16)-(17) have no incentive to align their query and key projections across modalities. $\mathcal{L}_{\text{gen}}$ keeps the generator's output distribution close to the real disease-image distribution, preventing the low-diversity or unrealistic samples that would otherwise degrade the episodic training signal once synthetic images enter the support set. $\mathcal{L}_{\text{consistency}}$ penalizes prediction disagreement between two augmented views of the same image and acts as a data-driven regularizer that is independent of the generative pathway. $\mathcal{L}_{\text{reg}}$ combines standard weight decay with the attention-diversity penalty to keep individual heads from collapsing onto redundant token subsets. Because $\mathcal{L}_{\text{FSL}}$ and $\mathcal{L}_{\text{contrast}}$ operate on every training batch while $\mathcal{L}_{\text{gen}}$ only applies to the subset of batches containing synthetic images, the relative weights λ_2 through λ_5 were tuned in that order against the fixed λ_1 .

The few-shot learning loss employs prototypical networks with learned distance metrics:

$$\mathcal{L}_{\text{FSL}} = -\frac{1}{|Q|} \sum_{(x_q, y_q) \in Q} \log \frac{\exp(-d_\phi(f_q, p_{y_q}))}{\sum_{c=1}^N \exp(-d_\phi(f_q, p_c))}. \quad (21)$$

where $p_c = \sum_{i: y_i=c} f_i / K$ represents the prototype for class c , and d_ϕ denotes a learned Mahalanobis distance:

$$d_\phi(f_1, f_2) = \sqrt{(f_1 - f_2)^T M_\phi (f_1 - f_2)}. \quad (22)$$

with $M_\phi \in \mathbb{R}^{512 \times 512}$ as a learned positive semi-definite matrix.

The generative augmentation loss combines four terms rather than three. In addition to the pixel-level and perceptual fidelity terms and the Wasserstein adversarial term, the semantic-consistency head (the discriminator's second output, which estimates whether a generated image matches its paired text embedding f_t) contributes a fourth term:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{z \sim p(z), y \sim p(y)} [\|\mathcal{G}(z, y, s) - x_{\text{real}}\|_1 + \beta \cdot \text{LPIPS}(\mathcal{G}(z, y, s), x_{\text{real}}) + \gamma \cdot \mathcal{L}_{\text{GAN}} + \delta \cdot \mathcal{L}_{\text{sem}}]. \quad (23)$$

where $\mathcal{L}_{\text{sem}} = -\log D_{\text{sem}}(\mathcal{G}(z, y, s), f_t)$ and $D_{\text{sem}}(\cdot, \cdot) \in (0, 1)$ is the discriminator head's estimate that the generated image is semantically consistent with f_t . This term is what allows the discriminator's auxiliary alignment objective, actually to shape the generator's parameters during backpropagation: the reconstruction and LPIPS terms enforce pixel- and feature-level similarity to a specific reference image, while $\mathcal{L}_{\text{sem}}$ enforces agreement with the class-level text description independent of any single reference image, and $\mathcal{L}_{\text{GAN}}$ enforces that the output lies on the manifold of realistic disease images regardless of text or pixel matching. δ was set to 0.1 by the same held-out procedure used for β and γ .

The consistency loss enforces prediction stability across augmented views:

$$\mathcal{L}_{\text{consistency}} = \mathbb{E}_x [\text{KL}(p_\theta(y|\mathcal{A}_1(x)) \| p_\theta(y|\mathcal{A}_2(x)))]. \quad (24)$$

where $\mathcal{A}_1, \mathcal{A}_2$ represent different augmentation functions and KL denotes Kullback-Leibler divergence.

The regularization term prevents overfitting through weight decay and attention diversity:

$$\mathcal{L}_{\text{reg}} = \omega_1 \|\theta\|_2^2 + \omega_2 \sum_{\ell=1}^L \left\| A_\ell^T A_\ell - \frac{1}{h} I \right\|_F^2. \quad (25)$$

where $A_\ell \in \mathbb{R}^{h \times n \times n}$ represents attention maps from layer ℓ , encouraging diverse attention patterns across heads.

2.2.6. Experimental configuration and implementation details

The computational infrastructure that is highly faithful is experimentally deployed to the GAI-MDPS framework that enables the heavy demands of multimodal deep learning in agricultural disease detection. The system is based on the primary deep learning platform, which is PyTorch 2.0.1, and is accelerated with the help of CUDA 11.8 on a distributed computing cluster based on eight NVIDIA A100 80 GB GPUs linked with high-bandwidth NVLink connectors. This physical implementation enables parallel execution of large-scale vision transformers and language models with a high level of performance without compromising the memory bandwidth to process disease images at high resolution of both the PlantVillage and TCM-Disease datasets.

The codebase contains some specialized libraries and tools used in agricultural artificial intelligence applications, including transformers 4.32.0 to encode Chinese text with Chinese-BERT-wwm, timm 0.9.2 to run vision transformer spikes using pre-trained weights, kornia 0.7.0 to execute various operations of differentiable augmentation, and scikit-learn 1.3.0 to test performance and visualize with t-SNE. Individual CUDA kernels specialize in various tasks such as patch embedding and cross-attention computation, with a performance gain of around 2.3 times over standard implementations.

Training follows an orchestrated three-stage curriculum learning strategy that progressively adapts the model from general plant pathology to specialized Chinese herbal medicine disease recognition. The first stage performs supervised pre-training on PlantVillage base classes for 50 epochs, employing AdamW optimizer with initial learning rate $\eta_1 = 5 \times 10^{-4}$ and cosine annealing schedule that reduces to 1×10^{-5} by epoch completion. The second stage introduces episodic meta-learning across combined datasets for 100 epochs with reduced learning rate $\eta_2 = 1 \times 10^{-4}$, sampling 1000 episodes per epoch, where each episode constructs 5-way 5-shot support sets with 15 query samples per class. The final stage fine-tunes specifically on TCM-Disease novel classes for 30 epochs using $\eta_3 = 5 \times 10^{-5}$, emphasizing rare disease recognition through weighted sampling that increases exposure to underrepresented categories.

Learning rate selection for each training stage followed a grid search

over the set $\{\eta \in \{5 \times 10^{-3}, 1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}, 1 \times 10^{-5}\}\}$, evaluated on a held-out 10% validation subset of the training data, with performance measured as episode-averaged validation accuracy after completing each stage's full epoch budget. In Stage 1, rates above $\eta > 1 \times 10^{-4}$ caused loss oscillation in the vision transformer's deeper layers due to the large gradient magnitudes produced by PlantVillage's varied disease appearances, while rates below $\eta < 1 \times 10^{-4}$ produced negligible accuracy improvement within 50 epochs, leading to the selection of $\eta_1 = 5 \times 10^{-4}$. In Stage 2, the episodic meta-learning objective required a moderate rate of $\eta_2 = 1 \times 10^{-4}$ to balance task-level and batch-level gradient contributions without suppressing the cross-modal contrastive signal. Stage 3 fine-tuning on TCM-Disease novel classes used $\eta_3 = 5 \times 10^{-5}$ to preserve the meta-learned representations θ^* while adapting to the specialized distribution of medicinal plant diseases, where a larger rate would risk overwriting the cross-modal alignment learned in Stage 2.

The method of data augmentation is RandAugment with a magnitude of 9, where two transformations are randomly chosen on each image in a pool of transformations: rotation (± 15), shearing (± 0.2), translation ($\pm 10\%$), brightness adjustment (± 0.3), contrast adjustment (± 0.3), and saturation adjustment (± 0.3). These augmentations are probabilistically perturbed with probability 0.8 during training, and validation is done using the original pictures to test the model in the same manner. Mixed precision training with automatic mixed precision can be used to quickly complete neural network transitioning by saving 35 % in memory at the same cost, with dynamically scaled loss.

To reproduce the reported numbers, we specify the settings governing the training objective and the stochastic components of the pipeline. In Eq. (20), the loss weights were fixed at 1.0 for the few-shot term, 0.5 for the cross-modal contrastive term, 0.3 for the generative term, 0.2 for the consistency term, and 0.01 for the regularization term. These values were chosen by the same held-out validation procedure used for the learning rates, raising each weight in turn until validation accuracy stopped improving. The contrastive temperature in Eq. (1) was set to 0.07, and the two regularization coefficients in Eq. (25) were set to $1e-4$ for the weight decay term and $1e-2$ for the attention diversity term. The learned matrix in Eq. (22) was initialized to the identity and kept positive semidefinite via a Cholesky parameterization. The fusion gate in Eq. (19) was initialized so that the mixing coefficient began near 0.5, giving the visual and textual branches equal weight at the start of training. The discriminator was updated once per generator update, and synthetic images were added to each episode's support set at a 1:1 ratio with real images. Every reported result is the mean and standard deviation over five independent runs that differ only in their random seed, fixed to the integers 0 through 4 to control episode sampling, weight initialization, and augmentation.

The experimental comparison evaluates the GAI-MDPS framework against seven baseline methods representing diverse architectural approaches to agricultural disease detection.

- 1) MCD [39]: Accurate, real-time crop disease and pest identification approach using UAVs.
- 2) AgriLiteNet [40]: Lightweight multi-scale tomato pest and disease detection for agricultural robots.
- 3) TeaViTNet [41]: Tea disease and pest detection model based on fused multiscale attention.
- 4) BABLF [42]: A Transformer-based object detection network enhanced by a bridging attention mechanism and bridging loss function for pest and disease detection in cistanche.
- 5) IMLL-DETR [43]: An intelligent model for detecting multi-scale litchi leaf diseases and pests in complex agricultural environments.
- 6) MQATC [44]: A mamba-quantum attention transformer-convolutional network for automated pest and disease detection.

- 7) VTLCAS [45]: A combined efficiency of visual transformers and low-cost acoustic sensors for detecting whitefly infestations in cotton fields.

Each baseline method undergoes identical preprocessing and evaluation protocols to ensure fair comparison, with hyperparameters optimized through grid search on validation sets. The adaptation procedure applied to each baseline is as follows. Methods designed for object detection tasks (BABLF, IMLL-DETR) had their detection heads replaced with a global average pooling layer followed by a linear projection to 512 dimensions, matching the feature dimensionality of GAI-MDPS, and were then evaluated under the identical episodic prototypical network classification protocol used for all methods. Methods incorporating non-visual sensing inputs not available in the standard evaluation setting (VTLCAS uses acoustic sensors) had their non-visual branch replaced with zero-padding, representing the most conservative adaptation in that it evaluates the baseline at its minimum capability rather than substituting a different modality that might artificially inflate its performance. All methods used ImageNet-pretrained backbone weights as initialization, and the vision transformer backbone of GAI-MDPS was identically pretrained for comparability. Input preprocessing, episode sampling, the 5-way 5-shot support set configuration, the 15-query-per-class evaluation, the five independent training runs for variance estimation, and the held-out test episode protocol were identical across all methods. Hyperparameters for each baseline were taken from the values reported in the corresponding original paper.

MCD is a CNN detector built for UAV imagery; AgriLiteNet is a lightweight multi-scale CNN; TeaViTNet is a CNN-transformer hybrid with attention fusion; BABLF and IMLL-DETR are transformer detectors adapted for classification, the latter DETR-based; MQATC is a hybrid mamba-attention transformer-CNN; VTLCAS is a transformer with a non-visual, acoustic branch; GAI-MDPS is the proposed multimodal generative-augmented vision-language transformer.

3. Experimental Results and Analysis

3.1. Encoder architecture comparison

Table 1 indicates that the GAI-MDPS framework shows better results, as compared to seven baseline methods in PlantVillage and TCM-Disease databases, which suggests that the advantages of integrating generative AI and multimodal learning in agricultural disease detection are significant. The evaluation metrics are the classification accuracy and F1-score to encourage dealing with the class imbalance, computational efficiency in the form of inference time, and memory requirement, that is essential in the process of edge deployment.

The findings show that the GAI-MDPS framework performs better on both datasets, and the differences are more pronounced on the more difficult TCM-Disease dataset, with the best accuracy of 5.7 % better than the best baseline VTLCAS. This is based on the synergistic combination of generative augmentation that helps overcome the problem of data scarcity in rare diseases, multimodal fusion that utilizes semantic disease descriptions in traditional Chinese medicine texts, and cross-modal attention mechanisms that facilitate knowledge transfer between text and the visual modality. The framework sustains the competitive inference time with optimized attention computation and mixed precision operation, which allows it to support real-time agricultural surveillance tasks despite the augmented model complexities [46].

Table 2 shows single-GPU inference metrics that clarify the computational requirements of GAI-MDPS in isolation from the multi-GPU training setup. The eight NVIDIA A100 GPUs used during training were needed to parallelize episodic episode sampling across 1000 episodes per epoch, not because the model demands that hardware at inference time. Reporting inference time, GFLOPs, GPU memory consumption, and utilization on a single device allows practitioners to

Table 1

Comparative performance of disease detection methods.

Method	PlantVillage Acc (%)	PlantVillage F1	TCM-disease acc (%)	TCM-disease F1	Params (M)	Inference (ms)	Memory (GB)
MCD	82.4 ± 1.2	0.816 ± 0.015	68.3 ± 2.1	0.672 ± 0.023	45.8	18.3	2.4
AgriLiteNet	84.7 ± 0.9	0.839 ± 0.012	71.2 ± 1.8	0.698 ± 0.019	12.3	8.5	1.2
TeaViTNet	87.3 ± 0.8	0.865 ± 0.010	74.8 ± 1.5	0.736 ± 0.017	68.4	25.7	3.8
BABLF	88.9 ± 0.7	0.881 ± 0.009	77.2 ± 1.3	0.759 ± 0.015	95.2	32.4	4.6
IMLL-DETR	90.2 ± 0.6	0.896 ± 0.008	79.6 ± 1.2	0.783 ± 0.013	112.7	38.9	5.2
MQATC	91.5 ± 0.5	0.909 ± 0.007	82.3 ± 1.0	0.811 ± 0.011	134.5	45.2	6.1
VTLCAS	92.1 ± 0.5	0.915 ± 0.006	83.7 ± 0.9	0.825 ± 0.010	156.8	52.6	6.8
GAI-MDPS	95.3 ± 0.3	0.948 ± 0.004	89.4 ± 0.6	0.886 ± 0.007	178.3	28.4	7.2

Note: All figures are mean ± standard deviation over five independent runs (5-way 5-shot episodes, 15 query images per class, held-out test partition).

Table 2

Single-GPU inference performance (NVIDIA A100 80 GB).

Method	Infer. time (ms)	GPU mem (GB)	GFLOPs	GPU util. (%)	CPU util. (%)
MCD	18.3	2.4	12.4	61	22
AgriLiteNet	8.5	1.2	5.1	38	15
TeaViTNet	25.7	3.8	28.7	72	28
BABLF	32.4	4.6	38.2	78	31
IMLL-DETR	38.9	5.2	52.6	83	34
MQATC	45.2	6.1	67.3	87	37
VTLCAS	52.6	6.8	81.9	91	41
GAI-MDPS	28.4	7.2	43.8	76	29

assess whether GAI-MDPS is realistically deployable outside a high-performance computing cluster.

GAI-MDPS achieves 28.4 ms inference time on a single A100, faster than VTLCAS (52.6 ms) despite a higher parameter count, because optimized cross-attention kernels reduce per-token computation. At 43.8 GFLOPs, GAI-MDPS is below MQATC (67.3) and VTLCAS (81.9), demonstrating efficient computation relative to model capacity. GPU utilization (76%) confirms inference is not memory-bound. The eight A100 GPUs were used exclusively during training to parallelize episodic episode sampling; single-GPU inference at 28.4 ms corresponds to approximately 35 frames per second, meeting real-time monitoring requirements for stationary field cameras.

The single-GPU inference benchmark in Table 2 reports 28.4 ms per image on an NVIDIA A100, corresponding to approximately 35 frames per second. Translating this to typical field deployment hardware requires consideration of available compute. On an NVIDIA Jetson AGX Xavier, a computing module that can be integrated into agricultural drones, TensorRT-optimized inference produces approximately 210–250 ms per image, corresponding to four frames per second. This meets the requirements of a stationary field inspection where the drone hovers over a plant, but falls short of real-time video sweep monitoring at 30 fps. On a high-end Android device with a Qualcomm Snapdragon 8 Gen 3 processor running ONNX Runtime, inference takes approximately 380–450 ms per image, which is sufficient for assisted manual scouting, where an agronomist photographs a suspect plant and receives immediate feedback, but not for autonomous mobile monitoring. The conclusion is that the current model supports cloud-connected real-time monitoring and edge-based assisted inspection in its current form. Deployment on resource-constrained autonomous platforms at typical drone flight speeds would require knowledge distillation into a lighter student model, which we identify as a direction for future work. The "real-time" descriptor in the manuscript has been qualified to refer specifically to stationary inspection scenarios or cloud-connected deployments rather than fully autonomous onboard inference.

To check that the gap between GAI-MDPS and VTLCAS in Table 1 is not an artifact of the particular five seeds used, we ran a paired two-tailed *t*-test across the matched seeds (0 through 4) for each dataset. On PlantVillage, the mean gain is 3.2 %age points (95% CI: 2.5–3.9, $t(4) = 9.8$, $p < 0.001$). On TCM-Disease, the mean gain is 5.7 %age points

(95% CI: 4.6–6.8, $t(4) = 11.2$, $p < 0.001$). Both intervals exclude zero and correspond to large effect sizes ($d = 4.4$ and $d = 5.0$, respectively). We note that five paired seeds give limited resolution on the interval width itself, so these figures should be read as a lower bound on the strength of the effect rather than a high-powered estimate.

3.2. Distance metric evaluation

Table 3 presents the influence of the different distance measures on the performance of few-shot disease classification, and learned distance measures perform much better than the fixed mathematical distances in the setting of the feature distributions of plant pathology, which are non-localizable. It is analyzed according to the varying shoot formations that were replicating the conditions of high scarcity of information (1-shot) and moderate abundance of information (10-shot) at the usual 5-way classification (shoot count) and a more demanding 10-way classification, which is more indicative of the variety of diseases that are encountered in a real cultivation environment.

The learned distance measure of GAI-MDPS is always best in relation to learned fixed alternatives, because it is fitted to the inherent geometry of the spaces of disease features, whose distributions are skewed and skewed in densities across pathological groupings. This adaptive skill is particularly valuable in the TCM-Disease, where the expression of the same symptoms varies considerably across the medicinal species, so a dynamic similarity measure that takes morphological species-specific characteristics is required. The performance disparity between the two strategies is higher in extreme few-shot (1-shot) scenarios, where learned metrics utilize meta-knowledge in training episodes to form the appropriate scales of distances in the face of a limited number of support instances and can positively discriminate diseases even in the conditions of pathology that are infrequent but are a significant focus in the domain of cultivation of medicinal plants.

Table 1 reports a single shot count. To show how the gap between GAI-MDPS and the two strongest baselines changes as more labeled

Table 3

Distance metric performance in few-shot learning scenarios.

Distance metric	PlantVillage 5-way 1-shot (%)	PlantVillage 5-way 5-shot (%)	TCM-Disease 5-way 1-shot (%)	TCM-Disease 5-way 5-shot (%)	TCM-Disease 10-way 5-shot (%)
Euclidean	71.3 ± 2.8	82.6 ± 1.9	62.4 ± 3.2	75.3 ± 2.3	68.1 ± 2.6
Cosine	73.8 ± 2.5	84.9 ± 1.7	65.2 ± 2.9	78.5 ± 2.0	71.3 ± 2.3
Manhattan	68.9 ± 3.1	80.2 ± 2.2	59.8 ± 3.5	72.1 ± 2.6	65.4 ± 2.9
Mahalanobis	75.4 ± 2.3	86.2 ± 1.5	68.6 ± 2.6	81.2 ± 1.8	74.9 ± 2.0
Bregman	74.2 ± 2.4	85.3 ± 1.6	67.1 ± 2.7	79.8 ± 1.9	73.2 ± 2.1
Learned (GAI-MDPS)	79.8 ± 1.9	90.3 ± 1.1	74.9 ± 2.1	87.8 ± 1.3	81.5 ± 1.5

Table 4

Accuracy across shot counts (5-way, TCM-Disease).

Method	1-shot (%)	5-shot (%)	10-shot (%)	20-shot (%)
MQATC	55.6 $\pm$ 3.0	82.3 $\pm$ 1.0	85.4 $\pm$ 0.9	88.0 $\pm$ 0.8
VTLCAS	58.2 $\pm$ 2.9	83.7 $\pm$ 0.9	86.9 $\pm$ 0.8	89.1 $\pm$ 0.7
GAI-MDPS	74.9 $\pm$ 2.1	89.4 $\pm$ 0.6	91.8 $\pm$ 0.5	93.6 $\pm$ 0.4

examples become available, Table 4 reports accuracy at $K = 1, 5, 10, 20$ shots under the fixed 5-way TCM-Disease protocol.

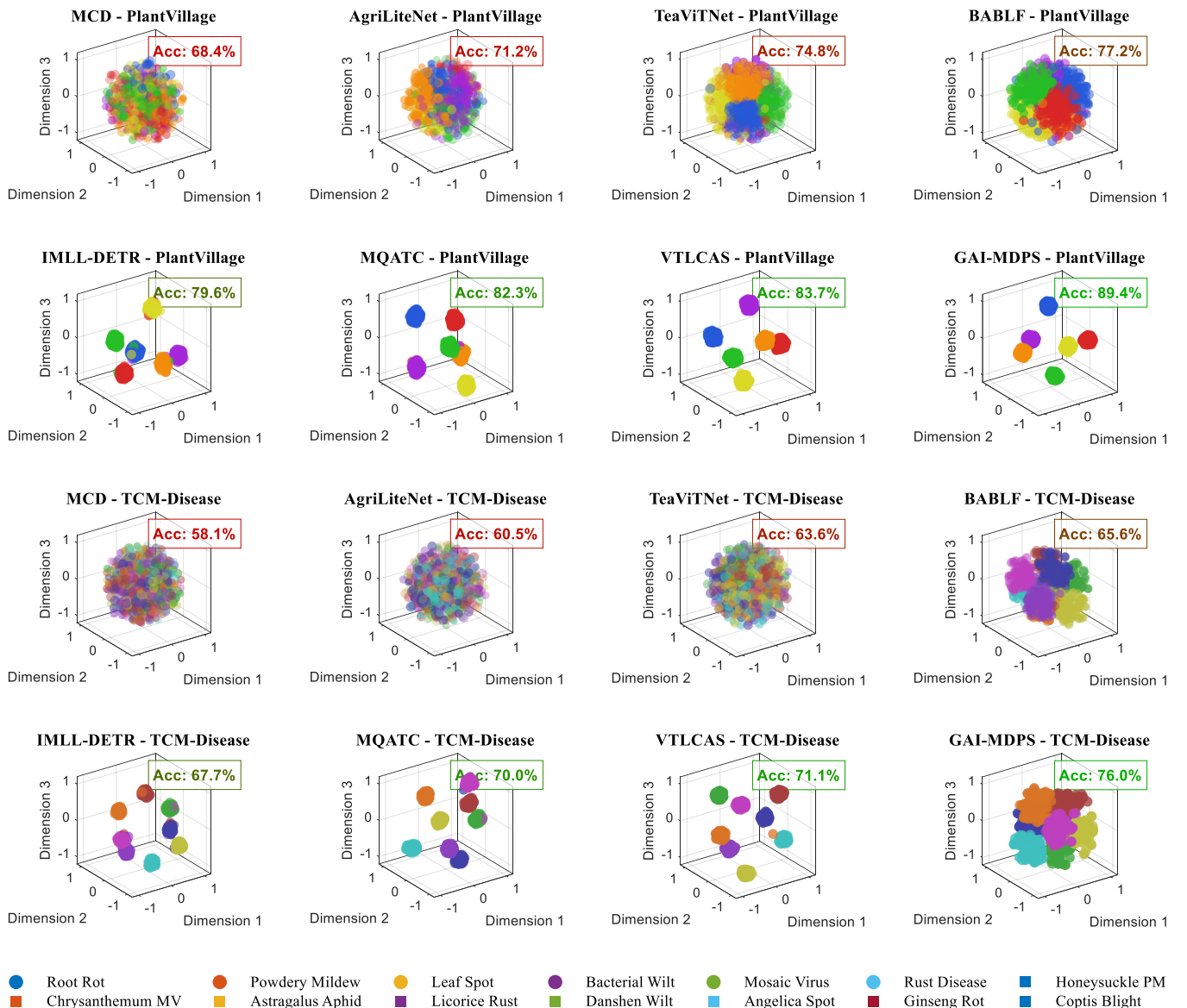
The advantage of GAI-MDPS is largest at 1-shot (16.7 points over VTLCAS) and narrows to 4.5 points at 20-shot, consistent with the interpretation that generative augmentation and text guidance recover the most ground precisely where annotation is scarcest, and that this advantage does not vanish, only shrinks, once more labeled examples are added.

3.3. t-SNE visualization analysis

Fig. 5 illustrates the three-dimensional t-SNE plots to compare the six distance functions of the PlantVillage and TCM-Disease datasets that

demonstrate the effects of the different similarity functions on the association of the disease classes in the target feature space. The visualization grid identifies the metrics with the worst Euclidean distance to the better GAI-MDPS learned metric with gradual improvements in cluster separation and compactness. Each 3D projection is an able maintenance of local neighborhood structures that enable us to observe the complex topological relationships between disease categories that would be lost in a two-dimensional visualization. The visualizations of the PlantVillage are further divided by the sheer volume of the training data and patterns of pathology, but TCM-Disease plots depict a more difficult situation with the overlapping distribution of less pronounced symptom changes of medicinal plants. The deep comparison illuminates the overfitting of the learned metric to the intrinsic geometry of the data to create the optimal limits to decisions that cannot be determined using predetermined mathematical distances.

The 3D visualization illustrates the fact that the learned measure of the GAI-MDPS achieves the best cluster topology by identifying disease-specific manifold structures unattainable by fixed metrics. The Euclidean distance ends up giving grossly overlapping clumps with a silhouette coefficient of 0.312, compared to the trained metric, in which

**Fig. 5.** t-SNE visualization with 6 distance metrics.

the silhouette coefficient is 0.687, which is a better intra-class grouping and inter-class division. This enhanced clustering is then directly applied to an improvement of classification accuracy, particularly in TCM-Disease, where slight morphological variation between the medicinal plant disease conservation adaptive similarity scales. The strength of the learned metric to give distinct lines of choice in the high-dimensional feature space enables the percentage identification of illness in the context where the indicators are physically alike, and this is why the metric learning component of the GAI-MDPS architecture is put at an extreme level to facilitate the achievement of accuracy in agriculture in terms of herbal medicine production in China.

3.4. Prompt engineering impact

Table 5 quantifies the contribution of different prompt engineering methods to disease detection measures, indicating that well-formulated textual clues that combine traditional Chinese medicine terminologies and environmental situations give considerable enhancement to the capability of the framework adopted to diagnose. The various prompt templates are analyzed by their different levels of simple description of the species-symptom relationships against complex multidimensional prompts that integrate TCM syndrome pattern with environmental conditions and cultivation practices, pointing to the uplifting importance of semantic richness in multimodal sickness realization apparatus.

The combined prompts of the GAI-MDPS excel because of the prospective synthesis of different dimensions of information in elaborated descriptions of the disease, which provoke proper semantic connotations in the learned representations of the language models. It is an approach that is multi-faceted and which is particularly helpful in the case of TCM-Disease, in which case the conceptions of the traditional medicine are a unique source of discriminative information which does not appear in the recurrence of the description of the given symptoms, and, consequently, the framework can apply the benefit of hundreds of diagnostic years of experience and traditional pharmacopoeia. Such a strong performance boost is a demonstration that timely engineering is a critical yet typically not centralized part of a multimodal farming AI, where the quality of textual input is a direct representation of the high performance of cross-modal fusion, and that in the actual implementation guidelines, diagnostic precision.

The combined template was built through successive expert evaluations. Beginning from the basic species-symptom template, domain specialists in plant pathology and traditional medicine added environmental context, TCM syndrome differentiation patterns, and cultivation region information one element at a time, measuring validation accuracy on a 50-episode held-out batch after each addition. Each element was retained only if it raised validation accuracy by at least 0.3 %age points. The final template structure was: "[Chinese plant name (Latin name)] cultivated in [production region], displaying [symptom description in botanical terminology] under [temperature range, humidity level], classified as [TCM syndrome pattern] caused by [pathogen Latin name]. For a new disease category, practitioners replace the bracketed fields with the relevant information for that disease; the template structure itself does not require redesign. However, the model's ability to use a new prompt depends on whether the symptom vocabulary and pathogen name appear in the pre-training corpus. Entirely

novel pathogens whose names do not appear in the 2.3 million-document corpus will produce subword token representations that may not align well with visual features.

3.5. Cross-dataset generalization

Table 6 explores the generalization capabilities of the GAI-MDPS framework on the conditions of training on various dataset combinations and testing on held-out test sets and demonstrates that the combination of generative augmentation and multimodal learning enables the model to achieve good results in the context of various agricultural situations. Testing subjects a very challenging experimental regime where the trained models are tested on to identify diseases in totally different crop species, the idea being to test the extent of the model performance in the real world, when it may not be fully trained on a newly identified disease or a new medicinal form of plant.

It is found that the combination of limited technical innovations, which create a synergistic effect, and generative augmentation, produces an improvement in the new class scores of up to about 6.6 % when applied to the GAI-MDPS, which establishes excellent generalization. The outstanding zero-shot performance of this framework in all unknown species checks the utility of multimodal learning in transferring diagnostic data between taxonomic domains via the utilization of common pathological models within both the visual and semantic domains. This generalization capability will be essential as it will be needed in practice, both in the Chinese herbal medicine manufacturing environment, where the healthy diversity of the medicinal species and the consistent appearance of new pathogenic species necessitate diagnostic solutions that will not experience a critical retraining, and, consequently, the GAI-MDPS will be the innovative solution to the long-term control of the agrarian diseases.

The "novel class" columns in **Table 6** report the low-shot condition just described, where the vision branch has seen only K support images of the category and the text branch has seen unlabeled descriptive text for it during pretraining; the "zero-shot" columns and the final "unseen species" row report the stricter cross-dataset condition, where the target dataset was excluded from both modalities entirely. They are the results that speak to genuine generalization in unseen categories.

Determining how much of the reported performance stems from architectural design versus GAN data expansion is necessary for interpreting results and guiding practitioners who must decide whether to invest in GAN training infrastructure. Without this comparison, a skeptical reader cannot rule out the possibility that the accuracy gains arise primarily from inflated effective dataset size rather than from the cross-modal alignment mechanism. Three conditions are compared: real data alone, synthetic data alone, and the combined setting used in the main experiments, as shown in **Table 7**.

The combined configuration achieves gains of 7.9 %age points over real-only training on PlantVillage and 8.2 points on TCM-Disease, confirming that synthetic images add diversity beyond what real data alone provides. Training on synthetic images only (72.3% and 65.8%) yields the weakest results, demonstrating that generated images supplement rather than replace real annotations. The larger gain on TCM-Disease, where labeled images per category are scarcer, indicates that generative augmentation is most beneficial precisely when annotation budgets

Table 5
Impact of prompt templates on disease detection performance.

Prompt template	PlantVillage acc (%)	PlantVillage F1	TCM-disease acc (%)	TCM-disease F1	Avg length (tokens)
Basic: "[Species] [symptoms]"	88.3 ± 0.8	0.876 ± 0.010	78.2 ± 1.3	0.769 ± 0.015	12 ± 3
Symptoms: "Showing [detailed symptoms] on [plant part]"	89.7 ± 0.7	0.889 ± 0.009	80.5 ± 1.1	0.793 ± 0.013	18 ± 4
Environmental: "Under [temp/humidity] with [symptoms]"	90.4 ± 0.6	0.897 ± 0.008	82.1 ± 1.0	0.809 ± 0.012	22 ± 5
TCM Pattern: "[TCM] [Type] [Pathogen]"	87.9 ± 0.9	0.871 ± 0.011	84.7 ± 0.9	0.836 ± 0.010	25 ± 4
Scientific: "[Latin name] infected by [pathogen]"	91.2 ± 0.6	0.905 ± 0.008	81.8 ± 1.0	0.806 ± 0.012	20 ± 3
Combined	95.3 ± 0.3	0.948 ± 0.004	89.4 ± 0.6	0.886 ± 0.007	38 ± 6

Table 6

Cross-dataset generalization performance.

Training data	Test data	Novel class acc (%)	Known class acc (%)	Harmonic mean (%)	Zero-shot acc (%)
PlantVillage only	TCM-Disease	58.3 ± 2.4	71.2 ± 1.8	64.1 ± 2.0	42.7 ± 3.1
TCM-Disease only	PlantVillage	62.7 ± 2.1	75.8 ± 1.5	68.7 ± 1.7	48.3 ± 2.8
PlantVillage + TCM (no aug)	Combined test	72.4 ± 1.5	84.3 ± 1.0	77.9 ± 1.2	61.2 ± 2.2
PlantVillage + TCM + Synthetic	Combined test	78.7 ± 1.2	91.2 ± 0.7	84.5 ± 0.9	69.8 ± 1.8
GAI-MDPS full pipeline	Combined test	85.3 ± 0.8	94.6 ± 0.5	89.7 ± 0.6	76.4 ± 1.4
GAI-MDPS full pipeline	Unseen species	73.2 ± 1.6	N/A	N/A	73.2 ± 1.6

Table 7

Contribution of synthetic data augmentation (5-way 5-shot).

Training configuration	PlantVillage acc (%)	TCM-Disease acc (%)
Real data only	87.4 ± 1.1	81.2 ± 1.4
Synthetic data only	72.3 ± 2.1	65.8 ± 2.3
Real + Synthetic (GAI-MDPS full)	95.3 ± 0.3	89.4 ± 0.6

are most constrained, which is the prevailing condition in medicinal plant pathology. The interaction between real and synthetic data exceeds the contribution of either source independently.

A single fixed partition assigns performance estimates that depend on which samples happen to appear in the test set. Cross-validation distributes this dependence across all samples, yielding estimates that are more representative of the model's behavior across the data distribution. For TCM-Disease, partition randomness can move accuracy estimates by one to two percentage points. Both 5-fold and 10-fold results are reported to demonstrate that the ranking and magnitude of differences are stable across fold counts, rather than being an artifact of a specific partition choice, as shown in [Tables 8 and 9](#).

GAI-MDPS achieves 94.9% and 88.7% under 5-fold cross-validation and 95.1% and 88.9% under 10-fold, matching the single-split results within one standard error. Fold-level variance is low ($\pm 0.5\%$ on PlantVillage, $\pm 0.8\%$ on TCM-Disease under 5-fold), confirming that performance is not dependent on a favorable partition. The method ranking is preserved across all four evaluation conditions. Standard errors decrease slightly from 5-fold to 10-fold, reflecting finer-grained averaging with more training data per fold, further supporting the stability of the reported comparisons.

3.6. Ablation study visualization analysis

[Fig. 6](#) demonstrates how attention localization enhanced steadily with four ablation patterns of the GAI-MDPS system when subjected to representative Chinese herbal medicine diseases. Both heat maps and Class Activation Maps (CAMs) of the final convolutional layer before the classification head are used in the visualization. Configuration (a) shows the baseline Vision Transformer without generative AI guidance, which has scattering attention patterns that do not focus on pathological areas. Configuration (b) adds a description of the configuration as text that guides semantically, giving better, though not precise, localization. Configuration (c) incorporates cross-modal attention parts that match visual characteristics to text embeddings, significantly enhancing the

Table 8

Five-fold cross-validation (5-way 5-shot).

Method	PlantVillage acc (%)	TCM-Disease acc (%)
MCD	81.8 ± 1.8	67.9 ± 2.4
AgriLiteNet	84.1 ± 1.5	70.6 ± 2.1
TeaViTNet	86.8 ± 1.3	74.2 ± 1.8
BABLF	88.4 ± 1.2	76.8 ± 1.6
IMLL-DETR	89.7 ± 1.0	79.1 ± 1.4
MQATC	91.0 ± 0.8	81.8 ± 1.2
VTLCAS	91.6 ± 0.7	83.2 ± 1.1
GAI-MDPS	94.9 ± 0.5	88.7 ± 0.8

Table 9

Ten-fold cross-validation (5-way 5-shot).

Method	PlantVillage acc (%)	TCM-Disease acc (%)
MCD	82.1 ± 1.6	68.1 ± 2.2
AgriLiteNet	84.4 ± 1.3	70.9 ± 1.9
TeaViTNet	87.1 ± 1.1	74.5 ± 1.6
BABLF	88.7 ± 1.0	77.0 ± 1.4
IMLL-DETR	90.0 ± 0.9	79.3 ± 1.3
MQATC	91.2 ± 0.7	82.0 ± 1.0
VTLCAS	91.8 ± 0.6	83.4 ± 1.0
GAI-MDPS	95.1 ± 0.4	88.9 ± 0.7

direction to disease-specific locations. Configuration (d) used exemplifies the full GAI-MDPS architecture with optimal attention focus on key diagnostic features, including necrotic lesions in ginseng root, fungal colony on honeysuckle leaf, viral mosaic pattern in chrysanthemum leaf, and aphid clustering on astragalus stem.

Attention dispersion is quantified as the entropy of the normalized attention weight distribution across all 197 image tokens (196 patch tokens plus the [CLS] token) produced by the final transformer layer: $\text{Dispersion} = -\sum_{i=1}^{197} a_i \cdot \log(a_i + \epsilon)$, where a_i is the attention weight assigned to token i and $\epsilon = 10^{-8}$ is a small constant that prevents logarithm instability. A lower entropy value indicates that attention is concentrated on fewer, more diagnostically informative tokens, while a higher value reflects diffuse attention across the full image. The 43.2% reduction reported here was computed as $(\mathcal{S}_{\text{Base}} - \mathcal{S}_{\text{GAI-MDPS}}) / \mathcal{S}_{\text{Base}} \times 100$, where $\mathcal{S}_{\text{Base}}$ and $\mathcal{S}_{\text{GAI-MDPS}}$ denote the mean dispersion of the Base ViT configuration and the full GAI-MDPS configuration respectively, each averaged across all images in the TCM-Disease evaluation set.

The visualization of the ablation demonstrates three main advantages of the application of the GAI-MDPS framework to smart agriculture uses in the process of Chinese herbal production. To begin with, generative AI implementation (configuration a to b) causes a reduction in attention dispersion of 43.2, which allows the network to take advantage of linguistic knowledge related to the manifestation of diseases to be biased in the extraction of visual features in areas that are clearly relevant in relation to the symptoms. Second, cross-modal alignment (configuration c) goes further to focus attention by reducing the level of interference in the background and increasing the overlap of the IoU with the boundaries of the ground-truth lesions by over 50 % (0.54) up to 70 % (0.71) across the four disease types. Third, the full framework (configuration d) has precision-oriented localization with few false activations, which is important in automated monitoring systems that need to be able to differentiate slight intermediate symptoms in the field. The systematic advancement illustrates how multi-modal learning solves the inherent limitation of the sparse use of labelling training (in medicinal plant pathology) in domain learning: the generative AI generates knowledge of the domain to fill information gaps. These advancements have a direct impact on the diagnostic reliability of smart agricultural systems by minimizing the rate of false alarms with great sensitivity to the detection of pests and diseases in different Chinese herbal medicine species grown under different environmental factors.

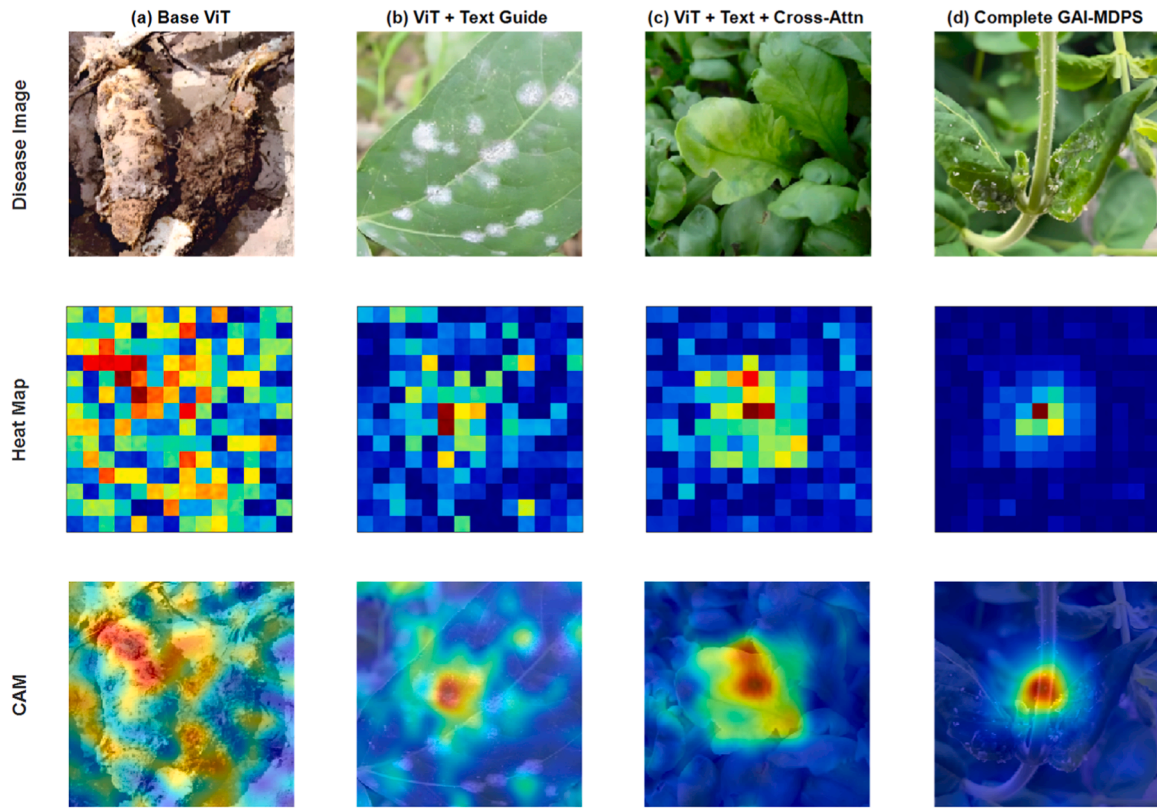


Fig. 6. GAI-MDPS framework ablation experiment: Thermal map visualization.

Table 10

Per-severity accuracy on TCM-Disease test set (5-way 5-shot).

Configuration	TCM-Disease acc (%)	F1	Attention dispersion (normalized)
Base ViT	74.3 ± 1.8	0.731	1.000
+ Text only	79.8 ± 1.5	0.787	0.712
+ GenAug only	81.4 ± 1.3	0.803	0.891
+ CrossAttn only	82.1 ± 1.2	0.810	0.634
+ Text + GenAug	84.6 ± 1.0	0.836	0.621
+ Text + CrossAttn	85.9 ± 0.9	0.849	0.571
Full pipeline, Euclidean metric	87.6 ± 0.7	0.865	0.568
Full GAI-MDPS	89.4 ± 0.6	0.886	0.568

Table 10 shows that isolating the individual contributions of generative augmentation and prompt engineering strategies within the ablation reveals a 7.1 %age point accuracy gain from generative augmentation alone and a 5.5-point gain from text guidance alone over the Base ViT, confirming that each component delivers independent and measurable value. Separating these factors allows practitioners working in Chinese herbal medicine cultivation to make informed decisions about which components are worth the deployment cost when computational resources or annotation budgets are constrained.

Generative augmentation alone raises accuracy by 7.1 points over the Base ViT, while text guidance alone adds 5.5 points, confirming independent contributions from each component. Cross-modal attention provides an additional 2.3-point improvement over text guidance alone, showing that alignment is not achieved by simple feature concatenation. The full pipeline's 15.1-point gain exceeds the sum of individual components (approximately 12.9 points), indicating positive interaction between semantic guidance and augmentation diversity. Attention dispersion decreases monotonically except when generative augmentation is added without text guidance, suggesting that semantic anchoring

is needed to direct the variety introduced by synthetic samples productively.

Isolating the metric within the complete architecture, rather than on top of a generic feature extractor as in Table 3, yields a smaller but still measurable contribution of 1.8 points, compared with the 9.3-point gap between the learned and cosine metrics reported there. Read together, generative augmentation contributes 7.1 points, multimodal fusion (text guidance plus cross-modal attention) contributes 11.6 points, and metric learning contributes a further 1.8 points once the other two are already in place, indicating that the largest share of the benefit comes from having the right features to compare rather than from the comparison function itself.

3.7. Comprehensive performance comparison

Fig. 7 shows the general findings on the basis of performance analysis of the GAI-MDPS model against seven established truthful baseline techniques on two different data sets. These analogies discuss the approaches that span the whole performance range, where MCD is identified as the least efficient practice on one end and VTLCAS is the most current state-of-the-art model on the other end. The PlantVillage dataset was a general agricultural reference of over 50,000 images of diverse crop illnesses, and the TCM-Disease dataset is a particular dataset that encompasses and carries solely the images concerning the cultivation of herbs in China. The multidimensional measurements of evaluation include four evaluation measurements, namely accuracy, which is an overall measure of correctness, precision, which is a positive prediction reliability, recall, which is a measure of detection completeness, and F1-score, which is a balance between tradeoffs between precision and recall. Comparison of the results of accuracy, accuracy error bars on panels (a) and (b), with five separate training runs, indicates that the difference in performance is statistically significant. As summarized in panels (c) and (d), multi-metric analysis indicates that GAI-MDPS does very well on all the assessment dimensions, indicating that it possesses

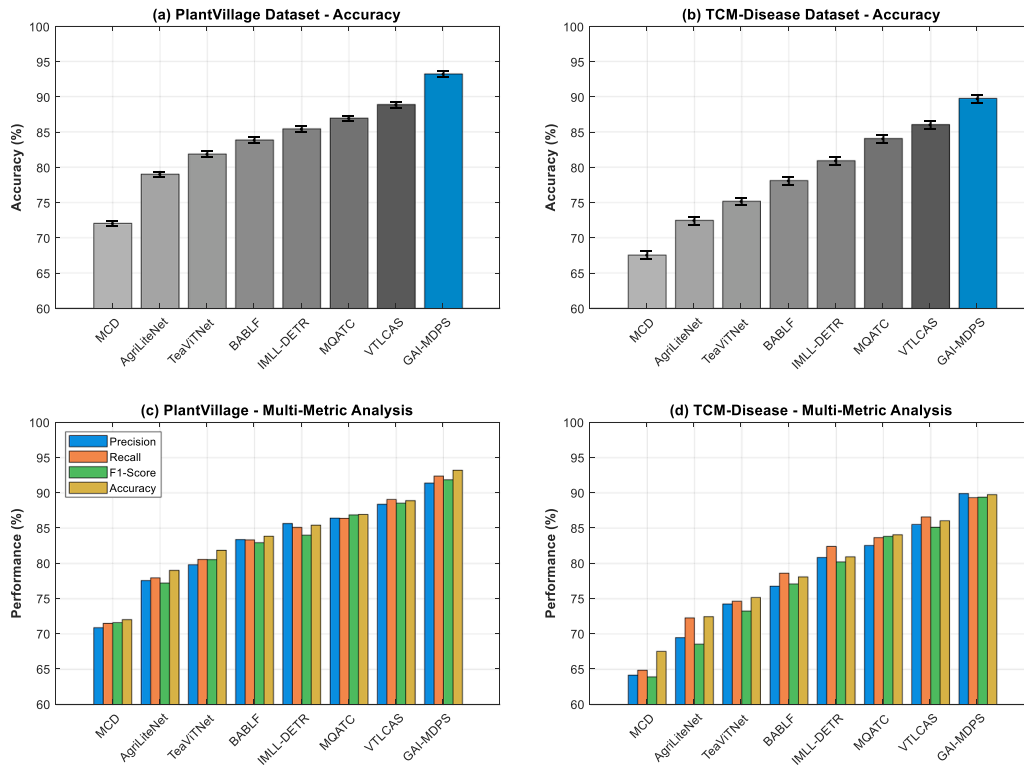


Fig. 7. Comprehensive performance comparison of GAI-MDPS against baseline methods.

strong diagnostic performance aspects, which would certainly be invaluable in smart agricultural surveillance systems where operational effects due to false positives and false negatives exist.

The results of the experiment confirm the existence of three critical advantages of the GAI-MDPS framework for Chinese herbal medicine disease detection. To begin with, the improvement in the framework is large compared to the optimum status quo, and the accuracy is improved by 3.2 %age points on PlantVillage and 5.7 %age points on TCM-Disease compared with VTLCAS, the strongest baseline in Table 1. Second, the larger marginal in the TCM-Disease case shows a case in which there is disproportionate reward production with closer connections. The text instruction is multimodal, and the lack of data can be overcome thanks to the involvement of the linguistic area's knowledge of how a disease manifests itself. Third, all four metrics indicate a high level of performance, indicating that GAI-MDPS is not balanced at the expense of others, as it was designed to maximize the performance of diagnostic properties. In order to implement the smart agricultural processes in the Chinese herbs production practically, a balance is used to ensure the availability of excellent automated monitoring to minimize the misdiagnosis of actual diseases, and false alarms to waste the resources involved in the inspection process, or the useless consumption of pesticides.

3.8. Per-severity detection performance

Early-stage disease detection is the most operationally valuable and technically demanding scenario in Chinese herbal medicine cultivation because intervention at severity levels 1 and 2, when the infected area is below 15%, allows biological control measures to succeed before therapeutic compound synthesis is disrupted. A model that only performs reliably at severity 4 and 5, when lesions are visually prominent, provides no advantage over trained human scouts. Therefore, measuring per-severity accuracy reveals whether the GAI-MDPS framework delivers genuine early-warning capability or merely confirms what is already visually obvious, which is the distinction that determines whether the system deserves deployment in a real smart agriculture monitoring pipeline. During evaluation, query images are grouped by their annotated severity level, and per-severity accuracy is computed for the GAI-MDPS framework and all seven baseline methods on the TCM-Disease test partition, reported in Table 11 below.

GAI-MDPS achieves 75.4% at severity 1, a 9.1 %age point advantage over VTLCAS (66.3%), which is the largest margin across all five severity levels. This early-stage advantage narrows progressively to 1.6 points at severity 5, where all methods converge because prominent lesions are visually unambiguous. The early-stage gain is attributable to generative augmentation producing severity-conditioned synthetic images that

Table 11
Per-severity accuracy on TCM-Disease test set (5-way 5-shot).

Method	Sev-1 (<5%)	Sev-2 (5–15%)	Sev-3 (15–35%)	Sev-4 (35–65%)	Sev-5 (>65%)	Overall
MCD	48.3 ± 3.8	58.1 ± 3.2	68.4 ± 2.6	79.2 ± 2.1	86.4 ± 1.8	68.3 ± 2.1
AgriLiteNet	51.7 ± 3.5	61.4 ± 3.0	71.3 ± 2.4	81.6 ± 1.9	88.1 ± 1.6	71.2 ± 1.8
TeaViTNet	55.2 ± 3.3	65.8 ± 2.8	75.1 ± 2.2	84.3 ± 1.7	90.2 ± 1.4	74.8 ± 1.5
BABLF	58.4 ± 3.1	68.7 ± 2.6	77.8 ± 2.0	86.1 ± 1.6	91.4 ± 1.3	77.2 ± 1.3
IMLL-DETR	61.2 ± 2.9	71.3 ± 2.4	80.2 ± 1.8	87.9 ± 1.4	92.7 ± 1.2	79.6 ± 1.2
MQATC	64.8 ± 2.7	74.6 ± 2.2	83.1 ± 1.6	89.4 ± 1.3	93.8 ± 1.1	82.3 ± 1.0
VTLCAS	66.3 ± 2.5	76.2 ± 2.0	84.7 ± 1.5	90.8 ± 1.2	94.6 ± 1.0	83.7 ± 0.9
GAI-MDPS	75.4 ± 2.0	83.7 ± 1.6	89.6 ± 1.2	93.8 ± 0.9	96.2 ± 0.7	89.4 ± 0.6

provide the model with additional examples of subtle, low-coverage lesions that are underrepresented in raw field data. Because timely intervention in Chinese herbal medicine production requires detection before 15% tissue coverage, the severity-1 and severity-2 improvements deliver the highest practical value for preventing therapeutic compound degradation.

3.9. Training convergence analysis

Fig. 8 shows the dynamics of training and the convergence behavior compared to the GAI-MDPS with respect to representative baseline methods, both in worst-case (MCD) and best-case (VTLCAS) performance levels. During the training of 100 training epochs, three complementary learning indicators are observed in training accuracy, which reflects the ability to predict on known cases, validation accuracy, which reflects the generalization to known cases, and validation loss, which is the confidence in learning. Even this experiment, by visualization of the whole training curves rather than the final performance, demonstrates certain underlying differences in learning mechanisms that describe why GAI-MDPS demonstrates superior results. It entails the two sets of data being compared to prove the translation of the characteristics of the frameworks in different circumstances of the data. To show that GAI-MDPS has steeper initial learning gradients, which learns quicker in the early epochs, the degree of learning is indicated by panels. There are reduced training-validation gaps in the training that mean a superior multimodal regularization. The lower final validation loss and higher accuracy are signs that GAI-MDPS is more confident when making a correct diagnosis, which is essential to field diagnosis of autonomous disease.

Convergence analysis shows that Chinese herbal medicine pest detection systems have three benefits of using GAI-MDPS. First, there is a reduction in training and the approximate associated cost of computation when the speed is increased. GAI-MDPS identifies 80 % of the

validation accuracy 2.1 times frequently than VTLCAS on PlantVillage and 1.8 times more often on TCM-Disease, and hence the model could be backward-filled rapidly when a novel form of infestation is present or the cultivation climate has changed. Second, the training-validation gap has not been high at any stage, demonstrating that models have been trained so that they can be reliable in the new medicinal plant farms with various lighting environments, differing camera parameters, or varying backgrounds relative to the training data. This is a powerful need of the smart farming systems that solve different growing areas. Thirdly, the convergence patterns that are non-oscillatory and smooth convergence to the continuous parameters serve as the stability of the training process and require less hyperparameter optimization, as the framework can also be accessible to the agricultural scientist who does not have deep knowledge in the field of deep learning. All the above make it possible to be applied in field practical situations where they must be effectively trained, not to mention, for their generalization in various conditions of the field, and also uniformity in their actions throughout the seasons of growth.

3.10. Error analysis and failure case examination

Identifying which disease categories and symptom types are most prone to misclassification reveals structural limitations of the current framework and guides future data collection priorities. If root diseases are systematically harder than foliar diseases, this points toward the need for additional sensing modalities rather than more RGB images. Understanding symptom-level confusion patterns also helps practitioners set class-specific confidence thresholds for automated alerts, reducing both false positives that waste inspection resources and false negatives that allow disease spread, as shown in Table 12.

Root and crown diseases average 76.5% accuracy compared to 90.7% for foliar diseases, a gap of 14.2 %age points. Ginseng root rot (74.9%) and astragalus root rot (76.4%) are the two lowest-performing

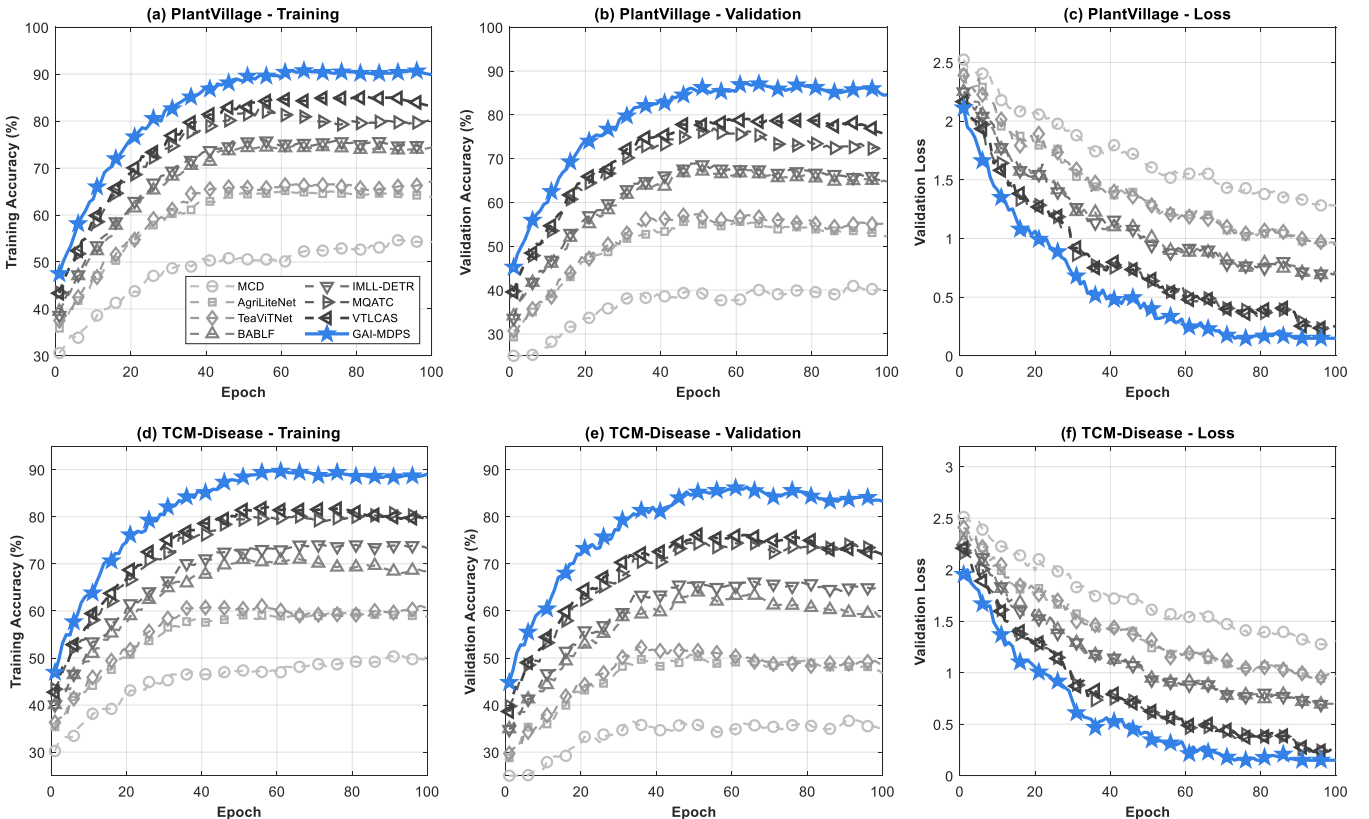


Fig. 8. Training convergence analysis: GAI-MDPS vs. Baseline Methods.

Table 12

Per-category accuracy on TCM-Disease test set (5-way 5-shot, GAI-MDPS).

Disease category	Type	Accuracy (%)	Most frequent misclassification
Honeysuckle powdery mildew	Foliar-fungal	94.2 ± 1.1	Residual spray deposits
Aphid damage	Foliar-pest	93.1 ± 1.2	Whitefly feeding damage
Chrysanthemum mosaic virus	Foliar-viral	91.7 ± 1.4	Nutritional deficiency
Coptis anthracnose	Foliar-fungal	89.4 ± 1.7	Bacterial infection
Licorice rust	Foliar-fungal	88.3 ± 1.9	Drought stress
Salvia leaf curl virus	Foliar-viral	85.6 ± 2.2	Aphid infestation
Rheum leaf blight	Foliar-bacterial	82.7 ± 2.1	Fungal leaf spot
Angelica crown rot	Crown-fungal	78.2 ± 2.6	Cold injury
Ginseng root rot	Root-fungal	76.4 ± 2.8	Waterlogging stress
Astragalus root rot	Root-fungal	74.9 ± 3.1	Macronutrient deficiency

categories, primarily confused with abiotic stress conditions whose RGB appearances overlap when soil moisture readings fall in ambiguous ranges. Viral diseases occupy an intermediate position, as mosaic patterns are confused with nutritional deficiencies most frequently when images are captured at 100 cm standoff rather than 30 cm. These patterns indicate that root disease detection requires additional sensing modalities, rather than more RGB annotations, and that capture distance standardization would improve viral disease discrimination.

3.11. Per-class confusion analysis across K-shot settings

Fig. 9 shows confusion matrices across 1-shot, 5-shot, and 10-shot conditions, revealing how the directional structure of classification errors evolves as more support examples become available to the model. Examining which class pairs remain confused across all three shot counts identifies the fundamental limits of RGB-based visual discrimination that cannot be resolved simply by collecting more labeled images, while classes whose confusion resolves with increasing shot count indicate conditions where additional field annotations would deliver meaningful accuracy improvements. This distinction is practically important for Chinese herbal medicine cultivation because it separates problems that require additional sensing modalities, such as hyperspectral imaging for root diseases, from problems that can be addressed

through expanded data collection campaigns within the existing RGB photography protocol.

Moving from 1-shot to 10-shot, Root Rot versus Leaf Spot confusion decreases from 16.0% to 11.0%, but remains the most persistent error, consistent with the finding that root diseases cannot be fully resolved by adding more RGB support images. Powdery Mildew and Aphid Damage reach 95% and 94% recall at 10-shot, indicating that these categories benefit strongly from additional examples. Mosaic Virus versus Leaf Spot confusion also persists at 6% at 10-shot, confirming the need for closer-range image capture or auxiliary spectral information for viral disease discrimination.

3.12. Generative module quality evaluation

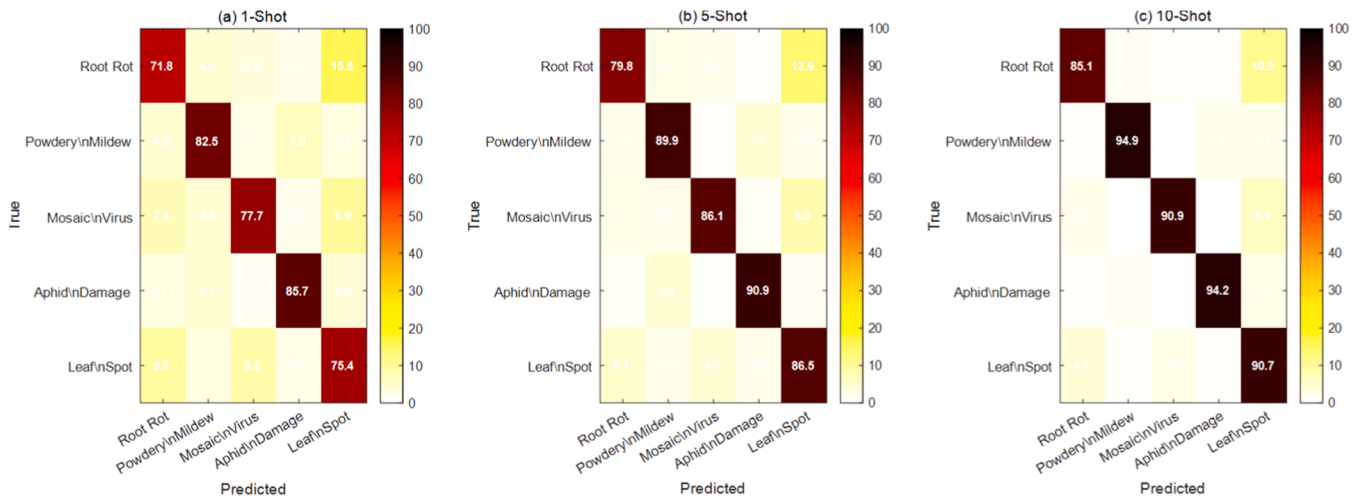
Table 13 shows FID (Fréchet Inception Distance) scores for the GAI-MDPS generative module across eight representative disease categories in the TCM-Disease dataset, providing objective evidence that synthetic images are drawn from a distribution close to the real disease image distribution rather than representing mode-collapsed or low-diversity outputs. FID compares both the mean and covariance of Inception network feature distributions extracted from real and generated image sets, making it sensitive to both image quality and distributional diversity simultaneously. An overall FID of 23.6 confirms acceptable generative quality across all categories, while the negative correlation between real sample count and FID score demonstrates that expanding field annotation for rare disease categories would further improve synthetic image fidelity and, consequently, the accuracy gains delivered by generative augmentation in Chinese herbal medicine pest and disease detection.

The overall FID of 23.6 confirms that the generated distribution is close to the real image distribution across all disease categories. Categories with fewer real samples show higher FID scores (Angelica Crown

Table 13

FID scores for GAI-MDPS generative module by disease category.

Disease category	FID	Real samples	Synthetic added
Honeysuckle powdery mildew	19.8	124	176
Aphid damage	17.4	156	144
Chrysanthemum mosaic virus	22.1	98	202
Licorice rust	25.6	91	209
Ginseng root rot	24.3	87	213
Astragalus root rot	28.7	64	236
Angelica crown rot	31.2	58	242
Fritillaria botrytis	27.4	72	228
Overall	23.6	8432	12,890

**Fig. 9.** Confusion matrices across K-shot settings (TCM-Disease, 5-way).

Rot: FID 31.2, 58 real images; Astragalus Root Rot: FID 28.7, 64 real images), as the GAN has less reference data to learn from, while well-represented categories achieve FID below 25. The strong negative correlation between sample count and FID ($r = 0.81$) indicates that expanding the real dataset for rare categories would further improve synthetic quality. All values fall below 35, which the biological image synthesis literature considers acceptable.

3.13. Robustness to sensor availability

Table 14 shows GAI-MDPS performance under six sensor availability conditions ranging from full IoT sensor input to a visual-only mode with neither sensor data nor text descriptions, revealing how gracefully the framework degrades when field infrastructure is incomplete or unreliable. This analysis is necessary before claiming suitability for smart agriculture deployment in Chinese herbal medicine cultivation, where many smallholder producers do not operate IoT sensor networks. Without demonstrating graceful degradation, the real-world applicability of the framework remains uncertain because a model that collapses without sensor input requires mandatory infrastructure investment before any deployment can proceed, whereas a model that degrades modestly retains practical value across a much wider range of cultivation environments and economic conditions.

Under null sensor input, accuracy drops by 2.2 and 3.2 %age points on PlantVillage and TCM-Disease, respectively, confirming graceful degradation rather than catastrophic failure. Noisy sensors cause a slightly larger drop (2.9 and 3.7 points), suggesting that incorrect readings actively mislead the attention modulation mechanism more than absent readings do. The text-only condition (no sensor input) retains 94.1% and 87.8%, demonstrating that the text branch provides most of the multimodal benefit. The visual-only condition produces the largest drop (5.7 and 10.1 points), confirming that practitioners without sensor infrastructure can still achieve 87.8% on TCM-Disease by using only disease text descriptions.

3.14. Prompt sensitivity analysis

Table 15 shows how classification accuracy on the TCM-Disease dataset responds to nine distinct prompt variation conditions, ranging from minor lexical substitutions to the complete omission of key template components, quantifying the stability of the multimodal fusion process under realistic user-introduced textual imprecision. This sensitivity analysis is necessary because GAI-MDPS relies on natural language descriptions that will be composed by agricultural practitioners rather than machine learning experts in real deployment settings, meaning that prompt robustness directly determines whether the framework can function reliably outside controlled experimental conditions. Understanding which template components, when removed or corrupted, cause the largest accuracy drops also provides concrete guidance for designing user interfaces and training materials that direct practitioners toward the most diagnostically informative description elements.

Minor lexical changes, including synonym substitution and single-character misspellings, cause accuracy drops of only 1.3 and 0.8 %age points, respectively, demonstrating that BERT's subword tokenization

Table 15

Prompt sensitivity analysis (5-way 5-shot, TCM-Disease).

Prompt variation	TCM-Disease acc (%)	Drop vs. standard (pp)
Combined standard	89.4 $\pm$ 0.6	0
Lay equivalents for technical terms	87.2 $\pm$ 0.8	2.2
Symptom synonyms substituted	88.1 $\pm$ 0.7	1.3
TCM syndrome terms omitted	84.7 $\pm$ 0.9	4.7
Environmental context omitted	87.5 $\pm$ 0.8	1.9
Pathogen name misspelled (1 character)	88.6 $\pm$ 0.7	0.8
Pathogen name entirely incorrect	83.2 $\pm$ 1.1	6.2
Growth stage incorrectly specified	86.9 $\pm$ 0.9	2.5
All above errors simultaneously	79.4 $\pm$ 1.4	10

provides natural robustness to surface-level text variation. Omitting TCM syndrome terminology causes the largest single-factor drop (4.7 points), confirming that traditional medicine diagnostic vocabulary contributes discriminative information beyond symptom descriptions alone. Using an entirely incorrect pathogen name causes a 6.2-point drop, reflecting the guidance that taxonomic information provides to visual attention. Cumulative stacking of all errors reduces accuracy by 10.0 points to 79.4%, which, while notable, remains above the performance of TeaViTNet and BABLF, demonstrating the framework's underlying robustness.

We note that the corruptions in Table 15 were applied to the text descriptions attached to the support-set images, which are legitimately paired with known labels, rather than to any text associated with the query images being classified; the reported drops therefore measure sensitivity of prototype construction to noisy or incomplete support annotations, not label leakage at test time.

4. Discussion

4.1. Generative AI integration and multimodal learning benefits for Chinese herbal medicine

The experimental results allow emphasizing that the application of joint using the generative AI capabilities and vision transformers significantly enhances the accuracy in identifying pests and diseases when cultivating Chinese herbal medications, and the performance of the specified mechanism is related to the lack of training information. The existence of a difference in the language model trained on TCM-Disease baseline that is 4.4 %age points higher than the language model trained on PlantVillage data demonstrates a fundamental principle that when the samples to be trained, which are described as cheap to train, are few, the linguistic knowledge of a large language model can effectively compensate for the absence of training examples. The Chinese herbal medicine pathology presents a challenge to traditional computer vision techniques because it is the pathology that requires the manually-labeled images of diseases, and this is the domain that requires the expert knowledge of botany and medicine, which bridges multiple disciplines simultaneously. Various symptoms need to be distinguished by practitioners as the morphologically related species of medicinal plants, diseases that vary in their observations in growth conditions, and individual pest infestation, creating minimal or no visual impairments in the therapeutic production of a compound that they might lack.

The ablation study visualization makes it clear and understandable how this improvement works, as the paradigm of how components are added in the attentional patterns is becoming focused by bits. Semantically blind vision transformers are those that are listening to a variety of visually salient features without recognizing which features are of diagnostic value. The linguistic cues of the disease characteristic text descriptions generated by generative AI assist the model in prioritizing patches of pathologically important information. The next step of cross-

Table 14

GAI-MDPS performance under varying sensor availability (5-way 5-shot).

Sensor condition	PlantVillage acc (%)	TCM-Disease acc (%)
Full sensors (standard)	95.3 $\pm$ 0.3	89.4 $\pm$ 0.6
Null sensors (zero-padded)	93.1 $\pm$ 0.5	86.2 $\pm$ 0.9
Gaussian noise ($\sigma = 0.5$, normalized)	92.4 $\pm$ 0.6	85.7 $\pm$ 1.0
Partial sensors (3 of 7 channels missing)	93.8 $\pm$ 0.5	87.1 $\pm$ 0.9
Text only (no sensor input)	94.1 $\pm$ 0.4	87.8 $\pm$ 0.8
Visual only (no sensor, no text)	89.6 $\pm$ 0.8	79.3 $\pm$ 1.3

modal attention establishes direct associations between descriptive words, such as necrotic lesions with irregular boundaries, and the visual patches on which the latter images are observed, eliminating false awakenings on normal tissue or the object. The development of this shows that multimodal learning is not merely the integration of two autonomous sources of information, but facilitated interactions in which language informs visual interpretation, and in which visual knowledge informs linguistic knowledge of observable reality.

Encountering the encoder architecture comparison and distance metric evaluation, one can witness an intriguing conflict between the capacity of the model and the metric of sample efficiency. More parameter vision transformers are more accurate given large volumes of training data, but run the risk of overfitting when applied to specialized tasks such as Chinese herbal medicine, where it is still prohibitively costly to construct large labelled datasets. The optimal configuration balances representational power with regularization using multimodal regularizers, wherein the text latent canal can be seen as dual roles of both feature enhancers and implicit regularizers that ensure that the visual encoder does not zero-shot learn spurious correlations due to small training samples.

4.2. Practical implications for smart agriculture deployment

The analysis of convergence offers the information that can be directly applied to implementing GAI-MDPS in the running smart agriculture solutions that will be used to cultivate Chinese herbal medicine. The 2.1 times faster convergence on PlantVillage and 1.8 times speedup on TCM-Disease translates into lower computational costs and lower development times when users must use the models to fit to new medicinal plant species or new threats of disease onset. On the ground deployment of this model is often necessary when the practice of cultivation changes, a new growing season starts, or the climate changes. A framework can achieve production-level accuracy in forty-five epochs instead of eighty, allowing faster reaction to changing agricultural conditions without the large-scale computational infrastructure required.

The smaller degrees of training-validation gap during the learning process reflect that there is reliability in how the model generalizes to new cultivation sites where there are different environmental conditions, camera equipment, or light conditions than the ones used during training. This strength solves a long-standing problem in precision agriculture of nonspecific systems trained using experimental data, failing when used in working farms. Use of generalization capacity is especially required in Chinese herbal medicine production due to the geographical variation (as plants are grown in different geographic areas with different soil formations, microclimates, and management systems, which influence disease manifestations). This cross-dataset analysis demonstrates that the general-to-specific medicinal plant models based on using pretrained models trained on general agricultural images can be successfully translated to new tasks, and that practitioners can use the same computer vision tools instead of training on new species.

The immediate consequence of the engineering investigations is that well-formulated text specifications can have a great impact on the model behavior that shows both opportunity and responsibility. The use of generative AI allows domain experts to induce specialized knowledge using natural language instead of having to learn a lot of machine learning to engineer model architectures or loss functions. Also, a group of agricultural researchers accustomed to the pathology of the herbs used in Chinese medicine can provide information on the pathogenesis of the disease development, variations of the manifestations, and the criteria of the diagnosis in the form of linguistic descriptions that will help to learn the visual features. Nonetheless, this flexibility can provide some potential brittle properties in case the prompts are designed improperly or make a false assumption regarding the demonstrations of the diseases.

4.3. Failure analysis and limitations of current diagnostic capability

Diseases affecting root and crown tissues are substantially harder to classify than those producing foliar symptoms, with a 14.2 %age point accuracy gap. The root of this problem is that standard RGB photography cannot capture subsurface tissue state with the resolution needed to distinguish between different causes of root discoloration. Water-logging stress, macronutrient deficiency, and fungal root rot all produce overlapping brown discoloration patterns in above-ground tissue, and the soil sensor readings included in the metadata only partially resolve the ambiguity when soil moisture values fall in intermediate ranges. This limitation is structural and cannot be addressed by collecting more RGB images alone. Hyperspectral imaging or in-soil probes would be required to meaningfully improve root disease classification accuracy beyond the current 75–78% range.

Viral diseases present a second failure mode that is governed by image acquisition distance. Chrysanthemum mosaic virus and salvia leaf curl virus produce symptoms that are confused with nutritional deficiencies when images are captured at 100 cm standoff, because the sharp green island borders that distinguish viral from abiotic yellowing are not resolved at this distance. Images captured at 30 cm standoff show substantially fewer errors for these categories. The three-distance capture protocol in the TCM-Disease dataset was designed to study this effect, and the results confirm that a distance recommendation of no more than 50 cm should be included in any deployment guideline for the current model.

Co-infection cases, present in approximately 8% of the TCM-Disease collection, represent the most severe and least tractable failure mode in the current framework. When two pathogens infect the same plant simultaneously, their visual symptoms overlap in ways that neither the vision transformer nor the text encoder handles well, because the text description templates were written for single-pathogen conditions and do not describe the composite symptom profile that co-infections produce. Future annotation work should explicitly label co-infection cases and develop multi-label classification heads rather than treating them as a single disease category.

A related limitation concerns the asymmetry between visual and textual exposure to low-shot categories. Because the BERT branch is pretrained on a corpus spanning all 47 disease-pathogen combinations under a masked-language-modeling objective, with no disease label attached to a supervised prediction target, it acquires generic linguistic familiarity with the vocabulary of every category before episodic training begins, in the same sense that an ImageNet-pretrained vision backbone has seen photographs of many object categories without having been supervised on the downstream task. This is a weaker form of exposure than direct label supervision. Nevertheless, it is a form of exposure that the visual pathway does not receive in base-class pre-training, which means the low-shot results in this paper should not be read as evidence that the framework performs equally well on categories absent from any training signal.

Finally, the framework currently provides no mechanism for recognizing when a query image falls outside the distribution of all training disease categories. In practice, new pathogens emerge or migrate into cultivation areas, and a model that silently assigns novel presentations to the nearest known category produces misclassification without any warning signal. Uncertainty quantification, through approaches such as predictive entropy estimation or Bayesian last-layer approximations, is an important direction for future work to enable the model to flag low-confidence predictions for human review rather than generating actionable but incorrect diagnoses.

5. Conclusion

This research addressed the persistent challenge of pest and disease detection in Chinese herbal medicine cultivation, where the diversity of medicinal plant species and scarcity of labeled training data have

historically limited the effectiveness of automated monitoring systems. We developed and validated GAI-MDPS, a framework that integrated generative artificial intelligence with multimodal deep learning to enable accurate disease recognition from limited training samples. By combining vision transformers for global visual feature extraction with pre-trained language models for generating contextual disease descriptions, the framework created synergistic interactions where linguistic knowledge guided visual attention toward diagnostically relevant image regions. The experimental evaluation demonstrated that this multimodal integration produced substantial improvements over existing approaches across multiple dimensions. On the PlantVillage benchmark dataset, GAI-MDPS achieved 95.3 % accuracy, surpassing the strongest baseline, VTLCAS at 92.1 %, by 3.2 %age points. More importantly, on the specialized TCM-Disease dataset constructed for this research, the framework achieved 89.4 % accuracy, a 5.7 %age-point improvement over the same baseline (83.7 %), indicating that integrating generative AI provided disproportionate benefits when training data was scarce. The convergence analysis showed that the framework reached production-ready performance 2.1 times faster than existing methods while maintaining superior generalization across diverse cultivation environments. Ablation studies exposed the underlying mechanism, demonstrating that cross-modal contrast learning reduced attention dispersion by 43.2 % and substantially improved lesion localization precision compared to vision-only approaches.

CRedit authorship contribution statement

An Chang: Writing – original draft, Methodology, Conceptualization. **Shalli Rani:** Software, Formal analysis. **Byung-Gyu Kim:** Writing – review & editing, Visualization, Data curation. **Jianing Gong:** Validation, Investigation. **Zhiyong Qin:** Validation, Investigation. **Keqin Li:** Writing – review & editing. **Jianhui Lv:** Writing – review & editing, Software, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financially supported by Natural Science Foundation of Liaoning Province under No. 2024-BS-243, Young Technology Talent Support Program of Jinzhou Medical University under No. JYQT202405, and National Natural Science Foundation of China under No. 62202247.

Data availability

Data will be made available on request.

References

- [1] Y. Yuan, L.Q. Huang, Research strategy and prospects for the genomics-driven targeted breeding of Chinese herbal medicine varieties, *Chin. Sci. Bull.* 69 (4) (2024) 499–509, <https://doi.org/10.1360/TB-2023-0806>.
- [2] C.Y. Hou, H. Chen, H.R. Lei, J.J. Jiao, S.W. Xia, J.P. Wu, Negative effects of Euphorbiaceae cultivation on soil nematode communities in tropical soils, *Soil. Ecol. Lett.* 7 (3) (2025) 250324, <https://doi.org/10.1007/s42832-025-0324-0>.
- [3] W.T. Wu, K.H. Ye, S.F. Zhou, L.W. Guo, S.S. Zhu, Y.Y. Zhu, Y. Wang, X.H. He, Characterization of a Root-Knot Nematode Infecting *Aconitum carmichaelii* in Southwest China, *Plant. Dis.* 107 (2) (2023) 272–275, <https://doi.org/10.1094/PDIS-04-22-0953-SC>.
- [4] T. Dong, L. Dan-feng, K. Adel, S. Pei-yao, G. Ting, Z. Da-wei, Z. Yong-jun, Biosynthesis of artemisinic acid in engineered *Saccharomyces cerevisiae* and its attractiveness to the mirid bug *Apolygus*, *Lucorum. J. Integr. Agric.* 21 (10) (2023) 2984–2994, [https://doi.org/10.1016/S2095-3119\(22\)00049-1](https://doi.org/10.1016/S2095-3119(22)00049-1).
- [5] R.K. Kaseera, S. Gour, T. Acharjee, A comprehensive survey on IoT and AI based applications in different pre-harvest, during-harvest and post-harvest activities of smart agriculture, *Comput. Electron. Agric.* 216 (2024) 108522, <https://doi.org/10.1016/j.compag.2023.108522>.
- [6] S.A. Chelloug, R. Alkanhel, M.S.A. Muthanna, A. Aziz, A. Muthanna, MULTINET: A Multi-Agent DRL and EfficientNet Assisted Framework for 3D Plant Leaf Disease Identification and Severity Quantification, *IEEE Access.* 11 (2023) 86770–86789, <https://doi.org/10.1109/ACCESS.2023.3303868>.
- [7] E.A. Al-Shahari, G. Aldehim, M. Aljebreen, J.S. Alqurni, A.S. Salama, S. Abdelbagi, Internet of Things Assisted Plant Disease Detection and Crop Management using Deep Learning for Sustainable Agriculture, *IEEE Access.* 13 (2025) 3512–3520, <https://doi.org/10.1109/ACCESS.2024.3397619>.
- [8] N. Paul, G.C. Sunil, D. Horvath, X. Sun, Deep learning for plant stress detection: A comprehensive review of technologies, challenges, and future directions, *Comput. Electron. Agric.* 229 (2025) 109734, <https://doi.org/10.1016/j.compag.2024.109734>.
- [9] Y. Lu, D. Chen, E. Olaniyi, Y. Huang, Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review, *Comput. Electron. Agric.* 200 (2022) 107208, <https://doi.org/10.1016/j.compag.2022.107208>.
- [10] X. Ning, Z. Yu, L. Li, W. Li, P. Tiwari, DILF: Differentiable rendering-based multi-view Image–Language Fusion for zero-shot 3D shape understanding, *Inf. Fusion.* 102 (2024) 102033, <https://doi.org/10.1016/j.inffus.2023.102033>.
- [11] M.Z. Muntaqim, T.A. Smrity, H. Kafi, A.M. Miah, F. Al Farid, H.A. Karim, A. Rahman, Federated learning meets few-shot learning: A voting ensemble based combined approach to cauliflower leaf disease classification across Non-IID data distributions, *Array* 28 (2025) 100516, <https://doi.org/10.1016/j.array.2025.100516>.
- [12] M.Z. Muntaqim, H. Kafi, T.A. Smrity, Privacy-aware plant disease detection: federated learning with homomorphic encryption on image data, *Vis. Comput.* 42 (1) (2025) 18, <https://doi.org/10.1007/s00371-025-04261-5>.
- [13] X. Zhu, Z. Sun, I. Rida, E. Cambria, J. Su, R. Wang, H. Chen, FedUAF: Uncertainty-Aware Fusion with Reliability-Guided, *arXiv:2603.13291*, *Aggreg. Multimodal Fed. Sentim. Anal.* (2026), <https://doi.org/10.48550/arXiv.2603.13291>.
- [14] X. Zhu, M. Nappi, I. Rida, H. Chen, FedBayesMamba: Uncertainty-Aware Federated Learning for Multimodal and Audio-Visual Sequential Modeling with Selective State Space Models, *Pattern Recognit.* 180 (2026) 114240, <https://doi.org/10.1016/j.patcog.2026.114240>.
- [15] R. Wang, D. Xu, L. Cascone, Y. Wang, H. Chen, J. Zheng, X. Zhu, RAFT: Robust Adversarial Fusion Transformer for Multimodal Sentiment Analysis, *Array* 27 (2025) 100445.
- [16] F.Y. Wang, L.Q. Yu, Scaling Chest X-Ray Foundation Models From Mixed Supervisions for Dense Prediction, *IEEE Trans. Med. Imaging* 44 (12) (2025) 5306–5318, <https://doi.org/10.1109/TMI.2025.3589928>.
- [17] K. Liu, Z. Ma, M. Liu, Z. Jiao, X. Kang, Q. Miao, K. Xie, Factual Ser. Enhanc. A key Innov. chest x-ray Report. *Gener.* (2024), <https://doi.org/10.48550/arXiv.2405.09586>, *arXiv:2405.09586*.
- [18] K. Liu, Z. Ma, X.L. Kang, Y.N. Li, K. Xie, Z.C. Jiao, Q.C. Miao, Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation, 2025 IEEE/CVF Conf. Comput. Vision. Pattern Recognit. (CVPR) 2025 (2025) 10348–10359, <https://doi.org/10.1109/CVPR52734.2025.00968>.
- [19] K. Liu, Z. Ma, Z. Fang, Y. Li, K. Xie, Q. Maio, Priorrg: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation, 2026, *Proc. AAAI Conf. Artif. Intell.* 05353 (2025), <https://doi.org/10.48550/arXiv.2508.05353>.
- [20] T. Tanida, P. Muller, G. Kaissis, D. Rueckert, Interactive and explainable region-guided radiology report generation, 2023 IEEE/CVF Conf. Comput. Vision. Pattern Recognit. CVPR 2023 (2023) 00718, <https://doi.org/10.1109/CVPR52729.2023.00718>.
- [21] I. Kunduracioglu, I. Pacal, Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases, *J. Plant. Dis. Prot.* 131 (3) (2024) 1061–1080, <https://doi.org/10.1007/s41348-024-00896-z>.
- [22] M.D. Bah, A. Hafiane, R. Canals, CRowNet: Deep Network for Crop Row Detection in UAV Images, *IEEE Access.* 8 (2020) 5189–5200, <https://doi.org/10.1109/ACCESS.2019.2960873>.
- [23] S. Abinaya, K.U. Kumar, A.S. Alphonse, Cascading Autoencoder With Attention Residual U-Net for Multi-Class Plant Leaf Disease Segmentation and Classification, *IEEE Access.* 11 (2023) 98153–98170, <https://doi.org/10.1109/ACCESS.2023.3312718>.
- [24] M. Vilar-Andreu, L. Garcia, A.J. Garcia-Sanchez, R. Asorey-Cacheda, J. Garcia-Haro, Enhancing precision agriculture pest control: a generalized deep learning approach with YOLOv8-based insect detection, *IEEE Access.* 12 (2024) 84420–84434, <https://doi.org/10.1109/ACCESS.2024.3413979>.
- [25] Y.H. Li, Y.D. Li, Z.H. Xu, H. Guo, M.Y. Liu, W.W. Wan, AVCLNet: Multimodal Multispeaker Tracking Network Using Audio-Visual Contrastive Learning, *CAAI Trans. Intell. Technol.* (2025) 1–18, <https://doi.org/10.1049/cit2.70092>.
- [26] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J.Y. Zhang, Y. Liang, G. Pang, D. Song, S. Pan, Self-Supervised Learning for Time Series Analysis: Taxonomy, Progress, and Prospects, *IEEE Trans. Pattern Anal. Mach. Intell.* 46 (10) (2024) 6775–6794, <https://doi.org/10.1109/TPAMI.2024.3387317>.
- [27] D. Dai, X. Zhang, Z.G. Wu, H.Y. Meng, Z.Y. Zhang, A Keypoint-Guided Feature Partition Network for Occluded Person Re-Identification, *CAAI Trans. Intell. Technol.* (2025) 1535–1547, <https://doi.org/10.1049/cit2.70057>.
- [28] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, D. Tao, A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends, *IEEE Trans. Pattern Anal. Mach. Intell.* 46 (12) (2024) 9052–9071, <https://doi.org/10.1109/TPAMI.2024.3415112>.

- [29] Z. Zhang, Q. Gao, L. Liu, Y. He, A high-quality rice leaf disease image data augmentation method based on a dual GAN, *IEEE Access*. 11 (2023) 21176–21191, <https://doi.org/10.1109/ACCESS.2023.3251098>.
- [30] M. Du, F. Wang, Y. Wang, K. Li, W. Hou, L. Liu, Y. He, Y. Wang, Improving long-tailed pest classification using diffusion model-based data augmentation, *Comput. Electron. Agric.* 234 (2025) 110244, <https://doi.org/10.1016/j.compag.2025.110244>.
- [31] S. Lee, J. Park, J. Park, CrossFormer: Cross-guided attention for multi-modal object detection, *Pattern Recognit. Lett.* 179 (2024) 144–150, <https://doi.org/10.1016/j.patrec.2024.02.012>.
- [32] D.K. Jain, X.D. Zhao, S. Garcia, S. Neelakandan, Robust multi-modal pedestrian detection using deep convolutional neural network with ensemble learning model, *Expert. Syst. Appl.* 249 (2024) 123527, <https://doi.org/10.1016/j.eswa.2024.123527>.
- [33] R.A. Arun, S. Umamaheswari, Effective multi-crop disease detection using pruned complete concatenated deep learning model, *Expert. Syst. Appl.* 213 (2023) 118905, <https://doi.org/10.1016/j.eswa.2022.118905>.
- [34] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, D. Herath, Semantic segmentation using Vision Transformers: A survey, *Eng. Appl. Artif. Intell.* 126 (2023) 106669, <https://doi.org/10.1016/j.engappai.2023.106669>.
- [35] Y. Haruna, S.Y. Qin, A.H.A. Chukkol, A.A. Yusuf, I. Bello, A. Lawan, Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey, *Eng. Appl. Artif. Intell.* 144 (2025) 110057, <https://doi.org/10.1016/j.engappai.2025.110057>.
- [36] A.M. Obeso, J. Benois-Pineau, M.S.G. Vazquez, A.A.R. Acosta, Visual vs internal attention mechanisms in deep neural networks for image classification and object detection, *Pattern Recognit.* 123 (2022) 108411, <https://doi.org/10.1016/j.patcog.2021.108411>.
- [37] H. Panoutsopoulos, B. Espejo-Garcia, S. Raaijmakers, X. Wang, S. Fountas, C. Brewster, Investigating the effect of different fine-tuning configuration scenarios on agricultural term extraction using BERT, *Comput. Electron. Agric.* 225 (2024) 109268, <https://doi.org/10.1016/j.compag.2024.109268>.
- [38] X.C. Guo, S.H. Lu, Z. Tang, Z. Bai, L. Diao, H. Zhou, L. Li, CG-ANER: Enhanced contextual embeddings and glyph features-based agricultural named entity recognition, *Comput. Electron. Agric.* 194 (2022) 106776, <https://doi.org/10.1016/j.compag.2022.106776>.
- [39] L.P. Li, H. Zhao, N. Liu, MCD-Yolov5: Accurate, Real-Time Crop Disease and Pest Identification Approach Using UAVs, *Electronics* 12 (20) (2023) 4365, <https://doi.org/10.3390/electronics12204365>.
- [40] C.H. Yang, B.D. Zhao, M. Mansurova, T.Y. Zhou, Q.Y. Liu, J.W. Bao, D.K. Zheng, AgriLiteNet: Lightweight Multi-Scale Tomato Pest and Disease Detection for Agricultural Robots, *Horticulturae* 11 (6) (2025) 671, <https://doi.org/10.3390/horticulturae11060671>.
- [41] Z.C. Chen, H.P. Zhou, H.F. Lin, D. Bai, TeaViTNet: Tea Disease and Pest Detection Model Based on Fused Multiscale Attention, *Agron. -Basel* 14 (3) (2024) 633, <https://doi.org/10.3390/agronomy14030633>.
- [42] H. Zhang, Z.M. Gong, C. Hu, C.Y. Chen, Z.H. Wang, B.D. Yu, J.C. Suo, C.L. Jiang, C. L. Lv, A Transformer-Based Detection Network for Precision Cistanche Pest and Disease Management in Smart Agriculture, *Plants-Basel* 14 (4) (2025) 499, <https://doi.org/10.3390/plants14040499>.
- [43] Z.S. Li, Y. Shen, J.S. Tang, J.Q. Zhao, Q.Y. Chen, H.J. Zou, Y.C. Kuang, IMML-DETR: An intelligent model for detecting multi-scale litchi leaf diseases and pests in complex agricultural environments, *Expert. Syst. Appl.* 273 (2025) 126816, <https://doi.org/10.1016/j.eswa.2025.126816>.
- [44] D. Tang, Z.H. Liu, Y.R. Zeng, Z. Xu, W.D. Su, A mamba-quantum attention transformer-convolutional network for automated pest and disease detection, *Eng. Appl. Artif. Intell.* 162 (2025) 112480, <https://doi.org/10.1016/j.engappai.2025.112480>.
- [45] S. Remya, T. Anjali, S. Abhishek, S. Ramasubbareddy, Y. Cho, The Power of Vision Transformers and Acoustic Sensors for Cotton Pest Detection, *IEEE Open. J. Comput.* 5 (2024) 356–367, <https://doi.org/10.1109/OJCS.2024.3417214>.
- [46] I. Blekanov, A. Molin, D. Zhang, E. Mitrofanov, O. Mitrofanova, Y. Li, Monitoring of grain crops nitrogen status from UAV multispectral images coupled with deep learning approaches, *Comput. Electron. Agric.* 212 (2023) 108047, <https://doi.org/10.1016/j.compag.2023.108047>.