MCDI-CVL: Multi-crop Disease Identification Model Based on Cross-modal Visual Language Feature Fusion

Jianlou Lou Aodi Ye Jianxun Lou*
School of Computer Science, Northeast Electric Power University
{loujianlou,2202301028,jianxunlou}@neepu.edu.cn
*Corresponding author

Abstract

Accurate crop disease recognition is essential for agricultural productivity and food security. Existing methods mainly focus on visual features extracted from crop leaf images, often insufficiently utilizing the complementary information embedded in textual symptom descriptions. To address this limitation, we propose a framework for multimodal data fusion and cross-modal alignment named MCDI-CVL, comprising a visual encoder and a text-enhanced dynamic encoder. The visual and textual representations are coupled via a bidirectional cross-modal attention mechanism and an adaptive fusion gating module, enabling effective alignment and integration of complementary features for fine-grained multi-crop disease recognition. We also introduce CropsDisease-5M16, a new multimodal dataset spanning five major crops and 16 representative disease types. Comparative experiments further show that MCDI-CVL achieves state-of-the-art performance on multimodal crop disease dataset, outperforming both unimodal and existing multimodal competitors. This study provides a practical and generalizable solution for real-world agricultural disease diagnosis.

***Keywords:** Crop disease identification, Multimodal learning, Cross-modal fusion, Agricultural AI.*

1. Introduction

Crop diseases pose a persistent threat to agricultural productivity and food security worldwide [30]. However, conventional identification rely heavily on expert knowledge and visual inspection, which are labor-intensive and often unreliable in field conditions [2]. These limitations have spurred increasing interest in automated crop disease recognition [14], a key task in agricultural visual intelligence that enables scalable and timely monitoring for field diagnosis and automated crop management.

Deep learning has substantially advanced crop disease recognition, with convolutional neural networks (CNNs) widely adopted for their effectiveness in capturing local lesion features [25, 8, 29]. However, standard CNNs often struggle with limited receptive fields and insufficient multi-scale representation, particularly under complex field conditions. To address these issues, several studies have explored structural enhancements for cross-scale feature fusion, such as dilated convolutions [34], receptive field modules [17], and cross-layer fusion strategies. Despite these improvements, CNNs remain inherently constrained in modeling long-range dependencies. Transformer-based architectures have been introduced to mitigate the locality bias of CNNs by modeling long-range dependencies via self-attention [7]. Building on this, Karthik et al. [18] proposed a dual-track fusion model that leverages Swin Transformer [21] as a global feature extractor alongside a local dual-attention pathway, achieving improved robustness to symptom scale variation. Similarly, Huang et al. [15] developed EConv-ViT, a two-branch model that synergistically combines context-aware and structure-aware features for apple disease identification. While these models demonstrate superior performance, they remain fundamentally limited by their reliance on unimodal visual inputs, particularly under noisy or ambiguous real-world conditions.

Despite their strong performance in many tasks, visual-only models face inherent limitations, particularly evident in real agricultural settings, where disease symptoms are often subtle and difficult to detect in the early stages. As the disease progresses, symptom expression can vary significantly depending on crop type, environmental conditions, and severity, and may still exhibit ambiguous visual similarity across different diseases [19]. In such cases, textual descriptions—derived from expert annotations, diagnostic manuals, or agricultural guidelines—offer complementary semantic cues that can enhance model robustness through cross-modal learning. Inspired by the success of vision–language pretraining models such as CLIP [24] and BLIP-2 [20], which demonstrate the potential of multimodal alignment, recent studies have begun exploring similar frameworks for agricultural disease recognition. Wang

et al. [33] developed a unified cross-modal system combining a wavelet-enhanced Mamba network [12], a vision–language alignment transformer, and a stochastic optimization module for joint feature refinement. While recent studies have demonstrated the potential of multimodal learning for crop disease recognition, most efforts remain limited in both data scale and modeling capacity. Existing methods often focus on individual crops such as cucumbers or tomatoes and rely on small-scale datasets with limited class diversity. In particular, these limitations in data coverage and diversity pose serious challenges for leaf-based disease recognition, where symptoms such as discoloration, deformation, and lesion distribution are often subtle and vary across species and disease stages. Moreover, many approaches overlook the semantic correspondence between disease descriptions and their visual manifestations on crop leaves, limiting the model's ability to interpret fine-grained lesion patterns across diverse conditions.

To address the above limitations, we construct a new multimodal dataset of leaf disease images paired with fine-grained symptom descriptions generated by a pretrained vision–language model, enabling aligned image–text supervision for real-world agricultural scenarios. Building on this resource, we propose MCDI-CVL, a multimodal fusion and cross-alignment framework for leaf-based disease recognition that jointly models visual and textual features via semantic-aware attention and gated fusion to support robust classification.

The main contributions of this paper are as follows:

- We construct CropDisease-5M16, a new multimodal dataset of 4,800 crop leaf images from five diverse crops and 16 disease types, each paired with a fine-grained symptom description.
- We propose a text-enhanced dynamic encoder that improves the representation of disease descriptions via keyword-aware masking and hierarchical semantic fusion. Experiments show that symptom descriptions offer complementary cues that benefit disease identification across diverse crops with varying disease types.
- We design bidirectional cross-modal attention and gated fusion mechanisms to align textual symptom descriptions with visual lesion regions, enabling the enhanced fusion of visual and textual modalities for accurate disease recognition.

2. Related Work

Crop Disease Identification. Conventional crop disease identification relies heavily on manual visual inspection by experts, which is subjective, labor-intensive, and lacks scalability. To move beyond manual assessment, early attempts at automated crop disease recognition adopted traditional machine learning techniques, which relied on handcrafted features and exhibited limited adaptability to complex symptoms. With the advent of deep learning, significant advances have been achieved in image-based disease classification. For instance, Tripathi et al. [28] introduced SoyaTrans, which improves small lesion detection and global context modeling through random shifting. Hoang et al. [32] proposed LGENetB4CA, which incorporates the LeafGabor filter to enhance texture features for chilli leaf disease classification. Arun et al. [1] designed a hierarchical lightweight CNN with pointwise convolution and cross-layer feature concatenation to enable efficient recognition under constraints. Despite these successes, existing approaches are limited to unimodal visual inputs and lack semantic-level supervision, leading to weak feature interpretability and vulnerability to background noise, symptom ambiguity, and inter-class similarity.

Multimodal Crop Disease Dataset. Recent efforts have explored the construction of multimodal datasets for crop disease diagnosis by integrating heterogeneous sources (e.g., crop images and symptom descriptions), and notably combining expert knowledge with generative techniques to produce semantically relevant image–text pairs. For example, Wang et al.[33] utilized generative models to produce descriptions aligned with visual symptoms. Zhou et al.[37] relied on multi-level expert annotations grounded in agricultural literature. Cao et al. [3] integrated web knowledge bases and applied modality-consistent augmentation through image enhancement and text variation. However, most existing datasets focus on single crops such as tomato or cucumber, and are limited in both semantic diversity and cross-crop generalization. There remains a lack of image–text paired datasets covering multiple crops to support the development of generalizable multimodal recognition models.

Multimodal Data Fusion for Disease Recognition. To address the limitations of unimodal recognition, recent efforts have explored cross-modal image–text modeling for crop disease identification. Zhou et al. [37] proposed SLIP, which combines image reconstruction with contrastive text learning using unlabeled image–text pairs. Feng et al. [10] developed ITC-Net to fuse image and text features for fine-grained classification of vegetable diseases. Cao et al. [3] introduced ITLMLP, which leverages contrastive and self-supervised learning to align multimodal features for cucumber disease recognition. Dai et al. [5] proposed ITF-WPI, which integrates CoTN and ODLS encoders to achieve semantic fusion for Lycium barbarum pest detection. Notably, these studies still face challenges in modeling rich pathological semantics and maintaining a balance between cross-modal alignment and robustness to noisy or ambiguous inputs. Moreover, most methods focus on single-crop

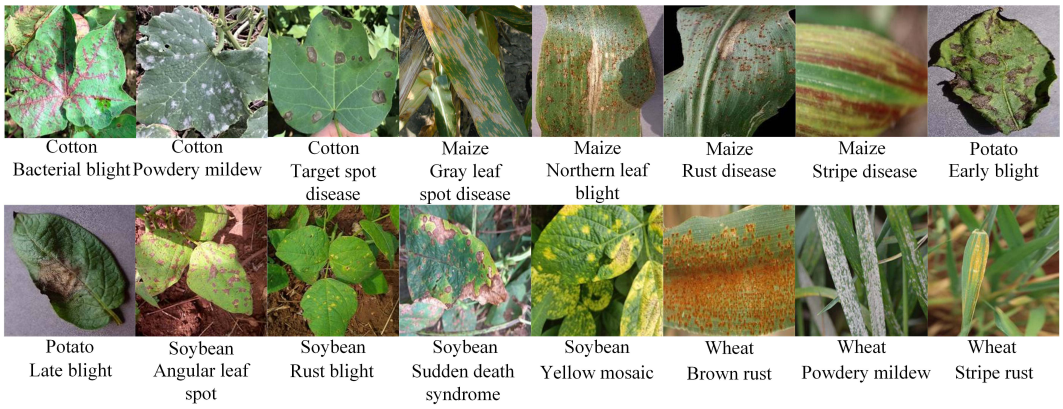

Figure 1: Example image–text pairs from the proposed CropsDisease-5M16 dataset, which includes five of the most widely cultivated crops (i.e., maize, wheat, soybean, cotton, and potato) and 16 representative plant diseases.

settings and exhibit limited generalization across diverse species and symptom types.

3. Method

Single-modal disease identification in agricultural settings faces several challenges, including interference from complex farmland backgrounds, high variability in symptom morphology, unclear features in early disease stages, and significant visual similarity across different disease types. Compared with other modalities, textual semantic information can offer more explicit and structured descriptions of disease symptoms, which helps guide visual models to focus on critical regions and improve classification accuracy [38]. Based on this insight, we propose a multimodal framework, MCDI-CVL, which integrates disease images with corresponding textual symptom descriptions to address the above issues effectively.

3.1. CropsDisease-5M16

Most existing crop disease datasets are limited to single-crop settings and lack paired image–text annotations, which constrains the exploration of multimodal learning in agricultural disease diagnosis. Inspired by recent progress in clinical and industrial domains [35], where generative models are increasingly used to augment expert annotation, we adopt a similar strategy for the agricultural setting.

The proposed CropsDisease-5M16 comprises five of the most widely grown food and cash crops: maize, soybean, cotton, wheat and potatoes [16], each crop contains multiple disease types. We collected annotated crop disease image datasets from an open-source platform (i.e., Kaggle), covering five major crops and 16 disease categories that are widely recognized as highly susceptible in agricultural studies [26]. The selected images include both laboratory settings and real farm environments, and the corresponding disease categories are detailed in Figure 1. For each disease, 300 representative images were curated under expert guidance to ensure intra-class diversity and visual relevance, while removing duplicates, mislabeled cases, or severely corrupted samples, resulting in a total of 4,800 images.

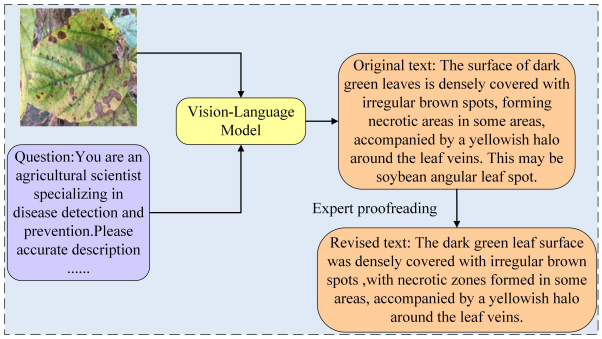

Figure 2: Example of symptom-level text generation and expert correction in the construction of the proposed CropsDisease-5M16 dataset.

To construct semantically aligned image–text pairs, we generated a textual description of the visible symptoms for each image using a vision–language model, following the recommendation from [35] and [11]. An example of the prompt design and generated description is shown in Figure 2. Symptom descriptions were generated using the Janus-Pro vision–language model, following the recommendation of [4], then refined and verified based on expert suggestions to ensure linguistic fluency and consistency with the visual content. To avoid label leakage, explicit disease names or category-indicative terms were removed during the revision process, and no additional pathological descriptions with strong disease specificity were introduced from the outside. This resulted in semantically aligned image–text pairs that serve as a reliable foundation for vi-

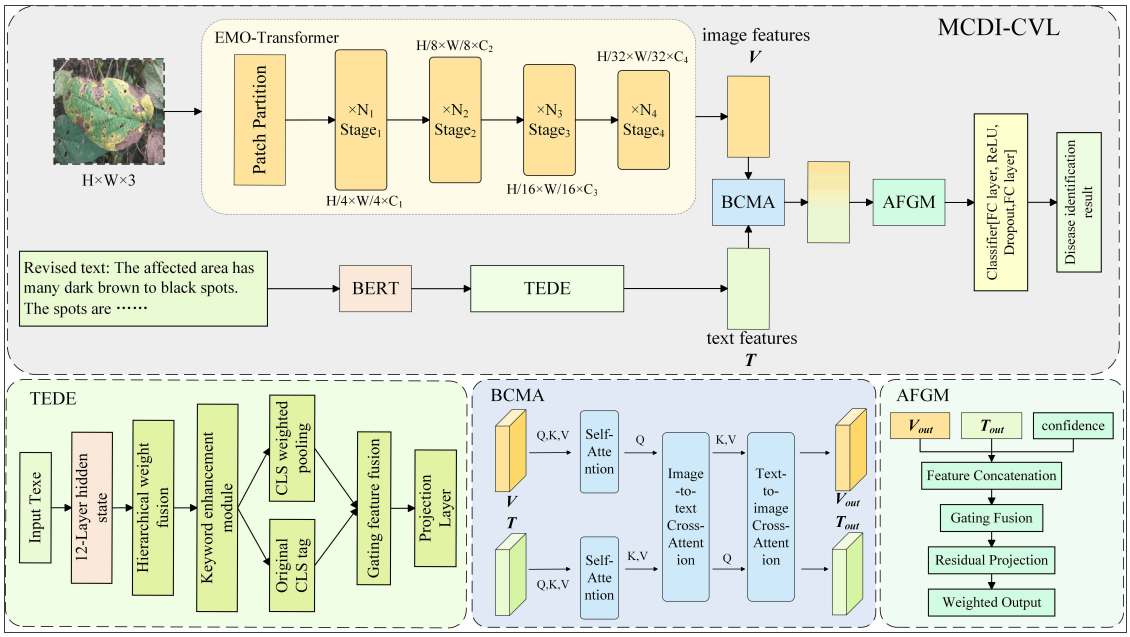

Figure 3: Overview and key components of the MCDI-CVL framework. The upper part illustrates the overall architecture, where a visual encoder extracts features from disease images and a text-enhanced dynamic encoder (TEDE) encodes symptom descriptions. Visual and textual features are aligned via bidirectional cross-modal attention (BCMA) and fused using an adaptive fusion gating module (AFGM) to produce disease predictions. The lower part illustrates the design of the three proposed modules: TEDE, BCMA, and AFGM.

sion–language modeling and analysis in crop disease recognition.

3.2. The Proposed MCDI-CVL Framework

As illustrated in Figure 3, the proposed MCDI-CVL framework integrates complementary visual and textual cues for robust crop disease recognition. We adopt the EMO-Transformer [36] as a backbone to extract multi-scale visual features, upon which three core modules are designed to enable fine-grained multimodal crop disease identification:

1. A text-enhanced dynamic encoder (TEDE) that provides enhanced representations of symptom descriptions via hierarchical fusion and keyword-aware masking;

2. A bidirectional cross-modal attention (BCMA) mechanism that aligns visual and textual representations through multi-head attention and residual refinement;

3. An adaptive fusion gating module (AFGM) that dynamically balances modality contributions using confidence scoring and residual projection.

The final fused features are passed to a classification head to produce disease probability distributions.

3.3. Text Enhanced Dynamic Encoder

To facilitate cross-modal alignment, the text encoder aims to generate hierarchical and symptom-aware representations that correspond to lesion-level visual features. In order to capture rich symptom semantics from disease descriptions, we adopt BERT [6] as the text encoder and design a multi-stage enhancement mechanism. Instead of using only the final BERT layer, we perform dynamic aggregation across all 12 transformer layers. Each layer output $H^{(i)} \in \mathbb{R}^{B\times L\times d}$ is assigned a learnable scalar weight α_i, which is normalized via softmax to produce the fused token representation:

$$w_i = \frac{e^{\alpha_i}}{\sum_{j=1}^{12} e^{\alpha_j}}, \quad H_{\text{fused}} = \sum_{i=1}^{12} w_i H^{(i)} \tag{1}$$

To emphasize symptom-related terms, we apply a keyword-guided attention mechanism. A soft attention map A_{kw} is computed from H_{fused} and modulated using a binary keyword mask $M_t \in \{0,1\}^L$, which identifies clinically important tokens:

$$A_{\text{kw}} = \text{Softmax}(H_{\text{fused}} W_{\text{attn}}) \odot M_t \tag{2}$$

$$H_{\text{enh}} = H_{\text{fused}} + A_{\text{kw}} \odot H_{\text{fused}} \tag{3}$$

where W_{attn} is a attention weighting matrix. We compute a token-level attention over H_{enh} to obtain a dynamic local

representation h_{dyn}:

$$\beta_t = \mathrm{Softmax}(H_{\mathrm{enh}} W_p), \quad h_{\mathrm{dyn}} = \sum_{t=1}^{L} \beta_t H_{\mathrm{enh}}^{(t)} \tag{4}$$

To balance this with global sentence-level information, we fuse h_{dyn} with the global category token embedding vector h_{cls} via a gated mechanism:

$$g_{\mathrm{sa}} = \sigma(W_p[h_{\mathrm{dyn}}; h_{\mathrm{cls}}] + b_g) \tag{5}$$

$$T = g_{\mathrm{sa}} \odot h_{\mathrm{dyn}} + (1 - g_{\mathrm{sa}}) \odot h_{\mathrm{cls}} \tag{6}$$

where W_{p} is dynamic weight projection matrix, b_{g} is the gated network bias term. The final output $T \in \mathbb{R}^{B\times d}$ integrates both fine-grained symptom expressions and global semantic context and serves as the textual representation in the subsequent cross-modal alignment.

3.4. Bidirectional Cross-Modal Attention

To align lesion features from images with symptom descriptions from text, we propose a bidirectional attention module that performs semantic-level interaction between modalities. Given visual features $V \in \mathbb{R}^{B\times d}$ and textual features $T \in \mathbb{R}^{B\times d}$, we first project them into a shared embedding space using linear transformations:

$$\tilde{V} = VW_v, \quad \tilde{T} = TW_t \tag{7}$$

where $W_v, W_t \in \mathbb{R}^{d\times d'}$ are learnable weights. We then apply multi-head scaled dot-product attention in both directions to facilitate fine-grained alignment. In the visual-to-text path, the query is $\tilde{V}$, while keys and values come from $\tilde{T}$:

$$V' = \mathrm{Softmax}\left(\frac{(\tilde{V}Q_v)(\tilde{T}K_t)^T}{\sqrt{d'}}\right) \cdot (\tilde{T}V_t) \tag{8}$$

where $Q_v, K_t, V_t \in \mathbb{R}^{d'\times d'}$ are projection matrices for query, key, and value. A symmetric process is performed in the text-to-visual direction:

$$T' = \mathrm{Softmax}\left(\frac{(\tilde{T}Q_t)(\tilde{V}K_v)^T}{\sqrt{d'}}\right) \cdot (\tilde{V}V_v) \tag{9}$$

In order to retain original modality information and stabilize training, residual connections and layer normalization are applied to both paths:

$$V_{\mathrm{out}} = \mathrm{LayerNorm}(\tilde{V} + V') \tag{10}$$

$$T_{\mathrm{out}} = \mathrm{LayerNorm}(\tilde{T} + T') \tag{11}$$

The output features V_{out} and T_{out} capture modality-specific patterns enhanced through cross-modal context, supporting downstream multimodal fusion.

3.5. Adaptive Fusion Gating Module

To adaptively balance contributions from visual and textual modalities, we design an adaptive fusion module that integrates confidence estimation, channel dynamic weighting, and residual projection to suppress cross-modal noise and enhance semantic aggregation. Given the core visual feature $V_{\mathrm{out}} \in \mathbb{R}^{B\times d}$ and an auxiliary instance feature T_{out} (e.g., from cross-modal interaction), we first estimate a modality confidence score $C_{\mathrm{score}} \in \mathbb{R}^{B\times 1}$ via a two-layer transformation:

$$C_{\mathrm{score}} = \sigma\left(W_{c2} \cdot \mathrm{ReLU}\left(W_{c1}V_{\mathrm{out}} + b_{c1}\right) + b_{c2}\right) \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid function. In order to achieve cross-modal dynamic weight assignment, we concatenate V_{out}, T_{out}, and C_{score} along the channel dimension to form a joint representation f_{joint}. This is passed through a gating network to produce a gating weight vector $g_{\mathrm{acw}} \in \mathbb{R}^{B\times d}$:

$$h_{\mathrm{acw}} = \mathrm{Dropout}\left(\mathrm{ReLU}(W_{g1}f_{\mathrm{joint}} + b_{g1})\right) \tag{13}$$

$$g_{\mathrm{acw}} = \sigma\left(W_{g2} \cdot h + b_{g2}\right) \tag{14}$$

Finally, we perform gated fusion by modulating V_{out} with g_{acw}, and projecting T_{out} to the core feature space via a residual mapping W_r. The fused feature f_{fused} is given by:

$$f_{\mathrm{fused}} = g_{\mathrm{acw}} \odot V_{\mathrm{out}} + C_{\mathrm{score}} \odot (W_r T_{\mathrm{out}}) \tag{15}$$

where $\odot$ denotes element-wise multiplication. This module enables the model to selectively integrate modality-specific features based on confidence and semantic reliability, enhancing robustness under noisy or ambiguous conditions.

4. Experiments

4.1. Experimental Setup

All experiments were conducted in a controlled computing environment equipped with the Intel Xeon Gold 6240 CPU and NVIDIA Tesla T4 GPU. We split the CropsDisease-5M16 dataset into a training-validation set (80%) and a test set (20%). The training-validation set is further used for five-fold cross-validation to obtain the optimal hyperparameters. The final model is retrained on the entire 80% set and evaluated on the held-out 20% test set. We utilized the Adam optimizer with an initial learning rate of 0.001, batch size of 32, and early stopping patience value 15 epochs. Cross-entropy loss function was used for training the model.

4.2. Comparative Experiments

CropsDisease-5M16 is a multimodal dataset proposed in this study, designed to support cross-species disease classification and multimodal learning via paired visual–textual

Table 1: Classification results of MCDI-CVL on 16 crop disease categories from CropsDisease-5M16. The reported p-value indicates the result of the Wilcoxon Signed-Rank Test on F1 scores between the proposed model (i.e., MCDI-CVL) and the compared methods.

Methods	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**	***p*-value**
MobileNetV3 [13]	93.65	93.82	93.65	93.57	<0.05
EfficientNetV2 [31]	93.75	93.75	93.75	93.73	<0.05
MobileViT [22]	94.48	94.74	94.48	94.33	<0.05
Swin Transformer [21]	94.79	94.82	94.79	94.75	<0.05
Conformer [23]	95.00	95.02	95.00	94.99	<0.05
Multiscale ViT [9]	95.31	95.31	95.31	95.31	<0.05
EMO Transformer [36]	96.88	96.85	96.94	96.87	<0.05
MCDI-CVL	**98.96**	**98.99**	**98.99**	**98.98**	–

symptom representations. Its diversity in crop types and symptom descriptions poses realistic challenges for agricultural disease recognition. To assess the advantages of our multimodal framework, we compare MCDI-CVL with a series of state-of-the-art unimodal classification models (i.e., models that use visual features only). As reported in Table 1, the EMO Transformer [36] achieves the strongest performance among the unimodal models, with an accuracy of 96.88%. MCDI-CVL further outperforms EMO Transformer by 2.08% in accuracy, 2.14% in precision, 2.05% in recall, and 2.11% in F1 score. Moreover, the p-values from the Wilcoxon Signed-Rank Test on F1 scores confirm that the proposed MCDI-CVL model significantly outperforms other methods ($p < 0.05$). Additionally, We evaluate the MCDI-CVL model on this dataset to assess its effectiveness under complex field conditions. As shown in Table 2, MCDI-CVL achieves robust performance across all categories, with perfect classification in 9 out of 16 disease types, and an overall accuracy of 98.96%. These results demonstrate the model's strong discriminative ability in handling diverse disease manifestations across multiple crop species.

Table 2: Classification results of MCDI-CVL on 16 crop disease categories from CropsDisease-5M16(%).

Disease Names	Accuracy	Precision	Recall	F1
Cotton bacterial blight	100.00	100.00	100.00	100.00
Cotton powdery mildew	97.83	100.00	97.83	98.90
Cotton target spot	100.00	95.56	100.00	97.73
Maize gray leaf spot	98.00	96.08	98.00	97.03
Maize northern leaf light	96.00	97.96	96.00	96.97
Maize rust disease	100.00	100.00	100.00	100.00
Maize stripe disease	100.00	98.36	100.00	99.17
Potato early light	97.78	100.00	97.78	98.88
Potato late blight	100.00	97.83	100.00	98.90
Soybean angular leaf spot	98.04	98.04	98.04	98.04
Soybean rust disease	98.25	100.00	98.25	99.12
Soybean sudden S	100.00	100.00	100.00	100.00
Soybean yellow mosaic	100.00	100.00	100.00	100.00
Wheat brown rust	98.00	100.00	98.00	98.99
Wheat powdery mildew	100.00	100.00	100.00	100.00
Wheat stripe rust	100.00	100.00	100.00	100.00

4.3. Comparison of visualization result

In order to verify that the semantic guidance of disease symptom text can make the model focus on the lesion area more accurately, we use Grad-CAM [27] to visually analyze the attention distribution of different models in the lesion area. As shown in Figure 4, by comparing the performance of MCDL-CVL, EMO Transformer, Multiscale ViT and Conformer on CropsDisease-5M16 data set, it can be seen that the attention area of MCDL-CVL can cover the lesion more comprehensively and accurately, while the attention to the non-pathological background is significantly reduced. This result directly confirmed the excellent performance of MCDL-CVL in lesion localization.

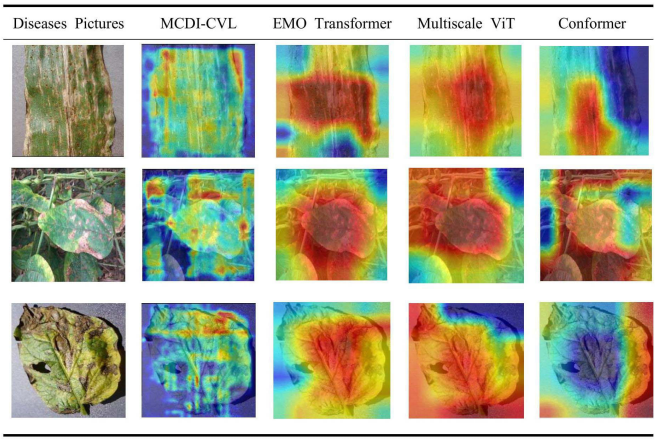

Figure 4: Comparison of visualization results.

4.4. Ablation experiments

To evaluate the individual and joint contributions of the proposed modules, we conduct a series of ablation experiments on three key components of MCDI-CVL: TEDE, BCMA, and AFGM. Table 3 summarizes the performance of different module combinations. Integrating TEDE alone leads to a noticeable improvement over the unimodal baseline, highlighting the value of incorporating adequate symptom-aware textual features for crop disease identification. When combined with AFGM, the model gains an additional 1.83% in accuracy by adaptively fusing visual and textual modalities. Pairing TEDE with BCMA yields a 2.35% improvement, underscoring the importance of fine-grained cross-modal alignment for capturing complementary information.

Table 3: Ablation Study of Module Combinations(%)

Modules			Metrics			
TEDE	BCMA	AFGM	Accuracy	Precision	Recall	F1
			93.88	94.57	94.29	94.10
✓			95.70	95.80	96.02	95.83
✓	✓		97.53	97.49	9.65	97.50
✓		✓	98.05	98.08	97.99	98.01
✓	✓	✓	**98.96**	**98.99**	**98.99**	**98.98**

In addition, the curve of the accuracy of each model with the training rounds in the ablation experiment is shown in Figure 5. Among them, TEDE module significantly improves the baseline performance, and the accuracy is further improved after combining AFGM or BCMA, and full MCDI-CVL model, which integrates all three components, achieves the highest accuracy of 98.96%. All the above experimental results verify the effectiveness of our multimodal crop disease recognition framework.

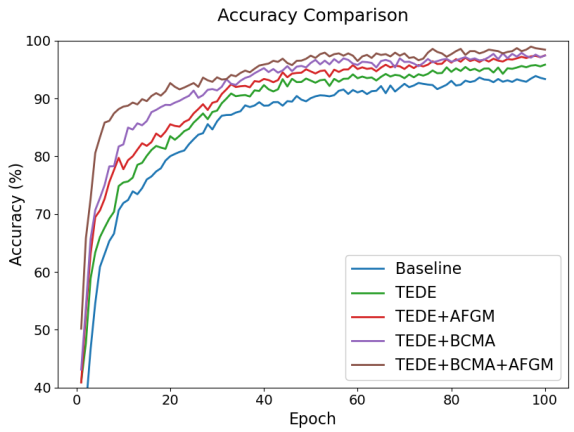

Figure 5: Accuracy curve of different groups in ablation experiment.

In order to evaluate the impact of text quality on multimodal recognition, two groups of comparative experiments were set up. In terms of text data, it is divided into the text not processed after Janus' generation and the text modified by experts. The two groups of experiments used the same image feature extraction and fusion architecture to ensure that the performance difference only came from the text quality variables. The experimental results are shown in Table 4, compared with the text without expert processing, the text revised by experts has significantly improved in various evaluation indicators, which verifies the effectiveness of expert revision in improving the quality of the text.

Table 4: Performance comparison between unprocessed text and revised text.

Method	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
Unprocessed	96.22	96.15	96.27	96.17
Revised	**98.96**	**98.99**	**98.99**	**98.98**

4.5. Comparison with other multimodal models

To further verify the performance of MCDI-CVL, a comparative experiment was conducted against WCG-Vmamba [33], a recently proposed multimodal classification model for corn diseases. WCG-Vmamba also introduced a corresponding multimodal dataset, which contains 4,633 images covering Fusarium wilt, rust, gray spot, and healthy corn leaves. Each image is paired with descriptive text generated by the LLaVA model, forming an image–text pair [33]. To ensure a fair and valid evaluation, all experiments were conducted on this dataset. As shown in Table 5, MCDI-CVL consistently outperforms the WCG-Vmamba model in identifying corn diseases. This demonstrates that MCDI-CVL can more effectively integrate and utilize multimodal features for accurate disease recognition.

Table 5: Performance comparison between WCG-VMamba and MCDl-CVL on the corn disease multimodal dataset.

Method	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
WCG-VMamba	96.97	95.94	96.04	95.99
MCDI-CVL	**97.52**	**97.57**	**96.51**	**97.01**

5. Conclusion

In this paper, we propose MCDI-CVL, a multimodal framework for crop disease recognition that integrates visual and textual cues through hierarchical encoding, cross-modal attention, and adaptive fusion, significantly improving disease recognition accuracy. We also introduce CropsDisease-5M16, a new image–text dataset spanning

multiple crops and disease types. Extensive experiments have demonstrated that MCDI-CVL is more effective than existing advanced image models and multimodal models across different datasets. Future work will focus on incorporating a broader range of crop diseases, developing lightweight optimizations to further enhance the applicability of the model in the real world as a human-machine collaborative intelligent agricultural system for auxiliary diagnosis, and releasing the CropsDisease-5M16 dataset to support further research.

Acknowledgement

Jianxun Lou acknowledges support from the PhD Research Start-up Fund of Northeast Electric Power University (Grant No. BSJXM-2025117).

References

[1] R. Arumuga Arun and S. Umamaheswari. Effective multi-crop disease detection using pruned complete concatenated deep learning model. *Expert Systems with Applications*, 213:118905, 2023. 2

[2] C. H. Bock, G. H. Poole, P. E. Parker, and T. R. G. and. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical Reviews in Plant Sciences*, 29(2):59–107, 2010. 1

[3] Y. Cao, L. Chen, Y. Yuan, and G. Sun. Cucumber disease recognition with small samples using image-text-label-based multi-modal language model. *Computers and Electronics in Agriculture*, 211:107993, 2023. 2

[4] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *CoRR*, abs/2501.17811, 2025. 3

[5] G. Dai, J. Fan, and C. Dewi. Itf-wpi: Image and text based cross-modal feature fusion model for wolfberry pest recognition. *Computers and Electronics in Agriculture*, 212:108129, 2023. 2

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 4

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 1

[8] A. Elhassouny and F. Smarandache. Smart mobile application to recognize tomato leaf diseases using convolutional neural networks. In *2019 International Conference of Computer Science and Renewable Energies (ICCSRE)*, pages 1–4, 2019. 1

[9] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6804–6815, 2021. 6

[10] X. Feng, C. Zhao, C. Wang, H. Wu, Y. Miao, and J. Zhang. A vegetable leaf disease identification model based on image-text cross-modal feature fusion. *Frontiers in Plant Science*, Volume 13 - 2022, 2022. 2

[11] Y. Gao, R. Li, E. Croxford, J. Caskey, B. W. Patterson, M. Churpek, T. Miller, and D. Dligach. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670, 2025. 3

[12] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. 2

[13] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. 6

[14] W.-J. Hu, J. Fan, Y.-X. Du, B.-S. Li, N. Xiong, and E. Bekkering. Mdfc–resnet: An agricultural iot system to accurately recognize crop diseases. *IEEE Access*, 8:115287–115298, 2020. 1

[15] X. Huang, D. Xu, Y. Chen, Q. Zhang, P. Feng, Y. Ma, Q. Dong, and F. Yu. Econv-vit: A strongly generalized apple leaf disease classification model based on the fusion of convnext and transformer. *Information Processing in Agriculture*, 2025. 1

[16] T. Iizumi and T. Sakai. The global dataset of historical yields for major crops 1981–2016. *Scientific Data*, 7(1):97, 2020. 3

[17] T. Ilyas, A. Khan, M. Umraiz, Y. Jeong, and H. Kim. Multi-scale context aggregation for strawberry fruit recognition and disease phenotyping. *IEEE Access*, 9:124491–124504, 2021. 1

[18] R. Karthik, A. Ajay, A. Singh Bisht, T. Illakiya, and K. Suganthi. A deep learning approach for crop disease and pest classification using swin transformer and dual-attention multi-scale fusion network. *IEEE Access*, 12:152639–152655, 2024. 1

[19] H. Li, B. Chen, J. Chen, S. Li, F. He, and Y. Hu. Itimca: Image-text information and cross-attention for multi-modal cassava leaf disease classification based on a novel multi-modal dataset in natural environments. *Crop Protection*, 189:106981, 2025. 1

[20] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. 1

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 1, 6

[22] S. Mehta and M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 6

[23] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, and Q. Ye. Conformer: Local features coupling global representations for recognition and detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8):9454–9468, August 2023. 6

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021. 1

[25] A. J. Rozaqi and A. Sunyoto. Identification of disease in potato leaves using convolutional neural network (cnn) algorithm. In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pages 72–76, 2020. 1

[26] S. Savary, L. Willocquet, S. J. Pethybridge, et al. The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3:430–439, 2019. 3

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 6

[28] V. Sharma, A. K. Tripathi, H. Mittal, and L. Nkenyereye. Soyatrans: A novel transformer model for fine-grained visual classification of soybean leaf disease diagnosis. *Expert Systems with Applications*, 260:125385, 2025. 2

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

[30] E. Stukenbrock and S. Gurr. Address the growing urgency of fungal disease in crops. *Nature*, 2023. 1

[31] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 10096–10106, 2021. 6

[32] H. T. Van, G. Van Vu, T. Thanh Tuan, B. Vo, and Y. S. Chung. Lgenetb4ca: A novel deep learning approach for chili germplasm differentiation and leaf disease classification. *Computers and Electronics in Agriculture*, 233:110149, 2025. 2

[33] H. Wang, M. He, M. Zhu, and G. Liu. Wcg-vmamba: A multi-modal classification model for corn disease. *Computers and Electronics in Agriculture*, 230:109835, 2025. 2, 7

[34] D. Wei, J. Chen, T. Luo, T. Long, and H. Wang. Classification of crop pests based on multi-scale feature fusion. *Computers and Electronics in Agriculture*, 194:106736, 2022. 1

[35] Y. Wu, F. Liu, L. Wan, and Z. Wang. Intelligent fault diagnostic model for industrial equipment based on multimodal knowledge graph. *IEEE Sensors Journal*, 23(21):26269–26278, 2023. 3

[36] J. Zhang, X. Li, J. Li, L. Liu, Z. Xue, B. Zhang, Z. Jiang, T. Huang, and Y. Wang. Rethinking mobile block for efficient attention-based models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1389–1400, 2023. 4, 6

[37] J. Zhou, J. Li, C. Wang, H. Wu, C. Zhao, and G. Teng. Crop disease identification and interpretation method based on multimodal deep learning. *Computers and Electronics in Agriculture*, 189:106408, 2021. 2

[38] H. Zhu, W. Shi, X. Guo, S. Lyu, R. Yang, and Z. Han. Potato disease detection and prevention using multimodal ai and large language model. *Computers and Electronics in Agriculture*, 229:109824, 2025. 3