PestVL-Net: Enabling Multimodal Pest Learning via Fine-grained Vision-Language Interaction
Appendix

Xueheng Li[1,2]*, Tao Hu[1,2]*, Ke Cao[1,2], Runsheng Qi[1], Huixin Zhang[1,2],
Rui Li[1], Jie Zhang[1,3], Chengjun Xie[1,3]†
[1]Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences
[2]University of Science and Technology of China
[3]Zhongke Hefei Institute of Technology Innovation Engineering
{lixueheng,ht_simon}@mail.ustc.edu.cn, {lirui,zhangjie,cjxie}@iim.ac.cn

1. Experiment

1.1. Dataset and Benchmark

Data Collection and Preprocessing. To advance the practical application of artificial intelligence in agricultural settings, we introduce two multi-species pest datasets, QianFSD and AgriInsect. We collected high-definition images of crop pests from real agricultural environments across multiple rural areas in China. We preprocessed the raw captured images by cropping and magnifying the central regions featuring the primary pest subjects. Ultimately, our constructed QianFSD dataset comprises 7,054 images across 141 categories, while the AgriInsect dataset contains 9,452 images spanning 200 categories. These images provide clear details of pest morphology, surface textures, and other characteristics, as shown in Figure 7. Prior to training, all images were uniformly resized to a resolution of 224 $\times$ 224 pixels. In contrast to existing datasets, which often feature limited pest species and low-quality images sourced from the internet, our dataset includes a broader range of pest species with high-quality images captured in actual production scenarios. We believe this work will contribute to the development of practical smart agricultural technologies in real-world applications.

Benchmark Visualization. Figures 1 and 2 illustrate the data distributions of the QianFSD and AgriInsect datasets, respectively. As observed, while the data exhibits a long-tail distribution, the class imbalance is relatively mild. Overall, our datasets provide an abundance of authentic, real-world pest images, serving as a robust foundation for the further training of deep learning models and MLLMs.

Knowledge Organization. In addition to pest images, we collaborated with experienced agricultural experts to compile relevant knowledge for each pest species, including core morphological features such as color, spots, and wings. Leveraging this expertise, we employed Chain-of-Thought (CoT) reasoning and Multimodal Large Language Model (MLLMs) priors to annotate the visual images, creating a multimodal pest dataset. All datasets and associated code will be made publicly available.

Baseline Selection. To ensure a more comprehensive evaluation of our framework, we selected a diverse set of deep learning models for training and testing across multiple pest datasets. This selection includes both classic network architectures and cutting-edge, state-of-the-art models, allowing us to assess the performance of our approach relative to a broad spectrum of methodologies. By incorporating a variety of models, we aim to provide a well-rounded comparison that highlights the strengths and versatility of our framework in handling effective and efficient pest recognition challenges.

1.2. Benchmark Comparison

As noted in main text, existing datasets (e.g., BIOSCAN-5M [1] and AgroBench [5]) focus on fundamental biological and agricultural research. In contrast, our dataset targets real-world pests responsible for agricultural damage and incorporates expert-defined features, as shown in Figure 3.

1.3. Expert Knowledge

Due to space constraints in the main text, we provide a more detailed explanation of the expert knowledge here. To enhance the accuracy of general MLLMs outputs in describing pest morphological features, we consulted agricultural experts with extensive experience to compile detailed descriptions of various pests, as shown in Figure 5. We organized and categorized the original knowledge provided by the experts into core visual features (i.e., Coloration, Spots and

*Equal contribution.
†Corresponding author.

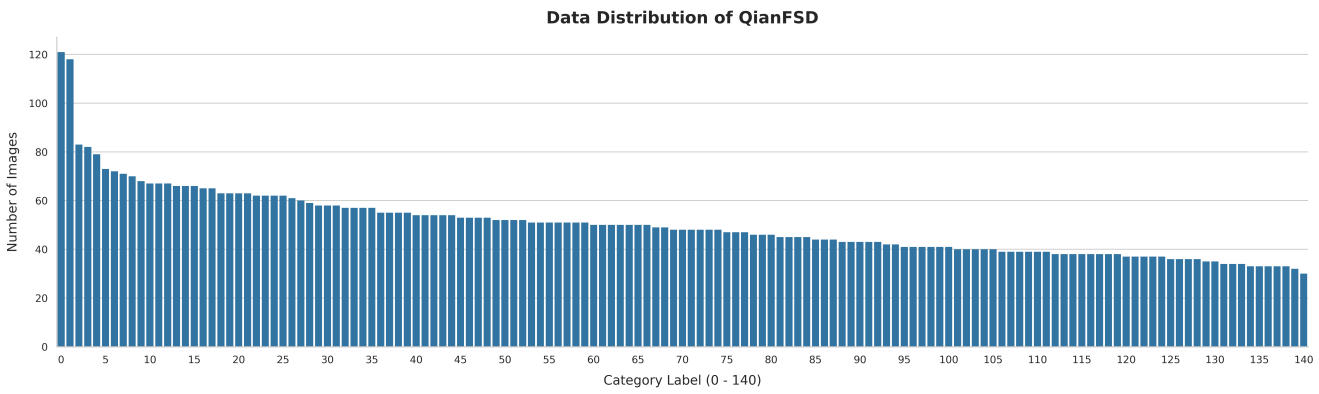

Figure 1. Data distribution of QianFSD.

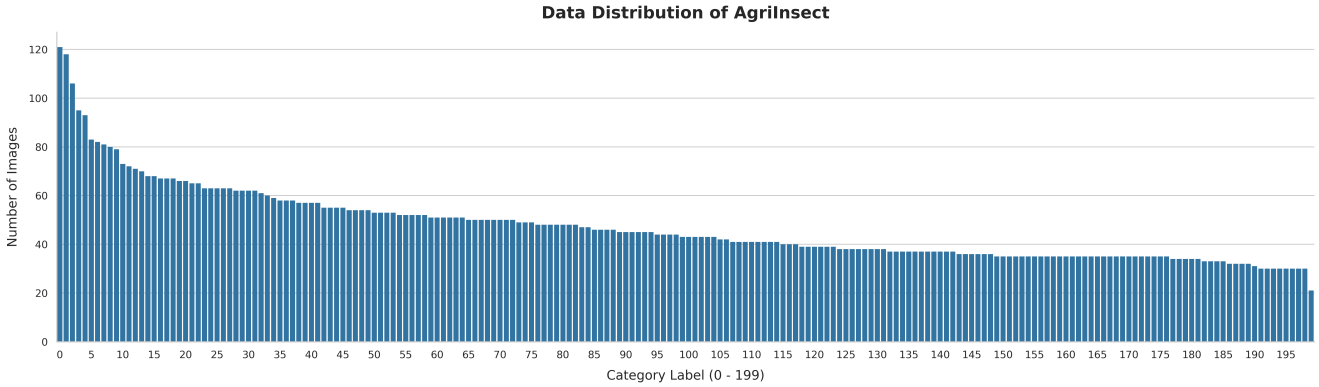

Figure 2. Data distribution of AgriInsect.

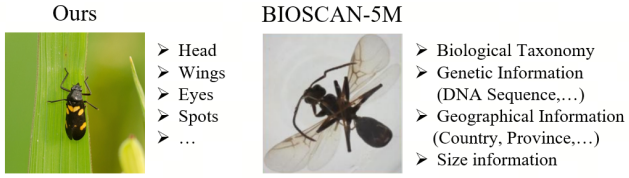

Figure 3. Comparison of dataset characteristics.

Surface Texture, Eyes and Antennae, and Wing Morphology) to guide the fine-grained understanding of MLLMs. Manual verification and related ablation experiments have shown that expert knowledge enhances the accuracy of pest descriptions. Beyond visual characteristics, we also gathered information on the taxonomic status, geographical distribution, and damage reports of these pests, etc. All expert knowledge will also be made public.

1.4. Architecture Comparison

Extensive research [3, 6, 7] has demonstrated the advantages of RWKV over Transformers and Mamba, particularly regarding global image perception and computational efficiency. To illustrate this, we provide a visual demonstration of these capabilities below. Figure 4 visualizes the Effective Receptive Fields across architectures. The wider distribution of dark regions implies larger receptive fields, clearly validating RWKV's advantage. Additionally, Table 1 details module-wise ablation studies to confirm the performance benefits of RWKV.

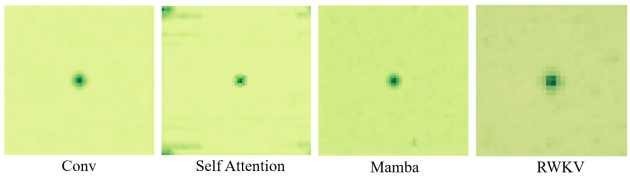

Figure 4. Receptive field comparison across architectures

Table 1. Comparison of architectural ablations on Li's dataset.

Config	Accuracy	Precision	F1 Score	GM
Vision Transformer	82.90	83.07	83.13	84.91
Linear Transformer	81.83	82.31	81.58	83.33
Vision Mamba	77.81	80.27	80.52	81.30
Ours	**88.49**	**89.36**	**88.52**	**88.56**

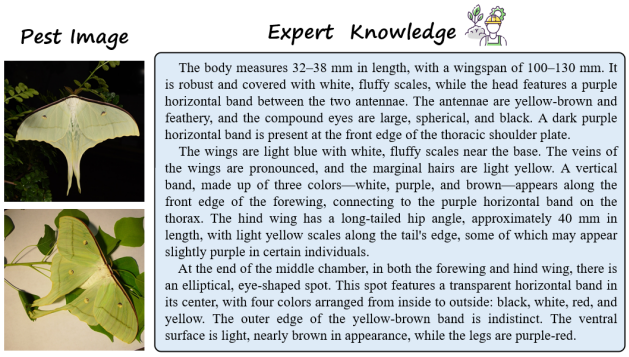

Figure 5. The example of expert knowledge, primarily focused on the visual characteristics of pests. Original agricultural expert knowledge is organized and utilized to guide MLLMs in generating captions for pest images.

1.5. Inference Setting

The MLLMs are excluded from the inference pipeline. The text inputs are pre-processed offline by MLLMs and simply retrieved during inference. The text utilized during inference describes morphological features rather than category identity, ensuring no label leakage. Alternatively, generic descriptions (e.g., "A photo of a pest") can be used to activate the text branch. Simple generic inputs remain competitive with advanced baselines, confirming the effectiveness of our text branch.

1.6. Ablation Experiments

To more thoroughly evaluate the effectiveness of our framework, we extend the experimental setup used in the main text and include the results of ablation studies on the QianFSD and AgriInsect datasets, as presented in Tables 3 and 4. These additional ablation experiments provide deeper insights into the contribution of each key component within our framework, shedding light on their individual impact on fine-grained pest recognition. The results underscore the significance of each element in enabling our framework to capture the subtle yet crucial features of pests, further validating its robustness to handle complex pest learning tasks.

1.7. Text Comparison

Although we selected proprietary MLLMs for the main experiments due to their robust multimodal capabilities, they can be replaced by open-source alternatives. We verify this by re-annotating Lidataset via Qwen2.5VL-32B. The results in Table 2 indicate that open-source models, despite a marginal performance drop, serve as viable alternatives to proprietary models.

In the main text, we investigated the impact of text generation methods (CoT). In Table 2, we further compare the performance impact of generic text (e.g., "A photo of a pest"), Qwen-generated text, and proprietary model outputs. The results (in Table 2) confirm that our method achieves best performance, indicating the generation of more precise pest descriptions.

Table 2. Performance comparison with different text inputs.

Config	Accuracy	Precision	F1 Score	GM
Generic Text	76.57	77.98	76.61	76.70
Qwen2.5VL-32B	86.32	85.98	86.55	86.07
GPT-4o (Ours)	**88.49**	**89.36**	**88.52**	**88.56**

1.8. Computational Complexity

Due to space limitations in the main text, we provide a supplementary comparison of the computational complexity for each network architecture, as shown in Table 5. It is important to note that the DFL algorithm [2], which primarily adjusts the model's training strategy and loss function, is not included here. The comparison results demonstrate that our method achieves both lower computational overhead and superior performance. Although the VMamba architecture has a lower computational cost, its pest recognition capabilities are limited. This further highlights the suitability of our approach for real-world agricultural pest recognition and management.

Table 3. Ablation study results of different components on the QianFSD dataset.

Config	**QianFSD**			
	Accuracy	Precision	F1 Score	GM
(I)	78.39	79.38	78.63	79.45
(II)	82.64	83.26	82.16	82.97
(III)	83.38	84.62	83.57	84.86
(IV)	82.10	83.47	82.52	82.69
(V)	85.92	86.31	85.72	86.31
Ours	**86.72**	**87.53**	**86.86**	**87.14**

1.9. Visualization

To demonstrate that our method effectively captures the core features of pests, we conducted a series of additional visualization experiments using Grad-CAM [4], as illustrated in Figure 6. For comparison, we selected VMamba, another linear model, which allows us to evaluate the performance of our approach relative to existing methods. The results clearly show that our model not only identifies the key morphological characteristics of pests but also excels in capturing their fine-grained details, which are crucial for accurate pest recognition and classification.

Table 4. Ablation study results of different components on the AgriInsect dataset.

Config	AgriInsect			
	Accuracy	Precision	F1 Score	GM
(I)	82.19	82.80	82.66	82.11
(II)	85.45	86.48	85.94	85.53
(III)	87.01	87.54	87.15	88.63
(IV)	86.56	86.37	87.28	86.97
(V)	89.52	89.87	88.73	89.82
Ours	**90.15**	**90.88**	**90.19**	**90.51**

Table 5. Comparison of computational complexity of different network architectures. Our framework strikes a balance between a small number of parameters and exceptional performance.

Methods	Flops (G)	Params (M)
AlexNet	**0.71**	61.1
VGG16	19.63	143.67
ResNet-18	1.82	11.69
ResNet-50	4.13	25.56
ResNeXt-50	4.29	25.02
Swin-T	2.98	19.62
UniFormer	7.77	49.33
SeaFormer	3.21	13.95
VMamba-B	0.99	**6.67**
TransXNet	4.57	25.54
Ours	4.16	12.26

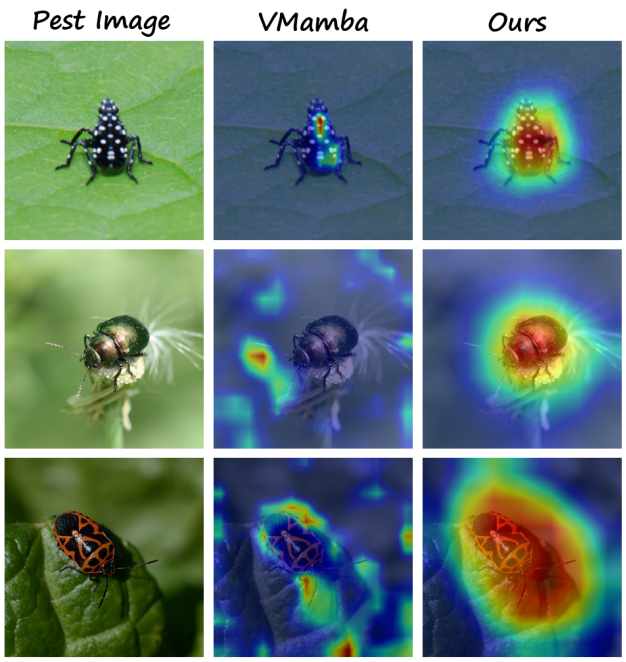

Figure 6. Comparison of Grad-CAM visualization results.

References

[1] Zahra Gharaee, Scott C Lowe, ZeMing Gong, Pablo Millan Arias, Nicholas Pellegrino, Austin T Wang, Joakim Bruslund Haurum, Iuliia Eyriay, Lila Kari, Dirk Steinke, et al. Bioscan-5m: A multimodal dataset for insect biodiversity. *Advances in Neural Information Processing Systems*, 37:36285–36313, 2024. 1

[2] Tao Hu, Jianming Du, Keyu Yan, Wei Dong, Jie Zhang, Jun Wang, and Chengjun Xie. Causality-inspired crop pest recognition based on decoupled feature learning. *Pest Management Science*, 80(11):5832–5842, 2024. 3

[3] Xueheng Li, Xuanhua He, Tao Hu, Jie Zhang, Man Zhou, Chengjun Xie, Yingying Wang, and Bo Huang. Freq-rwkv: Granularity-aware spatial-frequency synergy via dual-domain recurrent scanning for pan-sharpening. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1890–1899, 2025. 2

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[5] Risa Shinoda, Nakamasa Inoue, Hirokatsu Kataoka, Masaki Onishi, and Yoshitaka Ushiku. Agrobench: Vision-language model benchmark in agriculture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7634–7644, 2025. 1

[6] Zhiwen Yang, Jiayin Li, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-rwkv: Efficient and effective medical image restoration with rwkv. *IEEE Journal of Biomedical and Health Informatics*, 2025. 2

[7] Haobo Yuan, Xiangtai Li, Lu Qi, Tao Zhang, Ming-Hsuan Yang, Shuicheng Yan, and Chen Change Loy. Mamba or rwkv: Exploring high-quality and high-efficiency segment anything model. *arXiv preprint arXiv:2406.19369*, 2024. 2

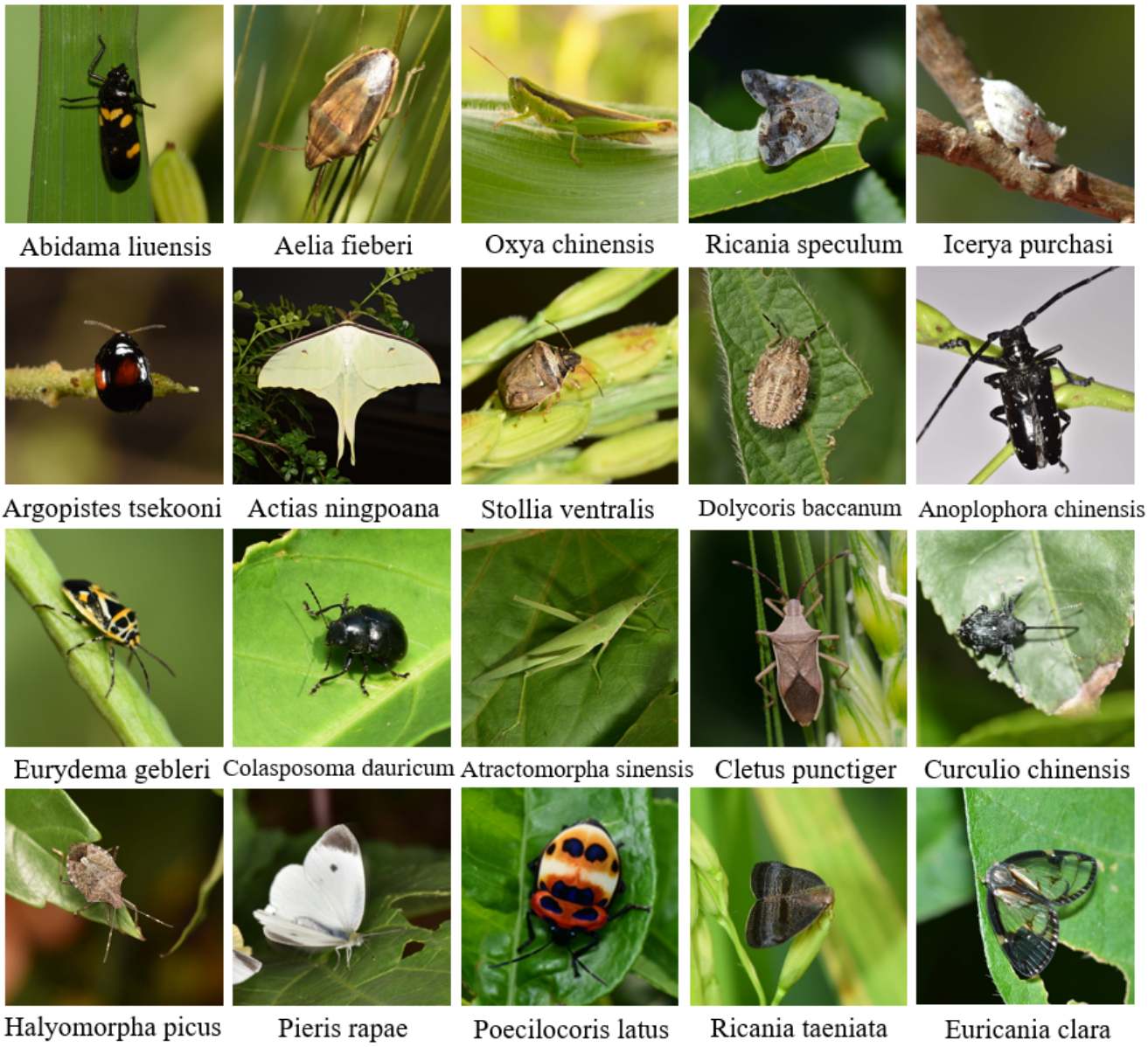

Figure 7. Examples of pest images from various categories in the QianFSD and AgriInsect datasets.