IP-VQA Dataset: Empowering Precision Agriculture with Autonomous Insect Pest Management through Visual Question Answering

Kairui Jin
School of Computer Science
University of Technology Sydney
Sydney, New South Wales, Australia
Kairui.Jin@student.uts.edu.au

Xing Zi
School of Computer Science
University of Technology Sydney
Sydney, New South Wales, Australia
Xing.Zi-1@uts.edu.au

Karthick Thiyagarajan
Smart Sensing and Robotics Laboratory (SensR Lab), Centre for Advanced Manufacturing Technology
Western Sydney University
Penrith, New South Wales, Australia
K.Thiyagarajan@westernsydney.edu.au

Ali Braytee
School of Computer Science
University of Technology Sydney
Sydney, New South Wales, Australia
Ali.Braytee@uts.edu.au

Mukesh Prasad
School of Computer Science
University of Technology Sydney
Sydney, New South Wales, Australia
Mukesh.Prasad@uts.edu.au

Abstract

Precision agriculture is essential for social good, global economy and food security, yet insect pests threaten productivity through crop damage, pathogen spread, and rising pest control costs. The overuse of pesticides leads to environmental issues and pesticide resistance. Advanced technologies like Visual Question Answering (VQA) provide solutions by integrating image processing with natural language understanding, facilitating efficient pest detection and crop health monitoring. While datasets like IP102 have enhanced pest recognition, they lack necessary question-answer pairs for VQA tasks in agriculture. To address this gap, we introduce the Insect Pest Visual Question Answering (IP-VQA) dataset, designed specifically for precision agricultural applications. This dataset includes a diverse collection of high-quality images annotated with detailed question-answer pairs related to crop health, pest identification, and agricultural practices. Our thorough data collection ensures reliability and relevance. We also utilize advanced multimodal large language models to set a benchmark for the dataset. The primary contribution of the IP-VQA dataset lies in its comprehensive coverage and VQA integration within agricultural contexts. By providing rich visual and textual information, it connects VQA techniques to practical agricultural needs, supporting ongoing research and paving the way for future studies in precision agriculture.

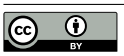
CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; *Visual inspection*; *Natural language processing*; *Natural language generation*;

Keywords

IP-VQA Dataset, Inspect Pest Management, Precision Agriculture, Visual Question Answering, Autonomous System.

ACM Reference Format:
Kairui Jin, Xing Zi, Karthick Thiyagarajan, Ali Braytee, and Mukesh Prasad. 2025. IP-VQA Dataset: Empowering Precision Agriculture with Autonomous Insect Pest Management through Visual Question Answering. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3701716.3718386



WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3718386

1 Introduction

Precision agriculture is essential for the global economy and food security. With recent global focus to address the food security concerns and promote sustainable agriculture for social good as outlined in the United Nations Sustainable Goal (SDG 2), the need for precision agriculture is immensely growing now than ever before. As modern agricultural techniques advance, the integration of innovative technologies, including precision agriculture, is increasingly crucial. Insect pests, however, pose a significant threat to agricultural productivity, resulting in considerable economic losses and a decline in crop quality. They cause direct harm by feeding on various plant parts, including leaves, stems, fruits, and roots, which leads to poor growth, reduced yields, and even plant mortality. Furthermore, insect pests can transmit pathogens, exacerbating crop diseases and diminishing yields.

The escalating need for pest control increases production costs due to frequent pesticide purchases and implementation of management strategies. The overuse of pesticides also raises concerns about residues in food products and the emergence of pesticide-resistant pests, complicating future management efforts. Additionally, the ecological impacts of pesticide usage threaten nontarget organisms and disrupt ecosystem balance. The challenges of pest identification and management are intensified by the varied knowledge levels among farmers, the diversity of pest species, and the absence of structured agricultural information. Thus, leveraging external

knowledge can assist farmers in identifying pests and selecting more effective, environmentally friendly control methods.

To tackle these challenges, advanced technologies such as Visual Question Answering (VQA) show promise in agricultural applications, including pest detection, crop health monitoring, and classification. VQA, an artificial intelligence task that merges image processing with natural language understanding, enables models to answer questions based on visual content. In pest control, a VQA system allows users to obtain critical pest information directly from images via natural language interaction, significantly enhancing diagnostic accuracy and efficiency.

While existing datasets like IP102 have made substantial contributions to pest recognition, comprising over 18,981 images across 102 pest categories, they lack the necessary question-answer pairs for effective VQA applications. Moreover, despite advancements in VQA technologies, there is a scarcity of comprehensive datasets tailored for agricultural purposes, particularly those that combine visual and textual data for VQA tasks. Many existing datasets are limited in scope or insufficiently represent the diverse challenges faced in agriculture.

To bridge this gap, we present the Insect Pest Visual Question Answering (IP-VQA) dataset, aimed at advancing research and development in this critical field. The IP-VQA dataset includes a diverse array of images and annotations covering a wide spectrum of crops, pests, and environmental conditions. Our dataset features high-quality images annotated with detailed question-answer pairs addressing various aspects of crop health, pest identification, and agricultural practices. A rigorous data collection and annotation process ensures the dataset's reliability and relevance.

The primary contribution of the IP-VQA dataset is its extensive coverage and innovative approach to integrating VQA with agricultural applications. By providing a rich source of visual and textual information, this dataset seeks to connect advanced VQA techniques with practical agricultural needs. It not only supports ongoing research efforts but also opens new avenues for future studies in smart farming and precision agriculture.

The potential applications of the IP-VQA dataset are broad, encompassing automated pest detection systems and intelligent crop monitoring tools. This dataset will enable researchers and practitioners to develop more effective and efficient solutions to various agricultural challenges. Furthermore, its extensibility ensures adaptability to evolving research requirements and the incorporation of new agricultural insights. The contributions of this article are as follows:

- Developed the first VQA dataset specifically for the insect pest domain and established a benchmark for its evaluation.[1]
- Applied a range of Generative AI techniques, including Vision-Language Models (VLM) and Large Language Models (LLM), to create the dataset using customizable combination structures that facilitate automated and controlled outputs at each stage.
- Designed a flexible pipeline that can be adapted for constructing VQA datasets in other fields after fine-tuning.

[1]https://drive.google.com/drive/folders/1ZW3hrLp-ByK1zTy9Uv31m11moXOQ5i-M?usp=drive_link

2 Related work

VQA is a significant interdisciplinary research field that merges computer vision and natural language processing, enabling models to answer questions based on visual content. The VQA dataset, introduced by [2], provided a foundational benchmark for VQA tasks. However, this initial dataset showed biases in language patterns. To address this, [8] released the VQA v2 dataset, balancing question-answer pairs to reduce language priors and strengthen visual comprehension. In model advancements, [1] proposed a bottom-up and top-down attention mechanism, which improved VQA performance by enabling models to focus on relevant image regions. [23] provided practical insights from the 2017 VQA Challenge, while [12] introduced Bilinear Attention Networks (BAN) to better capture complex interactions between visual and textual inputs.

Pretrained models have further advanced VQA capabilities. [18] introduced ViLBERT, a model for visio-linguistic representation pre-training, which enhanced performance across various vision-language tasks, including VQA. [27] presented Deep Modular Co-Attention Networks, using a modular architecture to flexibly co-attend to image and question features. In feature representation, [10] advocated for grid features over region features, simplifying models while maintaining performance. [14] proposed Relation-aware Graph Attention Networks to enhance object relationship modeling, boosting reasoning abilities. [7] contributed a dynamic fusion mechanism that uses intra- and inter-modality attention flows for better VQA performance.

Despite extensive research in general VQA, its application in agriculture is emerging. In agricultural imaging, [21] applied deep CNNs to plant disease recognition, and [6] used CNNs for plant disease diagnosis, highlighting deep learning's effectiveness in classifying plant health. [11] surveyed deep learning in agriculture, suggesting that techniques like VQA could improve agricultural decision-making. [24] introduced Deep Plant Phenomics for plant phenotyping, aligning with the objectives of agricultural VQA.

For multi-modal fusion, [26] used a fusion network for plant disease recognition, integrating image data with other modalities, akin to VQA's multi-modal nature. [19] developed real-time semantic segmentation for crop and weed classification, which VQA could expand by integrating question-based interaction. Remote sensing data has also benefited from artificial intelligence; [29] applied deep learning to crop classification via satellite imagery, and [3] highlighted the potential of combining visual and language understanding. Furthermore, domain-specific VQA datasets, such as those in medical imaging [13], set a promising precedent for creating VQA datasets in agriculture. Together, these developments underscore the potential of agricultural VQA as a promising area for future research.

3 Methodology

This paper introduces the first IP-VQA dataset specifically for insect pest detection, created by generating Image-QA pairs from an existing insect pest recognition dataset using a machine-human co-curation approach. The process involves three stages: data integrity enhancement, concept alignment, and instruction adherence.

Stage I: Insect Pest Data integrity. The original dataset is IP102 dataset. It consists of real-world images, insect object IDs and

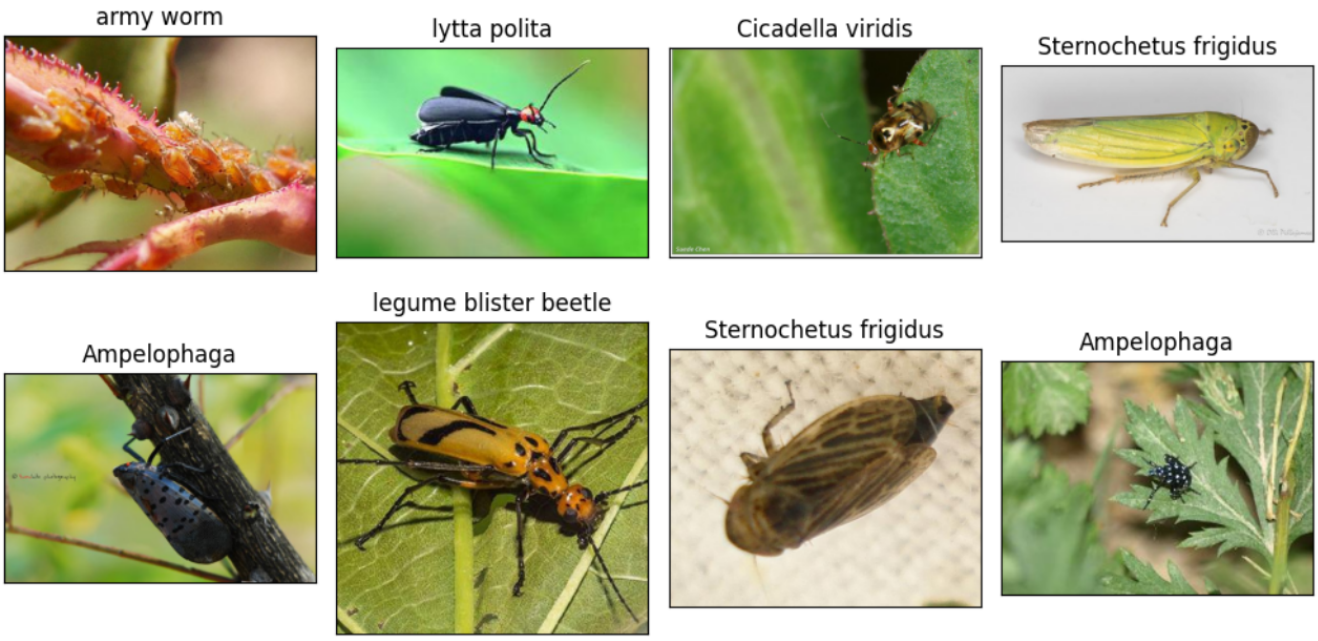


Figure 1: Sample images from the IP102 dataset.

position information for each image. To build a VQA dataset, text information is necessary, which will guide the LLM to generate the description of the image based on the given QA pairs. For each input image $\mathbf{X}_\mathrm{v}$, a prompt $\mathbf{X}_\mathrm{p}$ is designed to guide the LLM to generate answers to ensure the integrity of the data. The $\mathbf{X}_\mathrm{p}$ consists of 10 questions $\mathbf{X}_\mathrm{q}$ and the head of the answer:

$$\text{Question: } \mathbf{X}_{q_i} \langle \text{STOP} \rangle \text{? Answer: } \langle \text{STOP} \rangle$$
$$\mathbf{i} = [0, 1, 2, \ldots, 8, 9]$$

Each question will be answered using LLM based on the given image. In addition, the position information will also be included in the dataset to improve the accuracy of the description.

Stage II: Insect Pest Concept Alignment. Based on the internal information of the IP102 dataset, and consider of the real-world problems, we designed 10 questions to maximize the balance between data plurality and practical application.

(1) **Relevance to Farmers**: Ensuring that while the data covers a wide range of pests, the questions are still highly relevant to the problems faced by farmers. This includes prioritizing pests that are most damaging or widespread.
(2) **Scalability**: Designing questions that can be scaled and adapted to different agricultural regions and practices. This allows for the dataset and model to be useful in various global contexts.
(3) **User-Friendly**: Making sure the questions and the resulting model outputs are easy to understand and use by non-experts. This involves clear, concise questions and interpretable answers.

By carefully considering these aspects, we aim to create a balanced set of questions that leverage the comprehensive data in the IP102 dataset while addressing the practical needs of those in the agricultural sector. This approach ensures that the model trained on this data is both robust and practically useful.

Stage III: Insect Pest Instruction-Tuning Data. To align the model with a variety of instructions, we generate textual data supporting multiple dialogues using a two-phase prompt with GPT-4o mini.

In Phase I, this paper prompts GPT-4o mini to reformat captions into the same format as the COCO captions dataset [15]. To avoid unnecessary verbosity, we limit the length of the captions, avoiding overly long responses that might include superfluous sentences. Furthermore, because pests are closely related to the climatic conditions and ecology of specific geographical areas, we restrict the responses generated to avoid sensitive information.

In Phase II, this paper design prompts that instruct GPT-4o mini to describe the scene comprehensively in natural language, as if it could see the image, even though it only has access to textual information. To prevent redundancy and misrepresentation resulting from GPT's potential conjectures, we constrain the generated content to objective descriptions only, excluding modifiers, adjectives, and any commentary on the images.

3.1 Pipeline

Based on the three-stage methodology mentioned above, this article designed the pipeline.

BLIP2. This is a multimodal model architecture proposed by Salesforce Research to efficiently combine visual and linguistic information [16]. It achieves this by linking pre-trained image encoders (e.g., ViT [5], Swin Transformer [17]) with large pre-trained

language models (e.g., GPT, T5 [20]), enabling the training of only one lightweight cross-modal adapter while freezing both components. This drastically reduces the computational resources and time required for training.

The core idea of BLIP-2 is to introduce a bridge module between visual and linguistic models that is responsible for converting visual features into an input format understandable by the linguistic model. This design allows the model to perform well on a variety of tasks, including image description, visual quizzing, and image retrieval.

The advantage of the model is the following.

(1) Efficient training: Since the main components are frozen, only a small number of parameters need to be trained, saving computational costs.
(2) Flexible: can be combined with different image encoders and language models to adapt to various application scenarios.
(3) Excellent performance: leading results were achieved on several benchmarks.

The version of BLIP2 that we used was Salesforce/blip2-opt-6.7b-coco.As described in part **3.1**, the BLIP2 prompts were carefully designed to allow BLIP2 to answer only a limited number of generic questions. This has the advantage of making the best possible use of BLIP2's map-reading ability and generating answers about the questions in preparation for the next step of generating a COCO caption.

GPT-4o mini. This is a compact version of the GPT-4 model designed to provide efficient natural language processing capabilities in computationally limited environments. Through parameter compression and architectural optimization of the original model, GPT-4o mini retains key features of GPT-4 while significantly reducing model size and computational requirements. GPT-4o mini can perform tasks such as text generation, language translation, dialogue systems and sentiment analysis, providing a good balance between performance and efficiency.

In order to create reasonable datasets and eliminate the bias caused by a single model. We used a combination of multimodal models. The overview of the pipeline is shown in Figure 2.

(1) First, we use BLIP2 to generate initial answers for a set of 10 predefined questions. These questions are carefully crafted to align with the task, and the generated answers serve as prompts for generating captions.
(2) Next, we employ GPT-4o mini with specific instructions to rewrite the 10 QA pairs in the COCO caption format. The COCO captions dataset, widely used in image captioning and computer vision research, provides structured textual descriptions for images. By following this format and limiting the caption length, we enable the GPT model to better understand the image-question relationships, resulting in more accurate responses.
(3) We further use a structured GPT-4o mini model, which integrates the insect pest's bounding box information with instruction messages, to generate detailed descriptions based on the context established in the previous step. The inclusion of bounding boxes allows the description to specify the pest's location in the image. The instruction message provides a template and constraints for description generation, first offering an overview of the image content and then referencing the CABI database for information on the pest's impact. The CABI database [4], maintained by the International Center for Agricultural Biosciences (CABI), is a trusted resource on agricultural, biological, and environmental sciences.

We compared the performance of GPT-4o and GPT-4o mini in generating text constrained by specific content requirements. Our findings show that, while both models produced similar content, GPT-4o was more time-consuming and often generated additional, unintended details. For example, despite setting restrictions on content length and knowledge sources in Step (3), GPT-4o occasionally added speculative information, as observed in our manual review. In contrast, GPT-4o mini not only enhanced dataset-building efficiency but also reliably adhered to the specified format requirements. We plan to consider GPT-4o as a candidate for generating the final benchmark; however, using GPT-4o mini helps avoid inflating benchmark results, facilitating a fairer comparison of outcomes.

4 Dataset

This is the first VQA dataset tailored for the field of insect pests, encompassing 102 common agricultural pest species. It includes 18,981 data entries, each featuring an image, a caption, and five generated question-answer pairs. Both captions and QA pairs are generated automatically through a custom-designed pipeline by the authors. The dataset will be made publicly available on GitHub.

4.1 Data Collection

The original dataset, IP102 (Insect Pest 102), is an image dataset developed to advance agricultural pest identification research, specifically focusing on pest image recognition. It contains images from 102 common agricultural pest species, totaling 18,981 images. Information on pest species and bounding boxes within the dataset was extracted to facilitate the generation of additional data. A word cloud, shown in Figure 5, visually represents the dataset's content by displaying the names of pest species. Frequently occurring species are displayed in larger, more prominent fonts, highlighting key terms and themes for quick insight into the dataset's primary content and trends. This word cloud effectively illustrates the dataset's breadth and diversity across various pest species. The word cloud is shown in Figure 3.

Figure 3: The word cloud of the dataset.

4.2 Dataset Details

The dataset comprises a total of 18,981 entries, each consisting of an insect image, a caption, and 5 question-answer pairs.

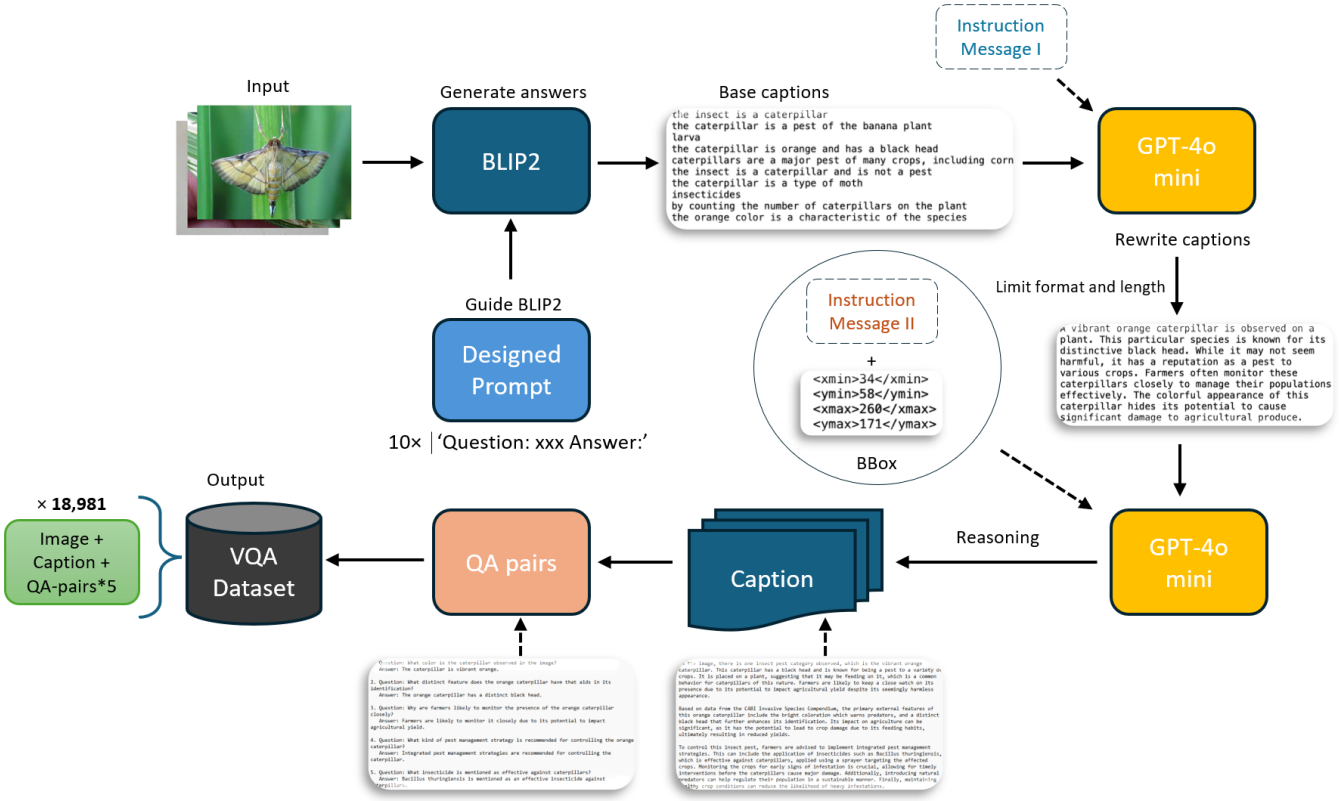


Figure 2: This is the full process of generating the VQA dataset. The details are mentioned in part 3.2.

Image. The images span 102 categories of common agricultural pests, with varying numbers of images per category, resulting in an uneven distribution. Pest types range from widely known crop pests like locusts, aphids, and leaf beetles to less common pests significant to specific crops. The image sources are diverse, reflecting variations in lighting, angles, backgrounds, and quality. Images were captured both in natural field conditions and in controlled settings like laboratories. Stored in standard JPEG format, the images vary in resolution, from high-definition to lower-resolution, simulating real-world conditions captured by different equipment.

Caption. The caption was generated sequentially from the image using multiple VLMs and LLM, along with manually designed sections. Further specifics will be discussed in Section IV. The caption comprises three main paragraphs: the first paragraph provides an objective description of the image, focusing on the number of pests, their names, and their spatial relationships within the image. The second paragraph outlines control measures for that particular insect species, detailing trapping methods, recommended insecticides, and guidelines on their application and timing. To ensure the authority of the information presented, the second and third paragraphs were generated using knowledge from the authoritative CABI database, rather than relying on sources like Wikipedia or other websites.

The CABI database is an essential resource in agriculture and life sciences, encompassing topics such as agriculture, environment, biosciences, and public health. It provides a comprehensive collection of academic literature, data resources, and expert recommendations, with significant applications in pest control, plant and animal health management, and ecological conservation. CABI supports farmers, researchers, and policymakers with reliable information to develop effective pest management strategies and sustainable agricultural policies. Through international collaboration and data sharing, CABI plays a vital role in advancing agricultural technology and biodiversity conservation on a global scale.

QA Pairs. The QA pairs were crafted by having GPTs engage in a rigorous self-dialogue, posing strict questions derived from the caption, and conducting multiple rounds of discussion before generating five QA pairs. These pairs relied exclusively on the content of the caption without incorporating any external knowledge. This approach ensured a diverse and randomized set of questions, mitigating the risk of inaccurate answers stemming from poorly designed questions. Furthermore, this method allows readers to generate as many QA pairs as needed without concerns about their accuracy and validity. An example of the IP-VQA dataset is shown in Figure 4.

Caption:

In the image, there is one visible insect pest identified as a moth. The moth is characterized by its distinctive wing structure, which features asymmetrical patterns and a coloration that blends into its natural habitat. Positioned on a leaf, it interacts with its environment primarily as an herbivore, feeding on the foliage.

Based on data from the CABI Invasive Species Compendium, the moth's impact includes potential damage to the vegetation it feeds on, which can lead to decreased plant health and agricultural yield losses. Such feeding behavior can also disrupt local ecosystems, as they may outcompete native species for resources.

To control this pest, integrated pest management strategies are recommended. One effective solution involves the use of insect traps to monitor and reduce the moth population. Additionally, applying insecticides such as Bacillus thuringiensis can be effective, particularly when targeting larvae stages. Timing the application early in the life cycle, right after egg hatch, maximizes effectiveness. Regular monitoring should continue to adjust control measures as necessary.

QA pairs

1. Question: What type of insect pest is identified in the image?
   Answer: The insect pest identified in the image is a moth.
2. Question: How does the moth interact with its environment according to the caption?
   Answer: The moth primarily interacts with its environment as an herbivore, feeding on the foliage.
3. Question: What potential impact does the moth have on vegetation?
   Answer: The moth can potentially damage vegetation, leading to decreased plant health and agricultural yield losses.
4. Question: What pest management strategy is recommended for controlling the moth?
   Answer: Integrated pest management strategies are recommended for controlling the moth.
5. Question: Which insecticide is mentioned as effective against the moth, specifically targeting which stage?
   Answer: Bacillus thuringiensis is mentioned as an effective insecticide, particularly targeting the larvae stages.

Figure 4: A sample content of the IP-VQA dataset.

4.3 Statistics

To analyze common question types in the dataset related to agricultural pest control, we employed the following approach.

(1) **Data Collection and Pre-processing:** We standardized the extraction of question texts from each file by matching the *Question* label using regular expressions.
(2) **Vector Representation:** Each question was converted into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) technique, which captures key features of the text [22].
(3) **Clustering Analysis:** We utilized the KMeans clustering algorithm [9] to categorize questions according to their similarity. By determining an optimal number of clusters, we identified five primary types of questions.
(4) **Question Type Description:** For each group, we selected representative questions and created a textual description that summarizes the main theme of each question type.

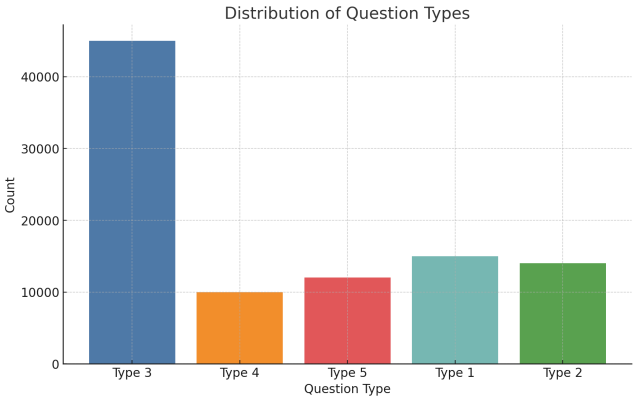


Figure 5: Statistics of question types.

Employing the clustering methodology outlined, we identified five primary question types associated with agricultural pest issues, with their distribution depicted in Figure 5. Each type is summarized below:

Type 1: Pest Identification. This category includes questions focused on identifying specific pest types, such as *"What type of pest is this?"* or *"Which species does this pest belong to?"* These questions aim to assist users in accurately recognizing observed pests to inform subsequent control actions.

Type 2: Pest Feeding and Damage. Questions in this category revolve around the pest's feeding habits and target plants, with examples like *"What does this pest primarily feed on?"* and *"Which crops are threatened by this pest?"* These questions help users understand the pest's feeding behavior and potential threat to crops.

Type 3: Pest Control Methods. This category includes questions recommending specific control methods, such as *"How can this pest be effectively controlled?"* and *"What is the recommended pesticide for this pest?"* These questions provide actionable pest control strategies for farmers and researchers.

Type 4: Optimal Timing for Pest Control. Questions here primarily inquire about the ideal timing for applying pesticides or other control measures, for instance, *"What is the best time to spray insecticides?"* Timely application can improve control efficacy and reduce pesticide usage.

Type 5: Pest Monitoring Techniques. This type includes questions focused on tools and methods for monitoring pest populations and activities, such as *"What tools should be set up to monitor pest presence?"* and *"How can pest activity be effectively tracked?"* These questions aid users in preventing pest outbreaks through effective monitoring.

5 Experiments and Evaluation

5.1 Experiment Settings

Dataset: The entire IP-VQA dataset comprises 18,981 images, and the models utilized in this article operate in a zero-shot manner, eliminating the need for training. These models employ the default parameters available on Hugging Face [25] or GitHub. This study uses BERTScore as the evaluation metric for the model [28]. All experiments were carried out on an NVIDIA GeForce RTX 4090 GPU, which is equipped with 24GB of video memory.

5.2 Models

5.2.1 BLIP2. The version of BLIP2 used is Salesforce/blip2-opt-6.7b-coco, and it operates with the default parameters. The only input provided is a prompt that directs the model to generate an answer based on the input image and the specified question, while also limiting the length of the generated response. To ensure consistency, all subsequent models utilize the same prompt.

5.2.2 LLaVA. The LLaVA model is an open-source multimodal AI system that integrates a large language model with visual capabilities. It not only comprehends textual information but also analyzes images to generate related textual descriptions. The version of LLaVA utilized is llava-hf/llava-v1.6-vicuna-13b-hf, and it is loaded in 4 bits to minimize both memory usage and computational demands.

5.2.3 GPT4o. GPT4o is an enhanced version of the advanced multimodal language model built upon the capabilities of GPT-4. It utilizes a unified Transformer architecture that processes and comprehends both visual and textual inputs, facilitating a profound integration of visual and linguistic information. Through pretraining on extensive multimodal datasets, GPT4o possesses a more comprehensive knowledge base and enhanced reasoning abilities.

5.3 Evaluation

In this paper, we employ BERTScore as a metric to assess the similarity between the generated text and reference answers. BERTScore is a text similarity evaluation method derived from the BERT (Bidirectional Encoder Representations from Transformers) model, which captures semantic information to measure the alignment between the generated content and reference answers. Unlike conventional metrics like BLEU and ROUGE, which focus primarily on exact n-gram matching, BERTScore utilizes embeddings from pre-trained language models to evaluate the semantic proximity of the generated text to the reference answers.

In practice, BERTScore calculates precision, recall, and F1 scores by analyzing the similarity of embeddings between the generated and reference texts, thereby capturing semantic accuracy, coverage, and overall alignment. This metric is especially effective for natural language generation (NLG) tasks where semantic quality is crucial, rather than just lexical overlap. Consequently, BERTScore offers a more semantically meaningful measure of similarity, providing a solid foundation for quantifying the quality of generated content in this context. The comparison of BERTScore results for answers generated by different models is presented in Table 1.

Table 1: BERTScore of Different Methods.

Method	Average Precision	Average Recall	Average F1-score
ViLT	0.4374	0.5539	0.4876
BLIP2	0.6158	0.6705	0.6405
LLaVA	0.7980	0.7438	0.7687
GPT-4o	**0.8172**	**0.7845**	**0.7903**

Table 2: Comparison of Model Outputs Using BERTScore Samples.

Model	Summary	BERTScore
True Answer	Farmers can implement integrated pest management (IPM) strategies to minimize the impact of the larva.	N/A
BLIP2	The larva is natural and not harmful to plants or crops.	**0.45**
LLaVA	Use strategies like IPM and natural predators to reduce larva impact.	**0.85**
GPT-4o	Use IPM, including crop rotation and natural predators, to minimize larva impact.	**0.92**

The analysis of content and BERTScore indicates that BLIP2 achieves a low similarity score due to its content being misaligned with the true answer's theme, as it fails to mention any strategies for mitigating the impact of the larva. In contrast, LLaVA obtains a high similarity score by referencing various strategies, including integrated pest management (IPM), to address the larval impact, providing relatively detailed information. GPT-4o records the highest similarity score by explicitly stating that farmers can implement integrated pest management strategies to lessen the larval impact, closely aligning with the true answer. The BERTScore results for the different methods tested across the entire dataset are presented in Table 2.

6 Conclusion and Future Work

In this study, we developed and released a novel Insect Pest Visual Question Answering (IP-VQA) dataset specifically tailored for the precision agricultural pest control sector. By providing a structured collection of images, questions, and corresponding answers, we aim to fill the gap in VQA datasets focused on agricultural applications, thereby establishing a comprehensive benchmark platform for researchers. We implemented various benchmark models to evaluate and compare different approaches on this dataset. The experimental results highlight that domain-specific datasets can significantly improve models' recognition and reasoning capabilities, particularly in complex agricultural scenarios.

Looking ahead, we plan to further expand the dataset by including a wider variety of scenes and question types to support more extensive agricultural applications. We anticipate that this dataset will propel the field of agricultural VQA forward, encouraging more researchers to explore and develop more accurate and intelligent models to aid decision-making and management in agriculture.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086. doi:10.1109/CVPR.2018.00636

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433. doi:10.1109/ICCV.2015.279

[3] Jun Chen, Dongxiao Zhang, and Zhongxian Lin. 2020. A Survey of Deep Learning Applications in Agriculture. *Journal of Physics: Conference Series* 1684, 1 (2020), 012098. doi:10.1088/1742-6596/1684/1/012098

[4] D. C. Church, S. M. Khan, and CABI. 2018. Invasive Species Compendium. Accessed: 2024-04-27.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. doi:10.48550/arXiv.2010.11929 Accessed: 2024-04-27.

[6] Konstantinos P. Ferentinos. 2018. Deep Learning Models for Plant Disease Detection and Diagnosis. *Computers and Electronics in Agriculture* 145 (2018), 311–318. doi:10.1016/j.compag.2018.01.009

[7] Peng Gao, Zhengkai Jiang, Hongjie You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic Fusion with Intra- and Inter-Modality Attention Flow for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6639–6648. doi:10.1109/CVPR.2019.00680

[8] Yash Goyal, Tejas Khot, Dustin Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913. doi:10.1109/CVPR.2017.729

[9] J. A. Hartigan and M. A. Wong. 1979. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108. doi:10.2307/2346830

[10] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In Defense of Grid Features for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10267–10276. doi:10.1109/CVPR42600.2020.01028

[11] Andreas Kamilaris and Francesc X. Prenafeta-Boldú. 2018. Deep Learning in Agriculture: A Survey. *Computers and Electronics in Agriculture* 147 (2018), 70–90. doi:10.1016/j.compag.2018.02.016

[12] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems*. 1564–1574.

[13] Jonah J. Lau, Swagata Gayen, and Aude Oliva. 2018. A Dataset of Pairwise Relative Attributes in Medical Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 209–217. doi:10.1109/CVPRW.2018.00036

[14] Liunian Harold Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Relation-Aware Graph Attention Network for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10313–10322. doi:10.1109/CVPR42600.2020.01029

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. Springer, 740–755. doi:10.1007/978-3-319-10602-1_48

[16] Junnan Liu, Kaiming Yao, Zexuan Li, Chenxin Wang, Shang Li, Xingchao Chen, Yuxin Xu, Chao Xia, Ruijie He, Erkang Xie, et al. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 Accessed: 2024-04-27.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. doi:10.1109/ICCV46641.2021.01123 Accessed: 2024-04-27.

[18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*. 13–23.

[19] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. 2018. Real-Time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2229–2235. doi:10.1109/ICRA.2018.8460668

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. doi:10.48550/arXiv.1910.10683 Accessed: 2024-04-27.

[21] Srdjan Sladojevic, Marko Arsenovic, Adam Anderla, Dubravko Culibrk, and Darko Stefanovic. 2016. Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification. *Computational Intelligence and Neuroscience* 2016 (2016), 1–11. doi:10.1155/2016/3289801

[22] Karen Spärck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28, 1 (1972), 11–21. doi:10.1108/EUM0000000000700289

[23] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4223–4232. doi:10.1109/CVPR.2018.00443

[24] Jordan Ubbens and Ian Stavness. 2017. Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. *Frontiers in Plant Science* 8 (2017), 1190. doi:10.3389/fpls.2017.01190

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and Mariama Drame. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Neural Information Processing Systems (NeurIPS 2020)*. doi:10.48550/arXiv.1910.03771

[26] Chuan Xie, Chengcai Yang, Yini He, Kaiguang Zhao, and Zhaobo Wang. 2020. A Multi-Modal Fusion Network for Plant Disease Recognition. *Computers and Electronics in Agriculture* 172 (2020), 105306. doi:10.1016/j.compag.2020.105306

[27] Zhou Yu, Jun Yu, Yueming Cui, Dacheng Tao, and Qi Tian. 2019. Deep Modular Co-Attention Networks for Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6281–6290. doi:10.1109/CVPR.2019.00644

[28] Tianyi Zhang, Varsha Kishan, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 8968–8978. doi:10.18653/v1/D19-1371

[29] Xiaoling Zhang, Liangyun Liu, Shengkui Wang, and Qiang Sun. 2019. Deep Learning-Based Crop Classification Using Multi-Temporal Sentinel-2 Data. *Sensors* 19, 20 (2019), 4377. doi:10.3390/s19204377