MLLM-ITV: A Multimodal Large Language Model for Image-to-Text Generation Based on Vertical Sector

Maolin Zhang[1], Xianyong Li[1,3†], Yajun Du[1], Songlin Chen[1], Xiaoliang Chen[1], Dong Huang[1], Shumin Wang[2]

[1]School of Computer and Software Engineering, Xihua University, Chengdu 610039, P.R., China
[2]China National Institute of Standardization, Beijing 100191, P.R., China
[3]Yibin Weite Ruian Technology Co., LTD, Yibin 644600, P.R., China

Keywords: Large language models; Multimodal large language models; Querying transformer recognition; Low-rank adaptation; Pest and disease recognition

Citation: Zhang M.L., Li X.Y., Du Y.J., et al.: MLLM-ITV: A Multimodal Large Language Model for Image-to-Text Generation Based on Vertical Sector. Data Intelligence, Vol. XX, Art. No.: 2025??XX, pp. 1-33, 2025. DOI: https://doi.org/10.3724/2096-7004.di.xxx

ABSTRACT

Integrating visual and textual data enables large language models (LLMs) to understand complex information better, broadening their applicability in real-world tasks. While multimodal large language models have advanced through large-scale image-text pre-training, few are optimized for Chinese forestry image comprehension and dialogue. To address this gap, we propose the multimodal large language model for image-to-text generation based on a vertical sector (MLLM-ITV), an efficient and adaptable multimodal large language model designed for the forestry sector. It is trained on a carefully curated Chinese forestry biology image-caption dataset and instruction-following data to support open-ended dialogues. A Querying Transformer (Q-Former) module connects a pre-trained vision encoder to the ChatGLM-6B (General Language Model for Chat, 6 Billion Parameters) language model, aligning visual features with domain-specific vocabulary. Fine-tuning based on Low-Rank Adaptation (LoRA) further adapts the model to forestry tasks. The resulting model performs strongly in species identification, image understanding, and visual dialogue, outperforming five state-of-the-art Chinese multimodal LLMs.

1. INTRODUCTION

Large language models (LLMs) [1-6] have exhibited remarkable performance across diverse domains, garnering substantial interest from both the academic and industrial spheres. While language models have

† Corresponding author: Xianyong Li (E-mail: lixy@mail.xhu.edu.cn).

demonstrated commendable performance, it is evident that possessing solely textual "comprehension" often inadequately communicates information with vividness. The integration of information from divergent modalities, notably textual and visual elements, coupled with the inclusion of perceptual acumen linked to "visual" comprehension, facilitates the attainment of a more comprehensive grasp of phenomena and amplifies our expressive proficiencies. Consequently, this expansion broadens the horizons of language model applications. Multimodal fusion provides us with richer ways to explore and solve complex problems. Artificial General Intelligence (AGI) [7] possesses efficient multimodal information processing capability. Multimodal artificial intelligence (AI) not only tackles tasks involving single data types but also establishes connections and fuses information across diverse data types, thus offering support for addressing intricate challenges. An open and transparent open-source base language model, ChatGLM [8], has caught our attention. It is an English and Chinese bidirectional dense model. It contains 130 billion parameters and is pre-trained using the general language model (GLM) algorithm. It unveils how the ChatGLM model with a 130 billion scale is successfully pre-trained. Hence, we contemplate the feasibility of transposing large language models into cross-modal domains to emulate human cognitive abilities.

Even with the notable performance showcased by the LLaMA [9] model through fine-tuning, its aptitude within the domain of the Chinese language remains constrained, stemming from its limited exposure to Chinese corpora during the pre-training stage. In contrast, the ChatGLM [8] model by Tsinghua University stands for its exceptional performance in the realm of the Chinese language. It is based on the general language model architecture, featuring 6.2 billion parameters. ChatGLM-6B employs methodologies analogous to those employed by ChatGPT, tailored to optimize performance within Chinese question-answering and dialogic interactions. However, it is noteworthy that frequent observation pertains to the suboptimal performance of numerous expansive language models in scenarios necessitating a profound grasp of domain-specific intricacies. Our supposition posits that this phenomenon may trace its origins to the paucity of data originating from distinct specialized domains during the formative pre-training phase of the model. Within an industrial purview, universal expansive models undertake the role akin to "comprehensive compendiums", typified by exemplars like GPT-3 [10], PaLM [11], MOSS [12], and ERNIE Bot [13], characterized by their capacity to furnish comprehensive responses. Meanwhile, domain-specific expansive models assume the semblance of mono-disciplinary experts, refined within the confines of their specific niches.

The expansion of LLMs into visual-linguistic multimodality and their judicious utilization constitute a subject of considerable significance. OpenAI's release of the multimodal version of GPT-3 [10], known as DALL.E [14], showcases remarkable capabilities in generating images from text. This achievement effectively bridges visual and linguistic domains, underscoring the potency of pre-trained multimodal models. Addressing the efficient synchronization of models with multimodal images and Chinese linguistic directives has spurred a resurgence of focus within the community toward refining Chinese language-enriched foundational visual models [15-16]. Concurrently, copious parallel image-text datasets across diverse domains provide a prolific resource. Empirical findings substantiate the capacity of generative pre-training to adeptly harness this parallel data for self-supervised visual-language modeling. Evidential

validation of this phenomenon has been showcased through open-source initiatives such as multimodal GPT-4 [17] and LLaVA [18]. The endeavor to equip ChatGLM with image recognition capabilities poses a formidable challenge, considering that LLMs do not integrate images during their pre-training phase.

To enable the interaction between text and image modalities, we have employed a Querying Transformer (Q-Former) [19]. Q-Former, a lightweight transformer, utilizes a set of trainable query vectors to extract visual features from the image encoder, serving as an information bottleneck between the image and text encoders. It supplies the most pertinent visual information required for text generation by LLMs. Q-Former predominantly consists of learned queries for both modalities, Cross-Attention, Self-Attention, and Feed Forward components. Concretely, during the training phase, we learn visual-language representation by feeding vectors processed through the image and text encoders into Q-Former, which bridges these two distinct modalities. Although natural image-text multimodal pre-training has shown impressive efficacy in numerous downstream applications, its seamless adaptation to the forestry domain faces challenges due to fundamental domain distinctions. Furthermore, collecting annotated datasets for forestry-related diseases and pests typically necessitates substantial domain-specific knowledge and resources, thereby making implementing our model in the specialized field of forestry a feasible proposition. Given the inherent constraints associated with the modest scale of ChatGLM-6B, substantial opportunities exist for further refining and optimizing the model.

In the past, the LLaMA-Adapter V2 [20] model effectively broadened the scope of language models to encompass multimodality by implementing adapter-based strategies. However, the method often introduces inference latency and increases memory demands. To address these limitations, Low-Rank Adaptation (LoRA) [21] has been proposed. LoRA reduces parameter storage and VRAM usage without adding inference overhead. When the rank r is much smaller than the original model dimensions, it eliminates the need to store optimizer states, significantly lowering memory consumption. Instead of updating all parameters, it fine-tunes only low-rank matrices, avoiding unnecessary gradient computations and improving efficiency. By aligning the rank of its matrices with pre-trained weights, it retains performance close to the original during fine-tuning. In multimodal scenarios, it also reduces the dimensionality of image features, transforming high-dimensional data into compact representations. To mitigate potential information loss, we adopt multiple parallel LoRA modules, combining their outputs to preserve semantic richness.

The scale of large models does not inherently equate to general artificial intelligence. As a result, fine-tuning in vertical domains has become a critical research direction to enhance their practical utility. In this work, we focus on forestry pest and disease applications. Using ChatGLM as the backbone, we accelerate scenario-specific iteration by integrating artificial intelligence into all forest pest detection and prevention stages. We will fine-tune ChatGLM with Chinese-language, domain-specific forestry data, enabling the model to adapt effectively during pre-training. This process steers the base language model towards better aligning with forestry scenarios' linguistic and semantic nuances.

Amidst the backdrop of climate change, the gravity of forest pest and disease calamities in China is escalating. The forestry ecosystem is confronting unparalleled trials, prompting an imperative need to advance automated and intelligent pest and disease detection technologies. These progressions are pivotal in guiding the biological and scientific strategies for preventing and managing forestry pests and diseases, thereby safeguarding ecological integrity. Consequently, the proficient management of detrimental organisms within forestry has acquired paramount significance and immediate attention. Certain pernicious entities exhibit deceptive characteristics in their external features despite harboring substantial peril. The forest-related biological catastrophes these detrimental organisms induce are frequently likened to "latent forest conflagrations". This analogy is drawn due to their shared resemblance with natural disasters and the distinctive intricacies and protracted challenges involved in their management.

Therefore, large-scale AI models known as "agricultural brains" will play a pivotal role in the practical implementation of pest and disease management solutions in the vertical domain of forestry. Fine-tuning in vertical domains will inevitably become a prevailing trend.

This paper introduces a multimodal large language model for image-to-text generation based on a vertical sector (MLLM-ITV). The primary purpose of this model is to expand the unadulterated language instruction model, ChatGLM, into a multimodal architecture, thereby endowing LLMs with the capability for generative production in alignment with visual modalities. This model represents the inaugural and successful initiative to incorporate multimodal instruction extension within the purview of forestry biology. The objective is directed at the comprehensive training of a forestry bio multimodal conversational assistant through an end-to-end approach, streamlining its implementation across practical application contexts. The efficacy of domain-specific pre-training has been demonstrated to be pertinent for Forestry Bio Natural Language Processing (NLP) applications and Forestry Bio Visual Language (VL) tasks alike.

- A multimodal Chinese-language forestry pest and disease (FPD) dataset is constructed. The dataset comprises forestry pest and disease images and their textual descriptions.
- A multimodal large language model for image-to-text generation based on a vertical sector (MLLM-ITV) model is an extension of the ChatGLM model using LoRA fine-tuning techniques.
- Experiments show that MLLM-ITV outperforms five state-of-the-art models, including VisualGLM [22], Ziya-BLIP2-14B-Visual [23], MiniGPT [24], VisCPM [25] and Qwen [26] models.
- Using Q-Former, a successful transformation from linguistic unimodality to image-linguistic multimodality has been realized.

The rest of this paper is organized as follows. In Section 2, we present related work. Subsequently, we provide a detailed exposition of our primary contributions in Section 3. Finally, in Section 4, we compare our model with relevant models. Section 5 offers conclusions and outlines directions for future work.

2. RELATED WORK

2.1 Language Models

The pre-training frameworks can be classified into three categories: autoregressive, autoencoding, and encoder-decoder models. Contemporary pre-trained language models, built upon the Transformer architecture like the GPT series [4, 27-28], BERT, and others, employ autoregressive Transformer models to pre-train expansive language models on extensive textual corpora. This practice demonstrates their prowess in few-shot learning capacities [4, 28]. Large Language Models (LLMs) have shown remarkable progress through training on extensive text corpora, gradually finding utility across diverse domains. The emergence of LLMs has initiated a technological paradigm shift, and a lineup of open-source large models, including LLaMA [9], BLOOM [29], and ChatGLM [8], has substantially propelled the advancement of the Natural Language Processing (NLP) field. In contrast, ChatGLM is a bilingual conversational language model proficient in accommodating Chinese and English. Having undergone training involving around 1 trillion tokens in both languages, bolstered by techniques like supervised fine-tuning and self-feedback, the ChatGLM model, boasting 6.2 billion parameters, demonstrates the capacity to generate responses that closely align with human preferences.

Based on the ChatGLM model, we enhance LLMs with the capacity to capture image features through fine-tuning. This endeavor lays the foundation for creating an open-source multimodal model. Within this study, we integrate domain-specific knowledge of forestry diseases and pests into the ChatGLM model, thereby reorienting the foundational language model towards a dedicated corpus specific to the field of forestry.

2.2 Vision-Language Models

In light of the emergence of expansive language models, scholarly investigations have been fervently delving into the application of LLMs for addressing multimodal challenges [20, 30], thereby culminating in the conception of Multimodal Large Language Model (MLLM) [17, 19, 31-34]. Various methodologies have entailed the infusion of visual data into LLMs and have meticulously refined these models through instructional directives. This strategic augmentation has facilitated their adeptness in generating textual content from visual inputs and has been shown to improve the generalization of language models to unknown tasks. In recent times, an evident transition has transpired within the landscape of image-language research, wherein the focus has shifted from expansive language models to substantial vision-language models. The Generative Pre-trained Transformer 4 (GPT-4) [17] has impressively showcased its prowess by adeptly handling inputs originating from diverse modalities, including images and text, fulfilling a wide spectrum of tasks. This exceptional adeptness has acted as a catalyst, giving rise to a fresh surge of investigation that extends the scope from singular language instruction models towards the realm of multimodal instructional models. Analogous to the principles behind LLaMA-Adapter [20], this emerging paradigm empowers LLMs with the faculty of visual reasoning, culminating in the proposal of LLaMA-Adapter V2 [20]. Conversely, BLIP2 [35] capitalizes on integrating Q-Former to facilitate the mapping of acquired image representations onto the textual embedding domain of LLMs.

In the pursuit of cultivating a directive comprehension akin to that exhibited by GPT-4, endeavors such as MiniGPT-4 [24] and LLaVA [18] have surfaced, embracing the utilization of datasets focused on image-guided tracking to cultivate the capacities of image-guided tracking within LMMs. MiniGPT-4 [24] embarks upon a trajectory of pre-training, encompassing a corpus of 134 million image-text pairs, to establish a connection between the static visual encoder and the LLM. This connection is subsequently reinforced through fine-tuning the model using well-aligned image-text datasets. LLaVA [18], in a similar vein, leverages the pairings of image and text to serve as a conduit for achieving congruence between visual models and LMMs. Video-chat [30] facilitates further expansion of the realm of comprehension, which extends the boundaries of image encoders to empower expansive models with the competence to decode the visual constituents embedded within videos. Although these methodologies have showcased commendable aptitude in comprehending multiple modalities, they require the adjustment of billions of model parameters and the assiduous aggregation of substantial quantities of training data encompassing multiple modalities. This dataset is sourced from human annotations or outputs produced by the OpenAI API. Furthermore, these models are predominantly designed for generic domains and have yet to be fine-tuned for the specific context of forestry pest management, diminishing precision in their generated responses.

Our endeavor is directed towards endowing foundational LLMs with the ability to comprehend visual attributes. In this context, our model introduces an innovative LoRA fine-tuning strategy, encompassing the immobilization of parameters inherent to the initial pre-trained model. Additionally, augmentation is achieved by integrating an auxiliary matrix to replicate the comprehensive fine-tuning of model parameters. This strategic implementation curtails computational requirements and orchestrates a gradual infusion of image-based visual attributes into the pre-existing ChatGLM model, facilitated by low-rank adaptive. The outcome is a model that showcases robust generalization capacities. Moreover, existing antecedent models have yet to attain the desired level of adeptness within forestry biology. Our model will demonstrate a high level of competitiveness in the forestry domain compared to previous multimodal models.

2.3 Querying Transformer

Visual and linguistic modalities represent two fundamental channels through which human beings apprehend and comprehend their external environment. The central predicament confronting image-language models revolves around the harmonious amalgamation of data from these heterogeneous modalities into a feature space that expansive language models can effectively apprehend. At present, the adoption of the Transformer architecture has ascended as the predominant methodology in the realm of multimodal algorithms for achieving the harmonious integration of information derived from diverse modalities into a feature space intelligible to LLMs and streamlining the process of feature fusion, owing to its remarkable aptitude in this regard. A new visual-language representation learning framework, Align before Fuse (ALBEF) [36], has been introduced, integrating multimodal contrastive learning into the domain of multimodal models. ALBEF encompasses an image encoder, a text encoder, and a multimodal encoder. It presents a straightforward, end-to-end, and highly proficient framework for acquiring visual-language representation skills.

An enhanced iteration of ALBEF, Q-Former, has been introduced. Q-Former is a streamlined model consisting of two transformer sub-modules. In contrast to ALBEF, the most salient divergence within Q-Former lies in the integration of Learned Queries. These Queries actively interact with image attributes through Cross-Attention and textual attributes through Self-Attention. Derived from information in both modalities, these Queries yield feature outputs of query length, irrespective of the visual backbone's scale, thereby substantially diminishing computational complexity. The image transformer is predominantly dedicated to extracting visual features, and the Text Transformer encompasses the roles of text encoder and text decoder. Q-Former incorporates three distinct training tasks, namely Image-Text Contrastive Learning (ITC), Image-grounded Text Generation (ITG), and Image-Text Matching (ITM) [19]. These tasks collaboratively enable the extraction and fusion of features.

BLIP-2 [35] efficiently utilizes frozen image encoders and frozen LLMs to achieve various visual-language tasks, yielding improved performance while minimizing computational overhead. Drawing from the Q-Former framework advanced in the BLIP-2 model, InstructBLIP [37] presents an instruction-aware visual feature extraction method. Q-Former serves as a lightweight bridge between the frozen vision encoder and the language model. Specifically, it employs a set of learned queries that interact with the visual features extracted by the image encoder through cross-attention. These queries generate compact visual embeddings projected into a token-level representation space. The resulting query outputs match the dimensional and semantic structure of language model input embeddings, allowing seamless integration into the frozen LLM without retraining its backbone. This alignment enables the language model to interpret visual semantics like textual tokens. The query outputs are prepended or interleaved with textual inputs and passed to ChatGLM's input layer during implementation. This mechanism ensures LLM can condition its generation on visual and linguistic contexts in a unified token space. Ultimately, the model performs better than GPT-4, attaining cutting-edge outcomes across diverse tasks. Recent research has also highlighted the potential of Q-Former in integrating audio-visual signals, denoted as Audio Q-Former [38]. Our model employs a strategy that involves encoding and decoding images and text, followed by their fusion within the Q-Former framework. Q-Former excels in extracting visual representations that are most informative for textual content. Subsequently, the combined data is fed into a language model, ensuring the model's adaptability with dynamic adjustments and enhanced learning capabilities. This approach aims to refine the training process for improved alignment.

2.4 Low-Rank Adaptation

For large models, full fine-tuning of all parameters of the retrained model becomes infeasible, and fine-tuning large models and large model deployments is also infeasible due to the massive number of parameters. The approach commonly used to adapt pre-trained models to multiple downstream tasks is fine-tuning, but fine-tuning involves updating all parameters with the trained model. The Low-rank structure is widespread in machine learning, and many machine learning algorithms have some inherent low-rank structure [39-41]. Moreover, it is well known that for many deep learning tasks, especially those with heavily overparameterized neural networks, the learned neural networks will have low-rank properties after training [42]. Some previous work has even explicitly imposed low-rank constraints when

training the original neural networks [43-47]. However, it was found that none of these works considered low-rank updates to the frozen model to adapt to downstream tasks. Therefore, LoRA [21] was proposed to indirectly train some dense layers in the neural network by optimizing the rank-decomposition matrix of the dense layers as they change during adaptation while keeping the pre-trained weights constant.

As such, we will adopt LoRA's adaptive strategy to enhance the effectiveness of the LLM fine-tuning for downstream tasks. It maintains high-quality model performance without introducing inference latency or reducing input sequence length. It can maintain high-quality model performance without introducing inference delays or reducing the length of input sequences. It also demonstrates its excellent capability in service deployment scenarios, achieving the goal of fast task switching by sharing most of the model parameters. The framework successfully optimizes the performance by approximating global training, thus effectively reducing the waste of resources. In achieving the best overall performance, LoRA cleverly employs attention-related matrices, including W^Q and W^V, while taking W^K into account. Experimental evidence from a related study [21] shows that the utility of the top singular vector direction is high when the matrix rank is set to 8, because the other directions usually contain most of the accumulated random noise during training. Therefore, during the training of the LoRA model, the rank is set to 8. The study shows that the neural network performs well when the underlying model concept has a low-rank structure. The most significant advantage of LoRA is that it is faster and uses less memory. Therefore, it can be run on consumer-grade hardware.

3. THE MULTIMODAL LARGE LANGUAGE MODEL FOR IMAGE-TO-TEXT GENERATION BASED ON A VERTICAL SECTOR

3.1 Descriptions of Pertinent Symbols and Parameters

Table 1 summarizes the pertinent symbols utilized in the multimodal large language model for image-to-text generation based on a vertical sector (MLLM-ITV), accompanied by their respective elucidations. In order to maintain a state of stability throughout the model training procedure, the vector dimensions resulting from the residual connections after the input of image-text pairs are consistently preserved. For the initially trained images, they can be systematically transcribed into a matrix array comprising n matrices, each delineated by 197 feature column vectors denoted as $A_a^{j'} = [a'_0, a'_1, a'_2, a'_3, \ldots, a'_i, \ldots, a'_{196}]$. Herein, $A_a^{j'}$ signifies the amalgamation of all feature vectors associated with the j-th training image. After applying residual connections, the initial training textual content can be projected to generate a matrix table mirroring the images. Each attribute delineated within the images possesses equivalent dimensions. In this paradigm, each matrix is also characterized by 197 text-based feature vectors, denoted as the augmented matrix $B_a^j = [b_0, b_1, b_2, b_3, \ldots, b_{196}]$, representing the text-based feature vectors corresponding to the j-th image. Subsequently, an alignment procedure is executed between the image-text vector pairs. Positive samples are maximized to achieve optimal similarity alignment, whereas negative samples undergo supplementary cross-attention mechanisms for fine-grained realignment. This supplementary step amplifies the alignment efficacy, enabling the model to furnish more precise responses throughout the text-generation process.

Table 1. Symbols.

Symbols	Symbol interpretation
A^j	Vector matrix of image features
B^j	Vector matrix of text features
A_a^j	An augmented matrix formed by splicing the overall vector of the picture
$A_a^{j'}$	Augmented matrix embedding location features
B_a^j	Text augmentation matrix formed by splicing text topics
A_{IS}	Fine-tuned image feature skill matrix
B_{TS}	Coded text feature skill matrix
H^j	Input description encoded feature matrix
H_a^j	An augmented matrix formed by splicing text features in the text encoding stage
$H_a^{j'}$	Augmented matrix formed by self-attention

The MLLM-ITV model embarks on the LoRA fine-tuning of the foundational VisualGLM model. The training process employs image-text data about forestry pest occurrences. Throughout this process, pertinent parameters within the image encoder and language model remain fixed, while LoRA-associated parameters in both components undergo refinement. Simultaneously, relevant parameters of Q-Former also undergo tuning. As a result, the refined multimodal large language model tailored for forestry applications is equipped to tackle issues surrounding forestry pest infestations. This development contributes to advancing research in the specialized forestry domain, facilitated by integrating multimodal large language models.

3.2 The Framework of the Proposed MLLM-ITV Model

The MLLM-ITV model comprises five principal constituents encompassing LoRA fine-tuning training, image encoding, text encoding, fine-grained hard sample alignment (fine-grained HSA), and answer testing, as shown in Figure 1. Given the suboptimal outcomes achieved with alternative fine-tuning methodologies, this study exclusively adopts LoRA for fine-tuning training. During the LoRA fine-tuning training, the critical action involves freezing parameters within the image encoder and the extensive language model. Q-Former connects the image encoder and the frozen large language model. As a result, the fine-tuned parameters encompass LoRA-related parameters in both the image encoder and the large language model, as well as the pertinent parameters within Q-Former. To align the model with the following instructions, we further train the model by prompting language-only high-quality dialogues. This training process culminates in acquiring and retaining multimodal proficiencies within the MLLM-ITV Model. Notably, this training approach substantially curtails the consumption of hardware resources throughout the training endeavor. During image encoding, the features of the image are subjected to representation learning, culminating in encoding image attributes into a feature vector. The primary

objective is to engage in comparative learning with the vector derived from the text encoding phase and to perform cross-attention fusion analysis with subsequent components.

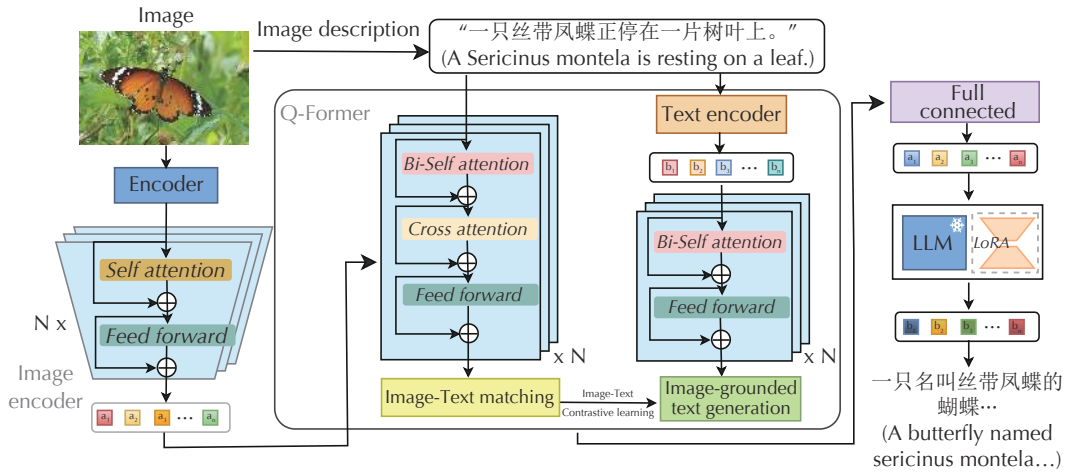

Figure 1. The architecture of the MLLM-ITV model.

In the text encoding phase, image text descriptions are encoded to generate a text vector. Subsequently, the vector dimensions undergo normalization through residual layers, aligning them with the dimensions of the image vector to facilitate convenient comparative learning. Within the Fine-grained HSA phase, an introduced cross-attention mechanism serves as the primary function, enabling the model to focus concurrently on the feature sequences of an alternative image while processing the input sequence. This endeavor is aimed at attaining a more meticulous alignment of image-text vectors. In the section dedicated to image description, the process encompasses the extraction of image vectors, followed by integrating cross-attention fusion with vectors derived from the text encoder. This process culminates in the task of text generation. During the answer testing phase, questions and images are fed into our well-trained MLLM-ITV model. Subsequently, the new model delivers exemplary responses grounded in the acquired skills and input image attributes. The following sections will give a detailed elucidation of the five constituents of the MLLM-ITV model.

3.3 LoRA Fine-Tuning Training

The LoRA fine-tuning training procedure can be delineated into two distinct stages, as shown in Figure 2. During the initial stage, the primary emphasis lies on training input images. The images undergo encoding, extracting features, and converting from multi-dimensional representations into one-dimensional column vectors. This extraction sequence proceeds from left to right and top to bottom, forming column vectors. Each column vector is denoted as a_i, where i = 1, 2, 3, …, 196. The features of the j-th image are encapsulated within the feature matrix $A^j = [a_1, a_2, a_3, \ldots, a_i, \ldots, a_{196}]$ (j = 1, 2, 3, …, n). Following the transformation of image vectors' dimensions, they are concatenated with the comprehensive information

vector a_0, which maintains an equivalent dimensionality as that of the image feature vectors. This amalgamation yields an argmented matrix $A_a^j = [a_0, a_1, a_2, a_3, \ldots, a_i, \ldots, a_{196}]$. After this, positional feature embedding is implemented on the amalgamated feature vectors, and then a residual connection layer is introduced. This iterative procedure culminates in the generation of a novel augmented feature matrix, denoted by $A_a^{j'} = [a_0', a_1', a_2', a_3', \ldots, a_i', \ldots, a_{196}']$. Subsequently, it is subject to a normalization process.

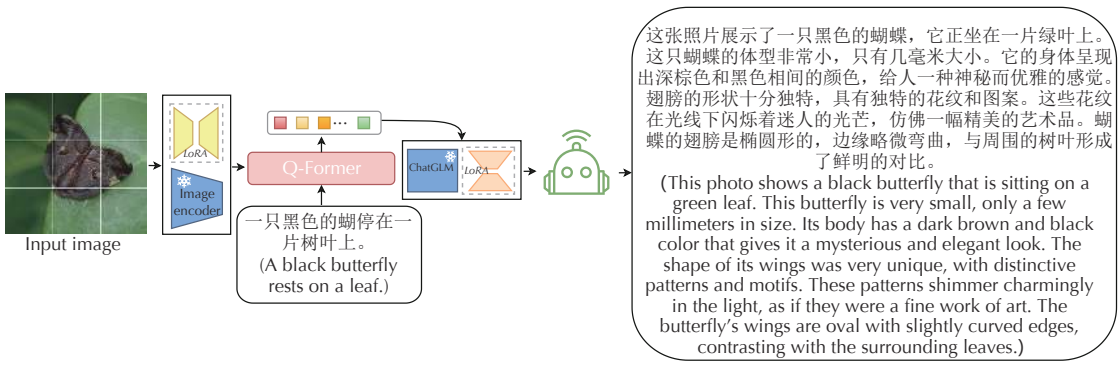

Figure 2. The flow chart of the LoRA fine-tuning training.

During the training process, the parameters in the graph encoder are frozen and fine-tuned in conjunction with LoRA. This process involves subjecting the image features to processing by both the image encoder model and LoRA, resulting in the establishment of the image feature skills matrix $A_{IS} = [A_a^{1'}, A_a^{2'}, A_a^{3'}, \ldots, A_a^{j'}, \ldots, A_a^{196'}]$. After this step, a phase of comparative learning between Q-Former and the image-text pairs ensued, aligning them with the text vectors. In this context, the text vector matrix B^j encapsulates the textual description of the features for the j-th image, formulated as $B^j = [b_1, b_2, b_3, \ldots, b_{196}]$. It is followed by the concatenation with the topic vector b_0, ensuring its dimensionality matches the text vectors. Consequently, this process forms the augmented text matrix $B_a^j = [b_0, b_1, b_2, b_3, \ldots, b_{196}]$. In Q-Former, text encoding leads to the formation of text features, resulting in a text feature skills matrix $B_{TS} = [B_a^{1'}, B_a^{2'}, B_a^{3'}, \ldots, B_a^{j'}, \ldots, B_a^{n'}]$. The skills matrices align within Q-Former, employing cosine similarity for the alignment analysis. Subsequently, the aligned vector features are fed into the ChatGLM language model for training. Throughout this training process, the parameters of the ChatGLM language model remain fixed while fine-tuning is carried out in conjunction with LoRA. The mathematical procedure for the LoRA fine-tuning is delineated as follows.

$$W = W_{pm} + tW_{LoRA} = W_{pm} + tE_{LoRA\text{-}zeros} \times F_{LoRA\text{-}gaussian} \tag{1}$$

where t is a random variable with an absolute value not exceeding 1, and W, W_{pm}, and W_{LoRA} represent the weight matrices of the trained model, the frozen model, and the LoRA fine-tuning process, respectively. During the model training, the $F_{LoRA\text{-}gaussian}$ matrix is initialized using a normal distribution, while the $E_{LoRA\text{-}zeros}$ matrix is initialized with zeros. It ensures that the bypass of the frozen model remains a zero matrix at the beginning of the training process.

In the process of fine-tuning LoRA, when LoRA is applied to the mapping matrices W of attention's Query and Value, the fine-tuning effect is further enhanced. The calculation process for the weights of the Query and Value mapping matrices in attention is illustrated below.

$$W^Q = W^Q_{pm} + tW^Q_{LoRA} \tag{2}$$

$$W^V = W^V_{pm} + TW^V_{LoRA} \tag{3}$$

When fine-tuning LoRA and passing through the multi-head self-attention layer for a training image-text data Y, the corresponding mapping produces calculation formulas for the Query matrix Q, Key matrix K, and Value matrix V. These formulas are shown below.

$$Query : Y \times W^Q = Y \times W^Q_{pm} + tY \times W^Q_{LoRA} \tag{4}$$

$$Key : Y \times W^k = Y \times W^k_{pm} + tY \times W^k_{LoRA} \tag{5}$$

$$Value : Y \times W^V = Y \times W^V_{pm} + tY \times W^V_{LoRA} \tag{6}$$

When Softmax is utilized, the computational inference for matrices Q and K with LoRA layers can be expressed as follows.

$$\begin{aligned} Softmax(Q,\ K^T) = softmax(&YW^Q_{pm}(W^k_{pm})^T Y^T + tYW^Q_{pm}(W^k_{LoRA})^T Y^T \\ &+ tYW^Q_{LoRA}(W^k_{pm})^T Y^T + t^2 YW^Q_{LoRA}(W^k_{LoRA})^T y^T) \end{aligned} \tag{7}$$

The final attention calculation can be represented as follows.

$$Head = softmax(Q,\ K^T)YW^V_{pm} + T \times softmax(Q,\ K^T)YW^V_{LoRA} \tag{8}$$

After undergoing fine-tuning with LoRA training, a final image-text understanding skill matrix is formed as $C = [A_{IS},\ B^T_{TS}]$. The corresponding understanding skills are stored in the newly trained model, achieving the LoRA fine-tuning process.

3.4 Image Encoding

The image encoding process commences with the initial segmentation of input images into smaller blocks, each signifying a distinct feature of the image. All images are uniformly divided into 14 × 14 blocks. The feature of each diminutive block is then embedded to formulate a comprehensive feature vector. Subsequently, a residual connection mapping is executed through a self-attention layer. The process further engages in residual connection mapping through a feedforward network, creating a feature column vector, denoted as d'_i, where i = 1, 2, 3, ..., 196. In this context, d'_0 represents the amalgamated vector encompassing the entirety of the image's information along with positional data. These feature vectors from the j-th image eventually amalgamate to form a freshly augmented feature matrix $D^{j'}_a = [d'_0,\ d'_1,\ d'_2,\ d'_3,\ \ldots,\ d'_i,\ \ldots,\ d'_{196}]$. The vectors contained within this feature matrix undergo alignment with subsequent-stage text encoding vectors. This alignment process is implemented to prevent the occurrence of erroneous correlated alignments. To achieve this, the feature

vectors of the images undergo a fusion analysis using a cross-attention mechanism in conjunction with the text. The primary objective is to validate further whether the image and text convey identical information. This approach aims to minimize alignment errors throughout the process. Additionally, these features offer an augmented representation of feature vectors for text generation within the model. The process of cross-attention fusion contributes to the generation of enhanced text descriptions. Consequently, this optimization ensures the best expression of the training effect of the model in text generation, yielding answers that more precisely align with the desired outcomes.

3.5 Text Encoding

Initially, the textual description of the image is input. During text encoding, the thematic content from the image description is assimilated and transformed into a vector representation mirroring the dimensionality of the image encoding. This vector is designated as h_0. The text gives rise to a vector representation, denoted by h_i (i = 1, 2, 3, ..., 196). The complete j-th image description is aggregated into a feature matrix of text descriptions, denoted as $H^j = [h_1, h_2, h_3, ..., h_i, ..., h_{196}]$. Ultimately, this feature matrix is merged with the thematic vector, resulting in a text augmented matrix $H_a^j = [h_0, h_1, h_2, h_3, \ldots, h_{196}]$. After encoding, a residual connection is established through a self-attention mechanism within the encoding section. A residual connection is created through a feedforward network involving the vector before the feedforward network layer. This procedure results in the generation of a novel augmented feature matrix for the text, denoted as $H_a^{j'} = [h_0', h_1', h_2', h_3', \ldots, h_j', \ldots, h_{196}']$. It promotes facilitative analysis in comparative learning. In comparative learning, cosine similarity analysis is employed as follows.

$$\cos\left\langle d_{i'}, h_{j'}\right\rangle = \frac{d_{i'} \cdot h_{j'}}{\left|d_{i'}\right| \cdot \left|h_{j'}\right|} \tag{9}$$

where $d_{i'}$ and $h_{j'}$ (i, j = 1, 2, 3, ..., 196) represent the i-th feature vector of an image and the j-th feature vector of the corresponding text, respectively. This calculation assesses the similarity between text and image features, where the closer the value of their similarity is to 1, the more they are similar. Nonetheless, misalignments with negative samples can arise in the context of similarity contrastive learning. Consequently, the model undertakes additional measures to address such challenges through fine-grained Hard Sample Alignment.

3.6 Fine-Grained Hard Sample Alignment

In the fine-grained Hard Sample Alignment process, we commence with the encoding of the j-th text to generate a matrix of text feature vectors, denoted as $H^j = [h_1, h_2, h_3, ..., h_i, ..., h_{196}]$, where h_i represents the i-th content feature vector of the text. Furthermore, the encoding process is separately applied to the thematic content of the text, resulting in a singular feature vector k_0. Ultimately, the feature vector corresponding to the text's theme is concatenated with the matrix of text feature vectors to create a new augmented feature matrix, denoted by $H_j' = [k_0, h_1, h_2, h_3, \ldots, h_i, \ldots, h_{196}]$. This matrix undergoes additional cross-fusion and alignment with the previously inputted image feature vectors. The goal is to

achieve a finer-grained matching between text and images, ensuring that the maximal correlation between positive and negative samples is leveraged during the matching process. In the image description phase of the model, the augmented matrix formed by the feature vectors of the images, denoted by $D_a^{j'} = [d'_0, d'_1, d'_2, d'_3, \ldots, d'_i, \ldots, d'_{196}]$, is synchronized with the inputted text feature matrix and subsequently input into the language model. Based on the feature vectors of the images, the language model elaborates on the textual inference description. Algorithm 1 outlines the procedure of encoding images to serve as prompts for the model.

Algorithm 1. Image encoding and model prompt process.

Input: Image vector feature matrix $D_a^{j'}$ and corresponding text feature vector matrix $H_a^{j'}$

Output: Model's description of the picture

function: Image feature matrix $D_a^{j'}$ and text feature vector matrix $H_a^{j'}$ embedding alignment

initialize D_a^j, H_a^j

for d_i *in* D_a^j **do**

    for h_i *in* H_a^j **do**

        d_i: Count the first section on relevant features.

        h_i: Count corresponding features in the text.

    end

    D_a^j . append(d_i)

    H_a^j . append(h_i)

end

initialize $D_a^{j'}$, $H_a^{j'}$

repeat

    for d_i in D_a^j

    for h_i in H_a^j

    after residual computation to obtain d'_i, h'_i

    vector alignment using cosine similarity

until convergence

using fine-grained hard sample alignment

return $D_a^{j'}$ and $H_a^{j'}$ embedding aligned feature vectors.

end function

3.7 Answer Testing

Figure 3 presents a comprehensive overview of the answer testing phase. The pre-trained MLLM-ITV model is initially supplied with the test image during answer testing. The image undergoes encoding through the image encoder, resulting in the generation of image feature vectors. These feature vectors, produced in the answer testing phase, are denoted as j_k for k = 0, 1, 2, ..., 196. The individual features are then aggregated to construct the feature matrix for the image, represented as $J = [j_0, j_1, j_2, ..., j_k, ..., j_{196}]$. At the same time, inquiries to extract information from the image are made. These queries are encoded to form a question vector. Subsequently, the image feature matrix and the question vector are fed into the Q-Former. A fresh vector matrix is created by extracting image-text information vectors. The dimensions of the feature matrix are then adjusted via a fully connected layer. This adapted matrix is subsequently input into the language model, which proceeds to respond to the presented questions. Throughout this response process, the language model generates answers based on the feature information extracted from the input image, resulting in a textual representation. Algorithm 2 provides the pseudo-code outlining the pertinent process.

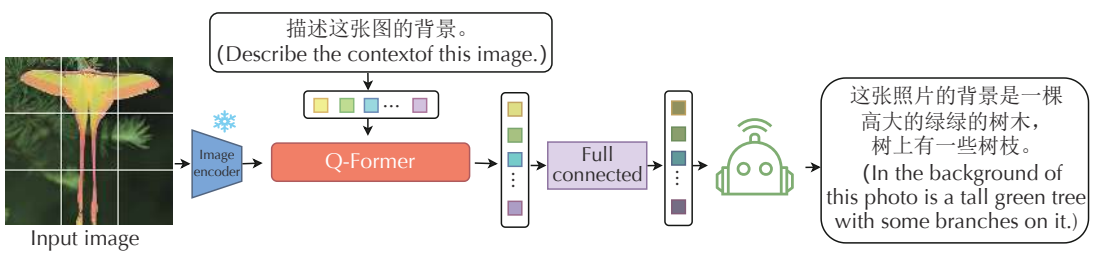

Figure 3. Answer testing formation.

Algorithm 2. Answer Testing.

Input: Image feature vector matrix J and problem vector matrix for the desired problem
Output: Modelling targeted responses to questions
function: Alignment of the image feature matrix J with the desired problem vector
initialize J
for $\forall j_k$ *in* J **do**
    for *problem vector in problem matrix* **do**
        j_k: Count Relevant features of responses to questions
        problem vector: Count computing key features in problem vectors
    end
    J. append(j_k)
    problem matrix.append (problem vector)
end
Alignment of features of computational questions with picture responses
Extract the required response feature vectors to pass to the language model
Language modeling for targeted responses to questions asked

4. EXPERIMENTS

4.1 Dataset

Due to the absence of multimodal forestry-related datasets for model training, we construct a multimodal Chinese-language forestry pest and disease (FPD) dataset[①] comprising forestry pest and disease images and their textual descriptions. To facilitate comprehensive learning from the images, 3, 4, or more relevant questions were generated for each image. The answers chiefly involve the image's presented content, the featured species, and the morphological attributes of said species. The FPD dataset encompasses approximately 4620 color forestry pest and disease images concerning 80 insect categories. This dataset includes information on pest categorization, temporal features, geographical distribution, damage caused, and pest control methods, as shown in Figure 4.

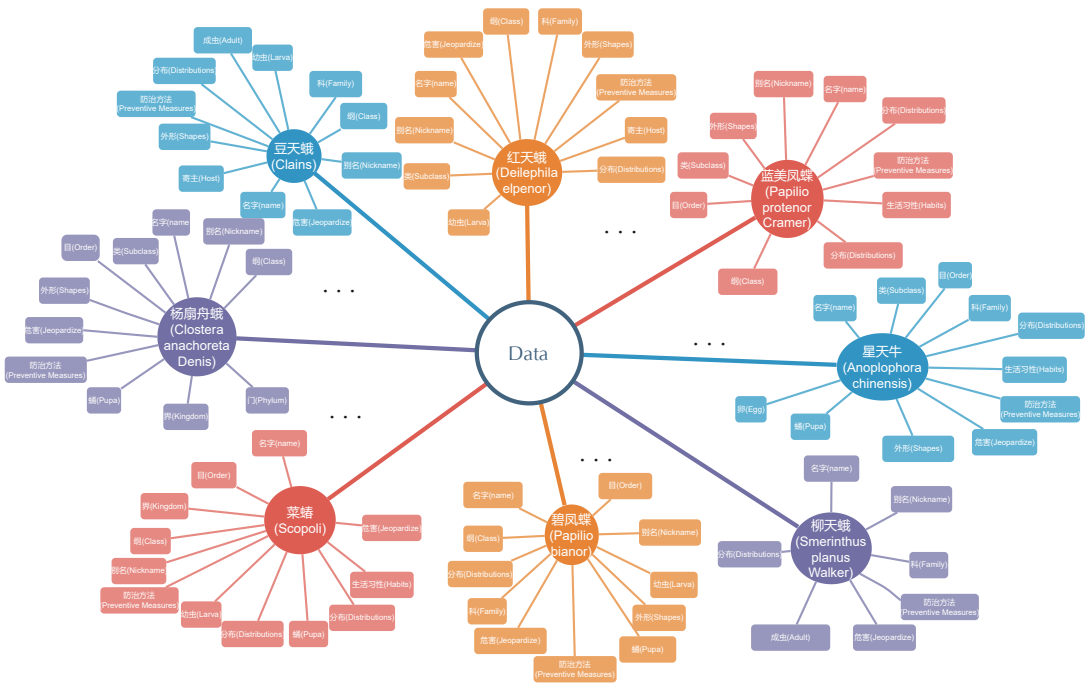

Figure 4. Overview of the FPD dataset.

To bolster the base model's conversational adeptness in forestry pests and diseases, we collect around 50,000 entries concerning various concepts within forestry pests and diseases. These entries are meticulously selected to constitute a pre-training corpus. Additionally, we perform web scraping to extract content related to forestry pests and diseases from Baidu Baike, incorporating it as an extra reservoir for our pre-training corpus. It is worth noting that when training the model, all experiments will be conducted

① https://github.com/motuomumu/MLLM-For.

on three NVIDIA A100 GPUs. The AdamW optimizer trains the model. Its learning rate, batch size, and training epochs are 5e-5, 64, and 10, respectively. A linear learning rate warm-up was applied over the first 500 steps. The maximum sequence includes 512 tokens. Input images are resized to 196×196 pixels. Standard augmentation techniques such as random cropping, resizing, and horizontal flipping are used to improve generalization. The training duration spans approximately 1 month and 5 days, including hyperparameter tuning and repeated experiments. The dataset is distinguished into 80% for training, 10% for validation, and 10% for testing. All experiments are repeated with three different random seeds, and average performance is reported to ensure robustness.

4.2 Comparison Models

We evaluate five leading open-source Chinese MLLM models, including VisualGLM-6B, Ziya-BLIP2-14B-Visual, MiniGPT, VisCPM, and Qwen-VL, as follows:

- **VisualGLM-6B** [22] is an open-source multimodal conversational language model that supports images, Chinese, and English. The language model is based on ChatGLM-6B, with a total of 6.2 billion parameters. The visual component is established by training the BLIP2-Qformer, bridging the gap between the visual and language models, resulting in a combined model with a total of 7.8 billion parameters. Pre-training is conducted on 30 million high-quality Chinese text-image pairs from the CogView dataset and 300 million carefully curated English text-image pairs.
- **Ziya-BLIP2-14B-Visual** [23] is crafted by the "Fengshen List" research team through the process of training on a subset of meticulously curated high-quality data sourced from open-access datasets. It is fashioned using an extensive corpus of around 20 million high-fidelity data instances designated for training.
- **MiniGPT** [24] originates from the King Abdullah University of Science and Technology. The team integrates a static visual encoder (Q-Former & ViT) with an immobile, extensive-scale text generation model, resulting in the development of MiniGPT-4.
- **VisCPM** [25] is a family of open-source large multimodal models that support multimodal conversational capabilities (VisCPM-Chat model) and text-to-image generation capabilities (VisCPM-Paint model) in both Chinese and English, achieving the state-of-the-art performance among Chinese open-source multimodal models. VisCPM is trained based on the large language model CPM-Bee with 10B parameters, fusing visual encoders including Muffin and Diffusion-UNet to support visual inputs and outputs. Thanks to the good bilingual capability of CPM-Bee, VisCPM can be pre-trained with English multimodal data only and well generalized to achieve promising Chinese multimodal capabilities.
- **Qwen-VL** [26] is a large-scale Vision Language Model (LVLM) developed by Alibaba Cloud. Qwen-VL can take images, text, and detection boxes as input and produce text and detection boxes as output. The distinguishing features of the Qwen-VL series models offer advanced capabilities in fine-grained recognition and understanding with robust performance.

4.3 Experimental Results

Extensive experiments are conducted on our benchmark. All comparative experiments are performed on an NVIDIA A100 GPU. We set the parameters for beam search to 1, set the temperature to 0.8, and set the top p to 0.4. For experimental testing, including the cognitive task, question and answer, count task, stages cognitive recognition, and common sense reasoning, we select images from our FPD dataset and non-dataset images for experiments.

4.3.1 Cognitive Task

We assess to ascertain whether MLLMs can engage in advanced logical reasoning following the perception of visual stimuli. For MLLMs to deduce accurate responses, they must adhere to instructions, comprehend the contents of visual stimuli, align text with images, and draw upon knowledge from LLM. This task presents a more formidable challenge compared to a singular perceptual assignment. The application of MLLMs to address specialized issues necessitates a profound exploration of this domain. It encompasses identifying fundamental problems and determining specific data essential for resolving them, a "modal completion" process. Prompt identification of pest and disease types and categories in forestry pests and diseases is paramount. Early detection and warning, facilitated through vigilant monitoring and swift identification, enable the timely revelation of pests and diseases, the issuance of alerts, and the implementation of control measures to contain their propagation. Swift and accurate identifications pinpoint the pest or disease types, facilitating targeted control interventions.

Because many AI image recognition datasets are cultivated within Petri dishes, the model settings are only conducive to real-world production requirements. As such, we randomly select 100 images from AI image recognition datasets and research libraries on agricultural pests and diseases from the IDADP dataset[②]. MLLMs are tasked with discerning the species depicted in each image. AI insect recognition entails a conventional object detection task, primarily discerning the positions and classifications of seven types of insects within monochromatic containers.

Perception is a fundamental capability of MLLMs, and its absence may lead to perceptual illusions. In this experiment, a single-turn dialogue format was employed, wherein the model was directly presented with an image and the prompt: "What species is shown in the picture?". Figure 5 depicts the outcomes of the baselines and the MLLM-ITV model. Within this task, none of the three comparative models could definitively ascertain the species depicted in the image. Notably, VisualGLM-6B demonstrates a laudable proficiency in scene depiction. It precisely delineates species patterns, spots, and the ambient milieu. Particularly striking are its portrayals of the image's backdrop, the habitat of the depicted animal, and finer features such as antennae, stripes, and spots on the animal. However, VisualGLM-6B occasionally refrains from answering queries or provides proverbial responses. The Ziya-BLIP2-14B-Visual model could discern the image's content, but not specify the exact species' name. Similar to VisualGLM-6B, MiniGPT could render an account of the overall environment and provide certain details about the insects. Nevertheless,

② https://www.heywhale.com/mw/dataset/63e50cfea2c1716e14fb9db6.

MiniGPT exhibits limited proficiency in adapting to Chinese; it responds in English to Chinese prompts, showing relatively superior performance when queries are in English. The MLLM-ITV model exhibits superior performance in this task. While it manages to identify some species, a discernible gap exists compared to seasoned human experts on an aggregate scale. Our model accommodates questions in Chinese and furnishes corresponding and accurate responses. Furthermore, upon further examination, we discover that when we persistently inquire: "What is presented in the picture?", the model capably generates corresponding descriptions for the provided image, encompassing particulars such as antenna length, the insect's developmental and growth stages, distribution regions, and more.

Figure 5. Experiment comparison chart of the cognitive task. In this figure, VisualGLM-6B, Ziya-BLIP2-14B-Visual, and Qwen-VL are shortened to VisualGLM, Ziya, and Qwen, respectively (The model abbreviations of the following figures are the same as those in this figure).

For the task of species recognition shown in Table 2, VisualGLM-6B, Ziya-BLIP2-14B-Visual, MiniGPT, VisCPM, and Qwen-VL exhibit minimal capability in identifying the species in the images. Our MLLM-ITV model outperforms these counterparts in this task. However, when it comes to identifying some species, there is still a discernible gap compared to experienced human experts.

Table 2. The Accuracy scores of evaluation models in the cognitive task.

Model	Acc
VisualGLM-6B	0.13
Ziya-BLIP2-14B-Visual	0
MiniGPT	0
Viscpm	0
Qwen-VL	0
MLLM-ITV	**0.21**

4.3.2 Question and Answer

The recommendations for preventing and controlling forestry pests and diseases should exhibit a scientific and rational foundation. They will furnish invaluable guidance for practical forestry production, mitigating the environmental repercussions of prevention and control measures, reducing costs, and bolstering economic efficiency. Furthermore, they will direct the refinement and advancement of prevention and control technologies, establishing a robust framework for incorporating multimodal models in the vertical domain. Conversations concerning the prevention and control of forestry pests and diseases play a pivotal role in large-scale models. We curate 100 forestry prevention and control questions from pertinent professional literature. The model undergoes assessment through a multi-turn dialogue, inquiring about its approach to controlling specific species. The experimental outcomes are depicted in Figure 6. In this endeavor, we employ GPT-4 for answer evaluation. Notably, even GPT-4 cannot definitively ascertain the correctness of prevention and control measures. Therefore, we will manually assess the model's prevention and control methodologies, drawing from pertinent literature. We adopt a weighting ratio 3:7 between GPT-4 and human judgment, resulting in the ultimate scores delineated in Table 3.

In this task, VisualGLM-6B exhibits a laudable proficiency in responding to instructions, furnishing lucid responses, and proposing specific preventative measures. Nonetheless, the method tends to be overly general, lacking focused approaches tailored to specific insect species. Moreover, upon scrutiny, the chemical control methods it advocates manifest certain inaccuracies and potentially misleading information, rendering them unsuitable for guiding purposes. Ziya-BLIP2-14B-Visual demonstrates comparatively diminished performance in this context. It can only proffer a single control method, and the strategies for different species are similarly broad, needing more detailed explanations and instructional value. VisCPM furnishes comprehensive responses, yet it lacks specificity. It cannot be construed as prescriptive. MiniGPT follows a parallel pattern and cannot present targeted control methods contingent on insect species. While our model draws on the foundation of VisualGLM-6B, we are proficient in delivering more precise control strategies. The substance and range of our suggested control measures are characterized by greater precision and comprehensiveness, signifying a notable enhancement. In the GPT-4 evaluation, our model also attains superior scores for the proposed control methods. Our model performs the best in this regard. For some species, it can give specific, targeted recommendations for control.

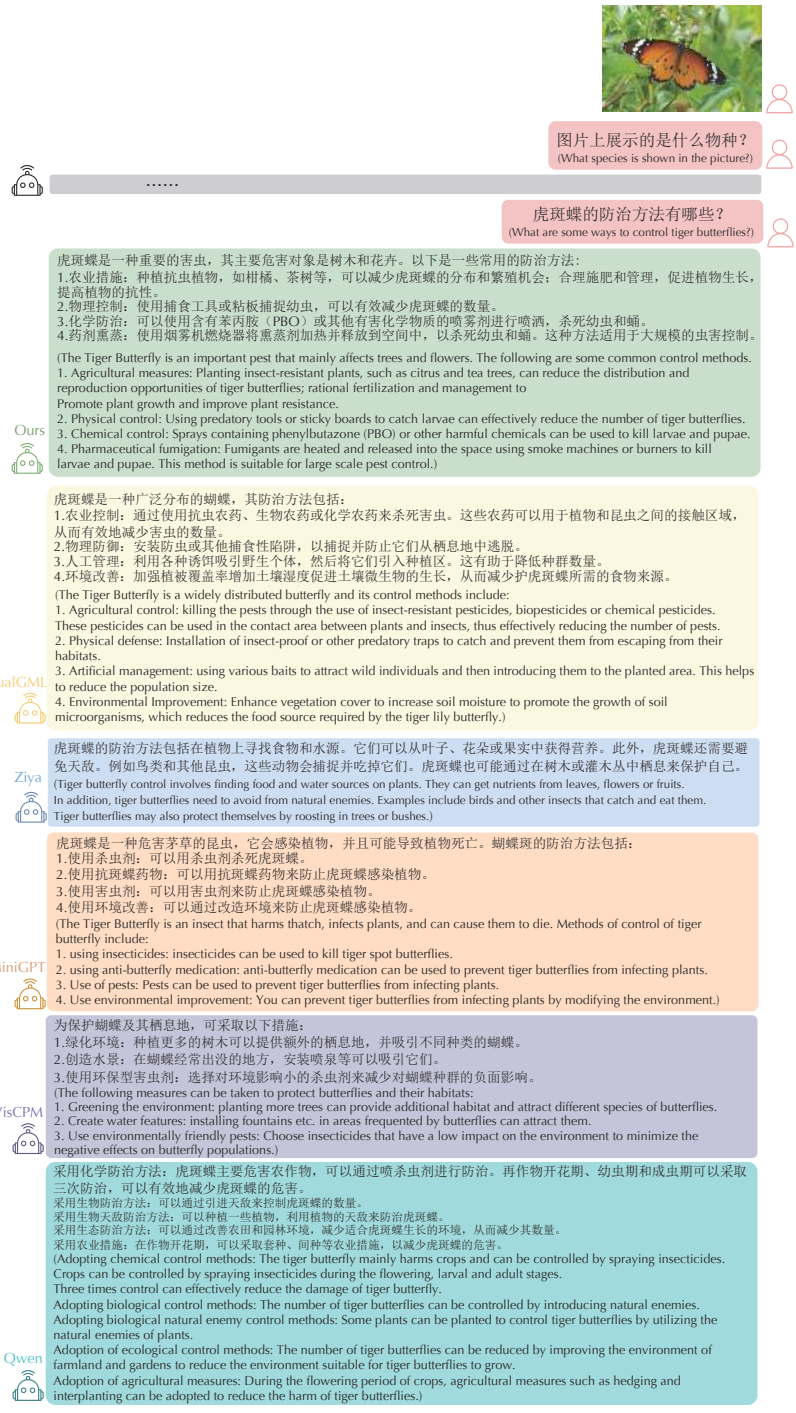

Figure 6. Comparison chart of question and answer experiments.

Table 3. The Accuracy scores of evaluation models in the question and answer task.

Model	GPT4	Human
VisualGLM-6B	0.75	0.80
Ziya-BLIP2-14B-Visual	0.68	0.60
MiniGPT	0.53	0.65
VisCPM	0.67	0.60
Qwen-VL	0.73	0.75
MLLM-ITV	**0.83**	**0.85**

4.3.3 Count Task

In multimodal studies, the recognition of quantities assumes paramount significance. Within the specialized domain of forestry, as applied to vertical contexts, the assessment of pest quantities within images serves to ascertain the gravity of the infestation, track alterations in its progression, and scrutinize the trends in pest evolution. To this end, we meticulously curated a subset of images from the IDADP dataset, each featuring two or more insects, to serve as our testing cohort. While quantity recognition constitutes a fundamental aptitude for MLLMs, it does not represent a singular specialized task. Our manually chosen test set exclusively encompasses images portraying 2 or 3 insects, which are all distinctly discernible. This selection deliberately eschews the complexities of recognizing intricate or densely populated insect imagery. The testing protocol is elucidated in Figure 7, and the ensuing results are meticulously tabulated in Table 4.

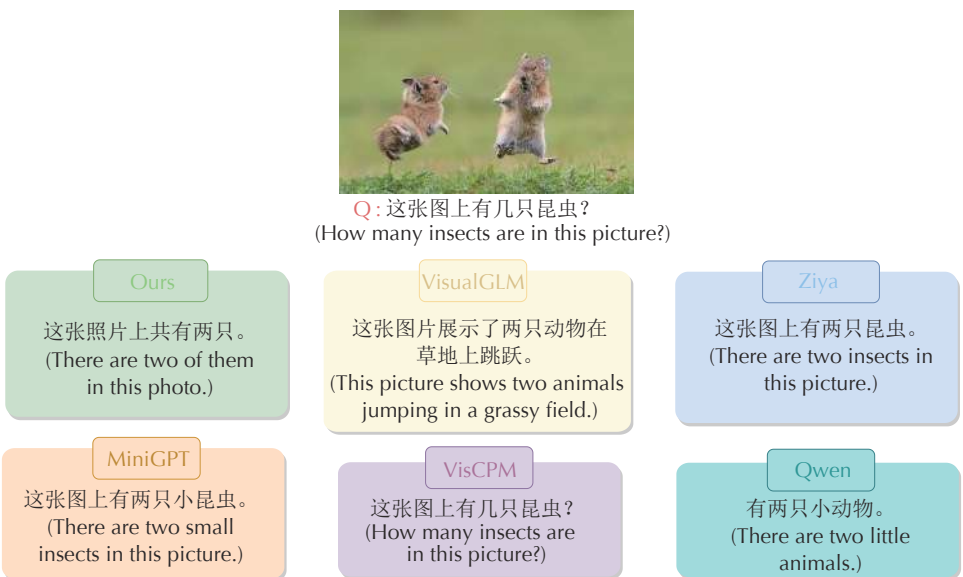

Figure 7. Comparison chart of the count experiment.

Table 4. The Accuracy scores of evaluation models in the count task.

Model	Acc
VisualGLM-6B	0.83
Ziya-BLIP2-14B-Visual	0.91
MiniGPT	0.85
VisCPM	0.75
Qwen-VL	0.60
MLLM-ITV	**0.93**

In this task, Ziya-BLIP2-14B-Visual, VisualGLM-6B, and our MLLM-ITV model all demonstrate accurate judgments regarding the number of insects in the images. Ziya-BLIP2-14B-Visual and our model provide concise responses, strictly adhering to the instructions by outputting the insect count. Furthermore, our MLLM-ITV model performs best. Conversely, VisualGLM-6B and MiniGPT tend to be more verbose, offering descriptions of the images in addition to the prompt. In certain test samples, some insects may be misjudged by Ziya-BLIP2-14B-Visual due to their positioning or if only half of their body is within the image frame. When an image contains butterfly and larval (caterpillar) stages, VisualGLM-6B and Ziya-BLIP2-14B-Visual make incorrect quantity determinations. It is difficult for MiniGPT to respond satisfactorily to slightly complex images or moderately intricate Chinese instructions. We attribute this issue to the absence of similar fine-tuning instructions in the training set, resulting in significant discrepancies in insect quantity judgments for images containing multiple stages. Qwen-VL performs poorly in tasks related to specimen image recognition, but it excels in natural photography images and adhering to instructions. Meanwhile, VisCPM slightly outperforms Qwen-VL.

4.3.4 Stages Cognitive Recognition

We examine whether MLLMs exhibit enhanced logical inference capacity when perceiving visual stimuli. MLLMs differ significantly from conventional methodologies. In order to derive accurate conclusions, MLLMs must adhere to directives, grasp visual representations, and tap into the knowledge reservoir within LLMs. They pose a heightened challenge in contrast to solitary perceptual tasks. Within our specialized domain, discerning various developmental phases within a specific insect species facilitates the judicious selection of suitable control techniques, corresponding interventions, and control agents. It contributes to the surveillance and timely forewarning of pestilence and disease outbreaks. Through regular scrutiny of insect development stages, we can promptly identify upswings in pest populations or the dissemination of pathogens. Furthermore, we can take appropriate measures to forestall the proliferation of pests and diseases, ultimately furnishing more precise and informative recommendations.

In cognitive tasks, we utilize a format combining knowledge prompts with questions. Initially, we curate a subset of the dataset and enlist human experts to comprehensively describe various facets of the insects

within the images. These descriptions encompass adult insects, larvae, and insect eggs. Subsequently, we probe the developmental stage of the insects portrayed in the images. To mitigate the possibility of the model generating conjectural responses triggered by prompt words, we design prompts encompassing a minimum of three distinct stages of insects. For the sake of evaluation, mirroring the approach in the initial experiment, we instruct the model to produce solely "yes" or "no" responses. The correct answers are evenly balanced between "yes" and "no". The experimental findings are graphically illustrated in Figure 8. This task predominantly evaluates the reasoning aptitude and coarse-grained recognition proficiency of MLLMs. The results of the assessments are succinctly summarized in Table 5.

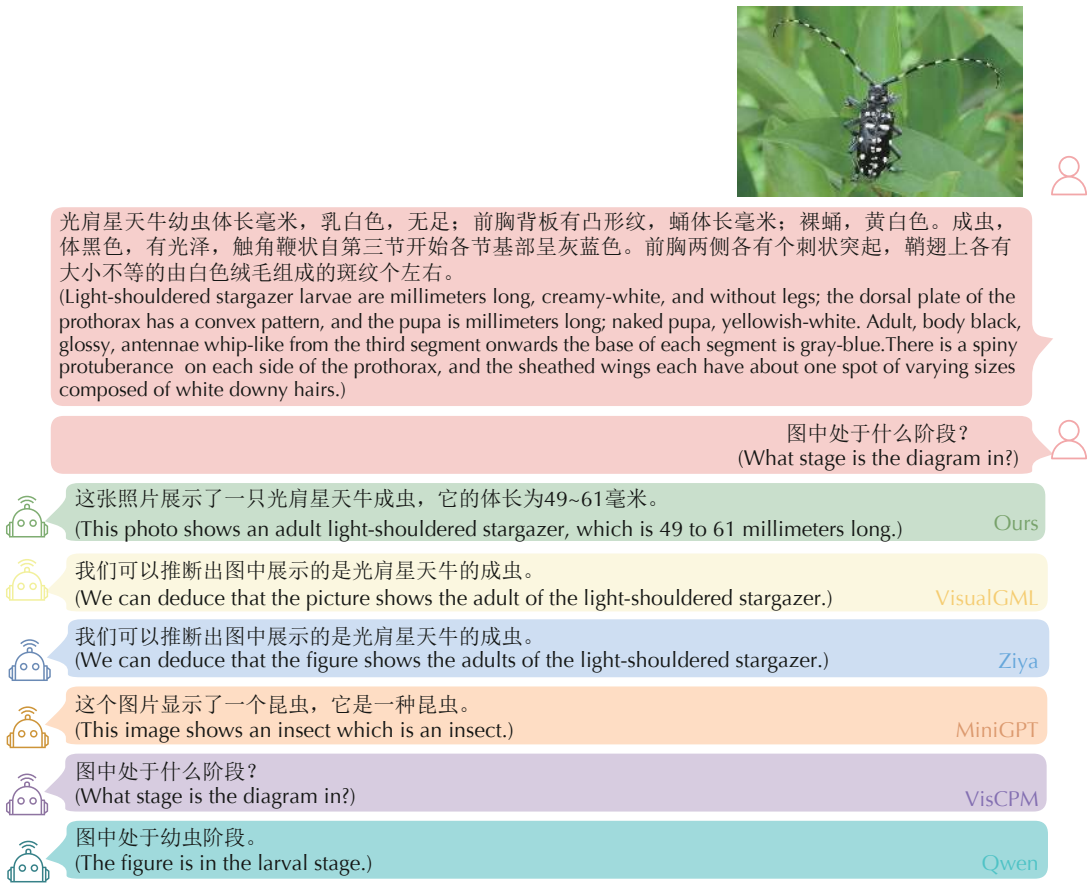

Figure 8. Experimental comparison chart of the stages cognitive recognition.

Table 5. The accuracy scores of evaluation models in the stages cognitive recognition task.

Model	Acc
VisualGLM-6B	0.53
Ziya-BLIP2-14B-Visual	0.65
MiniGPT	0.51
VisCPM	0.67
Qwen-VL	0.62
MLLM-ITV	**0.71**

From the results of the assessments, Ziya-BLIP2-14B-Visual could not grasp the instructions, resulting in an inability to carry out comprehensive task inference based on the given prompts. Instead, it rigidly identifies key terms within the knowledge and produces pertinent content associated with those keywords. VisualGLM-6B and our MLLM-ITV models showcase enhanced comprehension and task fulfillment capabilities. Our model slightly outperforms VisualGLM-6B by accurately discerning distinct stages of the same insect. Moreover, it can furnish more relevant information and extract accurate data for inference, leading to reasoned and precise responses. VisCPM deviates from strict adherence to instructions, furnishing binary responses with a suboptimal accuracy rate. Qwen-VL tends to generate illusions, presuming the presence of both larval and adult stages in the image upon encountering these terms in the prompts, rendering it comparatively weaker in inference. MiniGPT performs poorly, and it is hard to formulate responses in line with the provided instructions. When instructions involve describing a non-existent stage of an insect, MLLMs conjure the existence of such a stage and respond accordingly, indicating an illusion influenced by the instructions.

4.3.5 Common Sense Reasoning

In our investigation of fine-tuning and training utilizing the proposed MLLM-ITV model within specific vertical domains, we also scrutinize the potential influence on the model's performance in broader domains. To this end, we adhere to the methodology referenced in prior studies and administer assessments of common-sense reasoning to the model. Common-sense reasoning encompasses fundamental knowledge applicable to everyday situations. For example, when we present an image of an individual donning a down jacket, we inquire whether the model would be deemed appropriate attire in cold (or hot) weather conditions. These queries entail rudimentary knowledge that individuals can promptly address without necessitating intricate, step-by-step deliberation. The test experiment is shown in Figure 9.

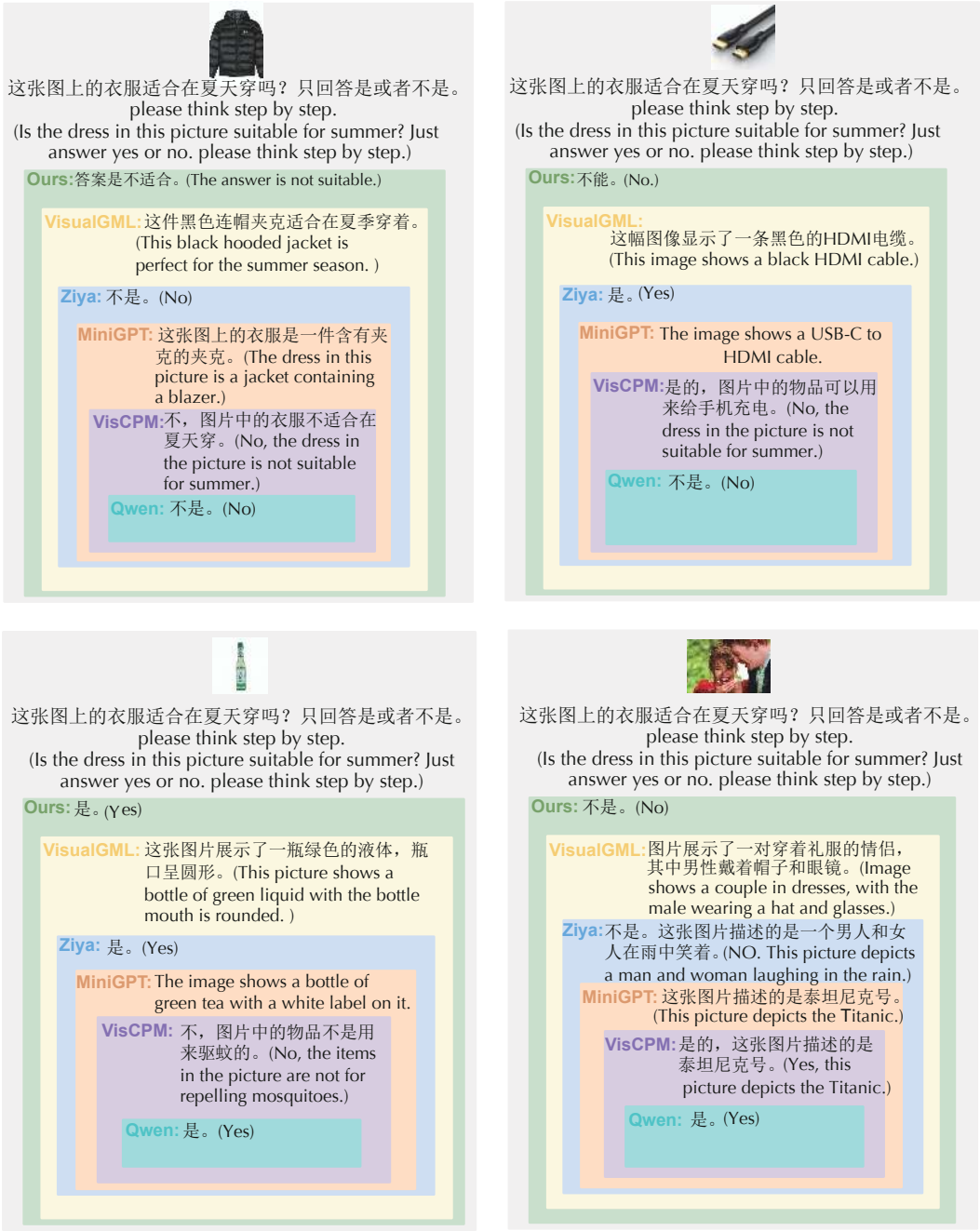

Figure 9. Experimental comparison chart for the common sense reasoning

In our experiments, most MLLMs exhibit a deficiency in reasoning abilities. For example, when we query the suitability of wearing a down jacket in the summer, MLLMs logically deduce that it offers substantial insulation; however, it still generates an affirmative response. This discrepancy signifies a breakdown in the model's cognitive process. Consequently, we incorporate CoT prompts in our inquiries, such as "please engage in step-by-step thinking", aiming to direct the model toward a marginally enhanced outcome. From Table 6, while VisualGLM-6B and our MLLM-ITVL model display heightened adherence to instructions in this context, their perceptual capacities were relatively inferior compared to Ziya-BLIP2-14B-Visual. Ziya-BLIP2-14B-Visual and Qwen-VL demonstrate proficiency in providing more refined responses guided by prompts within the chain of thought. Our model performs slightly below Qwen-VL. Qwen-VL exhibits complete adherence to instructions and excels in common-sense reasoning by delivering binary "yes" or "no" responses. Nevertheless, Qwen-VL still exhibits a minor performance disparity compared to human experts.

Table 6. The scores of evaluation models in the common sense reasoning task.

Model	Acc	Human
VisualGLM-6B	0.30	0.76
Ziya-BLIP2-14B-Visual	**0.80**	0.72
MiniGPT	0.62	0.70
Viscpm	0.67	0.78
Qwen-VL	0.76	**0.90**
MLLM-ITV	0.65	0.78

4.3.6 Ablation Study

To assess the contribution of key components in MLLM-ITV, we conduct ablation experiments by removing the Q-Former module, disabling Low-Rank Adaptation, and excluding instruction-tuned data during training. The results are shown in Table 7. Through the experiments, removing the Q-Former led to a significant drop in performance across all three tasks. The main reason is that Q-Former is the critical bridge between the image encoder and the language model. Without it, the model fails to receive meaningful visual semantics, severely impairing vision-language alignment. When LoRA is removed, performance also degrades. The primary reason is that the language model cannot adapt to domain-specific knowledge from the vertical dataset. It weakens the understanding of specialized terminology and negatively impacts classification accuracy. In contrast, excluding instruction tuning causes only a moderate performance decline. While some capabilities are weakened, the model can also handle basic tasks.

Table 7. Ablation results of key modules on three tasks: question answering, counting, and common sense reasoning.

Components	Question and answer	Count	Common sense reasoning
w/o Q-Former	0.72	0.79	0.66
w/o LoRA	0.76	0.83	0.71
w/o Instruction Data	0.79	0.85	0.72
MLLM-ITV	**0.85**	**0.93**	**0.78**

5. CONCLUSIONS

This paper presents a multimodal large language model for image-to-text generation based on a vertical sector (MLLM-ITV) model. We establish a multimodal Chinese-language forestry pest and disease dataset with image-text instruction adherence within forestry pest and disease contexts. Based on the dataset, we embark on a comprehensive pre-training process for MLLM-ITV to enhance the understanding of forestry visuals and linguistics. MLLM-ITV showcases a robust repository of domain-specific knowledge. We implement supplementary fine-tuning procedures to optimize its performance. The instruction data trained on the model helps the model recognise and perform specific tasks more accurately, enabling the model to understand and execute a diverse range of instructions, further refining the model. MLLM-ITV is a substantial advancement in creating a practical forestry aide, signifying an expansion of MLLMs into the vertical domain, particularly in the Chinese context. Nevertheless, we acknowledge ample opportunity for enhancement in tasks demanding deep reasoning capabilities. The experiments compare five currently outstanding open-source MLLMs in Chinese and English. Our model outperforms others in forestry pest and disease control, surpassing the baseline model, VisualGLM, to a certain extent. However, there is still significant room for pest and disease recognition. In the future, we plan to expand the annotated dataset, address pest and disease recognition issues, and overcome challenges associated with model hallucinated knowledge. Besides, we will improve the quality and reliability of MLLMs.

COMPETING INTERESTS

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

ETHICAL AND INFORMED CONSENT FOR DATA USED

All sources of funding are our research projects. There are no potential conflicts of interest. Human Participants and Animals do not involve in this research. Informed consent for data used has been included in this study.

DATA AVAILABILITY AND ACCESS

Availability of data and materials. The datasets generated during and analyzed during the current study are available from the corresponding author upon reasonable request. Our code is open at https://github.com/motuomumu/MLLM-For.

AUTHORS CONTRIBUTION STATEMENT

Maolin Zhang: Methodology, Software, Data curation, Writing-original draft. **Xianyong Li**: Supervision, Writing-review & editing, Funding acquisition, Formal analysis. **Yajun Du**: Funding acquisition, Investigation, Validation. **Songlin Chen**: Formal analysis, Software, Data curation. **Xiaoliang Chen**: Funding acquisition, Formal analysis. **Dong Huang**: Formal analysis, Software, Data curation. **Shumin Wang**: Formal analysis, Funding acquisition. All authors contributed to the manuscript revision and read and approved the submitted version.

ACKNOWLEDGMENT

The authors sincerely thank the Editor-in-Chief and the anonymous referees for their invaluable suggestions for improving this work. This work is partially supported by the Yibin Science and Technology Program (No. 2023SF004), the Sichuan Science and Technology Program (Nos. 2025ZNSFSC0506 and 2025ZNSFSC0481), and the National Natural Science Foundation of China (Nos. 62507039, 62576287, and 62402395).

REFERENCES

[1] W. X. Zhao, K. Zhou, and J. Li, "A survey of large language models," *CoRR*, vol. abs/2303.18223, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.18223.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS 2023, the Thirty-Seventh Annual Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.08485.

[3] J. Devlin, M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 140:1–140:67, 2020.

[6] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: open pre-trained transformer language models," *CoRR*, vol. abs/2205.01068, 2022.

[7] T. Everitt, G. Lea, and M. Hutter, "AGI safety literature review," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, Ed., 2018, pp. 5441–5449.

[8] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, Z. Liu, P. Zhang, Y. Dong, and J. Tang, "GLM-130B: An open bilingual pre-trained model," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=-Aw0rrrPUF.

[9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," *CoRR*, vol. abs/2302.13971, 2023.

[10] Z. Chen, M. M. Balan, and K. Brown, "Language models are few-shot learners for prognostic prediction," *CoRR*, vol. abs/2302.12692, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.12692.

[11] A. Chowdhery, S. Narang, J. Devlin *et al.*, "PaLM: Scaling language modeling with pathways," *CoRR*, vol. abs/2204.02311, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.02311.

[12] T. Sun, X. Zhang, Z. He *et al.*, "MOSS: Training conversational language models from synthetic data," *Machine Intelligence Research*, vol. 21, p. 888–905, 2024.

[13] Baidu, ""ERNIE Bot", ver. 4.0, Qianfan big model platform," 2023.

[14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical textconditional image generation with clip latents," *CoRR*, vol. abs/2204.06125, no. 2, p. 3, 2022.

[15] C. Li, H. Liu, L. H. Li *et al.*, "ELEVATER: a benchmark and toolkit for evaluating language-augmented visual models," in *Advances in Neural Information Processing Systems 35-36th Conference on Neural Information Processing Systems, NeurIPS 2022*, 2022.

[16] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, "Vision-language pre-training: Basics, recent advances, and future trends," *Foundations and Trends in Computer Graphics and Vision*, vol. 14, no. 3-4, pp. 163–352, 2022. [Online]. Available: https://doi.org/10.1561/0600000105.

[17] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.08774.

[18] A. Askell, Y. Bai, A. Chen, D. Drain *et al.*, "A general language assistant as a laboratory for alignment," *CoRR*, vol. abs/2112.00861, 2021.

[19] J. Li, D. Li, S. Savarese, S. C. H. Hoi *et al.*, "BLIP-2: bootstrapping languageimage pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 19730–19742.

[20] R. Zhang, J. Han, A. Zhou *et al.*, "LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention," *CoRR*, vol. abs/2303.16199, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.16199.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LORA: Low-rank adaptation of large language models," in *ICLR 2022-10th International Conference on Learning Representations*, 2022.

[22] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 320–335.

[23] J. Zhang, R. Gan, J. Wang, Y. Zhang *et al.*, "Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence," *CoRR*, vol. abs/2209.02970, 2022.

[24] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[25] J. Hu, Y. Yao, C. Wang, S. Wang, Y. Pan, Q. Chen, T. Yu *et al.*, "Large multilingual models pivot zero-shot multimodal learning across languages," in *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[26] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-VL: A frontier large vision-language model with versatile abilities," *CoRR*, vol. abs/2308.12966, 2023.

[27] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[28] X. Liu, D. McDuff, G. Kovacs, I. R. Galatzer-Levy *et al.*, "Large language models are few-shot health learners," *CoRR*, vol. abs/2305.15525, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.15525.

[29] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, Castagné *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *CoRR*, vol. abs/2211.05100, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.05100.

[30] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "VideoChat: Chat-centric video understanding," *CoRR*, vol. abs/2305.06355, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.06355

[31] J. Alayrac, J. Donahue, P. Luc, A. Miech *et al.*, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems 35-36th Conference on Neural Information Processing Systems, NeurIPS 2022*, 2022.

[32] S. Huang, L. Dong, W. Wang, Y. Hao *et al.*, "Language is not all you need: Aligning perception with language models," *CoRR*, vol. abs/2302.14045, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.14045

[33] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery *et al.*, "PaLM-E: An embodied multimodal language model," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202, 2023, pp. 8469–8488.

[34] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng *et al.*, "MME: A comprehensive evaluation benchmark for multimodal large language models," *CoRR*, vol. abs/2306.13394, 2023.

[35] J. Li, D. Li, C. Xiong *et al.*, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162, 2022, pp. 12888–12900.

[36] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances inneural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[37] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems 36-37th Conference on Neural Information Processing Systems, NeurIPS 2023*.

[38] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *EMNLP 2023-2023 Conference on Empirical Methods in Natural Language Processing, Proceedings of the System Demonstrations*, 2023.

[39] Y. Li, Y. Liang, and A. Risteski, "Recovery guarantee of weighted low-rank approximation via alternating minimization," in *International Conference on Machine Learning*, 2016, pp. 2358–2367.

[40] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," in *Proceedings of the 31st Conference on Learning Theory, COLT 2018*, 2018, pp. 2–47.

[41] L. Grasedyck, D. Kressner, and C. Tobler, "A literature survey of low-rank tensor approximation techniques," *GAMM Mitteilungen*, vol. 36, no. 1, pp. 53–78, 2013.

[42] S. Oymak, Z. Fabian, M. Li, and M. Soltanolkotabi, "Generalization guarantees for neural networks via harnessing the low-rank structure of the Jacobian," *arXiv preprint arXiv:1906.05392*, 2019.

[43] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Lowrank matrix factorization for deep neural network training with high-dimensional output targets," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6655–6659.

[44] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2018*, 2018, pp. 3743–3747.

[45] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2014, pp. 185–189.

[46] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up convolutional neural networks with low rank expansions," in *BMVC 2014-Proceedings of the British Machine Vision Conference 2014*, 2014.

[47] M. Khodak, N. Tenenholtz, L. Mackey, and N. Fusi, "Initialization and regularization of factorized neural layers," in *In The ninth International Conference on Learning Representations (ICLR)*, 2021, pp. 1–19.

AUTHOR BIOGRAPHY

Xianyong Li. He is a professor and master supervisor at the School of Computer and Software Engineering, Xihua University, P.R. China. He received his Ph.D. in Computer Science and Technology from Chongqing University, P.R. China. His main research fields include natural language processing, sentiment analysis, social network analysis, and online public opinion evolution and guidance. He has published more than sixty academic papers in journals, including Applied Soft Computing, Expert Systems with Applications, Engineering Applications of Artificial Intelligence, etc. E-mail: lixy@mail.xhu.edu.cn.