| **TITLE** | **AUTHOR** | **YEAR** | **OBJECTIVE** | **METHODOLOGY** | **DATA SET** | **FINDINGS** | **RELAVANCE** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Flamingo: A visual language model for few-shot learning** | Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc et al. | 2022 | Few-shot vision-language learning | Visual encoder + frozen LM with cross-attention | MultiModal MassiveWeb,ALIGN,LTIP,VTP | Strong few-shot reasoning | Basis for multimodal reasoning in agriculture |
| **Qwen-vl - A versatile vision-language model for understanding, localization, text reading, and beyond** | J Bai, S Bai, S Yang, S Wang, S Tan, P Wang, J Lin, C Zhou, and J Zhou | 2023 | Versatile vision-language understanding |  | COCO(Common Object In Context)  LVIS(Large Vocabulary Instance In Segmentation) | Robust multi-task ability | Useful for recognizing pests and text on leaves |
| **Sparks of artificial general intelligence: Early experiments with gpt-4** | Sebastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Jo- ´ hannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al | 2023 | Study emergent reasoning in GPT-4 | Evaluations across reasoning, vision,  coding, Medicine, Language,  Law | Specific Dataset is Not Disclosed publicly. | Early AGI behaviors | Inspires advanced reasoning in Agrillava |
| **X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages** | Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu | 2023 | Treat modalities as foreign languages | Multimodal pretraining via text-language mapping | Specific Dataset is Not Disclosed publicly. | Flexible cross-modality transfer | Helps integrate visual & text plant data |
| **Allava: Harnessing gpt4v-synthesized data for a lite vision-language model.** | Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, et al. | 2024 | Lightweight VL model using GPT-4V data | Uses synthetic GPT-4V captions for training | LAION (Large Scale web crawled image dataset)  Vision FLAN | Efficient with small compute | Ideal for low-resource agri setups |
| **Sharegpt4v: Improving large multi-modal models with better captions.** | Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, Dahua Lin | 2023 | Improve captions for multimodal training | Enhanced caption datasets for fine-tuning | COCO,LAION, CC 3M, SBU, SAM, TextCaps, WikiArt. | Better caption alignment | Improves image-text clarity in agri datasets |
| **Palm: Scaling language modeling with pathways** | Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. | 2023 | Scale language modeling via Pathways | Transformer-based, large distributed training | Massive web corpus | Strong generalization & reasoning | Foundation for Agrillava’s language core |
| **Instructblip: Towards generalpurpose vision-language models with instruction tuning** | Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, et al. | 2024 | Instruction-tuning on BLIP 2 models | Trains an instruction-aware Query Transformer on 26 publicly available vision-language datasets, enabling it to extract informative features tailored to specific instructions. | Specific Dataset is Not Disclosed publicly | Strong instruction following | Enables question answering for agri use |
| **Llama-adapter v2: Parameter-efficient visual instruction model** | Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang et al. | 2023 | Parameter Efficient visual instruction tuning | LLaMa  Adapter V2 unlocks more learnable patterns across LLaMa model. | COCO Caption  Dataset. | Strong instruction following. Open Ended  Visual Reasoning. | Easy  domain adaptation  for pests/ diseases |
| **Multimodal-gpt: A vision and language model for dialogue with humans.** | Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Kuikun Liu, Wenwei Zhang, et al. | 2023 | Multi round Dialogue with visual understanding | GPT backbone + image encoder | VQA v2.O,  OKVQA, GQA,  CLEVR, NLVR. | Visual-aware human-like chat | Chat-based pest diagnostic assistant |
| **A probabilistic interpretation of precision, recall and f-score, with implication for evaluation.** | Cyril Goutte and Eric Gaussier | 2005 | Clarify precision, recall, F-score | Probabilistic evaluation theory | Benchmark  IR datasets | Better metric interpretation | For evaluating Agrillava performance |
| **Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture** | Lakshay Goyal, Chandra Mani Sharma, Anupam Singh, and Pradeep Kumar Singh | 2021 | Detect wheat diseases | Deep CNN for wheat leaf/spike image classification with SVM and preprocessing. | LWDCD2020 (Leaf and Wheat Disease Classification Dataset 2020) | achieved high accuracy of 97.88% on segmenting and classifying 10 wheat disease classes | Relevant for crop disease module |
| **An open access repository of images on plant health to enable the development of mobile disease diagnostics** | David Hughes, Marcel Salathe, et al | 2015 | to support the development of mobile disease diagnostic tools using machine learning and crowdsourc-ing. | The authors collected and curated over 50,000 images of healthy and diseased plant leaves. | Plant  Village dataset | enhancing mobile diagnostics potential | Core training source for Agrillava |
| **Lisa: Reasoning segmentation**  **via large language model** | Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui  Yuan, Shu Liu, and Jiaya Jia | 2023 | To introduce a  LLM-based segmentation reasoning | LISA adds a <SEG> token and embedding-as-mask to enable segmentation from multimodal LLMs | COCO, ADE20K | LISA shows strong zero-shot segmenta-tion and improved performan-ce after fine-tuning. | Helps analyze infected leaf regions |
| **Bloom: A 176b-parameter open-access multilingual**  **language model** | Teven Le Scao, Angela Fan, Christopher Akiki, Ellie  Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagne, et al. | 2023 | To build and release an open-access, large-scale multilingual language model | Transformer trained on multilingual corpora | ROOTS corpus (341B tokens, 46 languages) | achieves strong competitive performance and serves as a transparent | Useful for regional agri-language support |
| **Llava-med: Training a large language**  **and-vision assistant for biomedicine in one day** | Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama,  Haotian Liu, Jianwei Yang, Jianfeng Gao, et al. | 2024 | Advances in  Neural Information Processing Systems | LLaVA-Med fine-tunes biomedical vision-language model in one day using GPT-4V-refined PubMedVision dataset. | PubMed  Vision dataset | Improving reasoning and interpreta-tion. | Template for Agri domain adaptation |
| **Mini-gemini: Mining the potential of multi-modality**  **vision language models.** | Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng  Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya  Jia | 2024 | Enhance VLMs by improving visual tokens, data quality, and guided generation. | Mini-Gemini uses dual vision encoders for refined high-resolution and low-resolution embeddings. | Synthetic multimodal data | Mini-Gemini leads zero-shot benchmarks, surpassing previous state-of-the-art models. | Efficient for limited agri data |
| **Improved Baselines with Visual Instruction Tuning** | Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee | 2024 | systematically investigate and improve the design of LMMs with visual instruction tuning. | LLaVA improved with MLP connector, CLIP-ViT-L, VQA data; fine-tuned fast. | COCO, VQA dataset | Improving data efficiency, training speed. | Improves Agrillava’s tuning efficiency |
| **Visual Instruction Tuning** | Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee | 2024 | To extend instruction tuning to multimodal language-image models. | GPT-4 generates multimodal instructions;  LLaVA trains end-to-end; benchmarks test. | COCO dataset | LLaVA matches GPT-4V, scoring 85.1% on synthetic data; Science QA fine-tuning achieves 92.53% accuracy. | Core idea for Agrillava’s design |
| **Efficient vision**  **language instruction tuning for large language models.** | Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai  Sun, and Rongrong Ji | 2024 | Efficient affordable vision-language tuning preserving LLM NLP capabilities using MMA. | MMA connects encoders and LLMs via lightweight adapters, enabling dynamic modality shifts and joint optimization. | COCO, VQAv2 | Competes with advanced multimodal LLMs, reducing training time and costs. | Fits resource-constrained agri setups |
| **Macaw-llm: Multi-modal language modeling**  **with image, audio, video, and text integration** | Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting  Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and  Zhaopeng Tu | 2023 | Unified model for image/audio/video/text | Cross-modal fusion | Multi-modal datasets | Multi-sensory understanding | Extend Agrillava to video pest detection |
| **Video-chatgpt: Towards detailed video**  **understanding via large vision and language models.** | uhammad Maaz, Hanoona Rasheed, Salman Khan, and Fa-  had Shahbaz Khan | 2023 | Video understanding via LLM | Frame encoder + GPT backbone | VideoQA datasets | Detailed temporal reasoning | Useful for drone-based crop monitoring |
| **Language models are few-shot learners** | Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal,  A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal,  et al | 2020 | Few-shot learning via large scale | Transformer pretraining | WebText, Books | Few-shot capabilities | Foundation for Agrillava’s language reasoning |
| **GPT-4 (Vision) System Card** | OpenAI | 2023 | Vision-capable GPT-4 | Multimodal architecture | Internal OpenAI data | Powerful visual reasoning | Benchmark for multimodal design |
| **Detgpt: Detect what you need via**  **reasoning** | Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong,  Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Ling-  peng Kong Tong Zhang | 2023 | Detection via reasoning | Visual reasoning + LLM | Detection datasets | Detects objects by reasoning | Helps identify pest regions on leaves |
| **Pandagpt: One model to instruction-follow them**  **all** | Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and  Deng Cai. | 2023 | Unified instruction following | Multimodal instruction alignment | Visual-language data | Universal instruction model | Consistent response generation for Agrillava |
| **Stanford alpaca: An instruction-following llama**  **model** | Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois,  Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B  Hashimoto | 2023 | Instruction following with LLaMA | Fine-tuned LLaMA with self-instruct data | Text instruction data | Strong alignment with small data | Basis for instruction-tuning Agrillava |
| **Open and efficient foundation language models** | Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier  Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste  Rozi`ere, Naman Goyal, Eric Hambro, Faisal Azhar, et al | 2023 | Efficient foundation LLM | Optimized transformer training | Web, books, code | Strong open model | Base language backbone for Agrillava |
| **To see is to believe: Prompting**  **gpt-4v for better visual instruction tuning** | Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan  Wu, and Yu-Gang Jiang | 2023 | Better prompts for visual tasks | Prompt engineering study | GPT-4V image tasks | Improved visual responses | Enhances prompt design for agri images |
| **Cogvlm: Visual expert for pretrained language**  **models** | Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji  Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan  Song, et al | 2023 | Visual expert for LLMs | Vision-aware fine-tuning | Image-caption datasets | Strong visual understanding | Adds visual reasoning depth to Agrillava |
| **Visionllm: Large language model is also an open-**  **ended decoder for vision-centric tasks** | Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu,  Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu  Qiao, et al | 2024 | Vision-centric decoder model | Unified vision decoder | Visual benchmarks | General visual task solver | For segmentation and leaf analysis |
| **Finetuned language models are zero-shot learn-**  **ers** | Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu,  Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and  Quoc V Le | 2021 | Zero-shot learners via tuning | Fine-tuned GPT models | Text corpora | Improves zero-shot learning | Used for few-shot pest Q&A |
| **Ip102: A large-scale benchmark dataset**  **for insect pest recognition** | Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng,  and Jufeng Yang | 2019 | Insect pest recognition dataset | Large-scale benchmark | 102 insect classes | Benchmark accuracy >90% | Primary dataset for Agrillava’s pest module |
| **Multiinstruct: Im-**  **proving multi-modal zero-shot learning via instruction tun-**  **ing** | Zhiyang Xu, Ying Shen, and Lifu Huang | 2022 | Improve multi-modal zero-shot | Instruction-tuned multi-modal data | COCO, VQA | Stronger cross-modal zero-shot | Template for multi-task tuning |
| **mplug-owl: Modularization empowers**  **large language models with multimodality** | Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan,  Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi,  Yaya Shi, et al | 2023 | Modular multimodal LLM | Modular plug-in vision encoders | Visual QA datasets | Flexible multi-modal LLM | Helps modular design of Agrillava |