

AgriBench: A Hierarchical Agriculture Benchmark for Multimodal Large Language Models

Yutong Zhou¹ and Masahiro Ryo^{1,2}

¹ Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder Str. 84, 15374, Müncheberg, Germany yutong.zhou@zalf.de

² Brandenburg University of Technology Cottbus-Senftenberg, Platz Der Deutschen Einheit 1, 03046, Cottbus, Germany

<https://github.com/Yutong-Zhou-cv/AgriBench>

Abstract. We introduce AgriBench, the first agriculture benchmark designed to evaluate MultiModal Large Language Models (MM-LLMs) for agriculture applications. To further address the agriculture knowledge-based dataset limitation problem, we propose MM-LUCAS, a multimodal agriculture dataset, that includes 1,784 landscape images, segmentation masks, depth maps, and detailed annotations (geographical location, country, date, land cover and land use taxonomic details, quality scores, aesthetic scores, *etc.*), based on the Land Use/Cover Area Frame Survey (LUCAS) dataset, which contains comparable statistics on land use and land cover for the European Union (EU) territory. This work presents a groundbreaking perspective in advancing agriculture MM-LLMs and is still in progress, offering valuable insights for future developments and innovations in specific expert knowledge-based MM-LLMs.

Keywords: Agriculture Benchmark · Dataset · Hierarchical Evaluations

1 Introduction

“Agriculture is the most healthful, most useful and most noble employment of man.”

– George Washington (1732–1799)

Agriculture is an important foundation for human existence. Nearly half of the terrestrial land is used for agriculture in Europe, contributing to food, fiber, and bio-energy resource production [32]. Agriculture depends on agroecosystems comprising subterranean soil, soil organisms, habitats for wild flora, and animals in and around the fields. Traditional agricultural practices heavily depended on the empirical knowledge and expertise of farmers to achieve productive yields. To increase efficiency and optimize production, digitalization has been an important agenda in agriculture for increasing efficiency and optimization, including the utility of Artificial Intelligence (AI). AI has rapidly developed and is widely used to promote agricultural automation. This advancement has significant promise

for enhancing agricultural processes, including efficient assessment, explanation, informed decision-making, and understanding of agricultural systems.

There have been several significant advancements in Machine Learning (ML) and Deep Learning (DL) domains within the agriculture and biodiversity research field in the past decade, such as species identification [21], wildlife monitoring and protection [39], plant disease detection [22,48], plant phenotyping [30], crop classification [42], weed detection [35], intelligent spraying [15,40], robotic harvesting [20, 41], *etc.* Despite these achievements, conventional ML and DL models still have certain limitations: require extensive, task-specific, and well-labeled datasets for effective training; only adapt to specific tasks but cannot generalize to other tasks or unseen data. Due to these limitations, several approaches have been examined, including transfer learning [34], few-shot learning [37], label-efficient learning [27], self-supervised learning [56], to name a few.

Recently, benefiting from the advancements in Large Language Models (LLMs) [8,11,44], MultiModal Large Language Models (MM-LLMs) have rapidly become a new paradigm bridging the fields of Natural Language Processing (NLP) and Computer Vision (CV). MM-LLMs preserve the inherent reasoning and decision-making capabilities of LLMs and showcase remarkable versatility and efficiency across a diverse range of multimodal (MM) tasks [55], such as emotional understanding [52], image captioning [28,38], Visual Question Answering (VQA) [19], *etc.* However, the agricultural domain involves more complex and expert tasks, including multimodality and human-nature interactions, which present significant challenges even for advanced MM-LLMs. Therefore, it is essential to develop benchmarks to evaluate the performance of the existing models specialized for agriculture. To our knowledge, no such MM-LLM benchmarks currently exist.

Here, we present four key questions about MM-LLMs in the agriculture domain, which will be discussed in detail in the following sections.

Q1: *How to evaluate the MM-LLM's capacities in agriculture?*

A1: [Sec. 3] Due to the unique needs and complexity of agricultural research, it is important to build an agriculture-specified benchmark to evaluate the MM-LLMs. Thus, we propose a novel **Agriculture Benchmark: AgriBench**.

Q2: *How to design an agriculture multimodal dataset?*

A2: [Sec. 4] To address the lack of agricultural datasets, we create **MM-LUCAS** with 1,784 annotated agricultural scenery images from 27 EU countries.

Q3: *How effective are advanced MM-LLMs in solving agricultural problems without extra fine-tuning?*

A3: [Sec. 5] MM-LLMs understand general agricultural content well but struggle with specific problems like diagnosing plant diseases without fine-tuning.

Our contributions can be summarized as follows:

1. AgriBench is the first hierarchical benchmark to evaluate the comprehension and reasoning abilities of existing MM-LLMs regarding agriculture.
2. An innovative MM-LUCAS dataset is also newly designed for the proposed AgriBench, which contains corresponding multi-modality annotations.
3. We evaluate the capability of 5 MM-LLMs on our benchmark and present some promising directions for future expert knowledge-based MM-LLMs.

2 Related Works

Multimodal Large Language Models (MM-LLMs) The rapid developments and remarkable achievements of LLMs have driven a growing research interest in MM-LLMs. These models enhance multimodal comprehension by aligning visual features from pre-trained image encoders with LLMs on image-text datasets. Groundbreaking MM-LLMs, such as Flamingo [9], GPT-4 [8], ModaVers [47], Cambrian-1 [43] and LLaVA-OneVision [26] have successfully fused visual data and text and adapted to various multimodal tasks. However, regarding specialized research fields and reality applications, such as industry, agriculture, and healthcare, existing advanced MM-LLMs still face significant challenges in accurately and comprehensively handling domain-specific tasks.

Benchmarks for Multimodal Large Language Models Current benchmarks mainly focus on the ability to predict the understanding of vision-text inputs. For instance, MMBench [31] creates extensive question sets to enhance the objective evaluation of MM-LLMs. GVTBench [46] is designed for two new tasks for MM-LLMs (object counting and multi-class identification), however, its evaluations are limited to some specific aspects of visual understanding. For specialized domains, evaluation requires a level of expertise and precision that current benchmarks often fail to achieve. Therefore, it is essential to develop MM-LLM benchmarks specifically customized for these research domains.

Land Use/Cover Area Frame Survey (LUCAS) The European Union (EU) encompasses a wide variety of landscapes and ecosystems, from densely populated urban centers to sparsely inhabited rural areas. Knowing the patterns in “Land Cover” and “Land Use” is important for human activity and geography.

The Land Use/Cover Area Frame Survey (LUCAS) is one of the most extensive and authoritative in-situ field surveys across Europe. Land cover (LC) denotes the visible physical and biological features, such as cropland and waterbody. Land use (LU) refers to how humans utilize the land for socio-economic purposes, such as residential living and agriculture [3]. From 2006 to 2018, LUCAS was conducted every three years across EU member states, providing a standardized framework for collecting statistics on LC and LU and other information such as soil physico-chemical parameters. Data was collected from 1,351,293 points at 651,780 unique locations, covering 106 variables, and including 5.4 million scenery photos [14]. The LUCAS data includes (1) Microdata on LC, LU, and environmental parameters; (2) Landscape images from the four cardinal directions (north, south, east, and west); and (3) Statistical tables aggregating estimates of LC and LU at the geographical level based on the microdata. This survey has significantly contributed to understanding agriculture, environmental conditions, and sustainable development across the EU. Previous studies published several datasets [24,54] based on the extensive information of LUCAS. Recently, Martinez-Sanchez *et al.* published a segmentation dataset [33], we use this database to develop a novel MM-LLM benchmark dataset.

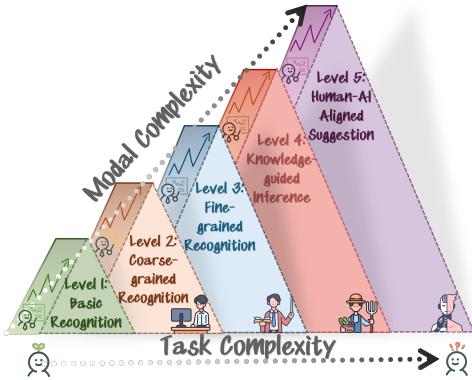


Fig. 1: Five levels of MM task difficulty in the agricultural domain.

Table 1: Abbreviations

| | |
|-------------------|----------------------------------|
| AgriBench: | Agriculture Benchmark |
| AI: | Artificial Intelligence |
| AS: | Aesthetic Score |
| EU: | European Union |
| I: | Image |
| LC: | Land Cover |
| LU: | Land Use |
| LUCAS: | Land Use/Cover Area Frame Survey |
| MM: | Multi-modal |
| MM-LLMs: | Multimodal Large Language Models |
| QS: | Quality Score |
| SOTA: | State-of-the-Art |
| T: | Text |

3 AgriBench

We introduce AgriBench, the first hierarchical benchmark designed to assess the visual comprehension capabilities of MM-LLMs in the agricultural domain. Most existing MM-LLM benchmarks mainly focus on multimodal complexity. However, given the unique characteristics and requirements of specific domains, we argue that task complexity must also be taken into account. Multimodal complexity and task complexity should be considered as two distinct but correlated axes. It is feasible to encounter complex tasks involving a single modality or simple tasks requiring multimodal data. To address this, AgriBench designs diverse scenario tasks that reflect real-world agricultural challenges, ensuring a robust assessment of model performance. Furthermore, AgriBench evaluates models across five levels of task complexity and various modalities, providing a comprehensive framework for advancements in agricultural AI.

3.1 Hierarchical Standard

Agriculture requires a broad range of tasks, from simple object detection to complex decision-making (*e.g.*, fertilization strategy). In particular, high-stakes decision-making (*e.g.*, planning and management) typically has no objectively single correct answer. There will be several answers equally convincing for domain experts. Therefore, this ambiguity demands to be addressed within the context of human-centered AI. As illustrated in Fig. 1, we define the capabilities of MM-LLMs in agriculture into hierarchical levels standards ranging from L1 (Basic Recognition) to L5 (Human-AI Aligned Suggestion).

3.2 The 5 Levels of Agriculture MM-LLMs Evaluation Strategy

We propose the 5 levels of MM-LLM capability that can assess to which extent the model can address multiple modalities and various tasks. For each

level, we first describe the definition and some agriculture task examples, ranging from perception to cognition, simple to complex. Then, we summarize totally 17 main tasks with detailed descriptions as follows, with the template: “ $T \rightarrow T / I \rightarrow T / I \rightarrow I / T + I \rightarrow T / T + I \rightarrow I$ agriculture tasks: details.”

Level 1: Basic Recognition At the lowest level, the model should accurately describe visual content in images. The textual descriptions must be immediately clear to the user based on human intuition, without additional justification to evaluate performance. This includes object detection (*e.g.*, identifying fruits on a tree, flowering plants, and weeds on soil), boundary delineation (*e.g.*, distinguishing soil from plants), and species classification (*e.g.*, categorizing visually distinct common crops such as sunflowers, lavenders, wheat, maize, and cotton).

- $T \rightarrow T$ *Basic Question Answering*: Describe the basic information on crops or scenes based on broad knowledge from LLMs without complex inference.
- $T + I \rightarrow T$ *Species Classification*: Detect and classify common crops.
- $T + I \rightarrow T$ *Image Captioning*: Describe the visible agriculture objects.

Level 2: Coarse-grained Recognition The model can describe clearly definable properties. These properties should be objectively extractable from the MM-LLMs, although the extraction process may require time for the user. While the task complexity is higher than Level 1, no inference is still necessary, and users should consistently agree on the results. This includes object counting (*e.g.*, counting the number of fruits in an image or on a specific branch), dominant class detection (*e.g.*, identifying the most prevalent crop type in the image if multiple crops are seen), and classification (*e.g.*, predicting widely recognized basic phenological growth stages such as seedling, vegetative, flowering, and ripening).

- $T + I \rightarrow T$ *Object counting*: Count the number of user-specified objects.
- $T + I \rightarrow T$ *Basic Scene Analysis*: Generate detailed textual descriptions of multiple visible agriculture objects and the overall scene accurately.

Level 3: Fine-grained Recognition The model at this level can describe fine-grained properties, though the estimation involves some subjectivity. Unlike Level 2, while a common consensus on the answer can be reached, there may still be some disagreement. This includes tasks such as image enhancement (*e.g.*, multiple enhancement methods are possible) and grounded image captioning (*e.g.*, the explanation focus could be various if multiple objects are present).

- $T \rightarrow T$ *Advanced Question Answering*: Describe the in-depth, detailed information on crops or scenes based on specific knowledge from existing LLMs that may require some degree of inference.
- $T + I \rightarrow T$ *Dense object counting*: Counting dense objects such as the numbers of individual plants, cherries, and flowers (Note that the number can be more than 50-100 in a single image).

- **T+I→T Contextual Scene Analysis:** Generate fine-grained descriptions of scenes considering various agricultural objects. Furthermore, evaluate the degree of aesthetics based on the analysis.
- **T+I→I Image Enhancement:** Recover high-quality images from low-quality (extreme weather, motion blur, or low light) images.

Level 4: Knowledge-guided Inference The model can describe elements that are not directly visible, requiring educated guesses based on expert insights or domain-specific knowledge. Experienced users can make informed guesses using visible properties as hints, often relying on empirical correlations with a high degree of flexibility. Thus, adding additional data sources can enhance the accuracy of these predictions. This includes tasks such as regression (*e.g.*, estimating crop yield, plant health, soil health), classification (*e.g.*, identifying plant disease, crop phenotype, and geographic region), and scene generation.

- **I→T Surrounding Scene Analysis:** Predict crop yield based on visible properties (plant size, leaf color, and density) with historical data and weathers.
- **I→I Automatic Image Enhancement:** Automatically enhance images from a comprehensive scene analysis, without relying on additional guidance.
- **T+I→T Environmental Impact Prediction:** Combine visual images with information on fertilizer use, irrigation practices, and local biodiversity to predict the environmental impact.
- **T+I→T Camouflaged Object Detection:** Identify camouflaged objects blended with surroundings, aiding in pest activity risk assessment.
- **T+I→I Scene Generation:** Generate scenes for future scenarios based on current observations.

Level 5: Human Aligned Suggestion Level 5 extends beyond inference based on expert insights. The model should be capable of suggesting actions or scenarios for future implementation. These suggestions could influence high-stakes decision-making, requiring the model to provide comprehensive justifications, which ensures that users can assess the validity and feasibility of the suggestions before taking action.

- **T→T Strategic Planning:** Suggest long-term strategies using continuously updated current and historical data, such as recommending crop rotation and irrigation schedules to maximize yield and soil health over multiple seasons.
- **I→T Scene Projection:** Evaluate various “what-if” scenarios, such as the impact of changing a particular farming practice or responding to an unexpected event (*e.g.*, sudden weather changes), and suggest the best course of action with a detailed analysis of potential outcomes and probabilities.
- **T+I→T Sustainability Recommendations:** Propose sustainable farming practices that balance productivity with environmental conservation, such as reducing chemical use, or adopting no-till farming, based on accurate prediction and detailed analysis results.

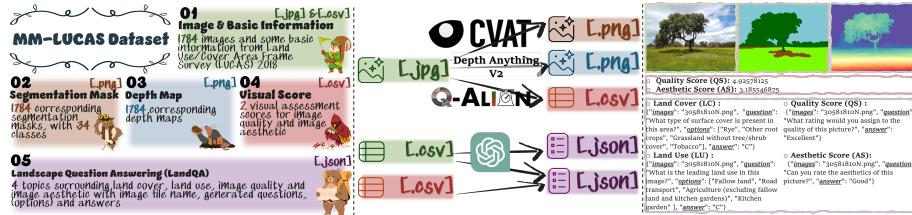


Fig. 2: *Left:* Overview of the MM-LUCAS dataset. *Middle:* The illustration of data collection and processing. *Right:* Data example in the MM-LUCAS dataset. *The box colors in the Middle and Right correspond to the colors in the Left.*

4 MM-LUCAS Dataset

Image acquisition and computer vision play a central role in advancing agricultural digitalization and optimization. Images are commonly captured using handheld cameras [16], Unmanned Aerial Vehicles (UAV)-mounted cameras [18], and earth observatories [45]. While multispectral (*e.g.*, near-infrared) and hyperspectral imaging techniques are increasingly popular for estimating plant and soil conditions across various wavelengths, RGB images still serve as the basis for various agricultural applications.

However, there is a lack of publicly available and well-annotated agricultural image datasets with multimodal information. To address this gap and enhance the agricultural knowledge understanding capabilities of MM-LLMs, we propose a novel multimodal agriculture dataset with specialized annotations, designed to follow the proposed AgriBench, named MM-LUCAS. As shown in Fig. 2, our dataset has the following properties: (1) It contains 1,784 scenery images and some basic information from the original LUCAS dataset. (2) It includes 1,784 corresponding semantic segmentation masks and depth maps. (3) For each image, we assess the quality and aesthetic. (4) We generate landscape question answering with 4 topics. MM-LUCAS was developed based on the existing segmentation LUCAS dataset [33]. We describe the detailed information below.

Table 2: Detail of microdata in MM-LUCAS dataset.

| Source Label | [Data Type] | Description | Example |
|---------------------------|-------------|------------------------------|--|
| file | [String] | Image file name | 30581810N.png |
| gps_long | [Double] | GPS observation longitude | -4.434544 |
| gps_lat | [Double] | GPS observation latitude | 38.29408 |
| date | [Date] | Date of observation | 04/06/2018 12:00:00 |
| [33] | nuts0 | [String] NUTS 2016 Level 0 | ES |
| | lc1 | [String] Land cover 1 | E20 |
| | lc1_label | [String] Label of land cover | Grassland without tree/shrub cover |
| | lu1 | [String] Land use 1 | U111 |
| | lu1_label | [String] Label of land use | Agriculture (excluding fallow land and kitchen gardens) |
| Sec. 4.1 Classes | [String] | Segmentation classes | Sky, Tree, Grass, Terrain, Stonewall |
| Sec. 4.2 Quality Scores | [Double] | | 4.92578125 |
| Sec. 4.2 Aesthetic Scores | [Double] | | 3.185546875 |

4.1 Data Collection

Image & Basic Information We collect 1,784 scenery images (1600×1200 pixels) from [33] for MM-LUCAS. These images were taken horizontally at the eye-height level across 1,784 different sites in 27 EU countries (see Tab. 3). The sites primarily represent agricultural and rural landscapes to ensure a diverse and comprehensive dataset. This selection aims to capture the variety and complexity of rural environments across Europe, providing a robust basis for precise analyses. Additionally, we selected specific microdata (geographical location, date, LC, LU) from [33], as shown in Tab. 2 (Top part).

Table 3: 27 EU Countries.

AT: Austria, **BE:** Belgium, **BG:** Bulgaria, **CY:** Cyprus, **CZ:** Czech Republic, **DE:** Germany, **DK:** Denmark, **EE:** Estonia, **EL:** Greece, **ES:** Spain, **FI:** Finland, **FR:** France, **HR:** Croatia, **HU:** Hungary, **IE:** Ireland, **IT:** Italy, **LT:** Lithuania, **LU:** Luxembourg, **LV:** Latvia, **NL:** Netherlands, **PL:** Poland, **PT:** Portugal, **RO:** Romania, **SE:** Sweden, **SI:** Slovenia, **SK:** Slovakia, **UK:** United Kingdom

Segmentation Mask Segmentation masks are provided with the same resolution as the corresponding RGB images. After thorough data cleaning, the final segmentation dataset includes 34 classes (see Tab. 4). The original segmentation masks [33] are single-layer grayscale images, where pixel values range from 0 to 33. To enhance intuition and visualization, we further provide color-coded segmentation masks, where each class is represented by a different color.

Table 4: 34 semantic segmentation classes (inherited from [33]) in MM-LUCAS.

'Sky', 'Tree', 'Background', 'Terrain', 'Plant_Bush', 'Flowerfield', 'Earth_Ground', 'Mountain', 'Poles', 'Tower', 'Automobile', 'Grass', 'Path', 'Dense_Woody_Features', 'Flower', 'Cropfield', 'Rail_Transport', 'Traffic_Sign', 'Wall', 'Crop', 'Fruit', 'Field_Margin', 'Road', 'Rock', 'Orchard', 'Waterbodies', 'Animal', 'Stonewall', 'Bridge', 'Lucas_Marker', 'Bark', 'Person', 'Well', 'Building'

4.2 Data Processing

Depth Estimation We adopt the advanced Depth Anything V2-Large model (335.3M parameters) [51], which achieves robust and fine-grained depth predictions by using synthetic images, with high efficiency and accuracy.

Visual Assessment We adopt Q-Align [49], which trained LLMs for visual rating by emulating human discrete-level rating processes. Compared with other similar models that rely on numerical score scaling, Q-Align aligns more with human cognition. We only evaluate image quality and aesthetics in our dataset.

Table 5: Examples of GPT-4o [7] generated question prompts for LandQA.

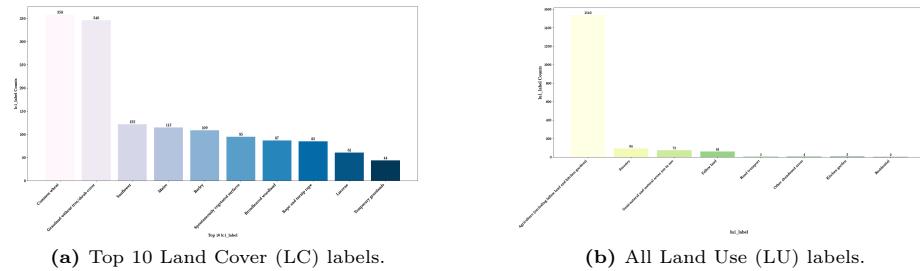
| Topics | Prompts |
|-----------------|---|
| Land Cover | What is the main type of land visible here? How would you describe the surface features in this image? Can you specify the type of land cover in this image? |
| Land Use | What kind of socio-economic activity is taking place in this image? How is the land used in this area? Can you identify the main use of this land? |
| Image Quality | Please evaluate the quality of this photo. How would you describe the quality of this photo? Can you give a rating on the quality of this image? |
| Image Aesthetic | Please provide a rating for the aesthetics of this picture. How do you evaluate the aesthetics of this picture? Can you provide your opinion on the aesthetics of this image? |

Landscape Question Answering (LandQA) Specifically, we utilize GPT-4o [7] to construct LandQA (see *Right* in Fig. 2) based on the LUCAS images and the corresponding annotations, land cover label (*lc1_label*) [33], land use label (*lu1_label*) [33], quality scores and aesthetic scores. The key idea is to query GPT-4o to generate diversity questions as shown in Tab. 5.

For land cover QA and land use QA, we set up multi-choice question-answering with 4 options: {“*images*:”, “*questions*:”, “*options*:”, “*answer*:”}. This format can effectively evaluate the model’s understanding of the land cover and land use annotations by providing distinct choices for each question.

For quality evaluation and aesthetic evaluation, we employ a basic question-answering format. This is structured as: {“*images*:”, “*questions*:”, “*answer*:”}. Note that, we design the standard of aesthetic scores as: “0-1:Bad, 1-2:poor, 2-3:fair, 3-4: good, 4-5: excellent”. In this setup, we suggest the model provide a straightforward answer based on the aesthetic annotations and effectively evaluate the model’s ability to interpret these subjective scores.

4.3 Data Analysis

**Fig. 3:** Distribution of the top 10 *lc1_label* and all *lu1_label* categories.

As shown in Fig. 3a, we present the top 10 land cover (LC) categories in our MM-LUCAS dataset. Tab. 6 list the detailed LC labels, which show a diverse range of agricultural and natural land cover types across the EU regions.

Common wheat is the most frequently occurring. The importance of crops like **wheat, sunflower, maize, and barley** highlights the agricultural focus.

Table 6: Classification of land cover (LC) in MM-LUCAS dataset.

- **A00**-Artificial Land: **A20**-Artificial non-built up areas
- **B00**-Cropland: **B10**-Cereals, **B20**-Root crops, **B30**-Non-permanent industrial crops, **B40**-Dry pulses, vegetables and flowers, **B50**-Fodder crops, **B70**-Permanent crops, **B80**-Other permanent crops
- **C00**-Woodland: **C10**-Broadleaved woodland, **C20**-Coniferous woodland, **C30**-Mixed woodland
- **D00**-Shrubland: **D10**-Shrubland with sparse tree cover, **D20**-Shrubland without tree cover
- **E00**-Grassland: **E10**-Grassland with sparse tree/shrub cover, **E20**-Grassland without tree/shrub cover, **E30**-Spontaneously re-vegetated surfaces
- **F00**-Bare land and lichens/moss: **F10**-Rocks and stones, **F20**-Sand, **F40**-Other bare soil
- **H00**-Wetlands: **H10**-Inland wetlands, **H20**-Coastal wetlands

As shown in Fig. 3b, we present all land use (LU) categories in our MM-LUCAS dataset, which indicate **Agriculture (excluding fallow land and kitchen gardens)**, overwhelmingly the most prevalent. **Forestry** is the second most common LU that is significantly less frequent than agriculture. Tab. 7 list the detailed LU labels, indicating that although agriculture is prevailing, there is still a variety of land use types represented, including forestry, natural areas, and fallow land, providing a diverse view of the landscape.

Table 7: Classification of land use (LU) in MM-LUCAS dataset.

- **U100**-Primary sector: **U110**-Agriculture, **U120**-Forestry
- **U300**-Tertiary sector, transport, utilities & residential: **U310**-Transport, communication networks, storage, protection works, **U370**-Residential
- **U400**-Unused and abandoned areas: **U110**-Abandoned areas, **U120**-Semi-natural and natural areas not in use

To discover how various factors influence the perceived quality and visual appeal of agriculture, we analyze the relationship between the microdata [33] (segmentation classes, geographical information and date) and quality scores (QS) / aesthetic scores (AS). This analysis provides insights that can help optimize agricultural practices, improve crop management, and enhance marketability.

Figure 4 presents the relationship between the number of different segmentation classes (see Tab. 4) and their average AS. Despite the **Orchard** is infrequent, have a relatively high average AS (2.04). Natural elements, like **flower** and **cropfield** have high AS (2.45 and 2.46), showing the importance of contributing to the scenery beauty.

Figure 5 illustrates the significant regional variability of QS and AS across the EU, which helps understand regional strengths and weaknesses, guiding local farmers and agricultural policymakers to focus on best practices that lead to higher quality and more aesthetically pleasing crops. **France, Germany, and**

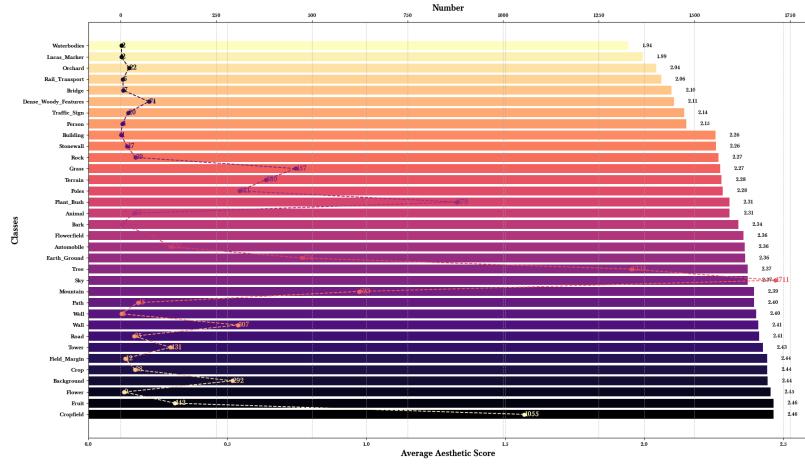


Fig. 4: Average aesthetic score (AS) for different segmentation classes.

the United Kingdom have higher QS. Romania and Bulgaria have a mix of high and low-quality images, which could be due to diverse geographical features and environmental conditions, as shown in Fig. 5 (left). France and parts of Romania and Bulgaria have higher AS, while Finland, Estonia, and Lithuania have lower AS, as shown in Fig. 5 (right).

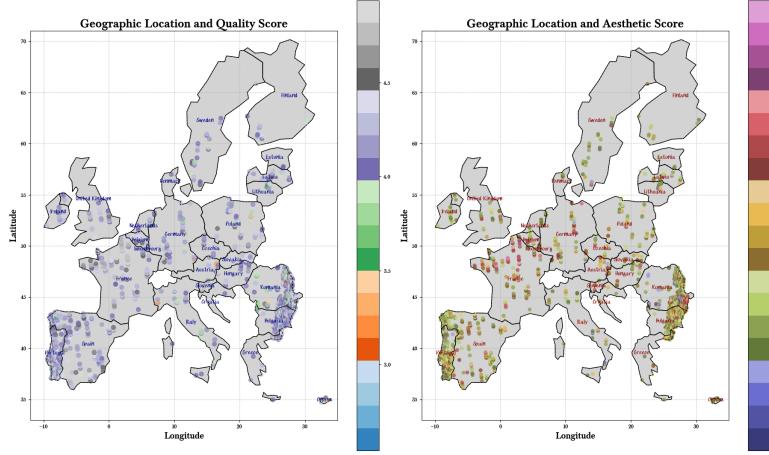


Fig. 5: Spatial distribution of quality score and aesthetic score across EU.

We further analyze seasonal changes in AS for the five most frequent and five least frequent classes, which provides insights into how different agricul-

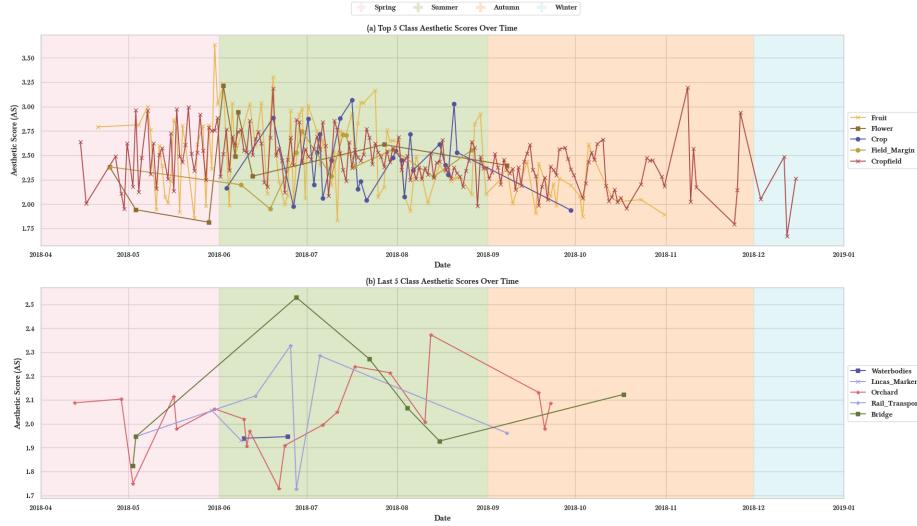


Fig. 6: Seasonal changes in aesthetic score (AS) for the five most frequent classes and five least frequent classes. Zoom in for details.

tural products and related surrounding environments are affected by seasonal changes. As demonstrated in Fig. 6(a), the top 5 classes are highly influenced by natural cycles of growth and harvest, with **spring** and **summer** generally showing higher and more variable scores. Strong seasonal dependency **flowers** and **fruit** show the highest variability, while **field margins** maintain more consistent scores. Compared to more frequent classes, the last 5 classes show lower variability in AS, indicating that their visual appeal is less correlated with seasonal changes. Overall, **summer** generally sees the highest AS due to the peak of natural beauty and crop maturity, while AS in **winter** are consistently lower across both frequent and less frequent classes. This information is necessary for optimizing planting and harvesting times, improving crop management throughout the year, and predicting market trends based on seasonality.

5 Experiments and Analysis

5.1 Mainstream MM-LLMs for Evaluation and Results

We evaluate 2 open-source MM-LLMs (InterVL2-26B [13], and mPLUG-Owl2 [53]) and 3 closed-source MM-LLMs (GPT-4o [7], Gemini-1.0 Pro [6], and Claude 3.5 Sonnet [5]) based on the official implementation sources in the following 2 tasks. Experimental results are shown as: “**Level X: Modal Task Question Models: Results**”. Some details are omitted due to page limitations.

Level 1: T→T Basic Question Answering Q1: Please briefly introduce the common vegetable crops in Europe.

* InterVL2-26B [13]: Tomatoes, cucumbers, peppers, onions, carrots, potatoes, and lettuce...

* mPLUG-Owl2 [53]: Potatoes, onions, carrots, tomatoes, peas, and lettuce.

* Claude 3.5 Sonnet [5]: Potatoes; Tomatoes; Carrots; Onions; Cabbage; Lettuce; Cucumbers; Bell peppers; Peas; Beans; Broccoli and cauliflower; Spinach; Asparagus; Zucchini (courgette); Garlic...

* Gemini-1.0 Pro [6]: Mass Production Leaders: Tomatoes; Onions; Carrots ... Other Widespread Crops: Brassicas; Solanaceae Family ...

* GPT-4o [7]: Tomatoes; Potatoes; Carrots; Onions; Cabbage; Cucumbers; Peppers (Bell and Chili); Spinach; Zucchini (Courgette) ...

Level 3: T+I→T Contextual Scene Analysis Q2: Rate the aesthetics (0-5) of this image and identify key features enhancing the visual appeal.



* InterVL2-26B [13]: **Score:** Composition (4/5), Lighting Conditions (4/5)→4/5. **Key features:** Central tree; Contrast between sky & ground.

* mPLUG-Owl2 [53]: **Score:** 4/5. **Key features:** the large tree in the middle of the field, the lush green grass, and the yellow flowers surrounding the tree.

* Claude 3.5 Sonnet [5]: **Score:** 4.5/5. **Key features:** Focal point; Composition; Color palette; Lighting; Depth perception; Seasonal beauty; Sky drama; Natural framing; Contrast; Symmetry; Rural charm.

* Gemini-1.0 Pro [6]: **Score:** 4/5. **Key features:** Simplicity and Minimalism: (4/5); Centralized Subject: (4/5); Contrasting Colors: (4/5); Rule of Thirds: (4/5); Natural Light: (4/5); Depth of Field: (4/5).

* GPT-4o [7]: **Score:** 4.5/5. **Key features:** Tree as the Focal Point; Color Contrast; Depth and Layers; Natural Symmetry and Balance; Texture and Detail; Light and Shadow; Serene Atmosphere.

5.2 Results Analysis

For **Basic Question Answering**, InterVL2-26B and mPLUG-Owl2 provide concise lists of 6-7 common vegetables, offering straightforward and relevant responses. In contrast, Claude 3.5 Sonnet and GPT-4o present more comprehensive lists, reflecting a broader knowledge. Gemini-1.0 Pro uniquely categorizes crops into mass production leaders and other widespread crops, highlighting the economic significance and scale of different crops.

For **Contextual Scene Analysis**, the five MM-LLMs consistently rate the AS between 4 and 4.5 out of 5. Key visual features include the tree, color contrast, and depth. While InterVL2-26B, mPLUG-Owl2, and Gemini-1.0 Pro focus on the tree and contrast as central elements. Claude 3.5 Sonnet and GPT-4o provide a more comprehensive analysis, especially highlighting seasonal beauty and light, offering a deeper understanding of the scene aesthetic.

6 Discussion

Correlation of MM-LUCAS and AgriBench. Semantic segmentation masks and depth maps play an important role in AgriBench, evaluating the agricultural MM-LLM tasks by offering detailed visual and spatial information. They provide an extra condition for accurately assessing tasks such as *species classification*, *object counting*, and *dense object counting* within complex scenes. They also improve *contextual scene analysis* and *environmental impact prediction* by offering deeper insights into the relationships between objects and their surroundings. Additionally, they support *scene projection*, *scene generation*, and *sustainability recommendations* by ensuring realistic spatial relationships and providing key data for informed decision-making.

Responsible AI. Another important challenge is the requirement for MM-LLMs to be interpretable and explainable, even in advanced models. Responsible AI considers various aspects, such as bias, fairness, transparency, human oversight, *etc.*, which ensures that users can understand how and why a model makes certain decisions, as this is crucial for decision-making and gaining the trust of stakeholders. For instance, in healthcare, understanding the decision-making process of an AI model can help doctors trust and effectively use AI recommendations. Similarly, in agriculture, farmers need clear insights into AI predictions to ensure regulatory compliance and to make informed decisions. This trust ensures that AI systems are effective, trustworthy, and ethical, which is foundational for the widespread implementation and acceptance of AI technologies in various research domains and real-world applications.

Limitations and Future Works. Our initial evaluation is primarily qualitative. We plan to extend our benchmark by including evaluation metrics, from perception (user satisfaction scores and expert reviews) to cognition (accuracy rate and consistency score), to provide a comprehensive evaluation that combines both qualitative and quantitative.

7 Conclusion

This paper introduces AgriBench, the first agriculture benchmark evaluating MM-LLMs across multi-modality and multi-tasks dimensions, supporting a broader range of agriculture scenarios and applications. We further introduce MM-LUCAS, the first multimodal agriculture dataset with multiple annotations. This initial exploration is still a work in progress and represents the first step designed to align with AgriBench. Finally, we compare 5 MM-LLMs on AgriBench and highlight significant insights in specialized domain MM-LLMs for further exploration. AgriBench represents the initial step and a significant contribution to developing agricultural MM-LLMs. We aim to boost the progress of AI technologies specifically for agricultural needs, benefiting both researchers and practitioners.

Acknowledgment. This study was supported by Bundesministerium für Bildung und Forschung (BMBF) project “KI und Citizen Science gestütztes Monitoring von zertifizierten Biodiversitätsprojekten” (16LW0441).

References

1. Eurostat - land cover and use 2018, <https://ec.europa.eu/eurostat/web/lucas/database/2018>, accessed: 2024-06-14
2. Eurostat, land cover and use-time series, <https://ec.europa.eu/eurostat/web/lucas/database/time-series>, accessed: 2024-06-16
3. Eurostat statistics explained, lucas - land use and land cover survey, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=LUCAS_-_Land_use_and_land_cover_survey, Accessed: 2024-06-14
4. Lucas 2018 (land use / cover area frame survey) technical reference document c3 classification (land cover & land use). Tech. rep., European Commission (2018), <https://ec.europa.eu/eurostat/documents/205002/8072634/LUCAS2018-C3-Classification.pdf>
5. Claude 3.5 sonnet (2024), <https://www.anthropic.com/news/clause-3-5-sonnet>, accessed: 2024-07-01
6. Gemini pro (2024), <https://deepmind.google/technologies/gemini/pro/>, accessed: 2024-07-01
7. Hello gpt-4o (2024), <https://openai.com/index/hello-gpt-4o/>, accessed: 2024-07-01
8. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
9. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
10. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023)
11. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
12. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
13. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023)
14. d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H.I., Joebges, C., Lemoine, G., et al.: Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union. Scientific data **7**(1), 352 (2020)
15. Hafeez, A., Husain, M.A., Singh, S., Chauhan, A., Khan, M.T., Kumar, N., Chauhan, A., Soni, S.: Implementation of drone technology for farm monitoring & pesticide spraying: A review. Information processing in Agriculture **10**(2), 192–203 (2023)
16. Han, J., Shi, L., Yang, Q., Huang, K., Zha, Y., Yu, J.: Real-time detection of rice phenology through convolutional neural network using handheld camera images. Precision Agriculture **22**, 154–178 (2021)

17. He, M., Liu, Y., Wu, B., Yuan, J., Wang, Y., Huang, T., Zhao, B.: Efficient multimodal learning from data-centric perspective. arXiv preprint arXiv:2402.11530 (2024)
18. Herrmann, I., Bdolach, E., Montekyo, Y., Rachmilevitch, S., Townsend, P.A., Karnieli, A.: Assessment of maize yield and phenology by drone-mounted super-spectral camera. *Precision Agriculture* **21**, 51–76 (2020)
19. Hu, Z., Yang, P., Jiang, Y., Bai, Z.: Prompting large language model with context and pre-answer for knowledge-based vqa. *Pattern Recognition* **151**, 110399 (2024)
20. Kok, E., Chen, C.: Occluded apples orientation estimator based on deep learning model for robotic harvesting. *Computers and Electronics in Agriculture* **219**, 108781 (2024)
21. Kong, J., Wang, H., Wang, X., Jin, X., Fang, X., Lin, S.: Multi-stream hybrid architecture based on cross-level fusion strategy for fine-grained crop species recognition in precision agriculture. *Computers and Electronics in Agriculture* **185**, 106134 (2021)
22. Kotwal, J., Kashyap, R., Pathan, S.: Agricultural plant diseases identification: From traditional approach to deep learning. *Materials Today: Proceedings* **80**, 344–356 (2023)
23. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al.: Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine* pp. 1–8 (2024)
24. Laso Bayas, J.C., See, L., Bartl, H., Sturm, T., Karner, M., Fraisl, D., Moorthy, I., Busch, M., Van Der Velde, M., Fritz, S.: Crowdsourcing lucas: Citizens generating reference land cover and land use data with a mobile app. *Land* **9**(11), 446 (2020)
25. Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., Li, C.: Llava-next: Stronger llms supercharge multimodal capabilities in the wild (May 2024), <https://llava-v1.github.io/blog/2024-05-10-llava-next-stronger-llms/>
26. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
27. Li, J., Chen, D., Qi, X., Li, Z., Huang, Y., Morris, D., Tan, X.: Label-efficient learning in agriculture: A comprehensive review. *Computers and Electronics in Agriculture* **215**, 108412 (2023)
28. Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al.: What if we recaption billions of web images with llama-3? arXiv preprint arXiv:2406.08478 (2024)
29. Li, X., Yue, C., Liu, X., Zhou, J., Wang, L.: Acwgan-gp for milling tool breakage monitoring with imbalanced data. *Robotics and Computer-Integrated Manufacturing* **85**, 102624 (2024)
30. Li, Z., Guo, R., Li, M., Chen, Y., Li, G.: A review of computer vision technologies for plant phenotyping. *Computers and Electronics in Agriculture* **176**, 105672 (2020)
31. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
32. Maes, J., Teller, A., Erhard, M., Liquete, C., Braat, L., Berry, P., Egoh, B., Puydarrioux, P., Fiorina, C., Santos, F., et al.: Mapping and assessment of ecosystems and their services. An analytical framework for ecosystem assessments under action **5**, 1–58 (2013)

33. Martinez-Sanchez, L., Hufkens, K., Kearsley, E., Naydenov, D., Czúcz, B., van de Velde, M.: Semantic segmentation dataset of land use/cover area frame survey (lucas) rural landscape street view images. *Data in Brief* **54**, 110394 (2024)
34. Nayak, A., Chakraborty, S., Swain, D.K.: Application of smartphone-image processing and transfer learning for rice disease and nutrient deficiency detection. *Smart Agricultural Technology* **4**, 100195 (2023)
35. Ong, P., Teo, K.S., Sia, C.K.: Uav-based weed detection in chinese cabbage using deep learning. *Smart Agricultural Technology* **4**, 100181 (2023)
36. Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., Alewell, C.: Soil erodibility in europe: A high-resolution dataset based on lucas. *Science of the total environment* **479**, 189–200 (2014)
37. Rezaei, M., Diepeveen, D., Laga, H., Jones, M.G., Sohel, F.: Plant disease recognition in a low data scenario using few-shot learning. *Computers and electronics in agriculture* **219**, 108812 (2024)
38. Rotstein, N., Bensaïd, D., Brody, S., Ganz, R., Kimmel, R.: Fusecap: Leveraging large language models for enriched fused image captions. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 5689–5700 (2024)
39. Roy, A.M., Bhaduri, J., Kumar, T., Raj, K.: Wildetect-yolo: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecological Informatics* **75**, 101919 (2023)
40. Seol, J., Kim, J., Son, H.I.: Field evaluations of a deep learning-based intelligent spraying robot with flow control for pear orchards. *Precision Agriculture* **23**(2), 712–732 (2022)
41. Tang, Y., Chen, M., Wang, C., Luo, L., Li, J., Lian, G., Zou, X.: Recognition and localization methods for vision-based fruit picking robots: A review. *Frontiers in Plant Science* **11**, 510 (2020)
42. Teixeira, I., Morais, R., Sousa, J.J., Cunha, A.: Deep learning models for the classification of crops in aerial imagery: A review. *Agriculture* **13**(5), 965 (2023)
43. Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860* (2024)
44. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023)
45. Tziolas, N., Tsakiridis, N., Chabrillat, S., Dematté, J.A., Ben-Dor, E., Gholizadeh, A., Zalidis, G., Van Wesemael, B.: Earth observation data-driven cropland soil monitoring: A review. *Remote Sensing* **13**(21), 4439 (2021)
46. Wang, G., Ge, Y., Ding, X., Kankanhalli, M., Shan, Y.: What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223* (2023)
47. Wang, X., Zhuang, B., Wu, Q.: Modaverse: Efficiently transforming modalities with llms. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26606–26616 (2024)
48. Wang, X., Liu, J.: Vegetable disease detection using an improved yolov8 algorithm in the greenhouse plant environment. *Scientific Reports* **14**(1), 4261 (2024)
49. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090* (2023)
50. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519* (2023)

51. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)
52. Yang, Q., Ye, M., Du, B.: Emollm: Multimodal emotional understanding meets large language models. arXiv preprint arXiv:2406.16442 (2024)
53. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration (2023)
54. Yordanov, M., d'Andrimont, R., Martinez-Sanchez, L., Lemoine, G., Fasbender, D., Van der Velde, M.: Crop identification using deep learning on lucas crop cover photos. Sensors **23**(14), 6298 (2023)
55. Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., Yu, D.: Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601 (2024)
56. Zhao, R., Zhu, Y., Li, Y.: Cla: A self-supervised contrastive learning method for leaf disease identification with domain adaptation. Computers and Electronics in Agriculture **211**, 107967 (2023)