The Knowledge Engineering Review

RESEARCH ARTICLE

Open Access

https://doi.org/10.48130/ker-0026-0010

The Knowledge Engineering Review **2026**, 41: e007

Exploring MLLMs for ultra-fine-grained agricultural classification

Chenhui Wang[1*] and Pengcheng Zhang[2]

[1] *School of Computing, Macquarie University, Sydney, New South Wales, 2109, Australia*
[2] *Centre on Security, Mobile Applications and Cryptography, Singapore Management University, 179873, Singapore*
* Correspondence: chenhui.wang@mq.edu.au (Wang C)

Abstract

Accurate identification of crop cultivars or pest species at subspecies granularity is essential for precision agriculture. Discriminative features are confined to subtle morphological details (leaf venation, surface texture, or reproductive organ structure), producing minimal inter-class visual distance alongside severe label scarcity, together defining the Ultra Fine-Grained Visual Classification (Ultra-FGVC) regime. Neither convolutional nor Vision Transformer advances have resolved this representational gap, suggesting a fundamental encoder bottleneck. Multimodal large language models (MLLMs) are a principled candidate: web-scale training encodes structured biological knowledge absent from image-label supervision, available at inference time without retraining, directly addressing both failure modes. We present a controlled evaluation of four state-of-the-art MLLMs across three Ultra-FGVC benchmarks, examining baseline few-shot accuracy, label semantics contribution, and prompt-design sensitivity under chain-of-thought decomposition, contrastive elimination, and few-shot in-context learning. Three fundamental failure modes limit MLLM deployment in Ultra-FGVC: the representational ceiling of CLIP-based visual encoders, which widens the accuracy gap as category count grows; output generation unreliability under structured-output requirements; and systematic confidence miscalibration that precludes raw scores from serving as deployment rejection thresholds. For each, we characterise its origin and propose remedies, including richer visual backbones, constrained decoding, and post-hoc calibration to advance MLLMs toward operational viability in precision agriculture.

Citation: Wang C, Zhang P. 2026. Exploring MLLMs for ultra-fine-grained agricultural classification. *The Knowledge Engineering Review* 41: e007 https://doi.org/10.48130/ker-0026-0010

1 Introduction

The proliferation of high-resolution imaging technology across modern agriculture has enabled the routine acquisition of large quantities of crop and animal imagery at field scale. Unlike general object recognition, species-level agricultural classification frequently demands discrimination at the subcategory level, a challenge formally characterised as Ultra Fine-Grained Visual Classification (Ultra-FGVC)[1], where inter-class differences are confined to highly localised morphological features such as leaf venation patterns, surface texture, or reproductive organ morphology.

The economic stakes of misclassification at this granularity are substantial: pathogens and pests cause global yield reductions for major food crops estimated at 20%–30%[2], and accurate subspecies identification is a prerequisite for targeted interventions, resistance monitoring, and precision agriculture workflows that minimise chemical inputs[2]. Erroneous weed identification can result in ineffective herbicide application[3,4], while failure to distinguish morphologically similar pest species may lead to inappropriate pesticide selection and consequent yield loss[5,6]. Such identifications have historically required trained entomologists or plant taxonomists whose availability does not scale with the volume of field imagery now routinely captured by UAVs and IoT-connected sensors[7,8].

The difficulty of automated Ultra-FGVC is twofold. First, discriminative features are highly localised and morphologically subtle: differences may manifest as minor variations in antennal segmentation, leaf margin serration, or seed coat patterning that occupy only a small fraction of image pixels, as illustrated in Fig. 1. Second, as classification granularity descends from species to cultivar level, the number of available labelled images per category decreases sharply[9], creating a data scarcity problem that undermines standard supervised learning. Together, these constraints define a regime where both convolutional and Vision Transformer architectures plateau well below operationally useful accuracy thresholds, as demonstrated on early Ultra-FGVC benchmarks[1].

Multimodal Large Language Models (MLLMs)[10,11] offer a fundamentally new avenue for addressing these limitations. Trained on web-scale corpora, MLLMs encode substantial structured knowledge about biological entities including morphology, ecology, phenology, and diagnostic attributes that is entirely absent from image-label supervision. This parametric knowledge is available at inference time without task-specific retraining, making it particularly attractive under the data-scarce conditions characteristic of Ultra-FGVC.

Despite this potential, MLLM performance on Ultra-FGVC remains largely unexplored. Existing multimodal benchmarks (MMBench[12], SEED-Bench[13], and MMMU[14]) evaluate visual reasoning at coarse to mid-level granularity, typically ImageNet-scale species-level discrimination. Agricultural vision benchmarks similarly operate at disease category or pest group level rather than cultivar or subspecies level. The few studies that do engage with standard FGVC granularity, such as discriminating aircraft variants[15] or car models[16] already document measurable accuracy shortfalls in MLLM performance relative to fine-tuned supervised models, even with careful prompt engineering[17,18], pointing to a fundamental visual encoder limitation[19] that the present work is the first to characterise at Ultra-FGVC scale. The transition from species to subspecies classification collapses inter-class feature distances and removes the coarse semantic anchors upon which MLLM visual encoders rely, yet no systematic study has characterised how this transition degrades model performance.

To address this gap, we present a systematic evaluation of state-of-the-art MLLMs: Kimi-K2.5[20], GPT-5.4[21], Gemini-3-Flash-Preview[22], Qwen3.5-397B[23] on Ultra-FGVC tasks at $K \in \{3, 10\}$, where K denotes the number of candidate categories presented at inference time, corresponding to operationally meaningful deployment scenarios in precision agriculture. Each experimental component is designed to isolate a distinct hypothesis: whether MLLMs can serve as viable retraining-free alternatives to supervised methods at these

https://doi.org/10.48130/ker-0026-0010

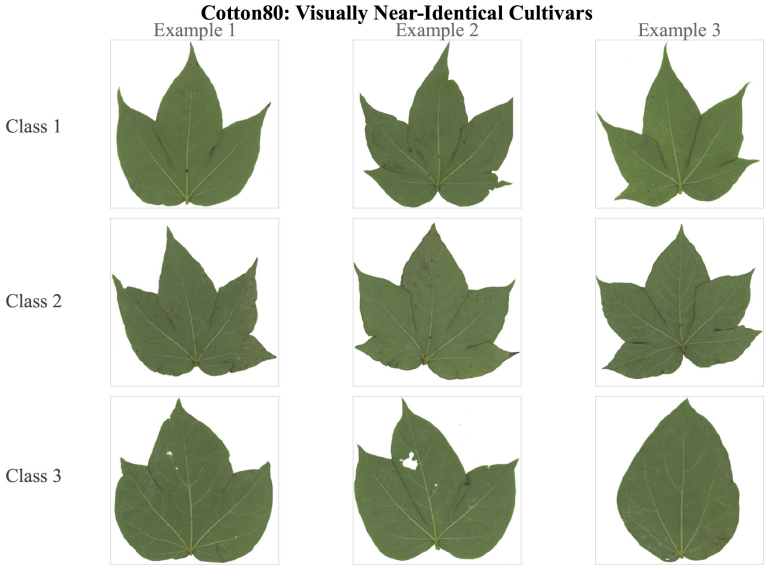

Fig. 1 Representative images from the Cotton80 cultivar classes (three images per class)[1]. Despite belonging to distinct cultivar categories, samples share near-identical macro-morphology: a defining characteristic of Ultra-FGVC that limits even state-of-the-art supervised methods to below 60% accuracy at K = 50. The figure is original and created by Chenhui Wang. © 2026 Chenhui Wang.

scales; whether accuracy at this granularity is driven by visual exemplars or by parametric knowledge activated by category names; and whether prompt engineering can materially close the gap to supervised performance, or whether the bottleneck lies upstream in the visual encoder itself.

We evaluate few-shot classification[10], where we provide a few reference images alongside structured prompts, as well as classification augmented with semantic category names, and a controlled ablation of six prompt design strategies including direct classification, chain-of-thought reasoning, and contrastive attribute prompting. A preliminary three-round evaluation at K = 50 was conducted for GPT-5.4; the remaining three models were subject to context-window and per-request image-count constraints at that scale, and systematic K = 50 results therefore remain an open empirical question. The main contributions are:

1. A systematic empirical evaluation of four state-of-the-art MLLMs on Ultra-FGVC benchmarks at $K \in \{3, 10\}$, establishing the first controlled baselines for MLLM few-shot ICL performance at this granularity. MLLMs exceed human expert accuracy at K = 3 but fall 12%–34% below fine-tuned supervised methods at K = 10, delineating the practical boundary of current MLLM capability.
2. A controlled blind-vs-named label experiment across twelve model–K–dataset combinations, providing the first statistical examination of whether label semantics contribute to Ultra-FGVC accuracy beyond visual exemplars. No condition yields a significant gain from name disclosure after Bonferroni correction (m = 12), though the K = 9 conditions remain underpowered and results should be treated as inconclusive rather than as evidence of visual primacy.
3. A ten-round, six-variant prompt ablation study on two model backends, revealing strongly model-dependent prompt sensitivity with sign-reversed effects across all four tested design dimensions. These findings caution against adopting prompt strategies validated on a single architecture as transferable defaults.
4. A characterisation of three fundamental failure modes limiting MLLM deployment in Ultra-FGVC—the visual encoder resolution, output generation unreliability, and confidence miscalibration—together with concrete technical directions (richer visual backbones, constrained decoding, and post-hoc calibration) well positioned to advance MLLMs toward operational viability.

2 Related work

2.1 Ultra fine-grained visual classification

Since the introduction of CUB-200-2011[24], FGVC has evolved into a well-defined research area. Early benchmarks, including FGVC-Aircraft[15] and Stanford Cars[16], established the task as discrimination among visually similar subcategories within a superordinate class (e.g., distinguishing a Boeing 737-700 from other 737 variants), driving progress towards modern ViT-based models[25] such as SM-ViT[26] and Cross-Layer Feature Fusion ViT[27], both of which substantially outperform CNN baselines including ResNet[28] and AlexNet[29].

Ultra-FGVC is qualitatively harder: distinguishing traits are highly localised and morphologically subtle; the expert knowledge required for their reliable identification contributes to data scarcity; and labelled images per category decrease sharply as granularity descends from species to cultivar level[9].

The first dedicated Ultra-FGVC dataset was established by Yu et al.[1], who released a large-scale leaf dataset across cotton and soy with baseline CNN evaluations. Even state-of-the-art supervised methods struggle: the CNN-based architecture of Pan et al.[30] and the ResNet-50/ViT hybrid of Yu et al. achieve top-1 accuracies of 62.08% and 60.42% on Cotton80, and 49.67% and 56.17% on SoyLocal. The more recent pure-Transformer CLE-ViT achieves similarly modest results[31], suggesting that the bottleneck is representational rather than architectural.

Beyond static recognition, real-world agricultural deployment introduces a temporal dimension: new cultivars are continually registered, and crops must be recognised across successive reproductive stages. Zhang et al.[32] formalise this as continual ultra-fine-grained visual recognition (C-UFG), decomposing the problem into class-incremental learning of new cultivars (VC-UFG) and domain-incremental recognition of the same cultivar across growth stages (HC-UFG). Their analysis of pre-trained model-based continual learning methods on large-scale soy benchmarks reveals that inter-task prompt interference and classifier recency bias emerge as the dominant failure modes in the continual regime, compounding the representational limitations already observed in static Ultra-FGVC. The core difficulty in this field thus lies not only in visual discrimination within a fixed

category set, but in maintaining that discrimination robustly as the task distribution evolves.

2.2 Multimodal large language models

MLLMs extend language models to operate jointly over images and text by coupling a CLIP-style vision encoder[33] with a language model backbone via a lightweight cross-modal connector[34], as shown in Fig. 2. Rather than using a fixed classification head, MLLMs treat recognition as instruction-following, enabling zero-shot and few-shot classification across open-ended label spaces within a single generative framework.

Flamingo[10] demonstrated few-shot visual learning from interleaved image-text sequences via gated cross-attention, achieving state-of-the-art performance across 16 benchmarks. BLIP-2[35] introduced the Q-Former, achieving stronger zero-shot VQA performance than Flamingo-80B with 54X fewer trainable parameters. InstructBLIP[36] extended this with instruction-aware query transformers across 26 datasets. At the proprietary frontier, GPT-4V[37], Gemini 1.5[38], and Claude 3[39] serve as key reference points for zero-shot visual reasoning.

MLLMs exhibit strong classification ability when prompted effectively: LLaVA-1.5 achieves 85% zero-shot accuracy on MNIST and 79% on skin-lesion classification without fine-tuning[17], and GPT-4 with generated textual descriptions rivals EVA-CLIP-ViT-E across 16 benchmarks[18]. Domain-adapted variants amplify performance further: LLaVA-Med[40] improves biomedical VQA by approximately 30 percentage points after instruction tuning, and Co-LLaVA[41] exceeds 85% on remote sensing scene classification. Persistent limitations include insufficient visual discriminability in CLIP-based encoders for fine-grained recognition[19], object hallucination[42], and the fact that ImageNet accuracy still lags behind specialised contrastive models without careful prompt engineering[17,18].

2.3 Prompting strategies for visual classification

Prompt engineering (the systematic design of natural language instructions supplied at inference time without parameter modification) directly governs how an MLLM frames a recognition task, which features it attends to, and how it produces output, as shown in Fig. 3. Prompt design has been shown to produce accuracy swings of tens of

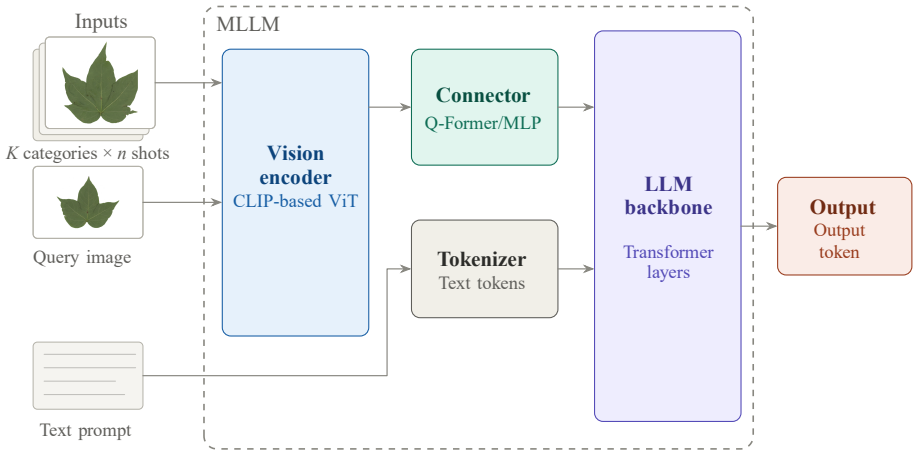

Fig. 2 Base architecture of a multimodal large language model (MLLM) for few-shot in-context learning. Reference images [1] (K categories, n shots each) and an unlabelled query image are encoded by a CLIP-based vision encoder[33], whose patch features are projected into the language model's token space by a lightweight cross-modal connector (Q-Former[35] or MLP bridge[11]). A text prompt encoding the task instruction and category identifiers is tokenised in parallel. Both token streams are consumed jointly by the LLM backbone, which generates a structured output containing the predicted category label. In this work, category identifiers are numeric only (Task 1) or real species names (Task 2); the backbone architecture is otherwise identical across conditions. The figure is original and created by Chenhui Wang. © 2026 Chenhui Wang.

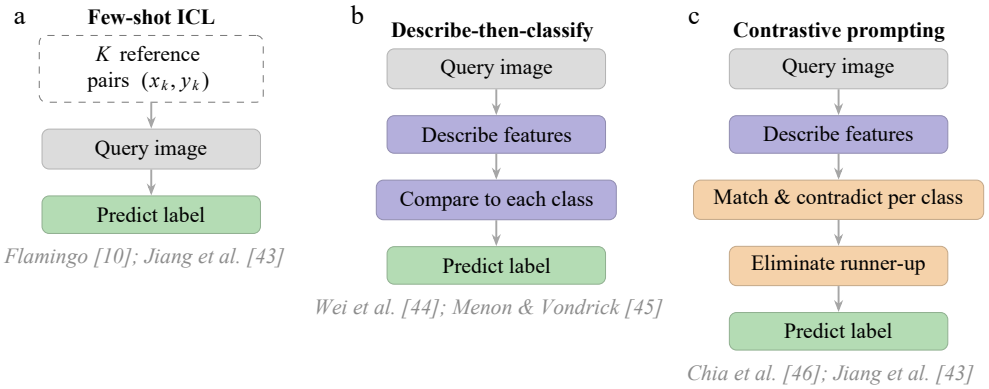

Fig. 3 Structural comparison of the three prompting paradigms reviewed in this section. (a) Few-shot ICL prepends K labelled reference image-label pairs $(\mathbf{x}_k, y_k)$ as in-context demonstrations before the query, enabling task adaptation without gradient updates[10,43]. (b) Describe-then-classify applies chain-of-thought decomposition: the model first articulates observable visual features, then compares them against candidate classes before committing to a label, reducing confirmation bias[44,45]. (c) Contrastive prompting extends (b) by requiring explicit enumeration of matching and contradicting evidence per class and a justified elimination of the runner-up[46,47]. Node shading: *dashed grey* = reference demonstrations; *solid grey* = query image; *blue* = chain-of-thought step; *orange* = contrastive reasoning step; *green* = predicted output. The figure is original and created by Chenhui Wang. © 2026 Chenhui Wang.

https://doi.org/10.48130/ker-0026-0010

percentage points on identical model weights and datasets[44], making it a primary engineering variable in the MLLM classification pipeline.

2.3.1 Chain-of-thought and describe-then-classify

Chain-of-thought (CoT) prompting[44] elicits step-by-step intermediate reasoning by including worked demonstration examples, decomposing complex decisions into traceable steps. Zhang et al.[48] extended this to the multimodal setting, showing that a two-stage rationale-then-answer pipeline enabled a sub-1B-parameter model to surpass GPT-3.5 on multimodal reasoning benchmarks while substantially reducing hallucination.

The most impactful variant for visual classification is the *describe-then-classify* paradigm, in which the model first generates a verbal description of the observable visual evidence before committing to a category label. Menon & Vondrick[45] formalised this by querying an LLM for discriminative per-class descriptors and aggregating CLIP similarity scores over those descriptors, yielding consistent gains on zero-shot benchmarks. Pratt et al.[49] showed that automatically generating category-specific descriptive prompts via an LLM consistently outperforms hand-crafted CLIP templates across diverse tasks. FineR[50] further instantiated this as a training-free framework that extracts part-level visual attributes via VQA, reasons over them with an LLM's world knowledge, and constructs multimodal classifiers surpassing state-of-the-art FGVC models on multiple benchmarks. Most directly relevant to the agricultural domain, ChatLeafDisease[51] applied structured CoT to crop disease classification, with disease-description scoring rules boosting GPT-4o accuracy from 45.9% to 88.9%; removing those rules caused accuracy to collapse to 51.8%.

2.3.2 Few-shot in-context learning

In-context learning (ICL) allows an MLLM to adapt to a new task at inference time by conditioning on interleaved image-label demonstrations without gradient updates. Flamingo[10] established this paradigm, showing that a handful of reference image-text pairs sufficed to match task-specific fine-tuned models across wide-ranging benchmarks. Jiang et al.[43] subsequently showed that classification accuracy scales log-linearly with demonstration count up to 2,000 shots for frontier models such as GPT-4o and Gemini 1.5 Pro, with this benefit absent in open-weight models.

Beyond shot count, demonstration selection and ordering are also consequential. Lu et al.[52] showed that permuting the order of otherwise identical few-shot examples can shift accuracy from near state-of-the-art to near chance. Zhang et al.[53] showed that semantically similar examples retrieved via nearest-neighbour search consistently outperform random sampling. Min et al.[54] found that ground-truth labels within demonstrations contribute less to ICL performance than the exposed label space, input distribution, and prompt format, implying that reference examples function primarily as task-specification signals. Ferber et al.[55] demonstrated that kNN-selected histopathology images brought GPT-4V to parity with specialist networks on cancer classification, a finding with direct implications for data-scarce agricultural classification.

2.3.3 Contrastive prompting

Contrastive prompting instructs the model to reason about categories *relative to one another*, identifying distinguishing features between visually similar classes. Chia et al.[46] proposed Contrastive CoT, augmenting demonstrations with paired invalid reasoning examples. Yao et al.[56] showed that a lightweight contrastive instruction, prompting consideration of both a correct and a plausible incorrect answer, raised GPT-4 accuracy on GSM8K from 35.9% to 88.8%, integrating seamlessly with existing CoT pipelines.

In the visual domain, Jiang et al.'s Discrepancy-aware Text Prompt (DaTP)[47] incorporates explicit subcategory-discriminative language into CLIP-based fine-grained prompts, highlighting the features that set each category apart from its closest neighbours. This reflects the view that fine-grained recognition is inherently a task of *relative* discrimination, and directly motivates the contrastive priming strategy adopted in the present work.

3 Benchmarks and experimental framework

3.1 Datasets and evaluation protocols

We evaluate on three Ultra-FGVC benchmarks spanning agricultural cultivar classification and ornithological species discrimination. All train/test splits are generated deterministically prior to inference. Following Yu et al.[1], for each dataset K categories are sampled without replacement per round using seed $s_r = s_{r-1} + 1$ with $s_0 = 42$, making any round exactly reproducible from its index.

SoyLocal is a subset of the UFG image dataset[1], containing 1,200 leaf images across 200 soybean cultivar categories with six images per category, acquired under controlled laboratory conditions with labels assigned from a genetic resource bank. The primary challenge is an extremely small inter-class variation paired with large intra-class variation: human breeding experts achieve only 63.89% top-1 accuracy on three-category evaluation and 12.84% on 50-category classification[1]. For Cotton80 and SoyLocal, category labels are integer identifiers from a shared annotation file; reference and query images come from non-overlapping dataset splits.

Cotton80 is a second subset of the same UFG dataset[1], containing 480 cotton leaf images across 80 cultivar categories. The dominant challenge is a severe self-overlapping problem arising from the palm-like morphology shared across categories, suppressing even the best supervised methods to below 60% accuracy on 50-way evaluation.

CUB-Crows is a curated Ultra-FGVC subset of CUB-200-2011 corresponding to the crow subset identified by Lu et al.[57], comprising nine visually confusable categories spanning four core crow species and five morphologically similar species that trained models systematically confuse with crows. State-of-the-art supervised methods achieve substantially lower accuracy on this subset than on the full CUB benchmark[57]. For CUB-Crows, where no pre-existing split is available, each category's images are sorted lexicographically and divided 50/50 into train and test pools; lexicographic ordering was validated to show no systematic correlation with acquisition metadata. Directory names beginning with `like_` or `similar_to_` are tagged as *confuser* categories for post-hoc stratified analysis.

CUB-Crows is included because its ecological confusion structure (visually near-identical corvid species that supervised models systematically confuse[57]) closely mirrors cultivar-level confusion in the crop datasets, while its real species names (e.g. *American Crow*, *Fish Crow*) provide a controlled testbed for Task 2's label-semantics experiment. In agricultural deployment, the analogous question is whether naming a pest species (e.g. *Helicoverpa armigera* vs *H. zea*) helps an MLLM beyond visual exemplars alone. All agricultural inferences drawn from CUB-Crows results are made explicit where stated.

CUB-Cormorant is a three-category subset of CUB-200-2011 comprising three morphologically similar cormorant species (Brandt's Cormorant, Red-faced Cormorant, and Pelagic Cormorant). It is used exclusively in Task 2. The dataset is included because standard $K = 3$ CUB-Crows conditions, where most models achieve 84%–92% accuracy, are subject to accuracy ceiling effects that suppress any

detectable semantic benefit regardless of whether one exists. CUB-Cormorant, where models achieve only 67%–74% under the blind condition, provides a $K = 3$ comparison free of ceiling effects. The same 50/50 lexicographic train/test split protocol applied to CUB-Crows is used here.

Split protocol. Two files are produced per configuration: `splits_simple_K.txt`, read directly by the inference pipeline, encoding role, path, and string category label per line (ground truth is therefore embedded in the file, requiring no separate annotation lookup); and `splits_detailed_K.txt` documenting seed, category roles, and per-category image lists for auditability. We evaluated $K \in \{3, 10\}$ for Cotton80 and SoyLocal, and $K \in \{3, 9\}$ for CUB-Crows ($K = 9$ exhausts the available categories), and $K = 3$ for CUB-Cormorant ($K = 3$ exhausts the available categories). The values $K = 3$ and $K = 10$ were selected to match two of the evaluation conditions of the benchmark paper[1], enabling direct comparison with its reported baselines at those levels.

Scope limitation. The $K = 50$ condition was not evaluated for Gemini-3-Flash-Preview, Kimi-K2.5, or Qwen3.5-397B, as supplying 150 reference images (50×3) exceeds their context-window or per-request image-count limits. For GPT-5.4, a preliminary evaluation comprising $N = 3$ rounds was feasible and is reported in Section 4.1. Results at $K \in \{3, 10\}$ correspond to the deployment-relevant triage and screening scenarios described in Section 1. The $K = 3$ condition should be interpreted with caution given the 33.3% random baseline, while $K = 10$ is the more demanding and informative result.

3.2 Models and experiment tasks

We selected four state-of-the-art MLLMs spanning different providers and architectural lineages based on the general vision capability analysis by arena.ai[58], as summarised in Table 1. All models receive identical prompts and reference images with no model-specific calibration beyond API-mandated formatting differences. Exact model version strings as used in API calls are reported in Table 1; temperature was set to the provider default, and no model-specific system prompts beyond the task instruction were applied. Because these models are accessed via commercial APIs whose weights and training data are not publicly disclosed, results should be interpreted as characterising the deployed versions at the time of evaluation rather than fixed architectural configurations.

We formulate Ultra-FGVC as a few-shot in-context learning problem for MLLMs. Given K categories, each with n reference images, and a set of unlabelled query images, the model assigns each query to one of the K categories by visual comparison against the provided references, without gradient-based adaptation or fine-tuning. This setup directly mirrors precision agriculture deployment constraints, where labelled data is scarce, making retraining impractical.

The evaluation is structured as three complementary tasks that decompose MLLM Ultra-FGVC performance into orthogonal factors: Task 1 establishes a semantics-free baseline that isolates raw visual encoding capacity, providing the anchor against which the remaining tasks are interpreted; Task 2 quantifies whether label semantics contribute accuracy beyond what visual exemplars alone provide, with Task 1's blind protocol as the reference; and Task 3 measures prompt-design sensitivity with model, dataset, and reference images held fixed, so that any accuracy differences are attributable solely to prompt structure. This decomposition is necessary because the three factors are confounded in typical MLLM evaluations; disentangling them is essential to identify which failure modes require architectural remedies versus engineering interventions.

Table 1. MLLMs evaluated in this study. Version strings are exact API identifiers used in all experiments.

Model (API version string)	Provider
`gemini-3-flash-preview`[22]	Google DeepMind
`gpt-5.4-2026-03-05`[21]	OpenAI
`qwen3.5-397b-a17b`[23]	Alibaba
`kimi-k2.5`[20]	Moonshot AI

Task 1: Baseline MLLM Performance on Ultra-FGVC. We evaluate all four models on SoyLocal and Cotton80 under the unified few-shot ICL protocol. In each round, the model receives three reference images per category alongside a structured prompt and classifies held-out query images into one of the K categories. We target $n = 20$ rounds at $K = 3$ and $n = 10$ at $K = 10$. GPT-5.4 was additionally evaluated at $K = 50$ for $n = 3$ preliminary rounds; the remaining models could not be evaluated at $K = 50$ due to context-window and image-count constraints. All categories are identified by numeric labels only $(1, 2, \ldots, K)$, conveying no semantic information about cultivar names and thereby isolating pure visual discrimination ability from any benefit of label semantics, which is investigated in Task 2.

Task 2: Effect of Category Name Disclosure. Task 2 investigates whether disclosing real species names (rather than the numeric identifiers of Task 1) improves accuracy, or whether MLLMs rely primarily on visual exemplars regardless of label semantics. We conduct a controlled two-pass evaluation on CUB-Crows and CUB-Cormorant, uploading reference and query images once per round and reusing them across both passes to eliminate API upload cost as a confound. The passes are structurally identical, differing only in label presentation. Within each round, the blind pass is administered before the named pass; pass order is fixed across rounds to allow cost-efficient image reuse, and any carryover effects within a round are assumed negligible given the stateless nature of the API inference calls.

- **Blind pass.** Categories are assigned anonymous numeric identifiers $\{1, 2, \ldots, K\}$, matching the protocol of Task 1.
- **Named pass.** Categories are presented with their real string identifiers (e.g. *American Crow*, *Fish Crow*), allowing the model to draw on any world knowledge associated with those names.

Query images are anonymised and randomly ordered within each pass. The primary quantity of interest is the per-round accuracy delta:

$$\Delta_r = \mathrm{acc}_r^{(\mathrm{named})} - \mathrm{acc}_r^{(\mathrm{blind})},$$

aggregated over $N = 20$ rounds for $K = 3$ conditions and $N = 10$ for the $K = 9$ CUB-Crows condition; CUB-Cormorant ($K = 3$) was evaluated over $N = 20$ rounds for all four models. This matches the number of rounds used in Task 1. A consistently positive Δ_r would indicate that label semantics provide a meaningful signal beyond visual similarity; values near zero would indicate visual-exemplar-driven behaviour.

Task 3: Prompt Ablation Study. Task 3 quantifies the contribution of prompt engineering to Ultra-FGVC performance by comparing six variants—Baseline, Direct, Instructions-Last, Contrastive, Interleaved, and Minimal—with model, dataset, category subset, and reference images held fixed, so any accuracy differences are attributable solely to prompt structure. All variants share the few-shot ICL scaffold (three reference images per category) following Flamingo[10], and differ along three dimensions: (i) whether CoT decomposition is applied; (ii) whether contrastive elimination is required; and (iii) whether instructions are co-located with the images they govern. Table 2 summarises these choices.

Scope of Task 3. The ablation is conducted on SoyLocal only, using two of the four evaluated backends (Gemini-3-Flash-Preview and Kimi-K2.5). GPT-5.4 and Qwen3.5-397B are not included due to API cost constraints. The choice of dataset, SoyLocal, is due to the

https://doi.org/10.48130/ker-0026-0010

results from Task 1 as shown in Section 4.1. SoyLocal at $K = 10$ puts MLLMs in the 45%–52% range, while Cotton80's accuracy at $K = 10$ is 33%–39%, dangerously close to the 10% random-chance baseline. The study is most informative when there is room for the prompt effects to manifest in both directions.

The **Baseline** implements describe-then-classify[45]: the model first articulates the query image's fine-grained visual features without reference to any category name, then compares those features against each category's references before deciding. This two-stage decomposition reduces confirmation bias by separating the descriptive and discriminative stages[48], and serves as the primary reference point.

The **Contrastive** variant extends the Baseline by requiring the model to enumerate both matching and contradicting features for every candidate category before reaching a verdict, and to name a runner-up with a justification for its elimination. This operationalises contrastive priming[47,46]: that fine-grained recognition is inherently a relative judgment, and surfacing contradictions reduces the rate at which plausible but incorrect categories escape elimination.

The **Interleaved** variant retains CoT structure but co-locates per-category discriminative cues with their corresponding reference image block rather than using a global preamble, testing whether local instruction placement improves feature binding.

The **Instructions-Last** variant presents all images before any task instruction, testing whether deferring instruction onset alters the model's initial attention allocation and produces different feature extraction behaviour compared to instruction-first designs.

The **Direct** variant removes CoT entirely, presenting a concise classification instruction with no intermediate reasoning steps, isolating the contribution of the two-stage decomposition by holding all other elements constant relative to the Baseline.

The **Minimal** variant reduces the prompt to its absolute minimum: a category list, reference images, and query images with no discriminative guidance, establishing the lower bound of prompt engineering contribution.

All six variants share an identical JSON output schema with six required fields, ensuring parsing differences cannot confound accuracy comparisons. Each variant is evaluated over $N_{\text{rounds}} = 10$ independent rounds on SoyLocal for both model backends, with accuracy reported as mean ± standard deviation across rounds. The actual prompts are shown in Fig. 4.

3.3 Inference protocol

Prompt structure. The prompt presents three sequential components, illustrated in Fig. 5: (i) a task preamble naming the K categories and directing attention to subtle discriminative cues; (ii) per-category reference image blocks, each prefaced by a category header; and (iii) query image blocks followed by a three-step classification instruction and a mandatory JSON output schema. Category identifiers are numeric only in Tasks 1 and 3; real species names replace numeric identifiers in the named condition of Task 2.

Structured output. Each model returns a strictly formatted JSON array. Every record must contain `filename`, `observed_features` three to five items), `category`, `runner_up`, `confidence` (float $\in [0, 1]$), and `reasoning`. The parser uses bracket-depth tracking and regex-based fence extraction to handle malformed outputs robustly, raising an exception if any required key is absent.

Filename resolution. After parsing, anonymised names are resolved to original filenames via the `name_mapping` dictionary before accuracy evaluation.

4 Performance analysis

4.1 Task 1: Baseline MLLM performance on ultra-FGVC

Table 3 reports top-1 accuracy for all four MLLMs under the blind numeric-label protocol alongside the supervised baselines of Yu et al.[1].

At $K = 3$, where the random-guessing baseline is 33.3%, cross-model mean accuracy reaches 82.4% on SoyLocal and 63.5% on Cotton80. With only three candidate categories, however, the discrimination task is substantially easier than the benchmark's primary evaluation condition, and even simple nearest-neighbour matching achieves high accuracy. The modest gap to supervised baselines at $K = 3$ therefore does not constitute a meaningful comparison of MLLM versus supervised performance.

Comparison with human experts. At $K = 3$, all four MLLMs substantially exceed human expert accuracy on SoyLocal (cross-model mean 82.4% vs 63.9%). On Cotton80, three of four models individually exceed human experts (Gemini 67.8%, GPT-5.4 71.7%, Kimi 67.2% vs 66.7%); the cross-model mean (63.5%) falls below the human expert level, pulled down by Qwen3.5-397B's substantially lower performance (47.2%). This reversal on SoyLocal is notable: domain experts whose labelling effort underlies the benchmark are outperformed by general-purpose MLLMs on the three-category task. The advantage does not persist at $K = 10$, however: cross-model MLLM means drop to 47.3% on SoyLocal and 35.3% on Cotton80, while human experts retain 32.5% and 63.3% respectively. On Cotton80 at $K = 10$, human experts substantially outperform all four MLLMs. This pattern is consistent with MLLMs benefiting from strong generic visual priors at very low category counts, while at higher category counts human experts' specialised morphological knowledge provides a decisive advantage.

The more informative condition is $K = 10$, where all four models fall substantially below fine-tuned supervised methods. The absolute gap widens sharply: DCL retains 63.3% and 73.3% at $K = 10$, while the best MLLM (GPT-5.4) reaches only 51.7% and 39.3%, deficits of 12 and 34 percentage points respectively. SoyLocal is consistently easier than Cotton80, a difference attributable to greater

Table 2. Summary of the six prompt variants.

Variant	CoT	Contrastive	Instruction placement	Role
Baseline	✓	–	Preamble (before images)	Primary reference; describe-then-classify[45]
Contrastive	✓	✓	Preamble (before images)	Adds per-category match/contradiction elimination[46,47]
Interleaved	✓	–	Co-located per category	Tests local vs global instruction placement
Instructions-Last	–	–	Appended (after all images)	Tests effect of deferred instructions on attention
Direct	–	–	Preamble (before images)	Isolates CoT contribution; no reasoning steps
Minimal	–	–	Inline with images	Lower bound; no discriminative guidance

CoT = chain-of-thought describe-then-classify decomposition; Contrastive = per-category match/contradiction elimination; Instruction placement = where task instructions appear relative to the images.

macro-morphological variation between soybean accessions compared to the subtler textural differences that distinguish cotton cultivars.

GPT-5.4 leads in three of four MLLM conditions; the sole exception is SoyLocal at $K = 3$, where Gemini achieves the highest top-1 accuracy (85.6% vs 81.7%). All models exhibit notable overconfidence, reporting mean confidence scores of 0.74–0.91 regardless of accuracy level.

Preliminary results at $K = 50$. A three-round preliminary evaluation of GPT-5.4 at $K = 50$ yields top-1 accuracy of 20.4% ± 3.8% on SoyLocal and 12.4% ± 5.1% on Cotton80. Both figures are well below all fine-tuned CNN baselines (AlexNet: 33.3% and 24.7%; ResNet-50: 62.0% and 52.7%) and, on Cotton80, fall below human expert performance (14.0%). On SoyLocal, GPT-5.4 at $K = 50$ slightly exceeds the human expert (12.8%), which may partly reflect the availability of semantic priors at higher category counts; however, three rounds are far too few to draw reliable conclusions. These preliminary results nonetheless support the visual encoder bottleneck hypothesis at benchmark scale: even the highest-performing MLLM degrades sharply from $K = 10$ to $K = 50$ in a manner consistent with CLIP-based encoder saturation rather than a failure of prompting or in-context learning.

Scaling behaviour. The degradation from $K = 3$ to $K = 10$ follows a consistent pattern across all four models and both datasets, with the gap to supervised methods widening at every step. The preliminary $K = 50$ data for GPT-5.4 are consistent with continued or accelerating degradation, though this must be confirmed with more rounds and additional models. At category counts representative of field-level cultivar triage ($K \leq 10$), MLLMs provide operationally useful accuracy, but they are not yet viable substitutes for fine-tuned supervised models in broader screening applications.

4.2 Task 2: Effect of category name disclosure

Statistical significance was assessed using a one-sample paired t-test applied to the per-round accuracy difference $\delta_r = A_{\mathrm{N}}^{(r)} - A_{\mathrm{B}}^{(r)}$, testing $H_0: \mu_\Delta = 0$ against a two-sided alternative. For $K = 9$ conditions ($n = 10$ rounds), where the Shapiro–Wilk test has insufficient power to confirm normality, results were verified with a Wilcoxon signed-rank test; all qualitative conclusions were unchanged. Cohen's d is computed as $\bar{\Delta}/s_\Delta$.

Table 4 reports paired-comparison statistics for all twelve model–K–dataset combinations. Applying Bonferroni correction for the twelve simultaneous tests (adjusted threshold $\alpha' = 0.05/12 \approx 0.00417$), no condition reaches statistical significance. The $K = 9$ conditions ($n = 10$ rounds) are underpowered: post-hoc power for a medium effect ($d = 0.50$) at $\alpha = 0.05$ is approximately 35%, rising to roughly 50% at $d = 0.70$. Null results for these conditions should therefore be interpreted as inconclusive rather than as positive evidence of equivalence. In particular, GPT-5.4 at $K = 9$ ($p = 0.050$, $d = 0.72$) narrowly misses the uncorrected threshold and exhibits a medium-to-large effect size; the absence of significance here reflects insufficient sample size rather than a reliably small effect. The four CUB-Crows $K = 3$ conditions ($n = 20$) yield $p \geq 0.23$ with small effect sizes ($d \leq 0.28$), but these conditions are subject to accuracy ceiling effects ($\bar{A} \geq 0.84$ for most models) that suppress any detectable semantic benefit regardless of whether one exists, and cannot therefore be interpreted as evidence of visual primacy.

CUB-Cormorant results. The four CUB-Cormorant conditions ($K = 3$, $n = 20$) provide the only $K = 3$ comparisons free of ceiling effects, with blind-condition means of 67.8%–73.9%. Despite the absence of ceiling suppression, no condition reaches statistical significance ($p \geq 0.11$). Effect sizes are inconsistent in direction across models: Gemini shows the largest positive trend ($d = 0.37$, $\bar{\Delta} = +10.0$ pp), GPT shows a small positive trend ($d = 0.15$), while Qwen shows a small negative effect ($d = -0.21$, $\bar{\Delta} = -5.0$ pp) and Kimi shows no effect ($d = 0.00$). The direction reversal for Qwen, where naming the species decreases accuracy, is inconsistent with a simple semantic enrichment hypothesis and suggests that name disclosure interacts with model-specific response tendencies rather than providing a uniformly additive signal. The cormorant results thus extend the CUB-Crows conclusion (no reliable semantic benefit) to a condition where that conclusion is no longer confounded by ceiling effects, while simultaneously revealing cross-model variability that resists a unified interpretation.

At $K = 3$ CUB-Crows, signed deltas are centred near zero with symmetric win/loss round counts. At $K = 9$, accuracy decreases globally, but the blind–named gap narrows rather than grows: Gemini, Kimi, and Qwen show mean deltas of at most +0.015.

Per-class analysis on CUB-Crows reveals a consistent two-tier structure. Visually distinctive species (Groove-billed Ani, Red-winged Blackbird, and Brandt's Cormorant) achieve $\bar{A} = 1.00$ across all four models at $K = 3$, confirming that label semantics add no marginal signal when exemplars are already discriminative. The corvid cluster (American Crow, Fish Crow, Common Raven) constitutes the persistent accuracy floor: Fish Crow is the hardest class for every model at $K = 9$, with named-condition means of 0.37, 0.33, 0.27, and 0.20 for Gemini, GPT, Kimi, and Qwen respectively, rankings identical to those in the blind condition. This per-class pattern is consistent with the corvid confusion reflecting a visual representational limitation rather than a vocabulary deficit resolvable by species names; however, the possibility of a moderate semantic benefit that the current design lacks power to detect cannot be excluded.

4.3 Task 3: Prompt ablation study

Table 5 reports mean top-1 and top-2 accuracy across the six prompt variants on SoyLocal for Gemini-3-Flash-Preview and Kimi-K2.5 ($n_{\mathrm{rounds}} = 10$ per variant per model). All rounds produced parseable output. The elevated standard deviations on the Direct and Minimal variants (17%–22% across both models) reflect genuine round-to-round variability driven by stochastic category sampling rather than output failures: these variants provide little discriminative scaffolding, making performance more sensitive to whether a given round happens to draw easily or poorly separated categories.

Overall level difference. Averaged across all six variants, Kimi-K2.5 achieves a mean top-1 accuracy of 45.0% versus 37.4% for Gemini, a 7.6 percentage point advantage that holds on every individual variant. This model-level gap is the single largest source of variance in the ablation, illustrating that architecture matters more than prompt design within the range of strategies tested here.

Prompt sensitivity is strongly model-dependent. The two models disagree on the optimal prompt variant: Gemini achieves its highest top-1 accuracy under Baseline (41.3%), while Kimi peaks under Interleaved (48.7%). The total span from best to worst variant is 6.3 percentage points for Gemini (41.3% vs 35.0%) and 9.7 percentage points for Kimi (48.7% vs 39.0%), indicating that Kimi is more sensitive to prompt structure within this dataset and condition. No single variant is consistently optimal across both backends.

Chain-of-thought decomposition. The CoT contribution (measured as Baseline minus Direct accuracy) is sign-reversed across the two models: +6.3 percentage points for Gemini and −3.3 for Kimi. For Gemini, the two-stage describe-then-classify pipeline plausibly reduces confirmation bias when categories are visually ambiguous[45,48]; for Kimi, the added verbosity introduces noise rather

https://doi.org/10.48130/ker-0026-0010

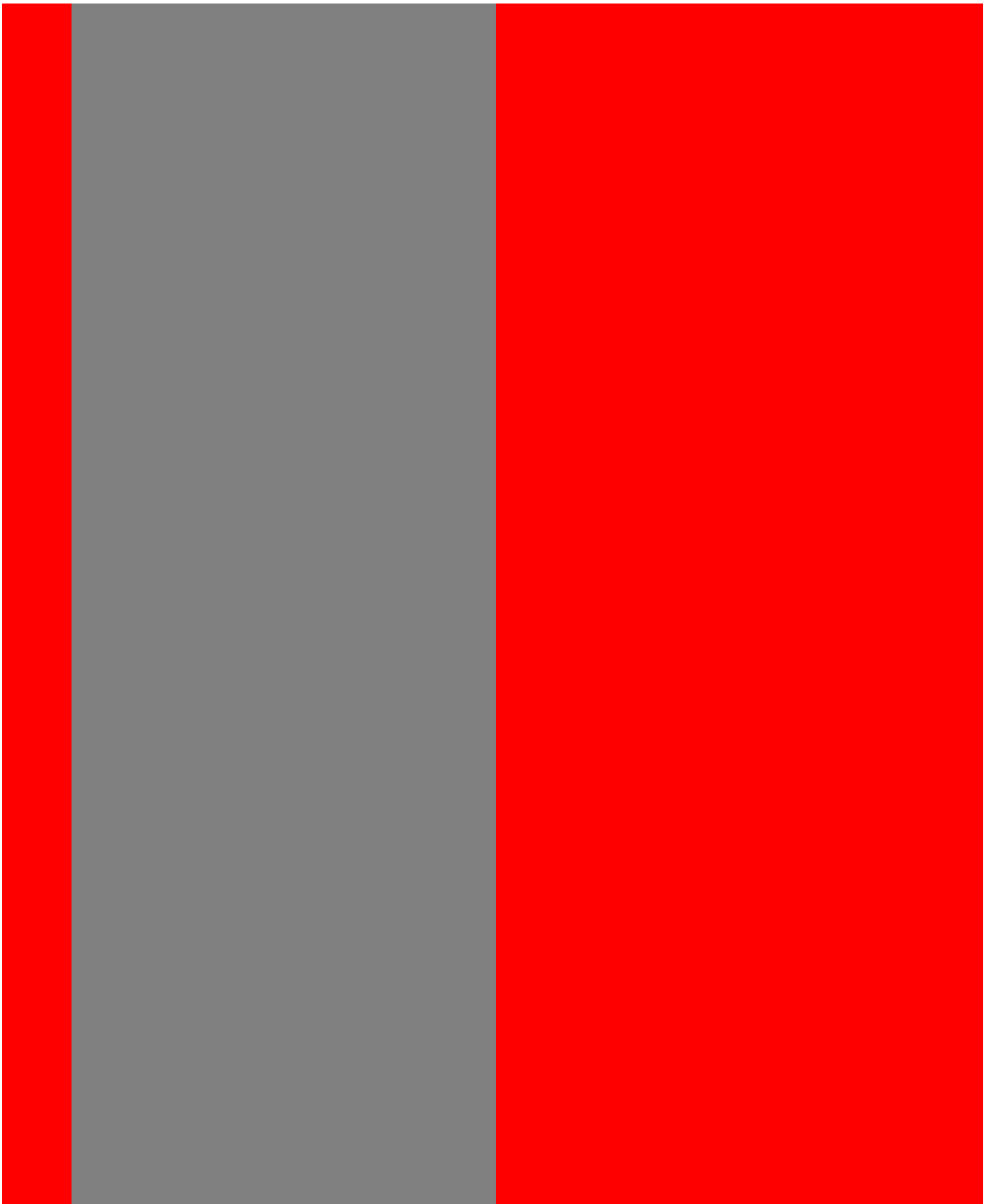

Fig. 4 All six prompt variants used in Task 3, shown schematically (part 1 of 2: variants a–c). Image blocks are denoted $\langle \cdot \rangle$; K categories with three reference shots each are assumed throughout. The shared JSON output schema is reproduced verbatim in every box so that differences in discriminative content are visible by direct inspection. See Fig. 4 for variants (d)–(f). (part 2 of 2: variants d–f.) Variants differ along three orthogonal dimensions: whether chain-of-thought (CoT) decomposition is applied ([a], [d], [e] vs [b], [c], [f]); whether explicit match/contradiction elimination is required ([d] only); and where task

Baseline prompt (schematic; image blocks denoted by ⟨·⟩; K categories, 3 shots each)

```
You are an expert on ultra-fine-grained image classification.
Below are reference images for K visually similar categories:
1, 2, ..., K. These categories differ in SUBTLE ways.  Focus
on:  minor colour/shade variations, fine texture patterns,
edge morphology, surface properties, structural proportions,
and spatial arrangement of subparts.  Do not rely on a single
feature -- use the conjunction of multiple fine-grained cues to
discriminate.

[ repeated for each category k = 1, ..., K ]
— Category:  k --
⟨reference image 1⟩ ⟨reference image 2⟩ ⟨reference image 3⟩

-- Query images to classify --
[ repeated for each query image i ]
Query image i (image_i):  ⟨query image⟩

For EACH query image, follow this process:
1. DESCRIBE the image's fine-grained visual features without
naming any category.
2. COMPARE those features against each category's references,
noting matches and mismatches.
3. DECIDE on the best category, explaining why it beats the
runner-up.

Respond ONLY with a JSON array.  Each element must have:
"filename":  query image identifier,
"observed_features":  list of 3-5 key features,
"category":  final classification,
"runner_up":  second most likely category,
"confidence":  float 0-1,
"reasoning":  why chosen category beats the runner-up.
No text outside the JSON.
```

Fig. 5 Full baseline prompt used in Tasks 1–3. Italicised K and indexed placeholders are instantiated at runtime. The mandatory JSON schema enforces structured output and eliminates free-text parsing as a confound across all six Task 3 variants. In the *named* condition of Task 2, numeric category identifiers are replaced with real species names; all other elements are identical. The figure is original and created by Chenhui Wang. © 2026 Chenhui Wang.

Table 3. Top-1 accuracy (%) on SoyLocal and Cotton80.

		Method	$K = 3$	$K = 10$	$K = 50$
SoyLocal	*Supervised methods (fine-tuned)*	Human Expert	63.89	32.50	12.84
		AlexNet[29]	88.89	53.33	33.33
		VGG-16[59]	77.78	50.00	56.00
		ResNet-50[28]	88.89	63.33	62.00
		DCL[60]	88.89	63.33	60.67
	MLLMs, few-shot ICL (this work)	gemini-3-flash-preview[22]	85.56 ± 12.01	46.00 ± 6.81	—
		gpt-5.4-2026-03-05[21]	81.67 ± 15.41	51.67 ± 9.72	20.44 ± 3.79 ‡
		kimi-k2.5[20]	82.78 ± 13.23	45.00 ± 7.24	—
		qwen3.5-397b-a17b[23]	79.45 ± 19.17	46.67 ± 10.30	—
Cotton80	*Supervised methods (fine-tuned)*	Human Expert	66.67	63.33	14.00
		AlexNet[29]	88.89	66.67	24.67
		VGG-16[59]	77.78	56.67	53.33
		ResNet-50[28]	88.89	66.67	52.67
		DCL[60]	77.78	73.33	54.67
	MLLMs, few-shot ICL (this work)	gemini-3-flash-preview[22]	67.78 ± 13.44	33.67 ± 12.42	—
		gpt-5.4-2026-03-05[21]	71.67 ± 14.63	39.33 ± 8.13	12.44 ± 5.09 ‡
		kimi-k2.5[20]	67.22 ± 16.31	34.00 ± 12.05	—
		qwen3.5-397b-a17b[23]	47.22 ± 26.46	34.33 ± 9.03	—

Supervised baselines from Yu et al.[1]; MLLM results are mean ± standard deviation over N independent rounds under the blind numeric-label protocol. '—' indicates conditions not evaluated. At $K = 3$, the random-guessing baseline is 33.3%. ‡ GPT-5.4 $K = 50$ results are from a preliminary evaluation of $N = 3$ rounds and should be treated with caution.

than precision at this category count. The sign reversal is sufficient reason not to treat prompt strategies validated on a single architecture as transferable defaults.

Contrastive elimination. The Contrastive variant trails Baseline by 4.0 percentage points for Gemini, suggesting that forced contradiction enumeration imposes generation overhead without a proportional discriminative return at $K = 10$. For Kimi the direction is reversed: Contrastive exceeds Baseline by 5.7 percentage points, consistent with contrastive priming theory[46,47] that explicitly surfacing contradictions tightens category boundaries in high inter-class similarity settings.

Instruction placement. Co-located per-category instructions (Inter-

https://doi.org/10.48130/ker-0026-0010

Table 4. Task 2 paired-comparison statistics (blind vs named label condition).

Dataset	Model	K	N	$\bar{A}_B$	$\bar{A}_N$	$\bar{\Delta}$	p (uncorr.)	d
CUB-Crows	gemini-3-flash-preview[22]	3	20	0.922	0.922	+0.000	1.000	0.00
	gemini-3-flash-preview[22]	9	10	0.807	0.822	+0.015	0.533	0.20
	gpt-5.4-2026-03-05[21]	3	20	0.883	0.889	+0.006	0.825	0.05
	gpt-5.4-2026-03-05[21]	9	10	0.678	0.730	+0.052	0.050	0.72
	qwen3.5-397b-a17b[23]	3	20	0.867	0.878	+0.011	0.629	0.11
	qwen3.5-397b-a17b[23]	9	10	0.743	0.754	+0.004	0.883	0.05
	kimi-k2.5[20]	3	20	0.845	0.883	+0.039	0.232	0.28
	kimi-k2.5[20]	9	10	0.789	0.793	+0.004	0.868	0.05
CUB-Cormorant (K = 3 exhausts available categories; no ceiling effects)	gemini-3-flash-preview[22]	3	20	0.678	0.778	+0.100	0.114	0.37
	gpt-5.4-2026-03-05[21]	3	20	0.689	0.722	+0.033	0.516	0.15
	qwen3.5-397b-a17b[23]	3	20	0.739	0.689	−0.050	0.359	−0.21
	kimi-k2.5[20]	3	20	0.683	0.683	+0.000	1.000	0.00

$\bar{A}_B/\bar{A}_N$: mean per-round accuracy; $\bar{\Delta}$: mean accuracy gain from name disclosure; d: Cohen's d ($= \bar{\Delta}/s_\Delta$). p-values are from a one-sample paired t-test on $\delta_r = A_N^{(r)} - A_B^{(r)}$, reported uncorrected; no condition survives Bonferroni correction ($\alpha/m \approx 0.00417$, $m = 12$). $K = 9$ CUB-Crows conditions ($n = 10$) are underpowered (post-hoc power $\approx 35\%$ for $d = 0.50$); null results are inconclusive. $K = 3$ CUB-Crows conditions ($n = 20$) are subject to accuracy ceiling effects ($\bar{A} \geq 0.84$) that preclude detection of any semantic benefit. CUB-Cormorant conditions ($K = 3$, $n = 20$) are free of ceiling effects ($\bar{A}_B \leq 0.74$) but show inconsistent effect directions across models.

leaved) cost Gemini 5.7 percentage points relative to the global preamble (Baseline), suggesting that global task context is more useful than local cues for this model. The direction reverses for Kimi: Interleaved outperforms Baseline by 6.3 percentage points, the largest single-step gain for either model, pointing to enhanced feature binding when discriminative cues appear alongside their corresponding reference images. Deferred instruction placement (Instructions-Last) follows the same sign-reversal pattern: −3.0 percentage points for Gemini relative to Baseline, but +4.0 for Kimi, suggesting that Kimi benefits from accumulating the full visual context before committing to a task frame. The cross-model spread is largest on Interleaved (13.0 percentage points) and smallest on Baseline (1.0 percentage point), making Baseline the most robust default across the two backends evaluated here.

High-variance variants and top-2 accuracy. Direct and Minimal produce the highest standard deviations (17%–22%), reflecting their minimal scaffolding rather than output failures. Notably, their top-2 accuracies are competitive with or exceed those of the structured variants (64.6% and 65.0% for Gemini; 68.9% and 69.2% for Kimi), indicating that both models generate informative visual feature comparisons even without explicit reasoning guidance; the discriminative deficit lies in the final category selection step rather than in feature extraction.

Lower bound and total prompt engineering contribution. The gap between each model's best-performing variant and Minimal quantifies the ceiling benefit of structured prompt engineering: 4.3% for Gemini (Baseline over Minimal) and 9.7% for Kimi (Interleaved over Minimal). These figures are upper bounds within this exploratory, single-dataset design.

4.4 Major challenges in ultra-FGVC with MLLMs

Three fundamental challenges currently limit MLLM deployment in Ultra-FGVC tasks.

Visual encoder resolution. The most pervasive challenge exposed across all three tasks is the representational ceiling of current CLIP-based visual encoders. The performance gap between MLLMs and fine-tuned supervised methods widens sharply as K grows (from a modest deficit at $K = 3$ to a 12–34 point deficit at $K = 10$), and the preliminary $K = 50$ GPT-5.4 data (20.4% and 12.4% on SoyLocal and Cotton80 respectively) are consistent with continued degradation at benchmark scale. This gap is not reduced by richer prompt structure (Task 3). The persistent corvid accuracy floor in Task 2, which does not improve meaningfully under named conditions, is consistent with a visual representational limitation, though the statistical design of Task 2 is insufficient to establish this definitively: the possibility of a moderate semantic benefit that current sample sizes cannot detect remains open. Current CLIP encoders were trained on image-text pairs at coarse semantic granularity and lack the resolution to reliably separate inter-class feature distances at subspecies level[19].

Output generation reliability. Although all Task 3 rounds produced parseable output under the current parser implementation, structured-output compliance remains a latent deployment risk. The elevated variance on Direct and Minimal ($\sigma \approx 17\%$–22%) shows that discriminative performance is sensitive to prompt scaffolding, and the schema compliance rates observed here reflect a parser that was hardened iteratively against model outputs. In broader deployment scenarios (different models, longer category lists, or schema changes), constrained decoding mechanisms or function-calling interfaces that enforce JSON schema at generation time remain advisable to eliminate compliance as a confound and ensure reproducible inference pipelines.

Confidence miscalibration. All four models display systematic overconfidence across Tasks 1 and 2, with mean reported confidence of 0.74–0.91 irrespective of accuracy level. This precludes the use of raw model confidence as a deployment rejection threshold or reliability signal in precision agriculture workflows where misclassification has operational consequences. For Gemini, the miscalibration extends across all six Task 3 prompt variants (0.85–0.89) and appears to be an intrinsic property of frontier model confidence estimation in discriminative tasks with subtle inter-class distances. Whether the same holds for Kimi-K2.5 in Task 3 cannot be determined from the current results, as confidence statistics for Kimi were not captured in the Task 3 experimental run.

5 Emerging directions and future work

5.1 Emerging techniques for ultra-FGVC

Several concrete technical directions are well-positioned to address the gaps documented here.

Richer visual representations. Both Task 1 and Task 2 point consistently to the visual encoder as a plausible primary bottleneck, though Task 2's statistical design is insufficient to conclusively rule out a role for label semantics at higher K. Replacing CLIP-based

Table 5. Prompt ablation results on SoyLocal (n = 10 rounds per variant per model).

Variant	CoT	Contr.	Instr.	Gemini-3-flash-preview		Kimi-K2.5	
				Top-1 (%)	Top-2 (%)	Top-1 (%)	Top-2 (%)
Baseline	✓	–	Pre	41.3 ± 6.9	58.3	42.3 ± 8.9	64.7
Instructions-Last	–	–	Post	38.3 ± 10.5	56.0	46.3 ± 7.1	68.7
Contrastive	✓	✓	Pre	37.3 ± 15.9	61.5	48.0 ± 6.9	66.7
Minimal	–	–	Inline	37.0 ± 20.0	65.0	39.0 ± 21.6	69.2
Interleaved	✓	–	Local	35.7 ± 13.7	61.1	48.7 ± 9.5	71.7
Direct	–	–	Pre	35.0 ± 19.7	64.6	45.7 ± 17.1	68.9
Mean (all variants)				*37.4*		*45.0*	

Top-1 and Top-2 accuracy: mean ± standard deviation. All rounds produced parseable output. The elevated standard deviations on Direct and Minimal (17%–22%) reflect genuine round-to-round variability under stochastic category sampling, not output failures. Instr. placement abbreviations: Pre = preamble before images; Post = appended after all images; Local = co-located per category; Inline = inline with images. Variants are ordered by Gemini top-1 accuracy (descending).

encoders with part-aware or region-grounded backbones[19] could increase the effective resolution at which fine-grained morphological differences are captured. Augmenting in-context exemplars with model-generated diagnostic captions via the describe-then-classify pipeline[45,50] may raise the visual resolution ceiling without requiring task-specific fine-tuning. The FineR framework[50], which extracts part-level visual attributes via VQA and reasons over them with world knowledge, is one concrete instantiation of this direction that could be adapted to cultivar-level agricultural classification.

Scale-up and fine-tuning. Should larger annotated datasets become available, supervised fine-tuning of the vision encoder, or lightweight parameter-efficient adaptation (e.g. LoRA) applied to the cross-modal connector, could significantly raise the accuracy ceiling documented here. The performance gap between MLLM few-shot ICL and fully supervised methods at K = 10 provides a clear and quantified target. Many-shot ICL[43], which scales log-linearly with demonstration count up to 2,000 shots in frontier models, represents a near-term option that does not require retraining and could be combined with kNN-based exemplar selection[53] to maximise the information content of each added shot.

Prompt robustness and extended ablation. Task 3 exposes sign-reversed responses to all four tested design dimensions across the two backends, revealing that architecture effects dwarf prompt effects in this setting. Extending the ablation to all four backends and the remaining datasets, with confidence statistics collected for each model, would substantially sharpen these conclusions. The elevated variance on unstructured variants (Direct, Minimal) also warrants investigation at higher round counts to narrow confidence intervals on the lower-bound estimates.

5.2 Outlook for future work

Several broader directions beyond the current scope merit systematic investigation.

Extended model and dataset coverage, including K = 50. The preliminary GPT-5.4 results at K = 50 establish that the scale is technically within reach for at least one frontier model. Future work should pursue systematic K = 50 evaluation across all four backends, via retrieval-augmented reference selection, multi-turn context management, or model-specific API adaptations, to enable the first complete MLLM–supervised comparison at benchmark scale.

Confidence calibration. All four models display systematic overconfidence (mean scores 0.74–0.91 irrespective of accuracy), precluding raw confidence from serving as a deployment rejection threshold. Post-hoc calibration methods such as temperature scaling or conformal prediction are prerequisites for operational deployment. Conformal prediction is particularly well-suited to the data-scarce conditions of Ultra-FGVC given its distribution-free coverage guarantees.

Many-shot and retrieval-augmented classification. Given the log-linear scaling of accuracy with shot count in frontier models[43] and the strong effect of exemplar selection[53,55], retrieval-augmented pipelines that dynamically select maximally informative exemplars may substantially close the gap to supervised methods without fine-tuning, while simultaneously addressing the K = 50 feasibility constraint. More broadly, active learning and semi-supervised methods that exploit the large pools of unlabelled field imagery already captured could substantially reduce the per-category annotation burden that currently bottlenecks Ultra-FGVC dataset construction. Combining such approaches with the few-shot ICL capabilities documented here offers a path toward operational pipelines that require only minimal expert labelling per new cultivar or pest species.

Statistical power in Task 2 follow-up. A conclusive test of the visual-primacy hypothesis requires K = 9 with $n \geq 30$ rounds to achieve > 80% power for a medium effect size (d = 0.50), extended to all four evaluated backends. The CUB-Cormorant design should be replicated with larger n and, if possible, extended to a K = 9 cormorant-analogous subset to avoid ceiling effects at higher category counts.

6 Conclusions

Accurate subspecies-level classification of crop cultivars and pest species is a prerequisite for targeted interventions in precision agriculture, yet it sits at the intersection of two compounding difficulties: the discriminative features that separate cultivars are highly localised and morphologically subtle, and the labelled data available per category at this granularity is too scarce to sustain standard supervised learning. MLLMs, which encode rich biological knowledge from web-scale training and adapt at inference time through in-context learning, are a principled candidate for closing this gap without retraining. Whether they actually do so, and exactly where they fail, has not previously been characterised at Ultra-FGVC scale.

This paper provides that characterisation through a controlled three-task evaluation of four state-of-the-art MLLMs across three Ultra-FGVC benchmarks. Task 1 establishes that MLLMs are practically viable for low-K field triage (K = 3), exceeding human expert accuracy on SoyLocal and matching it on Cotton80, but fall sharply below fine-tuned supervised methods at K = 10. Further still, in the preliminary K = 50 evaluation, tracing a degradation trajectory consistent with a visual encoder bottleneck rather than a prompting or in-context-learning failure. Task 2 finds no statistically reliable accuracy gain from disclosing real category names in place of numeric identifiers across any of the 12 evaluated conditions, though the design is underpowered at K = 9 and inconsistent effect directions across

https://doi.org/10.48130/ker-0026-0010

models at $K = 3$ preclude a strong visual-primacy conclusion. Task 3 shows that prompt-design effects are strongly model-dependent, with sign-reversed responses to every tested design dimension between the two evaluated backends, cautioning against treating prompt strategies validated on a single architecture as transferable defaults.

All three tasks converge on the same primary constraint: the representational ceiling of current CLIP-based visual encoders, which is insensitive to richer prompt structure and largely insensitive to label semantics within the conditions tested. Output generation reliability and confidence miscalibration are secondary challenges that compound the representational gap in deployment settings. Each has a concrete technical path forward (richer visual backbones, constrained decoding, and post-hoc calibration), and the quantified performance gaps reported here provide explicit targets against which progress on those directions can be measured.

Ethical statements

During the preparation of this work, the author(s) used Claude Sonnet 4.6 for language refinement and image enhancement. The author(s) reviewed and edited all content produced with the assistance of this tool, verified its accuracy, and take full responsibility for the integrity and originality of the final manuscript. This work represents the author(s)' own intellectual contribution, and no AI tool is credited as an author.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Wang C; data collection: Wang C; analysis and interpretation of results: Wang C; draft manuscript preparation: Wang C, Zhang P. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data that support the findings of this study are available at https://github.com/CWang0102/UFGVC-mllm-review in the repository. These data were derived from the following resources available in the public domain: https://github.com/XiaohanYu-GU/Ultra-FGVC, https://github.com/lucinda0love/SupCon-ViT-pytorch.

Acknowledgments

This research received no external funding.

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 31 March 2026; Revised 11 April 2026; Accepted 5 May 2026; Published online 30 July 2026

References

[1] Yu X, Zhao Y, Gao Y, Yuan X, Xiong S. 2022. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10–17, 2021, Montreal, QC, Canada*. USA: IEEE. pp. 10265–10275 doi: 10.1109/ICCV48922.2021.01012

[2] Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, et al. 2019. The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution* 3(3):430−439

[3] Gerhards R, Andújar Sanchez D, Hamouz P, Peteinatos GG, Christensen S, et al. 2022. Advances in site-specific weed management in agriculture — a review. *Weed Research* 62(2):123−133

[4] Jin X, Liu T, McCullough PE, Chen Y, Yu J. 2023. Evaluation of convolutional neural networks for herbicide susceptibility-based weed detection in turf. *Frontiers in Plant Science* 14:1096802

[5] Horowitz AR, Ghanim M, Roditakis E, Nauen R, Ishaaya I. 2020. Insecticide resistance and its management in *Bemisia tabaci* species. *Journal of Pest Science* 93(3):893−910

[6] Behere GT, Tay WT, Russell DA, Batterham P. 2008. Molecular markers to discriminate among four pest species of *Helicoverpa* (Lepidoptera: Noctuidae). *Bulletin of Entomological Research* 98(6):599−603

[7] Kim KC, Byrne LB. 2006. Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. *Ecological Research* 21(6):794

[8] Valan M, Makonyi K, Maki A, Vondráček D, Ronquist F. 2019. Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Systematic Biology* 68(6):876−895

[9] Yu X, Zhao Y, Gao Y, Xiong S. 2021. MaskCOV: a random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition* 119:108067

[10] Alayrac JB, Donahue J, Luc P, Miech A, Barr I, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), 2022, New Orleans, Louisiana, USA*. NeurIPS. https://openreview.net/attachment?id=EbMuimAbPbs&name=supplementary_material

[11] Liu H, Li C, Wu Q, Lee YJ. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36. December 10-16, 2023. New Orleans, Louisiana, USA*. NeurIPS. pp. 34892–34916 doi: 10.52202/075280-1516

[12] Liu Y, Duan H, Zhang Y, Li B, Zhang S, et al. 2025. MMBench: is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*. Cham: Springer. pp. 216–233 doi: 10.1007/978-3-031-72658-3_13

[13] Li B, Ge Y, Ge Y, Wang G, Wang R, et al. 2024. SEED-bench: benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 16–22, 2024, Seattle, WA, USA*. USA: IEEE. pp. 13299–13308 doi: 10.1109/CVPR52733.2024.01263

[14] Yue X, Ni Y, Zheng T, Zhang K, Liu R, et al. 2024. MMMU: a massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 16–22, 2024, Seattle, WA, USA*. USA: IEEE. pp. 9556–9567 doi: 10.1109/CVPR52733.2024.00913

[15] Maji S, Kannala J, Rahtu E, Blaschko M, Vedaldi A. 2013. Fine-grained visual classification of aircraft. *Technical Report*. University of Oxford, Oxford, UK. arXiv:1306.5151. 10.48550/arXiv.1306.5151

[16] Krause J, Stark M, Jia D, Li FF. 2013. 3D object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, December 2–8, 2013, Sydney, NSW, Australia*. USA: IEEE. pp. 554–561 doi: 10.1109/ICCVW.2013.77

[17] Islam A, Biswas MR, Zaghouani W, Belhaouari SB, Shah Z. 2023. Pushing boundaries: Exploring zero shot object classification with large multimodal models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), November 21–24, 2023, Abu Dhabi, United Arab Emirates*. USA: IEEE. pp. 1–5 doi: 10.1109/SNAMS60348.2023.10375440

[18] Wu W, Yao H, Zhang M, Song Y, Ouyang W, et al. 2024. GPT4Vis: what can GPT-4 do for zero-shot visual recognition? *arXiv* Preprint

[19] Tong S, Liu Z, Zhai Y, Ma Y, LeCun Y, et al. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 16–22, 2024, Seattle, WA, USA*. USA: IEEE. pp. 9568–9578 doi: 10.1109/CVPR52733.2024.00914

[20] Kimi Team, Bai T, Bai Y, Bao Y, et al. 2026. Kimi K2.5: visual agentic intelligence. *arXiv* Preprint

[21] OpenAI. ChatGPT. OpenAI, 2025. https://chatgpt.com
[22] Google. Gemini. Google, 2025. https://gemini.google.com
[23] Qwen Team. Qwen3.5: towards native multimodal agents. February 2026. https://qwen.ai/blog?id=qwen3.5
[24] Wah C, Branson S, Welinder P, Perona P, Belongie S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. *Technical Report CNS-TR-2011-001*. Pasadena, CA: California Institute of Technology. www.vision.caltech.edu/visipedia/CUB-200-2011.html
[25] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. 2021. An image is worth 16×16 words: transformers for image recognition at scale. *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy
[26] Demidov D, Sharif M, Abdurahimov A, Cholakkal H, Khan F. 2023. Salient mask-guided vision transformer for fine-grained classification. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, February 19–21, 2023. Lisbon, Portugal*. Portugal: Science and Technology Publications. pp. 27–38 doi: 10.5220/0011611100003417
[27] Zhang K, Cao J, Li HS, Cai Q. 2023. Cross-layer feature fusion vision transformer for fine-grained visual classification. In *2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS), September 22–24, 2023, Tianjin, China*. USA: IEEE. pp. 1–6 10.1109/DOCS60977.2023.10294512
[28] e K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA*. USA: IEEE. pp. 770–778 doi: 10.1109/CVPR.2016.90
[29] Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012), December 3–6, 2012, Lake Tahoe, Nevada, USA*. Curran Associates, Inc. pp. 1097–1105 https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
[30] Pan Z, Yu X, Zhang M, Gao Y. 2021. Mask-guided feature extraction and augmentation for ultra-fine-grained visual categorization. *arXiv* Preprint
[31] Yu X, Wang J, Gao Y. 2023. CLE-ViT: contrastive learning encoded transformer for ultra-fine-grained visual categorization. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, August 19–25, 2023, Macau, SAR China*. International Joint Conferences on Artificial Intelligence Organization. pp. 4531–4539 doi: 10.24963/ijcai.2023/504
[32] Zhang P, Yu X, Gu M, Wu Y, Gao Y, et al. 2025. Revisiting continual ultra-fine-grained visual recognition with pre-trained models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. ACM. pp. 9474–9482 doi: 10.24963/ijcai.2025/1053
[33] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. 2021. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021), July 18–24, 2021. Proceedings of Machine Learning Research*. Vol. 139. PMLR. pp. 8748–8763 https://proceedings.mlr.press/v139/radford21a.html
[34] Yin S, Fu C, Zhao S, Li K, et al. 2024. A survey on multimodal large language models. *National Science Review* 11(12):nwae403
[35] Li J, Li D, Savarese S, Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* 202:814
[36] Dai W, Li J, Li D, Tiong A, Zhao J, et al. 2023. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36. December 10–16, 2023. New Orleans, Louisiana, USA*. NeurIPS. pp. 49250–49267 doi: 10.52202/075280-2142
[37] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, et al. 2024. GPT-4 technical report. *arXiv* Preprint
[38] Gemini Team. 2024. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. *arXiv* Preprint
[39] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. *Technical Report*. www.anthropic.com/news/claude-3-family
[40] Li C, Wong C, Zhang S, Usuyama N, Liu H, et al. 2023. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023 Datasets and Benchmarks Track), December 10–16, 2023, New Orleans, LA, USA*. Curran Associates, Inc. pp. 28541–28564 https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html
[41] Liu F, Dai W, Zhang C, Zhu J, Yao L, et al. 2025. Co-LLaVA: efficient remote sensing visual question answering via model collaboration. *Remote Sensing* 17(3):466
[42] Li Y, Du Y, Zhou K, Wang J, Zhao X, et al. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, Singapore*. Stroudsburg, PA, USA: ACL. pp. 292–305 doi: 10.18653/v1/2023.emnlp-main.20
[43] Jiang Y, Irvin J, Wang JH, Chaudhry MA, Chen JH, et al. 2024. Many-shot in-context learning in multimodal foundation models. *arXiv* Preprint
[44] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), November 28 – December 9, 2022, New Orleans, LA, USA*. Curran Associates, Inc. pp. 24824–24837 https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
[45] Menon S, Vondrick C. 2023. Visual classification via description from large language models. *The Eleventh International Conference on Learning Representations (ICLR 2023), May 1–5, 2023, Kigali, Rwanda*. OpenReview.net. https://openreview.net/forum?id=41GFzHFc4zx
[46] Chia YK, Chen G, Tuan LA, Poria S, Bing L. 2023. Contrastive chain-of-thought prompting. *arXiv* Preprint
[47] Jiang X, Tang H, Gao J, Du X, He S, et al. 2024. Delving into multi-modal prompting for fine-grained visual classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 38(3):2570−2578
[48] Zhang Z, Zhang A, Li M, Zhao H, Karypis G, et al. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research 2024*. https://openreview.net/forum?id=y1pPWFVfvR
[49] Pratt S, Covert I, Liu R, Farhadi A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV), October 1–6, 2023, Paris, France*. pp. 15645–15655 doi: 10.1109/ICCV51070.2023.01438
[50] Liu M, Roy S, Li W, Zhong Z, Sebe N, et al. 2024. Democratizing fine-grained visual recognition with large language models. *The Twelfth International Conference on Learning Representations (ICLR 2024), May 7–11, 2024, Vienna, Austria*. OpenReview.net. https://openreview.net/forum?id=c7DND1iIgb
[51] Pan J, Zhong R, Xia F, Huang J, Zhu L, et al. 2025. ChatLeafDisease: a chain-of-thought prompting approach for crop disease classification using large language models. *Plant Phenomics* 7(3):100094
[52] Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. 2021. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, Dublin*. Ireland Association for Computational Linguistics. pp. 8086−8098 doi: 10.18653/v1/2022.acl-long.556
[53] Zhang Y, Zhou K, Liu Z. 2023. What makes good examples for visual in-context learning? In *Advances in Neural Information Processing Systems 36. December 10–16, 2023. New Orleans, Louisiana, USA*. NeurIPS. pp. 17773–17794 doi: 10.52202/075280-0780
[54] Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, et al. 2022. Rethinking the role of demonstrations: what makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, Abu Dhabi, United Arab Emirates*. Stroudsburg, PA, USA: ACL. pp. 11048–11064 doi: 10.18653/v1/2022.emnlp-main.759
[55] Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, et al. 2024. In-

https://doi.org/10.48130/ker-0026-0010

context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications* 15(1):10104

[56] Yao L, Yang Y. 2026. Large language models are contrastive reasoners. *Expert Systems with Applications* 301:130407

[57] Lu X, Yu X, Wang K, Wang Y, Wang P, et al. 2023. SupCon-ViT: supervised contrastive learning for ultra-fine-grained visual categorization. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2023, Port Macquarie, Australia*. USA: IEEE. pp. 281–288 doi: 10.1109/DICTA60407.2023.00046

[58] Chiang WL, Zheng L, Sheng Y, Angelopoulos AN, Li T, et al. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. *Proceedings of the 41st International Conference on Machine Learning (ICML 2024), July 21–27, 2024, Vienna, Austria*. PMLR. Vol. 235. pp. 8359–8388 https://proceedings.mlr.press/v235/chiang24b.html

[59] Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR 2015), May 7–9, 2015, San Diego, CA, USA*. arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556

[60] Chen Y, Bai Y, Zhang W, Mei T. 2019. Destruction and construction learning for fine-grained image recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, Long Beach, CA, USA*. USA: IEEE. pp. 5152–5161 doi: 10.1109/CVPR.2019.00530