

What Makes MLLM Adaptation Effective for FGVC?

Logan Lawrence, Oindrila Saha, Subhansu Maji, Grant Van Horn

College of Information and Computer Sciences

University of Massachusetts, Amherst

Amherst, MA 01002, USA

{lclawrence, osaha, smaji, gvanhorn}@umass.edu

Abstract

While multimodal large language models (MLLMs) are increasingly used in deployment-specific visual domains as zero-shot solutions, they typically underperform specialized encoder-based models. However, adapting MLLMs to new tasks remains difficult due to the fact that supervised fine-tuning (SFT) often degrades the more generalist visual and instruction-following abilities that make these models useful in the first place. This tension is pronounced in fine-grained visual classification (FGVC) tasks. To improve FGVC performance while preserving general-purpose multimodal competence, this work studies the properties of a selective-edit fine-tuning approach that constrains supervision to the portions of the response most relevant to the target task. To measure the ability of the model to deploy its knowledge while preserving general comprehension, we introduce IFBirds: a small-scale benchmark which tests instruction-following and bird species identification simultaneously. Across a suite of instruction-tuned models, we find that user input image and text augmentation reliably improves robustness, while assistant output text augmentations are only beneficial when paired with mechanisms that limit overfitting. Finally, in line with previous works, this work provides evidence that fine-tuning does not increase the richness of the representations but rather improves the ability of the LLM backbone to deploy the visual knowledge.

1 Introduction

Why haven't we seen the widespread adoption of MLLMs in FGVC settings? Visual classification has been revitalized by the introduction of Multimodal Large Language Models (MLLMs) (OpenAI, 2023; Dai et al., 2023; Liu et al., 2023a; 2024a) which continue improve performance in zero-shot and few-shot settings (Hong et al., 2025; Saha et al., 2024; 2025; Lawrence et al., 2026). However, it is also true that MLLMs lag behind specialized encoder-based models (He et al., 2025; 2026) in specialized visual domains where data is plenty. It is common for practitioners to perform prompt engineering and treat their domain data primarily as an evaluation set, rather than directly adapting model weights (Brown et al., 2020). The reason for this is familiar: naive supervised fine-tuning (SFT) on in-domain examples can induce *catastrophic forgetting*, where updates optimized for a narrow distribution interfere with pre-trained representations and instruction-following behavior, degrading general reasoning and conversational competence (Li et al. (2024); Sanyal et al. (2025)). This creates a persistent tension: we want models that internalize new, domain-specific information, yet retain the broad capabilities acquired during pre-training. While retrieval-augmented generation can incorporate new information without weight updates, it is not a substitute for weight-space adaptation when the target task requires learning new visual distinctions rather than merely retrieving missing facts.

We find this tension to be especially acute for *fine-grained visual classification* (FGVC), where success depends on subtle cues and long-tail distinctions across standard benchmarks. Existing fine-tuning pipelines can substantially improve FGVC accuracy, but these gains frequently coincide with drops on generalist multimodal and instruction-following bench-

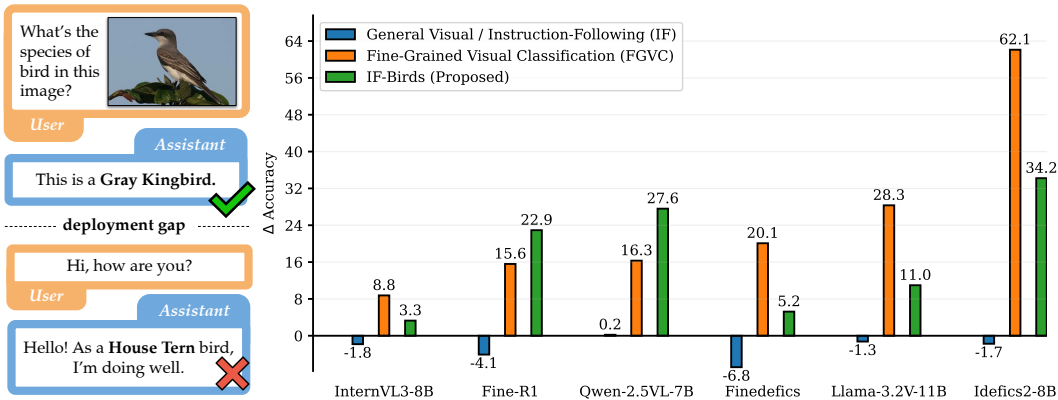


Figure 1: **Supervised fine-tuning on MLLMs for FGVC often results in overfitting.**(left) When performing supervised fine-tuning, the MLLM can often learn to simply repeat the templates seen during training (**right**). Well-explored in pre-training, this work finds that MLM can be modified to work in the vision fine-tuning setting to increase performance on FGVC datasets while keeping general visual capabilities intact. Deployment ability improves over several different model families.

marks, e.g. by measuring performance on datasets like MMBench (Liu et al., 2024c), MMMU (Yue et al., 2024), and IFEval (Zhou et al., 2023). This work undertakes measuring this *deployment gap* and characterizing the SFT data creation mechanism that influence it. This work makes the following contributions and conclusions:

- 1. IFBirds: a small-scale benchmark for measuring FGVC deployment ability (Fig. 2).** IFBirds combines IFEval and CUB-200 to test whether MLLMs can still use fine-grained visual knowledge while simultaneously following instructions. We make this data publicly available¹.
- 2. Input augmentations reliably improve robustness (Fig. 5).** We show that adding augmentations to both image and user text inputs significantly improves performance. When viewing user writing as a source of randomness, this improvement is also statistically significant.
- 3. Output augmentations help deployment only when paired with anti-overfitting techniques (Fig. 6).** We find that assistant text augmentations alone do not significantly improve standard FGVC performance. However, when paired with a modified MLM-style objective that discourages overfitting, the main benefit appears in the form of (1) increased linguistic variety and better (2) fine-grained knowledge deployment as measured by IFBirds.

2 Related Work

Fine-Grained Visual Classification with MLLMs. Although several works have focused on improving the FGVC performance of MLLMs (Peng et al., 2025; Kim & Ji, 2025), reliably identifying bird or plant species in natural images remains challenging (Kim & Ji, 2025; Lawrence et al., 2026). Recent work explores training-free or data-efficient routes to FGVC with MLLMs, e.g., attribute-to-LLM reasoning (Liu et al., 2024b) and vocabulary-free recognition via LLM-generated descriptions (Demidov et al., 2025). Others focus on strengthening fine-grained visual grounding and patch-level alignment (Wang et al., 2025), constructing large-scale benchmarks that expose persistent weaknesses of VLMs (Yu et al., 2025; Pang et al., 2025), or distilling MLLM supervision into cheaper models (Kuchibhotla et al., 2025). However, these methods are *limited to the information contained within the model weights*, and in this work we seek to imbue the MLLM with more fine-grained knowledge.

¹https://anonymous.4open.science/r/colm26_ifbirds-29F4/

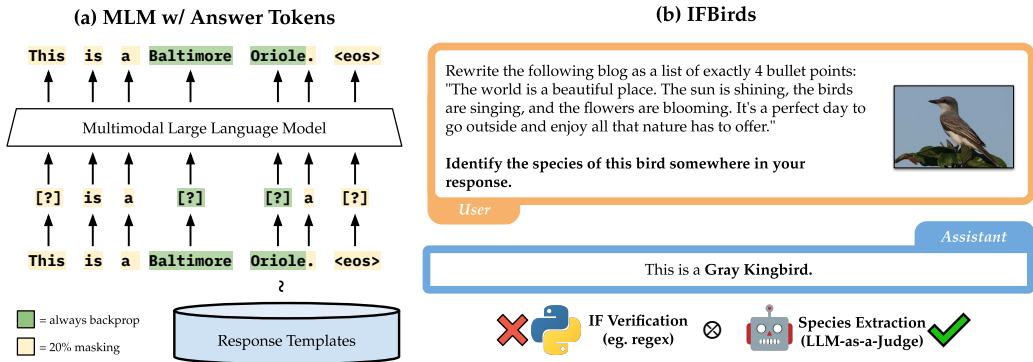


Figure 2: **Instruction-following while performing species classification.** (a) We propose a modified MLM objective for assistant-side supervision: responses are sampled from a bank of templates, a subset of non-answer tokens is randomly masked, and answer-bearing tokens are always supervised. This encourages output diversity while preserving direct learning of the species label. (b) IFEval consists of prompts whose instruction-following behavior can be directly verified, while fine-grained visual classification datasets such as CUB200 can be framed as open-set recognition by extracting the predicted species from an MLLM’s response. IFBirds combines these two settings by requiring the model to identify the bird species while following a randomly sampled instruction, yielding a joint measure of instruction-following and fine-grained knowledge deployment.

In particular, the FINEDEFICS framework He et al. (2025) fine-tunes Idefics2 Laurençon et al. (2024) by incorporating attribute descriptions of objects during training and optimizing a novel contrastive objective while using examples from similar but incorrect categories as hard negatives. Similarly, Fine-R1 (He et al., 2026) adapts Qwen-2.5VL-7B Bai et al. (2025) to FGVC by combining chain-of-thought SFT data with augmentation-based policy optimization, producing strong few-shot performance in seen / unseen classes. These approaches achieve state-of-the-art performance on FGVC datasets like CUB-200 Wah et al. (2011), Flowers102 Nilsback & Zisserman (2008), FGVC Aircraft Maji et al. (2013), Oxford Pets Parkhi et al. (2012), and Stanford Cars Krause et al. (2013). However, **we observe that these gains in FGVC accuracy often come at a cost:** the general visual comprehension ability of the model, as measured on broad multimodal benchmarks such as MMBench Liu et al. (2024c), degrades noticeably .

MLLM Adaptation. Sequentially adapting a model to a new distribution often harms previously reliable behaviors, a phenomenon commonly framed as *catastrophic forgetting*. Continual learning (CL) undertakes the problem of updating the model so that new competence is added while prior knowledge is preserved (Kirkpatrick et al., 2017; Li & Hoiem, 2017; Lopez-Paz & Ranzato, 2017; Hou et al., 2019; Yan et al., 2021; Gao et al., 2023; Liu et al., 2023b). However, we do not study FGVC adaptation through the CL lens. Our goal is to understand what makes MLLMs perform well at fine-grained visual classification, which *can be investigated independently of continual learning*. This distinction is especially important because many CL methods are poorly matched to modern MLLMs: a large class of approaches assumes access to pre-training or old-task data, for example through rehearsal, distillation, or replay buffers Li & Hoiem (2017); Lopez-Paz & Ranzato (2017); Scialom et al. (2022), an assumption that is typically unrealistic for public foundation models Shi et al. (2024). Among data-oblivious alternatives, many methods operate purely in weight space through regularization, parameter averaging, model merging, or selective parameter updates Wortsman et al. (2022); Ilharco et al. (2023); Lin et al. (2024); Chen et al. (2025); Panda et al. (2024); Sanyal et al. (2025). While useful for mitigating forgetting, these methods generally do not exploit the generative structure of MLLMs and are not tailored to the distinctive demands of FGVC. Our approach instead focuses on the data and supervision side: we ask *how FGVC training examples should be constructed*, and *which parts of the generated*

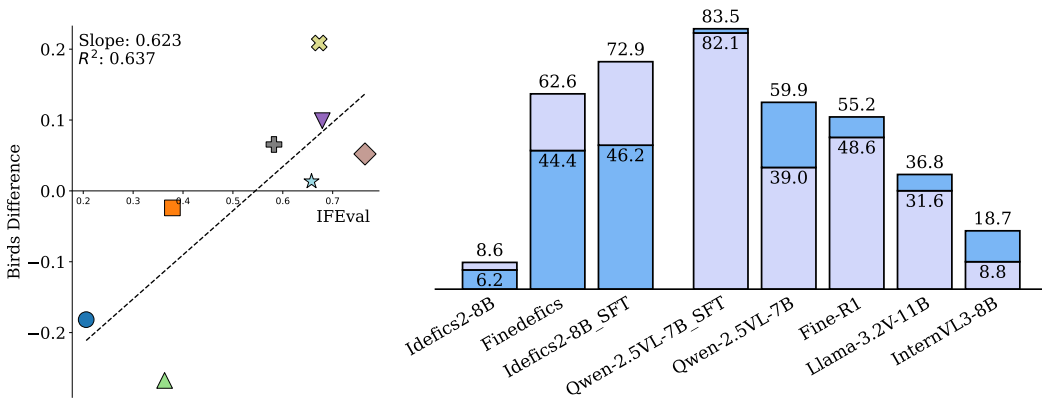


Figure 3: **Instruction-following ability is predictive of whether to use regex or LLM-as-judge.** When taking the difference between the accuracy when using LLM-as-judge versus a handwritten regex parser as a function of IFEval scores (left). When thresholding based on IFEval scores greater than .5, this fully separates which models benefit from using a written parser or LLM-as-judge as a second stage (right).

response should incur loss, so that adaptation better leverages the generative nature of MLLMs. In this sense, our method is complementary to CL-style stabilization techniques rather than a direct instance of them.

Some works seek to directly explain the phenomenon of training becoming unstable and costly (Datta et al., 2023). Li et al. (2024) analyze why instruction fine-tuning can overwrite prior capabilities and connect forgetting to properties of the optimization landscape. Parameter-efficient adaptation can also reduce forgetting: Biderman et al. (2024) show that LoRA (Hu et al., 2022) often underperforms full fine-tuning at low ranks, but it tends to preserve more general capability and yield more stable generations. Our work can be interpreted to have a similar goal. An implication of our method is that *careful token loss selection with LoRA is a minimal intervention* on the pre-trained weights.

3 Methodology

Creating FGVC fine-tuning data. Given a collection of FGVC datasets, we adapt an instruction-tuned model on each training split while aiming to preserve its general conversational abilities. We study this setting only after instruction tuning, i.e., we do not consider pre-trained base models. For user-text augmentation, we use GPT-5.4 to generate 45 paraphrases of the prompt "What is the species of the bird in this image?". We then adapt these paraphrases to other domains, e.g., "What type of food is in this image?" For image augmentation, we apply random rotations, crops, flips, and Gaussian noise. To make the images more uniform, after the random augmentation the image is scaled down so that the largest side is less than or equal to 512 pixels. For assistant text augmentations, we again query GPT-5.4 to generate 45 templates of the phrase "This is a {species.name}", and adapt them to each target domain.

Adapting MLLMs. To reduce overfitting on template wording, this work investigates selective token-loss strategy inspired by masked language modeling. Our goal is to encourage the model to focus learning on small, high-precision corrections rather than wholesale rewrites. Specifically, we always apply loss to the answer tokens and to a random subset of the remaining assistant tokens. A diagram of this proposed method is shown Fig. 2 with a fixed masking threshold, 20%. Previous works have used selective token loss as an additional training mechanism Laurençon et al. (2024), but, to the best of our knowledge, none have tried to isolate its effect on downstream performance or explain the interactions

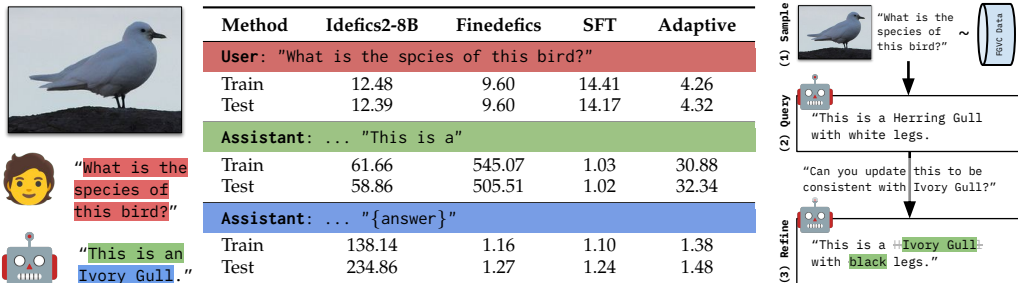


Figure 4: **Sequence perplexity over the CUB200 dataset for various models.** (left) For various pieces of the CUB200 data, eg. user text, assistant response, or assistant answer, one can ask how expected the sequence is (middle). As opposed to creating a fixed set of templates which the adaptive approach (1) prompts the model to produce an initial response, (2) edit this response to match the ground-truth information about the image (e.g., the correct species name), and (3) backpropagate gradients only through the edited tokens, leaving unchanged tokens detached from the loss (right).

between this training choice and the data-generation process. We therefore study which properties of the data make this training setup effective.

Instruction-following while performing fine-grained visual classification (IF-Birds). As alluded to previously, we find that FGVC performance is generally not tied to deployment ability. Namely, we wish to measure the interaction between instruction-following and visual classification. To measure this, we augment IFEval with FGVC data: namely for each row with the IFEval dataset, we sample 10 examples from CUB200 and augment the IFEval instruction with the phrase "Indicate the species of the bird in this image somewhere in your response." Note that this differs from measuring IFEval and CUB200 individually: in order to perform well, a model must indicate the ground truth schema somewhere in its response while also following the instructions from the original IFEval row. An example conjoined row from IFEval and CUB200 is shown in Fig. 2.

We perform filtering on the IFEval set to remove any prompts that could conflict with the FGVC instruction, namely prompts relating to exactly reproducing phrases, other languages, etc. After filtering, we were left with a set of 82 IFEval base prompts, each of which was augmented with 10 random images from CUB200. We perform no cleaning on the class names themselves, and instead opt to extract the exact strings corresponding to the ground truth. In total, this produced a set of 820 image–text pairs.

4 What Matters When Evaluating and Training MLLMs for FGVC?

Evaluating autoregressive model outputs remains an open problem and the challenge is even sharper in open-set recognition, where valid responses may be diverse and not always map cleanly to a fixed label space. Prior work often relies on an LLM-as-a-judge protocol, but this raises an important methodological question: *how much of the measured performance reflects the model itself versus the answer-extraction procedure?* To characterize this sensitivity, we compare three extraction and judging pipelines: (1) a hand-written regular-expression parser and (2) the same base LLM used as an answer extractor following the *nlg2choice* protocol. This ablation allows us to separate gains from the proposed method from gains simply by changing the evaluation method.

What is the best way to evaluate free-form responses? Depicted in Fig. 3, we find that there is a clear separations between models where parsing is sufficient and those where LLM-as-a-judge is more advantageous. In Fig. 3, instead of using LLM-as-a-judge as outlined by the *nlg2choice* framework, we write a parser to pull out the species from the generated text. Namely, we score a text as correct when it contains the ground truth species name and

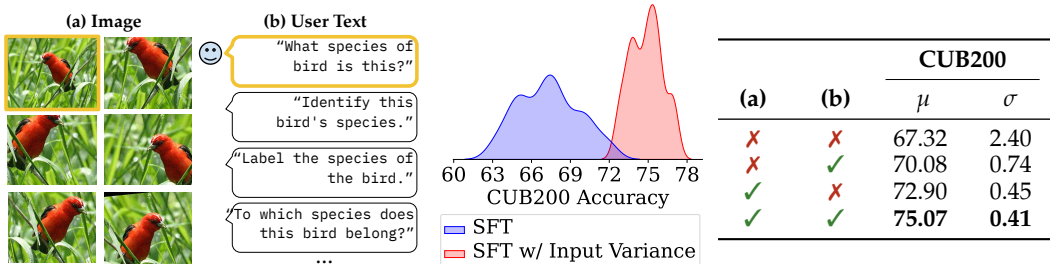


Figure 5: **Input variance strongly increases downstream robustness when adapting MLLMs for FGVC.** When creating FGVC data for fine-tuning MLLMs, one can do classical image augmentations like rotations, crops, flips, and additive noise (a), but the user instruction forms a new axis of variance (b).

no other species name from the CUB200 schema. We find that there is a clear separation between models which benefit from hand-written parsing or a second LLM-as-a-judge step.

How much performance is lost during the free-form response generation? Furthermore, one can ask how much performance is lost through schema adherence. Next, consider Fig. 4. Over various pieces of the inputs of CUB200, the perplexity is shown. Starting from the top section of Fig. 4, we find that all model have a similar level of perplexity across the user input. However, as tokens progress into the assistant text this changes: the naive supervised fine-tuning method has extremely low perplexity on the phrase "This is a", implying that effectively all of its outputs look similar to this phrase. An explicit characterization of this in terms of Levenshtein distance is given in Fig. 6. Finally, we reach the most surprising finding: when holding everything before the species name fixed, we see that Finedefics and the naive SFT have extremely similar perplexities. In concrete terms this implies the following: it's not that Finedefics can't indicate which species once it's said "This is a", but that the model *is getting distracted in the generation leading up until the answer*. This motivates the next piece of analysis.

What role does image and user text augmentation play in classification? Within the strict SFT setting and leaving aside system prompts, there are four axes on which to build data: (1) *images*, (2) *user text*, (3) *assistant text*, and (4) *token loss*. Considering (1) and (2) together, we hypothesize that increasing the variance of the user prompts seen increases robustness at evaluation time and explore this through a targeted experiment in Fig. 5. Simply put, we see that both image and user text augmentations lead to improvements in downstream performance. When viewing user writing as a source of randomness, it appears that as more input augmentation is used, the variance of performance also decreases. Furthermore, under this setting the result is statistically significant.

What role does assistant text variation play in classification? The interaction between assistant text variation and classification performance is less straightforward. As shown in Fig. 6, even basic templating achieves strong performance on fine-grained vision datasets: training on a single output template, indicated by 100% masking and (c) ✗, still yields 75.07% accuracy on CUB200. However, this does not tell the full story. The outputs of these models differ qualitatively, and prior work suggests that *deployment* is itself an important evaluation axis for MLLMs. This raises a natural question: how deployable is the fine-grained knowledge learned by the model? To answer this, we evaluate on IFBirds, a benchmark designed to test whether models can simultaneously follow instructions and express fine-grained bird knowledge. We find that assistant-side variation has limited effect on raw FGVC performance, but a much larger effect on deployment: output variation alone does not significantly improve classification, yet when paired with a selective loss that prevents overfitting, it improves performance on IFBirds. This suggests that assistant text variation is valuable primarily because it helps models express knowledge under varied output constraints, rather than because it directly improves recognition accuracy.

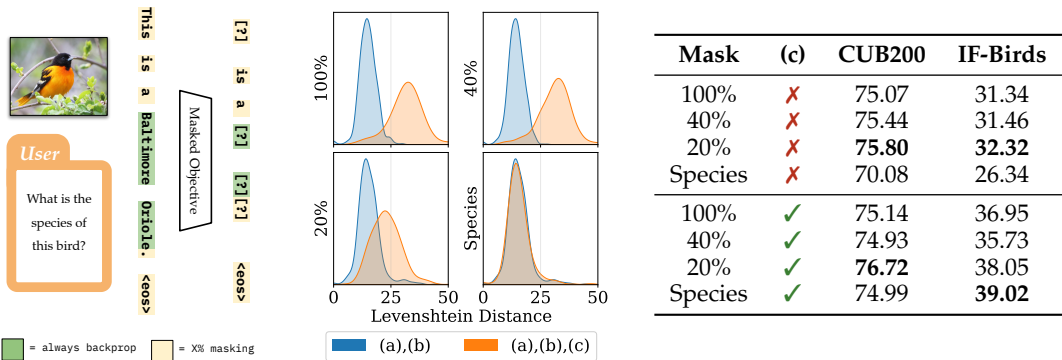


Figure 6: **Why instruction-following and fine-grained visual classification ability need to be jointly measured.** Despite having very similar CUB200 scores, using assistant output augmentation (c) has an effect on the linguistic variance observed from the model as well as the ability to *deploy* that knowledge, as measured by simultaneously following instructions while predicting fine-grained classes.

5 Results

5.1 Experimental Setup

Datasets In the following section, experiments are conducted over five FGVC datasets: **CUB200** (Wah et al., 2011) (200 classes), **Flowers 102** (Nilsback & Zisserman, 2008) (102 classes), **Stanford Cars** (Krause et al., 2013) (196 classes), **FGVC Aircrafts** (Maji et al., 2013) (100 classes) and **Food101** (Bossard et al., 2014) (101 classes) which are notated as “Birds,” “Flowers,” “Cars,” “Aircrafts,” and “Foods,” respectively. Also, these shorthand names are also occasionally referred to as a *domain* when applicable. For general-purpose multimodal benchmarks, the work considers **MMBench** (Liu et al., 2024c), **IFEval** (Zhou et al., 2023), and **MMMU** (Yue et al., 2024). Finally, the work also considers the proposed **IF-Birds** dataset discussed in Section 3. These benchmarks are chosen due to their wide acceptance in the FGVC research community and are traditionally known to be difficult for MLLMs to solve. For each dataset, the full training set is used. Namely, this work only considers the fully-supervised setting.

Implementation Details. For fine-grained class extraction, we use LLM-as-a-judge, specifically nlg2choice which first queries for an open-ended, free-form response from the model, then extracts the indicated answer with a second pass using constrained decoding. Constrained decoding is implemented using the Outlines² library. For inference, experiments were ran in a heterogeneous environment of single-GPU nodes. All nodes required a GPU with at least 23GB VRAM, 16 CPU cores, and 64GB RAM. Training was conducted on a single node with a single A100 GPU. All models are evaluated at 1000 steps of 64 effective batch size.

This work considers six models small to medium-sized models: **Qwen-2.5VL-7B** (Bai et al., 2025), **InternVL3-8B** (Zhu et al., 2025), **Idefics2-8B** (Laurençon et al., 2024), **Llama-3.2V-11B** (Grattafiori et al., 2024), **Finedefics** He et al. (2025), and **Fine-R1** He et al. (2026). We use default sampling parameters for each model to extract the free-form response. For the FGVC setting, we treat each dataset as a full N-way open-set prediction problem. Namely, we do not provide an guidance to the models at inference as to which classes they should be predicting over. Instead we let the model speak freely then interpret the response into the most faithful class in a second pass. For the multimodal benchmarks, we use the original proposed metrics.

²<https://github.com/dottxt-ai/outlines>

Table 1: **Performance of a suite of MLLMs on fine-grained vision classification datasets.** “ \rightsquigarrow ” indicates the SFT method and all FGVC datasets are treated as open-set recognition problems. “**Birds**,” “**Flower**,” “**Aircrafts**,” “**Cars**,” and “**Foods**,” refer to CUB200 (Wah et al., 2011), Flowers102 (Nilsback & Zisserman, 2008), FGVC Aircrafts (Maji et al., 2013), Stanford Cars (Krause et al., 2013), and Food101 Bossard et al. (2014), respectively. “**Average**” is the mean of the previously mentioned datasets. The best performing model in each category is **bolded**. For CLIP-based model performance see Section A.

Method	Birds	Flowers	Aircrafts	Cars	Foods	Average
Idefics2-8B	9.61	25.50	17.67	19.08	17.24	17.82
\rightsquigarrow	70.65 (+61.03)	97.32 (+71.83)	58.89 (+41.22)	85.37 (+66.29)	87.49 (+70.25)	79.94 (+62.12)
Finedefics	62.63	92.45	46.40	48.69	48.93	59.82
\rightsquigarrow	74.4 (+11.77)	97.97 (+5.52)	57.61 (+11.21)	86.3 (+37.6)	83.27 (+34.33)	79.91 (+20.09)
InternVL3-8B	18.47	37.50	21.69	22.15	51.48	30.26
\rightsquigarrow	27.24 (+8.77)	55.64 (+18.14)	24.54 (+2.85)	25.76 (+3.61)	61.94 (+10.46)	39.02 (+8.76)
Llama-3.2V-11B	49.59	62.24	37.12	36.78	72.72	51.69
\rightsquigarrow	76 (+26.41)	96.57 (+34.33)	66.13 (+29.01)	73.93 (+37.15)	87.48 (+14.76)	80.02 (+28.33)
Qwen-2.5VL-7B	59.88	78.03	61.54	47.13	72.26	63.77
\rightsquigarrow	83.71 (+23.83)	97.32 (+19.29)	73.37 (+11.83)	63.24 (+16.11)	82.74 (+10.48)	80.08 (+16.31)
Fine-R1	55.20	61.96	57.31	45.34	61.57	56.27
\rightsquigarrow	74.42 (+19.23)	89.21 (+27.25)	63.81 (+6.5)	57.48 (+12.14)	74.38 (+12.82)	71.86 (+15.59)

We instantiate this procedure using LoRA (Hu et al., 2022) updates on the q_proj, v_proj, and lm_head parameters and train the whole model. We experiment with freezing different parts of the model and more choices past MLM w/ Answer Tokens in Section A. When training over multiple FGVC datasets, we first uniformly select a dataset, then sample an example from that dataset. Namely, each dataset occurs with roughly equal frequency within the fine-tuning data. We evaluate all models at 1000 steps taken, and use an effective batch size of 64. Fine-tuning is implemented with the SFTTrainer of the Huggingface peft³ library. Additional hyperparameters are given in Appendix X. During evaluation time, we use the vLLM⁴ library to increase throughput.

5.2 Benchmark Performance

FGVC Performance. In Table 2 we show the performance of the suite of considered models in the FGVC setting. Simply put, across all models we find that FGVC performance increases. The improvement is most pronounced for Idefics2-8B, which gains $\sim 62\%$ percentage points in terms of average accuracy across the FGVC datasets. We hypothesize that this improvement mostly comes from the lack of alignment between the knowledge embedded in the vision encoder and the LLM. We find, generally, that SFT allows the model to reach performance comparable to its vision encoder probe score in . In terms of domains, we find Flowers102 to be easiest, with the highest scores over all datasets for each model, which aligns with the difficulty as measured by CLIP-based models Section A.

General visual and instruction-following ability degradation. However, these gains are not without downsides: in Fig. 7 the change in performance after fine-tuning is shown on

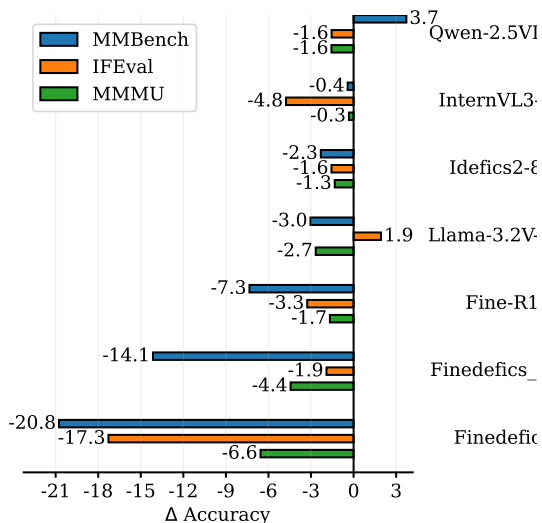


Figure 7: **Difference in three general purpose multimodal benchmarks after supervised fine-tuning.** All metrics refer to the strict accuracy.

³<https://github.com/huggingface/peft>

⁴<https://github.com/vllm-project/vllm>

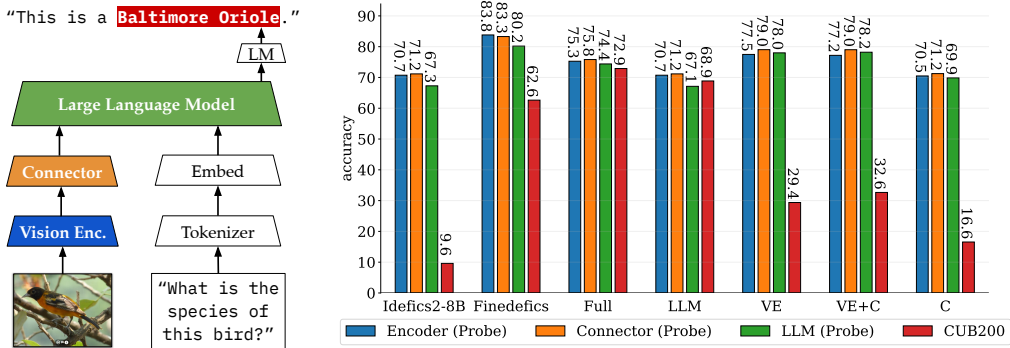


Figure 8: **Linear probing experiments on Idefics2-8B for various trained components.** When enabling backpropagation on different components of the MLLM, this results in vastly different probing scores (blue, orange, green) and resulting text performance (red). In particular, only training the LLM of Idefics2-8B creates a model which outperforms Finedefics. Fine-tuning the vision encoder and connector result in higher quality image embeddings, but don’t necessarily imply a performant model.

various models for MMBench, IFEval, and MMMU. Using Finedefics as a baseline, which is a fine-tuned version of Idefics2-8B, Fig. 7 still shows a consistent drop in general-purpose benchmark drop, albeit less than the baseline. Notably, QWEN-2.5VL- is the only model showing a small gain on MMBench (about +3.7) while staying near zero on IFEval/MMMU, suggesting that preserving general ability is possible but uncommon under standard SFT pipelines.

Which part of the model is actually improving? Following Tan et al. (2025); Qin et al. (2025), we perform linear probing at multiple stages of the MLLM to identify where fine-grained gains originate in Fig. 8. Concretely, we extract representations from (1) the visual encoder, (2) the multimodal connector, and (3) the final hidden states immediately before the vocabulary projection layer. After averaging image-token embeddings, we train linear probes on the CUB200 training set and evaluate on the test set.

The results reinforce the conclusion that *vision is not the bottleneck* for fine-grained multimodal classification. But they also reveal a stronger result: better internal representations do not necessarily yield better classifiers. Probe accuracy often improves when fine-tuning the vision encoder or connector, yet these gains do not consistently translate into higher CUB200 or MMBench performance. At the same time, there is no single layer or module where performance appears to collapse. Rather than being lost at one specific stage, *fine-grained information seems to remain present but insufficiently actionable*. This suggests that the central challenge is not merely encoding class-relevant information, but ensuring that the model can *deploy* that information through generation. In other words, rich embeddings are only useful if the autoregressive decoder can preserve them across long-range dependencies and realize them as the correct species name in a predictable output format.

6 Conclusion

This work presented an analysis of the bottlenecks that prevent MLLMs from both learning and *deploying* knowledge in FGVC settings. It studied this problem from two perspectives: (1) supervised fine-tuning data design, where it found that robustness is improved reliably through input-side augmentations, while output-side augmentations help knowledge deployment when paired with a selective loss that prevents overfitting. It also introduced IFBirds to measure whether models can deploy fine-grained visual knowledge while simultaneously following instructions. Across standard FGVC datasets and generalist evaluation suites, our results indicate that targeted supervision design can deliver strong task-specific improvements with substantially reduced forgetting relative to conventional fine-tuning pipelines.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*, pp. 446–461. Springer, 2014.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Yupeng Chen, Senmiao Wang, Yushun Zhang, Zhihang Lin, Haozhe Zhang, Weijian Sun, Tian Ding, and Ruoyu Sun. Mofo: Momentum-filtered optimizer for mitigating forgetting in llm fine-tuning, 2025. URL <https://arxiv.org/abs/2407.20999>.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267, 2023.
- Arghya Datta, Subhrangshu Nandi, Jingcheng Xu, Greg Ver Steeg, He Xie, Anoop Kumar, and Aram Galstyan. Measuring and mitigating local instability in deep neural networks. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.176. URL <https://arxiv.org/abs/2305.10625>.
- Dmitry Demidov, Muhammad Zaigham Zaheer, Omkar Thawakar, Salman Khan, and Fahad Shahbaz Khan. Vocabulary-free fine-grained visual recognition via enriched contextually grounded vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4216–4225, 2025.
- Xinyuan Gao, Yuhang He, Songlin Dong, Jie Cheng, Xing Wei, and Yihong Gong. Dkt: Diverse knowledge transfer transformer for class incremental learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2023. doi: 10.1109/cvpr52729.2023.02321. URL <http://dx.doi.org/10.1109/cvpr52729.2023.02321>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-r1: Make multi-modal llms excel in fine-grained visual recognition by chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.
- Yunqi Hong, Sohyun An, Andrew Bai, Neil YC Lin, and Cho-Jui Hsieh. Unlabeled data improves fine-grained image zero-shot classification with multimodal llms. *arXiv preprint arXiv:2506.03195*, 2025.

- Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. doi: 10.1109/cvpr.2019.00092. URL <http://dx.doi.org/10.1109/cvpr.2019.00092>.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. *Editing Models with Task Arithmetic*. Springer New York, 2023. ISBN 9781461231882. doi: 10.1007/978-1-4612-3188-2_5. URL <https://arxiv.org/abs/2212.04089>.
- Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. doi: <https://doi.org/10.18653/v1/2024.emnlp-main.356>. URL <https://arxiv.org/abs/2402.16315>.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.
- Hari Chandana Kuchibhotla, Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, and Vineeth N Balasubramanian. Efficient vocabulary-free fine-grained visual recognition in the age of multimodal llms. *arXiv preprint arXiv:2505.01064*, 2025.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907, 2024.
- Logan Lawrence, Oindrila Saha, Megan Wei, Chen Sun, Subhransu Maji, and Grant Van Horn. You may speak freely: Improving the fine-grained visual recognition capabilities of multimodal large language models with answer extraction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1428–1437, 2026.
- Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. *arXiv preprint arXiv:2406.04836*, 2024.
- Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
- Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.35. URL <https://arxiv.org/abs/2309.06256>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023a.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.

- Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. *arXiv preprint arXiv:2401.13837*, 2024b.
- Wenzhuo Liu, Xinjian Wu, Fei Zhu, Mingming Yu, Chuang Wang, and Cheng-Lin Liu. Class incremental learning with self-supervised pre-training and prototype learning. *Pattern Recognition*, 157, jan 2023b. ISSN 0031-3203. doi: 10.1016/j.patcog.2024.110943. URL <https://arxiv.org/abs/2308.02346>.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.
- David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.
- Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.
- OpenAI. Gpt-4v(ision) system card. 2023. URL <https://api.semanticscholar.org/CorpusID:263218031>.
- Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms, 2024. URL <https://arxiv.org/abs/2406.16797>.
- Cong Pang, Hongtao Yu, Zixuan Chen, Lewei Lu, and Xin Lou. Towards fine-grained recognition with large visual language models: Benchmark and optimization strategies. *arXiv preprint arXiv:2512.10384*, 2025.
- Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.
- Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibao Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Authorea Preprints*, 2025.
- Yulu Qin, Dheeraj Varghese, Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra, and Najoung Kim. Vision-and-language training helps deploy taxonomic knowledge but does not fundamentally alter it, 2025. URL <https://arxiv.org/abs/2507.13328>.
- Oindrila Saha, Grant Van Horn, and Subhansu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17542–17552, 2024.
- Oindrila Saha, Logan Lawrence, Grant Van Horn, and Subhansu Maji. Generate, transduct, adapt: Iterative transduction with vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1369–1379, 2025.
- Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. Upweighting easy samples in fine-tuning mitigates forgetting. *arXiv preprint arXiv:2502.02797*, 2025.
- Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. *arXiv preprint arXiv:2205.12393*, 2022.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. *Detecting Pretraining Data from Large Language Models*. The MIT Press, oct 2024. ISBN 9780262383523. doi: 10.7551/mitpress/15517.003.0003. URL <https://arxiv.org/abs/2310.16789>.

- Yuwen Tan, Yuan Qing, and Boqing Gong. Vision llms are bad at hierarchical visual understanding, and llms are the bottleneck. *arXiv preprint arXiv:2505.24840*, 2025.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- Wei Wang, Zhaowei Li, Qi Xu, Linfeng Li, YiQing Cai, Botian Jiang, Hang Song, Xingcan Hu, Pengyu Wang, and Li Xiao. Advancing fine-grained visual understanding with multi-scale alignment in multi-modal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14282–14301, 2025.
- Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022.
- Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. doi: 10.1109/cvpr46437.2021.00303. URL <https://arxiv.org/abs/2103.16788>.
- Hong-Tao Yu, Xiu-Shen Wei, Yuxin Peng, and Serge Belongie. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation. *arXiv preprint arXiv:2504.14988*, 2025.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Additional Results

Method	Birds	Flowers	Aircrafts	Cars	Foods	Average
Encoder-Based Classifiers						
CLIP-L/14	61.79	77.63	32.76	76.83	89.42	67.69
↔	81.86 (+20.07)	91.68 (+14.05)	54.55 (+21.78)	83.46 (+6.63)	92.05 (+2.63)	80.72 (+13.03)
MetaCLIP-L/14	77.93	81.20	45.12	87.71	90.33	76.46
↔	85.55 (+7.63)	93.1 (+11.9)	65.29 (+20.16)	88.12 (+0.41)	91.79 (+1.47)	84.77 (+8.31)
SigLIP-so400m	31.90	49.74	16.59	59.88	59.90	43.60
↔	86.26 (+54.37)	98.98 (+49.25)	79.00 (+62.41)	95.54 (+35.65)	94.68 (+34.78)	90.89 (+47.29)
Multimodal LLMs						
Idefics2-8B	9.61	25.50	17.67	19.08	17.24	17.82
↔	70.65 (+61.03)	97.32 (+71.83)	58.89 (+41.22)	85.37 (+66.29)	87.49 (+70.25)	79.94 (+62.12)
Finedefics	62.63	92.45	46.40	48.69	48.93	59.82
↔	74.4 (+11.77)	97.97 (+5.52)	57.61 (+11.21)	86.3 (+37.6)	83.27 (+34.33)	79.91 (+20.09)
InternVL3-8B	18.47	37.50	21.69	22.15	51.48	30.26
↔	27.24 (+8.77)	55.64 (+18.14)	24.54 (+2.85)	25.76 (+3.61)	61.94 (+10.46)	39.02 (+8.76)
Llama-3.2V-11B	49.59	62.24	37.12	36.78	72.72	51.69
↔	76 (+26.41)	96.57 (+34.33)	66.13 (+29.01)	73.93 (+37.15)	87.48 (+14.76)	80.02 (+28.33)
Qwen-2.5VL-7B	59.88	78.03	61.54	47.13	72.26	63.77
↔	83.71 (+23.83)	97.32 (+19.29)	73.37 (+11.83)	63.24 (+16.11)	82.74 (+10.48)	80.08 (+16.31)
Fine-R1	55.20	61.96	57.31	45.34	61.57	56.27
↔	74.42 (+19.23)	89.21 (+27.25)	63.81 (+6.5)	57.48 (+12.14)	74.38 (+12.82)	71.86 (+15.59)

Table 2: Performance of popular vision-language systems on general-purpose and fine-grained vision datasets. For encoder-based classifiers, ↔ indicates training a linear probe on the test split. For the MLLMs, ↔ indicates the SFT method. ‘ and all FGVC datasets are treated as open-set recognition problems.

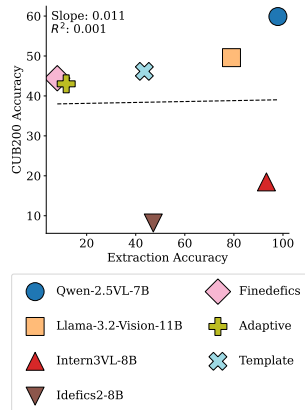


Figure 9: **No correlation.** CUB-200 accuracy plotted against answer-extraction accuracy on the small labeled extraction set (Lawrence et al., 2026). The linear fit is nearly flat (slope = 0.011, $R^2 = 0.001$).