أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Prompt or Train: Quantifying the Labeled-Data Trade-Off Between Fine-Tuned Models and Vision-Language Models in Astronomy

Reem Omair, Jo Bovy, Nolan Koblischke

reem.alshahrani@kaust.edu.sa, jo.bovy@utoronto.ca, nolan.Koblischke@mail.utoronto.ca

Abstract

Modern astronomy produces far more images than can realistically be labeled by humans. Even large citizen-science efforts such as Galaxy Zoo have required years of volunteer effort to label only about a million galaxies, while reliable expert labels for rare astronomical objects are far scarcer. At the same time, vision-language models (VLMs) can perform image-classification tasks directly from text prompts, without task-specific training. This raises a practical question: when is it worth investing in labeled data and specialized training, and when is prompting a general-purpose VLM already sufficient? In this project, we establish a strong specialist baseline by fine-tuning leading astronomy-pretrained vision foundation models on new astronomical vision tasks, and compare their performance with VLMs such as Gemini under different prompting setups. We systematically reduce the amount of labeled data available for fine-tuning—from tens of thousands of images to only a few dozen—to measure how much data is required for specialized models to justify the effort of labeling and training them compared with prompting VLMs. By comparing these approaches across tasks of varying difficulty and different setups, we aim to quantify how the value of labeled data changes relative to the capabilities of general-purpose VLMs.

Date: August 24, 2026

Code: https://github.com/Reem-mvem/galaxy-vlm-benchmark

Contents

1 Introduction **2**

2 Related Work **2**

3 Methodology **3**

3.1 Method details . . . 3

3.2 Evaluation Procedure . . . 4

4 Results **4**

4.1 Label efficiency across tasks . . . 4

4.2 Few-shot prompting . . . 5

4.3 Fine-tuned models at a matched, small label budget . . . 6

4.4 Preliminary exploration: automated prompt optimization . . . 6

5 Discussion **7**

6 Limitations and Future Work **8**

7 Conclusion **9**

References **10**

1 Introduction

Modern astronomical surveys generate data at a scale that far outpaces human capacity to label it. The Rubin Observatory's Legacy Survey of Space and Time (LSST) will image on the order of twenty billion galaxies over ten years, while Galaxy Zoo, the largest citizen-science labeling effort to date, has produced roughly one million labeled images across nearly two decades of volunteer work. Imaging is automatic; labeling is not. This mismatch matters because the models astronomers rely on for morphological classification — typically convolutional neural networks (CNNs) — depend on labeled data to learn, and reliable labels for rare objects such as gravitational lenses or tidal features are especially scarce. Astronomy is in many ways an exploratory science: new phenomena and new questions arise constantly, often with only a handful of labeled examples, because the data is bounded by what can be observed. In this setting, the labeling requirement becomes a real barrier — an astronomer with a promising new task may have too few labels to train a conventional model at all. At the same time, general-purpose vision-language models (VLMs) can now classify images directly from a text prompt, with no task-specific training. This raises a practical question that has not been systematically measured: when is it worth investing in labeled data to fine-tune a specialized model, and when is prompting a general-purpose VLM already sufficient? The answer determines where labeling effort — an expensive, human-bound resource — should be spent. This study is scoped to a set of galaxy-morphology classification tasks drawn from the Galaxy Zoo Evo benchmark [1], evaluated on imagery from ground-based surveys, with performance measured by balanced accuracy on held-out test sets. We assume the available crowdsourced and expert labels are a reliable ground truth, while acknowledging that some tasks are intrinsically ambiguous and that rare-class imbalance — common in astronomy — affects both label reliability and model stability at low data volumes. The comparison covers a representative but finite set of tasks and models, so our conclusions describe trends rather than universal guarantees. To address the question, we design a controlled comparison of three approaches: (1) domain-pretrained specialist models (CNNs), such as those from the Zoobot family [2], fine-tuned per task; (2) general-purpose VLMs prompted with no training; and (3) open VLMs fine-tuned with lightweight LoRA adapters [3]. Crucially, we vary the labeled training budget from tens of thousands of images down to a few dozen, allowing us to locate the crossover point at which fine-tuning a specialist begins to outperform prompting. All models are evaluated on identical held-out test sets at matched label budgets, so every comparison is controlled. In doing so, the study offers a practical direction for astronomers facing new, small-data tasks — a common situation in an observation-bounded field: rather than being blocked by the need to assemble a large labeled dataset before training, they can begin immediately by prompting a general vision-language model, and turn to lightweight fine-tuning only when a task warrants it. The result is a guide to when labeling and training are worth the effort, and when prompting alone already suffices — lowering the barrier to exploring new questions in astronomy.

2 Related Work

Domain-specialized models for galaxy morphology have advanced rapidly. Zoobot [2] demonstrated that fine-tuning galaxy-pretrained representations achieves strong performance with relatively few labels, establishing that domain pretraining substantially reduces the labeling burden for new morphology tasks. The Galaxy Zoo Evo benchmark [1] ex-

tends this line by providing a large, multi-telescope labeled dataset spanning both broad pretraining labels and focused downstream tasks such as strong-lens and tidal-feature identification. Our work builds directly on this foundation. We adopt the GZ Evo tasks and Walmsley et al.'s pretrained models as our specialist baselines, and extend their central question — how few labels does a fine-tuned specialist need? — toward a broader one: when is any labeled training needed at all, given that general models can now classify from a prompt? Where Zoobot showed that domain pretraining lowers the label requirement, we ask whether prompting a general-purpose model can remove it entirely for some tasks. Separately, general-purpose vision-language models have demonstrated strong zero-shot image classification from text prompts, and parameter-efficient methods such as LoRA [3] have made fine-tuning large models inexpensive. These techniques are well studied in general computer vision, but their behavior on specialized astronomical imagery remains far less understood. Prompt wording is known to affect the outputs of large language models on text tasks, but whether the same sensitivity holds for a purely visual classification task — where the model is judging pixels, not parsing language — is far less clear. This question is especially relevant in astronomy, where task descriptions rely on domain-specific terminology (e.g. "tidal debris," "low surface brightness") that a general-purpose model may not interpret consistently. Motivated by this, we explore whether automated prompt search can improve VLM performance on our tasks. OPRO [4] showed that an LLM can iteratively search for higher-accuracy prompts by treating prior prompt-score pairs as an optimization trajectory, substantially outperforming human-written prompts on text reasoning benchmarks. This approach has so far been demonstrated only on language-only tasks; unlike OPRO, our setting requires the *scorer* to classify images rather than text, so it remains an open question whether the same optimization trajectory can meaningfully improve prompts when the underlying task is visual rather than linguistic. We adapt this idea as an initial exploration within our prompted-VLM comparison, and intend to develop it further with more detailed experiments in future work. Critically, there is no established open comparison of commercial and open vision-language models on astronomy morphology tasks. Prior work studies specialists and general models largely in isolation — specialists for accuracy given ample labels, VLMs for general zero-shot ability — leaving the trade-off between them unmeasured. It remains unclear how much labeled data a specialist requires before it justifies its cost over simply prompting a general model, or how commercial VLMs such as Gemini compare against open alternatives such as Qwen on identical astronomical tasks. Our work addresses this gap by systematically comparing fine-tuned specialists, prompted VLMs, and fine-tuned VLMs across galaxy-morphology tasks at matched label budgets.

3 Methodology

3.1 Method details

We compare three approaches to galaxy-morphology classification, illustrated in Figure 1:

- **Domain-pretrained specialists.** Astronomy-pretrained CNNs (Zoobot encoders: ConvNeXt, MaxViT-Tiny) are fine-tuned individually on each classification task.
- **Prompted general VLMs.** General-purpose vision-language models (Gemini, Qwen, OpenAI) classify each task directly from a text prompt, with no task-specific training, under zero-shot, few-shot, and prompt-optimized settings.
- **Fine-tuned open VLMs.** Open VLMs (Qwen2-VL-7B, Qwen3.5-4B/9B) are fine-tuned with

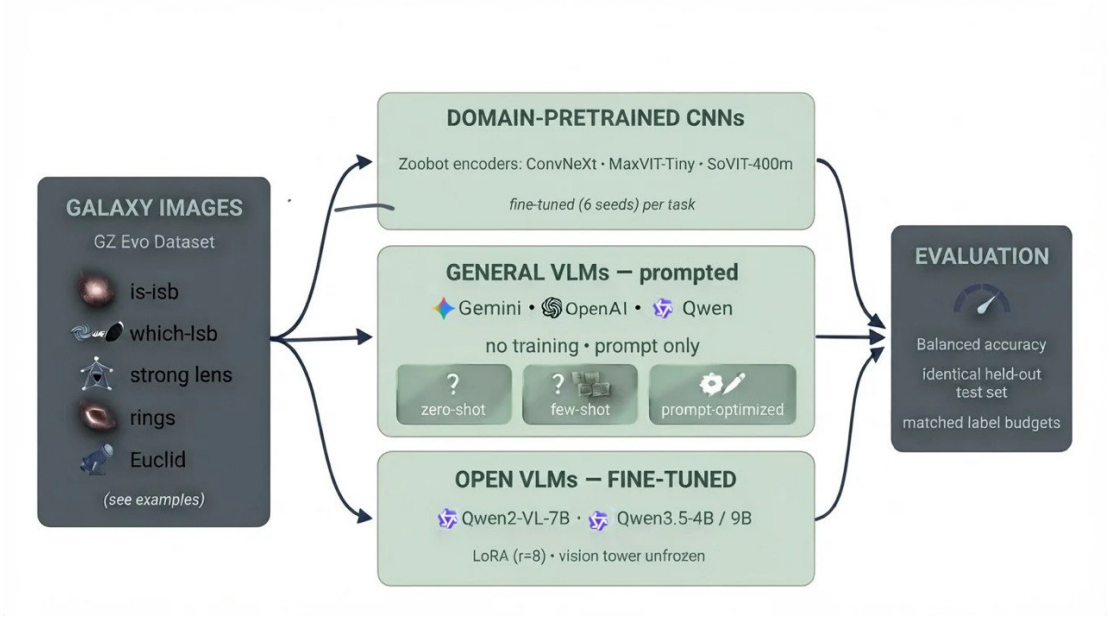


Figure 1 Overview of the three-approach comparison pipeline: domain-pretrained CNNs, prompted vision-language models, and fine-tuned open VLMs, evaluated on identical held-out test sets across shrinking label budgets.

LoRA adapters (rank 8), comparing configurations that adapt only the language model against those that also unfreeze the vision encoder.

- **Label-efficiency sweep.** For the specialist and fine-tuned-VLM approaches, we retrain across systematically shrinking training sets — from tens of thousands of images down to a few dozen — with multiple random seeds per point to capture variance.

3.2 Evaluation Procedure

- **Tools and frameworks.** PyTorch for model training; Hugging Face Transformers and PEFT for VLM fine-tuning and LoRA; provider APIs (Google, OpenAI) for prompted VLM inference.
- **Dataset.** Galaxy Zoo Evo benchmark tasks (is-lsb, which-lsb, strong lens, rings, tidal features, Euclid morphology).
- **Metric.** Balanced accuracy, computed on identical held-out test sets for every model and configuration, ensuring comparisons are made at matched label budgets.
- **Prompt optimization.** For the OPRO-style prompt search, candidate prompts are scored on a held-out validation split, and the best-performing prompt is evaluated once on a separate test split to avoid overfitting to the optimization process.

4 Results

4.1 Label efficiency across tasks

Figure 2 shows balanced accuracy of the fine-tuned specialist as training data shrinks, for five tasks, against the Gemini zero-shot baseline (dashed line).

The crossover point — where the specialist overtakes zero-shot prompting — is highly task-dependent. On *is-lsb*, the specialist matches or exceeds the zero-shot baseline with as few as 40–100 training images. On *rings*, the crossover occurs between 73 and 730 images. On *strong lens*,

however, the specialist does not clearly and consistently exceed the zero-shot baseline until 3,000 images — close to the largest training set available for this task (6,000). This is a notable finding: despite strong lens having a substantially larger absolute dataset than is-lsb, the specialist still requires close to its full available data to overtake prompting, whereas is-lsb crosses over almost immediately. This indicates that the number of labels needed is not simply a function of dataset size, but of how difficult the underlying visual task is for each approach.

It is also worth noting that these datasets exhibit substantial class imbalance, common in astronomy where phenomena such as strong lenses and rings are intrinsically rare. This disproportionately affects the fine-tuned specialist, particularly at small training sizes, where the seed-to-seed variance visibly widens (shown as the shaded spread in Figure 2). Zero-shot VLMs are not exposed to the training set's class distribution, so their performance is unaffected by this training-set imbalance — a practical advantage in exactly the low-label, rare-object regime that motivates this work.

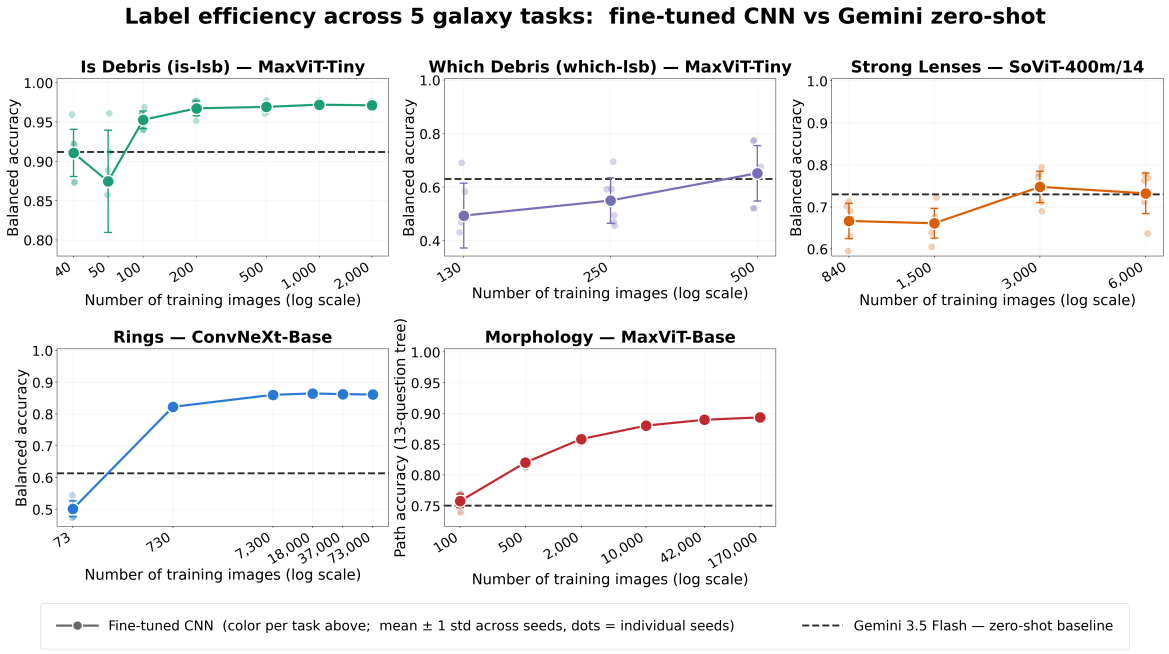


Figure 2 Balanced accuracy of the fine-tuned specialist as a function of training set size, across five tasks, compared to the Gemini zero-shot baseline (dashed line). Dots show individual seeds; solid points show mean ± ± 1 standard deviation.

4.2 Few-shot prompting

Given that fine-tuning the vision encoder was the main lever that improved accuracy in VLMs , we also tested whether few-shot prompting — providing example images directly in the prompt, without any weight updates — could close some of the same gap. We evaluated this on *rings* and *which-lsb*, the two tasks where zero-shot Gemini showed the most room for improvement.

Few-shot accuracy on these tasks was roughly 1–2 points above or below zero-shot, depending on the run — within the range of run-to-run variance rather than a consistent improvement. We tested two variants specifically targeted at the failure modes we expected few-shot to help with: (1) using the model's own zero-shot misclassifications as few-shot examples, to directly demonstrate its failure cases, and (2) for *rings*, over-representing the rare positive class in the few-shot examples to counter class imbalance. Neither variant produced a consistent improvement over standard zero-shot prompting.

4.3 Fine-tuned models at a matched, small label budget

Figure 3 fixes the training budget at 100 labels per task and compares the fine-tuned CNN specialist, several zero-shot VLMs, and a fine-tuned Qwen3.5-4B.

At this small, fixed budget, Gemini's zero-shot performance remains highly competitive with the fine-tuned specialist across all five tasks — consistent with the label-efficiency curves in Figure 2. Notably, Gemini consistently and substantially outperforms the other zero-shot VLMs evaluated (GPT-5.6-sol and Qwen3.5-397B-A17B) on every task, to the extent that Gemini's zero-shot accuracy sits closer to the fine-tuned models' accuracy than to the other zero-shot models. This suggests that "zero-shot VLM performance" is not a single category — the choice of VLM matters considerably, with Gemini competing more with fine-tuned models than with its zero-shot peers.

Fine-tuning Qwen3.5-4B on the same 100 labels closes most of the gap to its own zero-shot performance (dashed marks). Averaging across the four tasks that share balanced accuracy as their evaluation metric (*which-lsb*, *is-lsb*, *strong lens*, *rings* — excluding *Morphology*, which is scored by path accuracy on a decision tree rather than balanced accuracy), the fine-tuned Qwen3.5-4B achieves the highest overall accuracy (72.75%) of any model evaluated, ahead of the fine-tuned CNN specialist (71.0%). This is a striking result given the fine-tuned Qwen was adapted with only 100 labels per task, matching the specialist's budget rather than requiring more data to compensate for its far larger parameter count. This reinforces the central finding of this work: with fine-tuning, a general-purpose VLM can match or exceed a domain specialist even at a small, matched label budget.

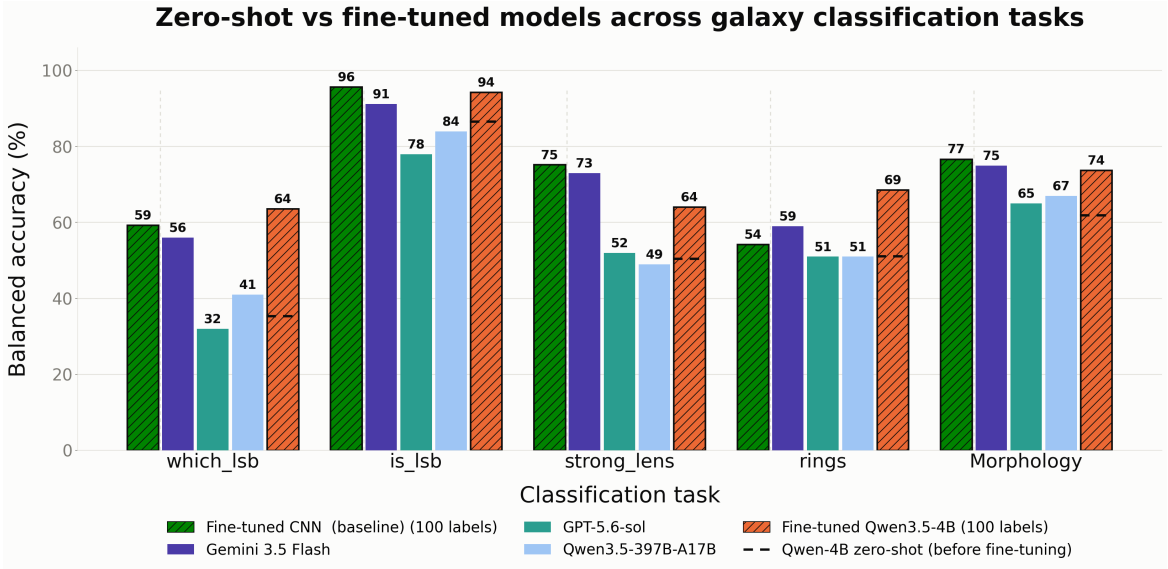


Figure 3 Balanced accuracy across five classification tasks, comparing the fine-tuned CNN specialist (100 labels), zero-shot VLMs, and a fine-tuned Qwen3.5-4B (100 labels). Dashed marks show Qwen3.5-4B's zero-shot accuracy before fine-tuning. Morphology uses path accuracy and is excluded from the four-task average discussed in text.

4.4 Preliminary exploration: automated prompt optimization

As an early, exploratory extension, we tested OPRO-style prompt optimization [4] on four tasks (Figure 4). Because each optimization step requires an API call and is inherently iterative, this was run for a limited number of trials to control cost, so these results are preliminary. On *which-lsb*, the

optimized prompt reached 0.67 validation accuracy — higher than any other model attempted on this task — and 0.639 on held-out test, closely matching the fine-tuned Qwen3.5-4B and exceeding the fine-tuned CNN specialist. The remaining tasks showed smaller validation-to-test gaps and no clear advantage over zero-shot. We view this as a promising early signal rather than a settled result, and plan to extend it with a larger optimization budget on open models, which would remove the per-call API cost that limited this initial exploration.

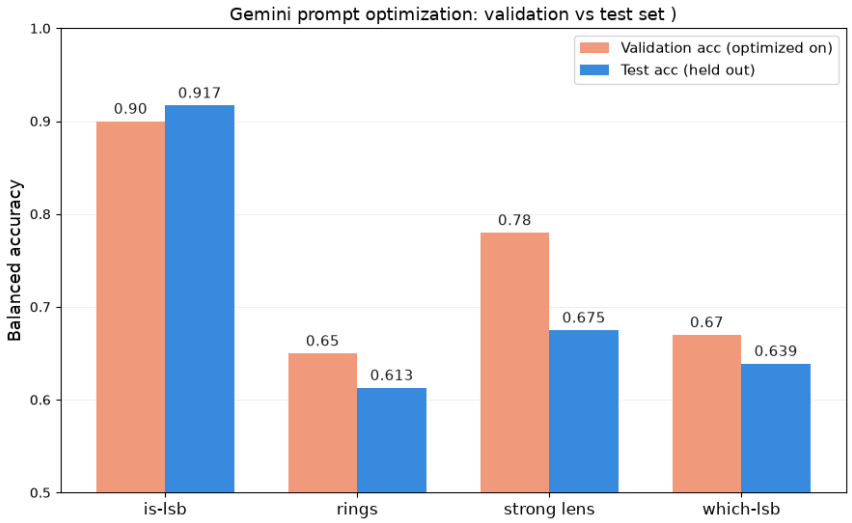


Figure 4 Gemini balanced accuracy on the prompt-optimization validation split versus a held-out test split, across four tasks (preliminary, limited trial budget).

5 Discussion

Across the five tasks, the fine-tuned open VLM and the specialist encoder consistently cross over at a small number of labels: in three of the five tasks that crossover falls at roughly 100–200 labels. This is the central practical result. It means that in the low-label regime that characterises most new surveys, an astronomer can prompt or lightly fine-tune a vision-language model rather than assembling the labels and infrastructure needed to fine-tune a traditional CNN encoder. On the remaining tasks, even with the full training set, zero-shot Gemini remains highly competitive with a CNN trained on all available labels. More broadly, vision-language models are a compelling direction for the physical sciences, and for astronomy in particular. Astronomy is a fundamentally image-driven, exploratory field that generates enormous volumes of imagery. Pre-trained specialist encoders are powerful but rigid: they impose a fixed architecture and a fixed output structure. A VLM, by contrast, can be queried through natural language, so the same model can address the data in many different ways and return outputs in whatever structure the question demands. The language component directly serves the domain specialist — the astronomer, in our case — which is a meaningful advantage over an encoder that only emits class scores. Among the models we evaluated, Gemini 3.5 Flash was the strongest vision-language model by a clear margin. We tested several models beyond those shown in the figures, spanning both Google and OpenAI systems, and Gemini 3.5 Flash consistently came out ahead. Notably, GPT-5.6-sol — a newer and nominally stronger model — recorded the worst average balanced accuracy of the models we compared. One plausible factor is that Google's models are trained with access to an exceptionally large corpus of images and literature;

however, competing labs have broadly comparable resources, so this alone does not fully explain the gap. Investigating cost led us to a more concrete explanation. GPT-5.6-sol is the most expensive API model per token — $5.00 per million input tokens and $30.00 per million output tokens, roughly three times Gemini 3.5 Flash's $1.50 / $9.00 — yet it was not the most expensive model to run on this task. The reason is that OpenAI and Google tokenize images very differently. Astronomical cutouts are small — ours are at most 424 px on a side — and OpenAI's tokenizer treats an image below its tiling threshold as a single tile, producing very few image tokens. This keeps GPT's cost low, but it also means the model receives comparatively little visual context, which may in turn explain its weak accuracy: with so few image tokens, there is simply little information about the image for the model to work with. Consistent with this, the average cost per 100 images was $0.17 for Gemini versus $0.15 for GPT — nearly identical, despite Gemini's per-token price being far lower and its accuracy far higher — which indicates that Gemini extracts substantially more tokens, and therefore more context, from the same image. This motivated us to test reasoning directly, since reasoning is often held up as a key strength of vision-language models. The result was clear and, if anything, ran counter to the prevailing narrative: additional reasoning did not help. For Gemini, spending more reasoning tokens did not improve accuracy, and in several cases slightly reduced it. We repeated the experiment with GPT to try to force more tokens out of each image — which is essentially what reasoning amounts to — but the cost and token counts were unchanged, because GPT applies a fixed size threshold: below roughly 768 px the entire image is treated as a single tile, so the token count (and cost) is identical regardless of whether the API's image-detail parameter is set to high, medium, or low. We report this deliberately, because there is considerable discussion in the scientific community about the value of "more reasoning"; in our setting, more reasoning did not translate into higher accuracy. Taken together, these observations indicate that the language side of the model is not the bottleneck. A second, independent piece of evidence points the same way: when we fine-tuned Qwen on only the attention projections (the query and value layers), accuracy barely moved; it was only when we fine-tuned all linear layers — including the vision tower — that the fine-tuned model's accuracy jumped. Perception, not language, is the limiting factor for these tasks. That the fully fine-tuned Qwen was able to edge past the specialist CNN in average accuracy — even by a modest margin — is a striking result: it shows that general-purpose vision-language models are genuinely capable of scientific tasks, and that they can perform well even in niche, highly specialised domains such as galaxy morphology. Finally, we are still experimenting with prompt optimization. This work is ongoing, and our intention is to run it with open models — we have already run Qwen (4B) locally — because open models are free to query and therefore permit the very large number of trials that a thorough prompt search requires. This should let us determine more definitively whether prompt optimization is a worthwhile direction for vision tasks. We also note that working in the small-data regime was itself a persistent challenge throughout this project, and designing methods around scarce labels was a recurring theme of the work.

6 Limitations and Future Work

Compute budget. All fine-tuning was single-GPU LoRA in bf16, and the cluster did not support bitsandbytes/quantization — which capped the model sizes we could load and train. As a result we did not evaluate open VLMs larger than 9B. Next step: with multi-GPU nodes or quantized training, extend to larger open VLMs (e.g. Qwen-VL 72B) to test whether the perception gains we observed continue to scale. Prompt optimization is incomplete. The

prompt search ran only on paid Gemini, so it was deliberately cost-capped (few rounds, small validation sets), and the best prompts partly overfit the validation set rather than transferring cleanly to test. Next step: rerun the search with a self-hosted open model (Qwen-4B), which has no per-call cost, allowing far more rounds, more candidate prompts, and larger validation sets — enough to tell whether prompt optimization is a real direction for vision tasks or just search noise. Toward a full astronomy VLM benchmark. This project is a focused study, but its evaluation core (vlm_eval — a provider abstraction, task specifications, a runner, and shared metrics) is already structured like the seed of a general benchmark. Turning it into one is a substantial engineering effort beyond the current scope: extending the provider layer to more models and backends (additional APIs, self-hosted open VLMs, local checkpoints) behind one interface; broadening task coverage across more surveys and problem types (multi-label, vote-fraction regression, detection) with standardized splits; and building an end-to-end pipeline (data staging → inference → scoring → report) so new models and tasks can be added and compared reproducibly. Next step: generalize vlm_eval into that pipeline and grow the task suite into a standard, reproducible benchmark for VLMs on real astronomical imaging. Promising directions. Beyond the above, the most exciting extensions are: (i) active learning — using the VLM's own uncertainty to choose which few galaxies to label next, pushing the low-label regime even further; (ii) treating the VLM's natural-language output as a scientific interface (free-form descriptions, follow-up questions, rationale) rather than just a classifier; and (iii) applying the same low-label recipe to new surveys at first light, where labels are scarcest and a promptable model is most valuable.

7 Conclusion

We benchmarked three families of approaches, specialist encoders (CNNs), zero-shot commercial VLMs (Gemini, GPT, Qwen-397B), and fine-tuned open VLMs (Qwen-VL, LoRA) — on galaxy-morphology tasks, focusing on the low-label regime that defines most new surveys. Our most striking result is that Gemini 3.5 Flash, used entirely zero-shot with no training at all, is already competitive with a fully-trained specialist CNN on several tasks — the strongest of every commercial model we tried, and a genuinely usable option when no labels exist. Building on that, we show that once even 100 labels are available, a small fine-tuned open VLM (Qwen-4B) takes the top spot, reaching the highest average balanced accuracy (72.7%) edging past the specialist CNN (71.1%) and every zero-shot API, while costing nothing to run, with the crossover happening at just 100–200 labels in three of five tasks, and the crossover point higher for the remaining ones. Beyond raw accuracy, a key advantage is that prompting makes these models flexible and exploratory: a researcher can pose open-ended questions, probe edge cases, and reshape a task on the fly through language alone with no retraining, no fixed output head, which is enormously valuable in a discovery-driven, image-rich field like astronomy. The takeaway a reader should remember: for image-rich sciences like astronomy, a vision-language model is now a practical tool at every label budget — Gemini 3.5 Flash off the shelf when labels are zero, and a small, free, fine-tunable open VLM that matches or beats a purpose-built CNN once a handful of labels exist — and because the method assumes nothing specific to galaxies, the same pattern is a strong signal for other image-rich scientific domains, from medical imaging and microscopy to remote sensing

References

[1] Mike Walmsley et al. "Galaxy Zoo Evo: 1 million human-annotated images of galaxies". In: *arXiv preprint arXiv:2512.23691* (2025).

[2] Mike Walmsley et al. "Practical Galaxy Morphology Tools from Deep Supervised Representation Learning". In: *Monthly Notices of the Royal Astronomical Society* 513.2 (2022), pp. 1581–1599. DOI: 10.1093/mnras/stac525.

[3] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *arXiv preprint arXiv:2106.09685* (2021).

[4] Chengrun Yang et al. "Large Language Models as Optimizers". In: *arXiv preprint arXiv:2309.03409* (2023).