Exploring the Universe with Artificial Intelligence (UniversAI)
Proceedings IAU Symposium No. 397, 2026
I. Liodakis & V. Efthymiou, eds.
doi:10.1017/S1743921325105085

Toward Vision-Language Assistants for Radio Astronomical Source Analysis

S. Riggi

INAF - Osservatorio Astrofisico di Catania, Via Santa Sofia 78, 95123 Catania, Italy
email: simone.riggi@inaf.it

Abstract. Vision-language models (VLMs) have recently shown promise in general-purpose reasoning tasks, yet their applicability to domain-specific scientific workflows remains largely unexplored. In this work, we evaluated a series of open-weight and commercial VLMs on six tasks relevant to radio astronomy, such as source morphology classification. We also introduced `radio-llava`, a fine-tuned multimodal assistant built on the LLaVA architecture and adapted for the radio domain through instruction fine-tuning.

In zero-shot mode, commercial models like GPT-4.1 outperform open-weight VLMs on most radio benchmarks. However, `radio-llava` significantly improves upon both base LLaVA and commercial models across nearly all tasks. Despite these gains, specialized vision-only models still deliver substantially better performance across the board. Additionally, we observed that fine-tuning introduces catastrophic forgetting on general multimodal tasks, with performance drops up to 40% that can be partly mitigated with a more diverse training dataset or shallow fine-tuning.

Keywords. Radio astronomy, Radio galaxies, Deep learning, vision-language models

1. Introduction

Vision-language models (VLMs) like `CLIP` (Radford et al. 2021) and `LLaVA` (Liu et al. 2023) have opened new directions in cross-modal learning by enabling natural language understanding grounded in visual inputs. Astronomy, with its rich collection of images, spectra, and tabular data, is an ideal domain for testing such capabilities. While `CLIP`-based works (Gupta et al 2025) primarily focused on static alignment for image-text retrieval tasks, this paper explores the application of generative, instruction-following VLMs to radio astronomical analysis.

2. Methodology

We used labelled radio image data from various surveys – including MeerKAT SMGPS (Goedhart et al. 2024), ASKAP EMU (Hopkins et al. 2025; Norris et al. 2021), and VLA FIRST (Becker et al. 1995) – to construct six radio classification tasks for VLM evaluation, summarised in Table 1.

We first assessed the zero-shot performance of several open-weight models – `LLaVA-OneVision` (Li et al. 2024), `Qwen2VL` (Wang et al. 2024), `InternVL` (Chen et al. 2024), `TinyLLaVA` (Zhou et al. 2024) – of different sizes (0.5B to 72B parameters) and a representative commercial closed-weight solution (OpenAI GPT-4.1).

We then evaluated a fine-tuned `LLaVA-OneVision` 7B model – hereafter referred to as *radio-llava* – trained on curated Q&A instructions derived from an independent set of 59k annotated radio images, as well as on curated 39k image-caption pairs extracted from a corpus of published radio astronomy papers (2000–2025). In all training runs we

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Astronomical Union. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1743921325105085 Published online by Cambridge University Press

Table 1. Radio benchmarks.

ID	Task Description	Survey	#Images	Classification	Labels
B1	extended/diffuse source detection	SMGPS, EMU	5,718	multi-label	EXTENDED, DIFFUSE DIFFUSE-LARGE
B2	source morphology classification	FIRST	3,835	multi-class single-label	1C-1P, 1C-2P, 1C-3P, 2C-2P, 2C-3P, 3C-3P
B3	extended radio-galaxy detection	SMGPS, EMU	5,718	binary class	YES, NO
B4	imaging artefact detection	SMGPS, EMU	5,718	binary class	YES, NO
B5	image peculiarity classification	SMGPS, EMU	5,718	multi-class	ORDINARY, COMPLEX, PECULIAR
B6	radio-galaxy morphology classification	FIRST	833	binary class	FR-I, FR-II

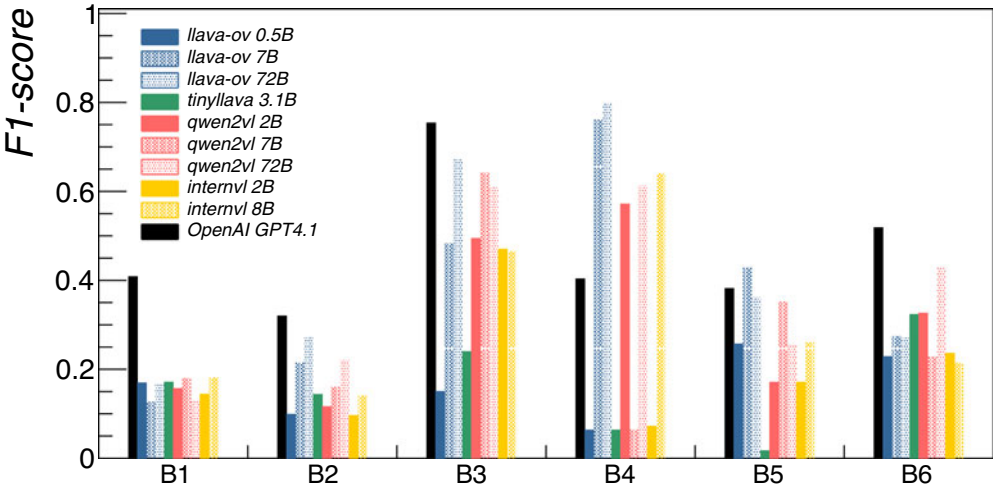


Figure 1. Classification "macro-averaged" F1-score obtained across tasks B1–B6 in zero-shot mode with open-weight VLMs of different sizes (0.5B, 2B, 3.1B, 7B, 8B, 72B), shown with coloured histograms (LLaVA: blue, TinyLLaVA: green, Qwen2VL: red, InternVL: orange), and OpenAI GPT-4.1 (black histograms).

kept the vision encoder (*siglip-so400m-patch14-384*) frozen, while we fully fine-tuned the LLM (*qwen2*) and adapter (*mlp2x_gelu*) components. Fine-tuning results were compared with both the base `LLaVA` model and a vision-only baseline.

Finally, we conducted additional diagnostic analyses to explore alternative `LLaVA` hyperparameter settings and to evaluate the extent of catastrophic forgetting of previously learned multimodal tasks following fine-tuning.

Further details on the training/evaluation data and strategy can be found in Riggi et al. (2025).

3. Results

Zero-shot benchmark results are presented in Fig. 1. GPT-4.1 (black histograms) outperforms the open-weight models in four out of six benchmarks, with a ∼20% improvement in tasks B1 and B6. However, it unexpectedly underperforms in tasks B4 and B5. Among the open-weight models, the largest versions of `LLaVA` match or slightly exceed the performance of their counterparts. Overall, all models demonstrate relatively low performance, indicating the need for domain-specific specialization on radio astronomical data.

https://doi.org/10.1017/S1743921325105085 Published online by Cambridge University Press

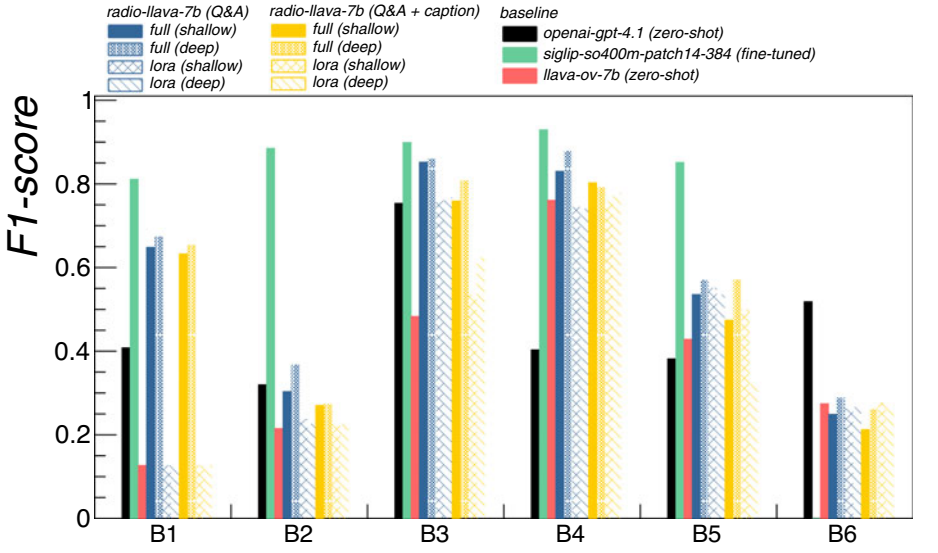


Figure 2. Classification "macro-averaged" F1-score obtained across tasks B1–B6 with the *radio-llava* model, fine-tuned on the Q&A training dataset (blue histograms) and the combined Q&A and caption datasets (orange histograms) with different training strategies (full vs. LoRA fine-tuning) and depths (shallow vs. deep). Results from different baseline models are also shown: base `llava-ov-7b` (red histograms), fine-tuned `siglip-so400m-patch14-384` vision encoder (green histograms), OpenAI GPT-4.1 (black histograms).

Results for the fine-tuned `radio-llava` models, reported in Fig. 2 as blue and orange histograms for Q&A-only and combined training datasets respectively, indicate substantial improvements over the base VLMs (red histograms), with gains of ∼30% in tasks B1 and B3. For the remaining tasks, the improvements are more modest − around 10% in B2, B4, and B5 − and negligible in B6. With the exception of B6, the fine-tuned models also outperform GPT-4.1 (black histograms), though with smaller margins. Notably, VLMs trained jointly on all tasks consistently underperform relative to vision-only classifiers trained individually per task, as shown by the green histograms in Fig. 2.

Models fine-tuned using alternative LLaVA hyperparameters − such as learning rate and relative scheduling, batch size, and LoRA configuration − exhibited performance broadly comparable to that obtained with the default settings. A modest improvement (∼2%) was observed when the vision encoder was also fully fine-tuned.

Deeper fine-tuning (e.g., three epochs vs a single epoch) had only a minor impact (few percent) on classification performance for most radio benchmarks. However, it led to substantial degradation in performance on standard multimodal tasks − such as *ChartQA*, *DocVQA*, and *InfoVQA* − with accuracy drops of up to 40%, indicative of catastrophic forgetting, as shown in Fig. 3. Incorporating more diverse training data, such as curated caption datasets, or adopting LoRA-based fine-tuning strategies partially mitigated this effect and helped recover some of the lost general multimodal capabilities.

4. Summary

This work investigates the feasibility of using small-scale Vision-Language Models (VLMs) as domain-specific assistants for radio astronomical image analysis. Our evaluation shows that large commercial models like GPT-4.1 deliver strong zero-shot results on some tasks but consistently underperform compared to specialized vision-only models trained specifically for radio astronomy.

https://doi.org/10.1017/S1743921325105085 Published online by Cambridge University Press

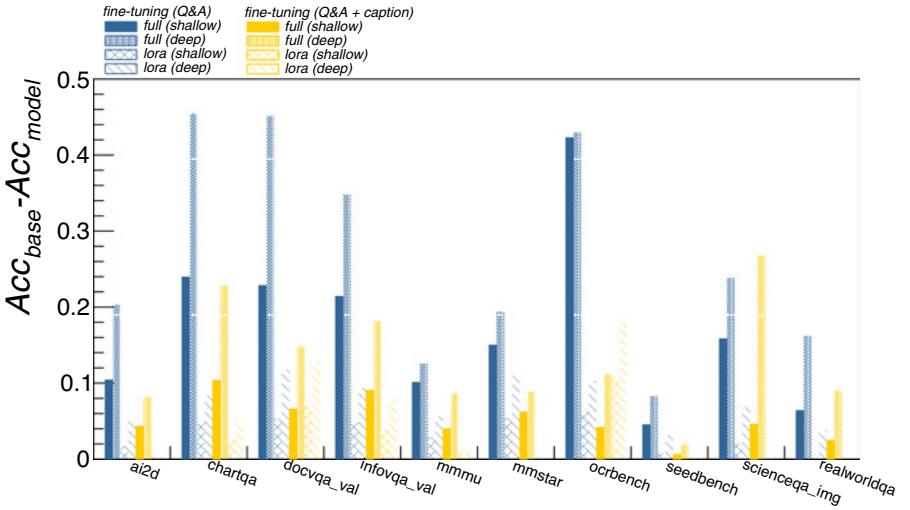


Figure 3. Classification accuracy differences between the `llava-ov-7b` base model and the fine-tuned `radio-llava` models (blue and orange histograms), evaluated on various multimodal (non-radio) benchmarks.

Fine-tuning smaller open-weight models like `LLaVA` on curated radio datasets (Q&A and captions) significantly improves performance over both base and commercial models, with up to ∼30% F1-score gains in certain benchmarks. However, these fine-tuned models still lag behind vision-only models trained per task and suffer from catastrophic forgetting, losing up to 40% accuracy on general multimodal benchmarks after full deeper fine-tuning. Incorporating caption data and adopting LoRA-based fine-tuning helps mitigate some of these issues, improving instruction-following ability and partially recovering lost generalization.

These results underscore both the potential and current limitations of VLMs in supporting astronomers through assistant-style interaction, and emphasize the need for better multimodal alignment, more diverse and high-quality training data, and improved fine-tuning strategies to fully realize their value in scientific workflows.

References

Becker, R. H., et al., 1995, ApJ, 450, 559, https://doi.org/10.1086/176166

Chen, Z., et al., Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 24185–24198, 2024, https://doi.org/10.48550/arXiv.2312.14238

Goedhart, S., et al., 2024, MNRAS, 531, 649, https://doi.org/10.1093/mnras/stae1166

Gupta, N., et al., 2025, PASA, in press, https://doi.org/10.1017/pasa.2025.10064

Hopkins, A. M., et al., 2025, PASA, 42, e071, https://doi.org/10.1017/pasa.2025.10042

Li, B., et al., *LLaVA-OneVision: Easy Visual Task Transfer*, 2024, https://doi.org/10.48550/arXiv.2408.03326

Liu, H., et al., *Visual instruction tuning*, 2023, https://doi.org/10.48550/arXiv.2304.08485

Norris, R. P., et al., 2021, PASA, 38, e046, https://doi.org/10.1017/pasa.2021.42

Radford, A., et al., In International Conference on Machine Learning, 2021, PMLR, 8748, https://doi.org/10.48550/arXiv.2103.00020

S. Riggi et al., 2025, PASA, in press, https://doi.org/10.48550/arXiv.2503.23859

Wang, P., et al., *Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution*, 2024, https://doi.org/10.48550/arXiv.2409.12191

Zhou, B., et al., *TinyLLaVA: A Framework of Small-scale Large Multimodal Models*, https://doi.org/10.48550/arXiv.2402.14289

https://doi.org/10.1017/S1743921325105085 Published online by Cambridge University Press