

# radio-llava: Advancing Vision-Language Models for Radio Astronomical Source Analysis

S. Riggi,<sup>1</sup> T. Cecconello,<sup>1,2</sup> A. Pilzer,<sup>3</sup> S. Palazzo,<sup>2</sup> N. Gupta,<sup>4</sup> A.M. Hopkins,<sup>5</sup> C. Trigilio,<sup>1</sup> and G. Umana<sup>1</sup>

<sup>1</sup>INAF - Osservatorio Astrofisico di Catania, Via Santa Sofia 78, 95123 Catania, Italy

<sup>2</sup>Department of Electrical, Electronic and Computer Engineering, University of Catania, Santa Sofia 64, 95123, Catania, Italy

<sup>3</sup>NVIDIA AI Technology Center, Italy

<sup>4</sup>CSIRO Space & Astronomy, PO Box 1130, Bentley WA 6102, Australia

<sup>5</sup>School of Mathematical and Physical Sciences, 12 Wally's Walk, Macquarie University, NSW 2109, Australia

Author for correspondence: S. Riggi, Email: simone.riggi@inaf.it.

---

## Abstract

The advent of next-generation radio telescopes is set to transform radio astronomy by producing massive data volumes that challenge traditional processing methods. Deep learning techniques have shown strong potential in automating radio analysis tasks, yet are often constrained by the limited availability of large annotated datasets. Recent progress in self-supervised learning has led to foundational radio vision models, but adapting them for new tasks typically requires coding expertise, limiting their accessibility to a broader astronomical community. Text-based AI interfaces offer a promising alternative by enabling task-specific queries and example-driven learning. In this context, Large Language Models (LLMs), with their remarkable zero-shot capabilities, are increasingly used in scientific domains. However, deploying large-scale models remains resource-intensive, and there is a growing demand for AI systems that can reason over both visual and textual data in astronomical analysis.

This study explores small-scale Vision-Language Models (VLMs) as AI assistants for radio astronomy, combining LLM capabilities with vision transformers. We fine-tuned the LLaVA VLM on a dataset of 59k radio images from multiple surveys, enriched with 38k image-caption pairs from the literature. The fine-tuned models show clear improvements over base models in radio-specific tasks, achieving  $\sim 30\%$  F1-score gains in extended source detection, but they underperform vision-only classifiers and exhibit  $\sim 20\%$  drop on general multimodal tasks. Inclusion of caption data and LoRA fine-tuning enhances instruction-following and helps recover  $\sim 10\%$  accuracy on multimodal benchmarks (e.g., ChartQA/DocVQA).

This work lays the foundation for future advancements in radio VLMs, highlighting their potential and limitations, such as the need for better multimodal alignment, higher-quality datasets, and mitigation of catastrophic forgetting.

---

## 1. Introduction

The upcoming Square Kilometer Array (SKA) (Dewdney et al., 2016) and its precursor telescopes (e.g., MeerKAT, ASKAP, LOFAR) are revolutionizing radio astronomy, enabling to probe the radio sky at unprecedented sensitivities and angular resolutions. SKA, once operational, is expected to produce exabytes of data annually. The immense volume and complexity of the generated data will challenge traditional data-processing methods, necessitating advanced computational and AI techniques to automate repetitive, resource-intensive tasks.

In this context, deep-learning methodologies have already shown promising results in various analysis tasks including: source detection in 2D radio maps (Mostert et al., 2022; Zhang et al., 2022; Yu et al., 2022; Riggi et al., 2023; Lao et al., 2023; Cornu et al., 2024; Stuardi et al., 2024), source and host galaxy detection from 2D radio+IR maps (Wu et al., 2019; Gupta et al., 2023) or radio+optical maps (Lou et al., 2023), source detection in HI cubes (Liang et al., 2023; Håkansson et al., 2023; Barkai et al., 2023), source classification (Aniyan & Thorat, 2017; Tang et al., 2019; Ma et al., 2019; Maslej-Krešňáková et al., 2021; Tang et al., 2022; Nair et al., 2022; Riggi et al., 2024a), search for objects with peculiar morphology (Ralph et al., 2019; Galvin et al., 2020; Mostert et al., 2021; Gupta et al.,

2022; Mesarcik et al., 2023; Lochner et al., 2023; Riggi et al., 2024b), fast radio burst detection (Connor & van Leeuwen, 2018; Agarwal, 2020), radio imaging (Schmidt et al., 2022; Geyer et al., 2023; Chiche et al., 2023), synthetic data generation (Rustige et al., 2023; Sortino et al., 2024; Martínez et al., 2024), and many others. The full potential of developed models, especially those using supervised learning techniques, is often hampered by the scarcity of large and balanced annotated radio datasets. Additionally, existing radio models typically employ data labelling schemes that vary widely across different analysis cases, hindering the integration of individual datasets into larger collections and restricting model usability beyond their initial applications.

Recent studies (Sljepcevic et al., 2024; Riggi et al., 2024b; Lastufka et al., 2024) have sought to address the challenges posed by limited annotated training datasets by applying self-supervised learning (SSL) techniques, which utilize the extensive collections of unlabelled radio images available in current and past surveys. Several foundational radio models have been developed to effectively enable feature extraction from radio maps for a variety of tasks, such as data inspection, source extraction and classification, anomaly detection, and image retrieval. Pretrained SSL models have also been fine-tuned on

smaller annotated datasets and specialized for these applications. Ongoing research is focusing on various areas: comparing alternative SSL methods on radio data (Cecconello et al., 2024), assessing the performance of SSL models pretrained on non-radio data (such as natural or optical images) for radio-specific tasks and vice versa (Riggi et al., 2024c; Lastufka et al., 2024b), exploring optimized dataset curation strategies, scaling up model training to larger architectures and millions of radio images, and defining more constraining downstream datasets and tasks.

While existing SSL models can be adapted or expanded for new use cases, their accessibility is often limited by the need for astronomers to write code for adaptation to similar or entirely new tasks. This requirement could hinder widespread adoption, as many astronomers may prefer more intuitive, user-friendly interfaces. An AI assistant with a more accessible, text-based interface would allow researchers to interact with the model by providing examples, querying specific tasks, and customizing output formats to suit their needs.

Large Language Models (LLMs) like GPT-4 (OpenAI, 2023), Claude3 (Anthropic, 2024), and open-source alternatives such as LLaMA (Touvron et al., 2023) or InternLM (Cai et al., 2024) have proven effective as AI assistants, showing remarkable zero-shot learning capabilities when prompted unseen data or tasks across a wide array of fields, including astronomy (Tanoglidis & Jain, 2024). Specializing and deploying very large open-source models is, however, currently prohibitive in terms of the required computing resources (high GPU requirements, memory demands, and power consumption). Furthermore, in addition to textual interaction, there is an increasing demand for models capable of processing visual data, facilitating multimodal reasoning for tasks like analysing complex astronomical images. Some initiatives, such as AstroLLaMA (Nguyen et al., 2023; Perkowski et al., 2024), have started to address this by developing astronomy-specific mid-size LLMs, though these efforts are still limited to text-based inputs. Commercial solutions add additional cost concerns, particularly in inference and fine-tuning expenses, as demonstrated by Sun et al. (2024) in the context of interpreting multi-band galaxy observations. Furthermore, while large models are well-suited for tasks requiring extensive general knowledge, this scope may be more than what is needed in astronomy, where specialized knowledge is essential. Given these considerations, current research has increasingly turned to adapting smaller LLMs (i.e., those with fewer than 10 billion parameters) for specific domains, as well as investigating multimodal models capable of processing combined data inputs, such as text, images, and videos.

In this context, we aim to explore recent, state-of-the-art, small-scale vision-language models (VLMs) to develop AI assistants tailored to radio astronomy. These models combine both visual and textual comprehension by integrating LLM capabilities with vision transformers (Dosovitskiy et al., 2021) for image processing. Typically, VLMs comprise two main components: a vision encoder transformer that extracts features from input images, and an LLM that generates textual responses from combined visual and textual input representations.

VLMs offer promising solutions to the outlined challenges in two ways. First, they can manage tasks through a text-based interface, allowing astronomers to specify task details and expected response formats in natural language. This flexibility enables the use of contextual image-text examples, making VLMs more adaptable than traditional vision models that are limited to predefined label outputs for specific tasks. Second, VLMs potentially support the integration of specialized vision encoder models that have been trained on unlabelled radio data through self-supervised learning (SSL) methods. This approach bridges the gap with ongoing SSL research, facilitating the reuse of existing radio SSL models and enabling the full utilization of extensive unlabelled image datasets from recent and past radio surveys. These smaller models offer a promising alternative, with lower computational costs and more manageable deployment requirements. Moreover, their specialized capabilities could be more than sufficient for the specific needs of radio astronomy, without the added complexity and resource demands of larger models. However, the suitability of these smaller models for astronomical tasks remains largely unexplored, particularly in radio source analysis.

This paper seeks to evaluate the current state of small multimodal language models as AI assistants for radio astronomy. By investigating their strengths, limitations, and applicability to radio astronomical source analysis tasks, we hope to familiarize the community with this emerging technology and its potential, as well as to highlight the challenges that need to be addressed in future developments. A tailored VLM for radio astronomy could eventually assist astronomers in efficiently analysing radio images without requiring extensive technical expertise in AI models. By leveraging a text-based interface, astronomers can perform complex and diverse image analysis tasks, even guiding the model using image-based examples. Key applications include the automated identification and retrieval of specific classes of radio sources in survey image data, as well as data quality assessment – enhancing the efficiency and accessibility of radio survey analysis. Additionally, a VLM-based assistant could be deployed to support less experienced users (e.g., students, citizens) in ongoing crowdsourcing projects like the EMU Radio Galaxy Zoo<sup>1</sup>. By providing real-time guidance and explanations, the model could help users classify radio sources, identify peculiar objects, and improve the reliability of crowdsourced annotations.

Multi-modal models have only very recently begun to gain traction in astronomy – with most developments emerging within the past year – as a means to bridge heterogeneous data modalities such as images, spectra, and natural language. These efforts have primarily focused on adapting the CLIP (Contrastive Language-Image Pretraining) model (Radford et al., 2021) and its derivatives to astronomical tasks, leveraging their ability to align visual and textual representations in a shared latent space. For instance, Gupta et al (2025) introduced EMUSE, a tool built on a fine-tuned OpenCLIP<sup>2</sup> model that

<sup>1</sup><https://www.zooniverse.org/projects/hongming-tang/radio-galaxy-zoo-emu>

<sup>2</sup>[https://github.com/mlfoundations/open\\_clip](https://github.com/mlfoundations/open_clip)

enables users to search EMU survey data using either textual queries or template image similarity. A similar application is PAPERCLIP (Mishra-Sharma et al., 2024), a CLIP-based model fine-tuned on Hubble Space Telescope (HST) proposal data – including image observations and proposal abstracts – which enables cross-modal retrieval based on image content or textual queries such as object names or scientific use cases. AstroCLIP (Parker et al., 2024) presents a powerful CLIP-style foundation model for galaxies in the optical domain, where the textual modality in the original CLIP framework is replaced by galaxy spectra. The model aligns image and spectrum modalities in a shared embedding space after self-supervised pre-training of each modality independently, and demonstrates impressive performance on tasks such as morphology classification, redshift estimation, and property inference.

Unlike these studies, our approach focuses on instruction-following vision-language models like LLaVA (*Large Language and Vision Assistant*) (Liu et al., 2023) to enable open-ended captioning, visual question answering, and multi-turn scientific dialogue grounded in domain-specific radio astronomy data. These tasks go beyond the static alignment capabilities of CLIP models, which lack generative, reasoning, and conversational abilities. Our model is thus particularly suited for exploratory analysis, educational interfaces, and assistant-style tools that can explain, summarise, or discuss diagnostic plots and observational data. Meanwhile, CLIP-based models remain better suited for scalable retrieval, zero-shot classification, and semantic similarity search over large datasets.

The paper is organized as follows. Section 2 provides an overview of vision-language models, with a focus on the architecture of a prominent model, LLaVA, which we aim to adapt for radio-astronomical data. In Section 3, we describe our adapted model, termed *radio-llava*, including the training datasets and methodology. Section 4 presents the evaluation of the specialized model across several radio-astronomy tasks. Finally, Section 5 summarises the results and discusses directions for future research.

## 2. Vision-Language Models

Multi-modal large language models (MLLMs) are designed to process and integrate data from multiple modalities, such as text, audio, images, and video. Vision-language models (VLMs) are a specific type of multi-modal system that focuses on combining visual and textual information. These models leverage large language models and vision transformers to align visual and textual representations, enabling them to perform complex tasks like image captioning, visual question answering (VQA), and object recognition in a descriptive context. A comprehensive review of MLLMs and VLMs is provided by Li et al. (2023b); Bordes et al. (2024); Yin et al. (2024). In this section, we focus on describing the current state, architecture, and training strategy of the VLM model *LLaVA*, which we have adapted for use with radio astronomical data in this work.

### 2.1 The LLaVA model

#### 2.1.1 Model overview

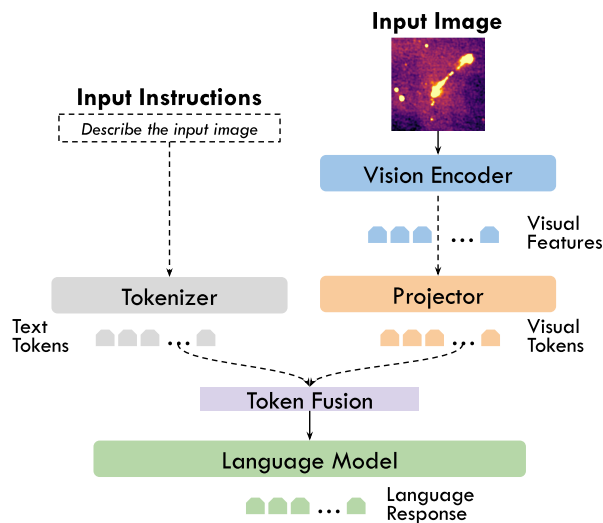
LLaVA (*Large Language and Vision Assistant*) (Liu et al., 2023) is a state-of-the-art multimodal model that integrates both visual and textual understanding, combining the capabilities of large language models (LLMs) with vision processing abilities. Its primary function is to interpret and generate responses to input that includes both images and text, making it ideal for tasks like visual question answering (VQA), image captioning, and other vision-language tasks. Since the first release, the model demonstrated exceptional multimodal conversational skills, often displaying behaviour comparable to GPT-4V when tasked with interpreting novel images and following new instructions for the first time.

Following releases (LLaVA 1.5, Liu et al. 2024a) greatly enhanced model capabilities by integrating a larger set of academic-focused instructional data, achieving state-of-the-art results on numerous benchmarks while utilizing a highly data-efficient strategy. Recent advancements in the LLaVA series, including models like LLaVA-NeXT (Liu et al., 2024b) and LLaVA-OneVision (Li et al., 2024), have significantly broadened the scope of input modalities they can handle, supporting both single or multiple images as well as video content. These improvements were driven by three key innovations: the *AnyRes* technique for processing high-resolution images, the expansion of high-quality instruction datasets, and the integration of the most advanced open-source LLMs available at the time, further enhancing model capabilities across diverse tasks. Various variants or specialization of the first LLaVA models have been produced so far. For example, TinyLLaVA (Zhou et al., 2024; Jia et al., 2024) is a compact refactored variant of the original LLaVA 1.5 model, designed to enable easier inclusion of alternative light vision and LLM models, thus significantly reducing overall model size and resource requirements. LLaVA-Med (Li et al., 2023a) is a specialized variant of the LLaVA model designed to assist in medical image analysis and diagnostics by fine-tuning its multimodal capabilities on medical datasets such as X-rays, MRIs, and other healthcare-related visual data.

#### 2.1.2 Model architecture

The LLaVA model, schematically represented in Figure 1, consists of these components:

- *Vision Encoder*: Processes image data using a pre-trained Vision Transformer (ViT) model with multiple transformer layers, such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023). It extracts visual features from input images;
- *Language Model*: Handles text processing, typically an LLM such as Qwen (Bai et al., 2023), capable of understanding, generating, and reasoning with natural language;
- *Projector*: Since the vision encoder and language model operate in different feature spaces, the projector transforms visual embeddings into a format compatible with the language model input space. This enables effective integration of visual and textual information. LLaVA employs a two-layer Multi-Layer Perceptron (MLP) trained to align the



**Figure 1.** A schematic representation of the LLaVA model architecture.

modalities, ensuring that the visual embeddings can be seamlessly used by the language model;

- **Multimodal Fusion Layer:** It aligns and fuses the visual features from the vision encoder with the text embeddings from the language model, enabling the model to process both modalities jointly through self-attention mechanisms.

The model processes multimodal image-text inputs as follows:

1. The input text (instruction or query) is tokenized into numerical tokens using a predefined vocabulary.
2. The input image is divided into patches and passed through the vision encoder, which extracts key features such as objects, colours, textures, and spatial relationships. The projector then converts these visual representations into language-compatible embeddings;
3. The multimodal fusion layer integrates the visual embeddings into the input sequence of the language model, allowing it to process both visual and textual data jointly. The model then generates an output response based on the given task, such as answering questions about the image (VQA) or generating a descriptive caption (image captioning).

### 2.1.3 Model training

The LLaVA model series is trained from pre-existing language and vision encoders through instruction fine-tuning on large-scale datasets. These datasets consist of image (or video) and text pairs, including captions, descriptions, and questions, enabling the model to learn associations between visual elements and natural language. The training process typically involves multiple stages, each potentially using different datasets, including pretraining on unimodal visual and textual data, aligning vision and language features, and fine-tuning with instructional data to address diverse visual tasks. Further details on training datasets and methodology can be found in the original model publications.

During instruction tuning, the model is optimized by minimizing a cross-entropy loss, which quantifies how closely the predicted text output matches the ground truth. The model generates text auto-regressively, predicting one token at a time based on previously generated tokens. At each step, it outputs a probability distribution over possible next tokens and is trained to minimize the difference between its prediction and the actual token. The cross-entropy loss is computed for each token while conditioning on prior tokens, accumulating over the entire sequence and penalizing incorrect predictions at each step. This iterative process ensures the model learns to generate coherent text in a structured manner.

## 3. The radio-llava model

The *radio-llava* model is a small multi-modal model fine-tuned from a base LLaVA model using radio astronomical image-text data. This section describes the training dataset, and the model fine-tuning procedure.

### 3.1 Training dataset

The training dataset consists of multiple conversations between a virtual assistant and a user regarding a given radio image. The dataset follows the standard JSON format required by multi-modal models:

```
[
  {
    "id": "image id",
    "image": "image path",
    "conversations": [
      {
        "from": "human",
        "value": "<image>\n Provide a brief description of the\n given image."
      },
      {
        "from": "gpt",
        "value": "The image is a radio astronomical cutout ..."
      },
      ...
    ]
  }
]
```

We constructed two training datasets. The first, referred to as the *Q&A dataset*, consists of a series of question-answer interactions related to the content of radio images, all extracted from radio continuum surveys. The second dataset, termed the *caption dataset*, contains a single user-assistant exchange per image, in which the assistant provides a description of the image content. In this case, images and their corresponding captions were sourced from a collection of scientific papers on radio astronomical topics available in the arXiv database. Details on both datasets are provided in the following sections.

#### 3.1.1 Q&A dataset

This dataset was assembled from multiple annotated radio datasets, each designed for different radio source classification or detection tasks:

- **Fine-Grained Datasets:** These datasets, typically used for training radio object detection and segmentation models like YOLO (Redmon et al., 2016) or Mask R-CNN (He et al., 2017), contain wide-field images (a few arcminutes



in size) with region- or pixel-level annotations, including object positions (centres, bounding boxes, segmentation masks), classification labels, and confidence scores.

- **Coarse-Grained Datasets:** Commonly used for radio source classification models, these datasets contain either zoomed-in source images or wide-field images with one or more assigned classification labels.

Details regarding the number of images and available classes for each dataset are provided in Appendix A.

It is important to note that classification schemes vary across datasets. Some provide astrophysical source-type labels (e.g., HII, SNR, GALAXY), while others focus on morphological classifications (e.g., FR-I vs. FR-II, COMPACT vs. EXTENDED vs. DIFFUSE). Before generating the conversational Q&A dataset, we aimed to standardize terminology whenever possible. In some cases (e.g., see Section A.2), we performed cross-matching and relabelling to augment the original datasets by adding additional labels to certain images or objects. However, variations in labelling schemes remain unavoidable due to the lack of annotation standards in the radio astronomy community. In this respect, our goal is to fine-tune an LLM model that is exposed to diverse classification schemes, making it potentially adaptable to different user domains.

The Q&A dataset was constructed from annotated datasets through the following steps:

1. **Automated Template-Based Descriptions:** For each image, we programmatically generated a template description based on the available annotations<sup>3</sup>.

Example: An image from the *radioimg-dataset* (see Section A.1) with the assigned labels COMPACT, EXTENDED, RADIO-GALAXY, ARTIFACT would be described as:

*The image is a radio astronomical image cutout extracted from a larger radio-continuum Stokes-I map produced by an interferometer telescope. The image contains various point-like or compact radio sources superimposed over the sky background noise. It also contains one or more extended radio sources. Some of them are likely extended radio galaxies. Some radio sources present in the image are poorly imaged and surrounded by imaging artefacts having a ring pattern.*

Fine-grained datasets include richer descriptions, specifying source positions and sizes.

2. **Automated Q&A Generation Using a Pretrained VLM:** We generated multiple Q&A interactions per image using a InternVL VLM model<sup>4</sup> (Chen et al., 2024). The model was fed with the image, template caption, and structured prompts to ensure that the generated conversations remained faithful to the original image and annotation content, and included at least the following questions:

- *Can you describe the image content?*
- *Can you provide the bounding box coordinates of all radio sources with class X (e.g., compact, extended, etc.) present in the image?*

<sup>3</sup>These descriptions are statically defined, meaning two images with identical annotations will have the same description.

<sup>4</sup>We used InternVL2\_5-8B-MPO model version, available here: [https://huggingface.co/OpenGVLab/InternVL2\\_5-8B-MPO](https://huggingface.co/OpenGVLab/InternVL2_5-8B-MPO), <https://github.com/OpenGVLab/InternVL>

- *Do you see any likely radio galaxy with an extended morphology in the image?*
- *Which of these morphological classes of radio sources do you see in the image?*
- *Do you see any imaging artefact around bright sources in the presented image?*
- *Is there any blank pixel region at the edges of the image?*
- *Is the image content ordinary or peculiar in terms of the objects it contains?*

To prevent excessive generalization, we constrained the VLM's output by using an intermediate temperature setting (0.5).

Overall, the final training dataset comprises 59,331 images and 1,590,202 user-assistant conversations.

Despite these efforts, the current annotated radio data – which primarily provide classification labels or bounding boxes – still limit the diversity and richness of generated image-based conversations. This constraint directly impacts model performance and its instruction-following capabilities, as discussed in Section 4.

### 3.1.2 Caption dataset

This dataset was compiled by extracting figures and their corresponding captions from a broad collection of arXiv scientific papers containing radio astronomy-related keywords, published between 2000 and 2025. To classify the image format and assess caption quality, we processed the extracted raw image-caption data using the same InternVL VLM model employed for generating the Q&A dataset. Specifically, we computed the following parameters for each image-caption pair:

- **n\_words:** number of words in the caption;
- **has\_multiplot:** binary flag set to true if the image contains multiple plots/frames, either as insets, side by side, stacked, or arranged in a grid layout;
- **is\_astromap:** binary flag set to true if the image and caption depict an astrophysical map with one or more sources superimposed on the sky background;
- **is\_corrupted:** binary flag set to true if the caption contains incomplete sentences or corrupted text;
- **caption\_score:** integer score assessing caption quality on a scale from 0 (low) to 10 (high), based on coherence, informativeness, completeness, clarity, and correctness of English style.

Only highly rated single-plot images were included in the training sample, applying the following selection criteria: **n\_words**>5, **has\_multiplot**=0, **is\_corrupted**=0, and **caption\_score**>7. This resulted in a final training sample of 38,545 images. The **has\_multiplot** criterion had a significant impact, removing approximately 62% of the images from the initial dataset. We opted not to apply the **is\_astromap** filter, as doing so would have further reduced the dataset size to approximately 8,700 images.

### 3.2 Model fine-tuning

The *radio-llava* model was trained using instruction fine-tuning on the Q&A radio dataset alone, as well as on the combined Q&A and caption datasets, starting from the pre-trained LLaVA-OneVision 7B model<sup>5</sup>. Keeping the vision encoder (siglip-so400m-patch14-384) frozen, we fine-tuned both the LLM (qwen2) and adapter (mlp2x\_gelu) components using either full fine-tuning or the Low-Rank Adaptation (LoRA) (Hu et al, 2021) fine-tuning strategy<sup>6</sup>. We set the LoRA rank and alpha scaling parameters to 64 and 128, respectively. The model was trained for either 1 epoch (*shallow* fine-tuning) or 3 epochs (*deep* fine-tuning). In all training runs, we used default hyperparameters, with a batch size of 1, a gradient accumulation step of 2, and a learning rate of  $10^{-5}$ .

On single-GPU servers with medium GPU memory (e.g., NVIDIA A30 24 GB or RTX6000 48 GB), we were only able to train the model using LoRA fine-tuning on the Q&A dataset, while full fine-tuning required more extensive computational resources. Consequently, all fine-tuning runs were conducted on a single node of the CINECA LEONARDO Booster infrastructure<sup>7</sup>, utilizing 4 GPUs (NVIDIA A100 64 GB) and 8 CPUs (Intel Xeon Platinum 8358, 2.60GHz) with 32 GB of memory allocated per CPU. Full fine-tuning required approximately 15 hours per epoch on the combined training dataset.

### 3.3 Implementation details

For training the LLaVA model, we utilized software and resources available in the LLaVA-NeXT repository<sup>8</sup>. Only minor modifications were applied to the original software to load our dataset and enable the loading of LoRA fine-tuned models using the Qwen LLM.

Inference scripts developed for LLaVA and other VLM models tested in this study are provided in the following repository: <https://github.com/SKA-INAF/radio-llava>. This repository also includes a Streamlit<sup>9</sup> application (see Figure C.1 in the Appendix), allowing users to load a LLaVA model, upload an image, and interact with the assistant via a web interface. Table C1 in the Appendix provides an example of user-assistant conversations for two sample radio images, comparing responses from the base and fine-tuned LLaVA-OneVision 7B models.

The fine-tuned models are publicly available in the *Hugging Face* repository: <https://huggingface.co/inaf-oact-ai>.

## 4. Model evaluation

Using independently annotated datasets, we defined six evaluation benchmarks (B1–B6) to assess the model's reasoning capabilities on radio image data. The benchmark datasets and

inference prompts are detailed in Section 4.1.1. Additionally, we evaluated our models on various standard non-astronomical benchmarks, listed in Section 4.1.2, to quantify the impact of fine-tuning on tasks previously learned by the base model.

Section 4.2 presents the zero-shot performance of the LLaVA base model, comparing it with alternative VLMs. The evaluation results for the fine-tuned *radio-llava* models are reported in Section 4.3.

We will consistently use these widely adopted metrics in classification problems:

- Recall ( $\mathcal{R}$ ): The fraction of sources (images) from a given class that are correctly identified by the model, out of all sources (images) that truly belong to that class:

$$\mathcal{R} = \frac{TP}{TP + FN}$$

- Precision ( $\mathcal{P}$ ): The fraction of sources (images) correctly predicted to belong to a given class, out of all sources (images) the model assigned to that class:

$$\mathcal{P} = \frac{TP}{TP + FP}$$

- F1-score: The harmonic mean of precision and recall, offering a balanced measure of a model's performance:

$$\text{F1-score} = 2 \times \frac{\mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} \quad (1)$$

- Accuracy ( $\mathcal{A}$ ): The overall fraction of correctly classified sources (images), regardless of class, over the total number of instances:

$$\mathcal{A} = \frac{TP + TN}{TP + TN + FP + FN}$$

Here,  $TP$ ,  $FP$ ,  $TN$ , and  $FN$  represent the number of true positives, false positives, true negatives, and false negatives, respectively.

## 4.1 Evaluation benchmarks

### 4.1.1 Radio benchmarks

**B1 - Extended/Diffuse Source Detection** We used the *radioimg-multilabel* test dataset (5,718 images) to evaluate the models' ability to detect extended or diffuse radio sources in input images.

For this task, we applied the following prompt:

```
### Context: Consider these morphological classes of radio astronomical sources, defined as follows:
EXTENDED: This class comprises either single-island compact objects with sharp edges, having a morphology and size dissimilar to that of the image synthesised beam (e.g. 10 times larger than the beam size or with elongated shape), or disjoint multi-island objects, where each island can have either a compact or extended morphology and can host single or multiple emission components. Typical examples are extended radio galaxies formed by a single elongated island or by multiple islands, hosting the galaxy core and lobe structures
DIFFUSE: A particular class of single-island extended objects with small angular size (e.g. smaller than few arcminutes), having diffuse edges and a roundish morphology;
```

<sup>5</sup>[lmsys-lab/llava-ov](https://github.com/lmsys-lab/llava-ov)

<sup>6</sup>LoRA is a lightweight training method that updates only small, low-rank matrices within the model instead of fine-tuning the entire model. This significantly reduces computational overhead and storage requirements while maintaining high performance

<sup>7</sup><https://www.hpc.cineca.it/systems/hardware/leonardo/>

<sup>8</sup><https://github.com/LLaVA-VL/LLaVA-NeXT>

<sup>9</sup><https://streamlit.io/>

DIFFUSE-LARGE: large-scale (e.g. larger than few arcminutes and covering a large portion of the image) diffuse object with irregular shape.  
An island is a group of 4-connected pixels in an image under analysis with intensity above a detection threshold with respect to the sky background level.

### Question: Which of these morphological classes of radio sources do you see in the image?  
EXTENDED  
DIFFUSE  
DIFFUSE-LARGE  
Answer the question using the provided context (and examples). Report the identified class labels separated by commas, without any additional explanation text. Report just NONE if you cannot recognise any of the above classes in the image.

**B2 - Source Morphology Classification** We considered the *rgz-smorph* test dataset, containing  $\sim 3,835$  images from the VLA FIRST survey, each zoomed and centred around radio sources belonging to six distinct morphological classes: 1C-1P, 1C-2P, 1C-3P, 2C-2P, 2C-3P, 3C-3P.

The model was evaluated using a single-label multi-class classification task with the following prompt:

### Context: Consider these morphological classes of radio astronomical sources:  
1C-1P: single-island sources having only one flux intensity peak;  
1C-2C: single-island sources having two flux intensity peaks;  
1C-3P: single-island sources having three flux intensity peaks;  
2C-2P: sources consisting of two separated islands, each hosting a single flux intensity peak;  
2C-3P: sources consisting of two separated islands, one containing a single peak of flux intensity and the other exhibiting two distinct intensity peaks;  
3C-3P: sources consisting of three separated islands, each hosting a single flux intensity peak.  
An island is a group of 4-connected pixels in an image under analysis with intensity above a detection threshold with respect to the sky background level.

### Question: Which of these morphological classes of radio sources do you see in the image?  
1C-1P  
1C-2C  
1C-3P  
2C-2P  
2C-3P  
3C-3P  
Answer the question using the provided context (and examples). Report only the identified class label, without any additional explanation text.

**B3 - Extended Radio Galaxy Detection** We used the *radioimg-multilabel* test dataset (5,718 images) to assess the models' ability to identify radio sources with morphologies characteristic of extended radio galaxies.

For this task, we applied the following prompt:

Do you see any likely radio galaxy with an extended morphology in the image?  
Answer concisely: Yes or No.

**B4 - Imaging Artefact Detection** We used the *radioimg-multilabel* test dataset (5,718 images) to evaluate the models' ability to detect imaging artefacts in input images.

For this task, we considered the following prompt:

Do you see any imaging artefact with a ring pattern around bright sources in the image?  
Answer concisely: Yes or No.

**B5 - Source Peculiar Morphology Classification** We used the *radioimg-multilabel* test dataset (5,718 images) to evaluate the models' ability to detect radio sources with complex or anomalous morphologies in input images. These sources were previously classified into three categories: PECULIAR (150 images), COMPLEX (1,978 images), and ORDINARY (3,590 images).

For this task, we applied the following prompt:

### Context: Consider these radio image peculiarity classes, defined as follows:  
ORDINARY: image containing only point-like or slightly-resolved compact radio sources superimposed over the sky background or imaging artefact patterns;  
COMPLEX: image containing one or more radio sources with extended or diffuse morphology;  
PECULIAR: image containing one or more radio sources with anomalous or peculiar extended morphology, often having diffuse edges, complex irregular shapes, covering a large portion of the image.  
### Question: Can you identify which peculiarity class the presented image belongs to?  
ORDINARY  
COMPLEX  
PECULIAR  
Answer the question using the provided context (and examples). Report only the identified class label, without any additional explanation text.

**B6 - Radio Galaxy Morphology Classification** We used the *Mirabest* (Porter & Scaife, 2023) *confident sample* dataset, which contains 833 images from the VLA FIRST survey, each zoomed and centred around radio galaxies belonging to two distinct morphological classes: FR-I (397 images) and FR-II (436 images).

For this task, we applied the following prompt:

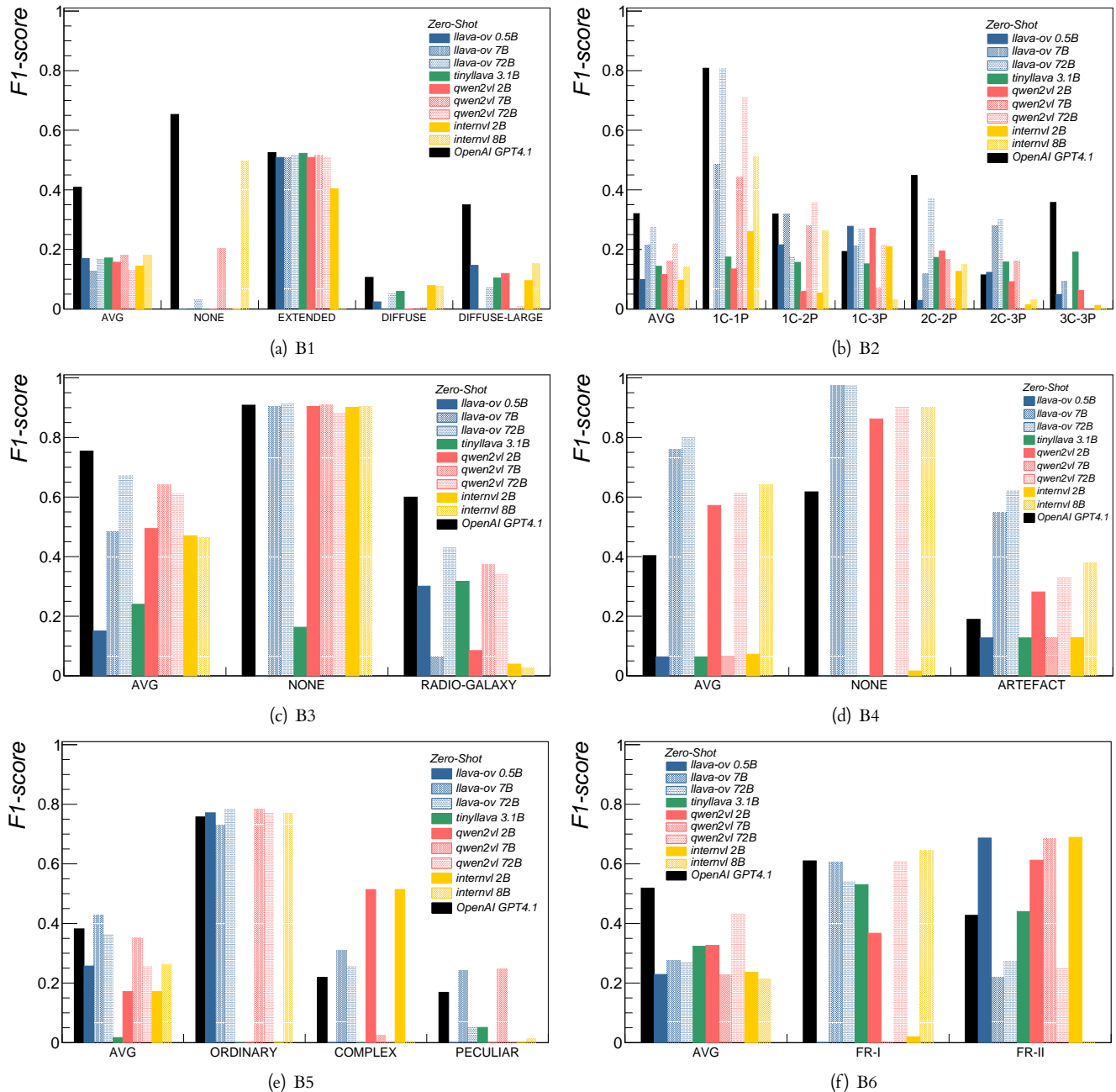
### Context: Consider these morphological classes of radio galaxies:  
FR-I: radio-loud galaxies characterised by a jet-dominated structure where the radio emissions are strongest close to the galaxy's centre and diminish with distance from the core;  
FR-II: radio-loud galaxies characterised by an edge-brightened radio structure, where the radio emissions are more prominent in lobes located far from the galaxy's core, with hotspots at the ends of powerful, well-collimated jets.  
### Question: Which of these morphological classes of radio galaxy do you see in the image?  
FR-I  
FR-II  
Answer the question using the provided context (and examples). Report only the identified class label, without any additional explanation text.

#### 4.1.2 Image standard benchmarks

We evaluated all *radio-llava* fine-tuned models on 11 image benchmarks (AI2, ChartQA, DocVQA, InfoVQA, MME, MMMU, MMStar, OCRBench, SEED-Bench, ScienceQA-IMG, RealWorldQA), which are widely used to assess multimodal model performance across various tasks, ranging from diagram, chart, and scene understanding to text extraction. Further details on each benchmark are provided in Appendix B.

#### 4.2 Zero-shot performance

We evaluated the zero-shot performance of LLaVA models of varying sizes on radio benchmarks, comparing with alternative open-weight VLMs and a representative commercial



**Figure 2.** Classification F1-scores obtained with VLMs of different sizes (0.5B, 2B, 3.1B, 7B, 8B, 72B) in zero-shot mode over B1-B6 evaluation benchmarks. We report the F1-score for individual classes, as well as the class-averaged F1-score (labelled as 'AVG'). LLaVA, TinyLLaVA, Qwen2VL, and InternVL models are respectively shown with blue, green, red, and orange histograms. OpenAI GPT4.1 model is shown with black histograms.

model (OpenAI GPT 4.1). Results are reported in Figure 2 and discussed in the following paragraphs. For each benchmark, we report the classification F1-score for individual classes, as well as the average F1-score across all classes (labelled as 'AVG' in the plots).

#### 4.2.1 Open-weight models

In Figure 2, we present the benchmark evaluation results for the base LLaVA-OneVision models (0.5B, 7B, 72B), shown in blue histograms, compared against alternative open-weight VLM models: TinyLLaVA 3.1B (green histogram), Qwen2VL

models (2B, 7B, 72B) (red histograms), and InternVL models (2B, 8B) (orange histograms).

As expected, smaller models (0.5B-3.1 B) perform consistently worse across most benchmarks, while larger models (Qwen2VL 72B, InternVL 8B, and LLaVA 72B) tend to achieve the best performance, particularly in B3 (radio galaxy detection), B4 (artifact detection), and B6 (FR-I vs. FR-II classification). In B1 (extended/diffuse source detection) and B2 (morphology classification), performance remains generally low across all models, with no significant advantage for any specific one. The best results are observed in B3 and B4,



where LLaVA 7B/72B models achieve competitive or slightly better performance compared to recently released VLMs. For instance, in artifact detection (B4), they attain a respectable 50–60% F1-score in a zero-shot setting. B5 (peculiar/complex morphology classification) and B6 (FR-I vs. FR-II classification) present significant challenges for all models, including the largest ones. Overall, the results indicate poor performance across all benchmarks, underscoring the need for models specialized in astronomical data.

#### 4.2.2 Commercial closed-weight models

Performing a comprehensive end-to-end benchmark evaluation across major proprietary solutions (e.g. OpenAI GPT, Google Gemini, Anthropic Claude) is not straightforward, as it would require academic institution to enter into contractual agreements with private providers to cover the cost of executing benchmarks via their APIs. Unlike public user interfaces, these APIs typically operate under separate pricing and access tiers. Nonetheless, we recognise the value of such an analysis for understanding the feasibility and cost-effectiveness of commercial LLM APIs in scientific benchmarking. Therefore, we made an effort to evaluate at least one commercial model – GPT-4.1 via the OpenAI API platform.

Benchmarks were split into multiple sub-tasks of approximately 500 images each to stay within the maximum batch file size limit (200 MB), with each mini-batch consisting of 80K–230K input tokens, costing around 0.23\$ – implying a total of ~2.5\$ per benchmark and under 20\$ for the full suite. As Tier 1 users, we were able to run one or two mini-batches per day without exceeding the token-per-day (TPD) limit of 900,000 tokens.

The GPT-4.1 benchmark results, shown in Figure 2 as black histograms, indicate superior performance in tasks B1–B3 and B6. In tasks B1 and B6, GPT-4.1 outperforms all open-weight models by a substantial margin – approximately 20% in average classification score. For tasks B2 and B3, the improvement is more modest, generally below 10%. Interestingly, GPT-4.1 underperforms in tasks B4 and B5, where its classification metrics fall below those of several open-weight models. These results may reflect both the advantage conferred by GPT-4’s significantly larger parameter count<sup>10</sup> and broader pretraining corpus, as well as limitations in its exposure to domain-specific astronomical concepts or visual patterns. While GPT-4.1 currently achieves the best overall performance in our benchmarks, the relatively small gap in several tasks – combined with the flexibility, transparency, and lower deployment costs of open-weight models – suggests there remains meaningful room for their development and application in specialized astronomical workflows.

### 4.3 Fine-tuning performance

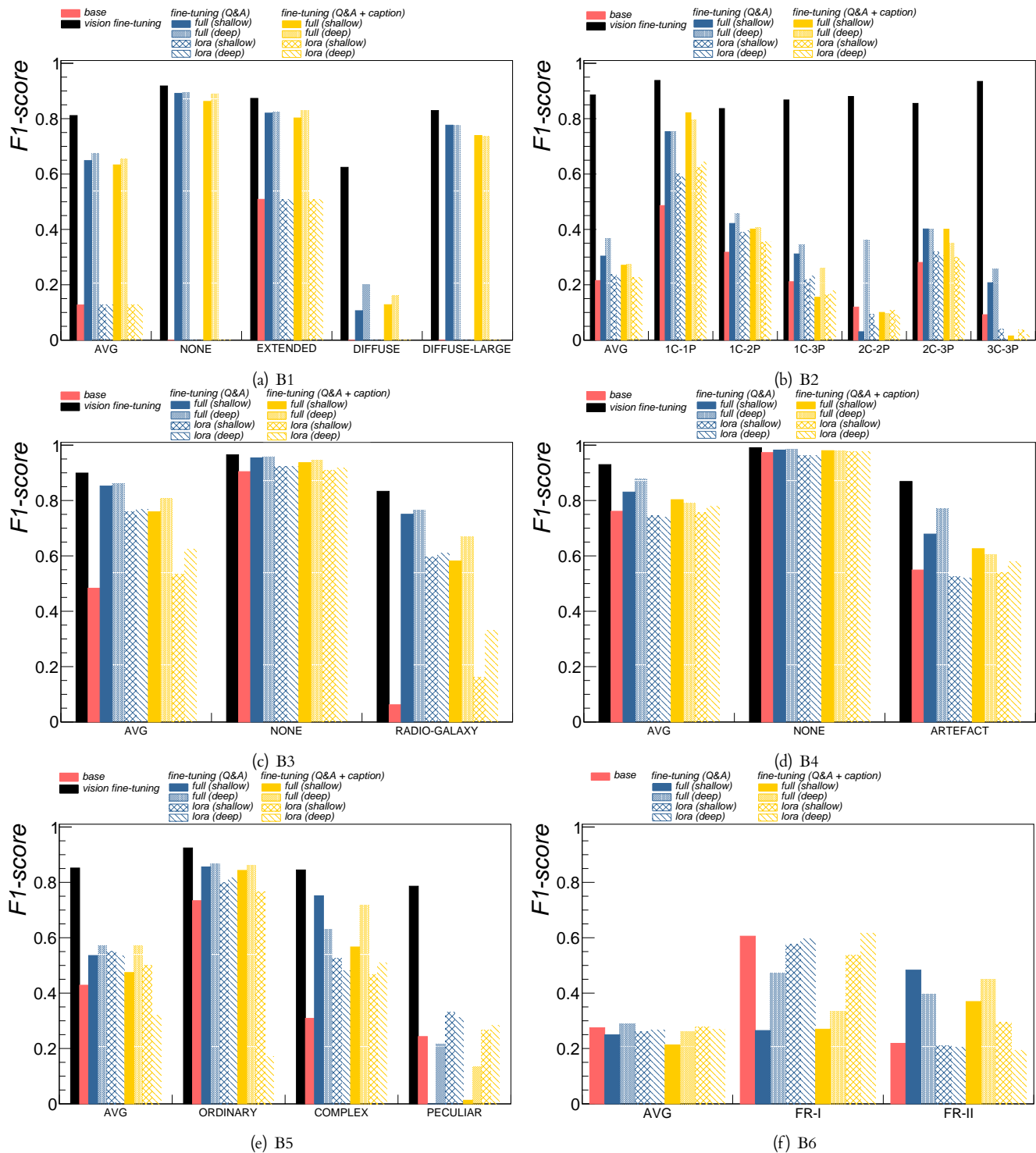
<sup>10</sup>Notably, GPT-4 models are estimated to have approximately 25 times more parameters (around 1.8 trillion parameters from various sources) than the largest open-weight models evaluated in this work

#### 4.3.1 Radio benchmarks

In Figure 3 we report the classification F1-score of *radio-llava* fine-tuned models obtained on radio benchmarks for each class and overall (labelled as ‘AVG’), compared to the base LLaVA-OneVision 7B model (solid red histograms). Blue histograms represent models fine-tuned on the Q&A dataset, using either deep/shallow full fine-tuning or LoRA fine-tuning. Orange histograms correspond to models fine-tuned on the combined Q&A and caption datasets. For comparison, we also report baseline metrics (shown as black histograms) obtained using a vision-only classifier that shares the same vision encoder as the LLaVA model (siglip-so400m-patch14-384). This classifier was fine-tuned and evaluated on the same training and test datasets.

With the exception of B6, we observe a general improvement in performance when fully fine-tuning the base model. The performance boost is particularly notable for B1 (extended/diffuse source detection) and B3 (radio galaxy classification), where average classification scores improve by more than 20–30%. For the remaining tasks, the improvement is more moderate (~10%). In contrast, LoRA fine-tuning leads to a clear improvement only in B3 and B5 tasks, with limited gains elsewhere. Deeper fine-tuning results in a modest improvement of only a few percentage points across all tasks, for both full and LoRA fine-tuning strategies. Fine-tuning on caption data (orange histograms) is observed to slightly decrease performance on radio benchmarks. This is somewhat expected, as all radio benchmarks are based on Q&A tasks rather than descriptive tasks. Caption data, on the other hand, have a positive impact on non-radio benchmarks, as discussed in the next section.

Overall, the achieved metrics remain well below those obtained using a vision-only model specialized for each task, which consistently reaches over 80–85% accuracy across all benchmarks – even after just 10 training epochs. The performance gap is especially notable in task B1, where the vision-only model attains an F1-score of approximately 60% for diffuse sources – class that multimodal models tend to struggle with, likely due to their underrepresentation in the training set (only 534 images). Similarly, in task B2, the model achieves over 80% accuracy across all morphological classes. These results also surpass our previous baseline of 74% average F1-score (Riggi et al., 2024c), which was obtained by training a LightGBM classifier on features extracted solely from the SigLIP vision encoder. At present, a fair comparison between the vision-only and LLaVA models cannot be provided for benchmark B6 due to the lack of a shared training dataset. Specifically, the FR-I/FR-II labels used in the LLaVA training were derived from object detection conversations on wide-field images, whereas training a vision-only classifier would require centred cutouts around FR-I/FR-II sources. Previous FR-I/FR-II classification studies (Cecconello et al., 2024; Slijepcevic et al., 2024) have achieved over 90% classification accuracy using smaller fine-tuned encoders (ResNet18/ResNet50). However, those models were trained on an independent subset of the same survey data (MiraBest dataset, VLA survey) used for test-



**Figure 3.** Classification F1-scores obtained with the *radio-llava* model on B1–B6 radio benchmarks, comparing fine-tuning on the Q&A training dataset (blue histograms) and the combined Q&A and caption datasets (orange histograms). For each training set, results are reported for different training strategies (full vs. LoRA fine-tuning) and training depths (shallow vs. deep). Results from the base model are shown as filled red histograms. Results obtained with a vision-only model (siglip-so400m-patch14-384 encoder) are shown as filled black histograms. The class-averaged F1-scores are labelled as 'AVG'.

ing. In contrast, our work trained on different survey data with varying source/image size ratios – specifically, using ASKAP EMU pilot data for training and VLA zoomed-in source images for testing. This difference likely contributes to the poor results observed on B6. We plan to update the dataset accord-

ingly in future work to enable a consistent evaluation across both models.

These findings suggest that the visual encoder provides a strong data representation, even for radio data, justifying our decision to keep it frozen during *radio-llava* fine-tuning. This

also indicates that the suboptimal performance of our multi-modal models is likely due to visual-language misalignment and the limited size and quality of the training dataset. Indeed, when we attempted to fully fine-tune *radio-llava*, including the vision encoder, we observed only a minor performance improvement ( $\sim 2\%$ ). However, it is important to note that, unlike specialized vision encoder models, *radio-llava* was trained to learn multiple radio tasks simultaneously.

### 4.3.2 Standard benchmarks

For comparison, Figure 4 reports performance on non-radio benchmarks, using the same colour labelling scheme of Figure 3. Consistent with previous studies (Pan et al., 2024) specializing LLMs for astronomy, we observed a notable decline in model performance on previously learned tasks compared to the base model (solid red histograms). This task forgetting effect is particularly pronounced in full fine-tuning, becoming significant (more than a 20% accuracy drop) in deeper training runs. In line with findings by Biderman et al. (2024), LoRA fine-tuned models achieve lower performance on radio benchmarks but are more robust against task forgetting.

Catastrophic forgetting remains a critical challenge when fine-tuning LLMs. Recent studies (Zhai et al., 2023; Zhang et al., 2024) have analysed this effect in multimodal models and proposed various strategies to mitigate it. One promising approach, successfully explored for language models by de Haan et al. (2024), involves expanding and curating the instruction-tuning dataset, followed by merging fine-tuned models with base models, using customizable balancing weights<sup>11</sup>. In this work, we nearly doubled the size of our initial Q&A dataset, enriching it with more diverse image captions extracted from a large collection of scientific papers. As shown in Figure 4, incorporating caption data (orange histograms) helped recover approximately 10 accuracy points across all standard benchmarks. This confirms trends observed in previous studies and underscores the importance of further curating our training dataset to enable future improvements.

### 4.3.3 Diagnostic analysis

To assess the impact of the default training configuration on model performance, we fine-tuned the model on the Q&A dataset with alternative choices of selected hyperparameters, resulting in various model variants, which are labelled and summarised in Table 1.

- Model v1 and v2 were trained with alternative learning rates ( $5 \times 10^{-5}$ ,  $5 \times 10^{-6}$ ) compared to the default  $10^{-5}$ ;
- Model v3 and v4 explored alternative learning schedulers: v3 used a "faster" warmup phase with `warmup_ratio=0.01` (compared to the 0.03 default), while v4 employed a `cosine_with_min_lr` scheduler with a minimum learning rate of  $5 \times 10^{-6}$ , instead of the unconstrained cosine scheduler;

<sup>11</sup>The merging tool used in de Haan et al. (2024) – MergeKit <https://github.com/arcee-ai/mergekit> – currently supports merging only the LLM components and requires extension to also include the LLaVA adapter and vision components.

- Model v5 was trained on 32 4 GPU nodes (`batch_size=1`, `gradient_accumulation_step=2`) to obtain a larger effective batch size of 256 (compared to the default 8);
- Model v6 used LoRA fine-tuning with larger ranks ( $r=128$ ,  $\alpha=256$ ) instead of the previously tested  $r=64$ ,  $\alpha=128$ .

Additionally, we examined the impact of the user prompt by repeating the benchmark evaluation with a more structured prompt version.

Figure C.2 in the Appendix compares average metrics across all radio tasks for the original and variant models: the black solid histogram represents the original model, the black dashed histogram represents the original model with the alternative prompt, and the coloured histograms correspond to the model variants. From the results, we conclude that the alternative configurations tested do not lead to performance improvements. Thus, the suboptimal performance is unlikely due to non-optimal hyperparameter selection but rather to dataset quality limitations. Furthermore, the reported metrics show minimal variation with the adopted prompt.

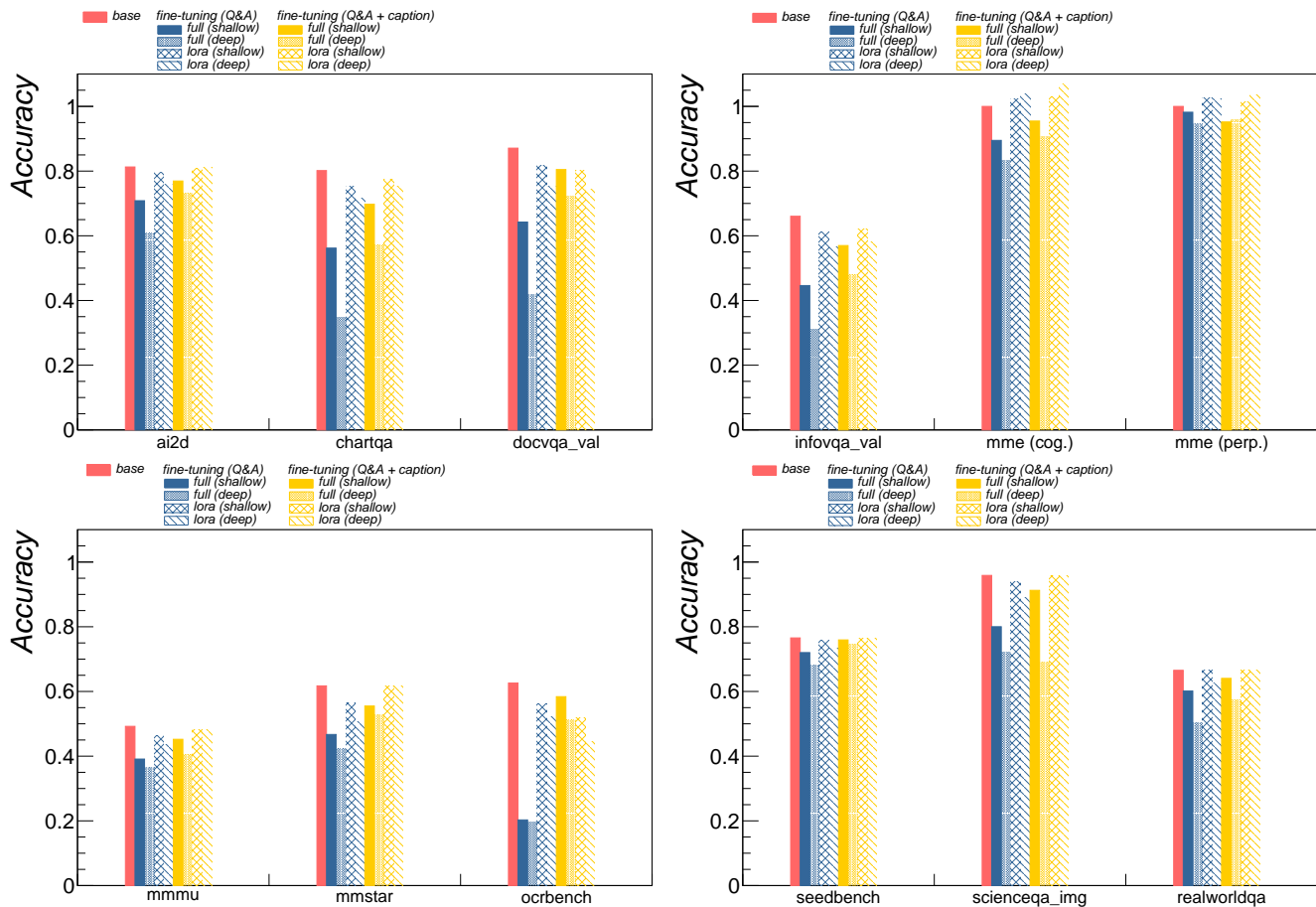
To assess the impact of using automated data curation via InternVL, we initially conducted experiments using a fully templated, uncured dataset without InternVL-generated variation. While benchmark performance metrics remained broadly similar, we observed a marked degradation in the model outputs, with responses frequently mirroring the rigid structure of the templates. This motivated us to setup an automated data curation to enhance linguistic diversity and model conversational ability.

**Table 1.** Summary of fine-tuned models with alternative hyperparameter configurations.

Model variant	Parameters
v1	<code>lr=<math>5 \times 10^{-5}</math></code>
v2	<code>lr=<math>5 \times 10^{-6}</math></code>
v3	<code>warmup_ratio=0.01</code>
v4	<code>cosine_with_min_lr (lr=<math>5 \times 10^{-6}</math>)</code>
v5	<code>effective_batch_size=256</code>
v6	<code>LORA rank=128, alpha=256</code>

## 5. Summary

In this work, we investigated the feasibility of using small-scale Vision-Language Models (VLMs) as AI assistants for analysing radio images, enabling tasks such as source classification, identification of specific object classes, and data exploration for quality assessment. Unlike conventional deep learning approaches, VLMs offer a more flexible, natural-language-driven interaction, reducing the need for complex coding or task-specific model adaptation. To this end, we fine-tuned LLaVA, a state-of-the-art VLM, on a custom dataset of over 59,000 radio images paired with instruction-based queries, along with an additional 38,000 image-caption pairs extracted from a large corpus of radio astronomical papers. The fine-tuning process leveraged both Q&A interactions and descriptive captions, enabling the model to handle a variety of radio analysis tasks, including



**Figure 4.** Classification accuracy obtained with the *radio-llava* model on standard non-radio benchmarks (Section 4.1.2), comparing fine-tuning on the Q&A training dataset (blue histograms) and the combined Q&A and caption datasets (orange histograms). For each training set, results are reported for different training strategies (full vs. LoRA fine-tuning) and training depths (shallow vs. deep). Results from the base model are shown as filled red histograms.

source morphology classification, extended source detection, and artifact identification. The resulting *radio-llava* model was evaluated across six radio-specific benchmarks (B1–B6) and compared against baseline VLMs on non-astronomical multimodal tasks. Fine-tuned models and the developed software have been publicly released.

Our key findings can be summarised as follows:

- *Fine-tuning improves performance:* Compared to the base model, *radio-llava* exhibits significant performance gains on radio benchmarks, particularly in extended source detection (B1) and radio galaxy classification (B3), with F1-score improvements exceeding 20–30%;
- *Challenges in multimodal alignment:* Despite fine-tuning, pure vision models still outperform VLMs, suggesting that visual-language alignment remains a limiting factor. Full fine-tuning of both vision and language components resulted in only marginal improvements (~2%);
- *Task forgetting effect:* While fine-tuned models improve in radio-specific tasks, they suffer from catastrophic forgetting when evaluated on general multimodal benchmarks. This effect is more severe for full fine-tuning (~20% accuracy drop), while LoRA fine-tuned models exhibit better

retention of prior knowledge. Fine-tuned models were also observed to exhibit degraded conversational capabilities;

- *Impact of caption data:* Incorporating descriptive captions from scientific literature into the training set enhances model generalization, helping recover ~10 accuracy points on standard multimodal benchmarks while also improving instruction-following abilities.

These findings highlight the potential of compact multimodal models for radio astronomy while also revealing key limitations that require further research to fully match the performance of specialized vision models. Future efforts should focus on improving vision-language alignment, curating larger, high-quality training datasets, and exploring hybrid fine-tuning strategies also for larger models (~70B) to mitigate task forgetting while maximizing domain-specific performance. Additionally, we plan to leverage the multi-image processing capabilities of the LLaVA-OneVision model for in-context learning of analysed tasks. Future investigations will also explore its performance on new tasks requiring comparative analysis across multiple images, such as image retrieval of known source classes.



## References

- OpenAI, Achiam, J., et al., *GPT-4 Technical Report*, 2023, <https://doi.org/10.48550/arXiv.2303.08774>
- Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku*, 2024, <https://api.semanticscholar.org/CorpusID:268232499>
- Agarwal, D., et al., 2020, MNRAS, 497, 1661, <https://doi.org/10.1093/mnras/staa1856>
- Aniyan, A. K., Thorat, K., 2017, ApJS, 230, 20, <https://doi.org/10.3847/1538-4365/aa7333>
- Bai, J., et al., *Qwen Technical Report*, 2023, <https://doi.org/10.48550/arXiv.2309.16609>
- Banfield, J. K., et al., 2015, MNRAS, 453, 2326, <https://doi.org/10.1093/mnras/stv1688>
- Barkai, J. A., et al., 2023, A&A, 670, A55, <https://doi.org/10.1051/0004-6361/202244708>
- Biderman, D., et al., *LoRA Learns Less and Forgets Less*, 2024, <https://doi.org/10.48550/arXiv.2405.09673>
- Becker, R. H., et al., 1995, ApJ, 450, 559, <https://doi.org/10.1086/176166>
- Bordes, F., et al., *An Introduction to Vision-Language Modeling*, 2024, <https://doi.org/10.48550/arXiv.2405.17247>
- Bordiu, C., et al., 2025, A&A, 695, A144, <https://doi.org/10.1051/0004-6361/202450356>
- Cai, Z., et al., *InternLM2 Technical Report*, 2024, <https://doi.org/10.48550/arXiv.2403.17297>
- Cecconello, T., et al., Proceedings of the 13th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), 1–5 December 2024, Kolkata, India, <https://doi.org/10.48550/arXiv.2411.14078>
- Chen, L., et al., *Are we on the right way for evaluating large vision language models?*, 2024, <https://doi.org/10.48550/arXiv.2403.20330>
- Chen, Z., et al., *Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 24185–24198, 2024, <https://doi.org/10.48550/arXiv.2312.14238>
- Chiche, B. N., et al., 2023, A&A 675, A116, <https://doi.org/10.1051/0004-6361/202245013>
- Connor, L., & van Leeuwen, J., 2018, AJ, 156, 256, <https://doi.org/10.3847/1538-3881/aae649>
- Cornu, D., et al., 2024, A&A, 690, A211, <https://doi.org/10.1051/0004-6361/202449548>
- de Haan, T., et al., 2025, Sci Rep, 15, 13751, <https://doi.org/10.1038/s41598-025-97131-y>
- Dewdney, P., et al., 2016, SKA1 SYSTEM BASELINE DESIGN V2, SKA-TEL-SKO-0000002, [https://www.skao.int/sites/default/files/documents/d1-SKA-TEL-SKO-0000002\\_03\\_SKA1SystemBaselineDesignV2\\_1.pdf](https://www.skao.int/sites/default/files/documents/d1-SKA-TEL-SKO-0000002_03_SKA1SystemBaselineDesignV2_1.pdf)
- Dosovitskiy A., et al., *An image is worth 16×16 words: Transformers for image recognition at scale*, 2021, Proc. Int. Conf. Learn. Representations, <https://doi.org/10.48550/arXiv.2010.11929>
- Håkansson, H., et al., 2023, A&A, 671, A39, <https://doi.org/10.1051/0004-6361/202245139>
- Hopkins, A. M., et al., 2025, PASA, 42, e071, <https://doi.org/10.1017/pasa.2025.10042>
- Chaoyou Fu, C., et al., *Mme: A comprehensive evaluation benchmark for multimodal large language models*, 2024, <https://doi.org/10.48550/arXiv.2306.13394>
- Galvin, T. J., et al., 2020, MNRAS, 497, 2730, <https://doi.org/10.1093/mnras/staa1890>
- Geyer, F., et al., 2023, A&A, 677, A167, <https://doi.org/10.1051/0004-6361/202347073>
- Goedhart, S., et al., 2024, MNRAS, 531, 649, <https://doi.org/10.1093/mnras/stae1166>
- Gupta, N., et al., 2022, PASA, 39, e051, <https://doi.org/10.1017/pasa.2022.44>
- Gupta, N., et al., 2023, PASA, 40, e044, <https://doi.org/10.1017/pasa.2023.46>
- Gupta, N., et al., 2024, PASA, 41, e027, <https://doi.org/10.1017/pasa.2024.25>
- Gupta, N., et al., 2025, *EMUSE: Evolutionary Map of the Universe Search Engine*, 2025, PASA, in press, <https://doi.org/10.1017/pasa.2025.10064>
- He, K., et al., *Mask R-CNN*, 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980–2988, <https://doi.org/10.1109/ICCV.2017.322>
- Hu, E. J., et al., *LoRA: Low-Rank Adaptation of Large Language Models*, 2021, <https://doi.org/10.48550/arXiv.2106.09685>
- Jia, J., et al., *TinyLLaVA Factory: A Modularized Codebase for Small-scale Large Multimodal Models*, <https://doi.org/10.48550/arXiv.2405.11788>
- Kembhavi, A., et al., *A diagram is worth a dozen images*, In ECCV, 2016, 10, 37, 39, <https://doi.org/10.48550/arXiv.1603.07396>
- Lao, B., et al., 2023, Astronomy and Computing, 44, 100728, <https://doi.org/10.1016/j.ascom.2023.100728>
- Lastufka, E., et al., 2024, A&A, 690, A310, <https://doi.org/10.1051/0004-6361/202449964>
- Lastufka, E., et al., *Bridging the Gap: Examining Vision Foundation Models for Optical and Radio Astronomy Applications*, 2024, <https://doi.org/10.48550/arXiv.2409.11175>
- Li, C., et al., *LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day*, 2023, <https://doi.org/10.48550/arXiv.2306.00890>
- Li, C., et al., *Multimodal Foundation Models: From Specialists to General-Purpose Assistants*, 2023, <https://doi.org/10.48550/arXiv.2309.10020>
- Li, B., et al., *SEED-bench: Benchmarking multimodal LLMs with generative comprehension*, 2023, <https://doi.org/10.48550/arXiv.2307.16125>
- Li, B., et al., *LLaVA-OneVision: Easy Visual Task Transfer*, <https://doi.org/10.48550/arXiv.2408.03326>
- Liang, R., et al., 2023, Res. Astron. Astrophys., 23, 115006, <https://doi.org/10.1088/1674-4527/acd0ed>
- Liu, H., et al., *Visual instruction tuning*, NeurIPS 2023, <https://doi.org/10.48550/arXiv.2304.08485>
- Liu, H., et al., *Improved baselines with visual instruction tuning*, CVPR 2024, <https://doi.org/10.48550/arXiv.2310.03744>
- Liu, H., et al., *LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*, January 2024, <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
- Liu, Y., et al., 2024, Sci. China Inf. Sci., 67, 220102 <https://doi.org/10.1007/s11432-024-4235-6>
- Lochner, M., et al., 2023, MNRAS, 520, 1439, <https://doi.org/10.1093/mnras/stad074>
- Lou, K., et al., 2023, Res. Astron. Astrophys., 23, 075012, <https://doi.org/10.1088/1674-4527/acd16b>
- Lu, P., et al., *Learn to explain: Multimodal reasoning via thought chains for science question answering*, In The 36th Conference on Neural Information Processing Systems (NeurIPS), 2022, 10, 39, <https://doi.org/10.48550/arXiv.2209.09513>
- Ma, Z., et al., 2019, ApJS, 240, 34, <https://doi.org/10.3847/1538-4365/aaf9a2>
- Martínez, T. V., et al., 2024, A&A, 691, A360, <https://doi.org/10.1051/0004-6361/202451429>
- Maslej-Krešňáková, V., et al., 2021, MNRAS, 505, 1464, <https://doi.org/10.1093/mnras/stab1400>
- Masry, A., et al., *Chartqa: A benchmark for question answering about charts with visual and logical reasoning*, In ACL, 2022, 9, 10, 36, 37, 39, <https://doi.org/10.48550/arXiv.2203.10244>
- Mathew, M., et al., *Docvqa: A dataset for VQA on document images*, In WACV, 2021, 9, 10, 36, 37, 39, 40, <https://doi.org/10.48550/arXiv.2007.00398>
- Mathew, M., et al., *InfographicVQA*, In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706, 2022, 9, 10, 36, 39, <https://doi.org/10.48550/arXiv.2104.12756>
- Mesarcik, M., et al., 2023, A&A, 680, A74, <https://doi.org/10.1051/0004-6361/202347182>
- Mishra-Sharma, S., Song, Y., & Thaler, J., 2024, *PAPERCLIP: Associating Astronomical Observations and Natural Language with Multi-Modal Models*, In First Conference on Language Modeling, <https://doi.org/10.48550/arXiv.2403.08851>
- Mostert, R. I.J., et al., 2021, A&A, 645, A89, <https://doi.org/10.1051/0004-6361/202038500>
- Mostert, R. I.J., et al., 2022, A&A, 668, A28, <https://doi.org/10.1051/0004-6361/202243478>

- Nair, A., et al., 2022, MNRAS, 511, 4946, <https://doi.org/10.1093/mnras/stac204>
- Nguyen, T. D., et al., *AstroLLaMA: Towards Specialized Foundation Models in Astronomy*, 2023, <https://doi.org/10.48550/arXiv.2309.06126>
- Norris, R. P., et al., 2021, PASA, 38, e046, <https://doi.org/10.1017/pasa.2021.42>
- Pan, R., et al., *AstroMLab 2: AstroLLaMA2-70B Model and Benchmarking Specialised LLMs for Astronomy*, 2024, <https://doi.org/10.48550/arXiv.2409.19750>
- Parker, L., et al., 2024, MNRAS, 531, 4990, <https://doi.org/10.1093/mnras/stae1450>
- Perkowski, E., et al., 2024, Res. Notes AAS, 8, 7, <https://doi.org/10.3847/2515-5172/ad1abe>
- Porter, F. A.M., Scaife, A. M.M., 2023, RAS Techniques and Instruments, 2, 293, <https://doi.org/10.1093/rasti/rzad017>
- Radford, A., et al., *Learning transferable visual models from natural language supervision*, In International Conference on Machine Learning, 2021, PMLR, 8748, <https://doi.org/10.48550/arXiv.2103.00020>
- Ralph, N. O., et al., 2019, PASP, 131, 108011, <https://doi.org/10.1088/1538-3873/ab213d>
- Redmon, J., et al., *You only look once: Unified, real-time object detection*, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788, 2016, <https://doi.org/10.48550/arXiv.1506.02640>
- Riggi, S., et al., 2023, Astronomy and Computing, 42, 100682, <https://doi.org/10.1016/j.ascom.2022.100682>
- Riggi, S., et al., 2024, PASA, 41, e029, <https://doi.org/10.1017/pasa.2024.26>
- Riggi, S., et al., 2024, PASA, 41, e085, <https://doi.org/10.1017/pasa.2024.84>
- Riggi, S., et al., Proceedings of the 34th Astronomical Data Analysis Software and Systems conference, 10–14 November 2024, La Valletta, Malta, <https://doi.org/10.48550/arXiv.2411.08519>
- Rustige, L., et al., 2023, RAS Techniques and Instruments, 2, 264, <https://doi.org/10.1093/rasti/rzad016>
- Schmidt, K., et al., 2022, A&A 664, A134, <https://doi.org/10.1051/0004-6361/202142113>
- Slijepcevic, V. I., et al., 2024, RAS Techniques and Instruments, 3, 19, <https://doi.org/10.1093/rasti/rzad055>
- Sortino, R., et al., 2024, IEEE Transaction on Artificial Intelligence, <https://doi.org/10.1109/TAI.2024.3436538>
- Stuardi, C., et al., 2024, MNRAS, 533, 3194, <https://doi.org/10.1093/mnras/stae2014>
- Sun, Z., et al., *Interpreting Multi-band Galaxy Observations with Large Language Model-Based Agents*, 2024, <https://doi.org/10.48550/arXiv.2409.14807>
- Tang, H., et al., 2019, MNRAS, 488, 3358, <https://doi.org/10.1093/mnras/stz1883>
- Tang, H., et al., 2022, MNRAS, 510, 4504, <https://doi.org/10.1093/mnras/stab3553>
- Tanoglidis, D., and Jain, B., 2024, Res. Notes AAS, 8, 265, <https://doi.org/10.3847/2515-5172/ad887a>
- Touvron, H., et al., *LLaMA: Open and Efficient Foundation Language Models*, 2023, <https://doi.org/10.48550/arXiv.2302.13971>
- Umata, G., et al., 2015, MNRAS, 454, 902, <https://doi.org/10.1093/mnras/stv1976>
- Umata, G., et al., 2021, MNRAS, 506, 2232, <https://doi.org/10.1093/mnras/stab1279>
- Yin, S., et al., 2024, National Science Review, 11, 12, nwae403, <https://doi.org/10.1093/nsr/nwae403>
- Yu, L., et al., 2022, MNRAS, 511, 4305, <https://doi.org/10.1093/mnras/stac281>
- Xiang Yue, X., et al., *MMM: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI*, In CVPR, 2024, 9, 10, 36, 38, <https://doi.org/10.48550/arXiv.2311.16502>
- Wu, C., et al., 2019, MNRAS, 482, 1211, <https://doi.org/10.1093/mnras/sty2646>
- Zhai, Y., et al., *Investigating the Catastrophic Forgetting in Multimodal Large Language Models*, 2023, <https://doi.org/10.48550/arXiv.2309.10313>
- Zhai, X., et al., *Sigmoid loss for language image pretraining*, In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 11975, <https://doi.org/10.48550/arXiv.2303.15343>
- Zhang, Z., et al., 2022, PASP, 134, 064503, <https://doi.org/10.1088/1538-3873/ac67b1>
- Zhang, Y., et al., *Wings: Learning Multimodal LLMs without Text-only Forgetting*, 2024, <https://doi.org/10.48550/arXiv.2406.03496>
- Zhou, B., et al., *TinyLLaVA: A Framework of Small-scale Large Multimodal Models*, <https://doi.org/10.48550/arXiv.2402.14289>

## Acknowledgement

This scientific work uses data obtained from Inyarrimanha Ilgari Bundara / the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. CSIRO's ASKAP radio telescope is part of the Australia Telescope National Facility (<https://ror.org/05qajvd42>). Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Research Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund.

We acknowledge ISCRa for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Additional computing resources for this work were provided by the INAF "CIRASA" (*Collaborative and Integrated platform for Radio Astronomical Source Analysis*) project, the Italian PON 2014–2020 "MOSAICO" project, and the Italian PNRR Project IR0000034 "STILES" (*Strengthening the Italian leadership in ELT and SKA*) project.

**Funding Statement** This work received funding from the INAF "SCIARADA" grant.

**Competing Interests** None

**Data Availability Statement** The software code used in this work is publicly available under the GNU General Public License v3.0<sup>12</sup> on the GitHub repository <https://github.com/SKA-Infra/radio-llava/>. The trained model weights have been made available in this *Hugging Face* repository: <https://huggingface.co/inaf-oact-ai>.

<sup>12</sup><https://www.gnu.org/licenses/gpl-3.0.html>

## A. Training Datasets

### A.1 Coarse-grained radio datasets

We describe below the annotated datasets used to create the conversational train dataset.

#### *radioimg-multilabel dataset*

The dataset currently includes a collection of 19,060 annotated radio images taken from multiple radio surveys, carried out both in the Galactic Plane and outside:

- SRAO MeerKAT Galactic Plane Survey (SMGPS) (Goedhart et al., 2024): 2,704 images (14.2%)
- ASKAP EMU main survey (Hopkins et al., 2025): 4,456 images (23.4%)
- ASKAP EMU pilot survey (Norris et al., 2021): 5,860 images (30.7%)
- ASKAP EMU pilot Galactic Plane surveys (Umana et al., 2021): 6,040 images (31.7%)

We manually assigned the following labels to each image:

- **BACKGROUND**: If the image is purely background noise, e.g. no sources are visible. Typically, this label is set for frames located at the map borders;
- **COMPACT**: if point sources or compact sources comparable with the synthesized beam size (say  $<10$  times the beam) are present. Double or triple sources with point-like components also fall into this category;
- **EXTENDED**: if any extended source is visible, e.g. a compact source with extension  $>10 \times$  beam;
- **RADIO-GALAXY**: if any extended source is visible with a single- or multi-island morphology, suggesting that of a radio galaxy (e.g. core + lobes);
- **DIFFUSE**: if any diffuse source is visible, typically having small-scale (e.g.  $<\text{few arcmin}$ ) and roundish morphology;
- **DIFFUSE-LARGE**: if any large-scale (e.g. covering half of the image) diffuse object with irregular shape is visible;
- **FILAMENT**: if any extended filamentary structures is visible;
- **ARTIFACT**: if any ring-shaped or ray-like artefact is visible, e.g. typically around bright resolved sources;
- **PECULIAR**: if any object is found with peculiar/anomalous morphology;
- **MOSAICKING**: if any residual pattern of the mosaicking process used to produce the image is present;
- **BORDER**: if the image contains blank/NaN regions along its borders.

More than one label can be assigned to each image, depending on the object/features the user recognises in the image. The number of images that have been assigned each specific label is reported in Table A1.

The dataset was split into two samples. The first sample, containing 13,342 images, was used to generate the user-assistant conversations for the training Q&A dataset starting from the template image description created from assigned class labels, as described in Section 3.1.1. The rest of the dataset, consisting of 5,718 images, was used to evaluate the performance of trained models.

**Table A1.** The number of images in the *radioimg-multilabel* dataset that have been assigned each specific label. Multiple labels can be assigned to a single image, as they are not mutually exclusive.

label	#images
BACKGROUND	116
COMPACT	18,671
EXTENDED	3,279
RADIO-GALAXY	3,269
DIFFUSE	757
DIFFUSE-LARGE	1438
FILAMENT	50
ARTIFACT	1283
PECULIAR	439
MOSAICKING	260
BORDER	453

#### *rgz-smorph dataset*

The dataset currently includes a collection of 9,570 radio images extracted from the VLA FIRST survey (Becker et al., 1995) and annotated in the Radio Galaxy Zoo (RGZ) crowd-sourced project (Banfield et al., 2015). Each image is centred and zoomed on radio sources of 6 different morphological classes, defined on the basis of the observed number of components (C) and peaks (P) as follows: 1C-1P, 1C-2P, 1C-3P, 2C-2P, 2C-3P, 3C-3P. The entire dataset was split into two samples. The first one, containing 5,735 images ( $\sim 1000$  per class), was used to create the conversational dataset, while the remaining sample (3,835 images,  $\sim 600$  per class) was reserved for model evaluation scopes.

#### *smgps-extcat dataset*

The dataset currently includes a collection of 17,062 radio images extracted from the SMGPS survey (Goedhart et al., 2024), each centred and zoomed<sup>13</sup> on radio sources listed in the SMGPS extended source catalogue (Bordiu et al., 2025). This includes single- or multi-island sources with morphologies classified as: **EXTENDED** or **DIFFUSE**. Furthermore, a fraction of the catalogued sources also have an astronomical class label obtained either through morphological considerations or cross-matching with various Galactic source catalogues (see Bordiu et al. 2025 for details). Available class labels are: **GALAXY** (radio galaxy), **HII** (HII region), **PN** (planetary nebula), **SNR** (supernova remnant), **PULSAR** (pulsar), **STAR** (generic radio star), **YSO** (young stellar objects), **LBV** (luminous blue variable star), **WR** (Wolf-Rayet star), **HMXB** (high-mass X-ray binary), **LMXB** (low-mass X-ray binary). Sources cross-matching to multiple catalogues have more than one label assigned. All the above source annotations are taken into account to generate the conversational dataset.

### A.2 Fine-grained radio datasets

We describe below the annotated datasets used to create conversational train datasets that contain precise object localization

<sup>13</sup>The original image crop size is set to 1.5 times the size of the source bounding box.



information.

### **caesar-mrcnn dataset**

The dataset currently contains 12,774 annotated radio images taken from different surveys, such as the VLA FIRST (Becker et al., 1995), ATCA Scorpio (Umana et al., 2015), and ASKAP-EMU Scorpio (Umana et al., 2021) surveys. The annotation data include bounding boxes, segmentation masks and classification labels for all radio object identified in the images (38,342 objects, including both real and spurious sources). Objects are classified into five possible classes: SPURIOUS, COMPACT, EXTENDED, EXTENDED-MULTISLAND, FLAGGED. A detailed explanation of the labelling scheme is provided in the reference publication (Riggi et al., 2023). The entire dataset was used to produce the Q&A training dataset.

### **emu-pilot-rgcat dataset**

The dataset currently contains 10,414 annotated radio images taken from the ASKAP EMU pilot survey (Norris et al., 2021), each containing at least one extended radio source. Annotation data have been extracted from EMU pilot RG-CAT catalogue (Gupta et al., 2024), including bounding boxes and classification labels for radio objects present in the images. Objects in the original catalogue are classified into six possible radio galaxy morphology classes:

- C: compact radio galaxies;
- FR-I: radio galaxies of Fanaroff-Riley type I;
- FR-II: radio galaxies of Fanaroff-Riley type II;
- FR-x: radio galaxies with mixed or hybrid morphology, showing characteristics of both FR-I and FR-II galaxy classes;
- R: radio galaxies with single-peak resolved morphology;
- Pec: radio galaxies with a peculiar morphology;

A total of 185,294 objects were annotated according to RG-CAT catalogue.

From a visual inspection of the data, we note that various objects classified as compact (C) should be rather considered as belonging to the EXTENDED class in the classification scheme adopted in the *caesar-mrcnn* dataset (see previous paragraph). To make the two fine-grained datasets more comparable, we applied the following processing steps. We first extracted objects from EMU pilot images using *caesar-mrcnn* trained model (Riggi et al., 2023). As a result, we obtained a list of detected objects classified with the *caesar-mrcnn* classification scheme, that was cross-matched with the original RG-CAT object collection. This was extended and complemented according to the match results. Objects with a match (~78%) were also given a *caesar-mrcnn* label. Objects detected by the *caesar-mrcnn* model but missed in the RG-CAT were added to the final collection, including a total of 231,439 objects. The obtained source annotations were taken into account to generate the conversational dataset.

## **B. Image multi-modal benchmarks**

### **AI2**

This benchmark<sup>14</sup> consists of 3,088 image-based Q&A pairs on annotated grade school science diagrams from the AI2 Diagrams (AI2D) dataset (Kembhavi et al., 2016).

### **ChartQA**

This benchmark<sup>15</sup> contains 2,500 image-based Q&A pairs on real-world charts in various formats (pie, bar) from the ChartQA dataset (Masry et al., 2022).

### **DocVQA**

This benchmark<sup>16</sup> contains 16,626 image-based Q&A pairs on document of various types and content, sourced from the DocVQA dataset (Mathew et al., 2021).

### **InfoVQA**

This benchmark<sup>17</sup> contains 2,801 image-based Q&A pairs on document infographics of various types and content, sourced from the InfographicVQA dataset (Mathew et al., 2022).

### **MME**

This benchmark<sup>18</sup> consists of 2,374 image-based Q&A pairs from the MME dataset (Fu et al., 2024), designed to evaluate multimodal models' perception and cognition abilities. Perception tasks include OCR, recognition of coarse-grained objects (e.g., object presence, count, position, and colour) and fine-grained objects (e.g., identification of movie posters, celebrities, scenes, landmarks and artworks). Cognition tasks cover common sense reasoning, numerical calculation, text translation, and code reasoning.

### **MMMU**

This benchmark<sup>19</sup> consists of 900 image-based Q&A pairs from the MMMU dataset (Yue et al., 2024), designed to assess multimodal perception and reasoning abilities across various image formats, including charts, diagrams, maps, tables, music sheets, and chemical structures. The images are sourced from college exams, quizzes, and textbooks spanning six disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering.

### **MMStar**

This benchmark<sup>20</sup> contains 1,500 image-based Q&A pairs from the MMStar dataset (Chen et al., 2024), designed to evaluate multimodal models across six core capabilities: Coarse Perception, Fine-grained Perception, Instance Reasoning, Logical Reasoning, Science & Technology, Mathematics.

<sup>14</sup><https://huggingface.co/datasets/lmms-lab/ai2d>

<sup>15</sup><https://huggingface.co/datasets/lmms-lab/ChartQA>

<sup>16</sup><https://huggingface.co/datasets/lmms-lab/DocVQA>

<sup>17</sup><https://huggingface.co/datasets/lmms-lab/DocVQA>, see InfographicVQA validation data split

<sup>18</sup><https://huggingface.co/datasets/lmms-lab/MME>

<sup>19</sup><https://huggingface.co/datasets/lmms-lab/MMMU>, see validation data split

<sup>20</sup><https://huggingface.co/datasets/Lin-Chen/MMStar>



**OCRBench**

This benchmark<sup>21</sup> consists of 1,000 image-based Q&A pairs from the OCRBench dataset (Liu et al., 2024c), designed to assess Optical Character Recognition (OCR) capabilities across various domains, including multilingual text, handwritten text, non-semantic text, and mathematical expression recognition.

**SEED-Bench**

This benchmark<sup>22</sup> contains 17,990 image-based Q&A pairs from the SEED-Bench dataset (Li et al., 2023c), designed to evaluate multimodal model capabilities across various domains, including scene understanding, instance identity/attribute/location/counting, spatial relations, instance interaction, visual reasoning, text recognition, action recognition/prediction, procedure understanding.

**ScienceQA-IMG**

This benchmark<sup>23</sup> consists of 2,017 image-based Q&A pairs from the ScienceQA dataset (Lu et al., 2022), featuring multiple-choice questions spanning diverse topics across three subjects: natural science, social science, and language science.

**RealWorldQA**

This benchmark<sup>24</sup> contains 765 image-based Q&A pairs from the RealWorldQA dataset, featuring real-world scenarios, including images captured from vehicles.

<sup>21</sup><https://huggingface.co/datasets/echo840/OCRBench>

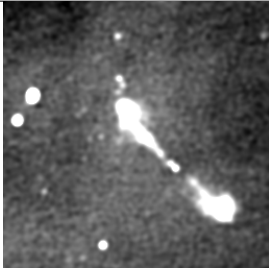
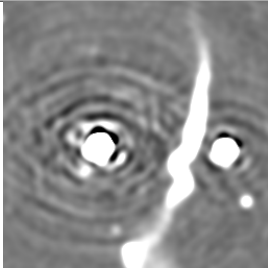
<sup>22</sup><https://huggingface.co/datasets/lmms-lab/SEED-Bench>

<sup>23</sup><https://huggingface.co/datasets/lmms-lab/ScienceQA>, see ScienceQA-IMG test data split

<sup>24</sup><https://huggingface.co/datasets/lmms-lab/RealWorldQA>

## C. Supplementary plots/tables

**Table C1.** User-assistant conversations on sample radio images for base and fine-tuned LLaVA-OneVision models.

		
User	<p>You are given a radio astronomical image as input. Answer to the question below, strictly following the provided instructions.</p> <p>## Question: Can you provide a brief description of the image?</p> <p>## Instructions:</p> <ul style="list-style-type: none"> <li>– Base your description solely on the objects and patterns visible in the image.</li> <li>– Do not include details that are not present in the image.</li> <li>– Use a scientific and astronomical writing style.</li> <li>– Limit your response to a maximum of four sentences.</li> </ul>	
llava-ov	The image depicts a section of the sky with several bright, white spots against a dark background. These spots are likely to be celestial objects such as stars or galaxies. The distribution and brightness of these objects suggest they could be part of a larger astronomical structure, possibly a galaxy cluster or a group of stars within a constellation.	The image depicts a radio astronomical observation, likely of a celestial object such as a galaxy or a star system. It shows two bright, circular regions that could represent the centers of these objects, surrounded by fainter, ring-like structures that might be indicative of surrounding material or disk-like formations. The central bright spots are possibly the cores of the objects, emitting intense radio signals. The faint lines and patterns suggest variations in intensity or different emission mechanisms within the observed region.
radio-llava (Q&A)	The image is a radio astronomical image cutout extracted from a larger radio-continuum Stokes-I map produced by an interferometer telescope, centred and zoomed on a multi-island extended radio source, consisting of 3 islands.	The image is a radio astronomical cutout featuring various point-like and compact radio sources, as well as extended radio sources, with some sources exhibiting poor imaging and ring-patterned artefacts.
radio-llava (Q&A + capt.)	The image shows a radio astronomical cutout of an extended radio galaxy candidate, characterised by its elongated structure and multiple bright spots.	The image shows a radio astronomical cutout with two prominent sources. One source appears as a bright, compact object near the center of the image, while the other is a more extended source located towards the right side. The background exhibits a complex pattern of concentric rings, likely resulting from the imaging process.
User	<p>## Question: Do you see any likely radio galaxy with an extended morphology in the image?</p> <p>## Context:</p> <p>– RADIO GALAXY: a type of active galaxy that emits an exceptionally large amount of radio waves, often extending beyond its visible structure. These galaxies host an active galactic nucleus (AGN), powered by a supermassive black hole (SMBH) at their center, which fuels the production of powerful radio jets and lobes.</p> <p>## Instructions:</p> <p>– Answer the question taking into account the provided context.</p>	
llava-ov	Yes, there is a likely radio galaxy with an extended morphology in the image.	Yes, there is a likely radio galaxy with an extended morphology in the image.
radio-llava (Q&A)	Yes, the image suggests the presence of an extended radio galaxy candidate due to its elongated and extended morphology.	Yes, the image contains candidate radio galaxies with extended morphology.
radio-llava (Q&A + capt.)	Yes, the image shows characteristics of a radio galaxy with an extended morphology.	No, the image does not show characteristics of a radio galaxy with an extended morphology.
User	<p>## Question: Do you see any imaging artifact in the provided image?</p> <p>## Context:</p> <p>– ARTEFACT: noise pattern with a ring-like or elongated morphology, typically found around bright compact sources in radio images. They are often mistaken for real radio sources but are actually spurious. These patterns arise from imperfections in the imaging processing stage of radio data.</p> <p>## Instructions:</p> <p>– Answer the question taking into account the provided context.</p>	
llava-ov	No	No
radio-llava (Q&A)	No, the image does not show any imaging artefacts.	Yes, the image contains imaging artefacts.
radio-llava (Q&A + capt.)	No, there are no imaging artefacts present in the image.	Yes, there are imaging artefacts present in the image.

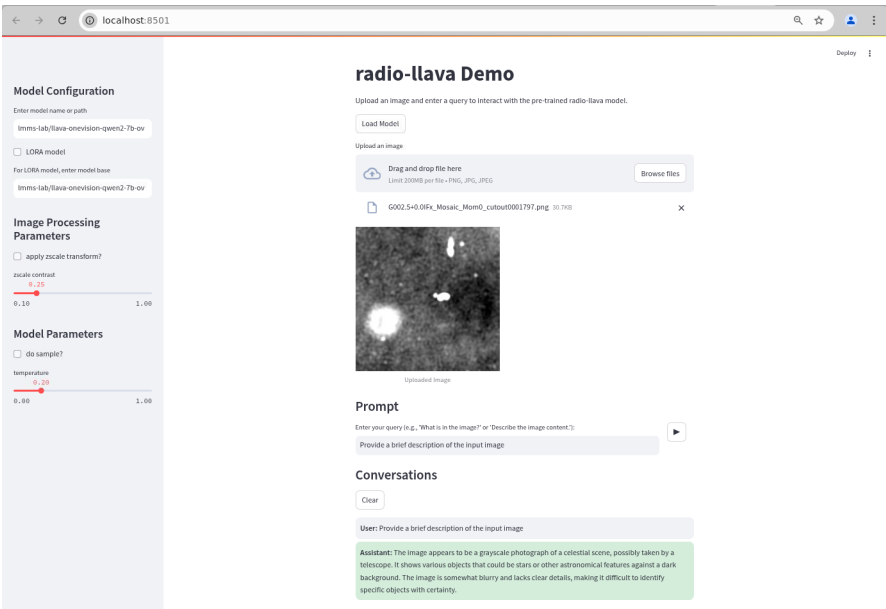


Figure C.1. A screenshot displaying the Streamlit web application developed for *radio-llava* demo purposes.

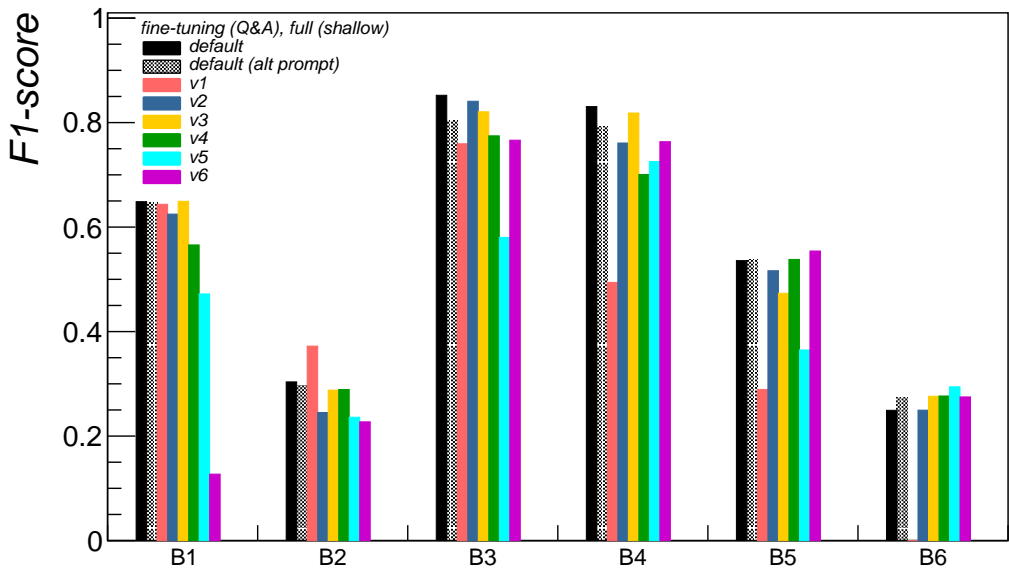


Figure C.2. Class-averaged classification F1-scores obtained with the *radio-llava* model on B1–B6 radio benchmarks, comparing fine-tuning on the Q&A training dataset (black solid histograms, labelled as "default") with model variants (v1–v6, coloured histograms), fine-tuned on the same dataset using alternative parameters (see text). The dashed black histogram represents the standard model evaluated on B1–B6 benchmarks using an alternative prompt.