UNIVERSITÉ DE GENÈVE

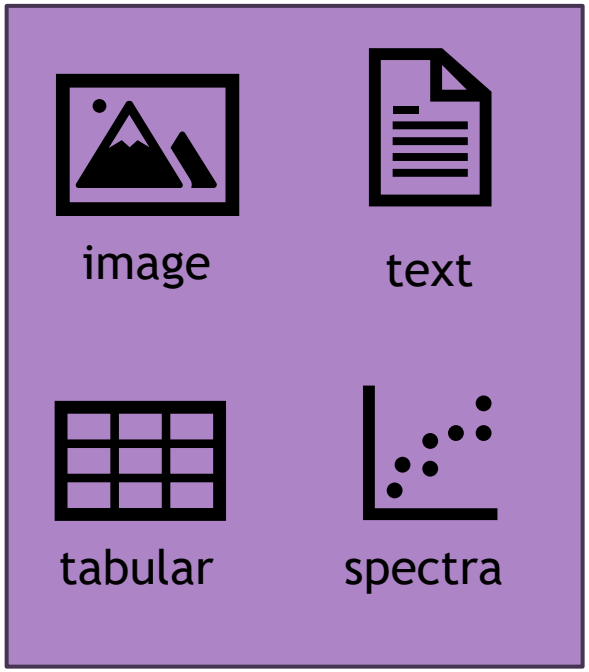
Aligning Astrophysical Images with Complementary Data Modalities

E. Lastufka

M. Dessauges-Zavadsky, M. Drozdova, T. Holotyak, V. Kinakh, D. Schaerer, S. Voloshynovskiy

SKACH

Multimodal Models for Astronomy



- Astrophysics data has many modalities
- Multimodal models (most commonly vision-language models) learn joint representations independent of the input modality
- Non-image data are useful as a grounding mechanism, helping the model to focus on scientifically relevant features

Multimodal Models for Astronomy – a recent history

2024 - Astro Mlab, Astro LLaMA (Pan et al): literature + metadata

2024 - AstroCLIP: optical images + spectra

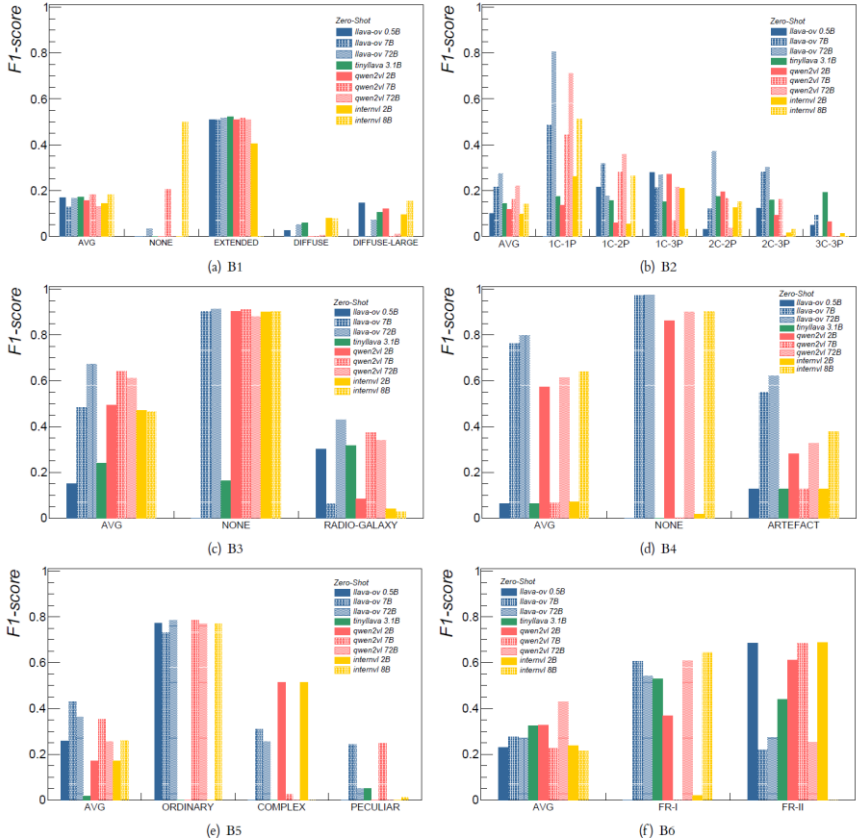
2024 – CosmoCLIP (Imam et al): images + generated captions

2025 – AstroLLaVA (Zaman et al): images + captions, interactive chat

2025 - AstroM3 (Rizkho & Bloom): photometry + spectra + metadata

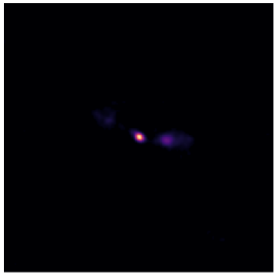
2025 - AION-1 (Parker et al): optical images + spectra + photometry + physical parameters

Vision-Language models and radio astronomy

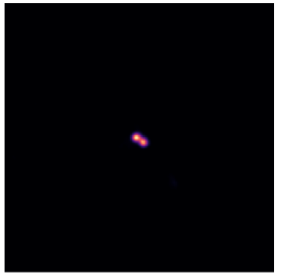


Riggi et al 2025: zero-shot classification generally poor
Fine-tuning leads to loss of generalization

Prompting Strategies

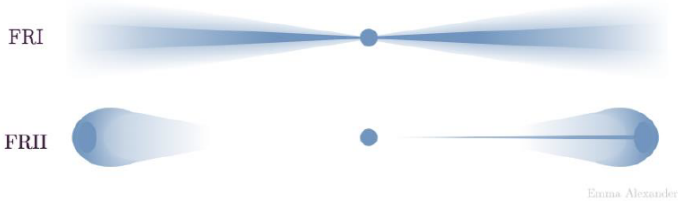


<image1>



<image2>

- **Zero-shot:**
  - **Text** – Natural language class descriptions
  - **Diagram** – Text + schematic class diagram
- **Exemplar-based (with retrieved examples):**
  - **Fixed-Imgs** – Same 4 labeled images for all test cases
  - **kNN-Imgs** – 5 nearest neighbors (CLIP-retrieved [5]), with labels
  - **kNN-Balanced** – 4 neighbors, balanced across classes (novel setup)



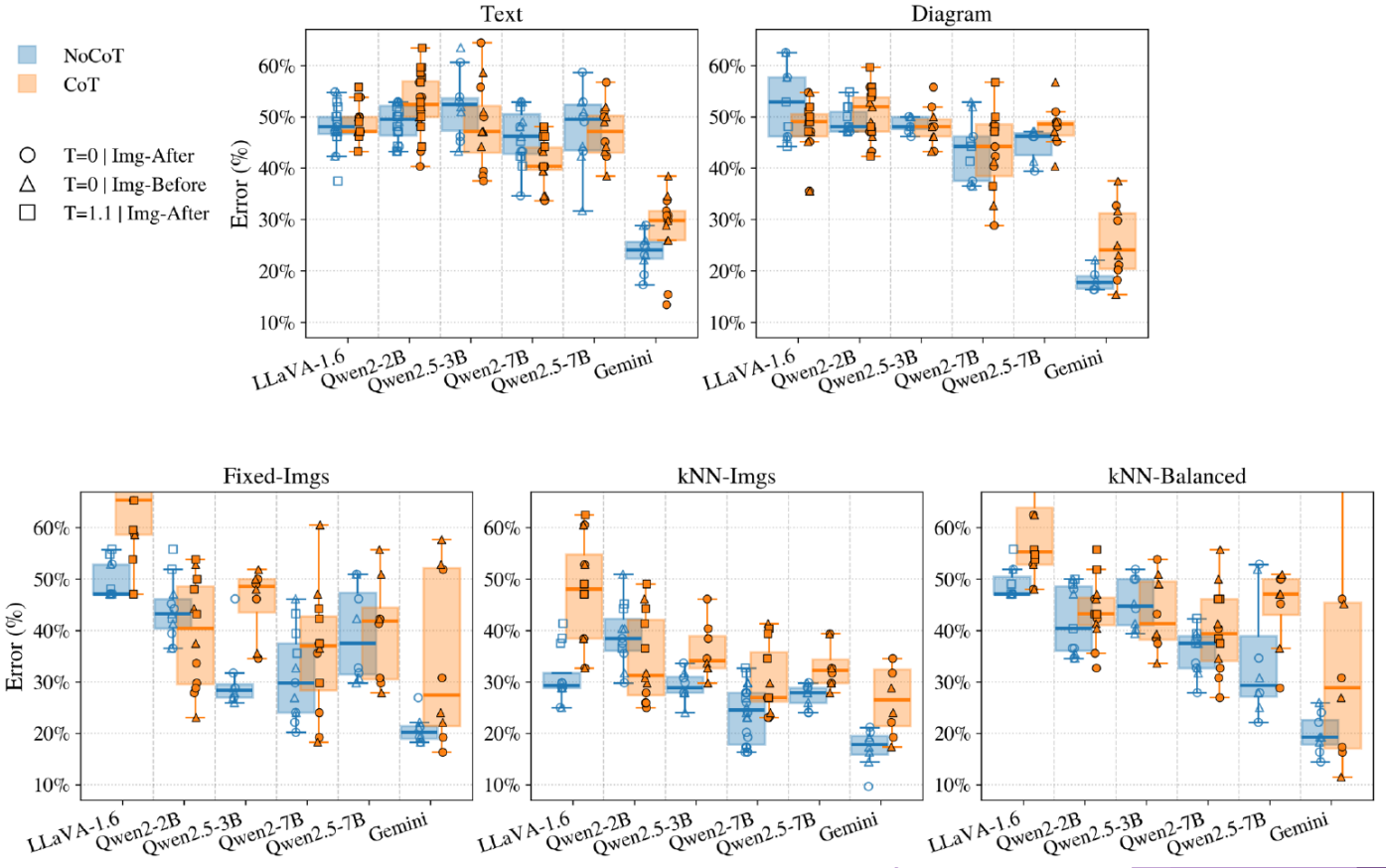
[4] Image by Emma Alexander, licensed under CC BY 4.0

Prompt:
<system>You are an astronomer tasked with classifying real images of radio galaxies.
There are two classes:
- FR-I: bright toward the center, fainter at the lobes' edges, often shows jets, etc.
- FR-II: edge-brightened, luminous lobes, bright hotspots at ends.

<user>Define a morphology class of this image. Respond only FR-I or FR-II. <image1>
<assistant>FR-I
<user> Define a morphology class of this image. Respond only FR-I or FR-II. <image2>
<assistant>FR-II
<user> Define a morphology class of this image. Respond only FR-I or FR II. <query image>
<assistant>

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021

Results: Prompting



Takeways

Full slides

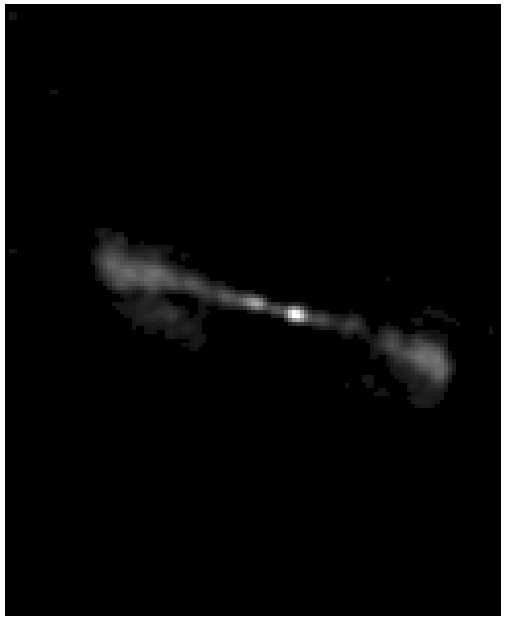
NeurIPS paper

- **Zero-shot** VLMs like Gemini can already perform well on scientific image tasks (error rate 13%)
- **Visual exemplars** can significantly boost performance in open models like Qwen2 (down to 16% for one of the prompts)
- **Balanced retrieval** reveals different behaviors—Gemini benefits most (error rate down to 11%)
- **Chain-of-thought** prompting can help, but often introduces instability (30% interquartile range for error rate)
- **LoRA fine-tuning** (just 15M params) brings Qwen2 close to domain-specific SOTA (SOTA 2% vs ours 3%)

Mariia.Drozdova@unige.ch

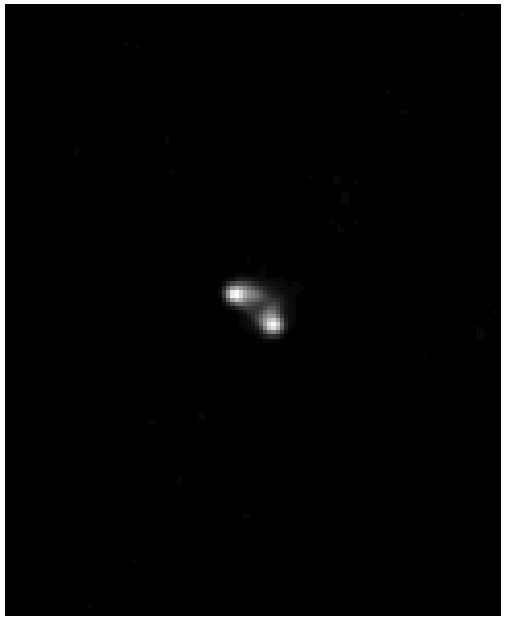
if VLMs can identify scientific features, will they benefit from scientific language?

Describing radio images via natural language



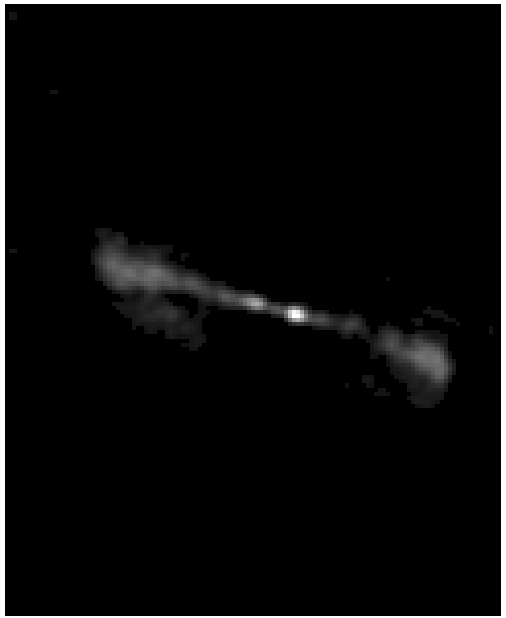
This radio image shows a powerful radio galaxy with a bright central active galactic nucleus (AGN). It emits two opposing jets of plasma that expand into large radio lobes, extending far into space.

Describing radio images via natural language



Three bright, blurry objects, seemingly connected, are visible against a dark background. This image could depict distant celestial bodies or an out-of-focus view of faint lights in space.

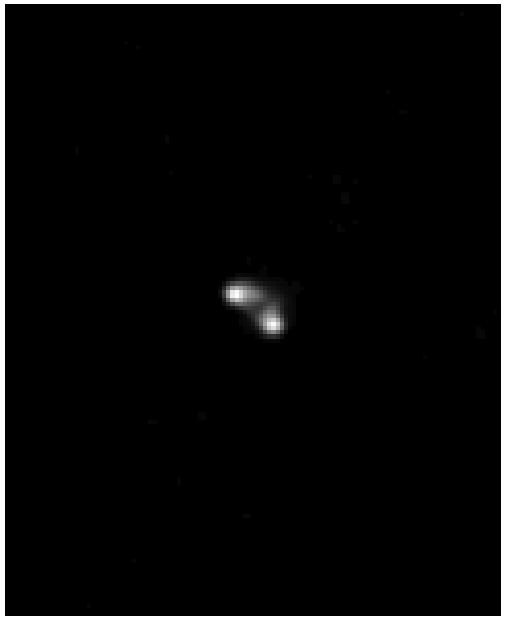
Describing radio images via natural language



This radio image shows a powerful radio galaxy with a bright central active galactic nucleus (AGN). It emits two opposing jets of plasma that expand into large radio lobes, extending far into space.

A bright, compact central source emits two opposing, collimated structures. These jets extend horizontally, broadening into diffuse, extended lobes terminating on the far left and right. No other distinct, unassociated sources are present.

Describing radio images via natural language

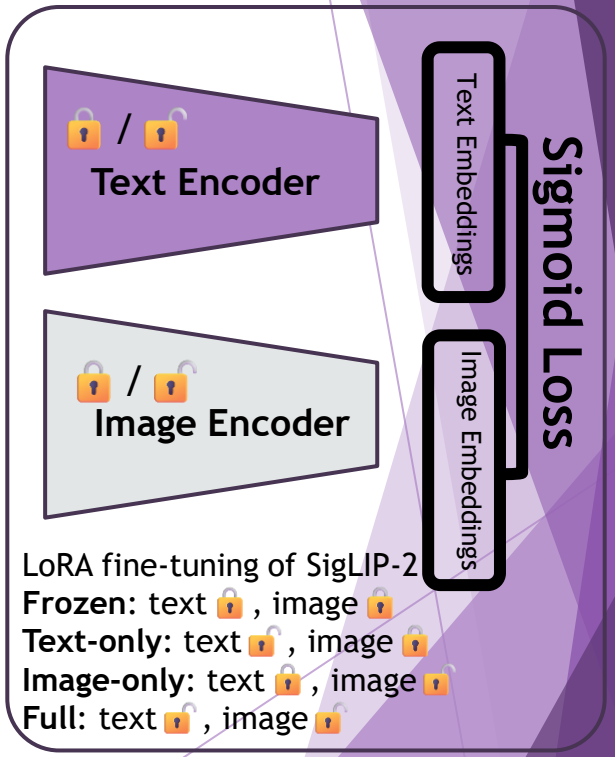


Three bright, blurry objects, seemingly connected, are visible against a dark background. This image could depict distant celestial bodies or an out-of-focus view of faint lights in space.

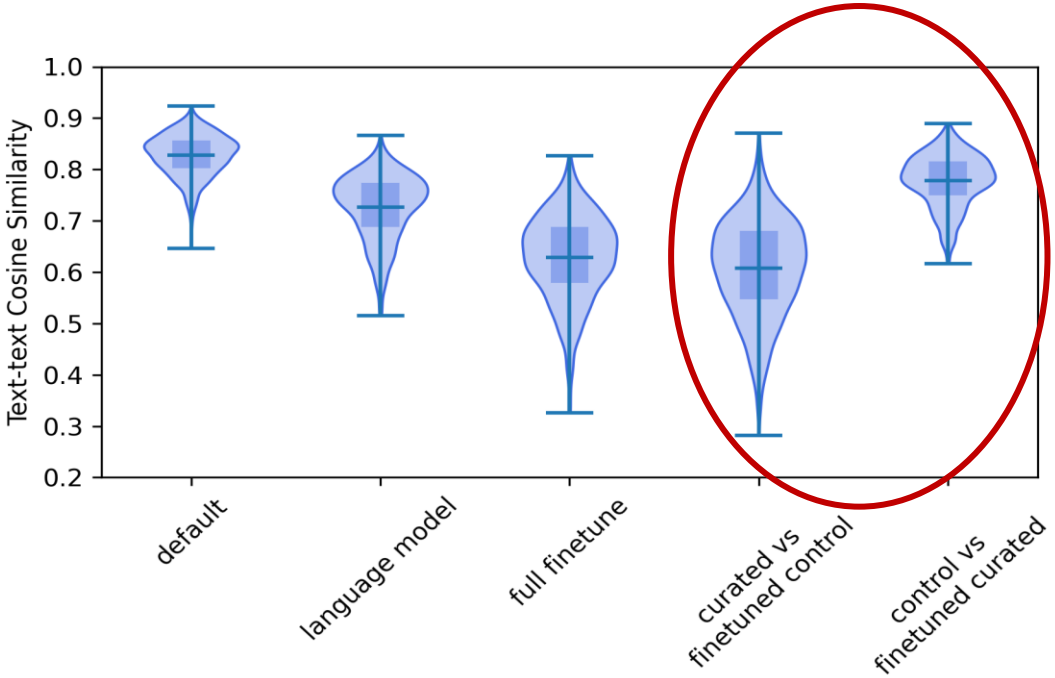
Two bright, closely spaced flux centroids are seen slightly top-right of center. They are surrounded by a fainter, diffuse, and irregular emission. Distinct lobes or collimated jets, with clear origins and ends, are not discernible from these centroids.

Experiment design

- Use SigLIP-2 VLM (sigmoid loss as opposed to contrastive loss)
- LoRA fine-tuning of image encoder, text encoder, or entire model
- FR-I/FR-II classification performance (higher F1 score is better) <- improves when fine-tuning text encoder or full model
- retrieval metrics – does a particular text match a particular image and/or its nearest neighbors? <- improves when fine-tuning image encoder
- Does cosine similarity (image/text alignment) increase or decrease? <- no significant change



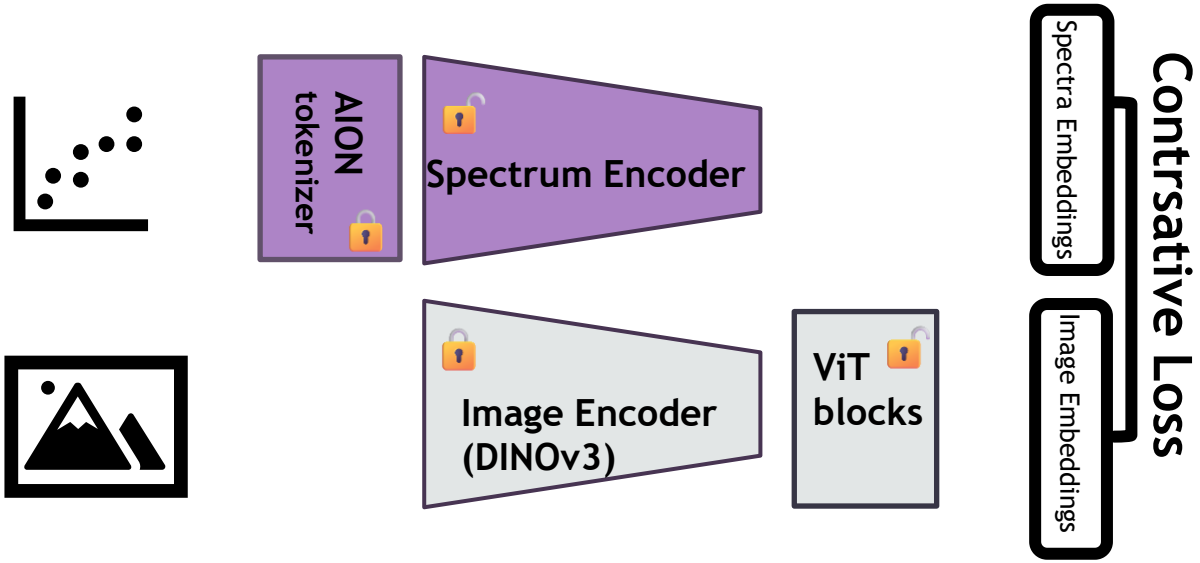
Cosine similarity



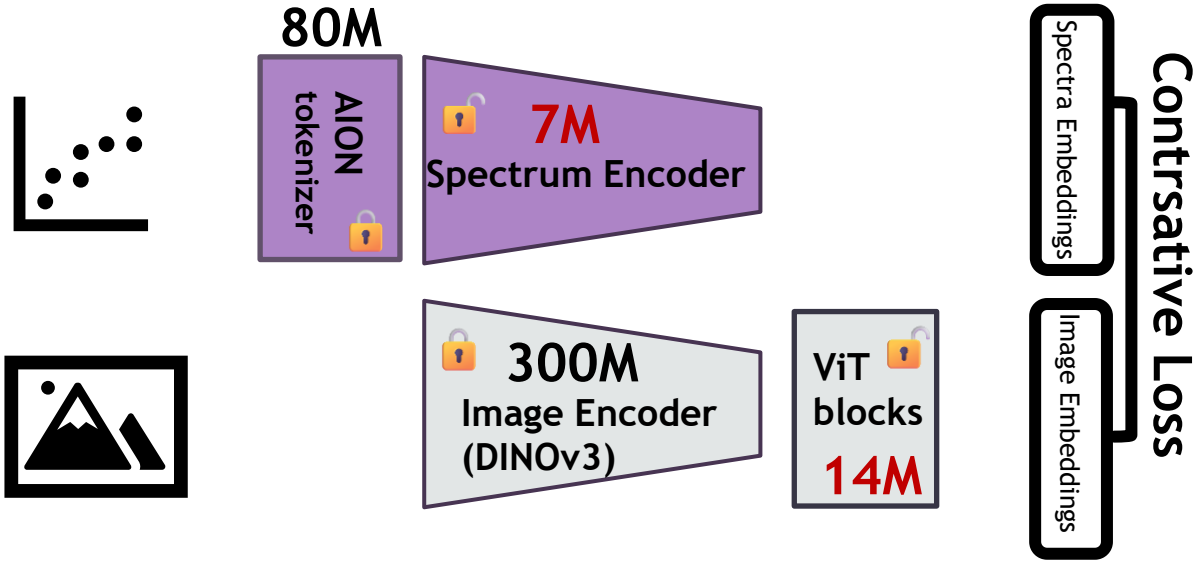
Lessons learned

- Curated captions with specialized vocabulary demonstrate the semantic control and structural coherence required to encode subtle image features
- However, the performance gain over control captions (prompted via “caption this image”) is very small!
  - Agrees with Drozdova 2025 zero-shot results, obtained using an empty prompt
  - Fine-tuning results in latent space that resembles the latent space distribution of control captions
- VLMs already encode useful priors for scientific domains

DINO.txt for images and spectra



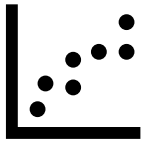
Align image and spectrum with small adapters



AION-base	300M
AION-large	800M
AION-Xlarge	3B

Pre-training data vs paired data

SDSS+DESI
100k
spectra



AION tokenizer

7M
Spectrum Encoder

HSC
86k *grz* images

Image Encoder (DINOv3)

ViT blocks
14M

Spectra Embeddings

Image Embeddings

Contrsative Loss

20k image - spectrum pairs

Preliminary results

Baseline Models

Model	GZ10 F1	Redshift prediction R2
DINOv3 large + LoRA	**0.757 ± 1.9e-2**	
DINOv3 large	0.722 ± 5.2e-3	
AION base	0.702 ± 1.2e-2	**0.910 ± 4.8e-3**
DINOv3 base	0.690 ± 1.1e-2	
AstroCLIP	0.462 ± 1.5e-2	

Experimental DINO.txt models

Pre-trained language model?	Frozen LM (first 50 epochs)?	Pre-trained vision blocks?	GZ10 F1	Redshift prediction R2
Y	Y	N + LoRA	**0.796 ± 5.8e-3**	0.879 ± 1.7e-2
Y	Y	N	0.779 ± 6.1e-3	0.882 ± 1.4e-2
N*	N	N	0.752 ± 5.7e-3	-
Y	N	N	0.747 ± 4.6e-3	0.884 ± 1.6e-2
N*	N	Y	0.731 ± 6.0e-3	-
Y	N	Y	0.724 ± 5.7e-3	**0.890 ± 3.0e-3**
N	N	Y	0.724 ± 5.6e-3	0.873 2.3e-3

*embedding layer only

Preliminary results

- Multimodal alignment possible with O(10M) parameters!
- Noticeable improvement in image task, no significant degradation in spectral task
- Strong spectral representations guide alignment; this is even more apparent when the spectral model is frozen at the start of joint training

summary

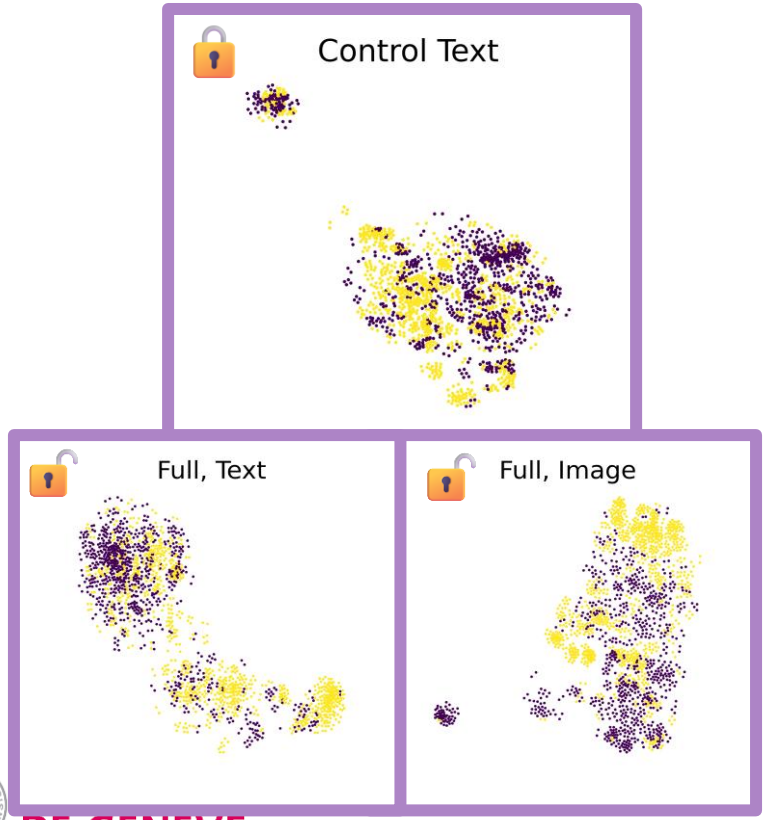
- Vision language models encode useful representations for radio astronomy images
- Generic natural language grounds concepts similarly to specialized language, especially after fine-tuning
- Joint representations between data modalities have great potential to encode scientific concepts that are independent of the input modality
- Joint representations can be efficiently learned using pre-trained encoders (even from different data distributions) and lightweight adapters!
  - Easy to extend to datasets from different instruments or entirely different modalities!

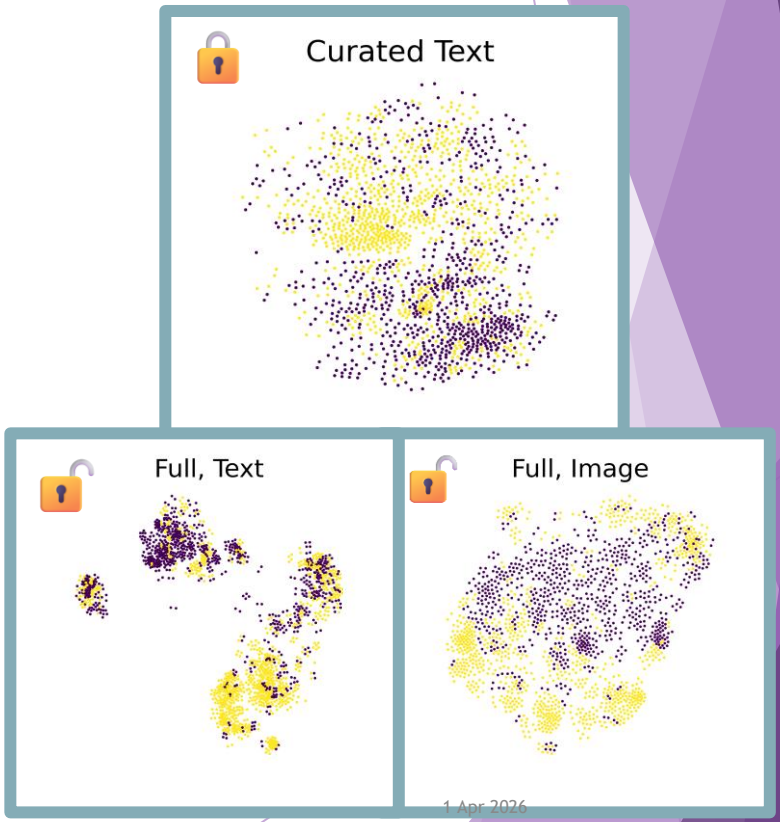
Classification F1

Linear Probe	Frozen	Text-only	Image-only	Full
Images	0.9 (-0.01)	0.93 (+0.03)	0.89 (-0.01)	**0.93 (+0.05)**
Text	0.9 (+0.01)	**0.92 (+0.03)**	0.89	**0.92 (+0.03)**
Images + Text	0.88 (+0.01)	0.88 (-0.01)	0.87	**0.88 (+0.02)**

Retrieval Metrics

	Frozen	Text-only	Image-only	Full
Recall@1	0.01 (-0.03)	0.03 (-0.08)	0.03	**0.07 (-0.01)**
Top-5 Recall	0.06 (-0.12)	0.13 (-0.23)	**0.23 (+0.04)**	0.23 (-0.18)
Class-level Recall	0.53 (-0.05)	0.66 (-0.12)	0.69 (+0.11)	**0.77**


Control Text
Full, Text
Full, Image


Curated Text
Full, Text
Full, Image