UNIVERSITÉ DE GENÈVE

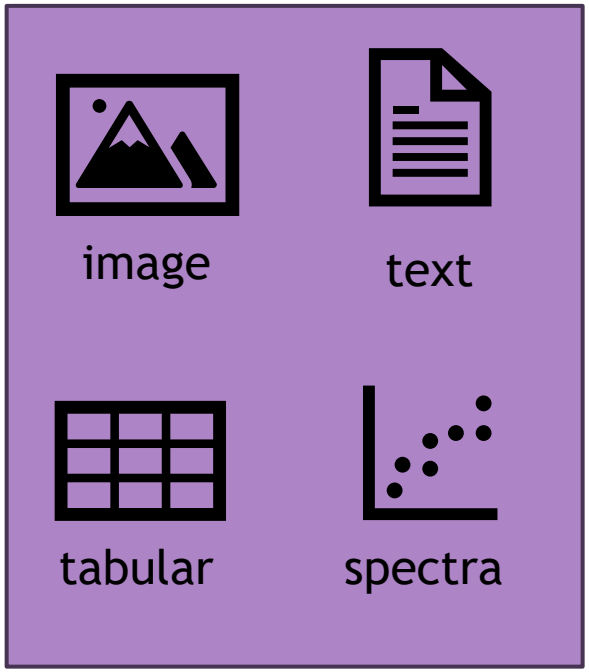
Aligning Astrophysical Images with Complementary Data Modalities

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

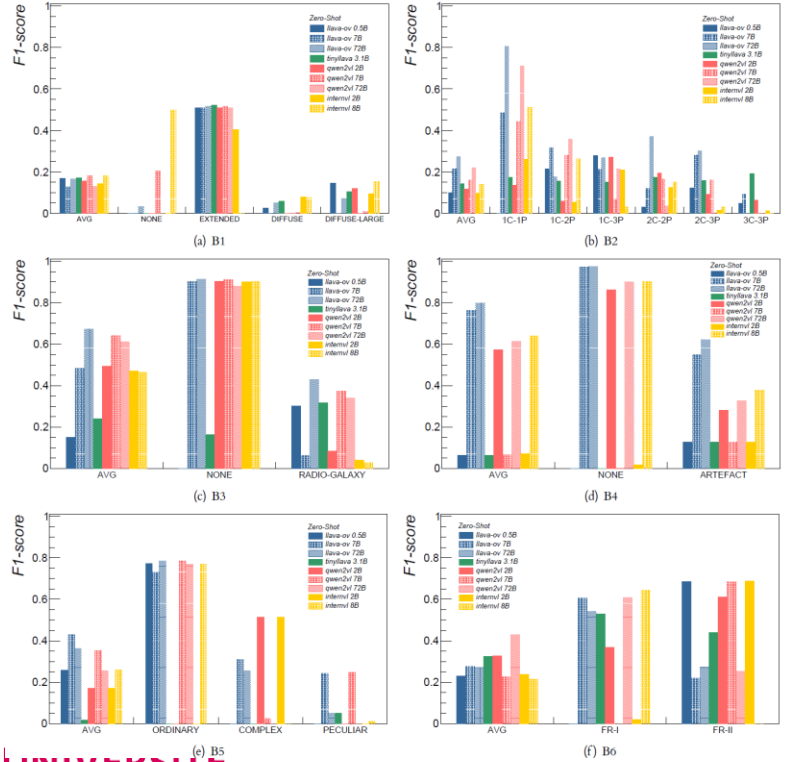
2024 – CosmoCLIP (Imam et al): images + generated captions

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

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

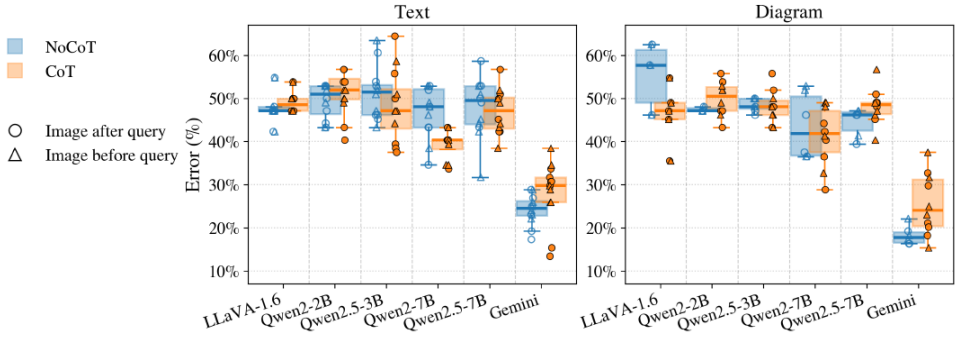
2025 - AION-1 (Parker et al): optical images + spectra + photometry + physical parameters

Vision-Language models and radio astronomy



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

Vision-Language models and radio astronomy

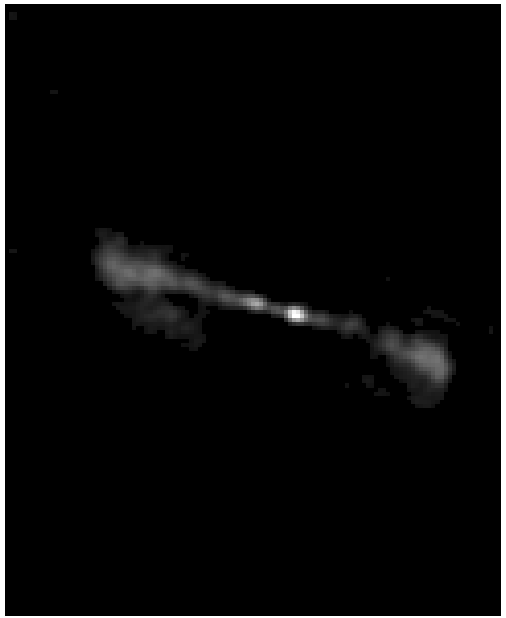


Drozdova et al 2025: zero-shot classification much better

Prompting strategy, such as use of visual in-context examples, is important

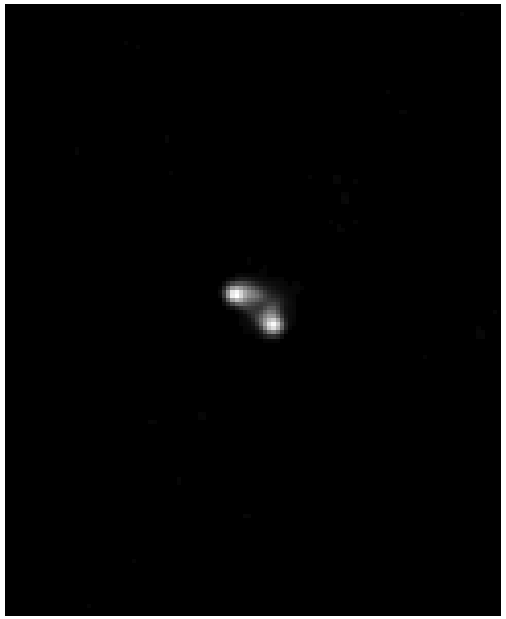
VLM outputs can be greatly improved but are unstable

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.

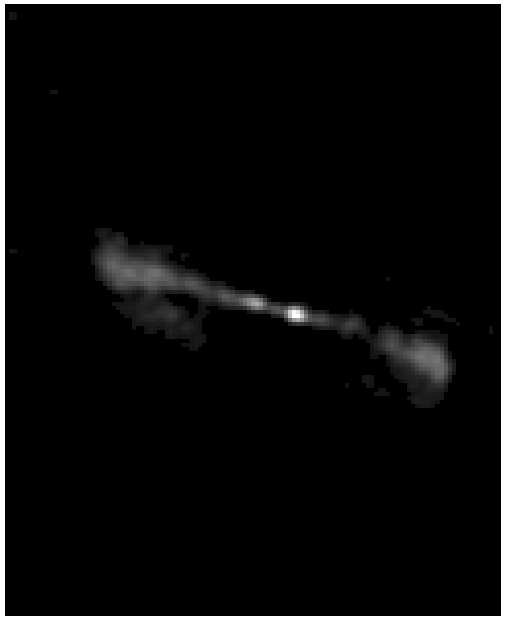
Describing radio images via natural language



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.

UNIVERSITÉ DE GENÈVE

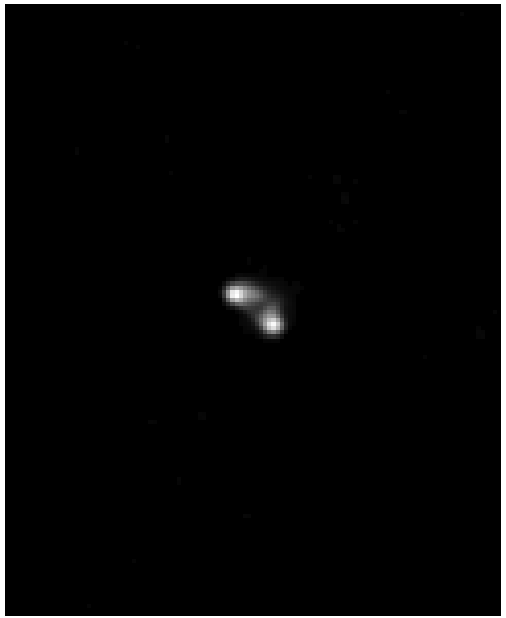
Describing radio images via natural language



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



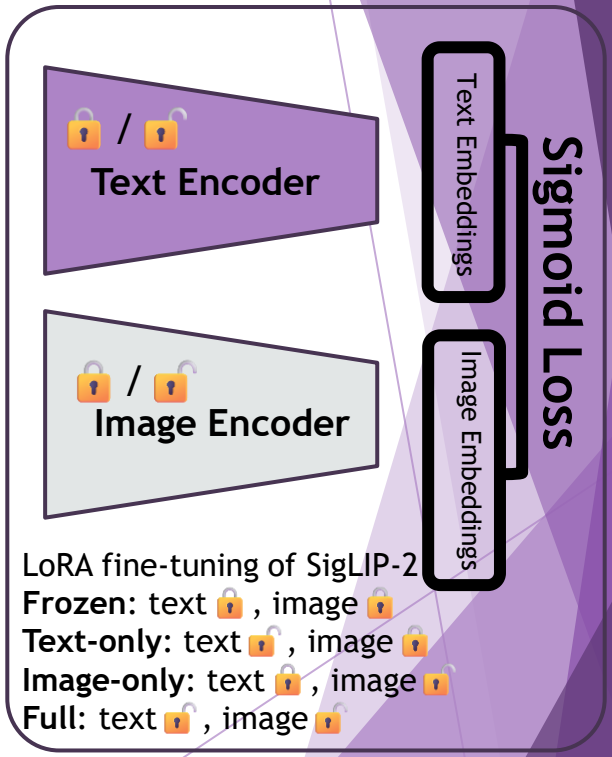
Three bright, blurry objects, seemingly connected, are visible against a dark background. This image could depict distant celestial bodies or an out-of-focus view of faint lights in space.

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.

Does image-text alignment benefit from scientific language?

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)
- retrieval metrics – does a particular text match a particular image and/or its nearest neighbors?
- Does cosine similarity (image/text alignment) increase or decrease?
- Examine changes in latent space distribution

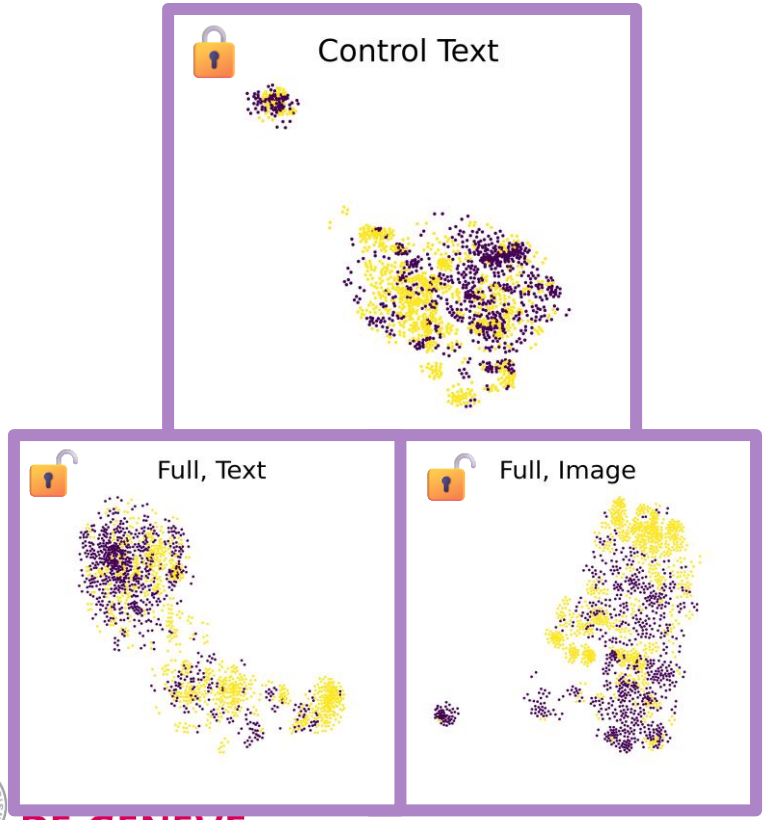


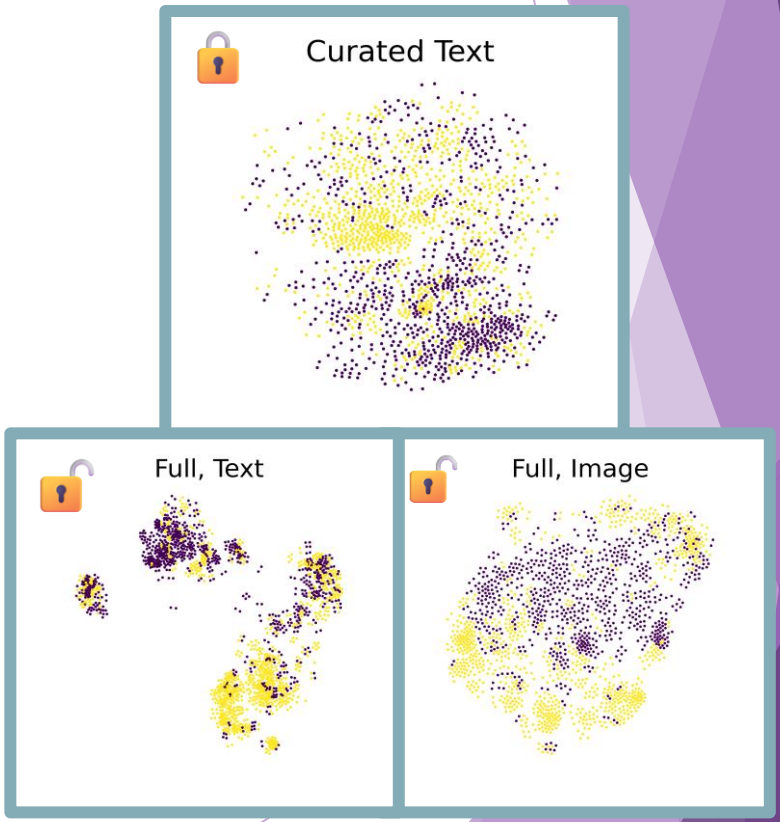
Classification F1

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)**

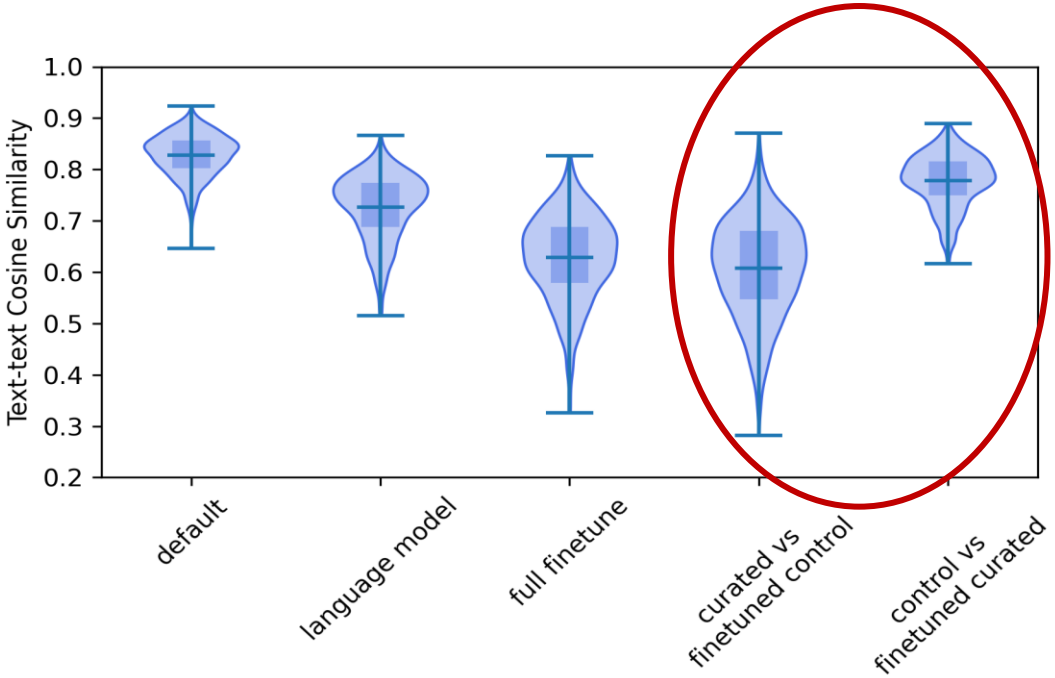
Retrieval Metrics

	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

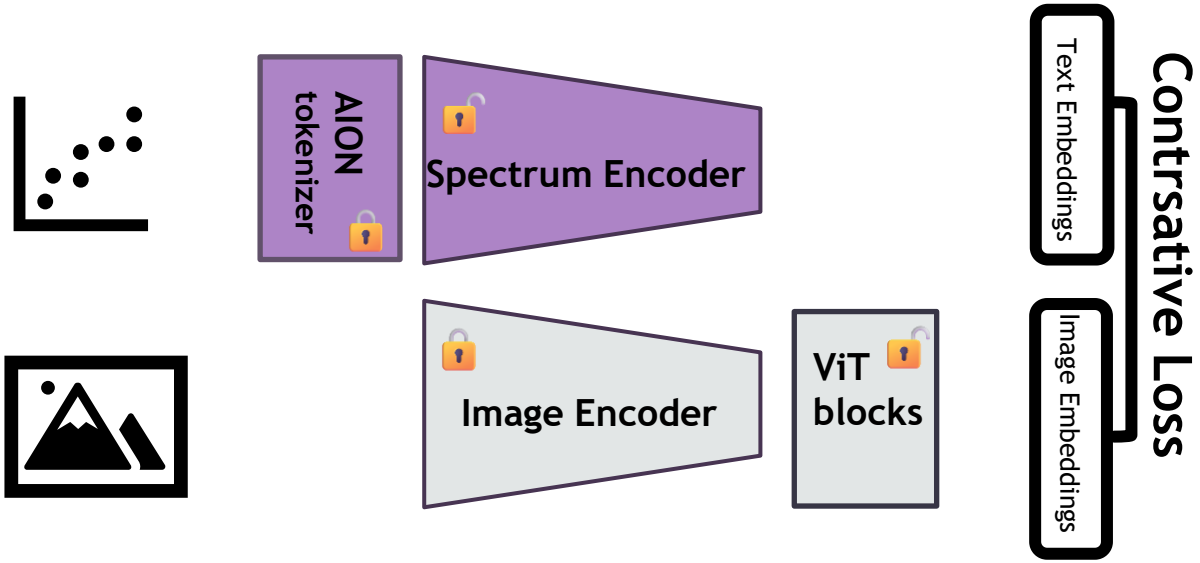
Cosine similarity



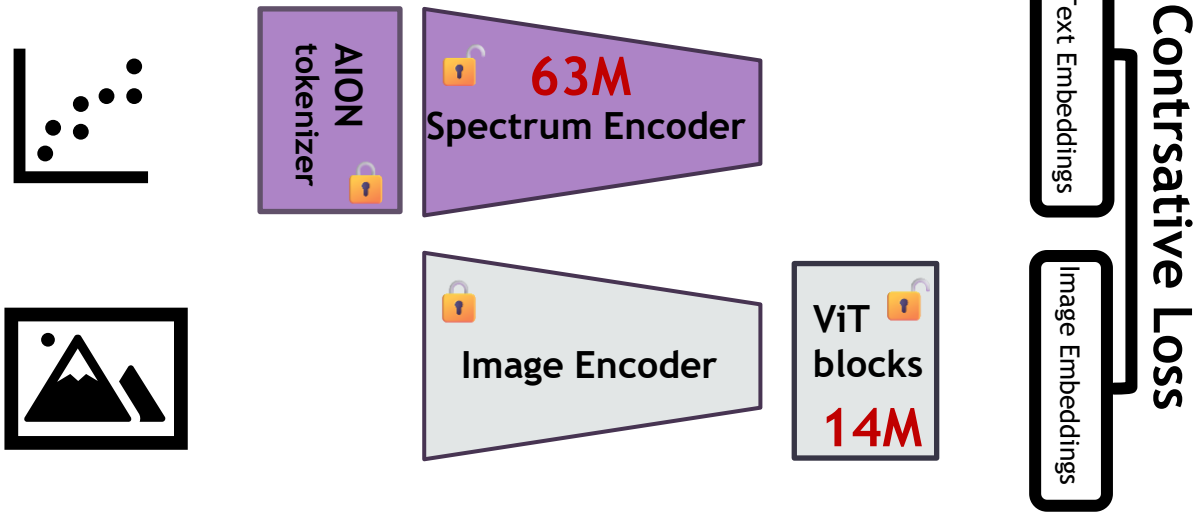
Lessons learned

- 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



DINO.txt for images and spectra



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

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
  - How to learn good joint representations without large amounts of pre-training data?

paper & code

https://github.com/elastufka/mirabest_captioned