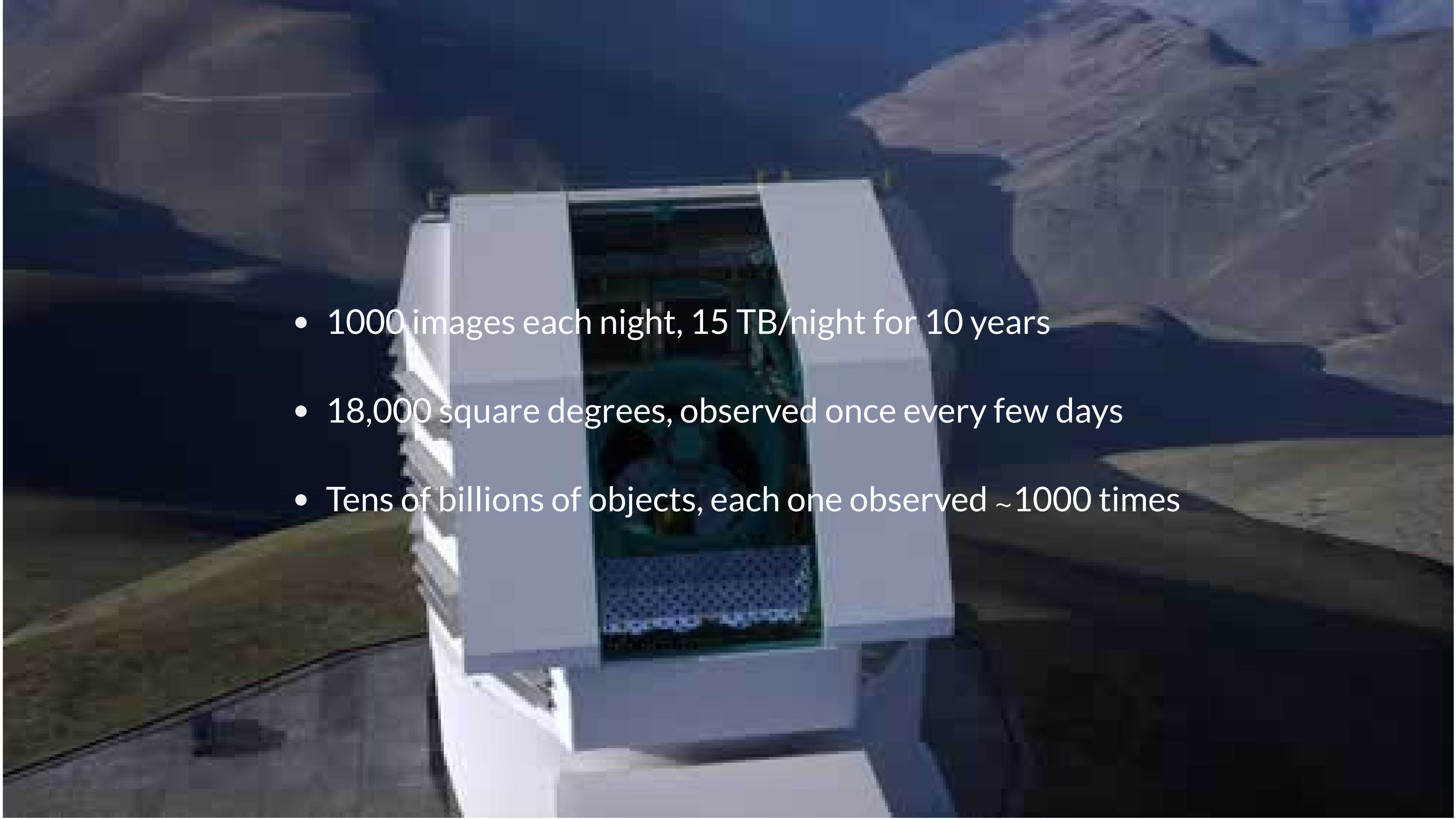
Multimodal Pretraining for Scientific Data

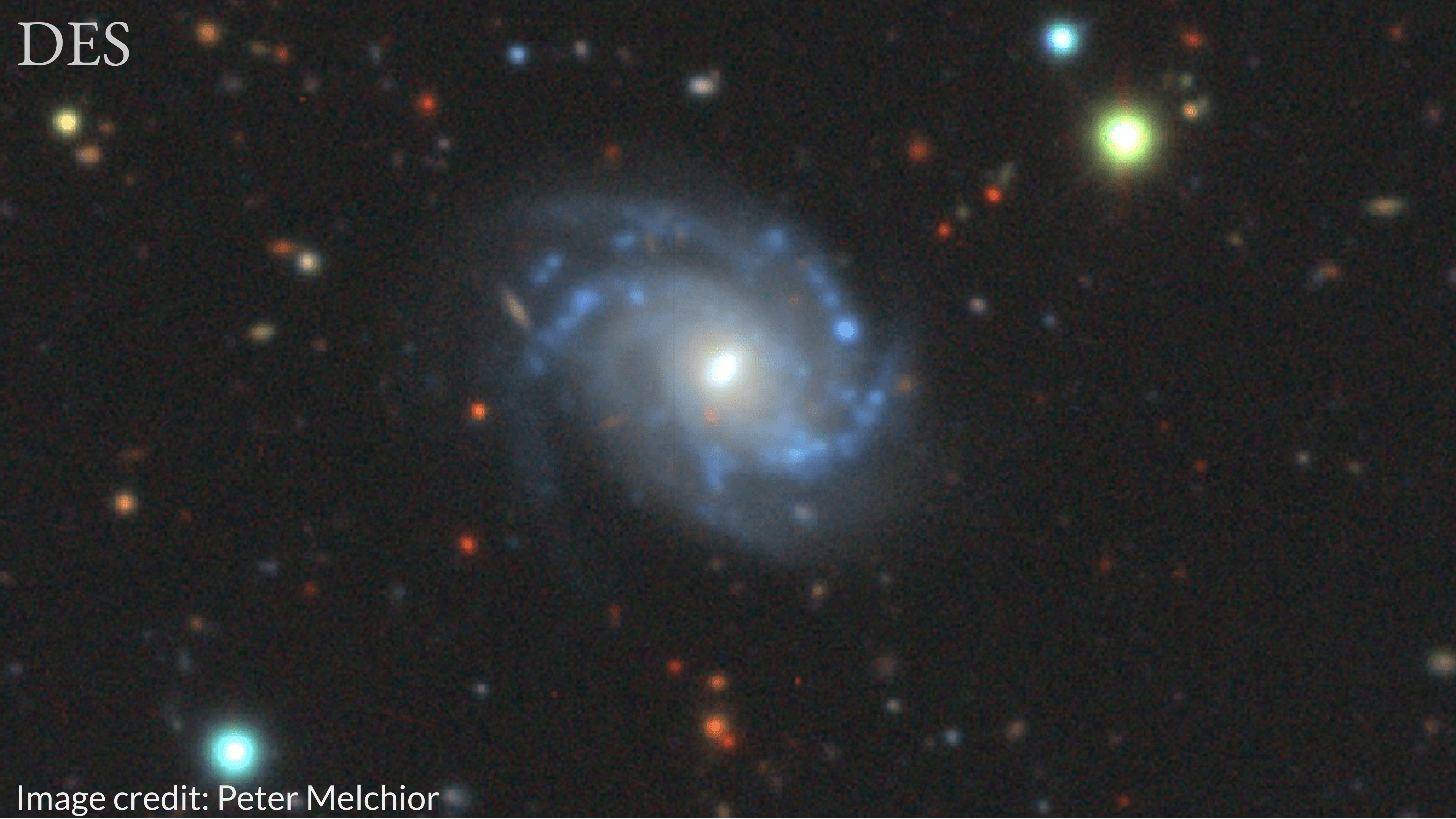
Towards Large Data Models for Astrophysics

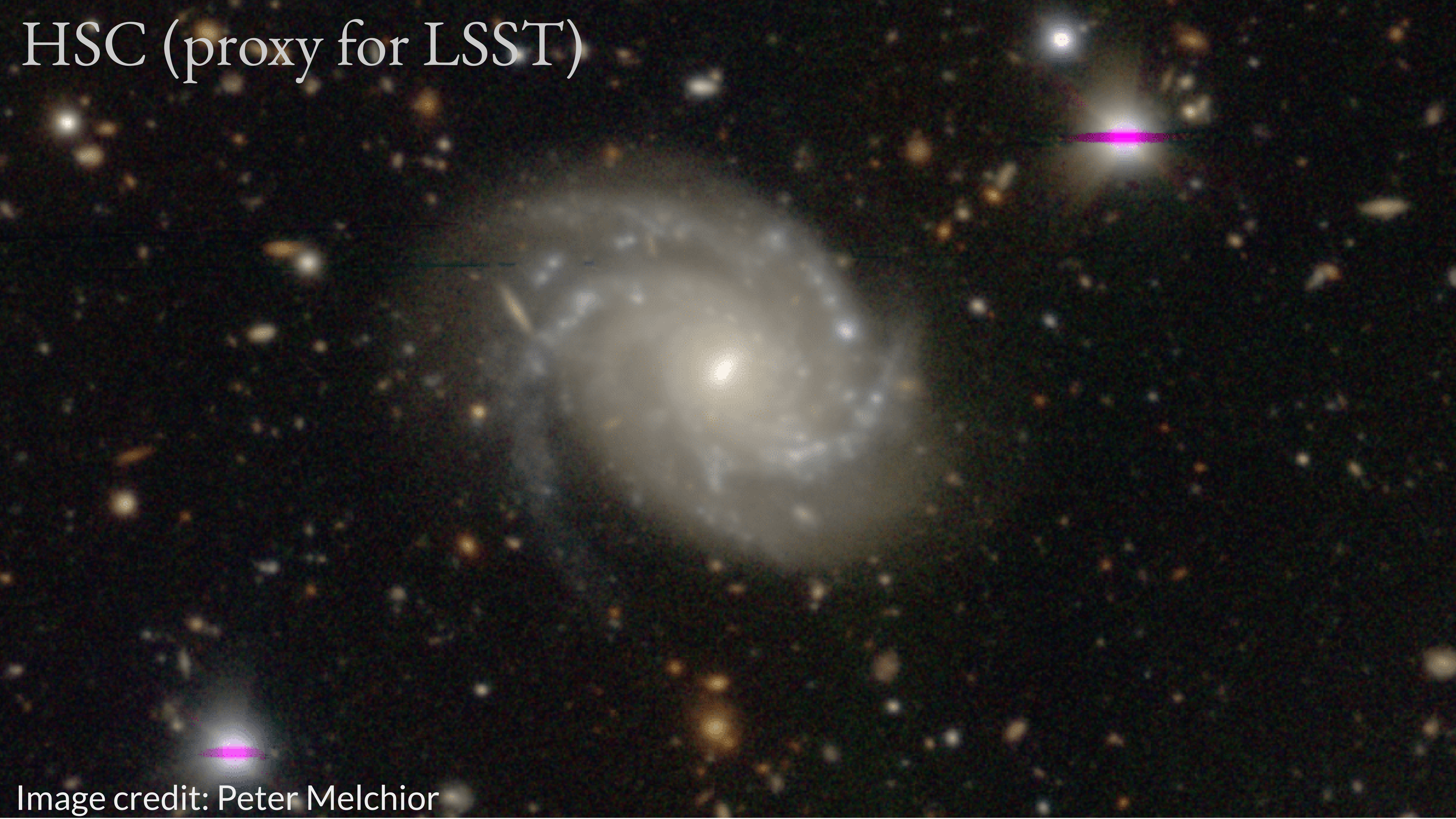
Francois Lanusse



- 1000 images each night, 15 TB/night for 10 years
- 18,000 square degrees, observed once every few days
- Tens of billions of objects, each one observed ~1000 times


SDSS
Image credit: Peter Melchior

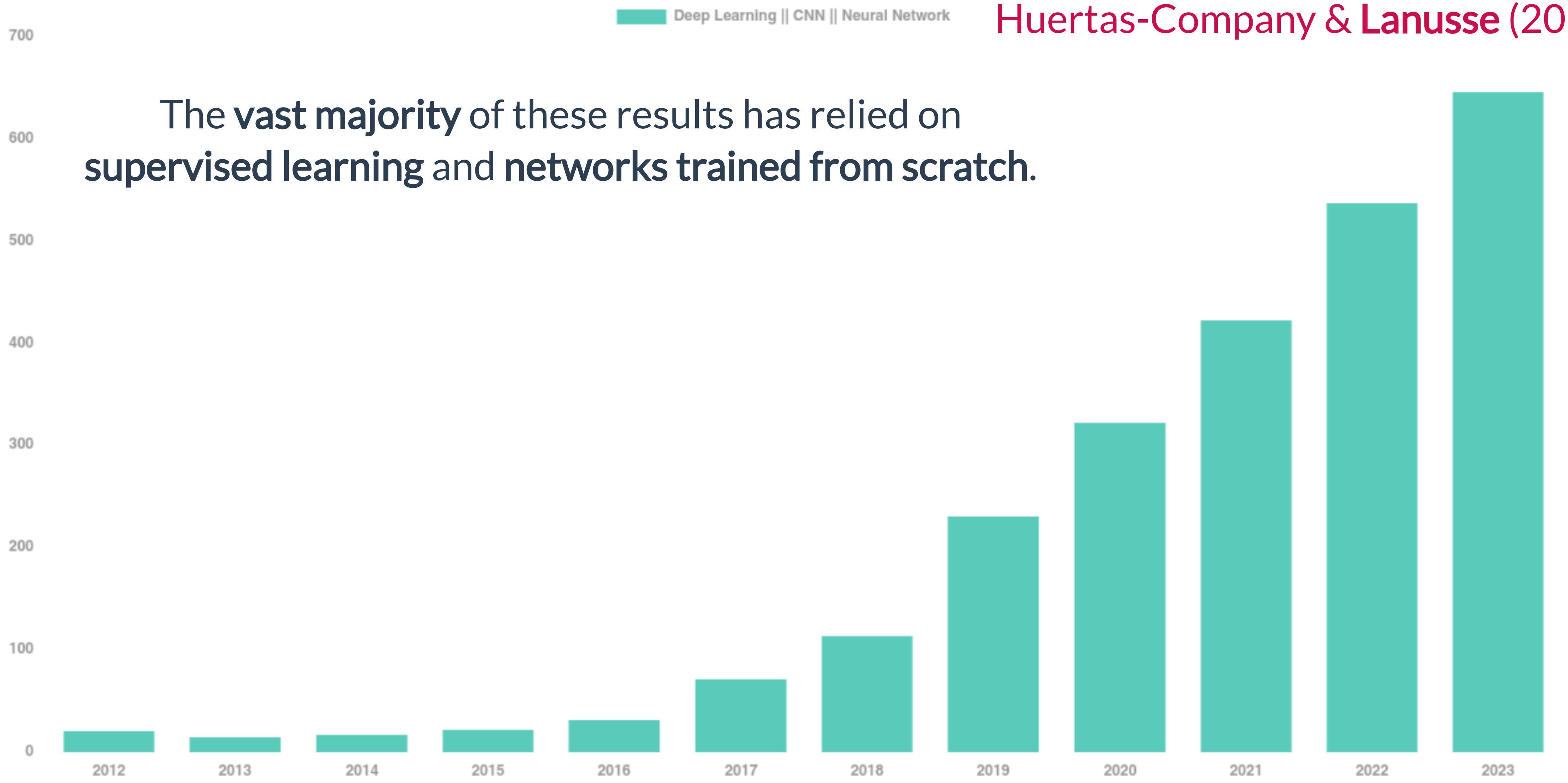

DES
Image credit: Peter Melchior


HSC (proxy for LSST)
Image credit: Peter Melchior

The Deep Learning Boom in Astrophysics

Huertas-Company & **Lanusse** (2023)

The **vast majority** of these results has relied on **supervised learning** and **networks trained from scratch**.



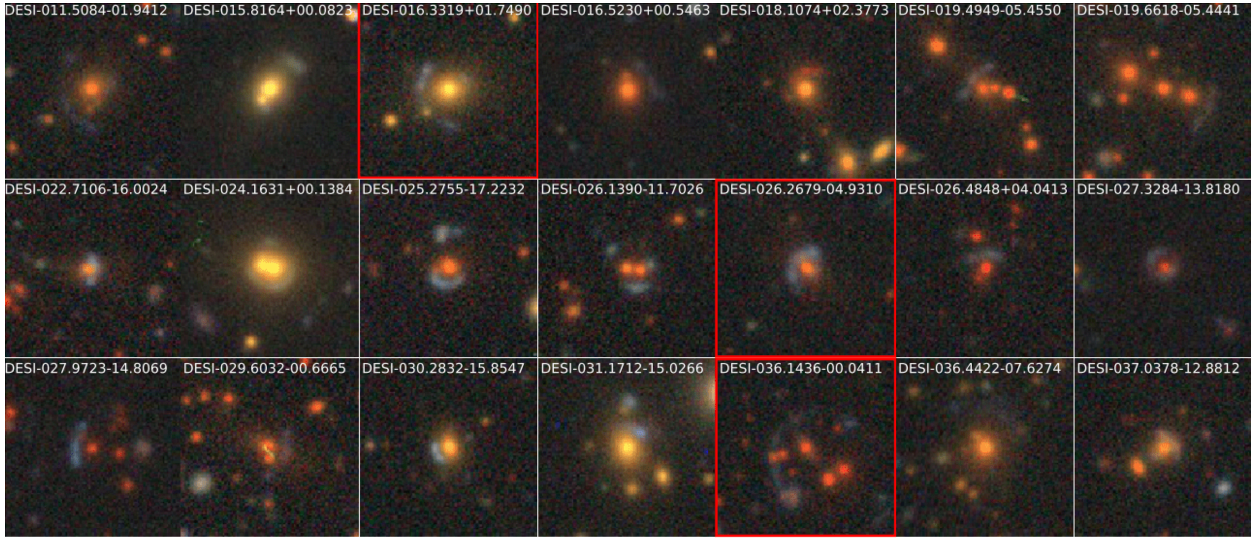
astro-ph abstracts mentioning **Deep Learning, CNN,** or **Neural Networks**

The Limits of Traditional Deep Learning

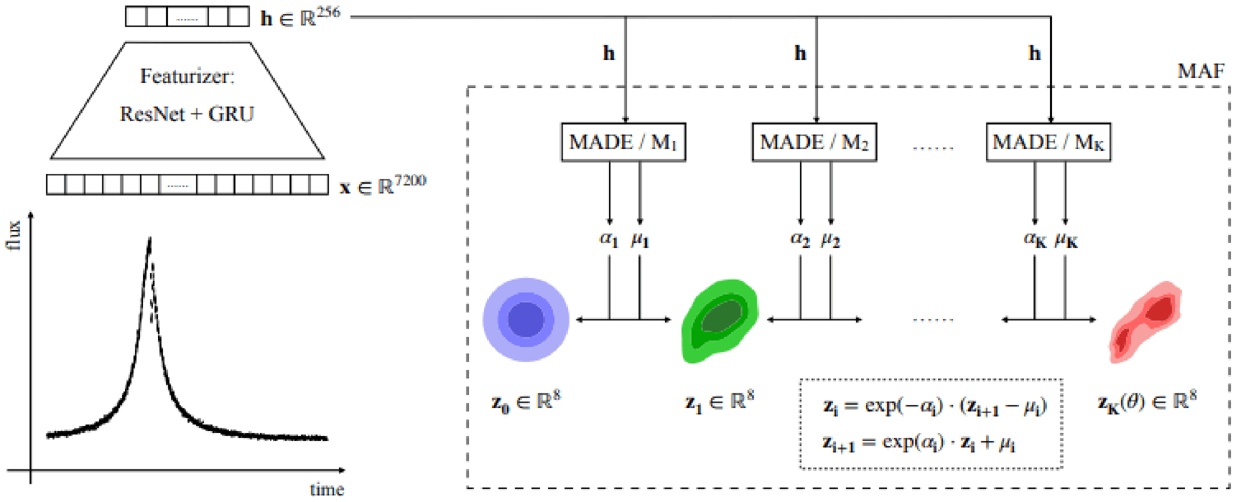
- **Limited Supervised Training Data**
  - Rare or novel objects have by definition few labeled examples
  - In Simulation Based Inference (SBI), training a neural compression model requires many simulations

- **Limited Reusability**
  - Existing models are trained supervised on a specific task, and specific data.

=> Limits in practice the ease of using deep learning for analysis and discovery



Huang et al. (2019)

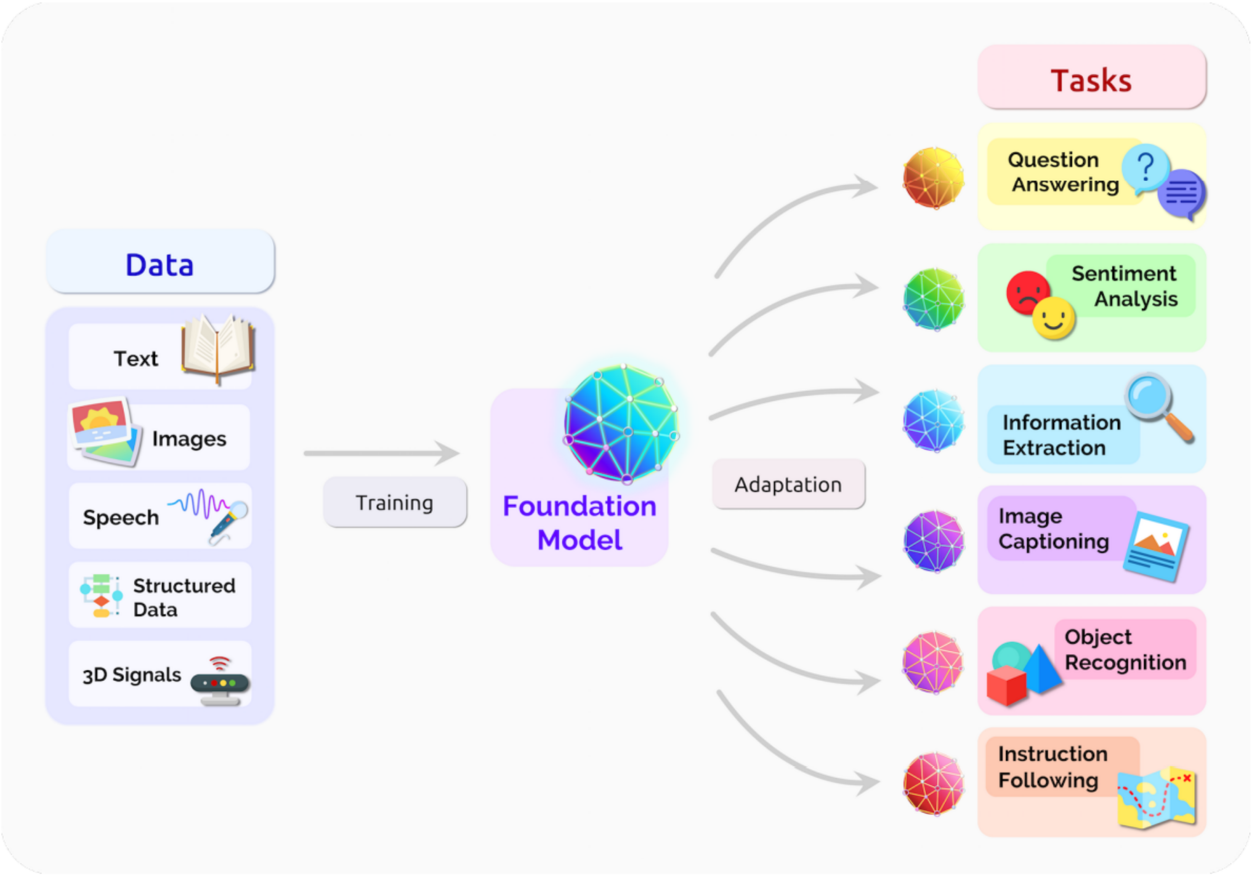


Zhang, Bloom, Gaudi, **Lanusse**, Lam, Lu (2021)

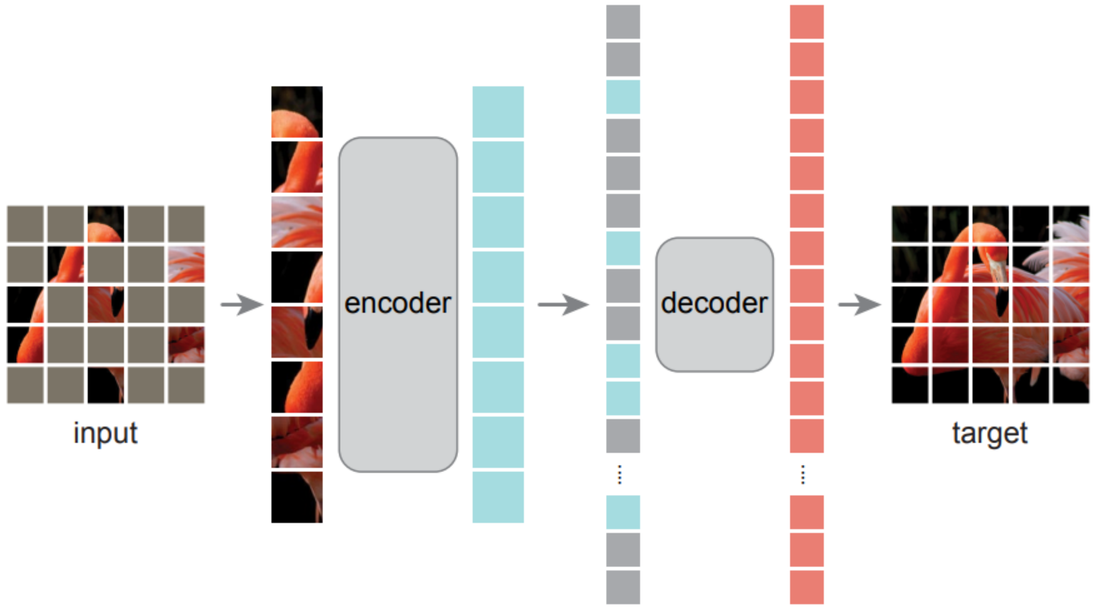
Meanwhile, in Computer Science…

The Rise of The Foundation Model Paradigm

- **Foundation Model approach**
  - **Pretrain** models on pretext tasks, without supervision, on very large scale datasets.
  - **Adapt** pretrained models to downstream tasks.
  - **Combine** pretrained modules in more complex systems.

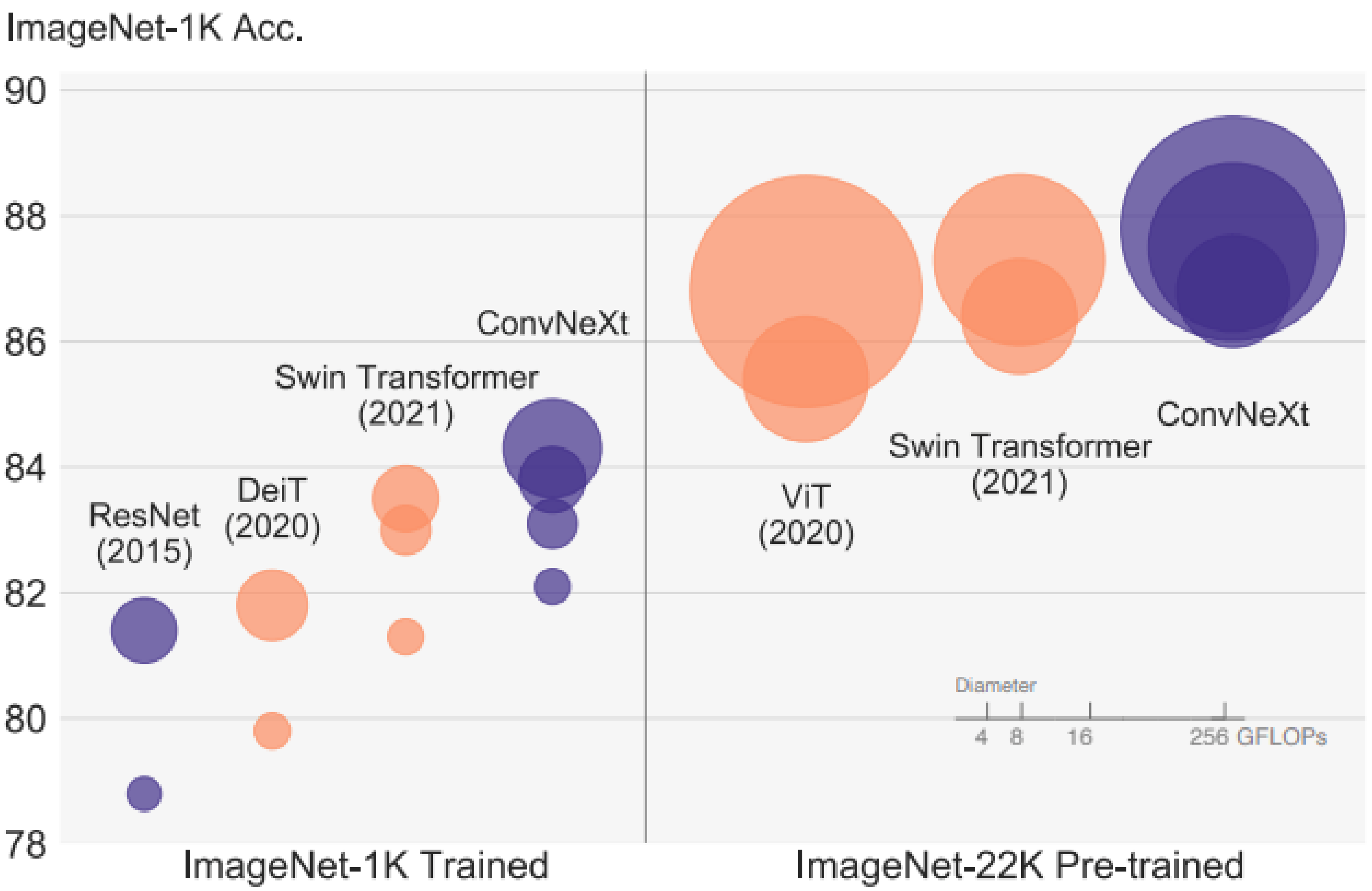


Bommasani et al. 2021



He et al. 2021

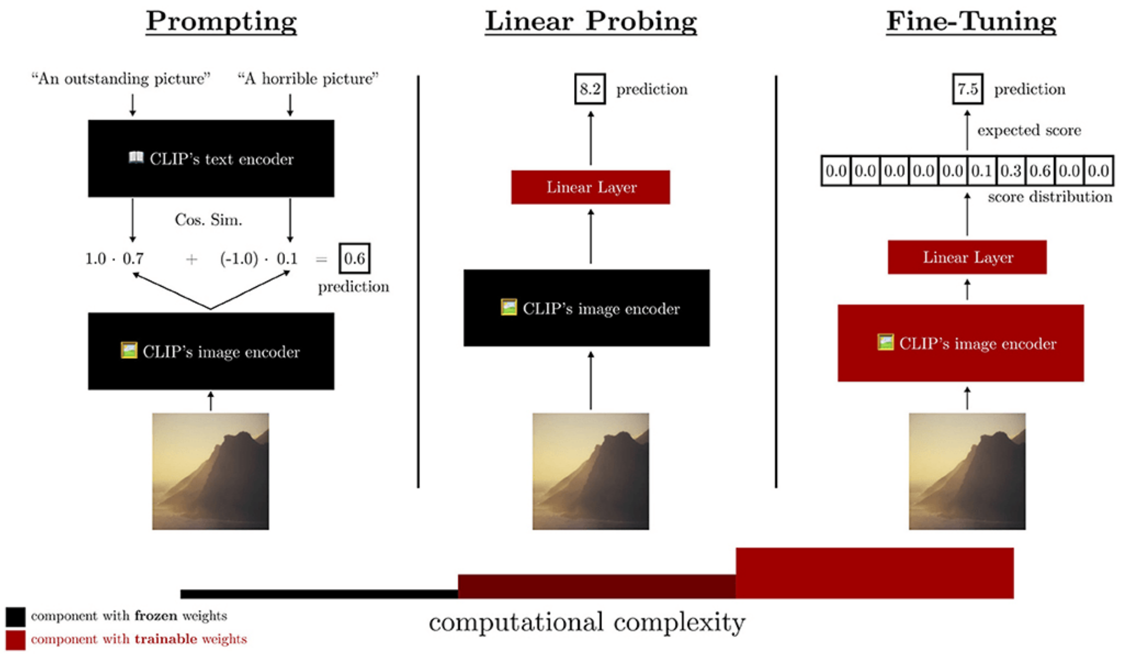
The Advantage of Scale of Data and Compute



Liu et al. 2022

Linearly Accessible Information

- Backbone of modern architectures embed input images as vectors in $\mathbb{R}^d$ where d can typically be between 512 to 2048.
- **Linear probing** refers to training a single matrix to adapt this vector representation to the desired downstream task.



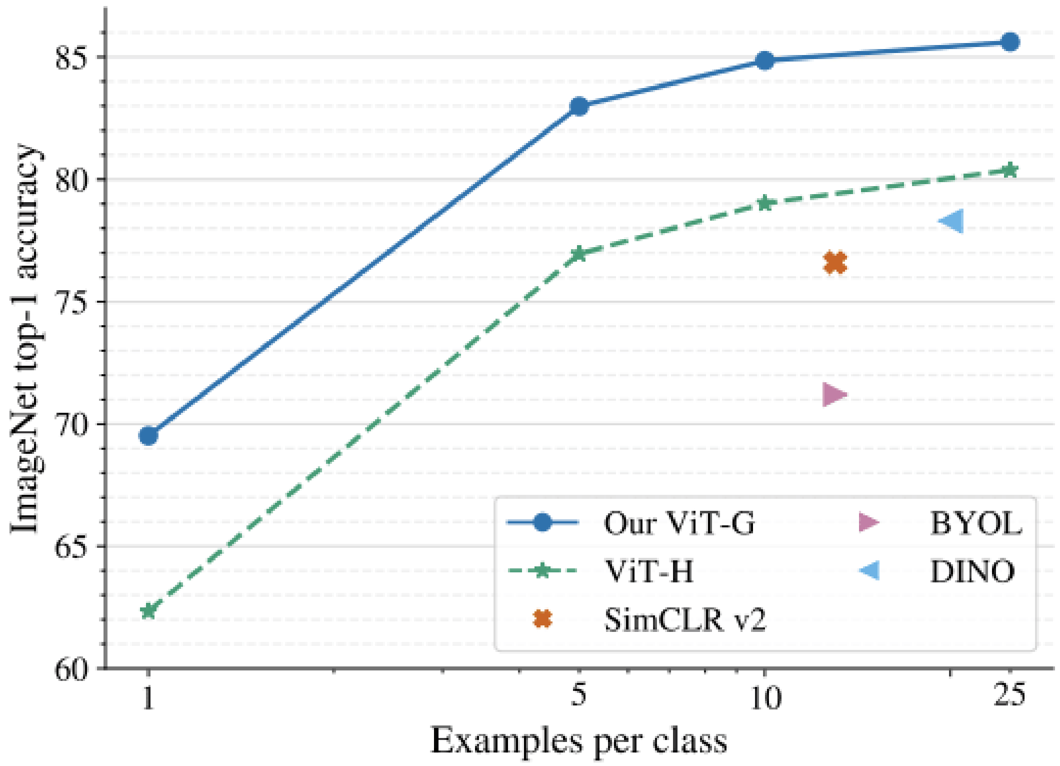


Figure 1. Few-shot transfer results. Our ViT-G model reaches 84.86% top-1 accuracy on ImageNet with 10-shot linear evaluation.

Zhai et al. 2022

What This New Paradigm Could Mean for Us Astrophysicists

- **Never have to retrain my own neural networks** from scratch
  - Existing pre-trained models would already be near optimal, no matter the task at hand
- Practical large scale Deep Learning even in **very few example regime**
  - Searching for very rare objects in large surveys like Euclid or LSST becomes possible
- If the information is embedded in a space where it becomes linearly accessible, **very simple analysis tools are enough** for downstream analysis
  - In the future, survey pipelines may add vector embedding of detected objects into catalogs, these would be enough for most tasks, without the need to go back to pixels

Can we translate these innovations into a similar **paradigm shift in deep learning for scientific applications?**

Polymathic


Alberto Bietti


Kyunghyun Cho


Miles Cranmer


Michael Eickenberg


Siavash Golkar


Keiya Hirashima


Shirley Ho


Geraud Krawezik


Francois Lanusse


Nick Lourie


Michael McCabe


Ruben Ohana


Liam Parker


Mariel Pettee


Bruno Regaldo


Leopoldo Sarra


Rudy Model

SIMONS FOUNDATION

NYU

UNIVERSITY OF CAMBRIDGE

SCHMIDT FUTURES

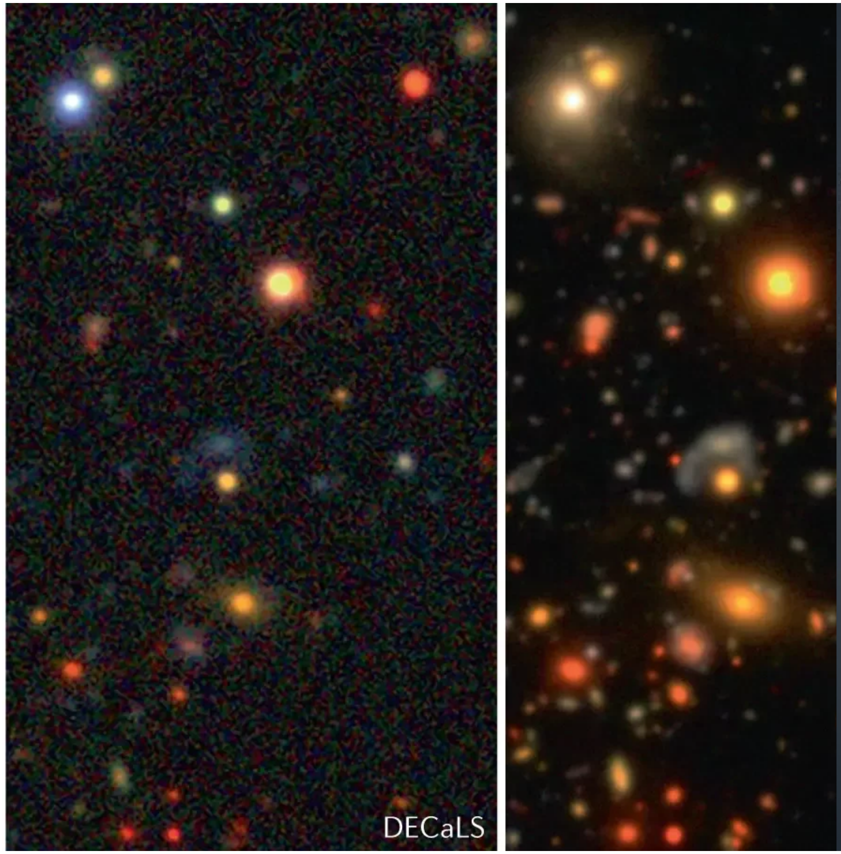
PRINCETON UNIVERSITY

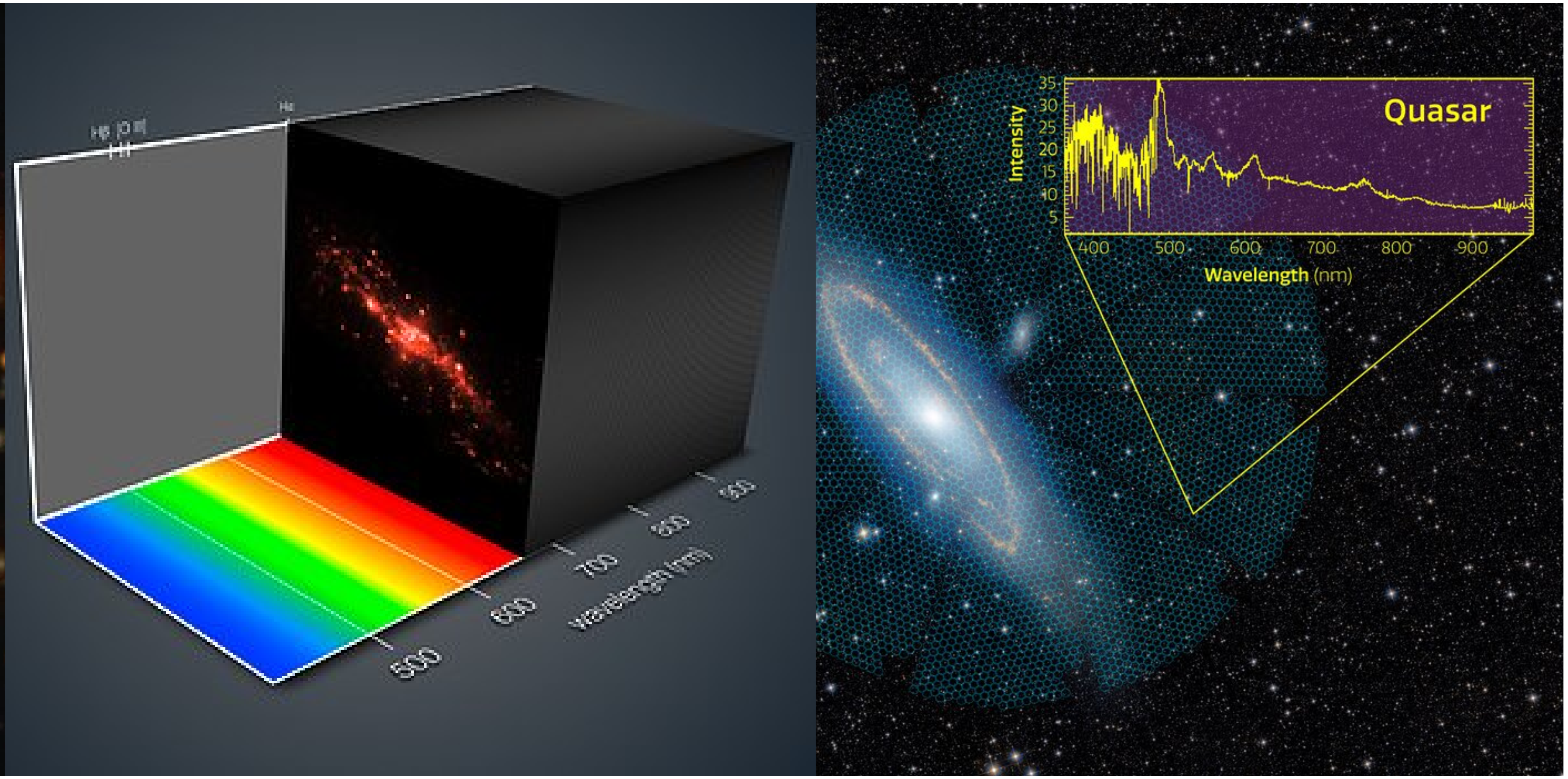
BERKELEY LAB

cnrs

dépasser les frontières

The Data Diversity Challenge





Credit: Melchior et al. 2021

Credit:DESI collaboration/DESI Legacy Imaging Surveys/LBNL/DOE & KPNO/CTIO/NOIRLab/NSF/AURA/unWISE

- Success of recent foundation models is driven by large corpora of uniform data (e.g LAION 5B).
- Scientific data comes with many additional challenges:
  - Metadata matters
  - Wide variety of measurements/observations

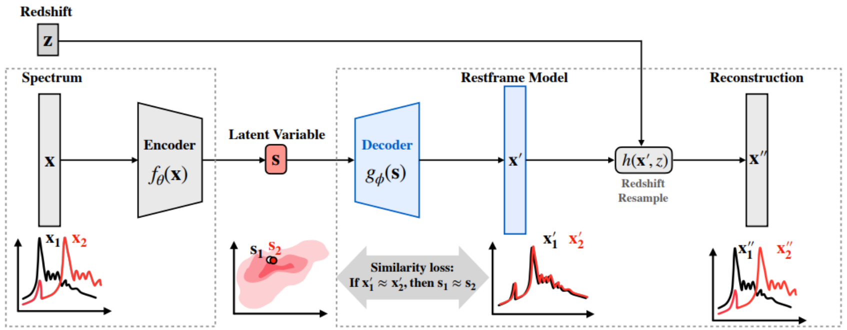
Towards Large Multi-Modal Observational Models

Most Specific

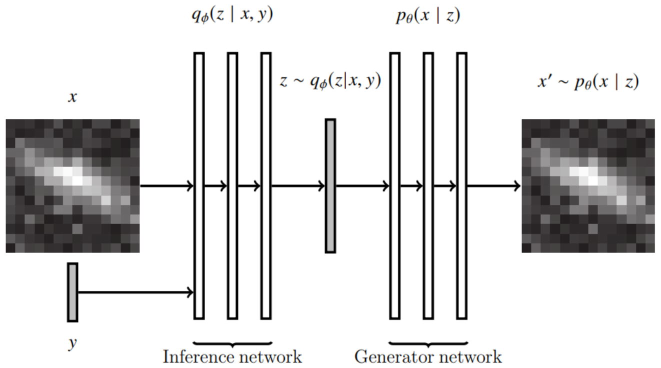
Most General

Independent models for every type of observation

Single model capable of processing all types of observations



Liang et al. 2023

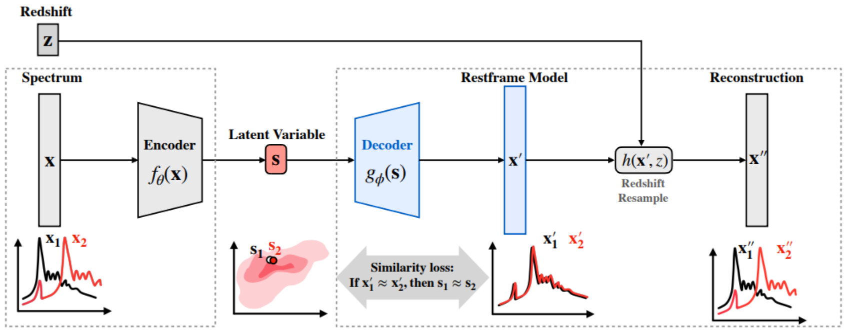


Lanusse et al. 2020

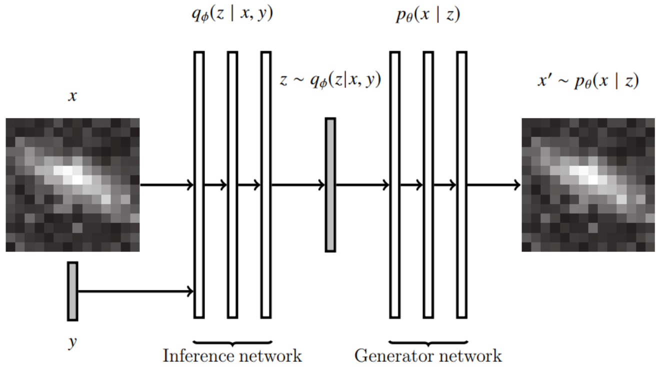
Towards Large Multi-Modal Observational Models

Most Specific | **Most General**

Independent models for every type of observation

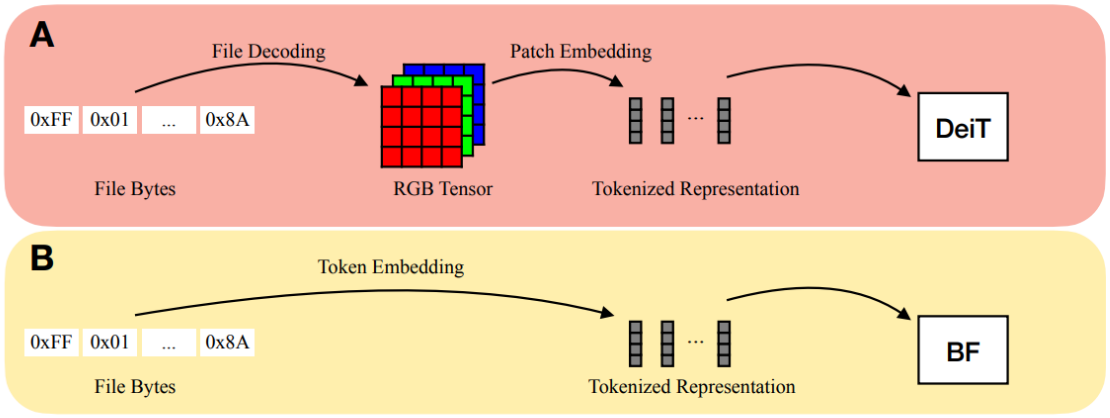


Liang et al. 2023



Lanusse et al. 2020

Single model capable of processing all types of observations



Bytes Are All You Need (Horton et al. 2023)

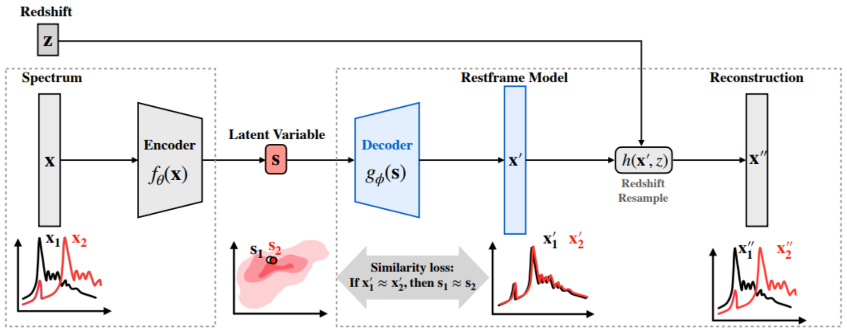
Towards Large Multi-Modal Observational Models

Most Specific **Most General**

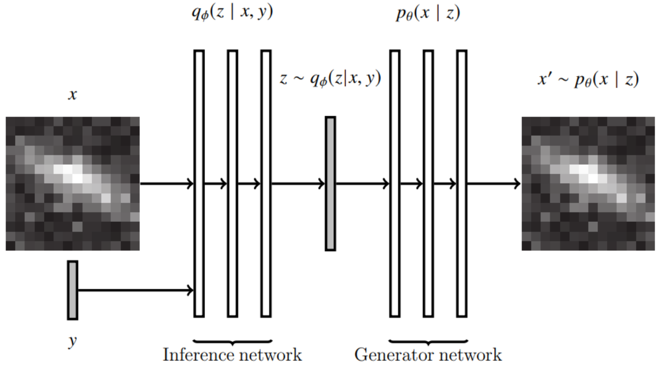
AstroCLIP

Independent models for every type of observation

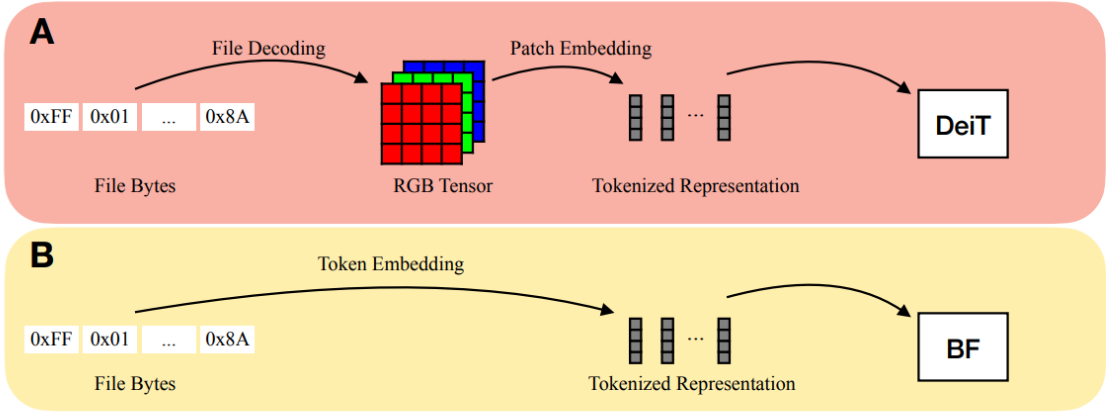
Single model capable of processing all types of observations



Liang et al. 2023



Lanusse et al. 2020



Bytes Are All You Need (Horton et al. 2023)

AstroCLIP

Cross-Modal Pre-Training for Astronomical Foundation Models

astro-ph.IM arXiv:2310.03024

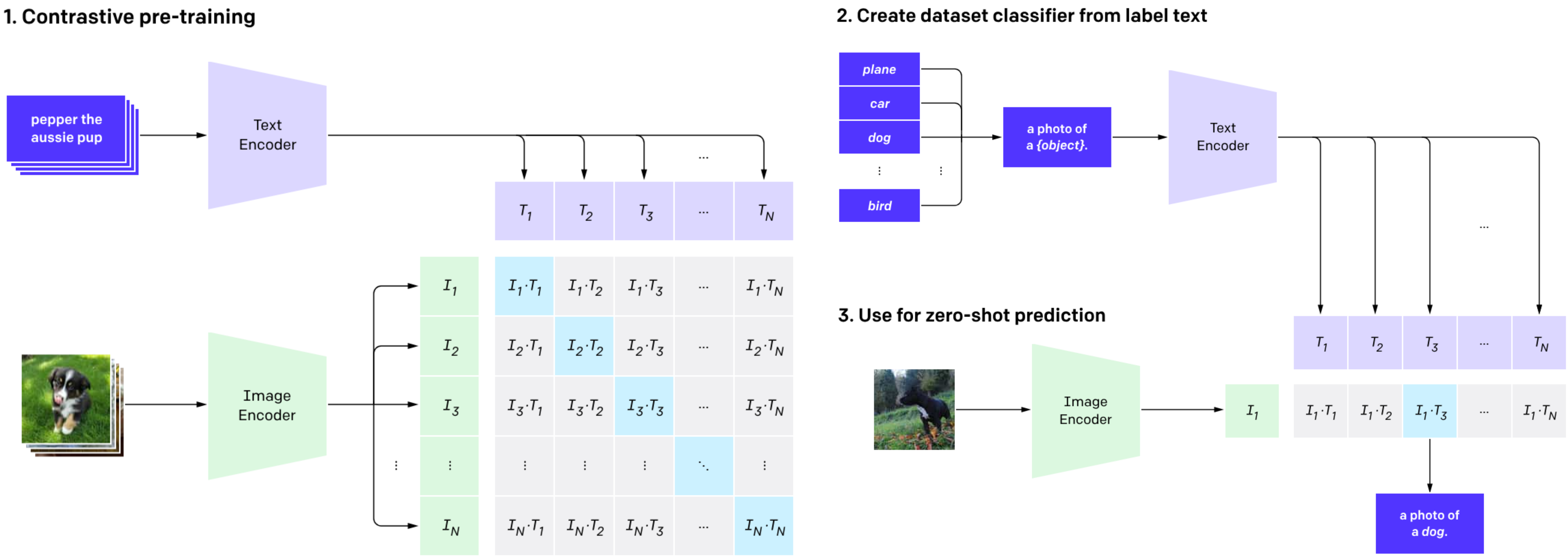
Polymathic

Project led by Francois Lanusse, Liam Parker, Leopoldo Sarra, Siavash Golkar, Miles Cranmer

Accepted contribution at the NeurIPS 2023 AI4Science Workshop

Published in the Monthly Notices of Royal Astronomical Society

What is CLIP?



$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

Contrastive Language Image Pretraining (CLIP)

(Radford et al. 2021)

One model, many downstream applications!

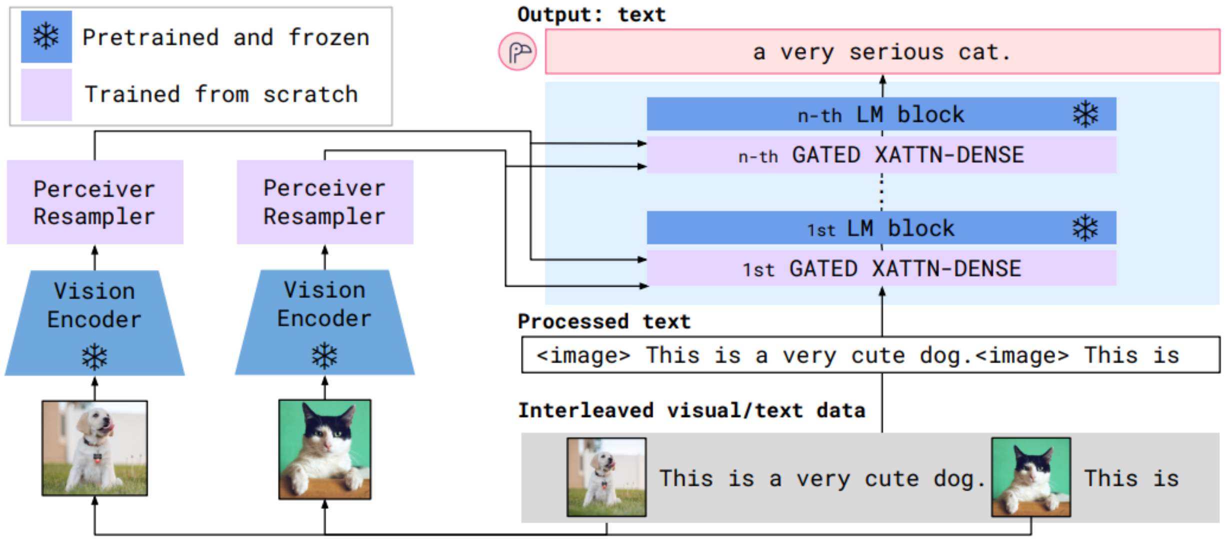
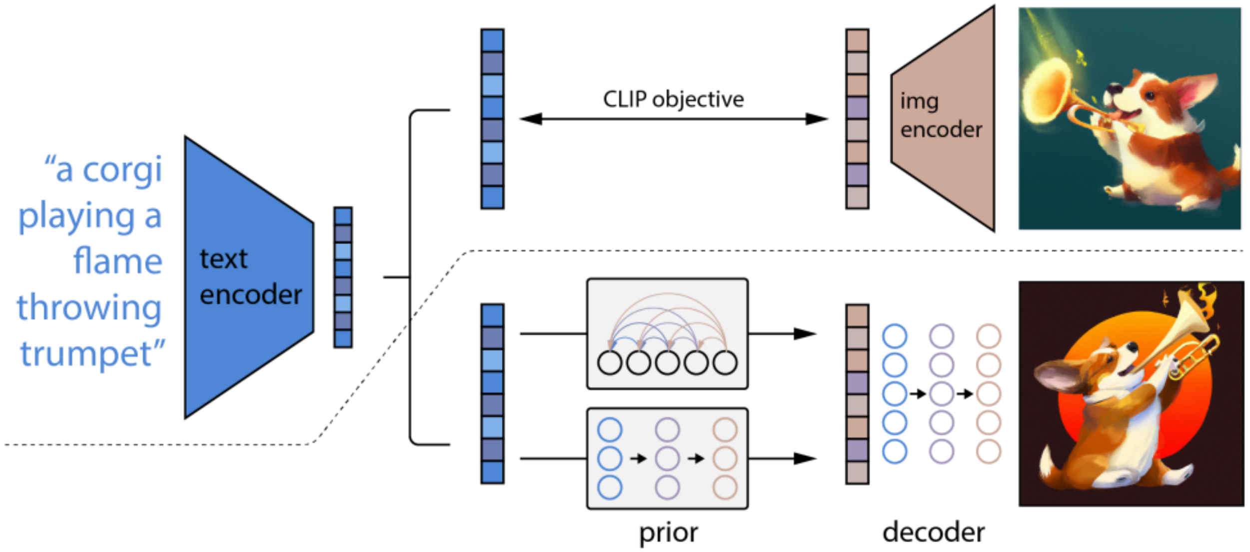


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

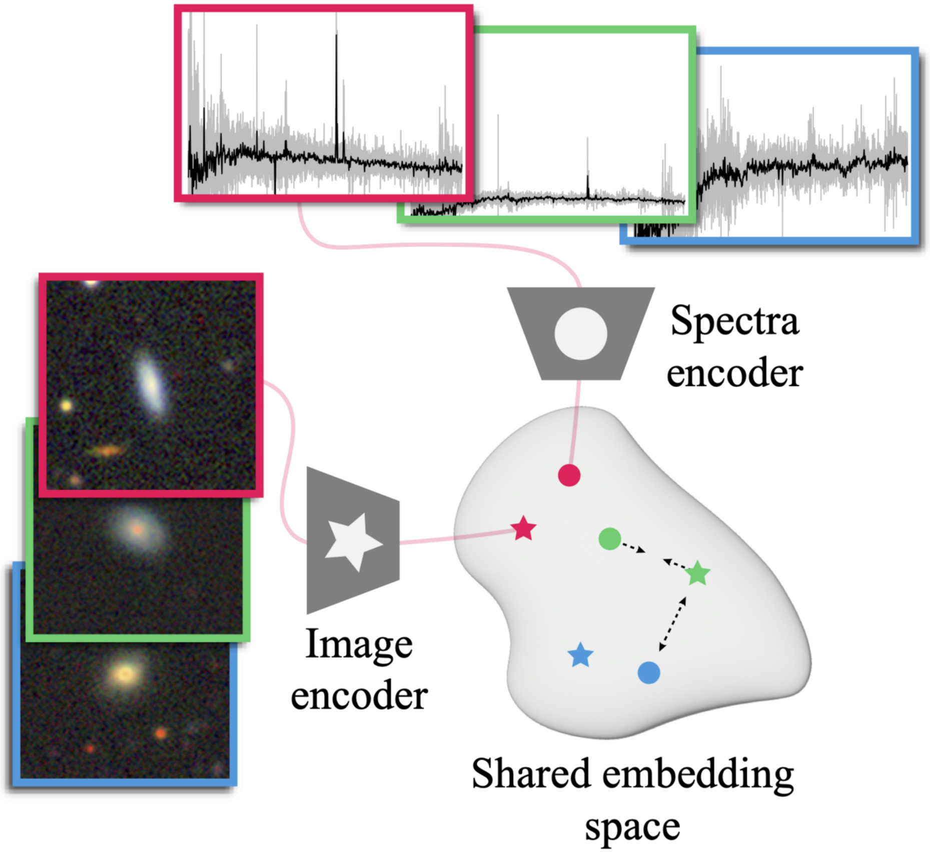


Flamingo: a Visual Language Model for Few-Shot Learning (Alayrac et al. 2022)

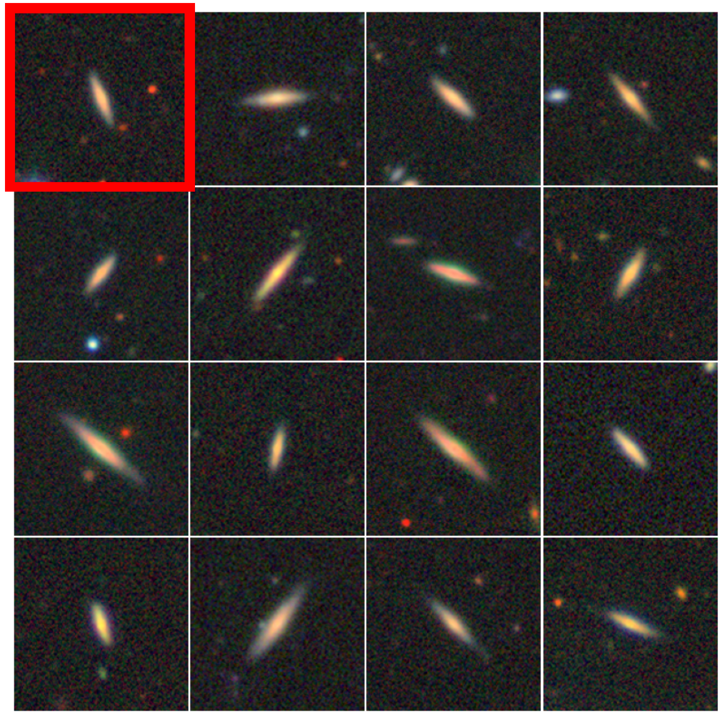
Hierarchical Text-Conditional Image Generation with CLIP Latents (Ramesh et al. 2022)

The AstroCLIP approach

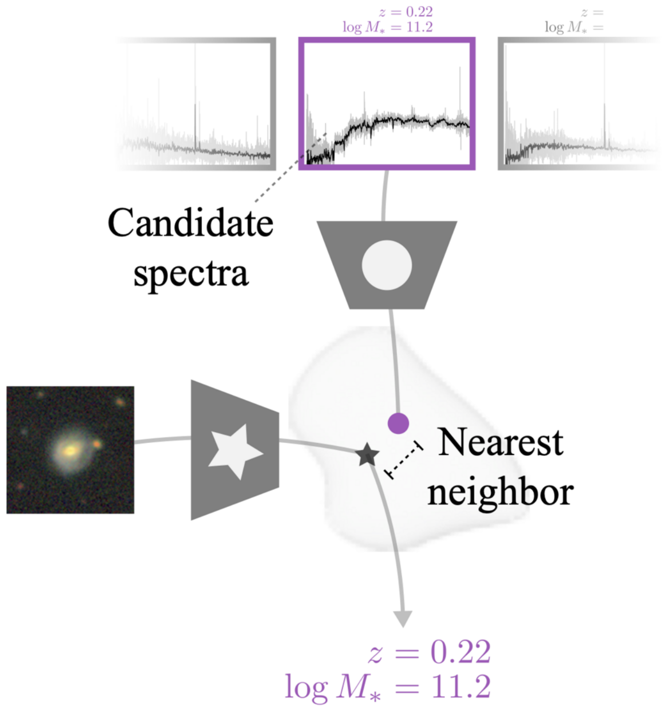
- We use **spectra** and multi-band **images** as our two different views for the same underlying object.
- DESI Legacy Surveys *(g,r,z)* images, and DESI EDR galaxy spectra.



$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$



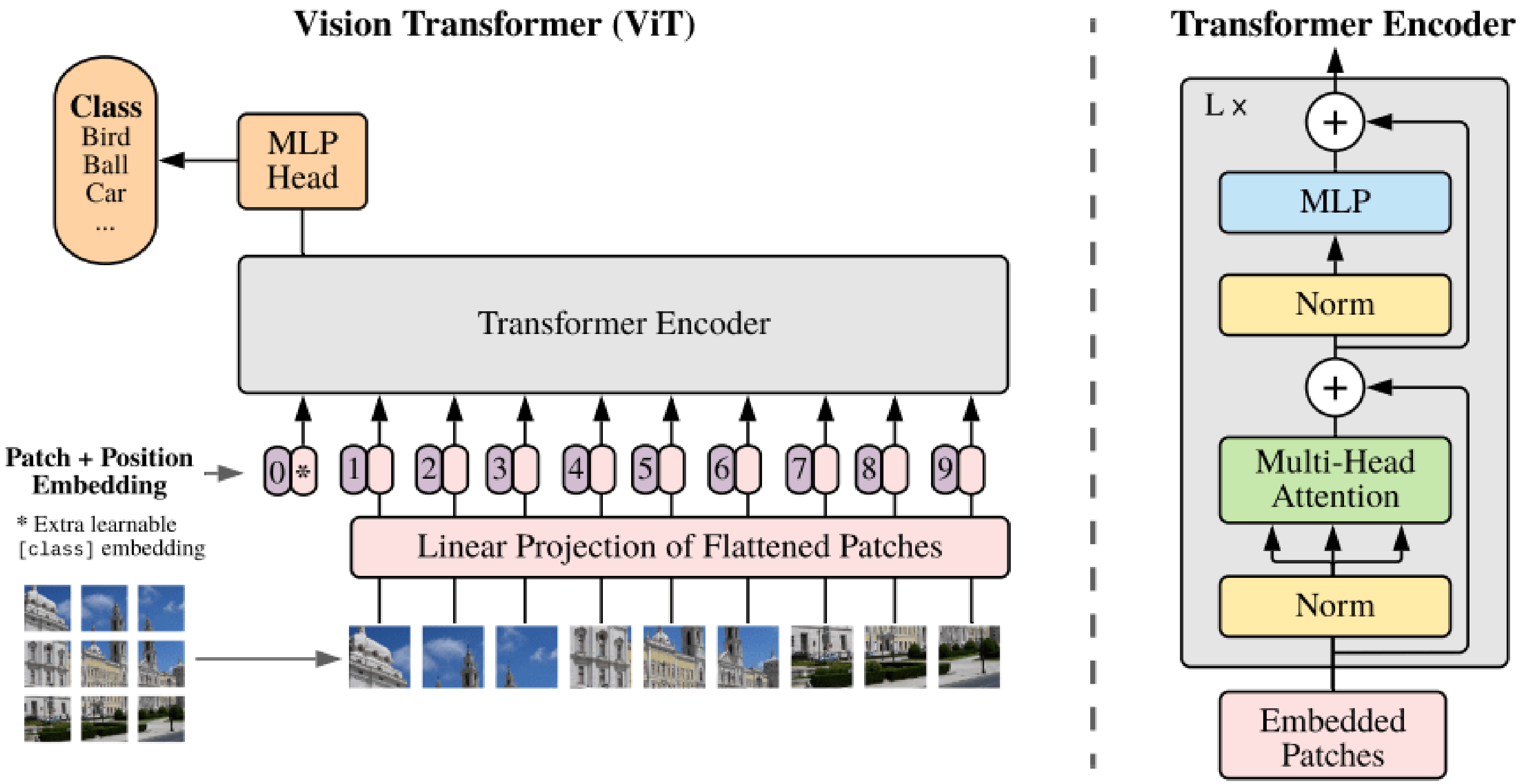
Cosine similarity search



Zero-shot prediction

The AstroCLIP Model (v2, Parker et al. in prep.)

- For **images**, we use a ViT-L Transformer (300M).
- For **spectra**, we use a decoder only Transformer working at the level of spectral patches.



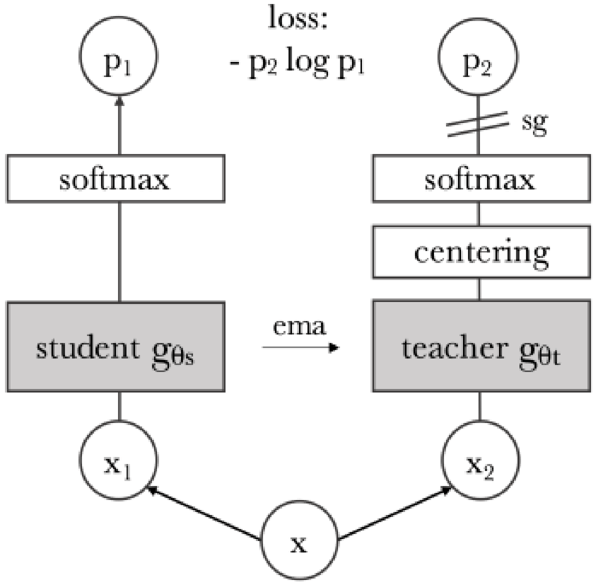
(Dosovitskiy et al 2021)

DiNOv2 (Oquab et al. 2023) Image Pretraining

- Common practice for SOTA CLIP models is to initially pretrain the image encoder before CLIP alignment
- We adopt the **DiNOv2** state of the art Self-Supervised Learning model for the initial large scale training of the model.

PCA of patch features

	INet-1k k-NN	INet-1k linear
iBOT	72.9	82.3
+(our reproduction)	74.5 ↑1.6	83.2 ↑0.9
+LayerScale, Stochastic Depth	75.4 ↑0.9	82.0 ↓1.2
+128k prototypes	76.6 ↑1.2	81.9 ↓0.1
+KoLeo	78.9 ↑2.3	82.5 ↑0.6
+SwiGLU FFN	78.7 ↓0.2	83.1 ↑0.6
+Patch size 14	78.9 ↑0.2	83.5 ↑0.4
+Teacher momentum 0.994	79.4 ↑0.5	83.6 ↑0.1
+Tweak warmup schedules	80.5 ↑1.1	83.8 ↑0.2
+Batch size 3k	81.7 ↑1.2	84.7 ↑0.9
+Sinkhorn-Knopp	81.7 =	84.7 =
+Untying heads = DINOv2	82.0 ↑0.3	84.5 ↓0.2



Dense Semantic Segmentation

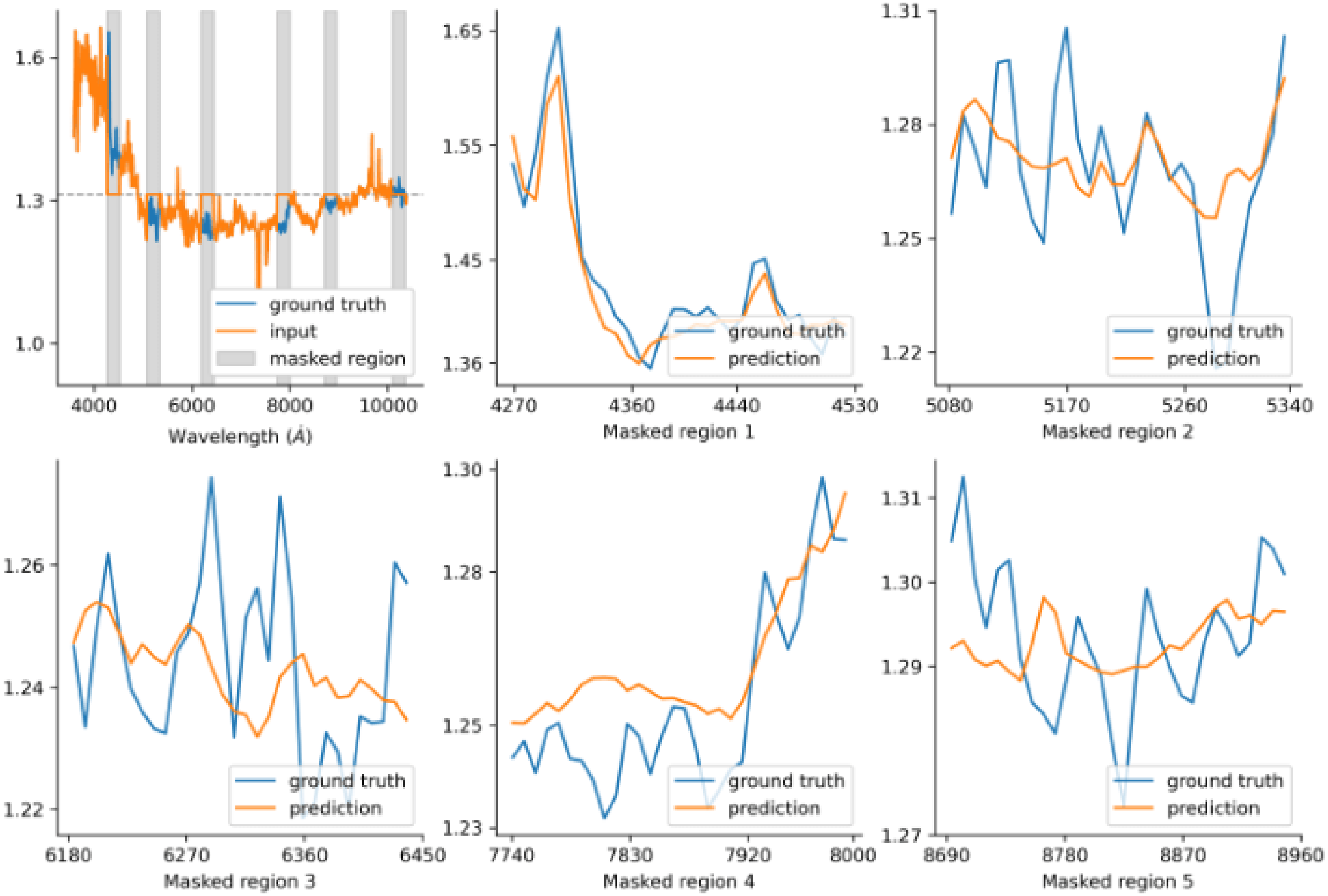
Dense Depth Estimation

- We **pretrain** the DiNOv2 model on ~**70 million postage stamps** from DECaLS

Spectrum Transformer Pretraining by Masked Modeling

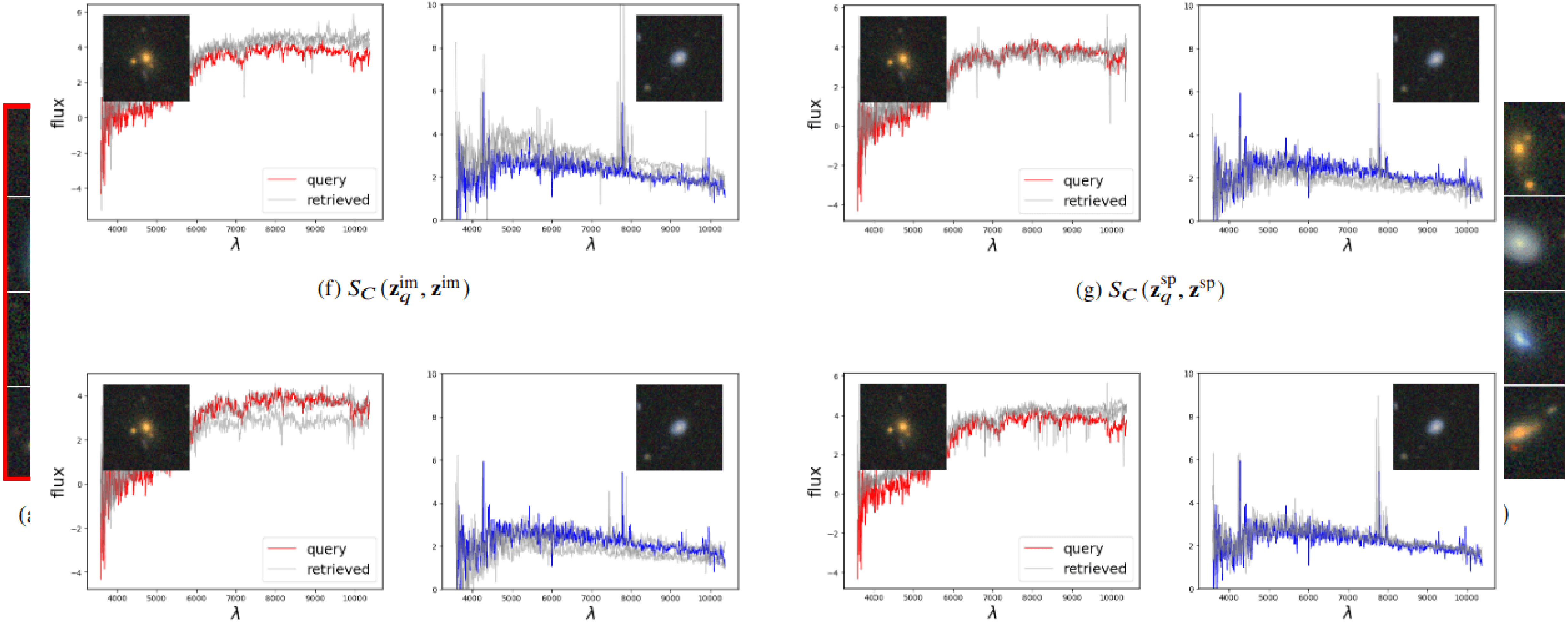
- To pretrain the spectrum embedder, we use a simple Masked Image Modeling strategy

$$\mathcal{L}_{\text{MM}} = \frac{1}{NK} \sum_{j=1}^{K} \sum_{i=1}^{N} \mathbf{m}_i \cdot (\mathbf{x}_i - \hat{\mathbf{x}}_i)^2,$$



Evaluation of the model: Similarity Search

- Cross-Modal similarity search



(f) $S_C(\mathbf{z}_q^{\text{im}}, \mathbf{z}^{\text{im}})$

(g) $S_C(\mathbf{z}_q^{\text{sp}}, \mathbf{z}^{\text{sp}})$

(h) $S_C(\mathbf{z}_q^{\text{sp}}, \mathbf{z}^{\text{im}})$

(i) $S_C(\mathbf{z}_q^{\text{im}}, \mathbf{z}^{\text{sp}})$

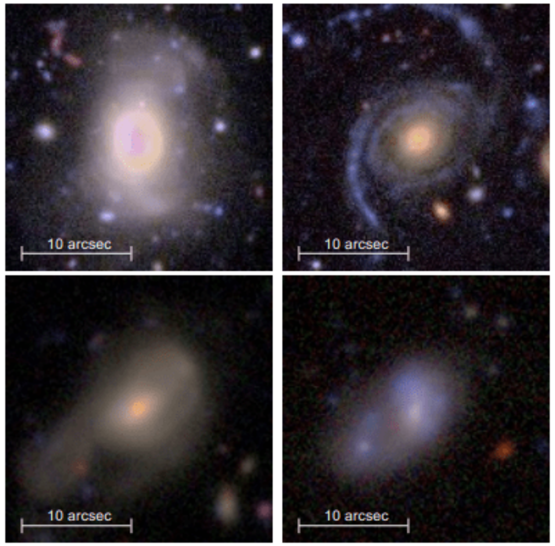
$$S_C(\mathbf{z}_i^{sp}, \mathbf{z}_i^{im}) = (\mathbf{z}_i^{sp} \cdot \mathbf{z}_i^{im}) / \parallel \mathbf{z}_i^{sp} \parallel \parallel \mathbf{z}_i^{im} \parallel$$

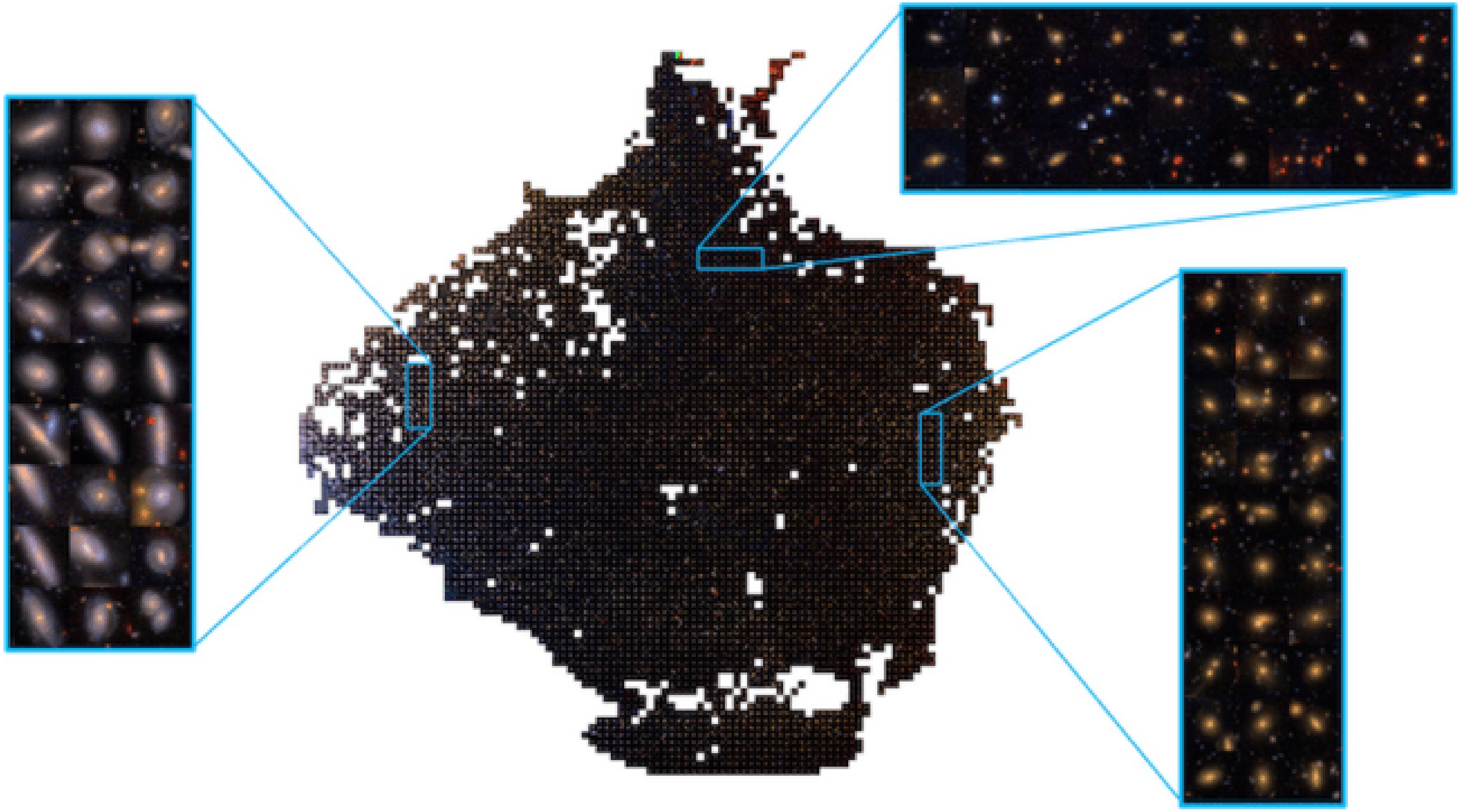
Image Similarity　Image-Spectral Similarity　Spectral Similarity

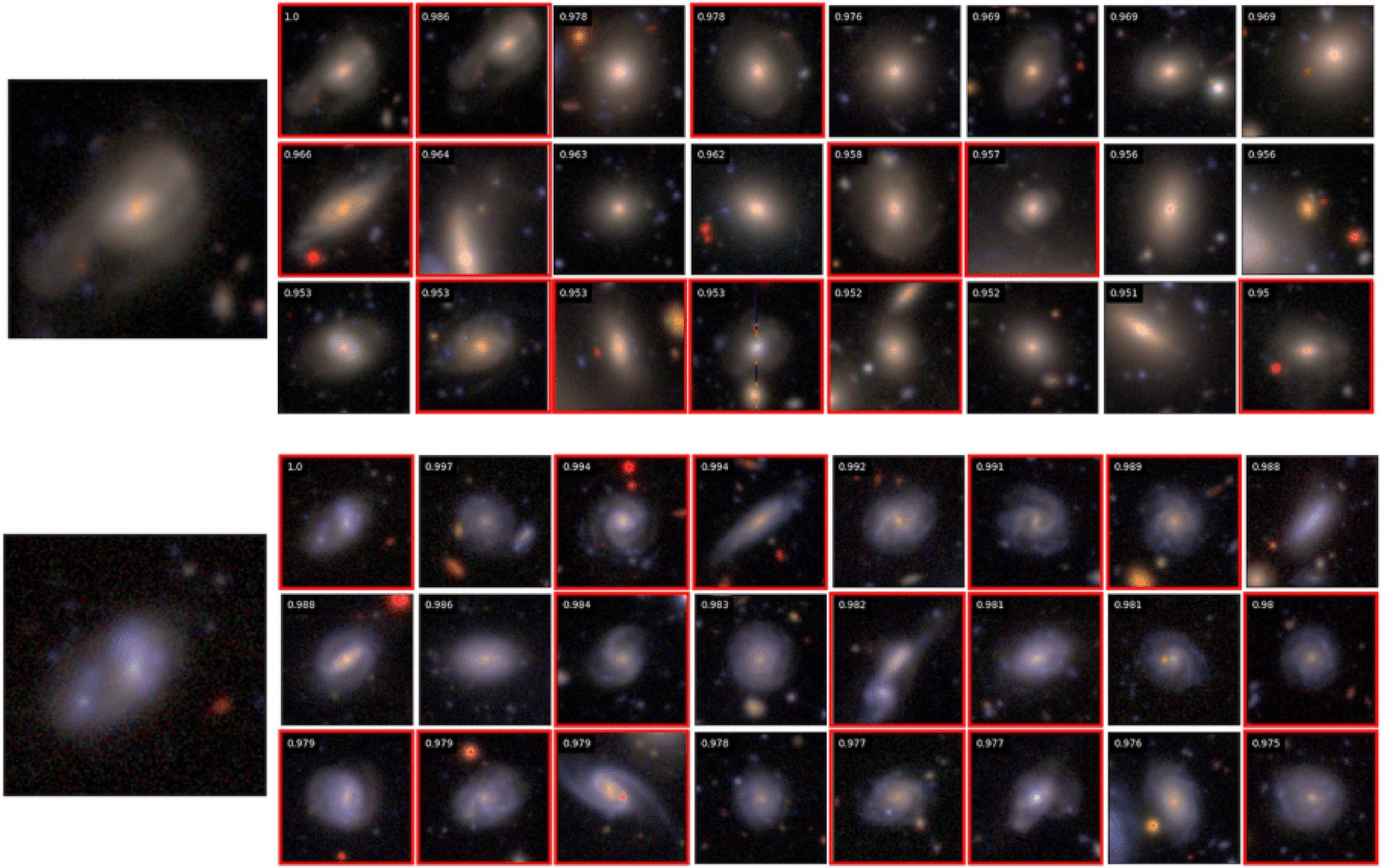
Detecting Galaxy Tidal Features Using Self-Supervised Representation Learning

astro-ph.GA arXiv:2308.07962

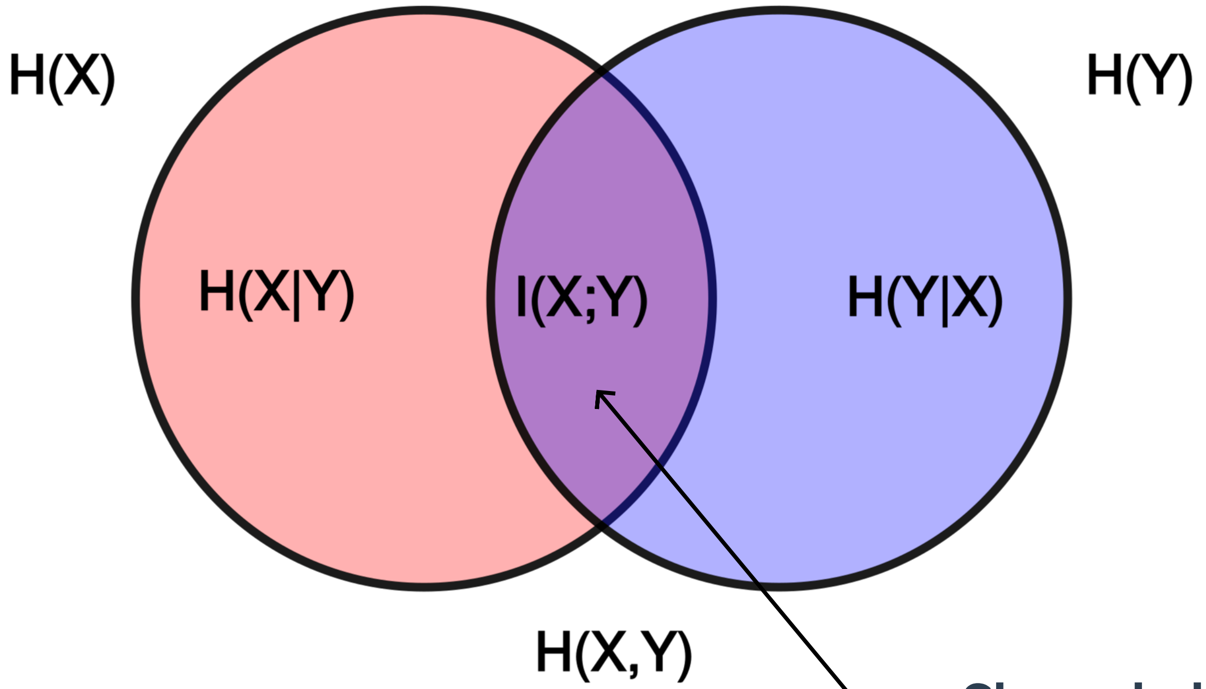
Project led by Alice Desmons, Francois Lanusse, Sarah Brough






1.0
0.986
0.978
0.978
0.976
0.969
0.969
0.969
0.966
0.964
0.963
0.962
0.958
0.957
0.956
0.956
0.953
0.953
0.953
0.953
0.952
0.952
0.951
0.95
1.0
0.997
0.994
0.994
0.992
0.991
0.989
0.988
0.988
0.986
0.984
0.983
0.982
0.981
0.981
0.98
0.979
0.979
0.979
0.978
0.977
0.977
0.976
0.975

The Information Point of View



- The InfoNCE loss is a lower bound on the **Mutual Information** between modalities

=> We are building **summary statistics for the physical parameters** describing an object in a completely **data driven way**

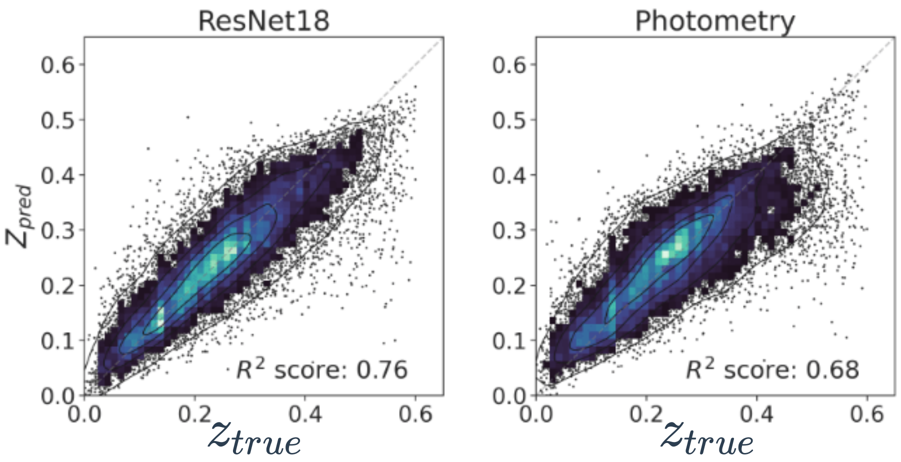
Shared physical information about galaxies between images and spectra

$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

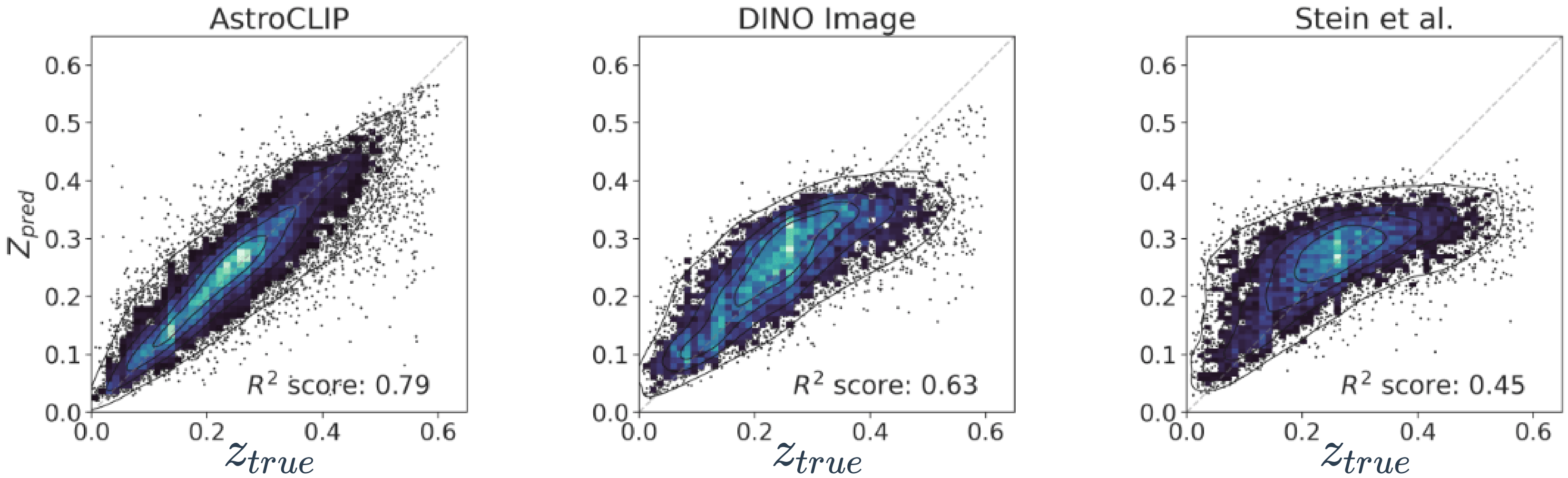
Daunhawer et al. (2023)
van den Oord et al. (2018)

Evaluation of the model: Parameter Inference

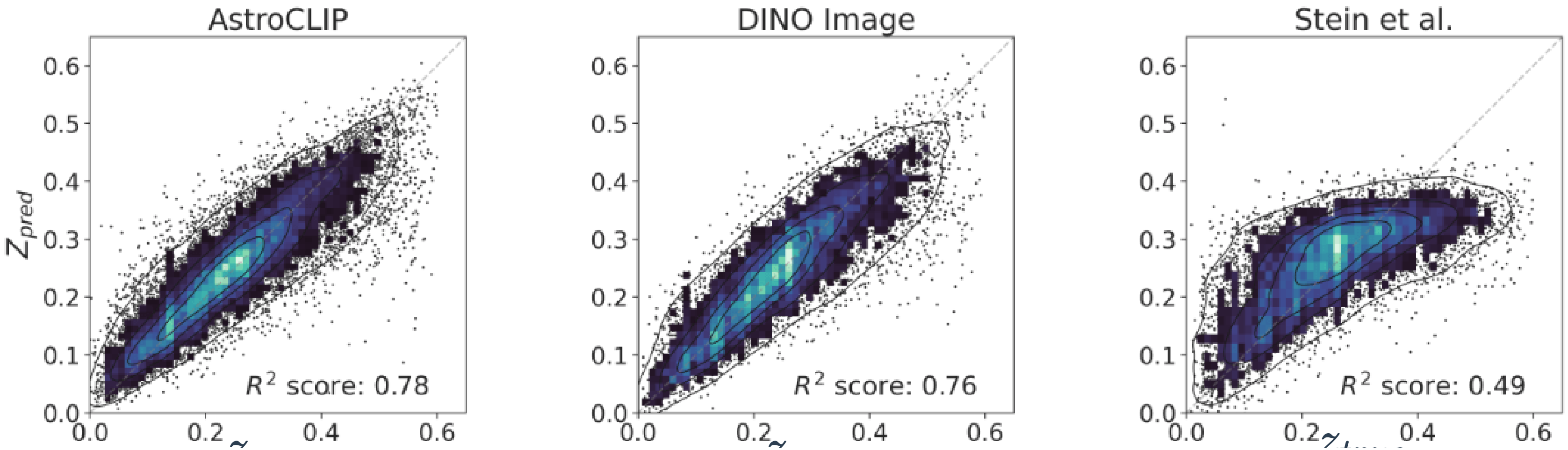
- Redshift Estimation From Images



Supervised baseline



- Zero-shot prediction
  - k-NN regression



- Few-shot prediction
  - MLP head trained on top of frozen backbone

- Galaxy Physical Property Estimation from Images and Spectra

We use estimates of galaxy properties from the PROVABGS catalog (Hahn et al. 2023) (Bayesian spectral energy distribution (SED) modeling of DESI spectroscopy and photometry method)

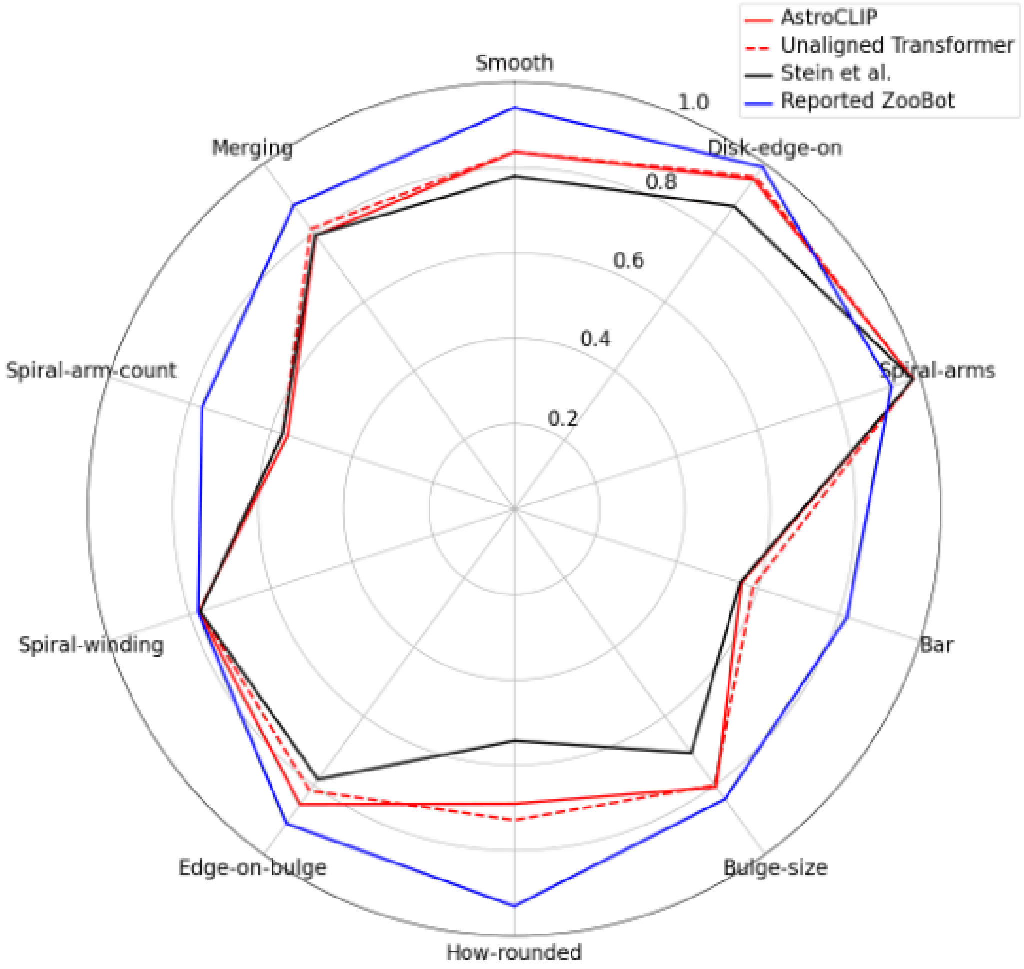
Source	Method	NLL
Images	AstroCLIP*	**0.77 ± 0.00**
	SSL Transformer*	0.82 ± 0.00
	Stein et al. (2021b)	1.02 ± 0.04
	ResNet18	0.84 ± 0.00
Spectra	AstroCLIP*	0.17 ± 0.04
	SSL Transformer*	**0.00 ± 0.04**
	Conv+Att	0.29 ± 0.000
Photometry	MLP	1.06 ± 0.05

Negative Log Likelihood of Neural Posterior Inference

Source	Method	M_*	Z_{MW}	t_{age}	$sSFR$
Images	AstroCLIP				
	Zero-Shot*	**0.73**	**0.43**	**0.25**	**0.42**
	Few-Shot*	0.71	0.42	0.25	0.42
	SSL Transformer				
	Zero-Shot*	0.62	0.37	0.14	0.22
	Few-Shot*	0.72	0.42	0.23	0.40
	Stein et al. (2021b)				
	Zero-Shot	0.30	0.22	0.10	0.23
	Few-Shot	0.36	0.24	0.11	0.21
	ResNet18	0.72	0.39	0.19	0.38
Spectra	AstroCLIP				
	Zero-Shot*	0.87	0.57	0.43	0.63
	Few-Shot*	**0.88**	0.58	0.43	0.64
	SSL Transformer				
	Zero-Shot*	0.84	0.57	0.38	0.62
	Few-Shot*	0.88	**0.64**	**0.47**	**0.69**
	Conv+Att	0.85	0.62	0.43	0.67
Photometry	MLP	0.67	0.40	0.26	0.34

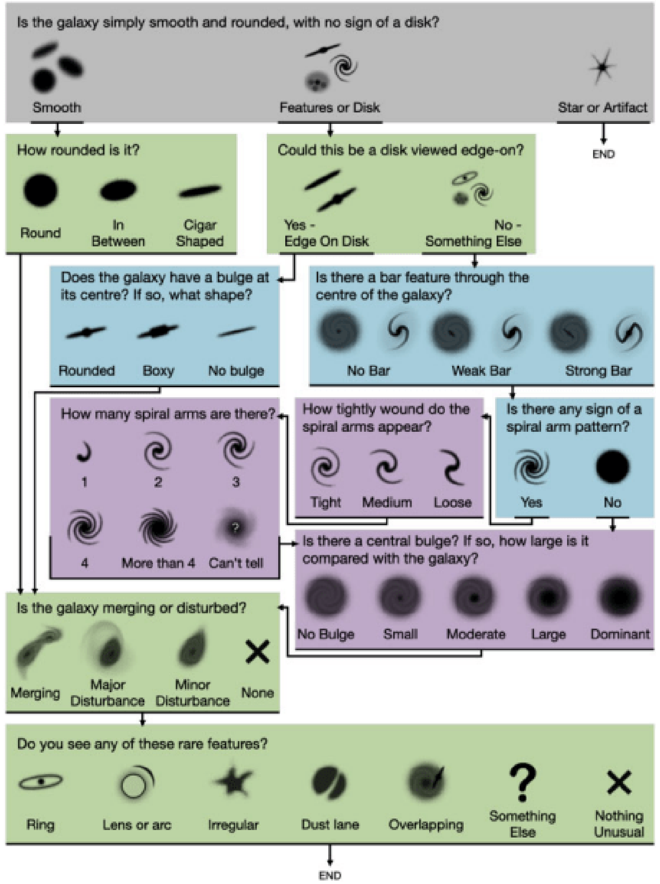
R^2 of regression

- Galaxy Morphology Classification



Classification Accuracy

We test a galaxy morphology classification task using as labels the GZ-5 dataset (Walmsley et al. 2021)



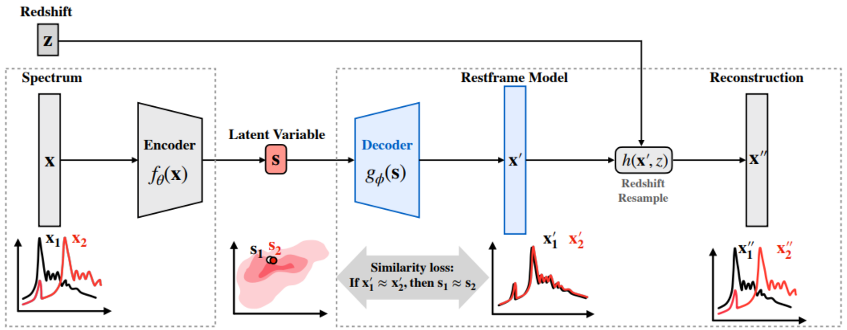
Towards Large Multi-Modal Observational Models

Most Specific **Most General**

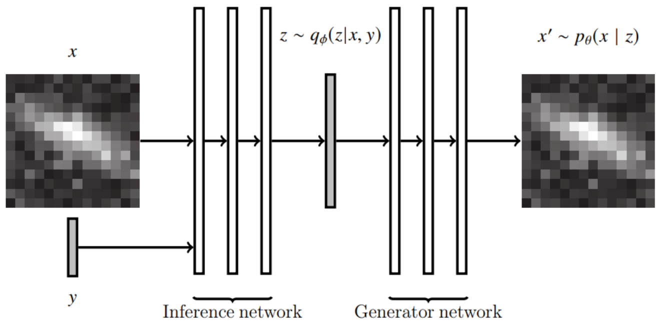
AstroCLIP

Independent models for every type of observation

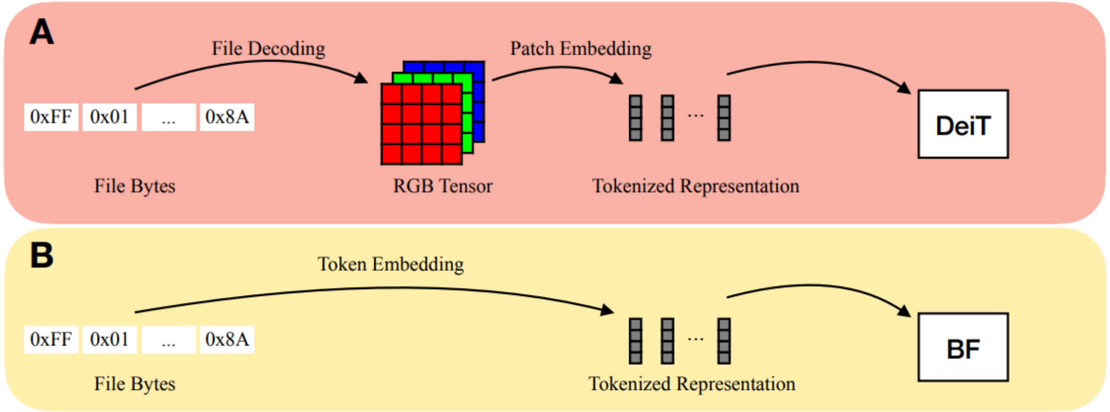
Single model capable of processing all types of observations



Liang et al. 2023



Lanusse et al. 2020



Bytes Are All You Need (Horton et al. 2023)

Towards Large Multi-Modal Observational Models

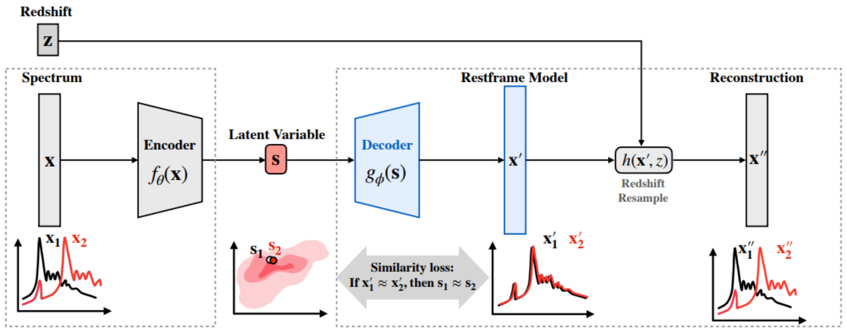
Most Specific

AstroCLIP

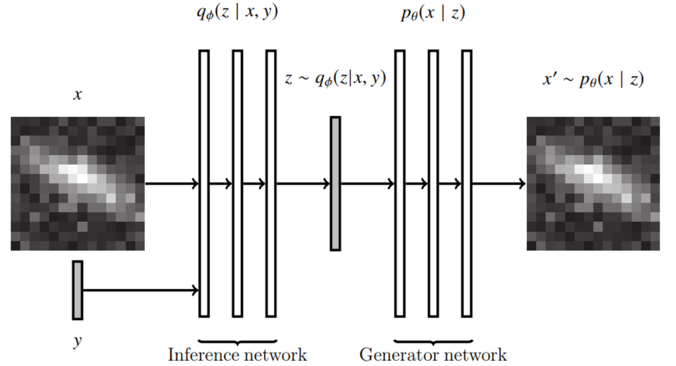
"Massively Multi-Modal Large Data Model for Astrophysics"

Most General

Independent models for every type of observation

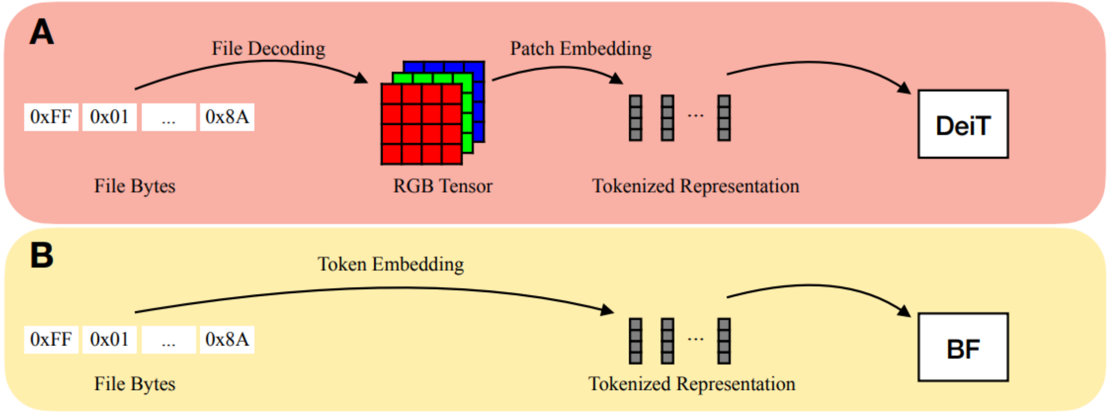


Liang et al. 2023



Lanusse et al. 2020

Single model capable of processing all types of observations



Bytes Are All You Need (Horton et al. 2023)

Towards Massively Multimodal Large Data Models for Astrophysics

cnrs

Institut du Développement et des Ressources en Informatique Scientifique

Polymathic

New Generation of Token-Based Multimodal Models

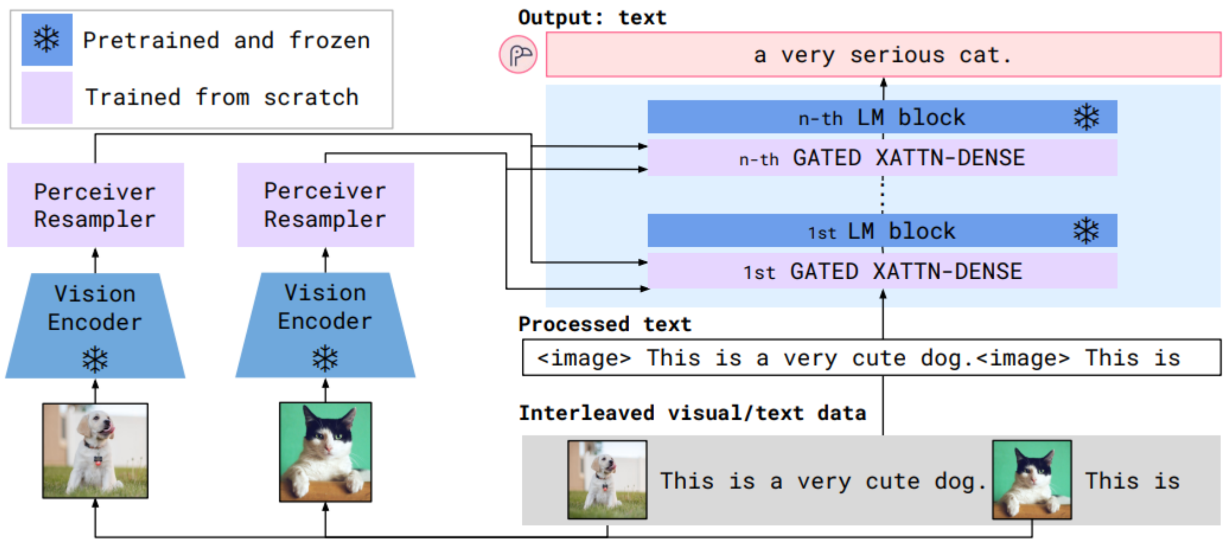
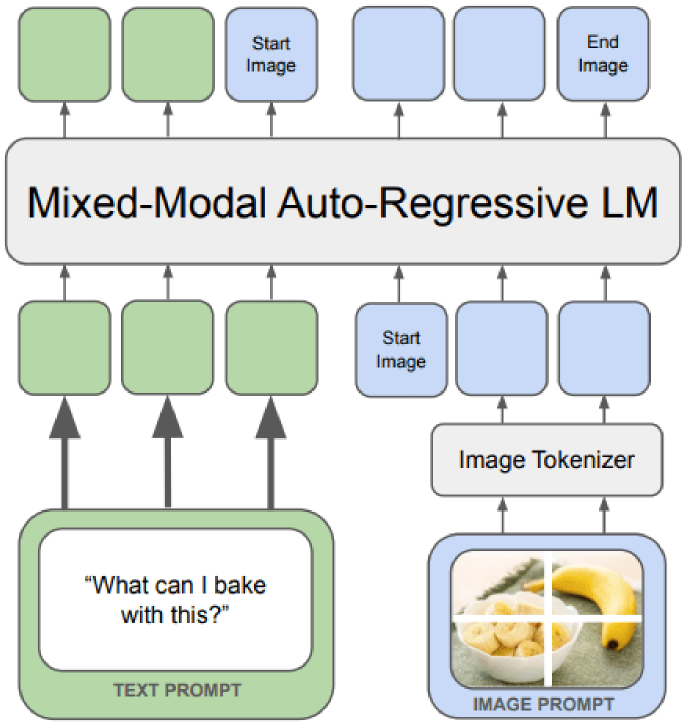


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

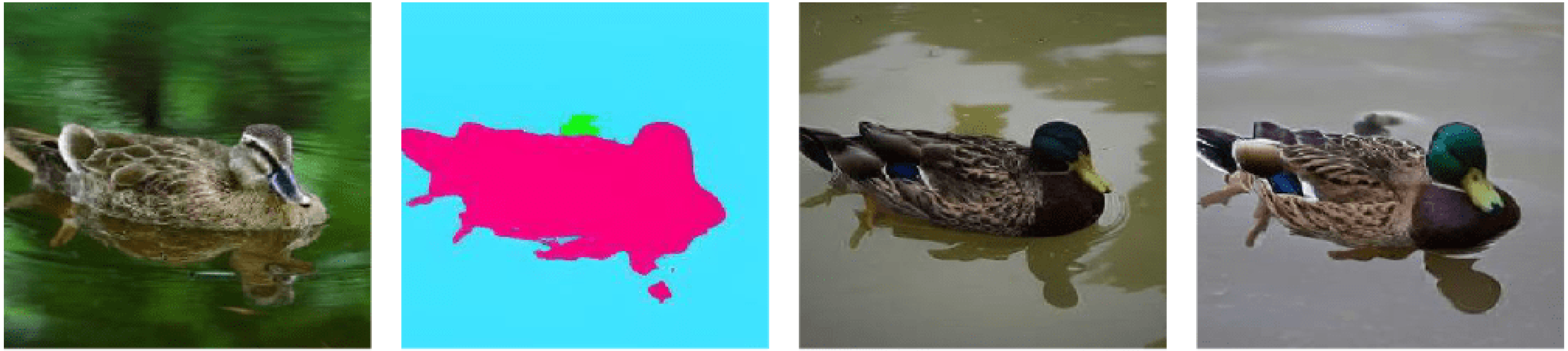


Flamingo: a Visual Language Model for Few-Shot Learning (Alayrac et al. 2022)

Chameleon: Mixed-Modal Early-Fusion Foundation Models (Chameleon team, 2024)

All-to-All Foundation Models

Generate high quality image of "a room that has a sink and a mirror in it" with bottle at location (199, 130) -> (204, 150) and with a sink at location (149, 133) -> (190, 154) and with bed at location (0, 169) -> (67, 255)



INPUT

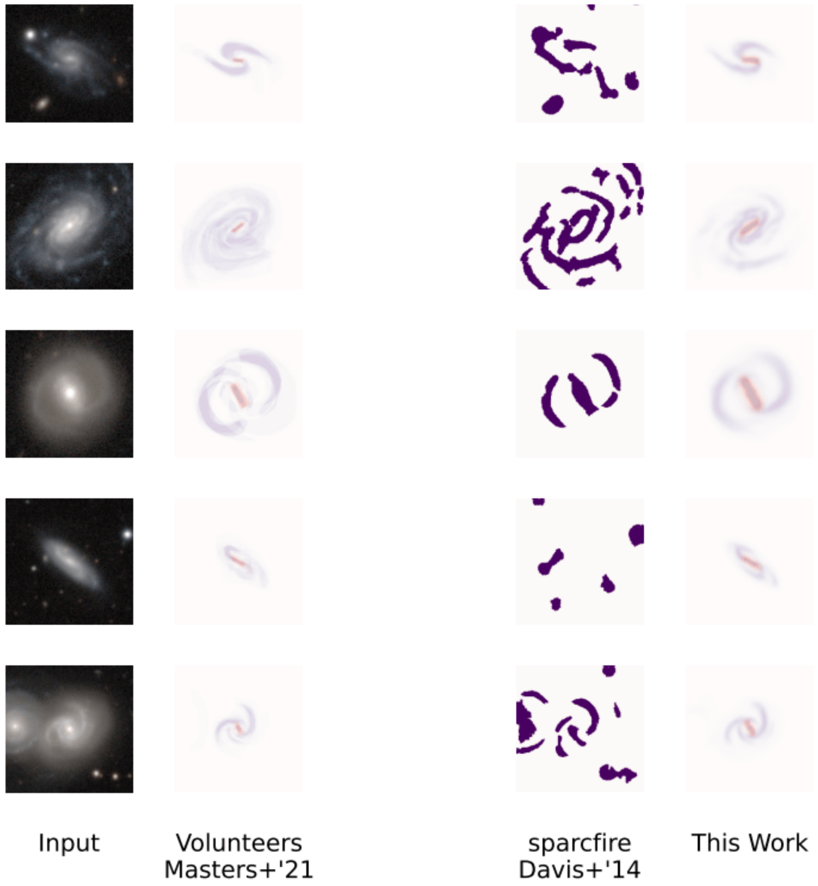
EXTRACTED SEGMENTATION (UNIFORMER)

GENERATION 1

GENERATION 2

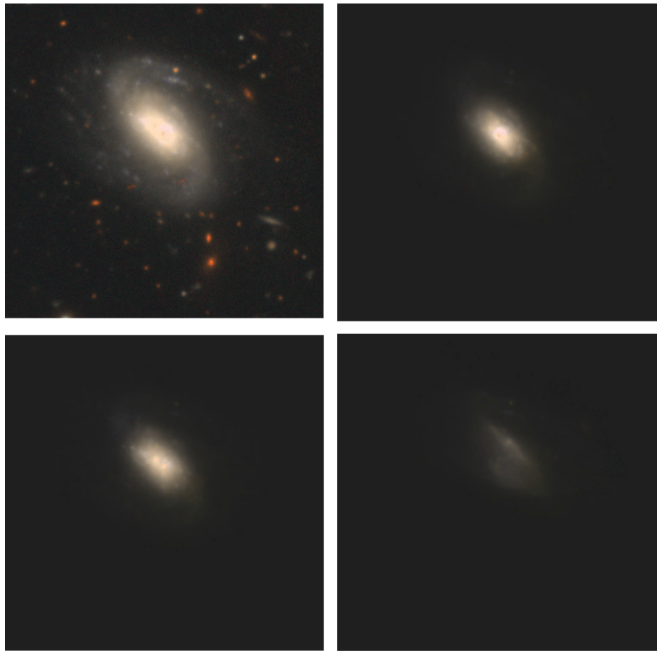
Yu et al. (2023)

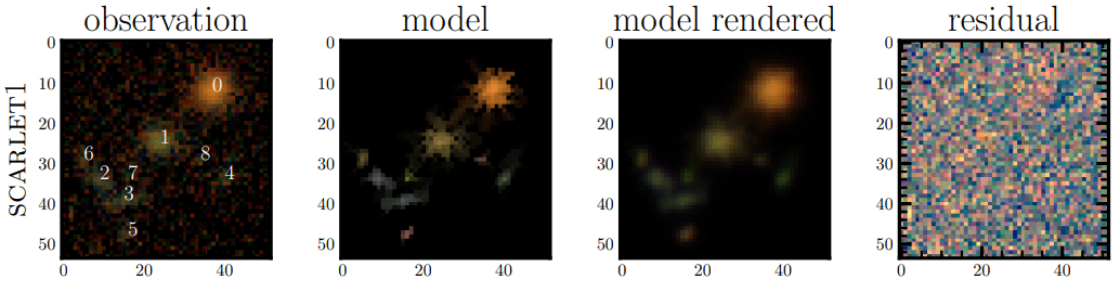
Why Is It Interesting to Us?



Galaxy Image Segmentation

Walsmley & Spindler (2023)





Galaxy Image Deblending

Bosch et al. (2017), Sampson et al. (2024)

=> Foundation Models that build a deep understanding of the data at the pixel level.

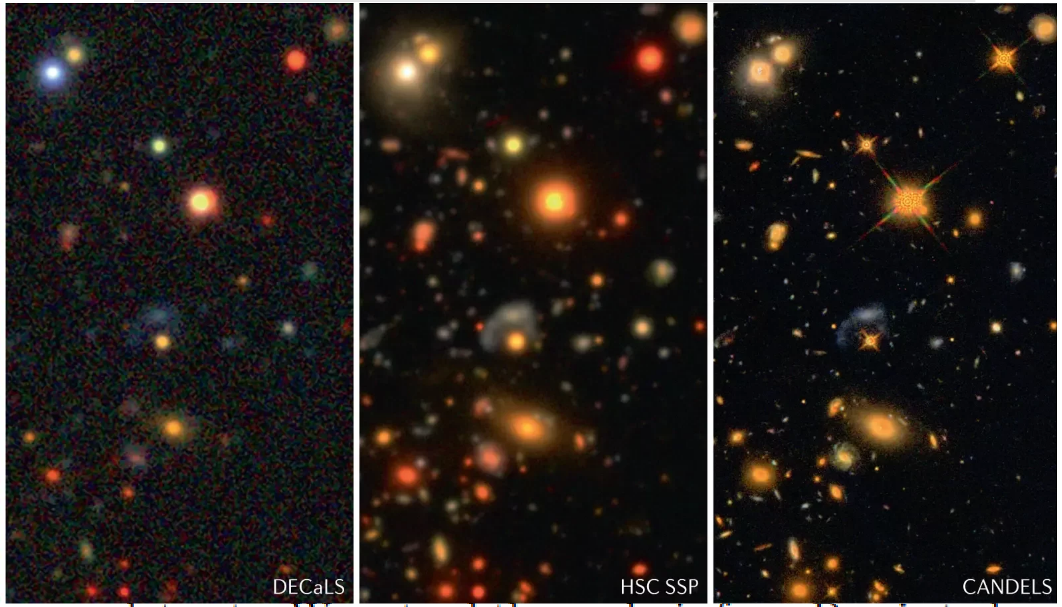
Going Further: Data Collection and Curation

- Development of large models requires access to "web scale" datasets

- Astrophysics generates large amounts of publicly available data,
  - **BUT**, data is usually not stored or structured in an ML friendly way.

- Accessing and using scientific data **requires significant expertise**, for each dataset.

Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K



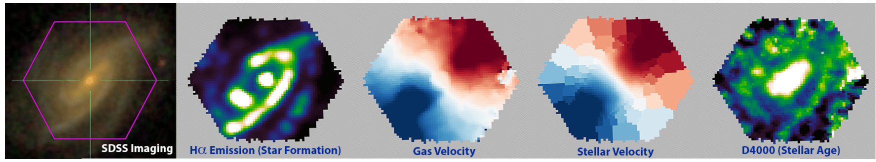
Credit: Melchior et al. 2021

datasets. We extend the analysis from Desai et al. [14] and compare the sizes of public and private image-text datasets.

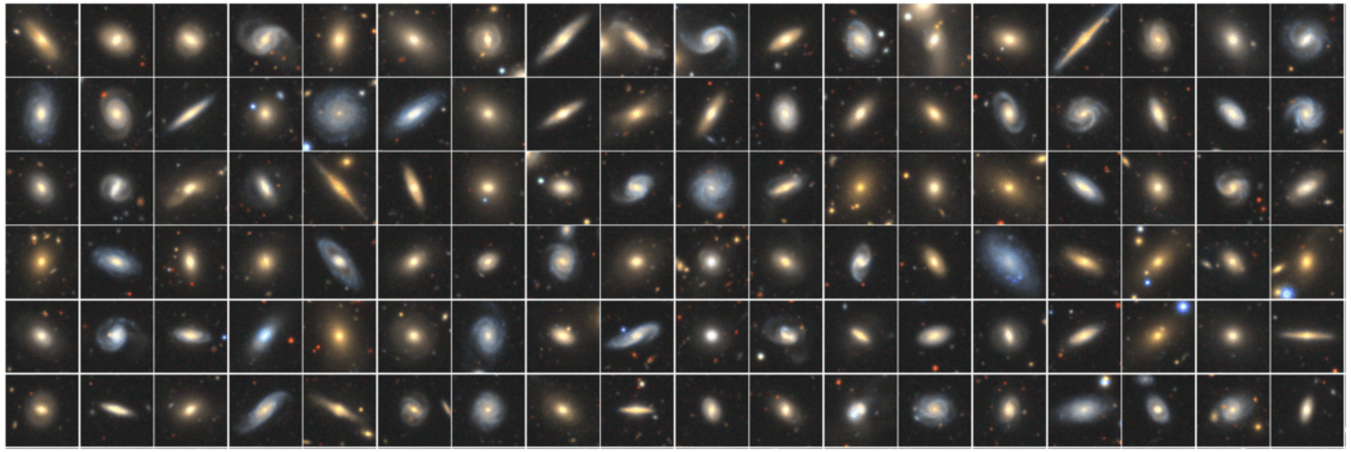
Schuhmann et al. (2022)

The MultiModal Universe Project

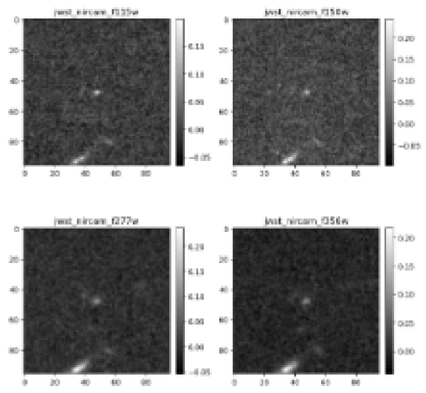
- **Goal**: Assemble the first large-scale multi-modal dataset for machine learning in astrophysics.
- **Main pillars**:
  - Engage with a **broad community of AI+Astro experts**.
  - Adopt **standardized conventions for storing and accessing data** and metadata through mainstream tools (e.g. Hugging Face Datasets).
  - Target large astronomical surveys, varied types of instruments, many **different astrophysics sub-fields**.



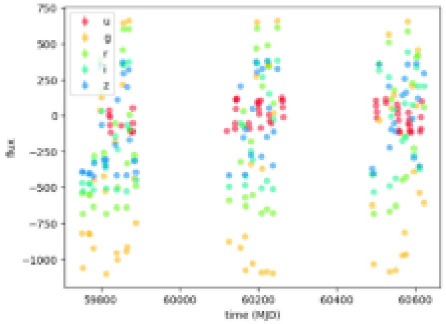
Hyperspectral Images from MaNGA



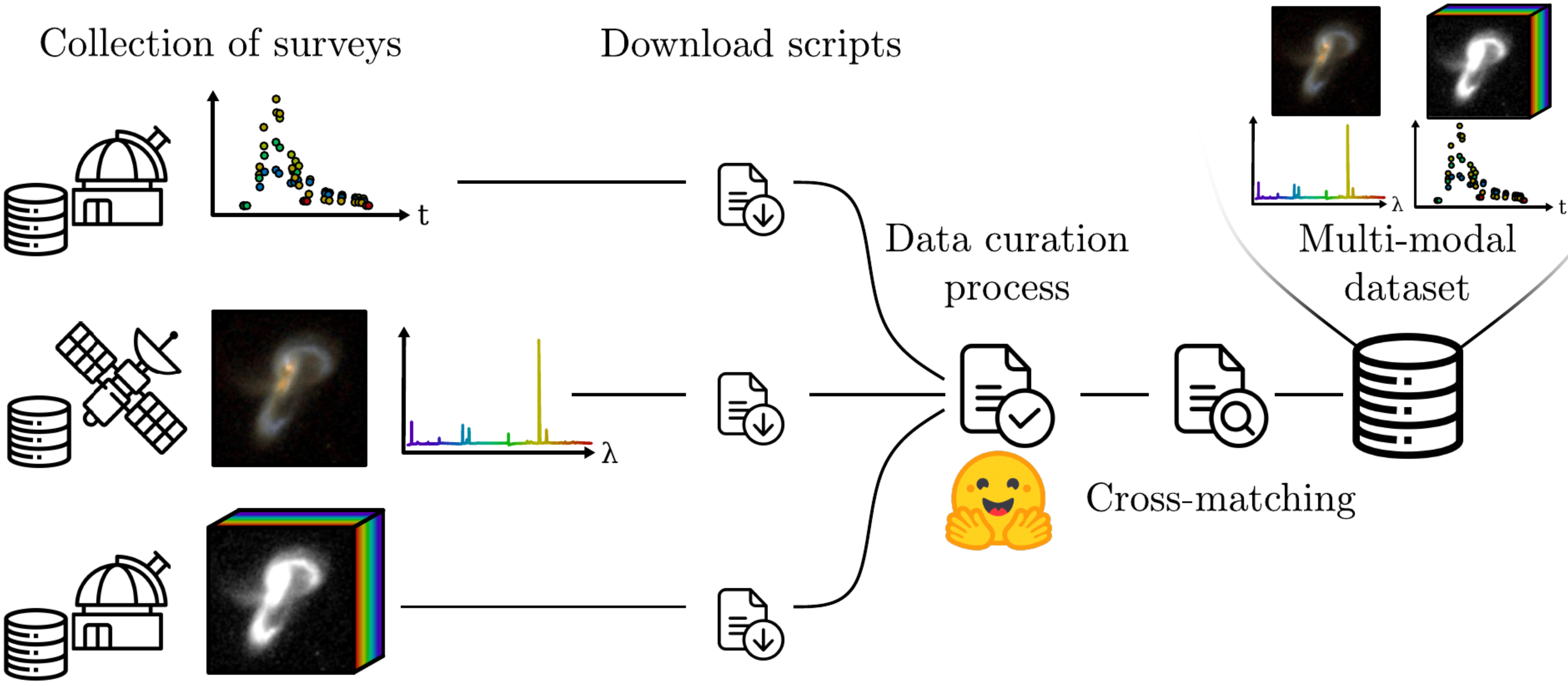
Multiband images from Legacy Survey



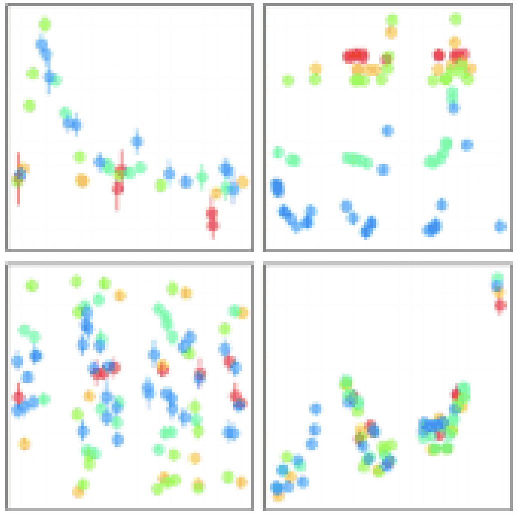
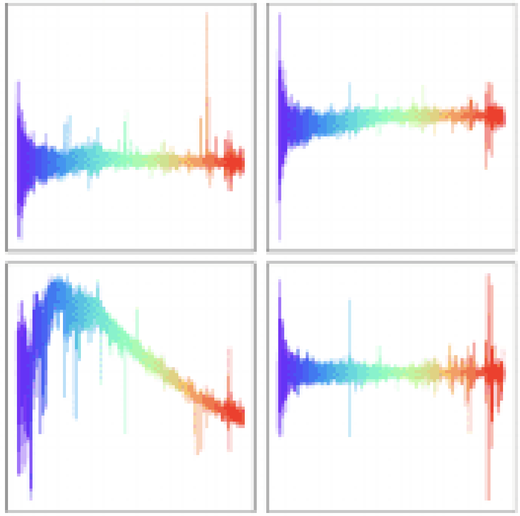
JWST Images



Time series of supernovae and stellar variability



Collection of surveys
Download scripts
Data curation process
Cross-matching
Multi-modal dataset
t
λ

	Images	Time-Series	Spectra
# examples	140M	3.6M	225M
Description	images in a variety of wavelength ranges, including optical and infrared	multivariate time-series of flux + uncertainty in different wavelength ranges	flux as a function of wavelength
Tasks	galaxy classification, physical property estimation	time-series classification, redshift estimation	physical property estimation
Examples			

Accepted at NeurIPS 2024 🎉

=> Official release October 2024

https://github.com/MultimodalUniverse/MultimodalUniverse

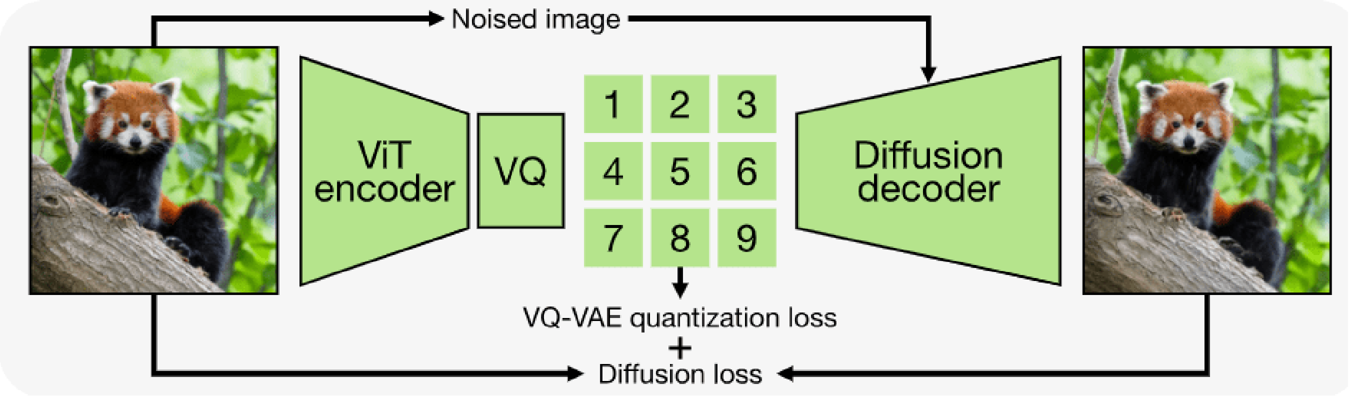
Multimodal Universe: Enabling Large-Scale Machine Learning with 70TBs of Astronomical Scientific Data

Dataset on 🤗 | Open in Colab | Testing datasets passing | License MIT | all contributors 28

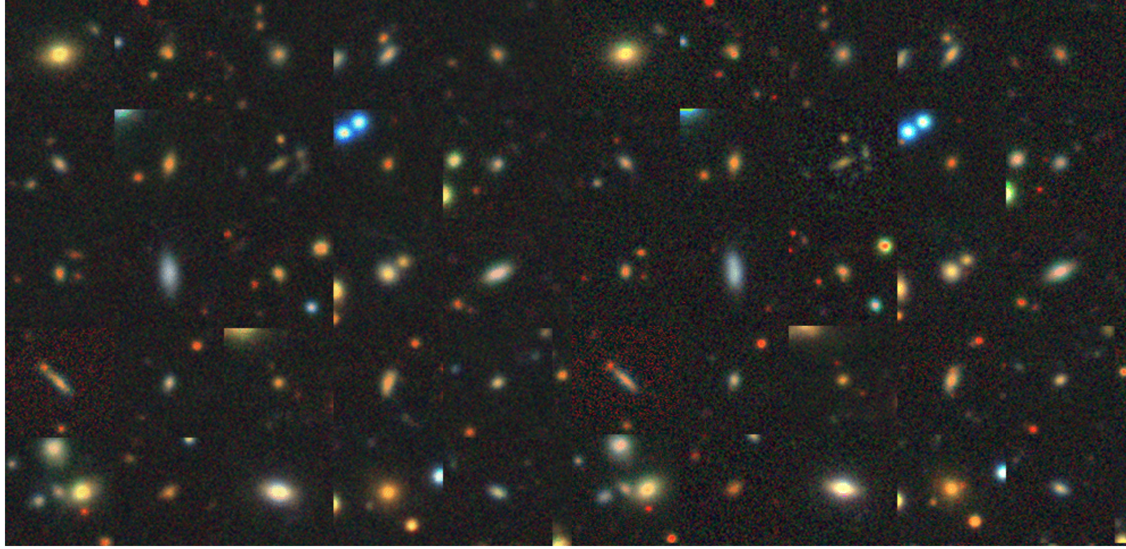
Overview

The Multimodal Universe dataset is a large scale collection of multimodal astronomical data, including images, spectra, and light curves, which aims to enable research into foundation models for astrophysics and beyond.

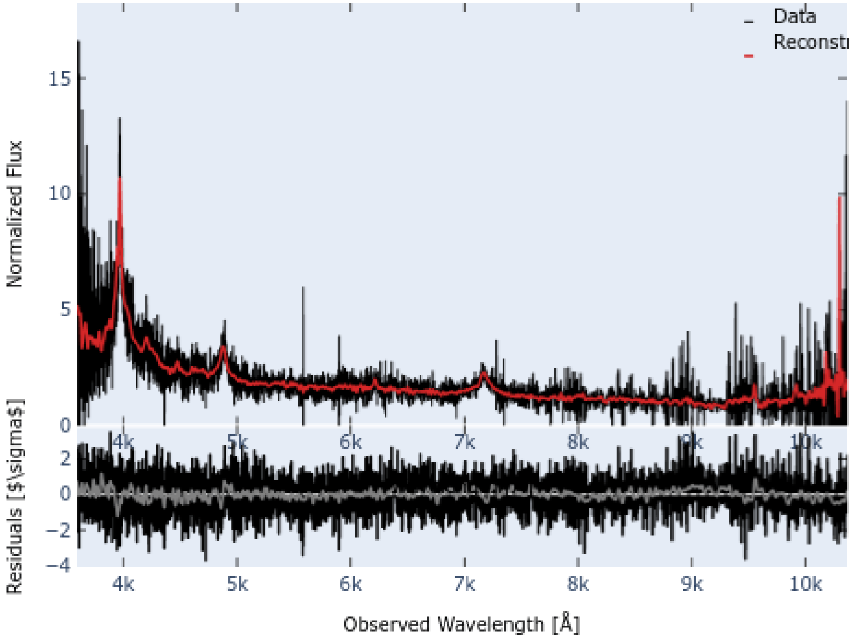
Scientific Data Tokenization



Mizrahi et al. (2023)



Input Reconstructed

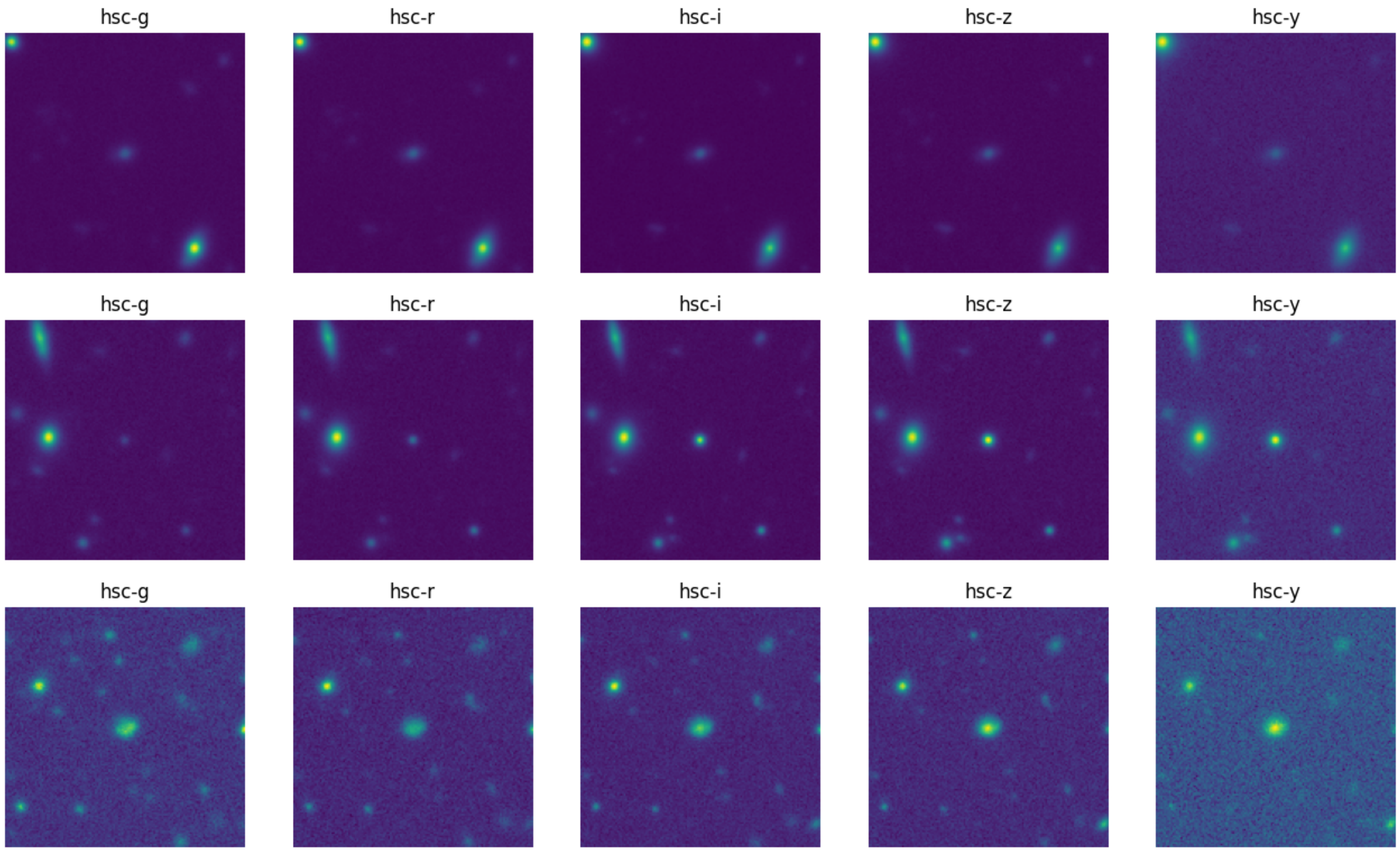


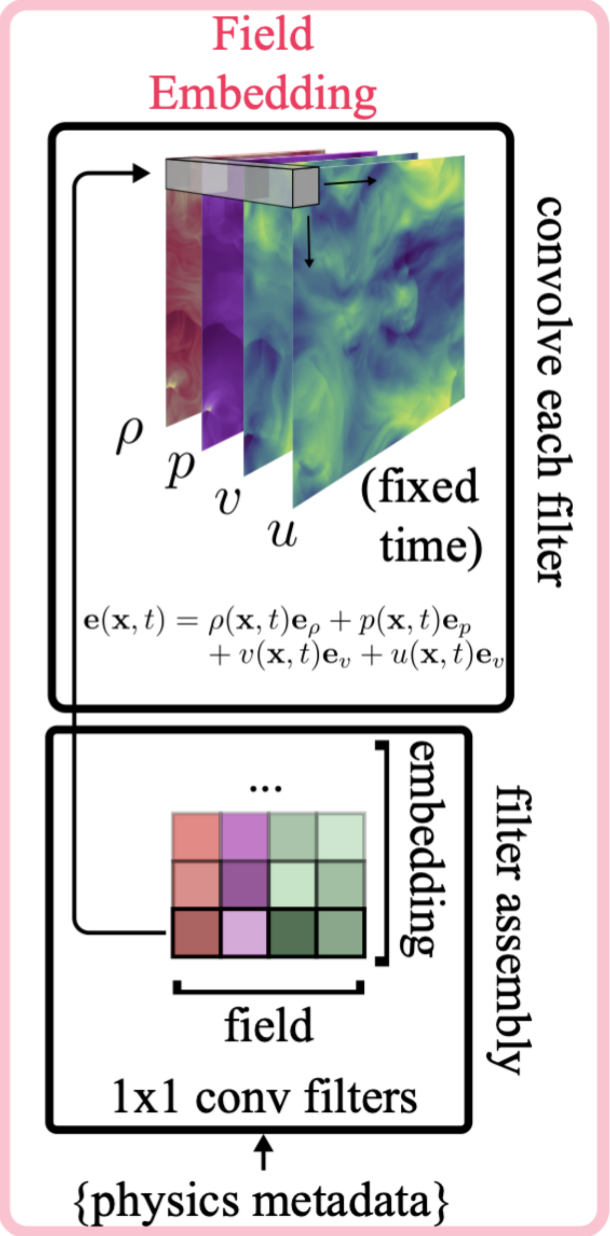
Our strategy:

- Develop **modality specific but universal tokenizers**, i.e. a single model to embed all type of astronomical images

- This requires specific innovations to take into account the metadata of observations.

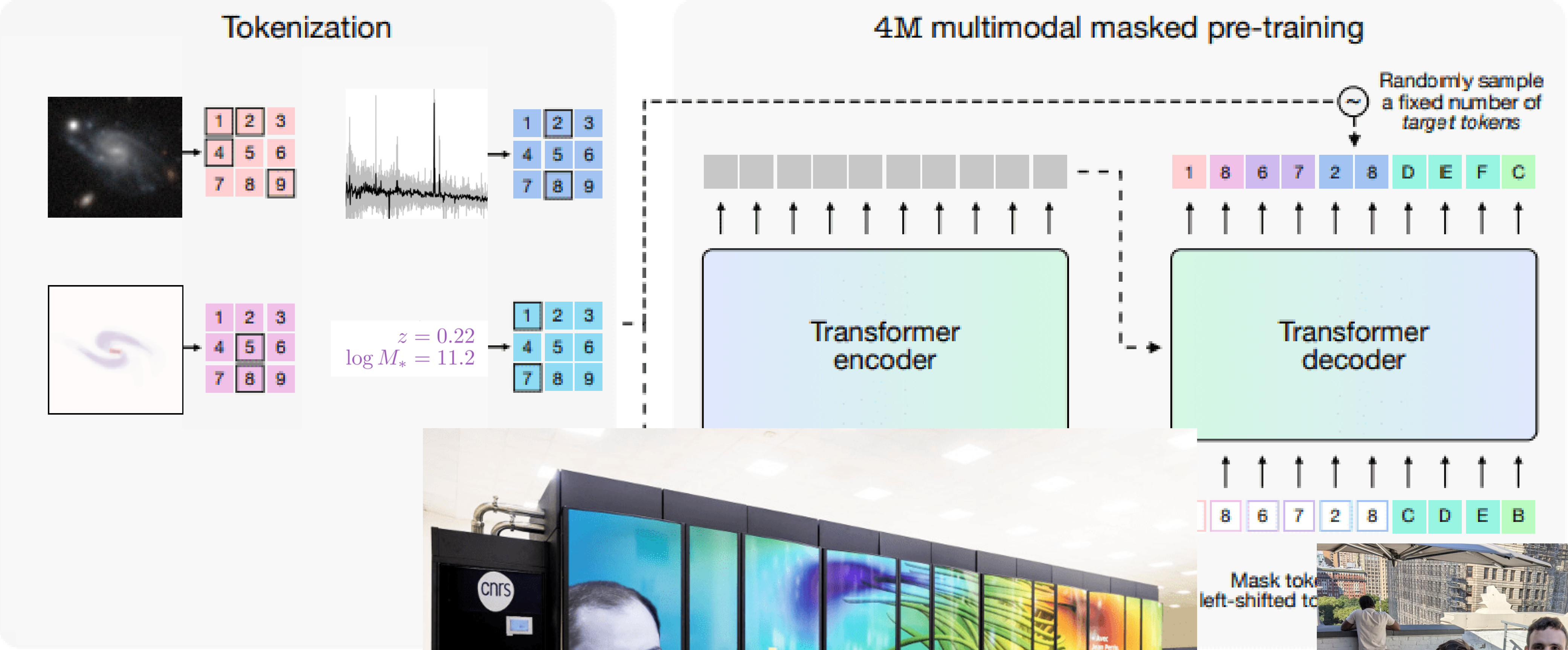
Example of strategy to embed different bands





Field Embedding Strategy Developed for Multiple Physics Pretraining (McCabe et al. 2023)

Next Step: Any-to-Any Modeling on Scientific Data



Jean Zay engineering team visiting Flatiron for a hackathon

- Learns the joint and all cond modalities: $\forall m, n$
- Can be further fine-tuned t

- Next year we are focusing on scaling up (more domains, more data, larger models) and developing the next generation of our models.

- **We are hiring!**
  - Postdoctoral positions
  - Research engineer positions

Polymathic

Follow us online!

Thank you for listening!