

AION-1: Omnimodal Foundation Model for Astronomical Sciences

Liam Parker^{*,1,2,3,4}, Francois Lanusse^{*,5,2}, Jeff Shen^{*,6}, Ollie Liu⁷, Tom Hehir⁸, Leopoldo Sarra³, Lucas Meyer³, Micah Bowles⁹, Sebastian Wagner-Carena^{2,3}, Helen Qu², Siavash Golkar^{2,3}, Alberto Bietti², Hatim Bourfoune¹⁰, Nathan Cassereau¹⁰, Pierre Cornette¹⁰, Keiya Hirashima^{2,11}, Geraud Krawezik², Ruben Ohana², Nicholas Lourie³, Michael McCabe^{2,3}, Rudy Morel², Payel Mukhopadhyay^{1,8}, Mariel Pettee¹², Bruno Regaldo-Saint Blancard², Kyunghyun Cho³, Miles Cranmer⁸, Shirley Ho^{2,3,6}

¹University of California, Berkeley, ²Flatiron Institute, ³New York University, ⁴Lawrence Berkeley National Laboratory,

⁵Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, ⁶Princeton University, ⁷University of Southern California, ⁸University of Cambridge, ⁹University of Oxford, ¹⁰IDRIS, CNRS, ¹¹RIKEN Center for iTHEMS, ¹²University of Wisconsin-Madison

*Equal Contribution.

While foundation models have shown promise across a variety of fields, astronomy still lacks a unified framework for joint modeling across its highly diverse data modalities. In this paper, we present AION-1, a family of large-scale multimodal foundation models for astronomy. AION-1 integrates heterogeneous imaging, spectroscopic, and scalar data using a two-stage architecture: modality-specific tokenization followed by transformer-based masked modeling of cross-modal token sequences. The model is pretrained on five large-scale surveys: Legacy Survey, Hyper Suprime-Cam (HSC), Sloan Digital Sky Survey (SDSS), Dark Energy Spectroscopic Instrument (DESI), and Gaia. These span more than 200 million observations of stars, galaxies, and quasars. With a single frozen encoder, AION-1 achieves strong results on a broad suite of downstream tasks, including galaxy and stellar property estimation, galaxy morphology classification, similarity-based retrieval, galaxy image segmentation, and spectral super-resolution. We release AION-1 model variants ranging from 300 M to 3.1 B parameters. Beyond astronomy, AION-1 provides a scalable blueprint for multimodal scientific foundation models that can seamlessly integrate noisy, instrument-specific observations. All code, tokenizers, pretrained weights, and a lightweight evaluation suite are released under an open-source license.

Date: October 22, 2025

Correspondence: Liam Parker: lharker@berkeley.edu; Francois Lanusse: francois.lanusse@cnrs.fr

Code: <https://github.com/PolymathicAI/AION/>

1 Introduction

Foundation models have transformed natural language processing and computer vision (Achiam et al., 2023; Dubey et al., 2024; Gemini Team et al., 2024). However, this class of models has not been fully explored in scientific domains where data are often complex and heterogeneous, combining multiple instruments, measurement protocols, and noise sources unique to real-world experiments. As a result, many scientific analyses employ bespoke models that treat each modality in isolation or rely on strict - often hand-crafted - schemas for cross-modal data fusion.

Within the broader scientific landscape, astronomy provides a particularly compelling testbed for the development of multimodal scientific foundation models owing to both the volume of publicly available data and its extraordinary diversity of measurements. Indeed, recent works have begun to explore multi-modal foundation models in astronomy (Mishra-Sharma et al., 2024; Parker et al., 2024; Rizhko and Bloom, 2024; Zhang et al., 2024); however, these approaches have been limited to single physical phenomena and relied primarily on contrastive objectives, which face fundamental limitations including generalization to arbitrary modalities and difficulty in capturing information beyond the mutual information between modalities.

In this paper, we introduce AION-1 (Astronomical Omni-modal Network), a large-scale multimodal foundation model for astronomy designed to handle arbitrary numbers of modalities across multiple physical phenomena. AION-1

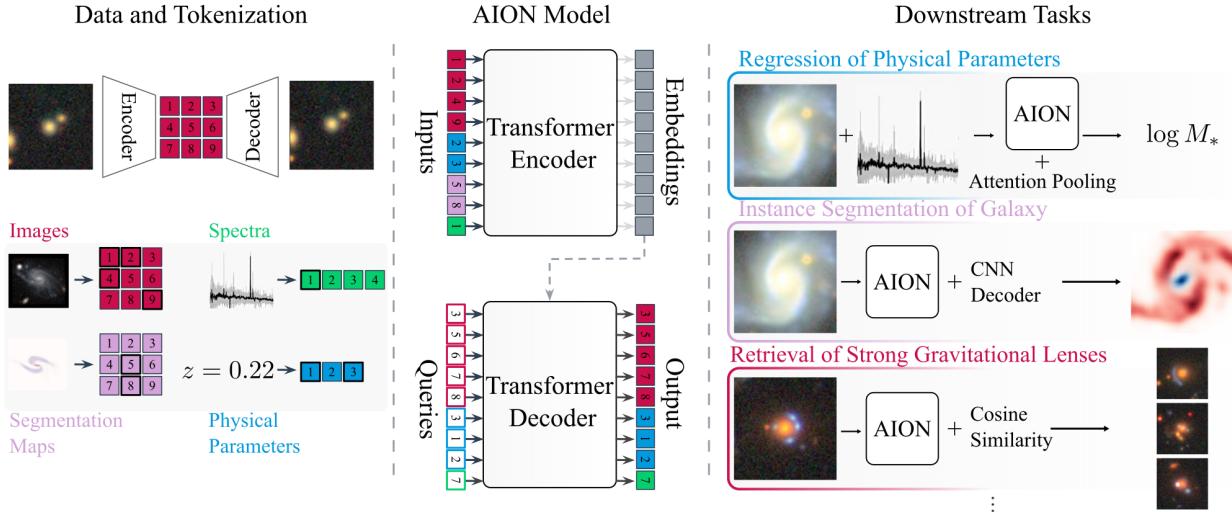


Figure 1: AION-1 integrates 39 different data modalities — multiband images, optical spectra, and various properties and measurements — into a single model usable for a wide range of downstream applications. It implements a two-step process: first, bespoke tokenization strategies that homogenize the diverse scientific data, followed by multimodal masked modeling that learns how different observations relate, inducing a deep understanding of the underlying physical objects. Astronomers can then leverage AION-1’s rich astrophysical understanding for a variety of downstream tasks.

unifies imaging, spectroscopy, photometry, and other object-level measurements from major ground- and space-based observatories into a single model for galaxies, stars, and quasars. By bridging these disparate data types, AION-1 addresses a key challenge in scientific machine learning: the integration of multiple heterogeneous datasets spanning different instruments, measurement protocols, noise sources, and physical phenomena into a single, unified framework.

At the heart of AION-1 lies a two-step approach: **Universal Tokenization of Diverse Data**, where we homogenize real-world scientific observations with discrete quantization across different data types, instruments, and observatories, followed by **Multimodal Masked Modeling**, where we train a single transformer encoder-decoder with a masked-token objective over all modalities simultaneously. Once trained, we demonstrate emergent behaviors in the AION-1 models that reflect the potential for multimodal scientific foundation models to capture non-trivial physical insights from raw data alone:

- **Emergent Physical Understanding.** AION-1 can solve non-trivial scientific tasks using only a simple linear head on top of its learned representations.
- **Superior Performance in the Low-Data Regime.** AION-1 can achieve competitive results on downstream inference tasks even with orders of magnitude less data than its supervised counterparts.
- **Flexible Data Fusion.** AION-1 can use arbitrary combinations of observations, enabling seamless data fusion on downstream tasks as well as cross-modal conditional generation.
- **Physical Structure of the Latent Space:** AION-1’s embedding space organizes objects along physically meaningful directions, enabling powerful retrieval of rare observations that surpasses current state-of-the-art retrieval methods in astronomy.

Beyond astronomy, the data tokenization strategies, masked modeling, and cross-modal generation strategies introduced address key challenges in real-world scientific data—namely, heterogeneity, noise, and instrument-specific idiosyncrasies. Moreover, by focusing on purely observational data, our approach is applicable in any data-rich field, even when strong physical models are not available.

1.1 Contributions

In summary, we present the following contributions:

- We present AION-1, a family of token-based multimodal scientific foundation models ranging in size from 300M to 3.1B parameters. AION-1 is a large-scale model designed for arbitrary combinations of highly heterogeneous scientific observations.
- We develop bespoke tokenization methods to homogenize a wide variety of astronomical data into a single coherent corpus. These innovations address the heterogeneity, noise, and instrument-specific peculiarities that challenge standard scientific modeling.
- We demonstrate that AION-1 achieves competitive to state-of-the-art performance on a broad range of scientific tasks with even simple probing, while significantly outperforming supervised baselines in low-data regimes, rendering the model highly usable by downstream researchers even without dedicated finetuning.

By tackling the challenges of data heterogeneity, noise, and diverse instrumentation, AION-1 offers a promising paradigm for future multimodal foundation models beyond astronomy, setting the stage for a new era of large-scale, cross-domain scientific exploration.

2 Related Work

Multimodal foundation models have become a cornerstone of modern self-supervised learning (Achiam et al., 2023; Anthropic, 2024; Dubey et al., 2024; Gemini Team et al., 2024; Liu et al., 2023; StabilityAI, 2022). Indeed, recent advances like GPT-4V (Achiam et al., 2023), Claude 3 (Anthropic, 2024), and LLaVA (Liu et al., 2023) have achieved human-level performance in visual reasoning, while models like Imagen (Saharia et al., 2022) and Stable Diffusion (StabilityAI, 2022) have enabled high-quality image generation from text. However, these models primarily rely on language to bridge modalities, which is often unavailable for scientific data. Recent work on early-fusion models, such as Chameleon (Team, 2024), 4M (Mizrahi et al., 2023), or PercieverIO (Jaegle et al., 2022), have demonstrated promising alternatives by learning mappings between modalities.

While these methodological advances in foundation models have transformed many fields, astronomy presents unique challenges, including heterogenous instruments, measurement protocols, and noise. As such, astronomy-specific efforts have emerged. For example, supervised pre-trained models like Zoobot (Walmsley et al., 2024) have leveraged 100M human annotations for galaxy morphology prediction obtained through extensive citizen science campaigns. Large-scale, self-supervised approaches trained on single-modal data have also emerged, including transformer-based models for Gaia stellar data (Leung and Bovy, 2024), APOGEE spectra (Koblischke and Bovy, 2024) and astronomical images (Smith et al., 2024) and contrastive approaches for astronomical images (Hayat et al., 2021; Stein et al., 2021, 2022). Finally, recent multimodal contrastive approaches have been introduced, starting with galaxy image-spectra pairs in Parker et al. (2024) and followed by galaxy images and text (Mishra-Sharma et al., 2024) and time-series and photometry (Rizhko and Bloom, 2024; Zhang et al., 2024).

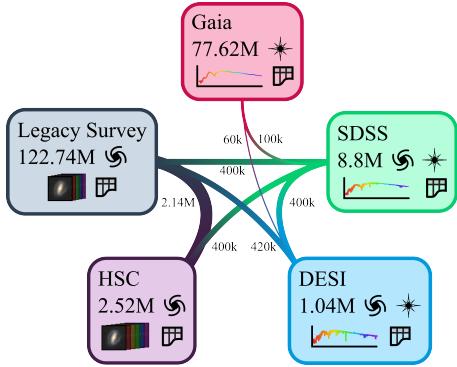
Relative to these methods, AION-1 represents an advance in both scale and scope: we train a multimodal model to billion-parameter scales and attempt to unify arbitrary modalites or different object types; in this case, 39 modalities across 200 million unique measurements spanning galaxies, stars, and quasars.

3 Data

AION-1 is pretrained on the publicly available data from The Multimodal Universe Collaboration et al. (2024) (hereafter MMU), a large-scale dataset of ML-ready, multimodal astronomical data. We use five surveys: Hyper Suprime-Camera (HSC) (Aihara et al., 2018) and Legacy Imaging Survey (Dey et al., 2019) for galaxy images; DESI (Aghamousa et al., 2016) and SDSS (York et al., 2000) for high-resolution spectra and cosmic distances on galaxies and stars; and Gaia (Collaboration et al., 2016) for low-resolution spectra and precise photometric and astrometric measurements for stars in the Milky Way. The relative contribution and modalities present in each survey, along with the size of the cross-matches,

is illustrated in [Figure 2](#).

AION-1’s pretraining emphasizes learning relationships between diverse observations of the same astronomical objects. Unlike 4M ([Mizrahi et al., 2023](#)), which requires all modalities simultaneously, we use pairwise associations across surveys. This flexibility accommodates uneven associations, and even enables the model to understand relationships between unpaired measurements, demonstrating transfer behavior. Below, we provide a brief overview of the details and data types from each survey, but refer the reader to [The Multimodal Universe Collaboration et al. \(2024\)](#) or the original source dataset paper for more exhaustive descriptions. We note that the per-survey magnitude/quality thresholds, footprint choices, Gaia XP availability, and our reciprocal cross-match together define the effective selection function of the pre-training corpus; this shapes which morphologies, redshifts, environments, and S/N regimes might be emphasized in AION-1’s embeddings (see [section 8.1](#)).



Survey	Modalities
Legacy Survey	4-band images; g, r, i, z fluxes; WISE $W1-W4$ fluxes; $E(B-V)$ extinction; ellipticity (e_1, e_2); half-light radius R_{eff}
HSC	5-band images; g, r, i, z, y fluxes and extinction; shape catalog ($\gamma_{11}, \gamma_{12}, \gamma_{22}$)
SDSS	Optical spectrum; redshift (z)
DESI	Optical spectrum; redshift (z)
Gaia	BP/RP spectra; parallax ϖ ; sky coords (α, δ); fluxes G, R, I, Z

Figure 2: AION-1 pre-training inputs. Left: visual representation of the various surveys used during pretraining, their total volume (in number of objects), and the size of the cross-matches between datasets. Right: A more detailed break-down of the modalities provided by each survey.

3.1 Legacy Surveys

The DESI Legacy Imaging Surveys combine three wide-field programs on the Blanco, Mayall, and Bok telescopes, delivering uniform $\{g, r, i, z\}$ imaging over $\sim 20\,000 \text{ deg}^2$ —roughly half the sky. The galaxies are imaged at a pixel scale of 0.262 arcsec. MMU provides 160×160 pixel postage stamp cut-outs centered on galaxies from the Data Release 10 ([Dey et al., 2019](#)), which we crop to 96×96 images; we use only objects in the Southern Galactic Cap, and retain objects with $\text{mag_z} < 21$ that pass the MMU quality cuts, corresponding to roughly 122 million galaxies.

Modalities Each object is packaged as: (1) calibrated $\{g, r, i, z\}$ integrated fluxes and their inverse-variance estimates, (2) mid-IR fluxes $W1-W4$ from WISE, (3) Milky-Way reddening $E(B-V)$, and (4) basic shape/size descriptors—the ellipticity components (e_1, e_2) and circularised half-light radius R_{eff} —derived from the Legacy tractor model fits.

3.2 Hyper Suprime-Cam (HSC)

The Hyper Suprime-Cam Subaru Strategic Program delivers deep, high-resolution $\{g, r, i, z, y\}$ imaging over $\sim 1\,200 \text{ deg}^2$. We use only the **wide** subset from PDR3 ([Aihara et al., 2018](#)). From the co-added calexps frames, MMU extracts 160×160 -pixel cut-outs at a pixel scale of 0.162 arcsec centered on catalog sources, which we crop to 96×96 , as with the Legacy Survey images. Objects are kept when they satisfy: $\text{mag_i} < 22.5$; at least three visits in every band (“full-depth full-colour”); and the standard HSC quality flags remove bright-star contamination, edge artefacts, saturation and unreliable cmodel photometry. The resulting sample contains roughly 2.5 million galaxies.

Modalities Each object is packaged as: (1) calibrated $\{g, r, i, z, y\}$ integrated fluxes and their inverse-variance estimates, (2) PSF-homogenised forced photometry in each band with extinction corrections, and (3) the moment-based SDSS shape tensor components ($\gamma_{11}, \gamma_{12}, \gamma_{22}$) computed by the HSC pipeline.

3.3 Sloan Digital Sky Survey (SDSS)

The Sloan Digital Sky Survey (SDSS) (Ahumada et al., 2020) has obtained medium-resolution ($R \sim 2\,000$) optical spectra for millions of objects. We use the aggregated public optical spectra from the Legacy, SEGUE-1/2, BOSS and eBOSS programs¹, covering 3650–10\,400Å with resolutions of $R = \lambda/\Delta\lambda = 1,500$ at 3,800Å and $R = 9,000$ at 9,000Å. We keep only primary, science-target spectra from plates flagged PLATEQUALITY='good'. This yields ~ 4 million galaxies and stars.

Modalities. Each object is packaged as: (1) the optical spectrum, its inverse-variance estimates, and its wavelength and (2) the pipeline redshift.

3.4 Dark Energy Spectroscopic Instrument (DESI)

The Dark Energy Spectroscopic Instrument (DESI Collaboration et al., 2016) survey is collecting spectra for ~ 40 million galaxies and quasars; MMU presently ingests the Early Data Release (EDR, 1% of the full survey) (DESI Collaboration et al., 2024). Each spectrum spans 3\,600–9\,800Å on a fixed 7,081-pixel grid at resolutions of $R = 2\,000$ at 3,600Å and $R = 5,500$ at 9,000Å and is distributed with flux, wavelength and inverse-variance arrays. We select spectra from the SV3 “one-percent” survey where SV_PRIMARY is true, OBJTYPE='TGT' and COADD_FIBERSTATUS=0, giving roughly 1 million galaxies, stars, and quasars.

Modalities. Each object is packaged as: (1) the optical spectrum, its inverse-variance estimates, and its wavelength and (2) the pipeline redshift.

3.5 Gaia

Gaia DR3 (Collaboration et al., 2016) provides low-resolution prism spectra from its blue (BP) and red (RP) photometers for 220 million Milky-Way sources in addition to precise astrometry and broad-band photometry. MMU stores each BP/RP spectrum as the 110 Gauss–Hermite coefficients released by the mission (55 BP + 55 RP), which can be resampled onto an 1\,101-pixel wavelength grid via GaiaXPy. We include all DR3 objects that have a mean BP/RP spectrum, retaining the full set of associated photometric, astrometric and stellar-parameter metadata.

Modalities. Each object is packaged as: (1) The 110 BP/RP spectral coefficients, (2) four-parameter astrometry (sky coordinates and parallax), and (3) mean fluxes in the G^2 , BP and RP bands.

3.6 Cross-matching strategy

For each pair of surveys we perform a nearest-neighbour match within a 1 arcsec radius on the sky and keep only reciprocal matches. Every resulting match is materialised as its own dataset. Each object in these datasets therefore aggregates all modalities from both parent surveys so that a single file read yields a fully fused, multi-survey view of the same astrophysical object. During AION-1 pre-training we draw samples both from the individual survey datasets and from these cross-matched sets, as detailed in the next sections. We note that this procedure may preferentially retain bright, isolated, well-centered sources and may de-emphasize blended or offset systems, introducing a further selection effect on the joint training distribution (see section 8.1).

4 Tokenization of Astronomical Data Modalities

Tokenization in AION-1 transforms heterogeneous data into a unified, transformer-compatible representation. Astronomical datasets present two key challenges: the variety of data types (2D images, 1D spectra, scalar values) and the diversity of sources within each type (different telescopes, resolutions, and instrument formats). We address this through modality-specific tokenizers that provide intra-modality standardization; each modality uses a dedicated tokenizer capable of handling multiple instruments, ensuring aligned representations within each data type. Moreover, the need to train multiple tokenizers for a modality with multiple survey inputs is removed.

¹We note that the SDSS and BOSS instruments have different fiber aperture sizes, but in the present work we include them in the same dataset.

² G is the mission’s very broad “white-light” band measured by the astrometric field CCDs.

4.1 Multi-Survey Image Tokenizer

4.1.1 Preprocessing

Each input from an imaging survey provides (i) a per-band flux map \mathbf{x} , (ii) a pixel-wise inverse variance map Σ , and (iii) a per-pixel mask \mathbf{m} for a given source. Our tokenizer ingests heterogeneous measurements drawn from both HSC (five filters $\{g, r, i, z, y\}$) and the Legacy Survey SGC (LS; four filters $\{g, r, i, z\}$). The two pipelines vary in central wavelength, pixel scale, zero-point, and noise. Therefore, we treat all bands from the surveys as distinct from each other; i.e. g from Legacy Survey is treated as a different band than g from HSC. We stack all distinct bands into a single fixed set of 9 channels (5 from HSC and 4 from LS), assigning a specific index to each channel. Next, we map every image into a 9-channel tensor, filling the subset of channels corresponding to that image's bands with flux values and setting any unused channels to zero; a binary mask $\mathbf{m}_c \in \{0, 1\}^C$ tracks which bands are zeroed out. The result of this process is that all images drawn have the same dimension, and can be stacked into a single, heterogenous batch, while maintaining survey-specific provenance information. We then normalize the zero-points between surveys by rescaling HSC to the Legacy Survey zero-point of 22.5 mag via $s = 10^{(\text{ZP}-22.5)/2.5}$, and multiply by the ratio of pixel scales. While these steps are not strictly necessary - as the bands are already separated above - we find that it helps with training stability. Finally, we apply an arcsinh normalization to the images to account for their high dynamic range, which we invert before computing the autoencoding loss. We find that adequate range compression is crucial for training stability.

4.1.2 Architecture and quantization.

Subsampled Linear Projection Given a batch of images, $\mathbf{x} \in \mathbb{R}^{B \times C \times H \times W}$ (batch by channels by height by width), we project it to a higher-dimensional space of size $\text{dim}_{\text{out}} \approx 6C$ using

$$\hat{\mathbf{x}} = \alpha(\mathbf{m}_c)(\tilde{\mathbf{x}}W + b),$$

with learnable $W \in \mathbb{R}^{C \times \text{dim}_{\text{out}}}$ and $b \in \mathbb{R}^{\text{dim}_{\text{out}}}$. The scale factor $\alpha(\mathbf{m})$ keeps the feature norm invariant to missing channels. The projection is inverted after decoding. We introduce the subsampled linear projection to expand each image into a higher-dimensional embedding that disentangles survey-specific channel information while preserving feature norms even when some bands are missing.

Autoencoder Once subsampled, we feed the output of the subsampled linear projection, which is now a 54-dimensional image, into a ResNet-based autoencoder. Specifically, we use the MagViT architecture adapted from [Yu et al. \(2023a\)](#), in which we remove transformer blocks. The encoder therefore consists of 2 downsampling ResNet blocks, which reduce the dimensionality of the input image by a factor of 16, resulting in a latent space that is $24 \times 24 \times 512$; this is compressed to $d = 4$ dimensions before being fed to the quantizer. The output of the quantizer is then projected back to 512 dimensions, before being upsampled in the decoder. In total, the ResNet-based autoencoder has roughly 50M learnable parameters.

Quantizer At the bottleneck of the tokenizer, we quantize features into a discrete set of codes. We experiment with multiple approaches, but empirically, we find that Finite Scale Quantization (FSQ; [Mentzer et al., 2023](#)) yields the best performance in terms of reconstruction fidelity and training stability. Further, to explore the trade-off between reconstruction loss and codebook utilization, we vary the codebook size from smaller (e.g., 2^4) to larger (e.g., 2^{14}) - following the recommended configurations in the FSQ paper - and observe that a size of 2^{12} offers a desirable balance: the reconstruction loss plateaus with larger codebooks, while code usage remains sufficiently high to avoid underfitting with smaller codebooks; see [Figure 3](#) for reference. Consequently, our final configuration employs FSQ with codebook levels of $n_i = \{8, 5, 5, 5\}$, equating to a rough size of 2^{12} codes.

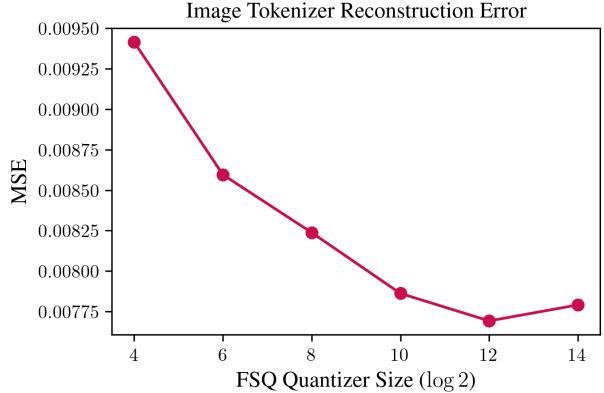


Figure 3: Tokenizer Codebook Scaling: We present the image tokenizer reconstruction error (MSE) as a function of FSQ quantizer size. We choose 2^{12} as our ultimate codebook size for images, as its reconstruction loss appears to plateau around this point.

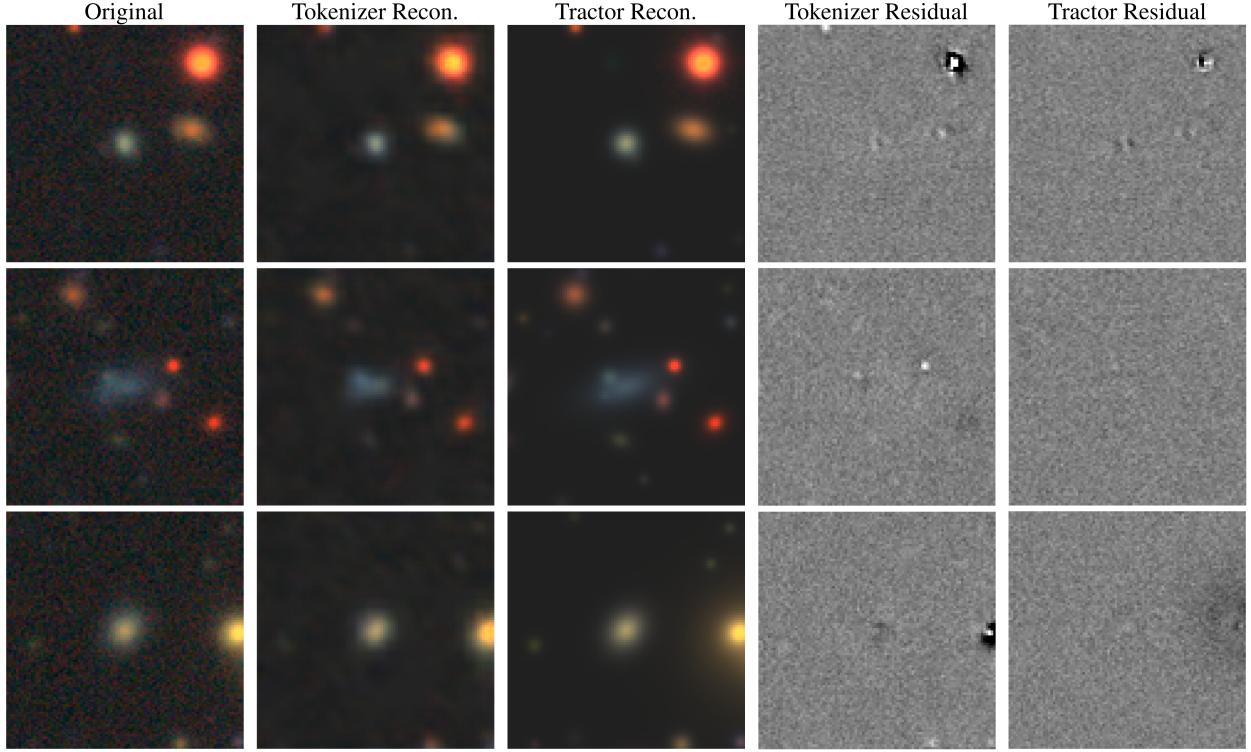


Figure 4: Image Tokenizer Performance: Reconstruction quality of the image tokenizer on three representative Legacy Survey images. The columns show, left to right, the original image, the reconstruction from the tokens, the reconstruction from the Legacy Survey tractor, the r -band residual for the tokenizer, and the r -band residual for the tractor.

4.1.3 Loss Function and Per-band Weighting

The tokenizer is trained using an inverse-variance-weighted Gaussian negative log-likelihood (NLL) that leverages our prior knowledge of the noise properties in each image, as reported by the data-generation pipelines. The NLL is given by:

$$\mathcal{L}_{\text{NLL}} = \sum_i \frac{1}{2} \| \Sigma_i^{-\frac{1}{2}} \mathbf{m}_i (\mathbf{x}_i - \text{Dec}_\theta(\text{Enc}_\phi(\mathbf{x}_i))) \|_2^2 \quad (1)$$

where \mathbf{x}_i is the input image, Σ_i is the diagonal noise covariance provided by the imaging pipeline, accounting for background and shot noise from bright sources, and \mathbf{m}_i is the survey pipeline mask which removes masked pixels in the image.

4.1.4 Training Details

We train with the image tokenizer using the Adam (Kingma et al., 2020) optimizer with a learning rate of 5×10^{-4} on batches of 256 images, sampling LS:HSC at a 20:1 ratio to reflect the relative size of the two datasets. The learning rate is warmed up over 1k steps before being decayed for 400k steps using a cosine decay. Training converges in ~ 5 days on $4 \times$ NVIDIA H100 GPUs, yielding a final reconstruction score of $\mathcal{L}_{\text{NLL}} = 0.00775$. We show some representative samples of the tokenizer’s reconstruction quality in Figure 4, and include reconstructions from the Legacy Survey pipeline tractor for comparison.

4.2 Multi-Survey Spectrum Tokenizer

4.2.1 Preprocessing

Each input spectrum provides (i) observed-frame flux density per unit wavelength $\mathbf{f}(\lambda)$, (ii) inverse standard deviation $\text{istd}(\lambda)$, and (iii) a per-pixel mask $\mathbf{m}(\lambda)$. Our tokenizer ingests heterogeneous measurements drawn from both DESI and SDSS. For each survey, we compute a robust median flux $\bar{\mathbf{f}}$ (ignoring masked pixels), use a \log_{10} range compression, and

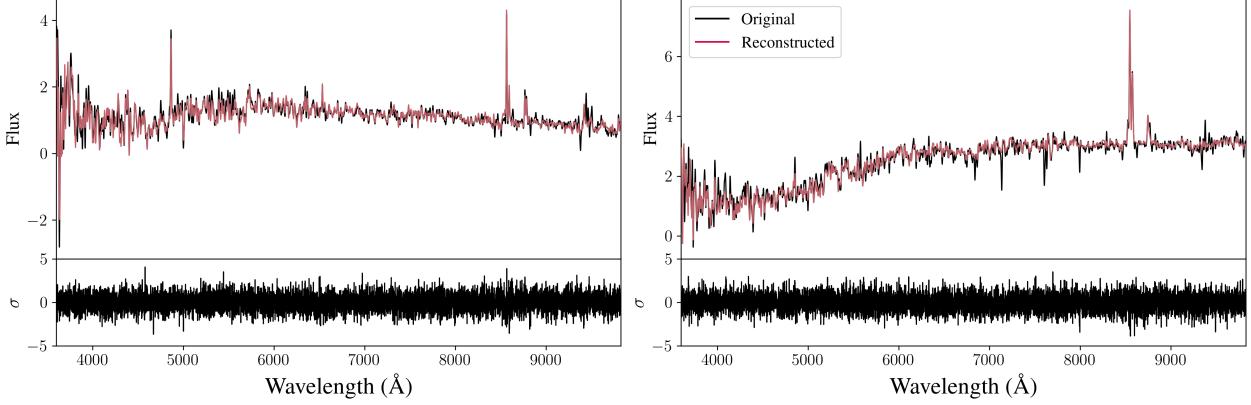


Figure 5: Spectrum Tokenizer Performance: Reconstruction quality of the spectrum tokenizer on two representative DESI EDR spectra. Top panels show the normalized flux as a function of wavelength, with original input (black) and reconstructed output (red) overlaid. Bottom panels show the residuals in units of the input uncertainty (σ).

quantize it with a 1-D scalar tokenizer (codebook size = 1024). We then normalize f and $istd$ by \tilde{f} and stack them into a 2-channel array. The array is linearly interpolated onto a fixed latent grid of 8704 points covering 3500–10462.4 \AA at 0.8 \AA spacing; this common grid is then shared between SDSS and DESI, and removes survey-specific wavelength/dispersion differences.

4.2.2 Architecture & Quantization

Autoencoder The stacked spectrum flux and inverse standard deviation $x \in \mathbb{R}^{B \times 2 \times 8704}$ is encoded by a 4-stage ConvNeXt-V2 backbone (Woo et al., 2023), consisting of an initial downsampling stack composed of a 4×4 convolution and LayerNorm, followed by three downsampling stacks of 2×2 convolutions and LayerNorms. Each of the four downsampling stacks is followed by multiple ConvNeXt V2 processing blocks. This compresses the spectrum into a 273×512 latent space, which is then further downsampled to 10 dimension before being fed to the quantizer; this dimensionality is chosen to conform to the 2^{10} codebook size used by the quantizer. Like with the image, these steps are inverted during the decoding part of the autoencoder.

Quantizer We use a Look-up-Free Quantizer (LFQ; Yu et al., 2023b) with an embedding dimension of ten (equating to a codebook size of 1024 codes) to convert the latent sequence into discrete codes. Contrary to the images, we find here that LFQ quantization slightly outperforms FSQ for the spectrum.

4.2.3 Losses

For each spectrum we project the decoder output back to its native wavelength grid and apply three losses:

1. **Flux likelihood.** Gaussian NLL weighted by inverse variance $w(\lambda)$, identical to Eq. (1).
2. **Mask accuracy.** Binary cross-entropy between the predicted reliability map $\hat{m}(\lambda)$ and the ground-truth mask $m(\lambda)$.
3. **Commitment.** LFQ commitment loss with weight $\beta_q = 0.25$.

4.2.4 Training Details

We train with using the AdamW optimizer with a constant 10^{-4} learning rate, a 0.1 weight decay penalty, and a global batch size of 128. Training for 215 k steps (\sim 24 hours on $4 \times$ NVIDIA H100) yields a token reconstruction $R^2 = 0.994$ and a mean mask AUC of 0.92. Reconstruction quality on two representative spectra from DESI are shown in Figure 5.

4.3 Scalar Tokenizer

While the scalar values could be quantized directly, equal width binning directly in the data space would lead to an uneven probability mass assignment and potentially imbalance training. Therefore, we map every scalar value to a unit normal Gaussian before quantization. To that end, we first need to tabulate the empirical cumulative distribution function F_x on the training set³. Once tabulated, we map a scalar value x_i to a standard normal variate via

$$z_i = \Phi^{-1}(F_x(x_i)), \quad (2)$$

where Φ^{-1} is the inverse CDF of $\mathcal{N}(0, 1)$. Because $z \sim \mathcal{N}(0, 1)$, equal-width binning in z -space allocates the same probability mass to every bin, automatically adapting to long tails or sharp peaks in the original distribution. Each Gaussianised scalar z_i is quantised independently with an FSQ codebook of $K = 1024$ centroids. Centroids are fixed a priori: we place them at the K equally spaced quantiles of the standard normal, i.e. $c_k = \Phi^{-1}((k - \frac{1}{2})/K)$. No parameters are learned and no loss is required. To recover an approximate scalar value \hat{x}_i from its token c_i ,

$$\hat{x}_i = F_x^{-1}(\Phi(c_i)).$$

With $K = 1024$ bins the median absolute reconstruction error is below typical measurement uncertainties, ensuring that tokenisation fidelity is sufficient for downstream tasks while keeping the representation compact and parameter-free. Note that for some of the scalars with large dynamic ranges, we also apply a \log_{10} or arcsinh transform before CDF mapping and tokenization. We apply the scalar tokenizer to the following scalars from each survey:

- **Legacy Survey:** $\{g, r, i, z\}$ fluxes, WISE W1-W4 fluxes, $E(B - V)$ extinction, ellipticity components (e_1, e_2), circularized half-light radius R_{eff} .
- **HSC:** $\{g, r, i, z, y\}$ fluxes, shape tensor components.
- **SDSS & DESI:** pipeline-reported redshift (z).
- **Gaia:** 110 BP/RP coefficients, parallax, sky coordinates (ra, dec), G , BP, RP fluxes.

4.4 Scalar Field Tokenizer

In addition to images, we included an additional tokenizer specialized for scalar maps with values in $[0, 1]$. This tokenizer is particularly adapted to handle segmentation maps, but could also be used to generate any property map scaled between 0 and 1, such as Star Formation Rate maps derived from Integral Field Spectroscopy.

4.4.1 Data & Preprocessing

The scalar field tokenizer was trained to autoencode a mixture of 5 categories of normalized single-channel images derived from Legacy Survey photometry: RGB cutouts converted to grayscale; individual red, green, and blue channels from the RGB cutouts; and an ‘object mask’ indicating the silhouettes of sources detected in each cutout. The object mask is generated from the Tractor model photometry included in the Legacy Survey data release. The Tractor classifies each detected source as one of 5 morphological types and fits a corresponding elliptical surface brightness model to the light emitted by the source. After fitting, parameters can be extracted from the surface brightness profile to define a centered ellipse enclosing 50% of the total emission from a given source. We generate an object mask for each cutout by painting such ellipses onto a null background for all sources detected in the cutout. The ellipses are filled with a constant value selected from $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ which corresponds to the morphological type.

4.4.2 Architecture & Quantization

Autoencoder We base our architecture on VQ-VAE (van den Oord et al., 2017); the encoder is comprised of a stack of 3 convolutional downsampling layers followed by 2 residual blocks, and this arrangement is mirrored in the decoder (with upsampling transpose convolutions replacing the downsampling convolutions). The downsampling convolutions have kernel size 4, stride 2, padding 1, and 128 / 256 / 512 kernels, respectively. Each residual block consists of a

³We estimate F_x with a fixed-size reservoir ($N \sim 10^6$ samples) maintained online during data streaming. Reservoir sampling (Algorithm R) produces an unbiased CDF while keeping memory $\mathcal{O}(N)$, independent of the full catalog size.

sequence of 2 convolutional layers separated by a batch norm layer, where the convolutions have kernel size 3 / 1, stride 1, and padding 1 / 0, respectively. All layers are ReLU-activated.

Quantizer We use a Finite Scale Quantizer (Mentzer et al., 2023) to quantize 4-dimensional codes with $n_{\text{dim}} = \{8, 5, 5, 5\}$ discrete levels available in the respective dimensions, yielding a codebook size of 1000.

4.4.3 Training Details

This tokenizer was optimized under a mean squared error objective using the AdamW optimizer with the weight decay parameter set to 0.01. The model was trained with batch size 256 for 114,000 steps at a base learning rate of 10^{-4} . The base learning rate was modified by a linear warmup phase in the first 1,000 steps and a cosine decay over the final 72,000 steps. The model weights were updated as an exponential moving average of previous values with a decay parameter of 0.9999. Under these conditions, the loss converged to $\mathcal{L}_{\text{MSE}} \approx 0.0017$.

5 Multimodal Masked Modeling

AION-1 is inspired by many who came before us, and builds on recent early fusion multimodal models (Girdhar et al., 2023; Jaegle et al., 2022; Team, 2024). In particular, it adopts the scalable multimodal masked modeling scheme proposed in 4M (Bachmann et al., 2024; Mizrahi et al., 2023) to learn from heterogeneous data (e.g., spectra, images, scalars) by randomly masking tokenized inputs across all available modalities and reconstructing the masked content.

Concretely, let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ be token sequences for M modalities available for a training example. During training, two disjoint subsets of \mathbf{X} are drawn: \mathbf{x}^{obs} (observed) and \mathbf{x}^{tgt} (target). Because these two subsets are sampled across the entire token pool, the model learns both intra- and cross-modal relationships in each training instance.

The loss for the model is then given by:

$$\mathcal{L}_{\text{4M}}(\theta) = - \sum_{t=1}^N \log p_{\theta}(\mathbf{x}_t^{\text{tgt}} \mid \mathbf{x}_t^{\text{obs}}), \quad (3)$$

where $p_{\theta}(\cdot \mid \mathbf{x}^{\text{obs}})$ is the categorical distribution over the predicted vocabulary, and N is the output token budget.

5.1 Architecture

We adopt a Transformer-based encoder-decoder framework suitable for the multi-modal masked prediction (see Figure 1). Beyond a standard encoder-decoder architecture, we emphasize below the modality specific embedding scheme needed at the input of both the encoder and decoder to implement our objective.

Concretely, each modality $i \in \{1, \dots, M\}$ has its own token embedding $\text{Embed}_i(\cdot)$, a learnable modality embedding \mathbf{m}_i , and positional embedding \mathbf{p}_t for token position t . Then, for an observed token from modality i , x_t^i , the full embedding is given by

$$\mathbf{e}_t^{(\text{enc})} = \text{Embed}_i(x_t^i) + \mathbf{m}_i + \mathbf{p}_t. \quad (4)$$

In the decoder, we feed information on the target we are querying tokens for, without providing their value:

$$\mathbf{e}_t^{(\text{dec})} = \mathbf{m}_i + \mathbf{p}_t, \quad (5)$$

omitting any direct lexical embedding $\text{Embed}_i(x_t)$.

In our implementation we use a different modality embedding \mathbf{m}_i for each modality and each source to identify the unique combination of data type and associated provenance metadata. In other words, two astronomical images from two different instruments will have two different modality embeddings even though they are both images. This is to provide the model with important provenance information which implicitly encodes aspects of data quality and resolution of the observations.

5.2 Modality Masking Strategy

A key consideration is to select which tokens become inputs (observed) vs. outputs (predicted) for each modality during training. We find that the dirichlet sampling from the original 4M implementation is inefficient when dealing with modalities that vary widely in length, and therefore results in a high frequency of mostly empty batches. Therefore, we follow a simplified approach:

Input Token Budget. We select a global input token budget B . To populate the budget, we first randomly pick one modality, and then uniformly randomly select a number of tokens for inclusion from that modality. We then fill the remaining budget B by uniformly sampling tokens from the other modalities.

Output Token Budget. For the remaining unselected tokens, we choose the number of tokens to predict for each modality by sampling from a Beta distribution skewed toward zero, which draws down the number of output tokens per sample, aligning with the eventual distribution of output tokens under a cosine schedule iterative sampling (e.g., in MaskGIT-style), ensuring that inference-time usage patterns are well covered during training. Similar to the input, one modality is chosen to draw an unconstrained number of tokens first, and the rest are filled by uniform random draws from the remaining modalities.

6 AION-1 Family of Models

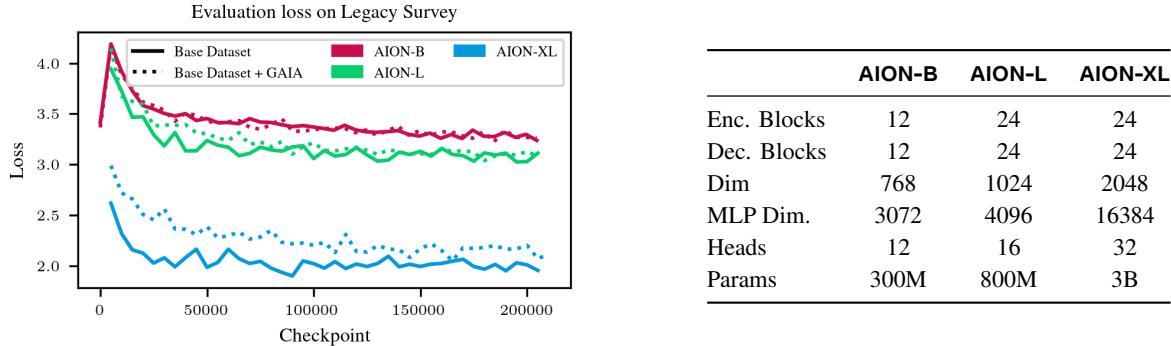


Figure 6: AION-1 Scaling Left: Legacy Survey test losses for three model sizes, with and without the Gaia stellar set. Note the increase in loss with the inclusion of Gaia. Right: AION-1 model variant breakdown of sizes, generally following the T5 model scaling convention (Raffel et al., 2020).

We train three model versions - Base (300M), Large (800M), and XLarge (3B) - using the AdamW (Loshchilov and Hutter, 2019) optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.05) for 205k steps with a global batch size of 8096. We use a linear warmup and cosine decay schedule, with a peak learning rate of 2×10^{-4} . We adopt an input budget of 256 tokens, and output budget of 128 tokens for all our models during pretraining. All models are trained with bfloat16 mixed precision, and model distribution under PyTorch’s Fully Sharded Data Parallel (FSDP) ZeRO-2 strategy. To achieve a batch size of 8192 in all cases, we train AION-1-B using 64 H100 GPUs for 1.5 days, AION-1-L using 100 H100 GPUs for 2.5 days, and AION-1-XL using 288 H100 GPUs for 3.5 days.

7 Evaluation on Downstream Tasks

This section evaluates AION-1’s performance within an exemplary set of astronomical workflows. Ultimately, we aim to demonstrate that AION-1’s streamlined foundation can significantly accelerate typical astrophysics tasks, facilitate data fusion, and provide superior results in low-data regimes, all while maintaining comparable or superior accuracy to typical supervised machine learning baselines. Although we present out-of-the-box generative capabilities, we propose using AION-1 primarily as a frozen backbone, as described in section 7.2.

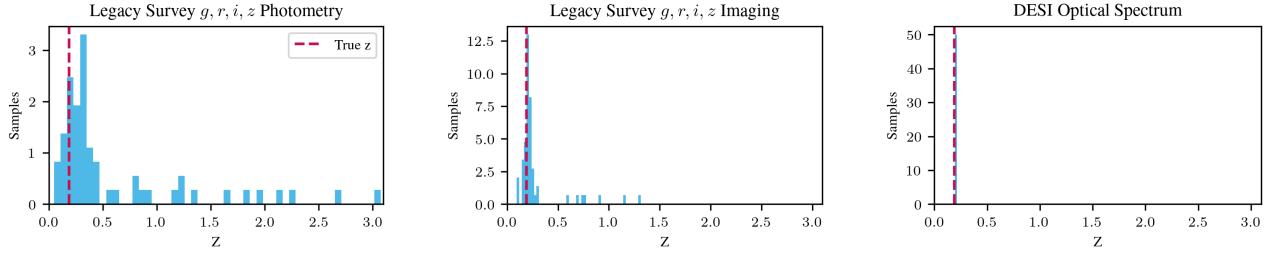


Figure 7: Redshift Posterior Estimation: Posterior samples for a single Legacy Survey galaxy under three conditioning scenarios. Left: broadband g, r, i, z photometry alone yields a broad $p(z)$. Middle: incorporating the corresponding 96×96 pixel multi-band cut-out (image+photometry) significantly tightens the credible interval. Right: conditioning on the high-resolution DESI spectrum yields a perfect redshift estimate.

7.1 Out-of-the-Box Capabilities

AION-1 is a generative model: once pretrained, it represents the joint distribution of all 39 tokenised modalities. At inference time we can therefore draw posterior samples of any modality in the training set by passing the appropriate query tokens and iteratively resampling them conditioned on the visible context. To generate these posteriors, we perform the following steps.

First, we pass the query modality through its appropriate tokenizer, producing a set of input tokens $\mathbf{x}^{\text{in}} = (x_1, \dots, x_{N_{\text{in}}})$. These are passed to the AION-1 encoder, while the decoder receives a sequence of query tokens $\mathbf{x}^{\text{qry}} = (x_{N_{\text{in}}+1}, \dots, x_N)$ whose values are to be inferred. At test-time we need samples from

$$p_\theta(\mathbf{x}^{\text{qry}} \mid \mathbf{x}^{\text{in}}) = \prod_{j \in \mathcal{Q}} p_\theta(x_j \mid \mathbf{x}^{\text{in}}), \quad (6)$$

where \mathcal{Q} indexes the query positions and p_θ is the categorical distribution produced by the frozen decoder. To perform this sampling, we follow the ROAR generation scheme introduced in 4M (Mizrahi et al., 2023): at each iteration t we

1. Draw a fresh random permutation $\pi_t : \mathcal{Q}_t \rightarrow \mathcal{Q}_t$ of the still-unknown query indices \mathcal{Q}_t ;
2. Reveal the first $\rho_t = \lfloor r^t |\mathcal{Q}_t| \rfloor$ positions of this permutation,

$$\mathcal{S}_t = \{\pi_t(1), \dots, \pi_t(\rho_t)\};$$

3. Sample those tokens once from the model,

$$x_j^{(t)} \sim p_\theta(\cdot \mid \mathbf{x}^{\text{in}} \cup \mathbf{x}_{\mathcal{Q}_{t-1} \setminus \mathcal{S}_t}^{(t-1)}), \quad j \in \mathcal{S}_t;$$

4. Promote them to inputs: $\mathbf{x}^{\text{in}} \leftarrow \mathbf{x}^{\text{in}} \cup \{x_j^{(t)}\}_{j \in \mathcal{S}_t}$ and update $\mathcal{Q}_{t+1} = \mathcal{Q}_t \setminus \mathcal{S}_t$.

With a decay factor $r \in [0, 1)$ the number of unresolved tokens drops exponentially, so the full sample is generated in $T = \mathcal{O}(\log |\mathcal{Q}|)$ decoder calls. After sampling is complete, every query token is routed back through its modality-specific tokenizer to recover the original data representation. Repeating the entire ROAR loop M times with a non-zero sampling temperature $\tau > 0$ yields M i.i.d. draws $\{\hat{\mathbf{x}}^{(k)}\}_{k=1}^M$ from the conditional distribution in Equation 6.

Importantly, we note that these draws are plausibility samples from the decoder’s categorical outputs under an iterative reveal schedule; they are not guaranteed to be well-calibrated joint posteriors for sequences of tokens longer than a single token. We discuss this limitation in further detail in section 8.1.

7.1.1 Redshift Estimation

Redshift z is one of the scalar channels quantized during pretraining, so the decoder can naturally output a categorical distribution over the 1,024 quantized redshift bins. In this example, we tokenize an input modality, and using the

scheme above, generate tokens (and corresponding redshifts) over 50 ROAR draws for each modality. Figure 7 displays posterior samples for a representative galaxy under three increasingly informative contexts: (1) Legacy Survey $\{g, r, i, z\}$ photometry only, (2) Legacy Survey $\{g, r, i, z\}$ photometry and multi-band imaging, and (3) high-resolution DESI spectra. It is clear how the posterior contracts as richer information is provided: starting from broadband Legacy Survey photometry, adding spatial morphology from multi-band imaging, and finally a full optical spectrum.

7.1.2 Spectral Super-Resolution

During pretraining, AION-1 learns to translate between multiple surveys, enabling conditional generation of one modality from another. Here, we demonstrate this capability by sampling query tokens corresponding to high-resolution DESI spectra while conditioning AION-1 on tokens from low-resolution GAIA BP/RP coefficients. Once sampled, the DESI tokens are passed through the spectrum decoder to produce realizations of the spectrum. Figure 8 illustrates this generation on a representative star: the red curve is the native Gaia spectrum, the black curve is the true DESI measurement, and the blue posterior samples trace the super-resolved features. AION-1 accurately recovers line centers, widths, and amplitudes within narrow posterior uncertainty bands, demonstrating its ability to impute high-frequency spectral structure from coarse observations.

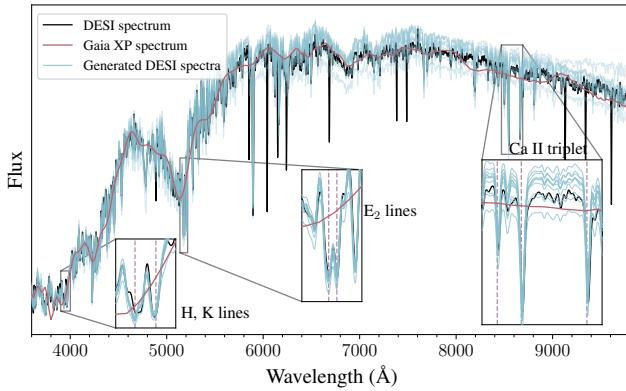


Figure 8: Spectral Super-Resolution: AION-1 can generate high-resolution posterior spectra samples (in this case, DESI; blue) conditioned on low-resolution spectra input (GAIA BP/RP coefficients; red), closely matching the ground-truth high-resolution measurements in line location, width, and amplitude. Several prominent lines are magnified in inset panels, with dashed lines marking their known locations.

Although the samples visually and quantitatively track key features, they should not be interpreted as calibrated joint posteriors over the full DESI spectrum; multi-token dependencies may be underrepresented by our current sampler as noted previously.

7.2 AION-1 Embeddings

AION-1’s primary practical benefit is its ability to produce powerful, physically meaningful, modality-agnostic embeddings that work out of the box for a wide range of tasks, avoiding the engineering and data costs of end-to-end supervised pipelines. At the same time, because foundation models carry an implicit prior from pre-training, we treat AION-1 as a frozen feature extractor and perform lightweight, task-specific calibration. In practice, we freeze the encoder, fit a small linear/MLP head on a representative calibration set that reflects the downstream selection function, and calibrate probabilities or continuous outputs. This workflow preserves AION-1’s representational power while letting researchers inject the scientifically relevant prior and maintain control over inference.

Extracting embeddings. To extract embeddings at inference time, we simply freeze the AION-1 encoder and discard the decoder. Given the contextualized vector sequence $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_T\}$, $\mathbf{z}_t \in \mathbb{R}^d$, produced by the input modality fed through its appropriate tokenizer, we form an object-level vector $\mathbf{e} \in \mathbb{R}^d$ with one of two pooling schemes:

1. Mean pooling:

$$\mathbf{e} = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t, \quad (7)$$

where the vectors are pooled deterministically by taking the average over all outlook vectors from the AION-1 encoder.

2. Attentive pooling:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right), \quad \mathbf{e} = \mathbf{A}\mathbf{V}, \quad (8)$$

where \mathbf{Q} , \mathbf{K} , \mathbf{V} are learnable query/key/value projections applied to \mathbf{Z} , which are learned online during model adaptation to the downstream task.

Both approaches yield a modality-agnostic embedding: the same encoder handles 2-D images, 1-D spectra, or scalar tokens without architectural changes. Moreover, AION-1 naturally supports multi-modality: one obtains a joint representation by concatenating the tokens from any subset of modalities and passing the union through the frozen encoder. No extra fusion module is required—the cross-modal context was learned during pre-training. Ultimately, these embeddings encode astrophysically meaningful structure as we demonstrate in the rest of this section.

7.2.1 Galaxy Parameter Estimation

For quantitative evaluation we adopt the PROBABilistic Value-Added Bright Galaxy Survey (PROVABGS; Hahn et al., 2023), which provides Bayesian spectral energy distribution (SED) fits for $\sim 140,000$ DESI Bright Galaxy Survey targets. We retain five global properties - redshift z , stellar mass M_* , stellar-population age t_{age} , gas-phase metallicity Z_{met} , and star-formation rate SFR - and cross-match the catalog against Legacy Survey DR10 $\{g, r, i, z\}$ imaging/photometry and DESI EDR spectra. Objects with $M_* < 0$ or non-physical magnitudes are discarded, leaving $\sim 120,000$ galaxies. To stabilize the dynamic range we predict $\log Z_{\text{met}}$ and $\log M_*$ and convert SFR to the specific rate $\text{sSFR} = \log(\text{SFR}/M_*)$.

Each input combination - photometry alone (Ph), photometry + imaging (Ph+Im), and photometry + imaging + spectra (Ph+Im+Sp) - is tokenised with the modality specific encoders described in §4.1–4.3. The resulting token sequence (which is simply stacked in the case of multimodal inputs) is passed through the frozen AION-1 encoder and compressed with a single learned cross-attention layer (2); a lightweight, two-layer multilayer perceptron (MLP, hidden size = 256, GELU) then maps the d -dimensional embedding to the target parameters. We quantify performance with the coefficient of determination, $R^2 = 1 - (\sum_j (y_j - \hat{y}_j)^2) / (\sum_j (y_j - \bar{y})^2)$, where y_j are the ground-truth values, \hat{y}_j the predicted values, and \bar{y} the mean of the ground-truth sample.

We train the cross-attention layer and the MLP probe with mean-squared error on 80% of the cross-matched sample and report the coefficient of determination R^2 on the held-out 20% split. We also benchmark against three modality-specific supervised networks trained end-to-end:

	z	M_*	t_{age}	$\log Z_{\text{Met}}$	sSFR
AION-1-B					
Ph	0.75	0.72	0.35	0.41	0.38
Ph+Im	0.93	0.89	0.45	0.49	0.64
Ph+Im+Sp	1.00	0.96	0.53	0.61	0.72
AION-1-L					
Ph	0.76	0.73	0.36	0.41	0.39
Ph+Im	0.94	0.89	0.45	0.50	0.64
Ph+Im+Sp	1.00	0.96	0.53	0.62	0.73
AION-1-XL					
Ph	0.79	0.76	0.31	0.38	0.48
Ph+Im	0.94	0.89	0.45	0.49	0.64
Ph+Im+Sp	0.99	0.95	0.53	0.62	0.73
Parker et al. (2024)					
Im*	0.78	0.73	0.29	0.36	0.42
Sp	0.99	0.90	0.52	0.60	0.70
Oquab et al. (2023)					
Im*	0.57	0.55	0.17	0.28	0.25
Supervised					
Ph ¹	0.71	0.69	0.30	0.30	0.38
Im ²	0.86	0.82	0.45	0.49	0.64
Sp ³	1.00	0.85	0.43	0.62	0.68

Table 1: R^2 (\uparrow) for galaxy property estimation. Inputs are photometry (Ph), photometry + imaging (Ph+Im), and photometry + imaging + spectra (Ph+Im+Sp). * AstroCLIP and DINOV2 use $\{g, r, z\}$ Legacy Survey images while AION-1 and supervised use $\{g, r, i, z\}$. Supervised models: ¹XGBoost, ²ConvNeXt, ³Conv + Attention network.

- **XGBoost on photometry**—a boosted trees regressor using calibrated fluxes as features;
- **ConvNeXt-Tiny on images**—the [The Multimodal Universe Collaboration et al. \(2024\)](#) vision baseline, trained from scratch on $\{g, r, i, z\}$ Legacy Survey 96×96 cut-outs;
- **Conv+Attention on spectra**—a 1-D CNN with gated attention pooling following [Melchior et al. \(2023\)](#) trained from scratch on the DESI optical spectra.

Additionally, we benchmark two strong self-supervised baselines. First, AstroCLIP ([Parker et al., 2024](#)), a previous state-of-the-art multimodal foundation model for galaxies; we follow the authors’ recommended protocol, extracting frozen embeddings from the CLIP image encoder and training a lightweight MLP ontop of the embeddings. Note that AstroCLIP was trained on Legacy Survey $\{g, r, z\}$ cut-outs only, so in our setting it has access to one fewer band (i) than AION-1. Second, DIONv2 ([Oquab et al., 2023](#)) represents a widely used vision model; we feed RGB-converted $\{g, r, z\}$ images to the ViT-g/14 backbone and again attach the same MLP probe. Full results are presented in [Table 1](#). We observe that across the board, AION-1 produces competitive results with minimal downstream adaptation required, and that its out-of-the-box multimodal fusion capabilities provide a powerful framework for downstream multimodal tasks.

7.2.2 Galaxy Morphology Classification

We consider here the problem of classifying galaxy images into ten distinct morphology classes (e.g. spiral arms, merging galaxies) defined by Galaxy Zoo 10 ([Leung and Bovy, 2018](#); [Walmsley et al., 2022](#), GZ10;). We construct the downstream sample by cross-matching the Galaxy Zoo 10 catalog with the Legacy Survey DR10 imaging footprint, yielding $\sim 8,000$ galaxies with $\{g, r, i, z\}$ cutouts. For AION-1, we tokenize each cutout with the multi-survey image tokenizer, mean-pool the resulting embeddings⁴ as in [Equation 1](#), and pass the 768-d mean vector to a two-layer MLP head (hidden size = 256, GELU, dropout = 0.1). The head is trained on 80% of the sample with class-stratified splits and evaluated on the remaining 20%.

Model	Accuracy (%)
AION-B	84.0
AION-L	87.2
AION-XL	86.5
Oquab et al. (2023)	71.4
EfficientNet	80.0
Walmsley et al. (2022)	89.6

Table 2: Galaxy Morphology Classification Accuracy (↑) on Galaxy Zoo 10. AION-1, DINOv2 ([Oquab et al., 2023](#)), and ZooBot ([Walmsley et al., 2022](#)) use an MLP head on frozen embeddings; EfficientNet-B3 ([Tan and Le, 2019](#)) is trained end-to-end from scratch.

We replicate this protocol with the DINOv2 baseline, replacing the tokenizer with the ViT-g/14 backbone and applying the RGB normalization recommended by [Oquab et al. \(2023\)](#). EfficientNet-B3 is trained end-to-end from random initialization using the same splits and standard data augmentations. Finally, we adapt ZooBot ([Walmsley et al., 2022](#)) by fine-tuning the penultimate layer on our 8,000 samples; although ZooBot was never exposed to GZ10 labels, it benefits from pre-training on $\sim 300,000$ images covering the broader, harder GZ-5 decision tree, and thus acts as an approximate upper bound on achievable accuracy.

As Table 2 shows, AION-1-L tops all baselines except ZooBot, exceeding EfficientNet by +7.2 pp and DINOv2 by +15.8 pp, while using only a lightweight MLP head. Moreover, it reaches close to the ZooBot accuracy, only under-performing by -2.4 pp, despite seeing two orders of magnitude fewer labeled galaxy images during its inference process.

7.2.3 Galaxy Image Segmentation

Going beyond broad morphology classification, we consider image segmentation based on human annotation of prominent galaxy structures obtained through the Galaxy Zoo 3D citizen science campaign ([Masters et al., 2021](#)). As above, we cross-match the sample with the Legacy Survey images, producing roughly 2,800 galaxy image-annotation pairs. We feed the $\{g, r, i, z\}$ cut-outs through the AION-1 model to produce embeddings using the mean-pooling⁵ aggregation scheme

⁴Although we experiment with attentive pooling in this setting, unlike with property estimation, we find that attentive pooling does not provide any meaningful gain in accuracy.

⁵As with galaxy morphology classification, we find no benefit from attentive pooling relative to mean pooling.

as in [Equation 1](#). Then, we train a lightweight convolutional upsampler to predict the image-level segmentation maps from the AION-1 embeddings. Our upsampler design is largely inspired by the mask decoder from [Kirillov et al. \(2023\)](#)⁶, but with a key modification: we do not include hypernetworks instantiated from additional register tokens. Instead, we use a single convolutional layer to project the upsampled output to the desired number of segmentation maps, simplifying the architecture while maintaining efficiency. Training is performed on 80% of the dataset, while 20% is held-out for validation. Additionally, we train a supervised, end-to-end U-Net baseline directly on the image-segmentation pairs, following the architecture from ([Walmsley and Spindler, 2023](#)). We measure performance on the held-out test set with the standard intersection-over-union

$$\text{IoU} = \frac{|M_{\text{pred}} \cap M_{\text{true}}|}{|M_{\text{pred}} \cup M_{\text{true}}|}, \quad (9)$$

averaged over images containing the structure. We present both the IoU on the held-out test set, as well as sample AION-1-B segmentations, in [Figure 9](#). AION-1’s frozen encoder plus a small decoder outperforms the supervised U-Net by +0.06 IoU on spiral arms and +0.04 on bars.

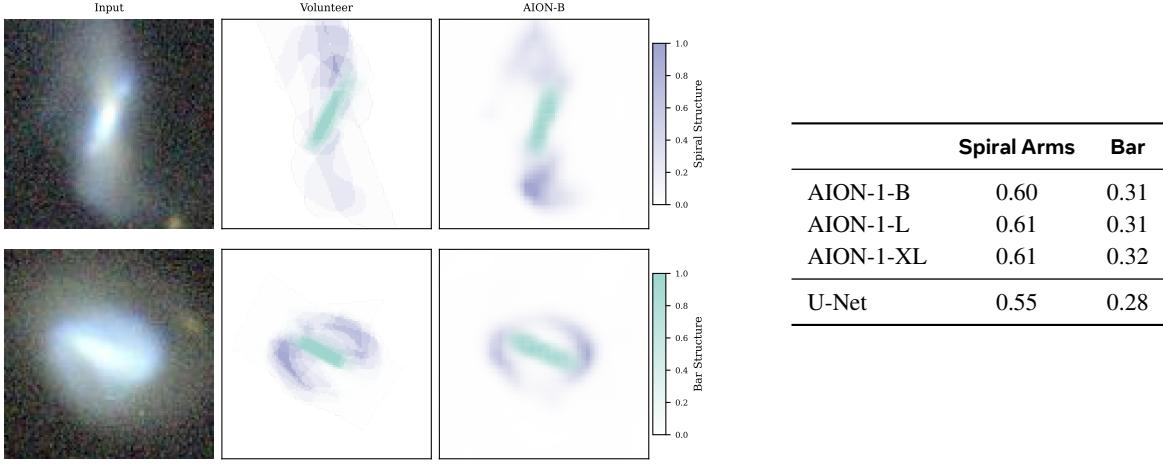


Figure 9: Galaxy structure segmentation. Left: Examples of galaxy image segmentation produced from the AION-1 embeddings with a lightweight convolutional neural network compared with the ground-truth true volunteer labels. Right: Mean IoU (\uparrow), given by [Equation 7.2.3](#), on the held-out test set.

7.2.4 Stellar Parameter Estimation

To evaluate AION-1 on stellar parameters we assemble a heterogeneous testbed that combines Gaia DR3 photometry and low-resolution XP spectra with high-resolution DESI optical spectra. Starting from the ~ 45 M stars used during pre-training, we cross-match them against the DESI EDR and retain only those sources that (i) fall inside the pre-training validation healpix tiles—thereby guaranteeing that AION-1 has never seen them before—and (ii) possess reliable stellar labels from the catalogue of [Zhang et al. \(2024\)](#), hereafter Z24. Z24 employs a data-driven regularised model to infer T_{eff} , $\log g$, [Fe/H], micro-turbulent velocity v_{mic} , and 18 elemental abundances directly from DESI spectra; here we focus on the four global parameters shared with the literature baselines. The resulting sample contains $\sim 240,000$ stars, of which 80% are used to train downstream heads and 20% are reserved for evaluation.

Each input configuration—photometry alone (Ph); Ph augmented with parallax and sky position (Ph + Plx + RA/Dec); Gaia XP spectra plus the previous channels (XP + Ph + Plx + RA/Dec); and DESI spectra with parallax (Sp + Plx)—is processed by the corresponding modality-specific encoders ([subsection 4.2](#)–[subsection 4.3](#)) to yield a single, mixed-modality token stream. Following the protocol adopted for galaxy properties, we pass the tokens through the frozen AION-1 encoder, apply a single learned cross-attention pooling layer [Equation 2](#) that compresses the variable-length sequence to a fixed d -dimensional vector; and map this vector to the target stellar labels with a linear probe.

In addition to the AION-1 results, we also train two baselines end-to-end for comparison:

⁶<https://github.com/facebookresearch/segment-anything/>

- **ConvNeXt regressor on raw spectra.** A stack of ConvNeXt-Tiny blocks and down-sampling layers identical to the AION-1 spectral encoder is trained end-to-end on DESI spectra and inverse-variance arrays, followed by gated attention pooling and a linear head which predicts the stellar properties.
- **XGBoost on token representations.** We mean-pool either (1) the 1-D Gaia photometry tokens or (2) the 1-D Gaia photometry tokens and XP coefficients and the DESI spectra produced by the frozen AION-1 encoder and fit a gradient-boosted decision-tree regressor to predict stellar properties, following the same general procedures as in [Leung and Bovy \(2024\)](#).

For all models, we minimize MSE on the training split and report the coefficient of determination R^2 on the held-out set. Results for all three AION-1 sizes, together with supervised and self-supervised baselines, are summarized in [Table 3](#). AION-1’s multimodal fusion yields state-of-the-art performance with minimal adaptation: adding geometric information (Plx+RA/Dec) noticeably improves metallicity, while incorporating XP spectra delivers a further $\sim 15\text{--}20\%$ absolute gain in [Fe/H] R^2 . When high-resolution DESI spectra are available, AION-1 matches the supervised ConvNeXt baseline despite using frozen weights, underscoring the quality of its spectral representations. Overall, the model’s ability to ingest heterogeneous inputs and deliver consistently strong predictions highlights its potential as a cross-survey foundation for stellar astrophysics.

In addition to the supervised baselines presented above, we also compare performance on the stellar parameter regression task with a current state-of-the-art baseline from [Leung and Bovy \(2024\)](#), who developed a Transformer-based foundation model for stellar data. More specifically, the task is to predict APOGEE-derived stellar parameters - namely T_{eff} , $\log g$, and [Fe/H] - from Gaia XP spectral coefficients. We use the same data as [Leung and Bovy \(2024\)](#), and cross-match APOGEE-derived stellar parameters with the MMU Gaia data, producing a set of roughly $\sim 10,000$ APOGEE parameter-Gaia XP spectral pairs. We feed as input to both AION-1 and the [Leung and Bovy \(2024\)](#) model only the first 32 BP coefficients and first 32 RP coefficients due to the fact that the [Leung and Bovy \(2024\)](#) model only has a context length of 64; we artificially handicap AION - which does not have this restriction - in order to perform a fair comparison. We note here that [Leung and Bovy \(2024\)](#) has been explicitly given APOGEE-derived stellar parameters and Gaia XP coefficients during its pretraining stage, and so this task is one that it has effectively been trained for. On the other hand, the pretraining dataset for AION-B does not contain APOGEE data, nor any other stellar parameters; we simply train a simple linear projection layer with cross-attention pooling on 5000 paired examples, and leave the weights of the AION model itself frozen. Despite this, we outperform the [Leung and Bovy \(2024\)](#) model across all parameters, as shown in [Table 4](#).

	T_{eff}	$\log g$	[Fe/H]	v_{mic}
AION-1-B				
Ph	0.94	0.95	0.58	0.86
Ph+Plx+RA/Dec	0.94	0.95	0.70	0.87
XP+Ph+Plx+RA/Dec	0.96	0.98	0.91	0.89
Sp+Plx	0.99	0.98	0.94	0.89
AION-1-L				
Ph	0.95	0.96	0.58	0.87
Ph+Plx+RA/Dec	0.95	0.96	0.71	0.88
XP+Ph+Plx+RA/Dec	0.97	0.98	0.92	0.89
Sp+Plx	0.99	0.98	0.94	0.89
AION-1-XL				
Ph	0.92	0.94	0.56	0.85
Ph+Plx+RA/Dec	0.93	0.95	0.68	0.87
XP+Ph+Plx+RA/Dec	0.97	0.97	0.91	0.88
Sp+Plx	0.98	0.98	0.92	0.89
XGB Baseline				
Ph ¹	0.94	0.95	0.59	0.87
XP+Ph+Sp ¹	0.99	0.98	0.89	0.89
ConvNeXT Baseline ²				
Sp	0.99	0.98	0.95	0.89

Table 3: R^2 (\uparrow) for stellar-label prediction. Inputs: Gaia photometry (Ph), low-resolution Gaia XP spectra (XP), parallax (Plx), celestial coordinates (RA/Dec), and high-resolution DESI spectra (Sp). ¹Gradient-boosted trees (XGBoost). ²Baseline convolution–attention network trained on spectra only.

Model	T_{eff} (K)	$\log g$ (dex)	[Fe/H] (dex)
AION-B	94.6	0.206	0.115
Leung and Bovy (2024)	99.1	0.229	0.143

Table 4: Performance on APOGEE stellar property predictions from Gaia XP coefficients (\downarrow), as measured by standard deviation of residuals. K is temperature units in Kelvin, and dex represents scatter on a logarithmic scale.

7.2.5 Performance in Low-Data Regime

In astronomy, many key classes (e.g. rare transients, metal-poor stars, high-redshift galaxies) come with only a handful of reliable labels. As a result, approaches that continue to perform well in the low-data regime are particularly valuable. We explore AION-1’s performance in this setting across three of the tasks described above: galaxy parameter estimation from images, stellar parameter estimation from spectra, and galaxy morphology classification from images. To that end, for these experiments, we keep the test set fixed at 20% of the total available data volume, but artificially reduce the size of the available training data. For each subset of training data, we retrain the lightweight head on top of the frozen AION-1 encoder as well as the supervised baselines from scratch. For galaxy and stellar tasks we average R^2 across the target parameters; for morphology we report overall accuracy. Figure Figure 10 shows two encouraging trends:

- 1. Low-Data Performance:** With only 10^2 – 10^3 labels, AION-1 already reaches $R^2 \sim 0.5$ for physical properties and $\geq 80\%$ for galaxy morphology, while end-to-end supervised models remain near-zero R^2 or below 70% accuracy, highlighting AION-1’s performance even in extremely low-data regimes.
- 2. Faster Data Saturation** While AION-1 plateaus by $\sim 10^3$ – 10^4 training examples; the baselines need an order of magnitude more data to approach similar performance.

Together, these results demonstrate that a strong, multimodal foundation model can transfer rich physical knowledge into downstream tasks long before a sizable, task-specific training set exists. In practical terms, observatories and survey teams can deploy a frozen AION-1 backbone and achieve competitive results with only a few hundred carefully vetted labels—orders of magnitude fewer than conventional supervised pipelines require. For future tasks, this may reduce the need for expensive spectroscopic follow-up or large volunteer-led annotation campaigns and make it feasible to apply machine learning to data-sparse niches.

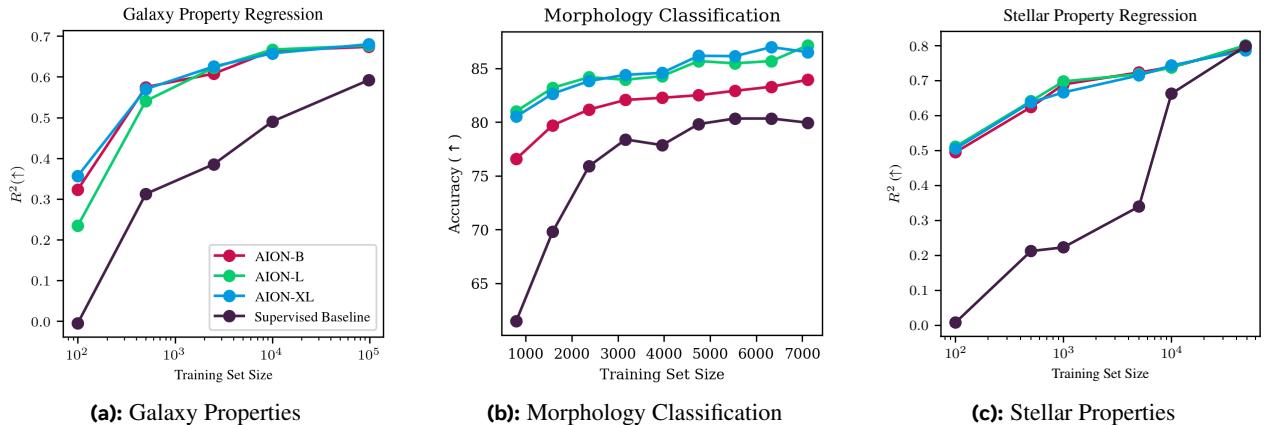


Figure 10: Model performance vs downstream task training set size. We regress (a) galaxy physical properties from images, (b) classify galaxy morphology, and (c) regress stellar properties from spectra on the same held-out test sample equating to 20% of the available data. However, we artificially reduce the training set size, and train a lightweight head on top of the frozen AION-1 encoder and a supervised model on the raw input data for each training set size. For galaxy and stellar properties, we report the R^2 averaged over all the properties, while for morphology classification we report the average accuracy.

7.3 Rare Object Detection

Modern astronomical surveys catalog hundreds of millions of sources, however, the discoveries that most advance cosmology often lie in the distribution’s extreme tail—for example, strong gravitational lenses that constrain dark-matter substructure and cosmic expansion. Because such phenomena appear so infrequently, assembling large, well-annotated training sets is inherently difficult. The small catalogs that do exist are usually built with hand-tuned selection cuts, baking those choices—and their attendant biases—into any supervised model trained on them. As a result, effective search for rare objects requires methods that minimize reliance on task-specific labels and remain robust to selection bias. AION-1 addresses this challenge by semantic retrieval in latent space, in which AION-1 is used to embed both query galaxies and all candidate images into a shared space. Cosine similarity then ranks the corpus, enabling zero-shot discovery without specialist tuning.

Type	Dataset	Image Type	Total Number	Frequency in Dataset
Spiral	GZ-DECaLS	Legacy Survey	24 622	26%
Merger	GZ-DECaLS	Legacy Survey	726	2%
Lenses	HSC Strong Lenses	Legacy Survey	758	0.1%

Table 5: High-confidence query sets used for our retrieval experiments. Citizen-science morphologies (spirals, mergers) come from the Galaxy Zoo–DECaLS catalog (Walmsley et al., 2022), while strong-lens candidates are drawn from previous lens-finding catalog in HSC crossmatched with Legacy Survey (Jaelani et al., 2024). All images used are from the Legacy Survey in the ensuing results.

Set-Up Specifically, given a query galaxy—which, for example, could be a known strong gravitational lens-AION-1—is used to produce an embedding of the galaxy’s image, $\mathbf{z}_q \in \mathbb{R}^d$; in this case, we use mean-pooling over AION-1’s output vectors, as in Equation 1. For all other galaxies in a search corpus, which we will refer to as candidates, a series of candidate embeddings, \mathbf{z}_c , are also produced. For each candidate, we compute the cosine similarity between its embedding and the query embedding, given by

$$S_c(\mathbf{x}_q, \mathbf{x}_c) = \frac{\mathbf{x}_q^\top \mathbf{x}_c}{\|\mathbf{x}_q\|_2 \|\mathbf{x}_c\|_2}. \quad (10)$$

All candidates can then be sorted in descending order of S_c , and the top N galaxies can be returned as the most suitable candidates for that type of query.

In order to quantify the performance of AION-1 on such a task, we use the normalized Discounted Cumulative Gain (nDCG). Specifically, let r_i be the relevance label of the candidate object ranked at position i . The discounted cumulative gain (DCG) at rank k is

$$\text{DCG}@k = \sum_{i=1}^k \frac{2^{r_i} - 1}{\log_2(i + 1)}. \quad (11)$$

The ideal DCG (IDCG@ k) is computed with the candidates ordered by descending relevance; finally

$$\text{nDCG}@k = \frac{\text{DCG}@k}{\text{IDCG}@k}.$$

We adopt $k = 10$ in all experiments.

Galaxy-Zoo Retrieval Experiments We construct a benchmark from the Galaxy Zoo–DECaLS catalog (GZ-DECaLS; Walmsley et al., 2022), which provides citizen-science morphology votes for Legacy Survey galaxies (Dey et al., 2019). After removing objects with fewer than three volunteer votes and cross-matching the GZDECaLS with the Legacy Survey SGC, the sample contains $\sim 171\,000$ galaxies. We focus on two visually distinctive classes—mergers and spirals, and focus on high-confidence exemplars to form our query galaxies, which we define as galaxies for which the fraction of volunteers who agreed on the class f exceeds 90%. This yields roughly 700 merger queries and 25,000 spiral queries. For each query (merger or spiral), we then return the top $k = 10$ candidate galaxies according to Equation 10, and compute the nDCG@10 score, where the relevance score of the retrieved galaxies r_i is given by the vote fraction for the query class; e.g., a galaxy for which 70% of the volunteers labeled the galaxy a spiral receives a relevance score of $r = 0.7$ for a spiral query. This soft-label strategy rewards the retrieval of unambiguous examples while still giving partial credit to visually ambiguous cases. Ultimately, the nDCG scores are averaged over all query vectors and reported for that class.

Strong-Lens Retrieval Experiments For the strong lensing retrieval task, we start by filtering the cross-matched catalog of objects within the Legacy Survey and HSC datasets to approximately reproduce the parent sample used in the HSC strong lensing searches Jaelani et al. (2024). Specifically, we impose three additional cuts: (1) objects with photometric redshifts between 0.2 and 1.2, (2) objects with an estimated stellar mass above $5 \times 10^{10} M_\odot$, and (3) objects with a star formation rate to stellar mass ratio less than 1×10^{-10} . In order to identify the strong gravitational lenses within the resulting parent sample, we cross-match with previous lens-finding catalogs Bina et al. (2016); Bolton et al. (2006, 2008); Cañameras et al. (2021); Faure et al. (2008); Goulaud et al. (2018); Jacobs et al. (2019); Jaelani et al. (2020, 2024); Li et al. (2019); More et al. (2014); Oguri et al. (2012); Ruff et al. (2011); Schuld et al. (2024); Shu et al.

(2022); Sonnenfeld et al. (2018, 2019, 2020); Talbot et al. (2021); Wong et al. (2018, 2022). This yields roughly 700 strong lenses. Most strong lensing catalogs offer a grade for each candidate. Since the criteria for this grading varies between catalogs, we ignore these grades and instead assign a relevance score of 1.0 to all the strong lensing candidates found within each catalog. All other objects in our parent sample are given a relevance score of 0.0. For each strong lens Legacy Survey image, we perform the same steps as above, and return the top $k = 10$ candidates, after which we compute the nDCG@10 score and average over all queries.

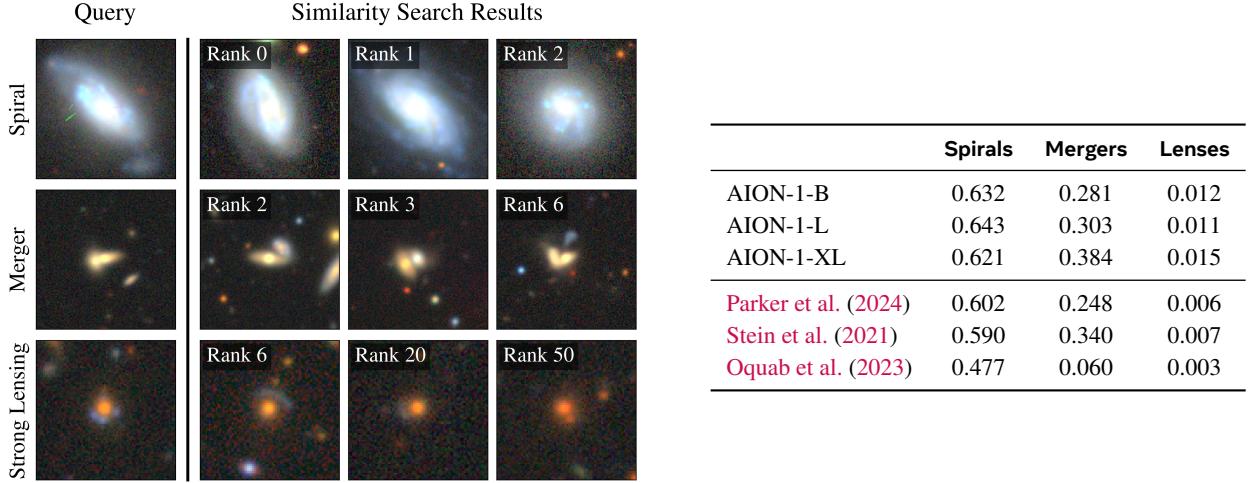


Figure 11: Galaxy Image Retrieval for three astronomical classes of decreasing prevalence; spirals, mergers and strong lenses. Left: Example candidates for a given query image. Correct candidates are shown, along with their rank among all retrieved objects sorted by cosine similarity. Right: The nDCG@10 score for each of the classes averaged over all queries in the given dataset.

Results Figure 11 demonstrates the performance of using AION-1’s encoder as an embedding model for rare object retrieval tasks. We include for comparison the results from two state-of-the-art astrophysics foundation models, AstroCLIP (Parker et al., 2024) and a self-supervised galaxy image model on Legacy Survey images (Stein et al., 2021), which are executed and evaluated in the same way as AION-1, except for the fact that they both ingest $\{g, r, z\}$ Legacy Survey imaging rather than $\{g, r, i, z\}$ imaging. We also demonstrate the relative performance of a DINOv2 (Oquab et al., 2023) vision model’s embeddings on these tasks. Qualitatively, the nearest neighbours returned by AION-1-XL are visually convincing for all three classes—even for strong lenses, whose true abundance is below 0.1%. Quantitatively, AION-1 outperforms every baseline across the board. Taken together, these results demonstrate that AION-1’s latent space is sufficiently discriminative to enable zero-shot retrieval of both common and vanishingly rare phenomena.

7.4 Emergent Transfer Properties

7.4.1 Generative Transfer of Multimodal Understanding

Our training mixture contains many pairwise matches, e.g. HSC images \leftrightarrow SDSS spectra and SDSS spectra \leftrightarrow DESI spectra. However, the model is never trained to on the DESI-HSC pairs, and so never learns to produce HSC images from DESI spectra directly. Nevertheless, when we condition use AION-1 to sample a DESI spectrum from an HSC image, the generated spectrum (blue) closely tracks the ground-truth DESI spectrum (black); see Figure 12. Both a quiescent (red) galaxy and a star-forming (blue) galaxy are reproduced with realistic absorption and emission features, demonstrating that AION-1 has learned a transitive mapping across modalities. This is likely due to the fact that AION-1 already understands the mapping between HSC and other spectroscopic (SDSS) or imaging (Legacy Survey) surveys, as well as the mapping between those intermediate surveys and the final target survey, DESI.

7.4.2 Survey-to-survey Transfer in Embedding Space

Beyond conditional generation, we ask whether AION-1’s frozen image encoder produces survey-invariant representations that let us port knowledge from one telescope to another. Specifically, we train a single linear classifier on Legacy

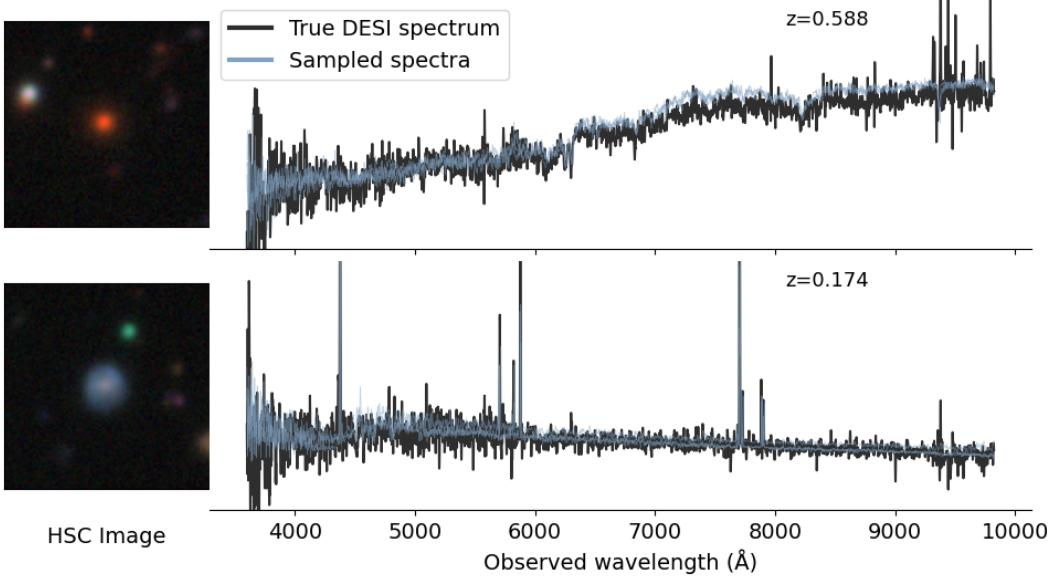


Figure 12: Out-of-distribution conditional generation. DESI spectra sampled from AION-1 (blue) conditioned on HSC images (insets), overlaid on the true DESI spectra (black). Even though HSC–DESI pairs were never seen during pre-training, the model reproduces key spectral features, demonstrating emergent transitive understanding.

Survey embeddings to predict the ten Galaxy Zoo-10 morphology classes (see [subsubsection 7.2.2](#)). The encoder weights remain fixed; only the 10-way soft-max layer is optimized. We then apply this exact head—with any fine-tuning—to embeddings of Hyper Suprime-Cam (HSC) images. To create the evaluation set we cross-match the HSC wide catalog (see [subsection 3.2](#) with Galaxy Zoo-10 (GZ10) volunteer votes and explicitly remove any targets that overlap with the Legacy Survey-GZ10 training set to prevent test leakage. The resulting sample contains $\sim 1,000$ galaxies. As reported in [Table 6](#), the zero-shot classifier attains 84 – 86% accuracy across all AION-1 scales, essentially matching its performance on the native Legacy Survey domain. This robustness holds despite factor-of- ~ 2 differences in depth, distinct filter sets ($\{g, r, i, z, y\}$ vs. $\{g, r, i, z\}$), and a different pixel scale. The result underscores that AION-1 embeddings capture morphology in a way that is largely agnostic to survey-specific imaging characteristics, enabling workflows that recycle scarce labeled data from one survey to bootstrap science in another.

	Legacy Survey (Train)	HSC (Eval.)
AION-1-B	83.95	84.15
AION-1-L	87.16	85.66
AION-1-XL	86.99	85.91

Table 6: Zero-shot morphology-classification accuracy (%). A classifier trained on Legacy Survey images transfers directly to HSC, indicating that AION-1 produces survey-invariant representations.

8 Conclusion

In this work we have introduced **AION-1**, a billion-parameter, omni-modal foundation-model family for the astronomical sciences. Leveraging a series of tokenizers to homogenise 39 heterogeneous data modalities from five of the largest public surveys, and a multimodal masked-modeling objective to learn their joint distribution, AION-1 demonstrates—for the first time—the feasibility and utility of training a single, encoder-decoder architecture that:

- **Integrates imaging, spectroscopy and scalar metadata** drawn from instruments that differ widely in resolution, noise properties and wavelength coverage, without bespoke per-task architecture changes.
- **Achieves or surpasses state-of-the-art performance** on a diverse suite of downstream tasks—including galaxy

and stellar property estimation, morphology classification, image segmentation and spectral super-resolution—using only lightweight probes or small task-specific heads.

- **Excels in the low-data regime**, maintaining high R^2 and classification accuracy with two to three orders of magnitude fewer labels than fully supervised baselines, a critical capability when high-quality annotations are scarce or expensive.
- **Enables zero-shot semantic retrieval**, discovering rare objects such as strong gravitational lenses with better scores than previous state-of-the-art astronomical foundation models.
- **Enables transfer learning** across surveys by producing survey-invariant representations allowing us to port downstream tasks from one telescope to another.

Ultimately, AION-1 offers the community a new backbone that collapses traditional, siloed pipelines into a single model. Astronomers can now fuse heterogeneous observations, prototype new analyses and mine extreme outliers with only modest computational resources. By releasing all code, tokenisers, pretrained weights and evaluation suites under an open-source licence, we hope to accelerate the adoption of foundation-model approaches across current and forthcoming surveys.

8.1 Limitations and future work

Despite its versatility, AION-1 inherits several caveats. We discuss these in detail below.

Architectural Limitations There are two main architectural limitations to the AION-1 model: (1) Discretisation limits: The use of quantization/discretization during the tokenization process intrinsically limits the information content captured, which may impact downstream analyses; and (2) Naïve embedding aggregation: the mean-pool strategy used for retrieval is a pragmatic first pass but leaves performance on the table; contrastive post-training or other more intelligent approaches could furnish richer, task-aware representations.

Selection Functions & Representativeness AION-1 inherits the selection functions of its pre-training data: magnitude and quality cuts (e.g., LS magnitude thresholds, HSC “full-depth full-colour” + flags), Gaia XP availability, DESI EDR SV3 filters, sky-footprint choices (e.g., SGC-only), and our reciprocal cross-match. These selection functions may influence downstream predictions in the absence of recalibration (Gallegos et al., 2024; Kumar et al., 2022; Thaler et al., 2024). For this reason, we primarily propose using AION-1 as an embedding model to extract relevant features from a given data source, on top of which we train a simple projection head which is effectively re-calibrated during task-specific adaptation on a specific downstream dataset that defines both a task and a selection function for that task. Nonetheless, future work on exploring selection functions in pretraining would be an exciting direction for the astronomy community.

Generative Capabilities Our pre-training uses multimodal masked modeling. While this is effective and simple, it does not furnish a principled joint sampler for modalities with many output tokens (images, spectra, segmentation maps). In practice, samples can look convincing and even score well on task metrics, yet still be mis-calibrated and/or under-correlated across tokens. This is not a new problem—masked-model decoders and, more broadly, many ML generative methods face calibration gaps in multi-token settings, including in astronomy. Accordingly, we mainly suggest that the astronomy community recalibrate generative tasks using AION-1 embeddings instead of performing posterior draws over multi-token outputs so that downstream samples are better calibrated and preserve the correct token-to-token correlations prior to scientific use.

For future work, there exist alternatives to masked modeling. For example, one can replace masked decoding with models that define a coherent joint distribution over output tokens—e.g., (i) autoregressive token models (LM/decoder-only) with proper likelihoods, or (ii) diffusion models over continuous or tokenized representations conditioned on observed modalities. Both choices typically yield better-behaved samplers and correlations across tokens. However, they still encode a pre-training prior that may not match a new survey or selection; principled adaptation is then necessary; see, for instance, Rozet et al. (2024) for an expectation-maximization approach to refit a diffusion prior from incomplete or shifted observations.

8.2 Broader scientific impact

Although developed for astronomy, the AION-1 recipe—data-driven tokenisation, scale-appropriate masked modelling and modality-aware provenance embeddings—addresses challenges endemic to many experimental sciences: heterogeneity, noise and instrument-specific idiosyncrasies. We therefore envisage direct extensions to adjacent domains in which multi-instrument data proliferation similarly outpaces bespoke modeling.

9 Contributions

The author contributions are summarized below. In each category, authors are listed alphabetically.

- **Project Leads:** Francois Lanusse and Liam Parker
- **Data Team:** Micah Bowles, Tom Hehir, Lucas Meyer, Francois Lanusse, Liam Parker, Jeff Shen, Helen Qu, Sebastian Wagner-Carena
- **Tokenization Team:**
 - **Image:** Francois Lanusse, Liam Parker
 - **Spectrum:** Francois Lanusse, Jeff Shen
 - **Scalar:** Jeff Shen, Sebastian Wagner-Carena
 - **Catalog:** Ollie Liu
 - **Segmentation maps:** Micah Bowles, Tom Hehir
- **Pretraining Team:** Francois Lanusse, Siavash Golkar, Liam Parker, Leopoldo Sarra
- **Downstream Evaluation:** Micah Bowles, Tom Hehir, Francois Lanusse, Ollie Liu, Lucas Meyer, Liam Parker, Jeff Shen, Helen Qu, Sebastian Wagner-Carena
- **Computing and Optimization:** Hatim Bourfoune, Nathan Cassereau, Pierre Cornette, Geraud Krawezik, Lucas Meyer, Ruben Ohana
- **Manuscript Writing:** Micah Bowles, Tom Hehir, Francois Lanusse, Ollie Liu, Lucas Meyer, Liam Parker, Jeff Shen, Helen Qu, Sebastian Wagner-Carena
- **Advisory:** Alberto Bietti, Kyunghyun Cho, Miles Cranmer, Shirley Ho

10 Acknowledgments

We would like to acknowledge the support of the Simons Foundation and of Schmidt Sciences. This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2024-GC011015468 on the supercomputer Jean Zay’s H100 partition. Additionally, some of the computations in this work were run at facilities supported by the Scientific Computing Core at the Flatiron Institute, a division of the Simons Foundation. Liam Parker also acknowledges support from the National Science Foundation Graduate Research Fellowship Program. Jeff Shen is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number 587652. We would like to thank Andrew Engel, Marc Huertas-Company, Stephanie Juneau, Andy Morgan, and Mike Smith for their valuable feedback during beta testing and Sophie Barstein for her help with writing the corresponding blog post.

10.1 Data

10.1.1 Legacy Survey

The Legacy Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS; Proposal ID #2014B-0404; PIs: David Schlegel and Arjun Dey), the Beijing-Arizona Sky Survey (BASS; NOAO Prop. ID #2015A-0801; PIs: Zhou Xu and Xiaohui Fan), and the Mayall z-band Legacy Survey (MzLS; Prop. ID #2016A-0453; PI: Arjun Dey). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF's NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. Pipeline processing and analyses of the data were supported by NOIRLab and the Lawrence Berkeley National Laboratory (LBNL). The Legacy Surveys project is honored to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation. LBNL is managed by the Regents of the University of California under contract to the U.S. Department of Energy.

This project used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo a Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Cientifico e Tecnologico and the Ministerio da Ciencia, Tecnologia e Inovacao, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenossische Technische Hochschule (ETH) Zurich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciencies de l'Espai (IEEC/CSIC), the Institut de Fisica d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig Maximilians Universitat Munchen and the associated Excellence Cluster Universe, the University of Michigan, NSF's NOIRLab, the University of Nottingham, the Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University.

BASS is a key project of the Telescope Access Program (TAP), which has been funded by the National Astronomical Observatories of China, the Chinese Academy of Sciences (the Strategic Priority Research Program “The Emergence of Cosmological Structures” Grant # XDB09000000), and the Special Fund for Astronomy from the Ministry of Finance. The BASS is also supported by the External Cooperation Program of Chinese Academy of Sciences (Grant # 114A11KYSB20160057), and Chinese National Natural Science Foundation (Grant # 12120101003, # 11433005).

The Legacy Survey team makes use of data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), which is a project of the Jet Propulsion Laboratory/California Institute of Technology. NEOWISE is funded by the National Aeronautics and Space Administration.

The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.

10.1.2 Hyper Suprime-Cam

The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University.

This paper makes use of software developed for the Large Synoptic Survey Telescope. We thank the LSST Project for making their code available as free software at <http://dm.lsst.org>

The Pan-STARRS1 Surveys (PS1) have been made possible through contributions of the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation under Grant No. AST-1238877, the University of Maryland, and Eotvos Lorand University (ELTE) and the Los Alamos National Laboratory.

10.1.3 Dark Energy Spectroscopic Instrument

This research used data obtained with the Dark Energy Spectroscopic Instrument (DESI). DESI construction and operations is managed by the Lawrence Berkeley National Laboratory. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High-Energy Physics, under Contract No. DE-AC02-05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Science and Technology of Mexico (CONACYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: www.desi.lbl.gov/collaborating-institutions. The DESI collaboration is honored to be permitted to conduct scientific research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation, the U.S. Department of Energy, or any of the listed funding agencies.

10.1.4 Sloan Digital Sky Survey

Funding for the Sloan Digital Sky Survey V has been provided by the Alfred P. Sloan Foundation, the Heising-Simons Foundation, the National Science Foundation, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. SDSS telescopes are located at Apache Point Observatory, funded by the Astrophysical Research Consortium and operated by New Mexico State University, and at Las Campanas Observatory, operated by the Carnegie Institution for Science. The SDSS web site is www.sdss.org.

SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration, including Caltech, The Carnegie Institution for Science, Chilean National Time Allocation Committee (CNTAC) ratified researchers, The Flatiron Institute, the Gotham Participation Group, Harvard University, Heidelberg University, The Johns Hopkins University, L'Ecole polytechnique fédérale de Lausanne (EPFL), Leibniz-Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Extraterrestrische Physik (MPE), Nanjing University, National Astronomical Observatories of China (NAOC), New Mexico State University, The Ohio

State University, Pennsylvania State University, Smithsonian Astrophysical Observatory, Space Telescope Science Institute (STScI), the Stellar Astrophysics Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Illinois at Urbana-Champaign, University of Toronto, University of Utah, University of Virginia, Yale University, and Yunnan University.

10.1.5 Gaia

This work has made use of data from the European Space Agency (ESA) mission Gaia (<https://www.cosmos.esa.int/gaia>), processed by the Gaia Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. The Gaia data are open and free to use, provided credit is given to ‘ESA/Gaia/DPAC’. If you use Gaia DR3 data in your research, please acknowledge it as above.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balland, Otger Ballester, et al. The desi experiment part i: science, targeting, and survey design. arXiv preprint arXiv:1611.00036, 2016.
- Romina Ahumada, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, Scott F Anderson, Brett H Andrews, Borja Anguiano, Riccardo Arcodia, Eric Armengaud, Marie Aubert, et al. The 16th data release of the sloan digital sky surveys: first release from the apogee-2 southern survey and full release of eboss spectra. *The Astrophysical Journal Supplement Series*, 249(1):3, 2020.
- Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, Stéphane Arnouts, Neta A Bahcall, Steven Bickerton, James Bosch, Kevin Bundy, Peter L Capak, James HH Chan, et al. The hyper suprime-cam ssp survey: overview and survey design. *Publications of the Astronomical Society of Japan*, 70(SP1):S4, 2018.
- Anthropic. Claude. <https://anthropic.com/clause>, 2024. Large language model.
- Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities, 2024. <https://arxiv.org/abs/2406.09406>.
- D. Bina, R. Pelló, J. Richard, J. Lewis, V. Patrício, S. Cantalupo, E. C. Herenz, K. Soto, P. M. Weilbacher, R. Bacon, J. D. R. Vernet, L. Wisotzki, B. Clément, J. G. Cuby, D. J. Lagattuta, G. Soucail, and A. Verhamme. MUSE observations of the lensing cluster Abell 1689. *A&A*, 590:A14, May 2016. doi: 10.1051/0004-6361/201527913.
- Adam S. Bolton, Scott Burles, Léon V. E. Koopmans, Tommaso Treu, and Leonidas A. Moustakas. The Sloan Lens ACS Survey. I. A Large Spectroscopically Selected Sample of Massive Early-Type Lens Galaxies. *ApJ*, 638(2):703–724, February 2006. doi: 10.1086/498884.
- Adam S. Bolton, Scott Burles, Léon V. E. Koopmans, Tommaso Treu, Raphaël Gavazzi, Leonidas A. Moustakas, Randall Wayth, and David J. Schlegel. The Sloan Lens ACS Survey. V. The Full ACS Strong-Lens Sample. *ApJ*, 682(2):964–984, August 2008. doi: 10.1086/589327.
- R. Cañameras, S. Schuldt, Y. Shu, S. H. Suyu, S. Taubenberger, T. Meinhardt, L. Leal-Taixé, D. C. Y. Chao, K. T. Inoue, A. T. Jaelani, and A. More. HOLISMOKES. VI. New galaxy-scale strong lens candidates from the HSC-SSP imaging survey. *A&A*, 653:L6, September 2021. doi: 10.1051/0004-6361/202141758.
- Gaia Collaboration et al. The gaia mission. arXiv preprint arXiv:1609.04153, 2016.
- DESI Collaboration, Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E. Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balland, Otger Ballester, Charles Baltay, Lucas Beaufort, Chris Bebek, Timothy C. Beers, Eric F. Bell, José Luis Bernal, Robert Besuner, Florian Beutler, Chris Blake, Hannes Bleuler, Michael Blomqvist, Robert Blum, Adam S. Bolton, Cesar Briceno, David Brooks, Joel R. Brownstein, Elizabeth Buckley-Geer, Angela Burden, Etienne Burtin, Nicolas G. Busca, Robert N. Cahn, Yan-Chuan Cai, Laia Cardiel-Sas, Raymond G. Carlberg, Pierre-Henri Carton, Ricard Casas, Francisco J. Castander, Jorge L. Cervantes-Cota, Todd M. Claybaugh, Madeline Close, Carl T. Coker, Shaun Cole, Johan Comparat, Andrew P. Cooper, M. C. Cousinou, Martin Crocce, Jean-Gabriel Cuby, Daniel P. Cunningham, Tamara M. Davis, Kyle S. Dawson, Axel de la Macorra, Juan De Vicente, Timothée Delubac, Mark Derwent, Arjun Dey, Govinda Dhungana, Zhejie Ding, Peter Doel, Yutong T. Duan, Anne Ealet, Jerry Edelstein, Sarah Eftekharzadeh, Daniel J. Eisenstein, Ann Elliott, Stéphanie Escoffier, Matthew Evatt, Parker Fagrelius, Xiaohui Fan, Kevin Fanning, Arya Farahi, Jay Farihi, Ginevra Favole, Yu Feng,

Enrique Fernandez, Joseph R. Findlay, Douglas P. Finkbeiner, Michael J. Fitzpatrick, Brenna Flaugher, Samuel Flender, Andreu Font-Ribera, Jaime E. Forero-Romero, Pablo Fosalba, Carlos S. Frenk, Michele Fumagalli, Boris T. Gaensicke, Giuseppe Gallo, Juan Garcia-Bellido, Enrique Gaztanaga, Nicola Pietro Gentile Fusillo, Terry Gerard, Irena Gershkovich, Tommaso Giannantonio, Denis Gillet, Guillermo Gonzalez-de-Rivera, Violeta Gonzalez-Perez, Shelby Gott, Or Graur, Gaston Gutierrez, Julien Guy, Salman Habib, Henry Heetderks, Ian Heetderks, Katrin Heitmann, Wojciech A. Hellwing, David A. Herrera, Shirley Ho, Stephen Holland, Klaus Honscheid, Eric Huff, Timothy A. Hutchinson, Dragan Huterer, Ho Seong Hwang, Joseph Maria Illa Laguna, Yuzo Ishikawa, Dianna Jacobs, Niall Jeffrey, Patrick Jelinsky, Elise Jennings, Linhua Jiang, Jorge Jimenez, Jennifer Johnson, Richard Joyce, Eric Jullo, Stéphanie Juneau, Sami Kama, Armin Karcher, Sonia Karkar, Robert Kehoe, Noble Kennamer, Stephen Kent, Martin Kilbinger, Alex G. Kim, David Kirkby, Theodore Kisner, Ellie Kitanidis, Jean-Paul Kneib, Sergey Koposov, Eve Kovacs, Kazuya Koyama, Anthony Kremin, Richard Kron, Luzius Kronig, Andrea Kueter-Young, Cedric G. Lacey, Robin Lafever, Ofer Lahav, Andrew Lambert, Michael Lampton, Martin Landriau, Dustin Lang, Tod R. Lauer, Jean-Marc Le Goff, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Su-Jeong Lee, Daniela Leitner, Michael Lesser, Michael E. Levi, Benjamin L'Huillier, Baojiu Li, Ming Liang, Huan Lin, Eric Linder, Sarah R. Loebman, Zarija Lukic, Jun Ma, Niall MacCrann, Christophe Magneville, Laleh Makarem, Marc Manera, Christopher J. Manser, Robert Marshall, Paul Martini, Richard Massey, Thomas Matheson, Jeremy McCauley, Patrick McDonald, Ian D. McGreer, Aaron Meisner, Nigel Metcalfe, Timothy N. Miller, Ramon Miquel, John Moustakas, Adam Myers, Milind Naik, Jeffrey A. Newman, Robert C. Nichol, Andrina Nicola, Luiz Nicolati da Costa, Jundai Nie, Gustavo Niz, Peder Norberg, Brian Nord, Dara Norman, Peter Nugent, Thomas O'Brien, Minji Oh, and Knut A. G. Olsen. The DESI Experiment Part I: Science, Targeting, and Survey Design. arXiv e-prints, art. arXiv:1611.00036, October 2016. doi: 10.48550/arXiv.1611.00036.

DESI Collaboration, A. G. Adame, J. Aguilar, S. Ahlen, S. Alam, G. Aldering, D. M. Alexander, R. Alfarsy, C. Allende Prieto, M. Alvarez, O. Alves, A. Anand, F. Andrade-Oliveira, E. Armengaud, J. Asorey, S. Avila, A. Aviles, S. Bailey, A. Balaguera-Antolínez, O. Ballester, C. Baltay, A. Bault, J. Bautista, J. Behera, S. F. Beltran, S. BenZvi, L. Bernaldo e Silva, J. R. Bermejo-Climent, A. Berti, R. Besuner, F. Beutler, D. Bianchi, C. Blake, R. Blum, A. S. Bolton, S. Brienden, A. Brodzeller, D. Brooks, Z. Brown, E. Buckley-Geer, E. Burtin, L. Cabayol-Garcia, Z. Cai, R. Canning, L. Cardiel-Sas, A. Carnero Rosell, F. J. Castander, J. L. Cervantes-Cota, S. Chabanier, E. Chaussidon, J. Chaves-Montero, S. Chen, X. Chen, C. Chuang, T. Claybaugh, S. Cole, A. P. Cooper, A. Cuceu, T. M. Davis, K. Dawson, R. de Belsunce, R. de la Cruz, A. de la Macorra, J. Della Costa, A. de Mattia, R. Demina, U. Demirbozan, J. DeRose, A. Dey, B. Dey, G. Dhungana, J. Ding, Z. Ding, P. Doel, R. Doshi, K. Douglass, A. Edge, S. Eftekharzadeh, D. J. Eisenstein, A. Elliott, J. Ereza, S. Escoffier, P. Fagrelius, X. Fan, K. Fanning, V. A. Fawcett, S. Ferraro, B. Flaugher, A. Font-Ribera, J. E. Forero-Romero, D. Forero-Sánchez, C. S. Frenk, B. T. Gänsicke, L. A. García, J. García-Bellido, C. García-Quintero, L. H. Garrison, H. Gil-Marín, J. Golden-Marx, S. Gontcho A Gontcho, A. X. González-Morales, V. González-Pérez, C. Gordon, O. Graur, D. Green, D. Gruen, J. Guy, B. Hadzhiyska, C. Hahn, J. J. Han, M. M. S. Hanif, H. K. Herrera-Alcantar, K. Honscheid, J. Hou, C. Howlett, D. Huterer, V. Iršič, M. Ishak, A. Jacques, A. Jana, L. Jiang, J. Jimenez, Y. P. Jing, S. Joudaki, R. Joyce, E. Jullo, S. Juneau, N. G. Karaçaylı, T. Karim, R. Kehoe, S. Kent, A. Khederlarian, S. Kim, D. Kirkby, T. Kisner, F. Kitaura, N. Kizhuprakkat, J. Kneib, S. E. Koposov, A. Kovács, A. Kremin, A. Krolewski, B. L'Huillier, O. Lahav, A. Lambert, C. Lamman, T. W. Lan, M. Landriau, D. Lang, J. U. Lange, J. Lasker, A. Leauthaud, L. Le Guillou, M. E. Levi, T. S. Li, E. Linder, A. Lyons, C. Magneville, M. Manera, C. J. Manser, D. Margala, P. Martini, P. McDonald, G. E. Medina, L. Medina-Varela, A. Meisner, J. Mena-Fernández, J. Meneses-Rizo, M. Mezcua, R. Miquel, P. Montero-Camacho, J. Moon, S. Moore, J. Moustakas, E. Mueller, J. Mundet, A. Muñoz-Gutiérrez, A. D. Myers, S. Nadathur, L. Napolitano, R. Neveux, J. A. Newman, J. Nie, R. Nikutta, G. Niz, P. Norberg, H. E. Noriega, E. Paillas, N. Palanque-Delabrouille, A. Palmese, Z. Pan, D. Parkinson, S. Penmetsa, W. J. Percival, A. Pérez-Fernández, I. Pérez-Ràfols, M. Pieri, C. Poppett, A. Porredon, and S. Pothier. The Early Data Release of the Dark Energy Spectroscopic Instrument. AJ, 168(2):58, August 2024. doi: 10.3847/1538-3881/ad3217.

Arjun Dey, David J Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, et al. Overview of the desi legacy imaging surveys. The Astronomical Journal, 157(5):168, 2019.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Cecile Faure, Jean-Paul Kneib, Giovanni Covone, Lidia Tasca, Alexie Leauthaud, Peter Capak, Knud Jahnke, Vernesa Smolcic, Sylvain de la Torre, Richard Ellis, Alexis Finoguenov, Anton Koekemoer, Oliver Le Fevre, Richard Massey, Yannick Mellier, Alexandre Refregier, Jason Rhodes, Nick Scoville, Eva Schinnerer, James Taylor, Ludovic Van Waerbeke, and Jakob Walcher. First Catalog of Strong Lens Candidates in the COSMOS Field. ApJS, 176(1):19–38, May 2008. doi: 10.1086/526426.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024. <https://arxiv.org/abs/2309.00770>.

Petko Gemini Team, Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023. <https://arxiv.org/abs/2305.05665>.

- Charles F. Goulaud, Joseph B. Jensen, John P. Blakeslee, Chung-Pei Ma, Jenny E. Greene, and Jens Thomas. The MASSIVE Survey. IX. Photometric Analysis of 35 High-mass Early-type Galaxies with HST WFC3/IR. *ApJ*, 856(1):11, March 2018. doi: 10.3847/1538-4357/aab1f3.
- ChangHoon Hahn, KJ Kwon, Rita Tojeiro, Małgorzata Siudek, Rebecca EA Canning, Mar Mezcua, Jeremy L Tinker, David Brooks, Peter Doel, Kevin Fanning, et al. The desi probabilistic value-added bright galaxy survey (provabgs) mock challenge. *The Astrophysical Journal*, 945(1):16, 2023.
- Md Abul Hayat, George Stein, Peter Harrington, Zarija Lukić, and Mustafa Mustafa. Self-supervised Representation Learning for Astronomical Images. *ApJ*, 911(2):L33, April 2021. doi: 10.3847/2041-8213/abf2c7.
- C. Jacobs, T. Collett, K. Glazebrook, E. Buckley-Geer, H. T. Diehl, H. Lin, C. McCarthy, A. K. Qin, C. Odden, M. Caso Escudero, P. Dial, V. J. Yung, S. Gaitsch, A. Pellico, K. A. Lindgren, T. M. C. Abbott, J. Annis, S. Avila, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, L. N. da Costa, J. De Vicente, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. A. Goldstein, D. Gruen, R. A. Gruendl, J. Gschwend, D. L. Hollowood, K. Honscheid, B. Hoyle, D. J. James, E. Krause, N. Kuropatkin, O. Lahav, M. Lima, M. A. G. Maia, J. L. Marshall, R. Miquel, A. A. Plazas, A. Roodman, E. Sanchez, V. Scarpine, S. Serrano, I. Sevilla-Noarbe, M. Smith, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, V. Vikram, A. R. Walker, Y. Zhang, and DES Collaboration. An Extended Catalog of Galaxy-Galaxy Strong Gravitational Lenses Discovered in DES Using Convolutional Neural Networks. *ApJS*, 243(1):17, July 2019. doi: 10.3847/1538-4365/ab26b6.
- Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs, 2022. <https://arxiv.org/abs/2107.14795>.
- Anton T. Jaelani, Anupreeta More, Masamune Oguri, Alessandro Sonnenfeld, Sherry H. Suyu, Cristian E. Rusu, Kenneth C. Wong, James H. H. Chan, Issha Kayo, Chien-Hsiu Lee, Dani C. Y. Chao, Jean Coupon, Kaiki T. Inoue, and Toshifumi Futamase. Survey of Gravitationally lensed Objects in HSC Imaging (SuGOHI) - V. Group-to-cluster scale lens search from the HSC-SSP Survey. *MNRAS*, 495(1):1291–1310, June 2020. doi: 10.1093/mnras/staa1062.
- Anton T. Jaelani, Anupreeta More, Kenneth C. Wong, Kaiki T. Inoue, Dani C. Y. Chao, Premana W. Premadi, and Raoul Cañameras. Survey of gravitationally lensed objects in HSC imaging (SuGOHI) - X. Strong lens finding in the HSC-SSP using convolutional neural networks. *MNRAS*, 535(2):1625–1639, December 2024. doi: 10.1093/mnras/stae2442.
- Diederik P Kingma, J Adam Ba, and J Adam. A method for stochastic optimization. arxiv 2014. arXiv preprint arXiv:1412.6980, 106: 6, 2020.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023.
- Nolan Koblischke and Jo Bovy. SpectraFM: Tuning into Stellar Foundation Models. arXiv e-prints, art. arXiv:2411.04750, November 2024. doi: 10.48550/arXiv.2411.04750.
- Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022. <https://arxiv.org/abs/2202.10054>.
- Henry W Leung and Jo Bovy. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, November 2018. ISSN 1365-2966. doi: 10.1093/mnras/sty3217. <http://dx.doi.org/10.1093/mnras/sty3217>.
- Henry W. Leung and Jo Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *MNRAS*, 527(1): 1494–1520, January 2024. doi: 10.1093/mnras/stad3015.
- Rui Li, Yiping Shu, Jianlin Su, Haicheng Feng, Guobao Zhang, Jiancheng Wang, and Hongtao Liu. Using deep Residual Networks to search for galaxy-Ly α emitter lens candidates based on spectroscopic selection. *MNRAS*, 482(1):313–320, January 2019. doi: 10.1093/mnras/sty2708.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019. <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Karen L Masters, Coleman Krawczyk, Shoaib Shamsi, Alexander Todd, Daniel Finnegan, Matthew Bershady, Kevin Bundy, Brian Cherinka, Amelia Fraser-McKelvie, Dhanesh Krishnarao, Sandor Kruk, Richard R Lane, David Law, Chris Lintott, Michael Merrifield, Brooke Simmons, Anne-Marie Weijmans, and Renbin Yan. Galaxy Zoo: 3D - crowdsourced bar, spiral, and foreground star masks for MaNGA target galaxies. *Monthly Notices of the Royal Astronomical Society*, 507(3):3923–3935, 08 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2282. <https://doi.org/10.1093/mnras/stab2282>.

- Peter Melchior, Yan Liang, ChangHoon Hahn, and Andy Goulding. Autoencoding galaxy spectra. i. architecture. *The Astronomical Journal*, 166(2):74, 2023.
- Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. arXiv preprint arXiv:2309.15505, 2023.
- Siddharth Mishra-Sharma, Yiding Song, and Jesse Thaler. PAPERCLIP: Associating Astronomical Observations and Natural Language with Multi-Modal Models. arXiv e-prints, art. arXiv:2403.08851, March 2024. doi: 10.48550/arXiv.2403.08851.
- David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling, 2023. <https://arxiv.org/abs/2312.06647>.
- A. More, R. Cabanac, S. More, C. Alard, M. Limousin, J. P. Kneib, R. Gavazzi, and V. Motta. The CFHTLS-Strong Lensing Legacy Survey (SL2S): Investigating the group-scale lenses with the SARCS sample. In *Journal of Physics Conference Series*, volume 484 of *Journal of Physics Conference Series*, page 012041. IOP, March 2014. doi: 10.1088/1742-6596/484/1/012041.
- Masamune Oguri, Matthew B. Bayliss, Håkon Dahle, Keren Sharon, Michael D. Gladders, Priyamvada Natarajan, Joseph F. Hennawi, and Benjamin P. Koester. Combined strong and weak lensing analysis of 28 clusters from the Sloan Giant Arcs Survey. *MNRAS*, 420(4):3213–3239, March 2012. doi: 10.1111/j.1365-2966.2011.20248.x.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.
- Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, Ruben Ohana, Mariel Pettee, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, Shirley Ho, and Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. *MNRAS*, 531(4):4990–5011, July 2024. doi: 10.1093/mnras/stae1450.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Maria Rizhko and Joshua S. Bloom. Astrom³: A self-supervised multimodal model for astronomy. 2024. <https://arxiv.org/abs/2411.08842>.
- François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization. *Advances in Neural Information Processing Systems*, 37:87647–87682, 2024.
- Andrea J. Ruff, Raphaël Gavazzi, Philip J. Marshall, Tommaso Treu, Matthew W. Auger, and Florence Brault. The SL2S Galaxy-scale Lens Sample. II. Cosmic Evolution of Dark and Luminous Mass in Early-type Galaxies. *ApJ*, 727(2):96, February 2011. doi: 10.1088/0004-637X/727/2/96.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Imagen: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487, 2022.
- Stefan Schuldt, Raoul Cañameras, Irham T. Andika, Satadru Bag, Alejandra Melo, Yiping Shu, Sherry H. Suyu, Stefan Taubenberger, and Claudio Grillo. HOLISMOKEs XIII: Strong-lens candidates at all mass scales and their environments from the Hyper-Suprime Cam and deep learning. arXiv e-prints, art. arXiv:2405.20383, May 2024. doi: 10.48550/arXiv.2405.20383.
- Yiping Shu, Raoul Cañameras, Stefan Schuldt, Sherry H. Suyu, Stefan Taubenberger, Kaiki Taro Inoue, and Anton T. Jaelani. HOLISMOKEs. VIII. High-redshift, strong-lens search in the Hyper Suprime-Cam Subaru Strategic Program. *A&A*, 662:A4, June 2022. doi: 10.1051/0004-6361/202243203.
- Michael J. Smith, Ryan J. Roberts, Eirini Angeloudi, and Marc Huertas-Company. AstroPT: Scaling Large Observation Models for Astronomy. arXiv e-prints, art. arXiv:2405.14930, May 2024. doi: 10.48550/arXiv.2405.14930.
- Alessandro Sonnenfeld, James H. H. Chan, Yiping Shu, Anupreeta More, Masamune Oguri, Sherry H. Suyu, Kenneth C. Wong, Chien-Hsiu Lee, Jean Coupon, Atsunori Yonehara, Adam S. Bolton, Anton T. Jaelani, Masayuki Tanaka, Satoshi Miyazaki, and Yutaka Komiyama. Survey of Gravitationally-lensed Objects in HSC Imaging (SuGOHI). I. Automatic search for galaxy-scale strong lenses. *PASJ*, 70:S29, January 2018. doi: 10.1093/pasj/psx062.
- Alessandro Sonnenfeld, Anton T. Jaelani, James Chan, Anupreeta More, Sherry H. Suyu, Kenneth C. Wong, Masamune Oguri, and Chien-Hsiu Lee. Survey of gravitationally-lensed objects in HSC imaging (SuGOHI). III. Statistical strong lensing constraints on the stellar IMF of CMASS galaxies. *A&A*, 630:A71, October 2019. doi: 10.1051/0004-6361/201935743.

Alessandro Sonnenfeld, Aprajita Verma, Anupreeta More, Elisabeth Baeten, Christine Macmillan, Kenneth C. Wong, James H. H. Chan, Anton T. Jaelani, Chien-Hsiu Lee, Masamune Oguri, Cristian E. Rusu, Marten Veldthuis, Laura Trouille, Philip J. Marshall, Roger Hutchings, Campbell Allen, James O'Donnell, Claude Cornen, Christopher P. Davis, Adam McMaster, Chris Lintott, and Grant Miller. Survey of Gravitationally-lensed Objects in HSC Imaging (SuGOHI). VI. Crowdsourced lens finding with Space Warps. *A&A*, 642:A148, October 2020. doi: 10.1051/0004-6361/202038067.

StabilityAI. Stable diffusion, 2022. <https://github.com/CompVis/stable-diffusion>.

George Stein, Peter Harrington, Jacqueline Blaum, Tomislav Medan, and Zarija Lukic. Self-supervised similarity search for large scientific datasets. arXiv preprint arXiv:2110.13151, 2021.

George Stein, Jacqueline Blaum, Peter Harrington, Tomislav Medan, and Zarija Lukic. Mining for strong gravitational lenses with self-supervised learning. *The Astrophysical Journal*, 932(2):107, 2022.

Michael S. Talbot, Joel R. Brownstein, Kyle S. Dawson, Jean-Paul Kneib, and Julian Bautista. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: a catalogue of strong galaxy-galaxy lens candidates. *MNRAS*, 502(3):4617–4640, April 2021. doi: 10.1093/mnras/stab267.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. <https://arxiv.org/abs/2405.09818>.

Marion Thaler, Abdullatif Köksal, Alina Leidinger, Anna Korhonen, and Hinrich Schütze. How far can bias go? – tracing bias from pretraining data to alignment, 2024. <https://arxiv.org/abs/2411.19240>.

The Multimodal Universe Collaboration, Jeroen Audenaert, Micah Bowles, Benjamin M. Boyd, David Chemaly, Brian Cherinka, Ioana Ciucă, Miles Cranmer, Aaron Do, Matthew Grayling, Erin E. Hayes, Tom Hehir, Shirley Ho, Marc Huertas-Company, Kartheik G. Iyer, Maja Jablonska, Francois Lanusse, Henry W. Leung, Kaisey Mandel, Juan Rafael Martínez-Galarza, Peter Melchior, Lucas Meyer, Liam H. Parker, Helen Qu, Jeff Shen, Michael J. Smith, Connor Stone, Mike Walmsley, and John F. Wu. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data. arXiv e-prints, art. arXiv:2412.02527, December 2024. doi: 10.48550/arXiv.2412.02527.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. CoRR, abs/1711.00937, 2017. <http://arxiv.org/abs/1711.00937>.

Mike Walmsley and Ashley Spindler. Deep learning segmentation of spiral arms and bars, 2023. <https://arxiv.org/abs/2312.02908>.

Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin Gal, et al. Galaxy zoo decals: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 2022.

Mike Walmsley, Micah Bowles, Anna M. M. Scaife, Jason Shingirai Makechemu, Alexander J. Gordon, Annette M. N. Ferguson, Robert G. Mann, James Pearson, Jürgen J. Popp, Jo Bovy, Josh Speagle, Hugh Dickinson, Lucy Fortson, Tobias Géron, Sandor Kruk, Chris J. Lintott, Kameswara Mantha, Devina Mohan, David O’Ryan, and Imigo V. Slijepevic. Scaling Laws for Galaxy Images. arXiv e-prints, art. arXiv:2404.02973, April 2024. doi: 10.48550/arXiv.2404.02973.

Kenneth C. Wong, Alessandro Sonnenfeld, James H. H. Chan, Cristian E. Rusu, Masayuki Tanaka, Anton T. Jaelani, Chien-Hsiu Lee, Anupreeta More, Masamune Oguri, Sherry H. Suyu, and Yutaka Komiyama. Survey of Gravitationally Lensed Objects in HSC Imaging (SuGOHI). II. Environments and Line-of-Sight Structure of Strong Gravitational Lens Galaxies to $z \sim 0.8$. *ApJ*, 867(2):107, November 2018. doi: 10.3847/1538-4357/aae381.

Kenneth C. Wong, James H. H. Chan, Dani C. Y. Chao, Anton T. Jaelani, Issha Kayo, Chien-Hsiu Lee, Anupreeta More, and Masamune Oguri. Survey of Gravitationally lensed objects in HSC Imaging (SuGOHI). VIII. New galaxy-scale lenses from the HSC SSP. *PASJ*, 74(5):1209–1219, October 2022. doi: 10.1093/pasj/psac065.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16133–16142, 2023.

Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The Sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10459–10469, 2023a.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737, 2023b.

Meng Zhang, Maosheng Xiang, Yuan-Sen Ting, Jiahui Wang, Haining Li, Hu Zou, Jundan Nie, Lanya Mou, Tianmin Wu, Yaqian Wu, and Jifeng Liu. Determining Stellar Elemental Abundances from DESI Spectra with the Data-driven Payne. ApJS, 273(2):19, August 2024. doi: 10.3847/1538-4365/ad51dd.