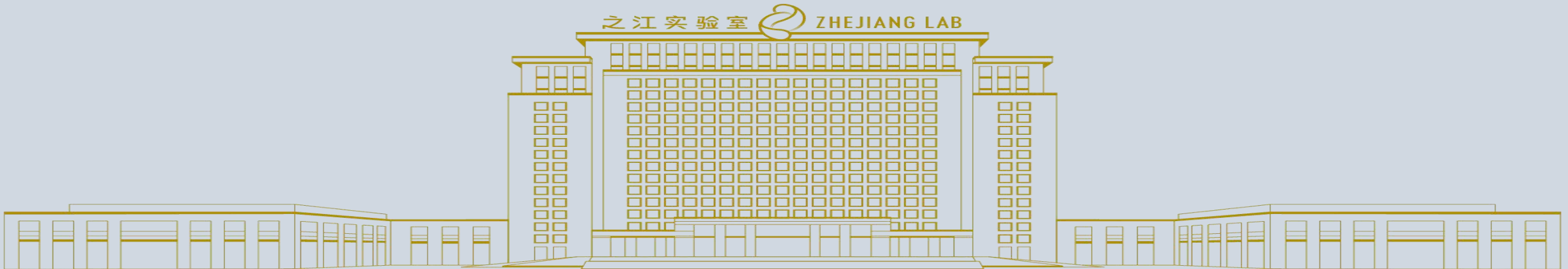
OneAstronomy

Building an AI Foundation Model for Astronomy

that Serves as a Global Public Good

Yongfu Hao · Zhejiang Lab
3rd International Workshop on AI + Astronomy
20 Oct 2026



Survey Era: Data Deluge & Resource Crunch

We observe more and more — yet our ability to understand has stalled



1 Data-processing costs keep climbing

PB-to-EB deluge; storage & compute costs grow linearly

2 Long-tail science targets are ever harder to find

Transients, strong lenses and other rare signals drown in the data flood

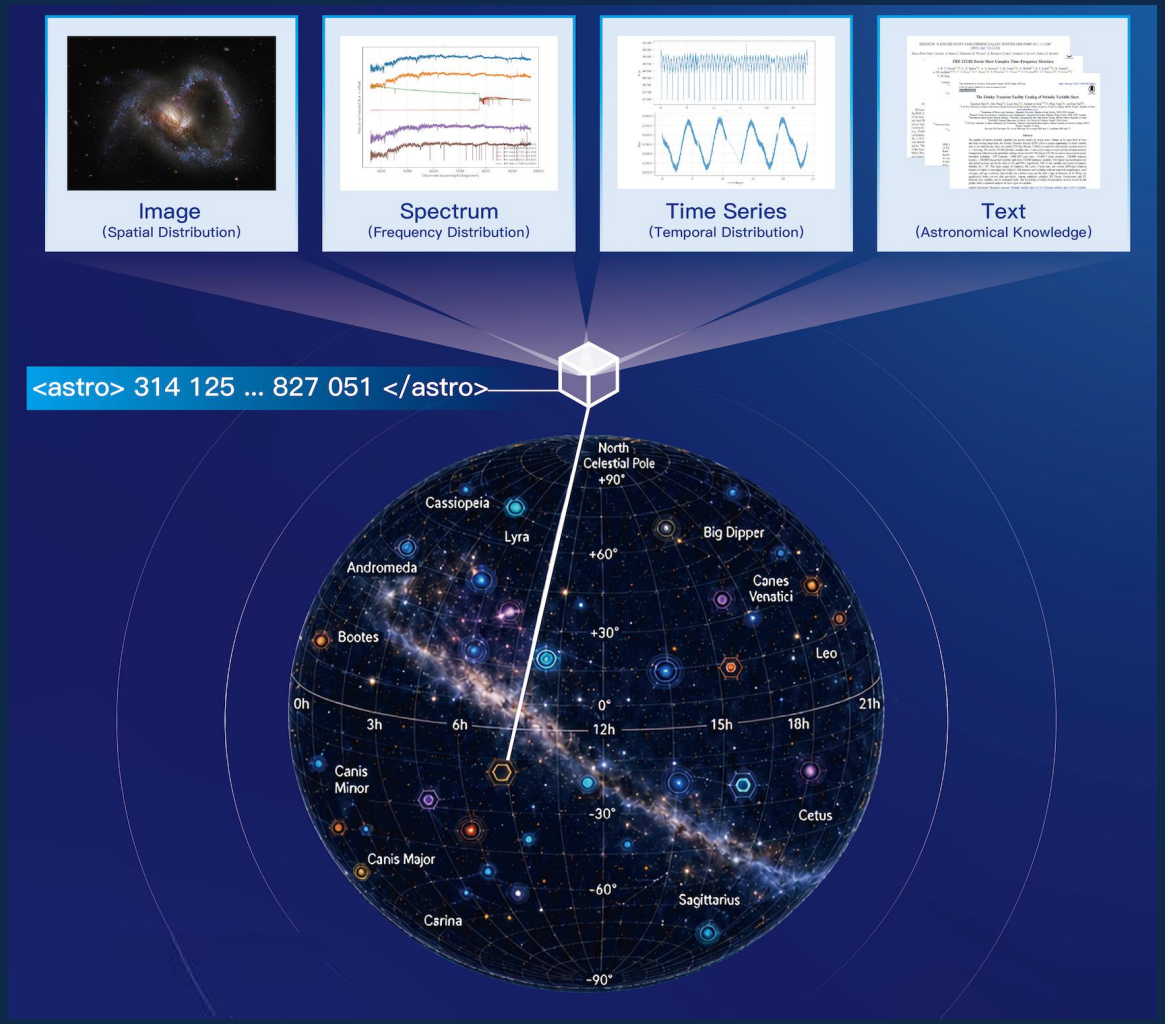
3 Traditional pipelines scale ever harder

Siloed, purpose-built pipelines — hard to reuse across modalities & tasks

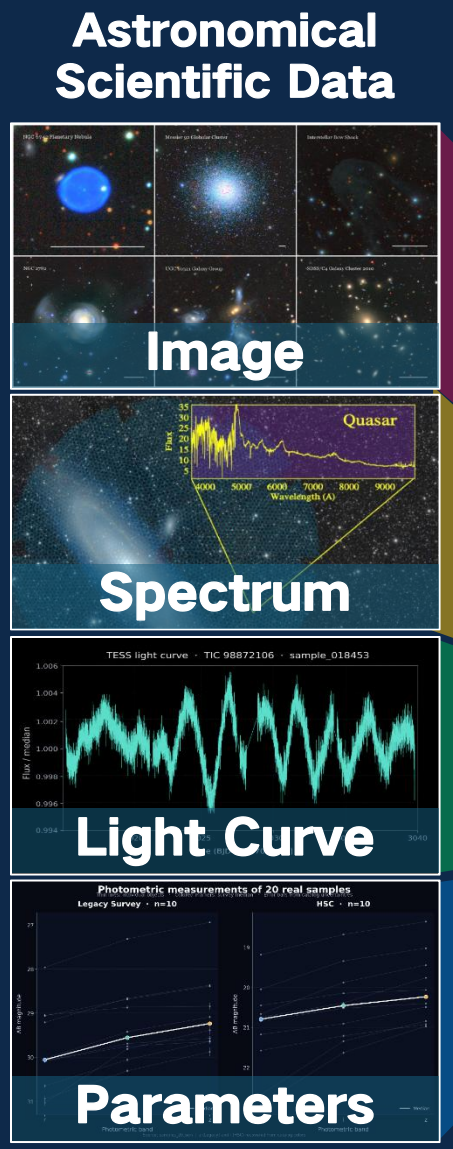
AI foundation models are the key to breaking this systemic bottleneck

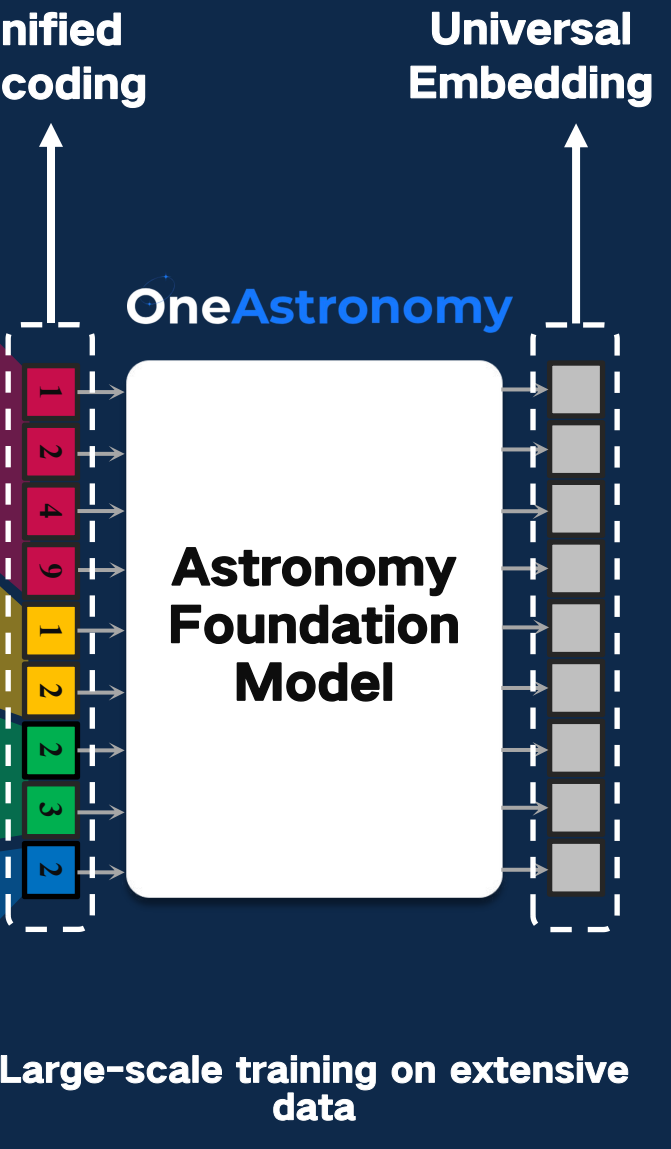
OneAstronomy: Astronomy Foundation Model

之江实验室 ZHEJIANG LAB



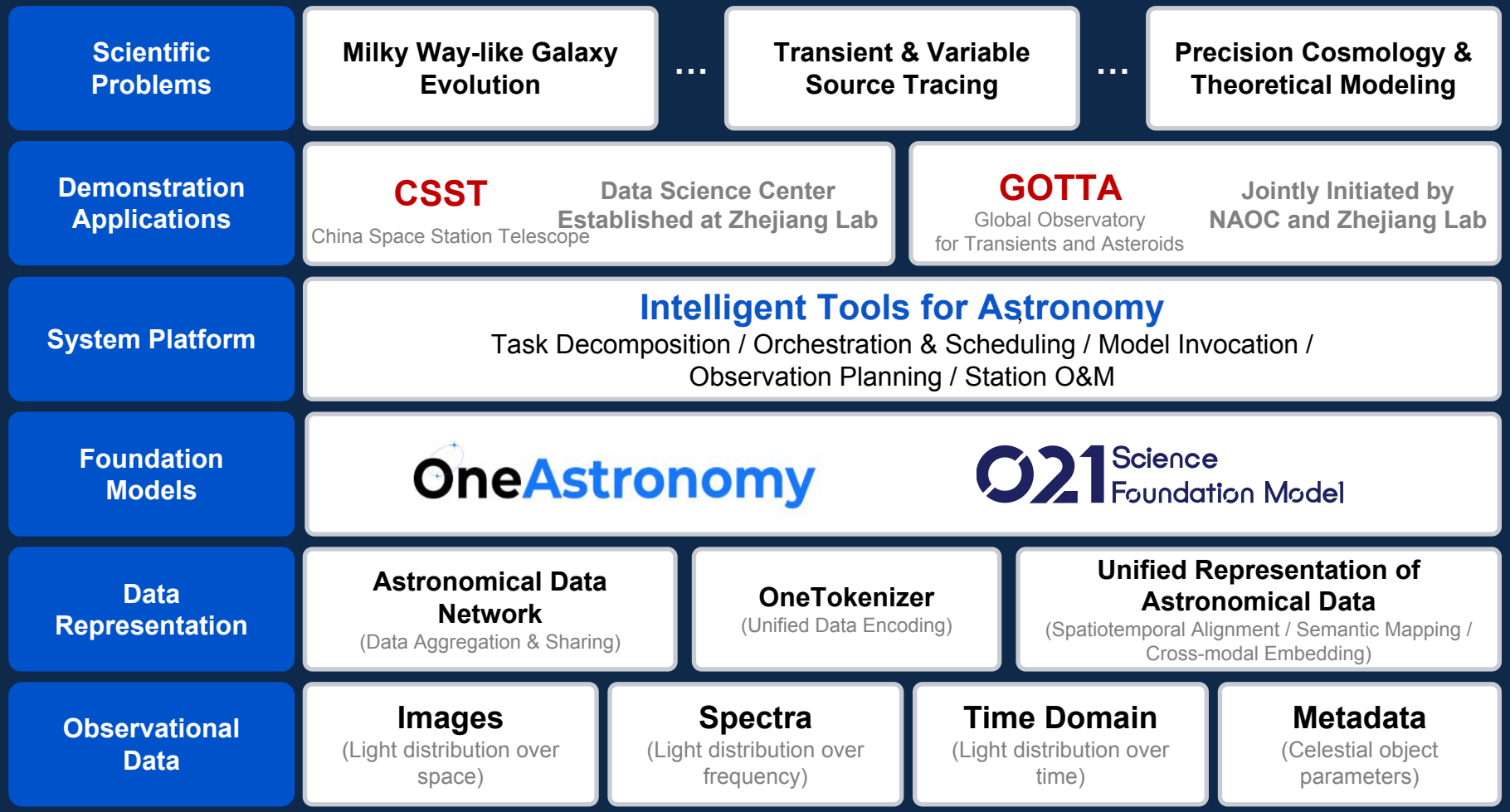
High-precision support for various cross-modal understanding
and complex astronomical discovery tasks





Large-scale training on extensive data

OneAstronomy: Overall Framework



Develop an astronomy foundation model with joint cross-modal representation of 'image-spectrum-time-parameters' to support major astronomical science missions such as CSST and GOTTA.

OneAstroDB: one object, one ID, across surveys

126 M objects with unified IDs, 100% aligned: organised for models from the source

Heterogeneous surveys → Unified entity sky partition → Epoch propagation proper motion → Probabilistic cross-match → master_view north + south

126 M objects, unified ID

24,587 HEALPix sky cells

3.49 M multimodal matches

Survey	Observations	Alignment	Matched / orphan
DESI spectra	171,844	0.012″	100% / 0
BOSS	535,798	0.091″	100% / 0
eBOSS	172,216	0.138″	100% / 0
SDSS legacy	304,998	0.078″	100% / 0
HSC images	2,301,115	—	100% / 0

Tokenized Astronomy Data Asset

91.3B
multimodal tokens

284.7 TB
observations

126M
objects

7
surveys

TOKEN COMPOSITION



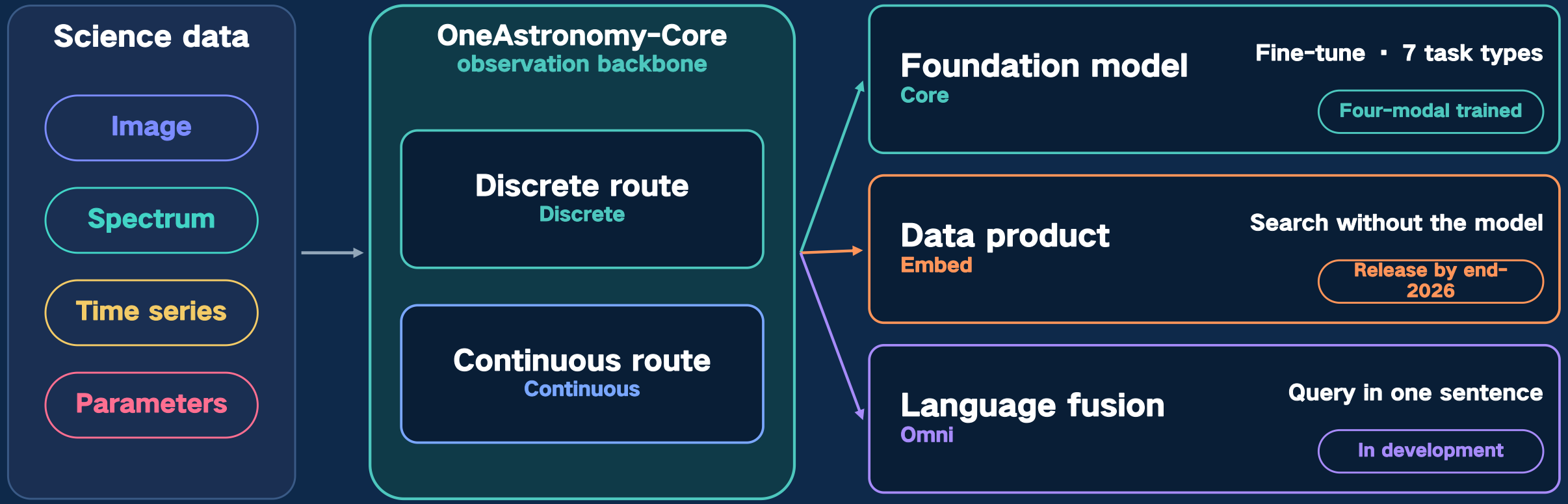
Imaging	72.6B
Gaia XP	13.4B
Scalars	2.5B
Spectra	2.2B
Light curves	0.57B

MODEL-READABLE SCIENTIFIC DATA

Every token remains linked to its object, modality, survey and position.

OneAstronomy model family: one foundation, three outputs

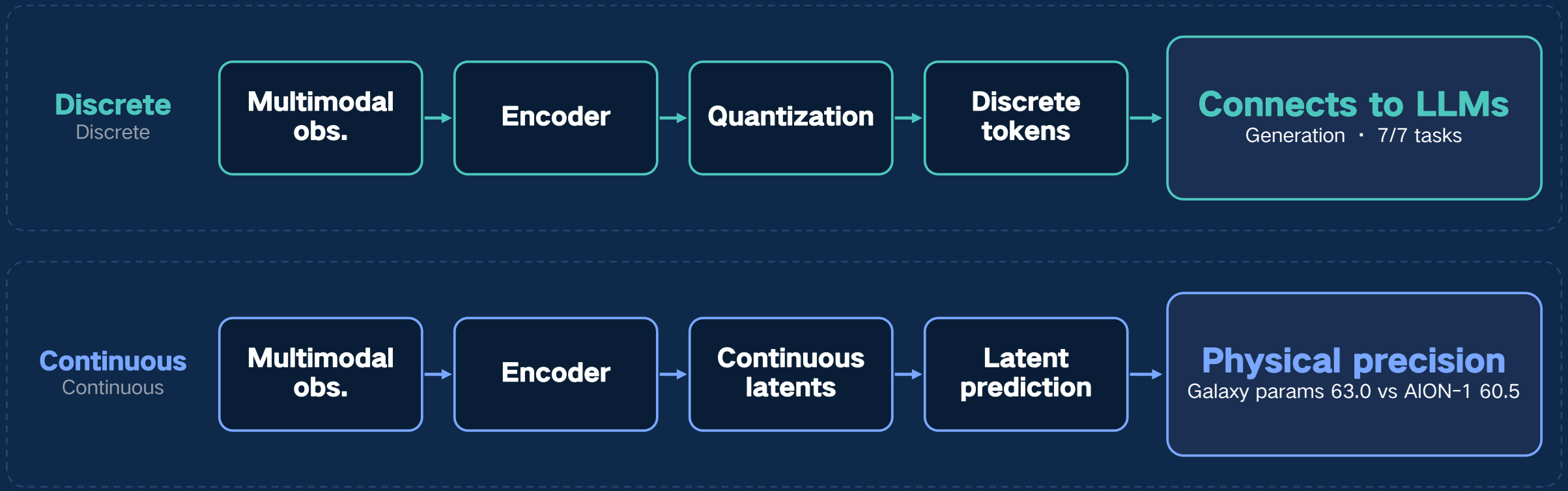
One object-centric dataset, one observation backbone, three outputs



Value Process the data once; models, data products and agents share one representation

Architecture: discrete + continuous routes

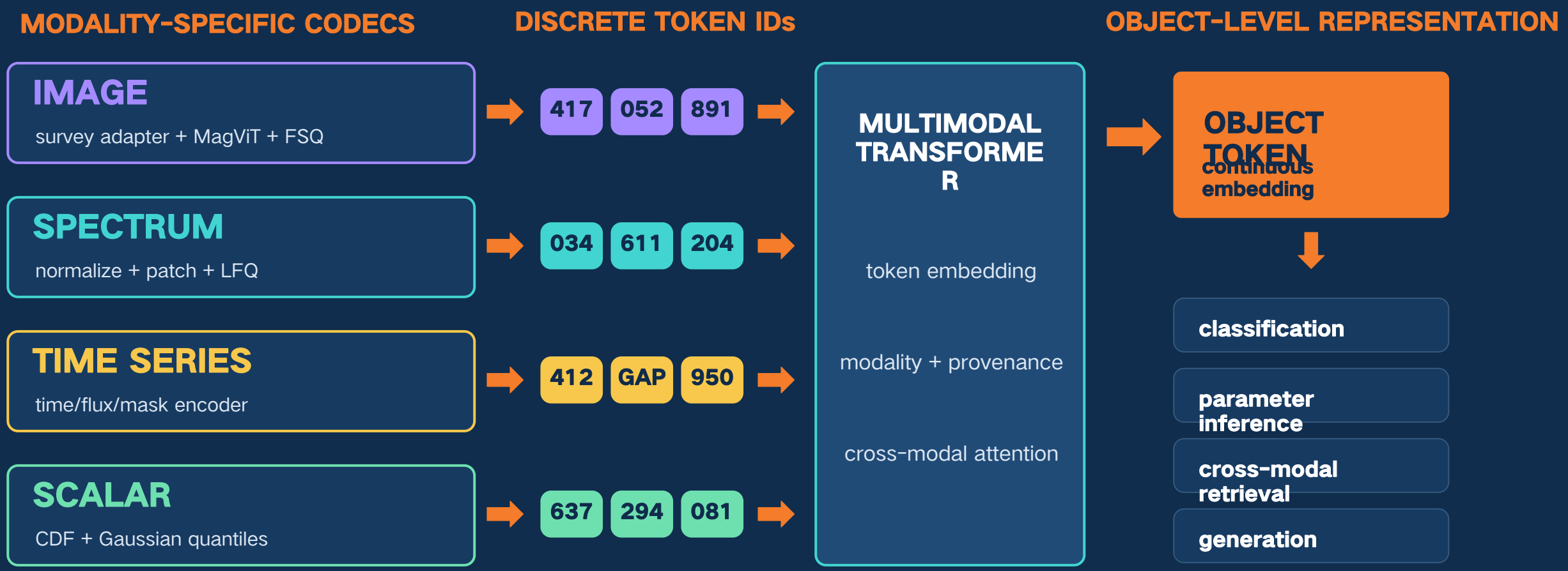
Discrete tokens connect to language; continuous latents keep physical precision



Same data | **Same exam: OneAstroBench** | **Each other's baseline** | **Best route feeds Embedding**

Discrete route ①: two-level tokenization

DISCRETE ROUTE



Discrete tokens describe observations.
The continuous token describes the astronomical object.

Discrete route ②: tokens to one object embedding

DISCRETE ROUTE

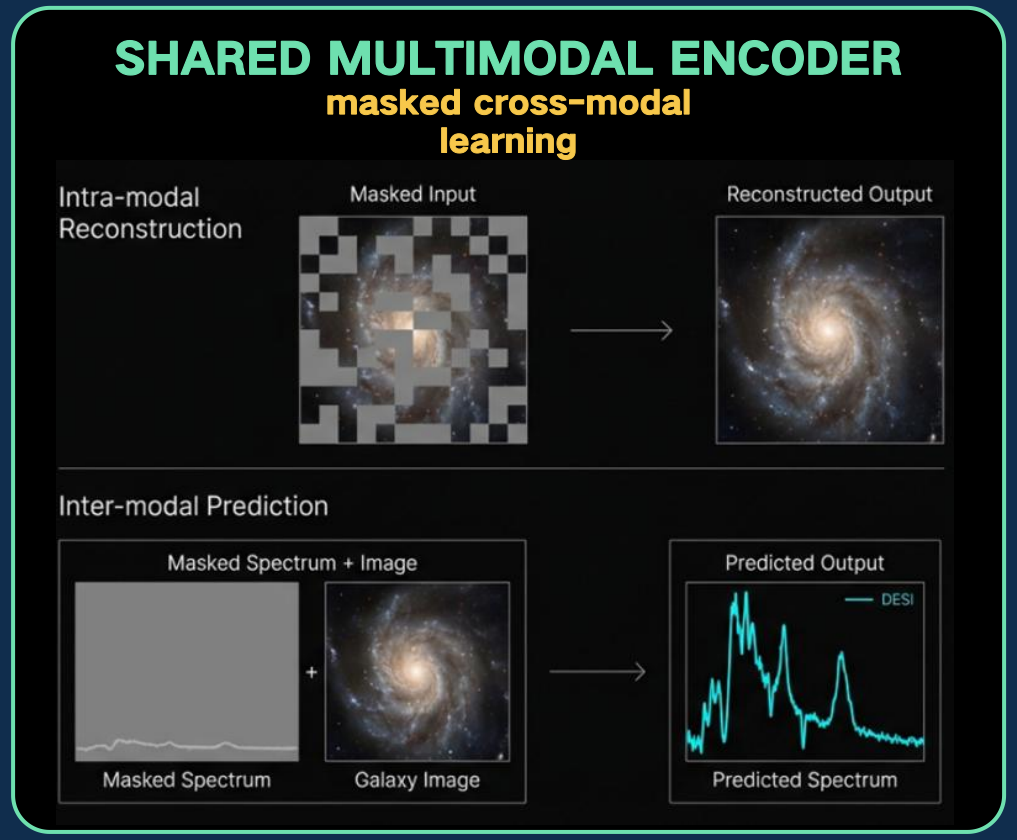
Masked cross-modal training aligns all observations of one object

IMAGE
417 052 891

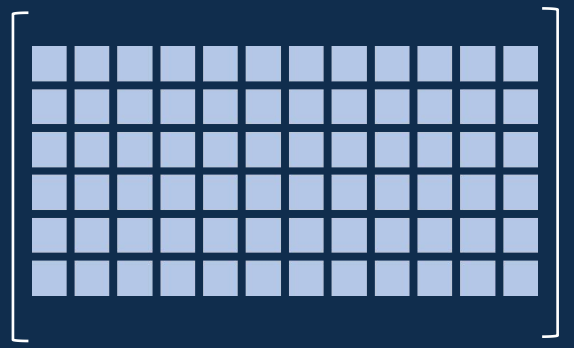
SPECTRUM
034 611 204

TIME SERIES
412 GAP 950

SCALAR
637 294 081



OBJECT EMBEDDING



N_tokens × N_dim

TRAINING

cross-matched modalities pull the same object into a shared latent neighborhood

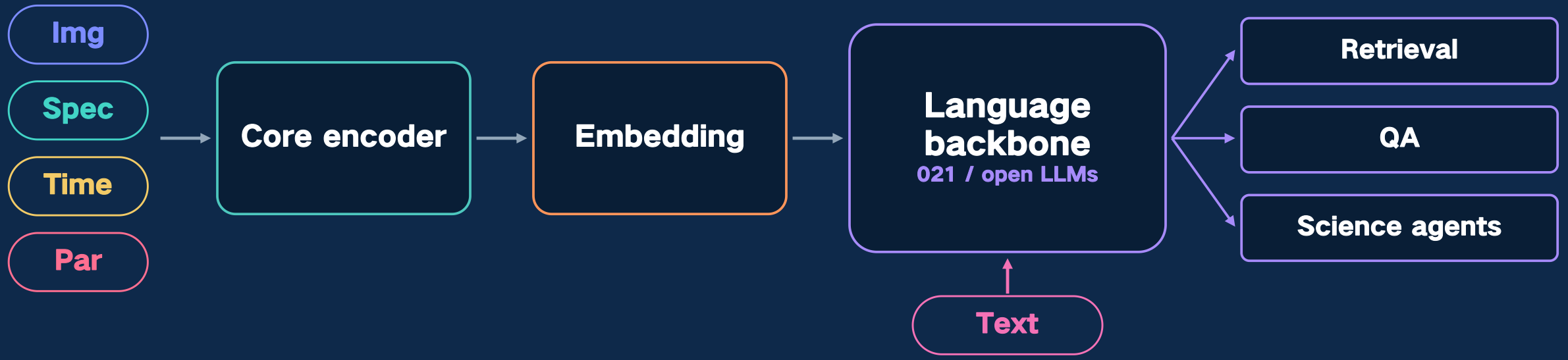
MISSING MODALITIES
still produce an embedding

CROSS-MODAL SEARCH
compare objects across surveys

DATA FUSION
richer evidence when available

Five-modal: observations and language in one model

Add text to image, spectrum, time and parameters: one-sentence queries



Traditional workflow — **5 steps**

Shape → Spectra → Time → Catalog → Cross-match

Five-modal model — **1 sentence**

“Find elliptical-looking objects with recently quenched star formation at z ≈ 0.1”

One year of progress

Data foundation built, four-modal model trained, discovery validated on strong lenses

Data

284.7 TB

Four-modal training data · 91.3 B tokens

2.8× the AION-1 training data

126 M objects, unified IDs, 100% aligned

Model

Four-modal

Foundation model trained · two routes

Covers all 7 benchmark tasks

Embedding data product validated

Discovery

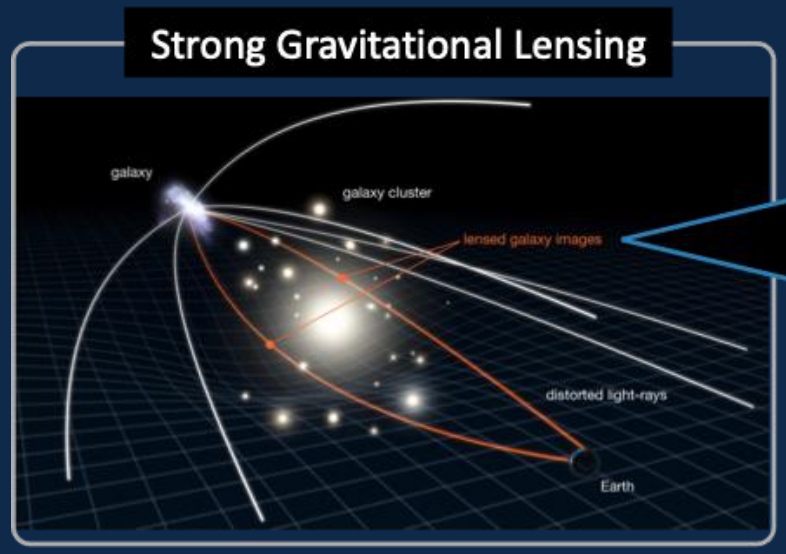
88.7%

Top-1 hits, strong-lens blind test

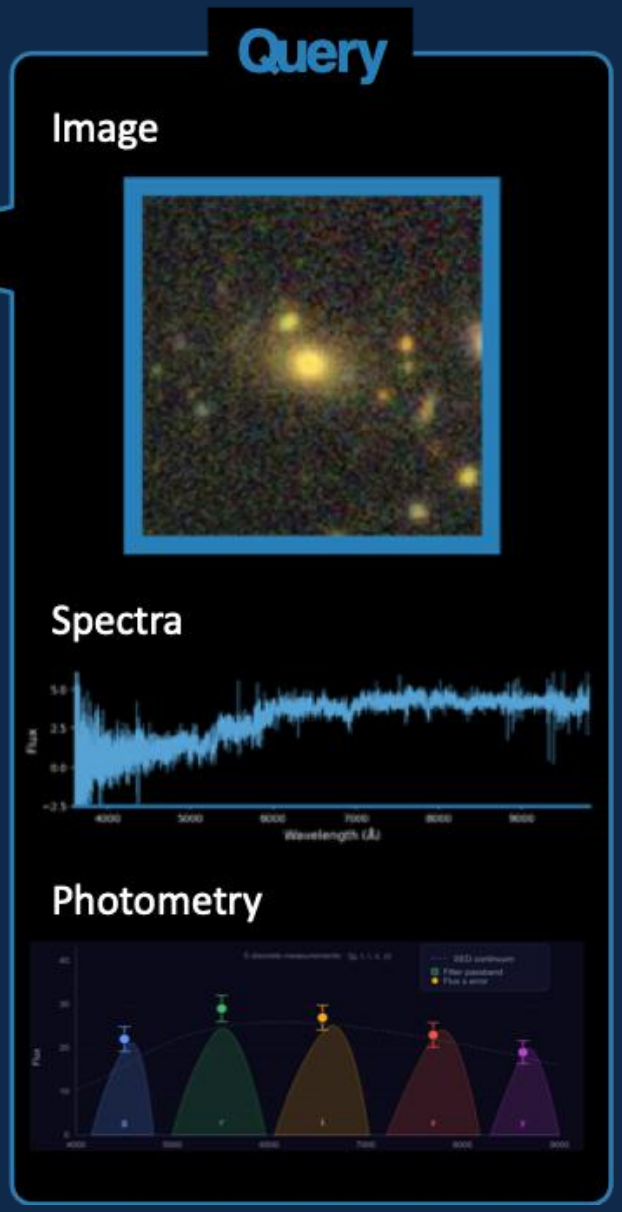
ResNet-50 / Zoobot: 0.6%

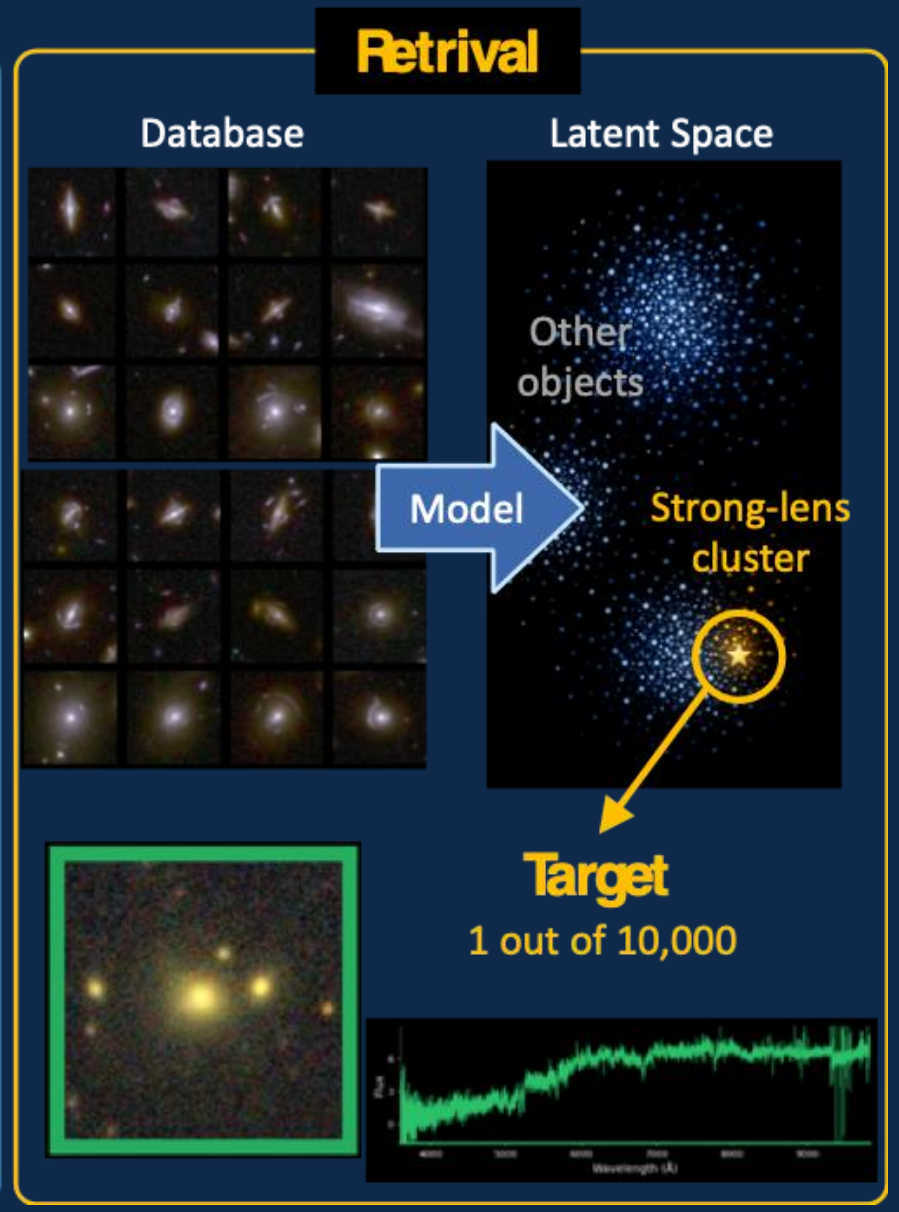
Supporting 4 CSST science centers

Strong lenses: retrieval in embedding space



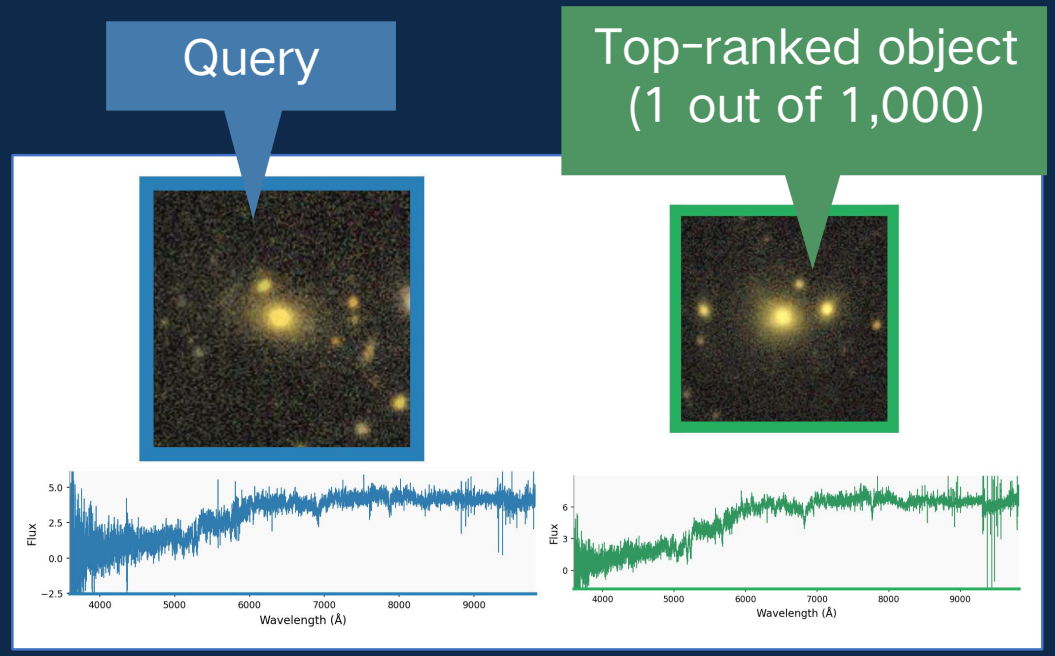






Initial result: 88.7% top-1 in a blind test

Standard image models: 0.6%. All-sky candidates generated; vetting has just begun



modality	Top-1 Rank Rate of true lens	Median Rank of true lens
ResNet-50: Image Only	0.6%	2285
GZ-Zoobot: Image Only	0.6%	1268
OneAstronomy: Image Only	25.2%	47
OneAstronomy: Spectrum Only	41.5%	4
OneAstronomy: Image + Spectrum	88.7%	1

53 blind sets: 1 true lens + 999 look-alike galaxies each · with Jean-Paul Kneib (EPFL)

Unified evaluation — Done
OneAstroBench lens task

→ **Blind test** — Done
Top-1 88.7% · baseline 0.6%

→ **Candidate generation** — Done
from 120 M objects

→ **Candidate vetting** — Started
help wanted, see next slide

Time domain: multimodal variable stars

之江实验室 ZHEJIANG LAB

Language Variable-type classification criteria, text descriptions

TESS 2-min cadence light curves

Gaia DR3 Multi-band photometry / parallax / astrometry

021SFM + Tools

- Text semantic understanding
- Time-series feature extraction
- Static feature extraction
- Cross-modal fusion

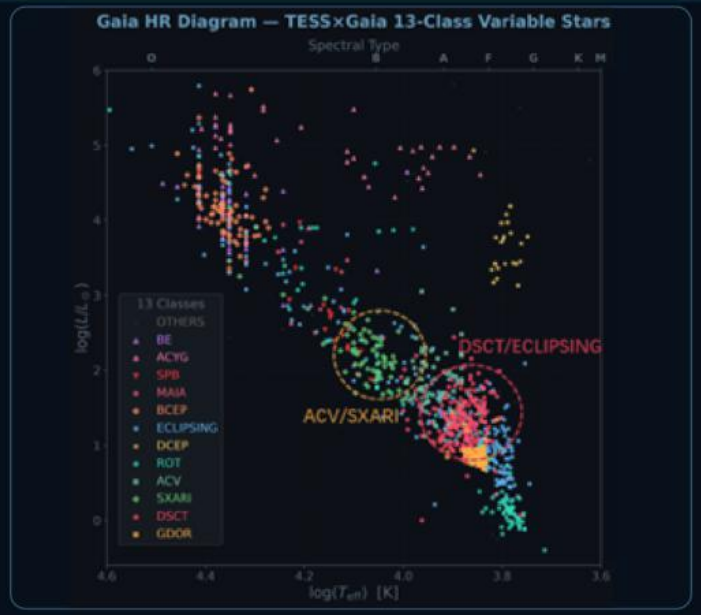
13-class candidate prediction

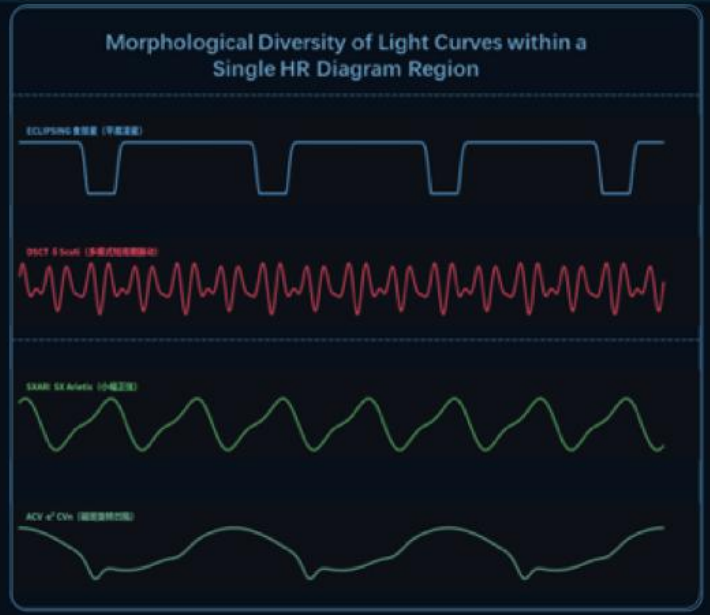
ROT · DSCT · GDOR
ECLIPSING · BCEP · SPB
ACV · SXARI
ACYG · BE · MAIA · DCEP

Astronomer review → scientific discovery

Exoplanet Search – Building a High-Purity Catalog of Exoplanet Candidates

...and other directions (magnetic field evolution, stellar interiors, distance ladder)





Classification Gains from Multi-Modal Fusion
Gaia Unimodal vs. Joint TESS+Gaia (Macro-F1)

Type	Gaia F1	TESS+Gaia F1	Gain Δ
ECLIPSING	0.5263	0.7945	+0.268
SPB	0.4615	0.6667	+0.205
SXARI	0.5714	0.7000	+0.129
ROT	0.7640	0.8506	+0.087
ACV	0.6061	0.6875	+0.081
BCEP	0.6957	0.7619	+0.066
DSCT	0.8485	0.9067	+0.058
GDOR	0.5750	0.6374	+0.062
MAIA	0.5455	0.5714	+0.026
Macro Avg.	0.6850	0.7420	+0.057

Fusion = Joint TESS + Gaia Multi-Modal; Gain Δ = TESS+Gaia F1 − Gaia F1

Key Gain: ECLIPSING

Gaia Unimodal F1 = 0.526 → Multi-Modal Fused F1 = 0.795
AUC% lacks time-domain support → Gaia unimodal upper limit is 68.5%
Time-domain fusion breaks the static performance ceiling, achieving statistically useful classification accuracy.

Data are a public good; AI capability is not — yet

Data: already open

SDSS, Gaia, DESI, TESS, Legacy Surveys ··· release their archives to everyone

Anyone can download petabytes, in principle

AI capability: concentrated

Compute, curated multimodal training sets and ML engineering sit with a few groups

Most teams rebuild their own pipelines, one survey at a time

OneAstronomy: make the capability itself open, reusable, verifiable and co-governed

Zhejiang Lab is a non-profit research institute committed to open science

What makes an AI model a public good

Open

Anyone can get it

Open weights by end of 2026

Reusable

Processed once, used by all

Embeddings for 126 M objects

Verifiable

Anyone can check the claims

OneAstroBench, open to every model

Co-governed

Community shapes the roadmap

Open roadmap and shared credit

What we don't do yet

Simulations

SBI, emulators, sim-to-real

Radio

SKA, FAST

More surveys

LAMOST, Euclid DR1, 4MOST, PFS, MUST, CSST

Uncertainty

Calibrated errors and abstention

Systematics

Fibre apertures, cross-survey shift

Thank you

Let's build it together.

Explore the Platform

Scan to open the OneAstronomy web workbench

https://oneastronomy.zero2x.org/

Contact: support.oneastronomy@zhejianglab.org