之江实验室
ZHEJIANG LAB

OneAstroBench

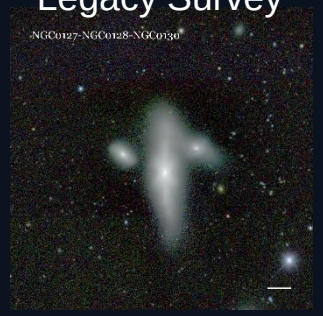
An Open, Science-Driven Benchmark for Multimodal AI in Astronomy

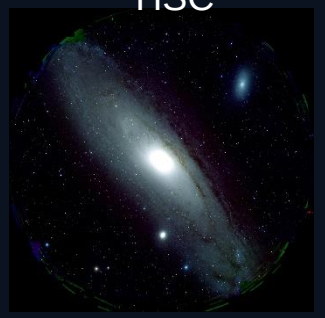
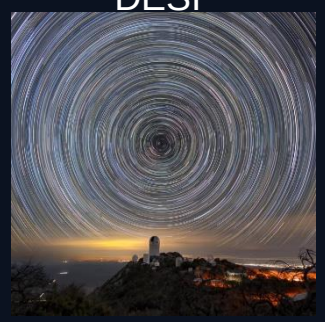
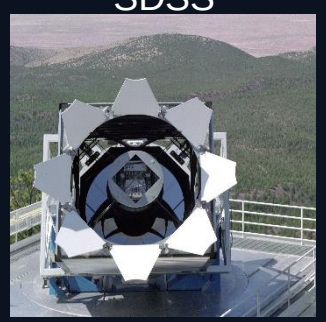
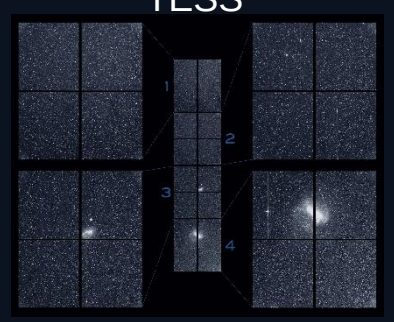
A collaborative work by the OneAstronomy Team
Presented by Jiaqi Zhao
Zhejiang Lab
Hangzhou, China · 20 October 2026

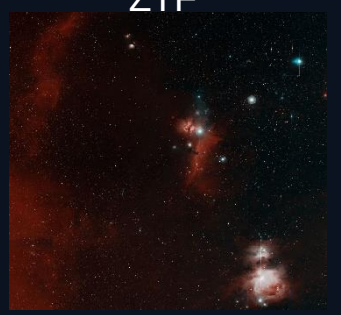
Astronomy in the Multi-Survey, Multimodal Era

之江实验室 ZHEJIANG LAB

Legacy Survey | HSC | DESI | SDSS | TESS | ZTF | Gaia






Images

Euclid · Perseus cluster · VIS + NISP colour composite

Spatial structure and morphology

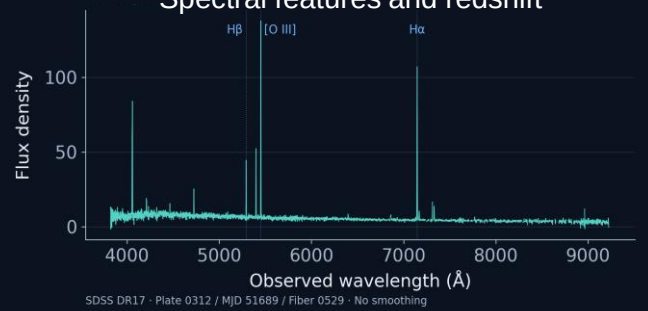


ESA/Euclid/Euclid Consortium/NASA · Processing: J.-C. Cuillandre, G. Anselmi

Spectra

SDSS · Galaxy · z = 0.08864

Spectral features and redshift

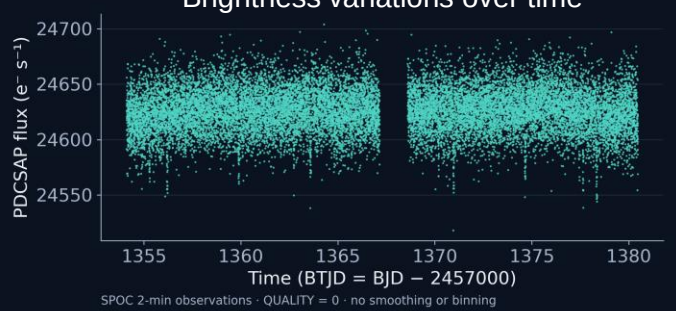


SDSS DR17 · Plate 0312 / MJD 51689 / Fiber 0529 · No smoothing

Light curves

TESS · L 98-59 / TIC 307210830 · Sector 2

Brightness variations over time

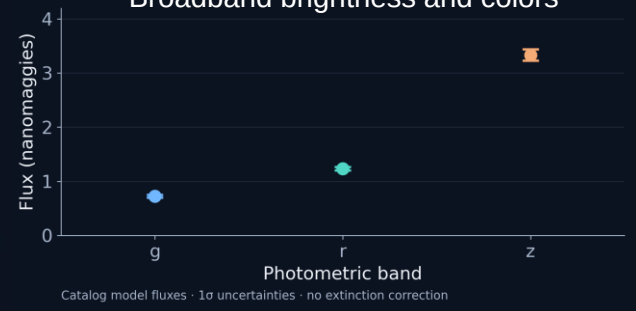


SPOC 2-min observations · QUALITY = 0 · no smoothing or binning

Photometry

Legacy Surveys DR10 · 0001p000-4 · g, r, z

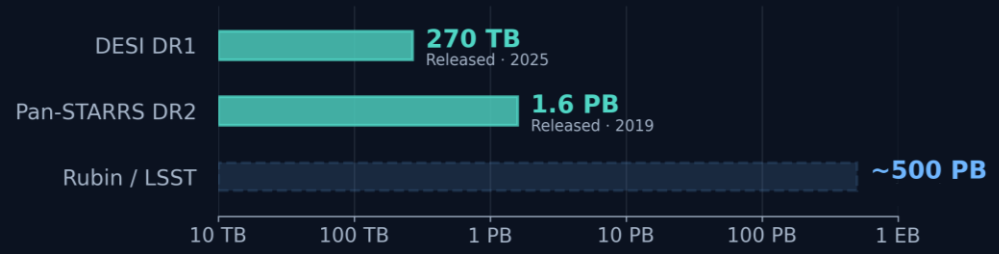
Broadband brightness and colors



Catalog model fluxes · 1σ uncertainties · no extinction correction

Survey data volumes: from TB to hundreds of PB

Released datasets and projected survey products



- **Growing datasets, richer observations**

Surveys expand both data volume and modality coverage.

- **Increasing use of multimodal inputs**

Astronomy models increasingly use multimodal observations.

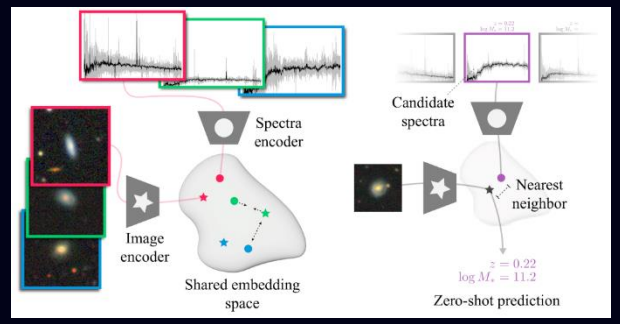
- **Scientific capabilities need systematic testing**

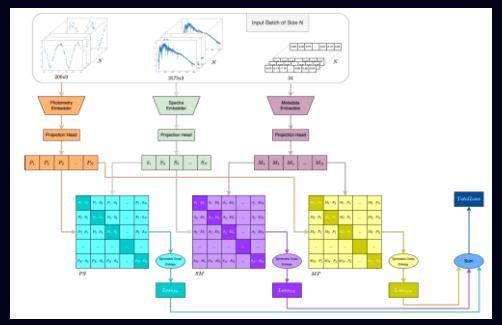
Cross-survey transfer, fusion benefits and rare-object discovery.

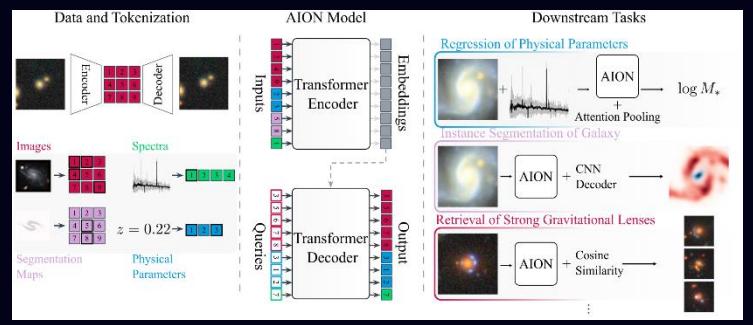
OneAstroBench: evaluating astronomy AI through scientific tasks.

Multimodal Models, Datasets, and Benchmarks

- **Multimodal Models & Agents**







Claude
deepseek
ChatGPT
智谱 BigModel
... ...

AstroCLIP
Images + Spectra
Cross-modal alignment

AstroM³
Light curves + Spectra + Metadata
Stellar representation learning

AION-1
Images + Spectra + Scalars
Cross-modal masked modeling

General-purpose Models & Agents
Data analysis and tool use
Scientific workflows

- **Datasets & Benchmarks**

Multimodal Universe	PLAsTiCC	**Astrobench**	**AstroMMBench**
Multimodal dataset	Single-task challenge	**Text-based benchmark**	**Vision-language benchmark**
Images, spectra, time series and metadata	Light-curve classification	Astronomy knowledge and reasoning	Astronomical image understanding and QA

OneAstroBench organizes astronomy tasks, multimodal data, models, and science-driven evaluation in one framework.

From Scientific Questions to Benchmark Design

Start with the scientific claim, then define the data, task and evidence needed to test it

01 Scientific question

What scientific claim should the model support?

Transfer	Across surveys
Fusion	Added value
Discovery	Rare targets

02 Task package

What data and controls make the task valid?

Objects	Target population
Inputs	Surveys + modalities
Protocol	Labels + fixed splits

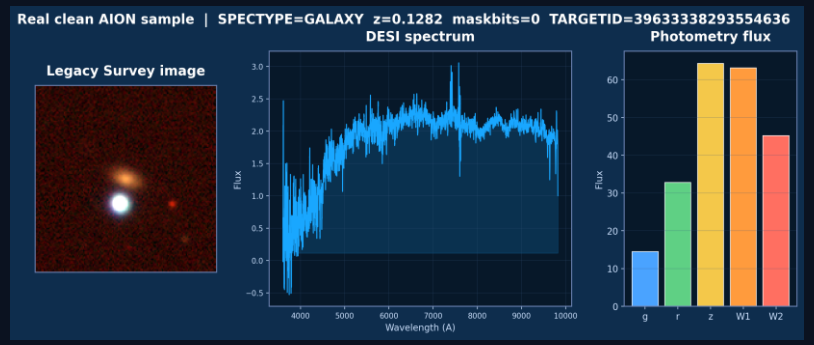
03 Evaluation evidence

What result would count as useful evidence?

Comparison	Right baseline
Metric	Matches the goal
Uncertainty	Limits the claim



- Spectrum Generation from Images



- Estimate physical properties



- Retrieve rare objects from large samples

Cross-survey transfer

Hold out another survey

Measure performance after the survey changes

Multimodal value

Use the same test objects

Compare with the strongest single-input model

Rare-object discovery

Build a low-prevalence gallery

Report precision and enrichment

Scientific questions define the task, the comparison and the metric.

OneAstroBench

- **29 registered tasks spanning 9 scientific domains**

01 SURVEYS AND OBSERVATIONAL INPUTS

Legacy Surveys / DECaLS · DESI · SDSS / SEGUE · HSC · Euclid · Gaia · TESS · Kepler · ASAS-SN · LAMOST · ZTF · SDO

Images	Spectra	Photometry / SEDs	Light curves	Astrometry / Catalogs	Text

Also used where the task requires them: Gravity Spy, RadioGalaxyDataset, GREAT3, AstroVLBench, astronomy QA.

Coverage varies by task.

02 SCIENTIFIC DOMAINS AND TASKS

Cosmology and large-scale structure
- **T08** Legacy–DESI photo-z
- **T10** Weak-lensing shear
- **T17** HSC photo-z
- **T18** Euclid photo-z

Stars and Galactic archaeology
- **T05** SSPP stellar parameters
- **T19** DESI MWS stellar parameters
- **T27** Euclid stellar radial velocity

Astrometry and celestial mechanics
- **T24** Gaia cluster membership

Galaxy formation and evolution
- **T03** Galaxy physical parameters
- **T04** GZ10 galaxy morphology
- **T20** Euclid galaxy morphology
- **T25** Metal-poor galaxy retrieval
- **T28** HSC Sérsic index
- **T29** HSC morphology classification

High-energy and compact objects
- **T12** Radio morphology and host matching
- **T13** GW detector glitch classification

Solar physics and space weather
- **T14** Solar flares and active regions

Time-domain and multi-messenger
- **T02** TESS variable-star classification
- **T09** Transient and supernova classification
- **T21** Multimodal variable-star classification
- **T22** Time-series representation and anomalies

Exoplanets and habitability
- **T11** Transit vetting

Solar system and small bodies
- **T23** Small-body classification and retrieval

Cross-domain tasks
T01 Star / galaxy / QSO · **T06** Strong-lens retrieval · **T07** Cross-modal generation and retrieval · **T26** Euclid object classification

These four are not placed inside a single domain.

Knowledge and reasoning
T15 Astronomical knowledge and reasoning · **T16** Observational reasoning

03 EVALUATION FOCUS

Physical-property estimation
Infer physical parameters from observations

Object classification
Identify astronomical sources and events

Cross-modal prediction
Predict one observation from another

Cross-survey transfer
Generalization under survey shifts

Multimodal gains
Compared with the strongest single modality

Rare-object retrieval
Recall, enrichment, and candidate purity

Shared task definitions, controlled comparisons, and task-specific metrics. Agentic science is a cross-cutting track, not a pooled score.

Scientific Questions and Task Design

- **Three examples of turning scientific goals into measurable tasks**

	Spectral Prediction	Galaxy Properties	Strong-Lens Retrieval
Scientific question	Can other observations predict a spectrum?	How well can observations constrain galaxy properties?	Can we prioritize rare lens candidates?
Inputs	Images + photometry	Images / spectra / photometry	Query + candidate images
Model output	Generated Spectrum  Illustrative	Estimated physical parameters 	Ranked candidate list  Illustrative ranking
Evaluation	Spectral reconstruction error	Parameter-estimation accuracy	Recall, purity and enrichment

The scientific question determines the inputs, targets and evaluation metrics.

Representative tasks; detailed protocols and results follow.

OneAstroBench: Data and Task Packages

- **Data Contribution and versioned task packages**

01 Public data resources

Survey observations

Representative sources

Legacy Surveys · DESI · SDSS

HSC · Euclid · Gaia

TESS · ASAS-SN · LAMOST · ZTF

Scientific task datasets

Representative sources

GZ10 · ProVABGS · SSPP

AstroM3 · StarEmbed · Gravity Spy

Record sources, releases and access conditions

02 Sample construction



- Associate observations
- Apply quality selection
- Record label provenance
- Track available modalities

Enable multimodal tracks only with reliable pairings

03 Versioned task packages

- Scientific objective
- Inputs and reference labels
- Object manifests and fixed splits
- Input / output schemas
- Scoring rules and metrics

Separate tracks for different input conditions

Task cards · Baselines · Result records

Each task links a scientific objective to defined inputs, reference labels, and evaluation conditions.

Galaxy-Property Estimation



之江实验室
ZHEJIANG LAB

ProVABGS: mean R^2 over five prediction targets

Mean R^2 over redshift, stellar mass, metallicity, age, and specific star-formation rate. Axis starts at 0.

OneAstronomy-LC | OneAstronomy-v1.0 | AION | Task baseline

Image

ResNet18	0.591
AION-L	0.569
AION-B	0.567
OA-v1.0	0.565
OA-LC	0.563
AION-XL	0.559
AstroCLIP	0.559

Spectrum

OA-LC	0.754
OA-v1.0	0.740
AION-L	0.739
Spectrum CNN	0.737
AstroCLIP	0.737
AION-B	0.733
AION-XL	0.722

Photometry

Photometry MLP	0.512
XGBoost	0.495

Image + photometry

AION-L	0.666
AION-B	0.664
OA-v1.0	0.663
OA-LC	0.646
AION-XL	0.635
AstroCLIP	0.571

Spectrum with another input

AstroCLIP I+S	0.743
AstroCLIP S+P	0.750

Image + spectrum + photometry

OA-LC	0.772
OA-v1.0	0.762
AION-L	0.758
AION-B	0.757
AION-XL	0.747
AstroCLIP	0.746

Five targets, OneAstronomy-LC

Spectrum only 0.754 Full inputs 0.772

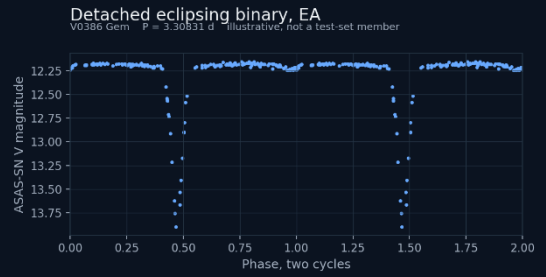
	Spectrum	Full
Redshift	0.995	0.994
Stellar mass	0.901	0.958
Metallicity	0.639	0.639
Age	0.501	0.521
sSFR	0.737	0.749

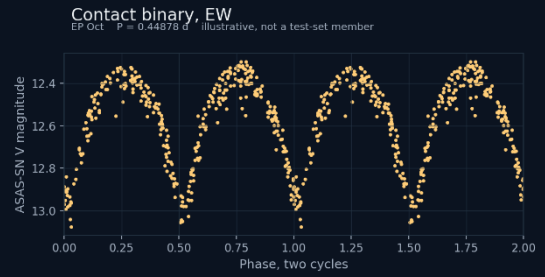
Spectrum carries most of the score. Adding image and photometry raises the mean by 0.018, mainly through stellar mass.

Time-Domain Tasks

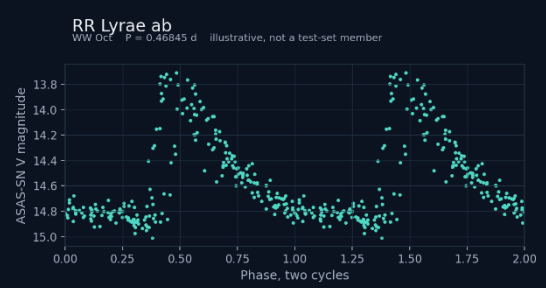
之江实验室 ZHEJIANG LAB

Variable-star classification with complementary observations









6, 7 and 12 classes · macro-F1

Hand-crafted forest, OneAstronomy-LC, and four Chronos-2 heads.

	Forest	OA-LC	LogReg	MLP	RF head	XGBoost
6 classes	0.816	0.820	0.993	0.818	0.808	0.802
7 classes	0.660	0.665	0.611	0.613	0.531	0.547
12 classes	0.519	0.524	0.432	0.387	0.353	0.347

ASTRAFier · TESS

8 classes, aux test, N = 1,859.

Macro-F1	**0.883**
Accuracy	**0.879**

Registered score 0.886.

MAVEN · ZTF supernovae

5 classes, 4,168 objects. Macro-F1.

Recorded score	**0.417**

Light curve 0.419. Spectrum 0.414. The 3-class kNN, 0.687.

BTSbot · ZTF, SN vs not

AUC, separate from the MAVEN score.

Image CNN	0.974
Metadata	0.992
Image + metadata	0.992
Three-seed ensemble	0.993

StarEmbed · ZTF

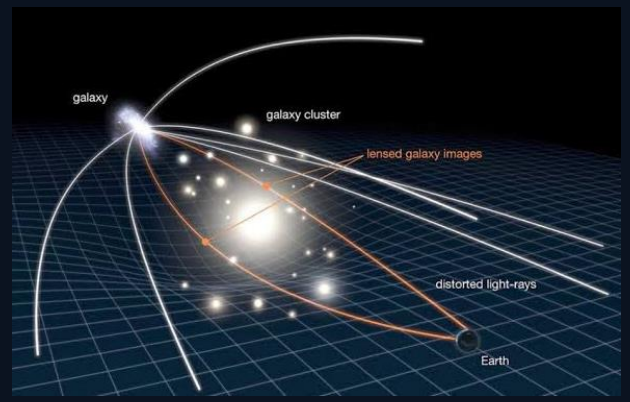
Top-10 purity, axis 0–1

Hand-crafted	Time-MoE	Chronos-Bolt	Moirai	Chronos-T5	Astromer-1	Astromer-2
0.230	0.204	0.191	0.154	0.135	0.127	0.121

Forest and Chronos-2 share their own two-split mean. OA-LC is a third split of the same class lists.

Strong-Lens Retrieval: Baselines and OA

Left: mean P@10 over image, image+photometry and image+photometry+spectrum. Right: Legacy S1 and HSC S1 are separate protocols.





Spectral library · P@10 × 100

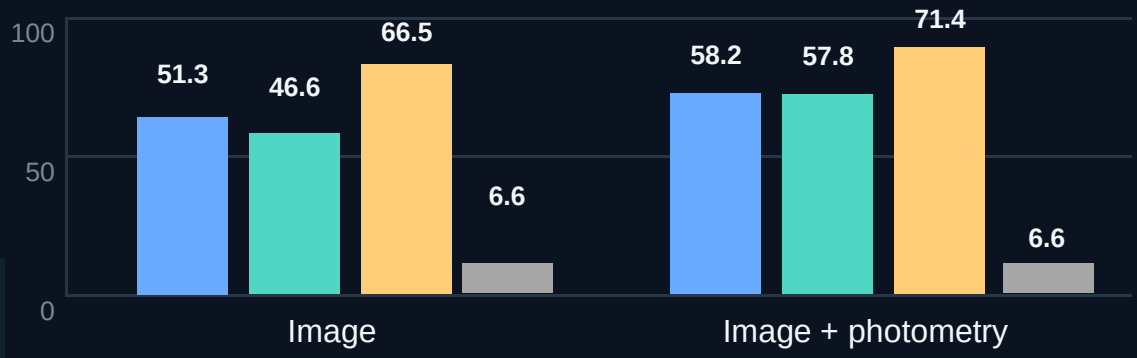
Model	Image	+ photometry	+photometry+spectrum
MJEPA	41.9	40.2	38.3
OneAstronomy-v1.0	20.2	30.0	38.3
OneAstronomy-LC	21.7	26.4	36.4
AION-XLarge	20.8	23.0	32.6
AION-Base	17.9	23.0	29.8
AION-Large	17.4	22.1	29.3

Image-only baselines on the same gallery. Axis 0–50.

ZooBot	5.5		
ResNet-50	2.3		
Random rank	0.6	0.6	0.6

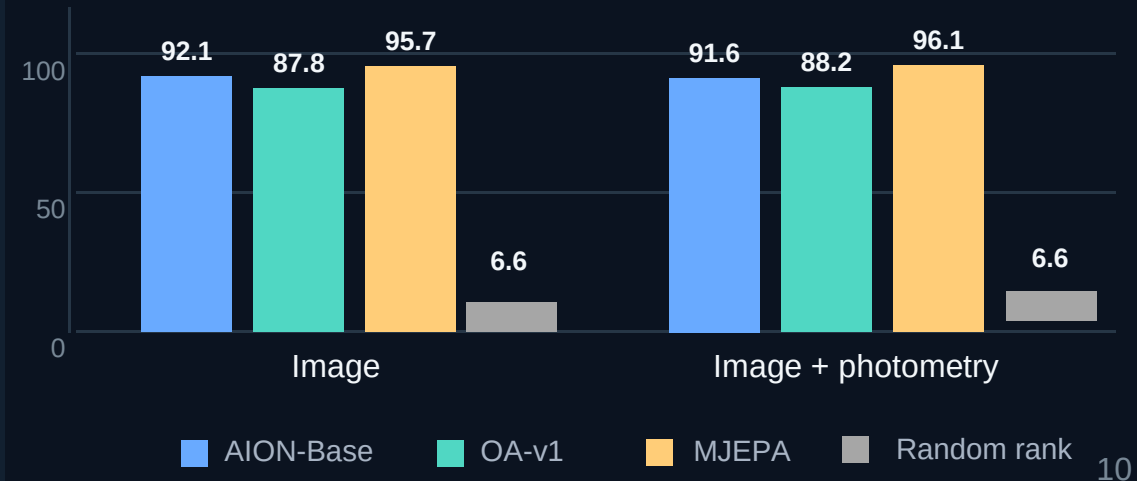
Precision@10 on Legacy S1

500 queries, 107,028 objects
7,028 lenses + 100,000 field sources



Precision@10 on HSC S1

100 queries, 21,417 objects
1,417 lenses + 20,000 field sources



Cross-survey prediction and Euclid transfer

Generate a DESI spectrum from images and photometry. Euclid uses a frozen encoder and a newly trained probe.

之江实验室 ZHEJIANG LAB

Generate DESI spectra

Pearson measures shape. T7 = 0.6 r + 0.4 NCC, scaled to 100.

	Pearson	R^2	T7
Legacy Survey · 3,185 pairs			
AION-Base	0.641	0.161	64.3
OA-v1.0	0.645	0.173	64.9
OA-LC	0.552	0.133	56.3
MJEPA	0.638	0.269	64.2

MJEPA uses the flow decoder on 3,185 Legacy pairs.

HSC · 72,731 pairs			
AION-Base	0.598	0.176	60.2
OA-v1.0	0.656	0.272	65.7
OA-LC	0.517	−0.136	52.9

Frozen encoder. The probe is trained on Euclid labels.

Morphology · macro-F1 · N = 8,086

ZooBot	0.831
OA-v1.0	0.787
AstroPT	0.728

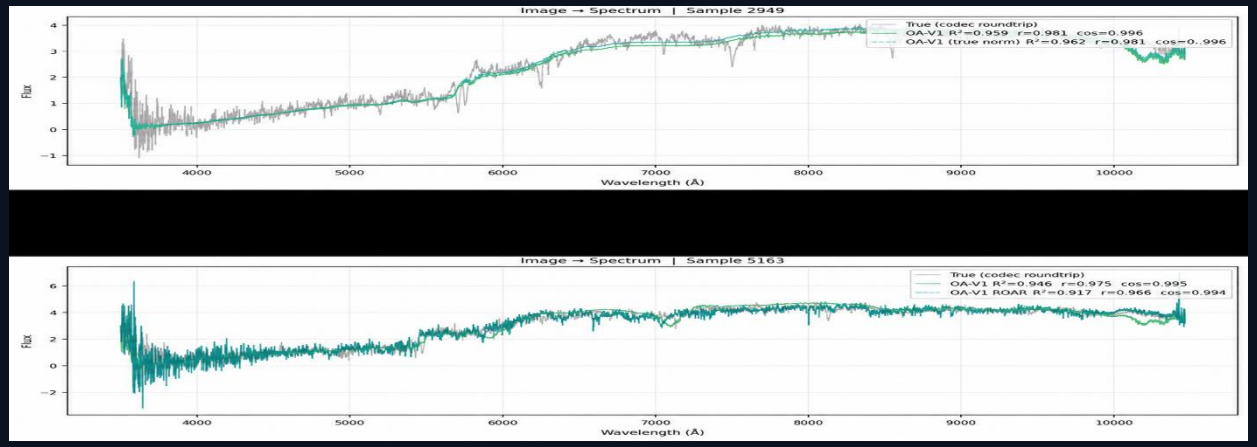
Star / galaxy / QSO · macro-F1 · N = 15,000

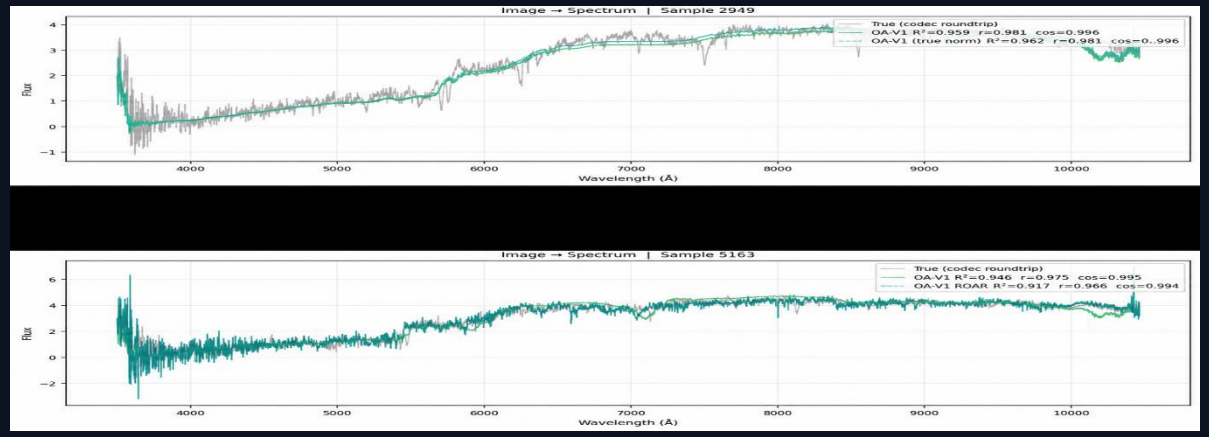
OA-v1.0	0.430
AstroPT	0.421

Photo-z · NMAD, lower is better · N = 15,000

TPZ 0.673 AstroPT 0.681 OA-v1.0 0.682

The outlier rate is about 80% for all three models.





Selected high-score examples. Gray is the DESI spectrum after codec reconstruction. The medians above are for the full sample.

Multimodal Gains Vary Across Scientific Tasks

Add complementary observations, keep the protocol fixed, and compare within each task

REGRESSION Galaxy properties



Mean R^2 · ProVABGS

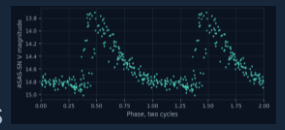
Spectrum	0.7371
Spectrum + phot.	0.7504
Image + S + P	0.7461

+0.013 mean R^2

AstroCLIP-based comparison

Images do not improve the best fusion.

VARIABLE STARS AstroM3



Macro-F1 · same 2,225 objects

Photometry	0.9009
All modalities	0.9340

+0.0331 Macro-F1

All available inputs are combined.

RETRIEVAL Strong lenses



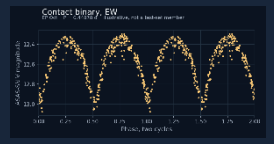
Precision@10 · fixed S1 gallery

Image	0.4662
Image + phot.	0.5778

+11.16 percentage points

Photometry improves ranking.

VARIABLE STARS TESS + Gaia



Macro-F1 · OA-USFM-LC · seed 42

Wang-6	TESS 0.815 → +Gaia 0.820	+0.0047
Wang-7	TESS 0.646 → +Gaia 0.665	+0.0193
Gao-12	TESS 0.479 → +Gaia 0.524	+0.0452

TRANSIENTS ZTF BTSBot



AUC · SN versus non-SN

Image CNN	0.9736
Metadata HGB	0.9917
Image + metadata	0.9924

+0.0188 AUC

vs image-only

Same BTSbot protocol.

AUC axis: 0.95–1.00

Multimodal value is task-dependent: compare gains within a task, not across different metrics.

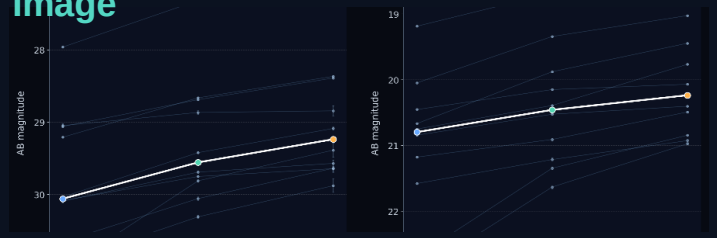
Protocols for Fair and Reproducible Evaluation

Six checks define which scientific claims a result can support

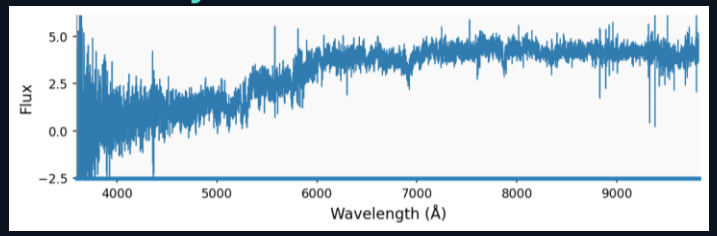
Observed inputs



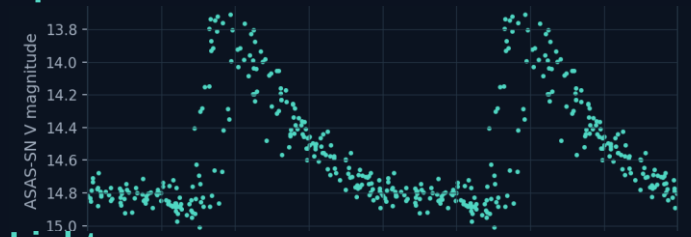
Image



Photometry



Spectrum



Light curve

Examples illustrate the input types. They are not one matched object.

Evaluation checks

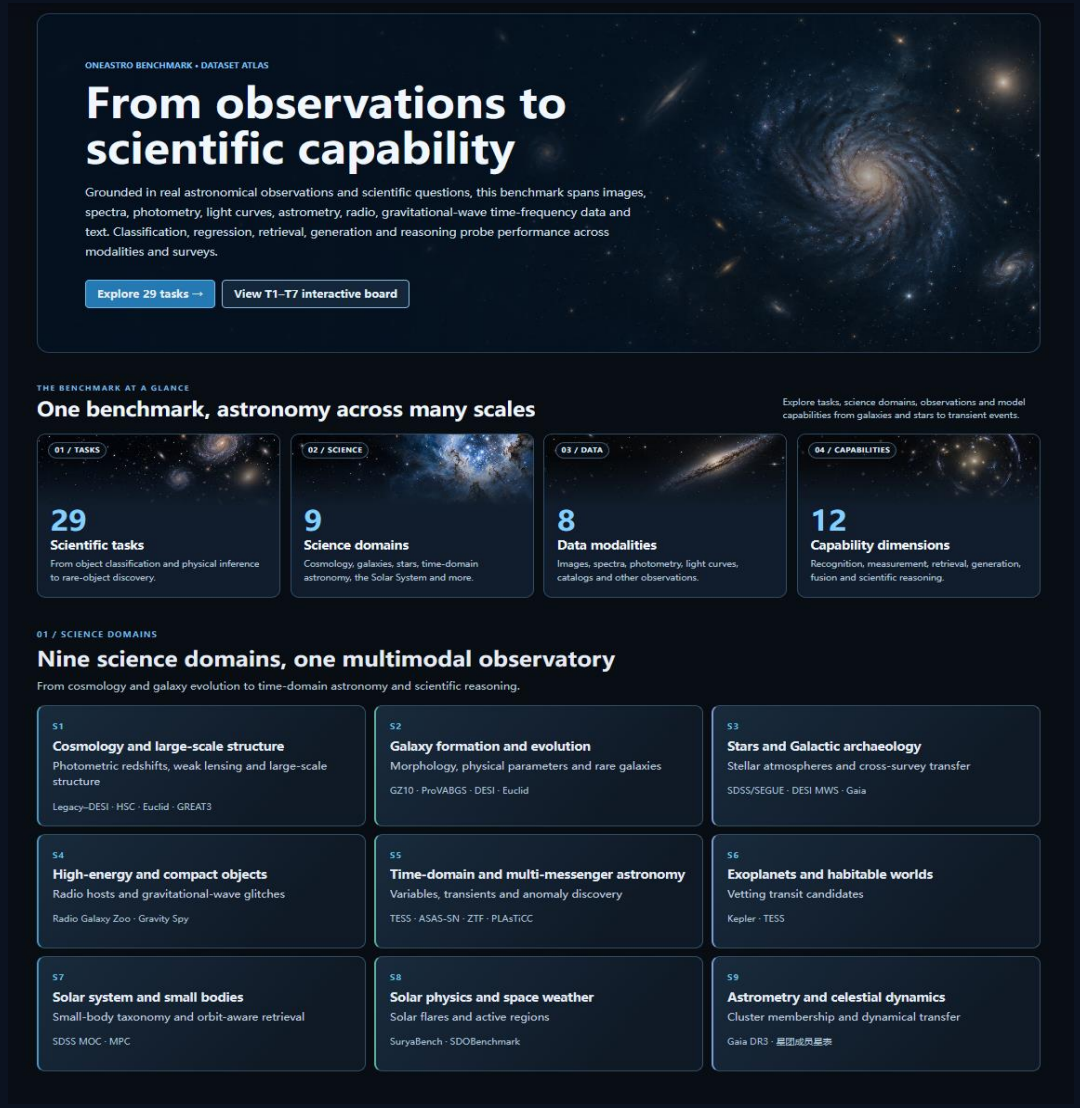
01	**Bind observations**	Group bands, repeats, spectra and light curves by object.
02	**Audit the split**	Hold out complete objects, clusters or regions.
03	**Check pretraining overlap**	Label unknown overlap instead of treating it as clean.
04	**Lock the test sample**	Use the same IDs, labels and retrieval gallery.
05	**Fix the adaptation track**	Keep frozen probes, fine-tuning and native recipes separate.
06	**Report uncertainty**	Use repeated seeds or paired intervals with the right random baseline.

Publish every score with its task, protocol, adaptation track and domain.

A failed check narrows the claim or keeps the result off the strict board.

Contributions and the Next Release

What OneAstroBench adds, and what the next public snapshot must complete



A working interface for tasks, protocols and recorded results

Four contributions

01 Scientific organization
Tasks begin with a scientific question.

02 Reusable task packages
Inputs, labels, splits and scorers stay versioned.

03 Comparable evidence
Transfer, fusion and retrieval use explicit controls.

04 Visible boundaries
Each score stays attached to its task, survey, and range of use.

Next release

01	**Release**	Freeze a labeled snapshot with access paths and licences.
02	**Expand**	Complete missing tasks, baselines and modality controls.
03	**Tighten**	Calibrate Euclid and document transfer conditions.

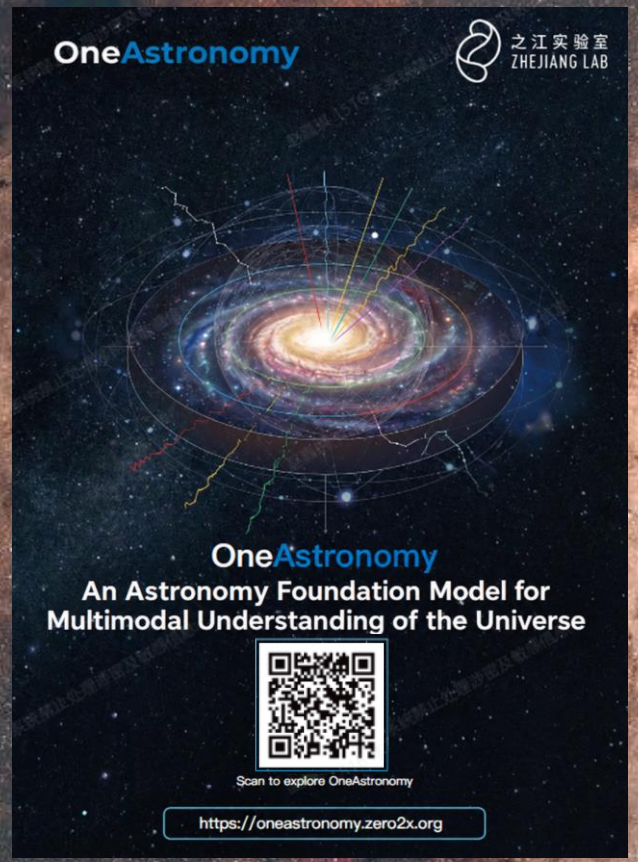
The next release should increase coverage and comparability together.

THANK YOU

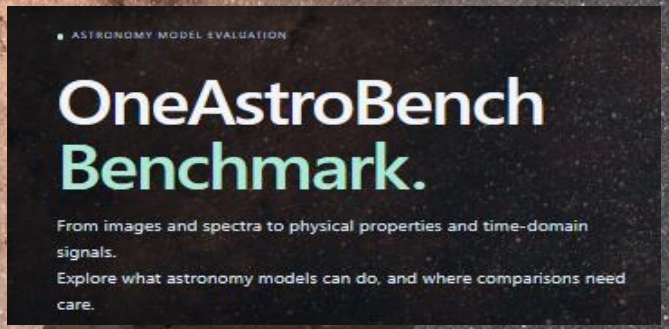
OneAstroBench

An Open, Science-Driven Benchmark for Multimodal AI in Astronomy

Use the benchmark · Submit your results · Contribute new tasks







Coming soon to the zero2x platform

Jiaqi Zhao
zhaojq@zhejianglab.org

Credit: J. Rodrigues/ESO