On the hunt for XMPGs

Introducing a three-stage pipeline for eXtremely Metal-Poor Galaxy screening

EPFL team: Aurélien Verdier, Antoine Rocher, Timour Jestin, Jean-Paul Kneib

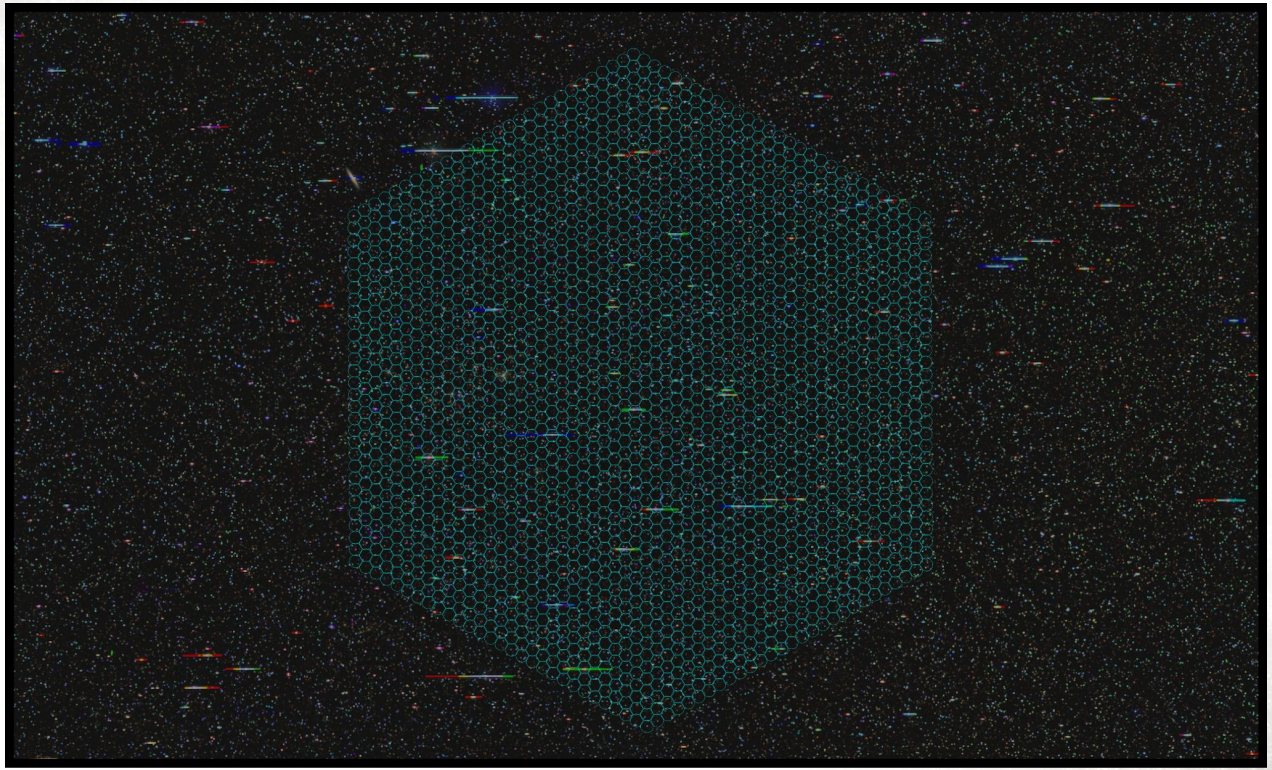
ZJ Lab team: Zhichao Zheng, Chiye Li, Hao Chen, Jiaqi Zhao

EPFL

之江实验室 ZHEJIANG LAB

Why eXtremely Metal Poor Galaxies ?

A surprising encounter



Some characteristics of XMPGs

- **Low metallicity** $Z < 0.1Z_{\odot}$
  (in some extreme cases $Z \lessapprox 0.02Z_{\odot}$)
  Sui et al. 2026; Kunth & Östlin 2000

- **Low mass,** $10^7 - 5 \cdot 10^9 \ M_{\odot}$
  Zou et al. 2024; Senchyna & Stark 2019

- **High specific SFR,** $10 - 100 \ Gyr^{-1}$
  Zou et al. 2024; Senchyna & Stark 2019

Their usefulness

- **Low metallicity** $Z < 0.1Z_{\odot}$
  (in some extreme cases $Z \lessapprox 0.02Z_{\odot}$)
  Sui et al. 2026; Kunth & Östlin 2000

- **Low mass,** $10^7 - 5 \cdot 10^9 \ M_{\odot}$
  Zou et al. 2024; Senchyna & Stark 2019

- **High specific SFR,** $10 - 100 \ Gyr^{-1}$
  Zou et al. 2024; Senchyna & Stark 2019

- **Thought to be analogues for early galaxies**
  Act as local laboratories for early galactic evolution.
  Anchors to interpret JWST data.

- **Extreme emission line XMPGs as cosmological tracers**
  As part of DESI Run 2 target selection

- **Rare objects** - rare properties ?
  Current XMPG counts are so low, finding more might uncover unknown unknowns

Finding XMPGs photometrically

Three-Stage Screening and Ranking of XMPG Candidates

Design rationale: retain potential targets at low cost, refine selection with complementary image and photometric information, then prioritize for limited observing slots.

Stage 1: Efficient large-scale screening

32-dimensional catalog features
XGBoost high-recall screening

- Filter out non-target while retain potential XMPGs
- Reduce the Stage 2 workload

Stage 2: Foundation-model multimodal fine screening

g/r/z images + catalog features
Cross-attention

- Combine morphology and photometry
- Improve identification with detailed image features

Stage 3: VLM-assisted re-ranking

Vision–language model
review and recommendation

- Improve purity at the top of the list
- Adapt criteria to the scientific task

Output: candidates ranked by priority for expert review and spectroscopic follow-up.

Three-Stage Screening and Ranking of XMPG Candidates

Stage 1 — High-recall photometric screening

32D photometric feature vector
Input features: photometry, colors, morphology, Gaia, S/N, photo-z

↓

XGBoost binary classifier

↓

P_1D (probability of XMPG)

↓

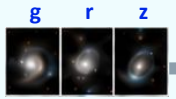
screening threshold: **P_1D threshold**

↓

candidate pool (high-recall)

Stage 2 — Multimodal model

Image branch (g/r/z images)



DR11 g/r/z image [B, 3, 128, 128]

→ **AION-1 image tokenier**

→ 576 spatial tokens [B, 576, 512]

→ Linear projection 512 → 128 + 24 × 24 2D positional encoding

→ image tokens [B, 576, 128]

Scalar-condition branch

Full 32D scalar features → 32 condition tokens [B, 32, 128]

Cross-Attention

Query = image tokens
Key/Value = condition tokens

→ **Residual addition**

→ Restore 24 × 24 spatial grid [B, 128, 24, 24]

24 × 24

Spatial refinement and feature extraction

Two 3 × 3 convolution layers (spatial refinement)

Extract five 128D representations:
- Central 8 × 8
- Outer region
- Global average
- Global maximum
- Mean-pooled condition tokens

→ Concatenate 640D

→ Fusion MLP 640 → 256 → 128

→ Classification head 128 → 64 → 1

→ Sigmoid output P_2D

Model-level fusion

$$P_{final} = c \times P_1D + (1 - c) \times P_2D$$

→ **Frozen threshold or Top-K ranking**

Stage 3 — VLM-assisted review and re-ranking

Top-300 candidates

↓



Diagnostic g/r/z images
+ image-quality information
+ nearest-neighbor references

↓

VLM review

Nature Science Foundation Model

↓

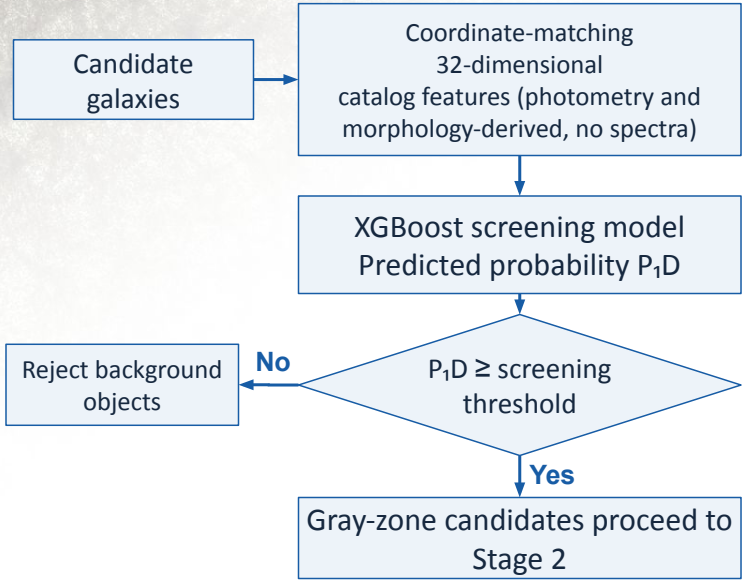
Rubric: centeredness, cross-band consistency, morphology, contamination, image quality

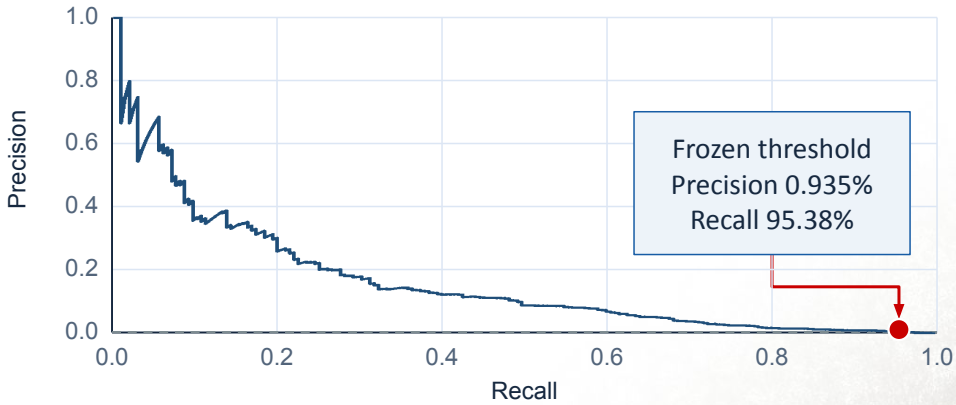
↓

Re-ranked Top-100 high-priority candidates

Stage 1: Photometry-Only High-Recall Screening

Fast screening with a low threshold to retain potential XMPGs





A low threshold retains targets: recall 95.38%

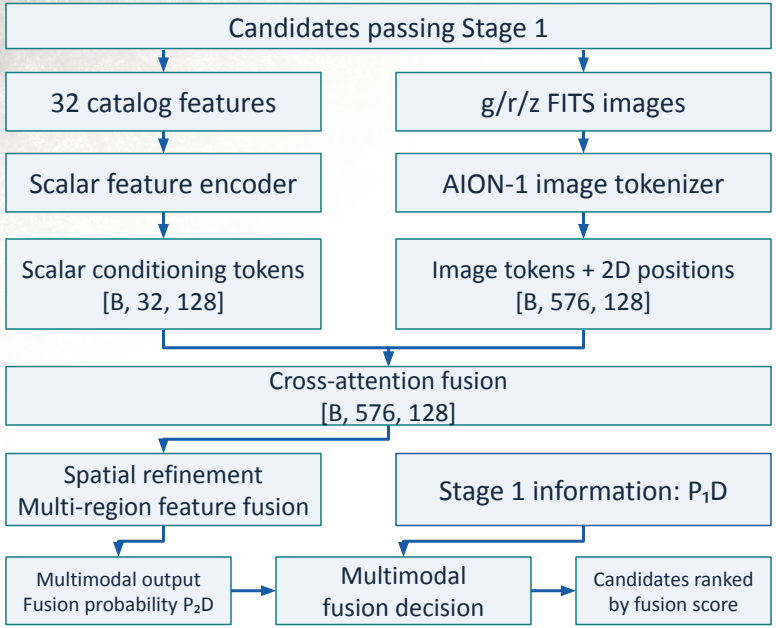
Excludes ~90% of non-target galaxies

Training set: DESI Main Survey, 1,214 positives + 60,700 negatives (negative/positive = 50:1).
Test set: DESI Survey Validation (SV1–SV3), 195 positives + 194,805 negatives (negative/positive = 1000:1).
Positives are based on spectral analysis report in Jipeng Sui *et al* 2026 *AJ* **171** 351
Negatives include direct-T_e-confirmed negatives and regular spectroscopic reference negatives.

TP	FN	FP	Recall	FPR	Precision	Candidates	Selected fraction
186	9	19,706	95.38%	10.12%	0.935%	19,892	10.20%

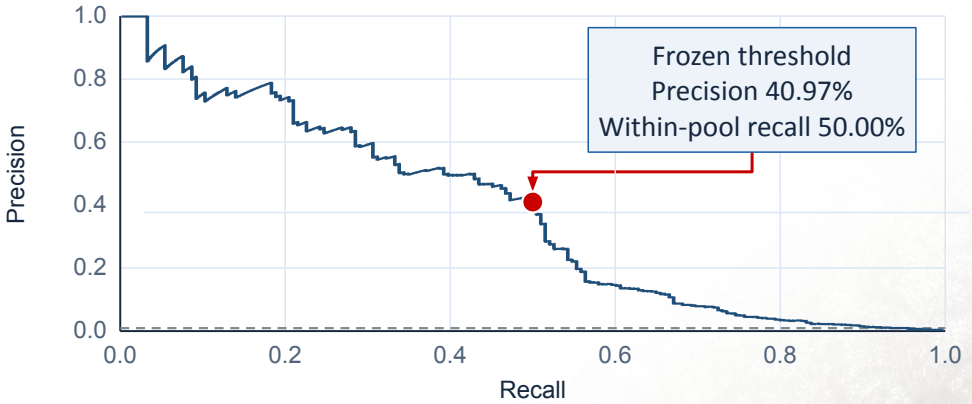
Stage 2: Multimodal Fine Screening

Combine photometric and image features; select and rank by the multimodal fusion score



PR curve AP = 0.3770

After-stage 1: 19,892 objects; 186 positives (0.935%)



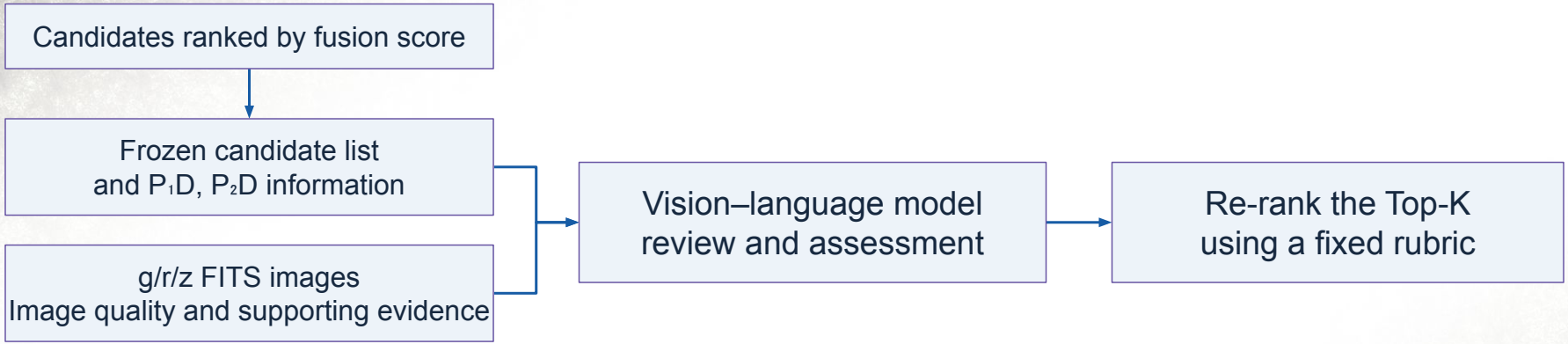
Within-pool (Stage 2) recall = 93 / 186 = 50.00%

End-to-end recall (Stage 1&2) = 93 / 195 = 47.69%

Objective	TP	FN	FP	Precision	End-to-end recall	F1	F2
F2 priority	93	102	134	40.97%	47.69%	0.4408	0.4618
F1 priority	80	115	79	50.31%	41.03%	0.4520	0.4260

Stage 3: VLM Review and Top-K Re-ranking

Images and model evidence enter the VLM; fixed-rubric scoring re-ranks candidates



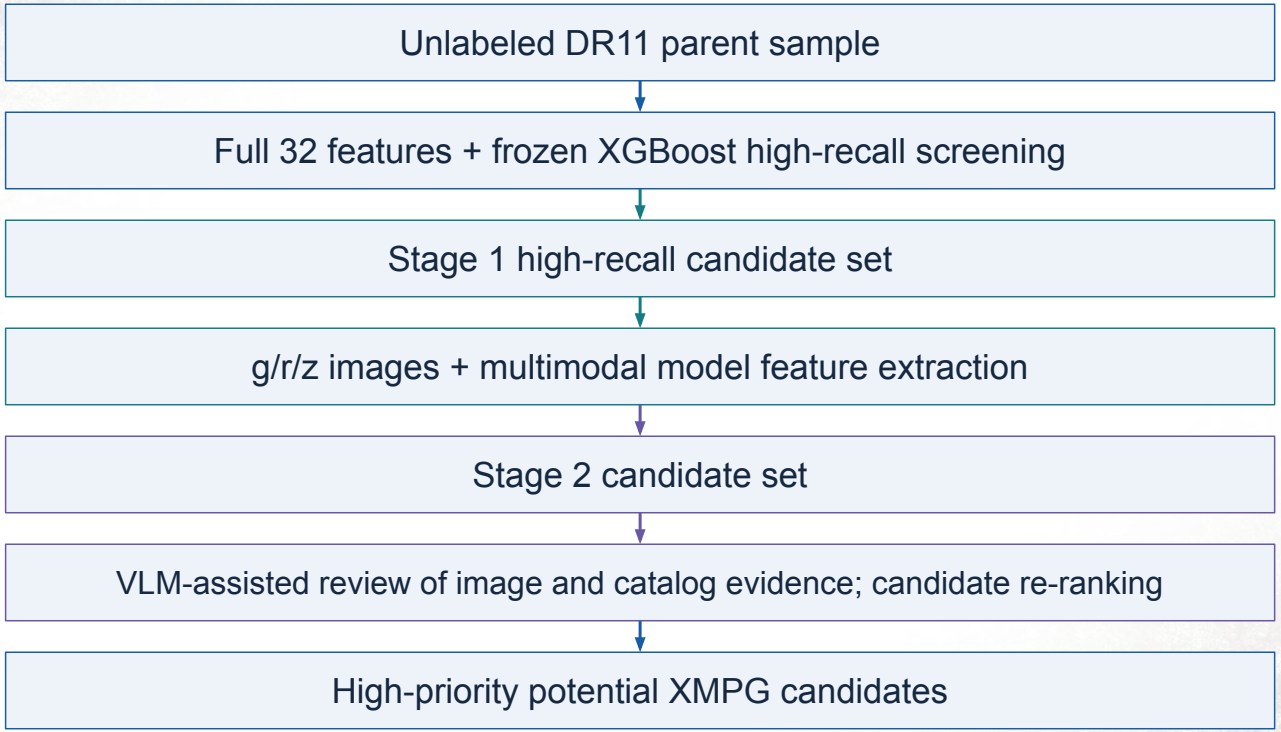
For candidates retained by Stages 1–2, the VLM reviews image quality, morphology, cross-band consistency and contamination using a fixed rubric. It combines quality score with model evidence to re-rank the list, enriching targets in the limited Top-K sample for expert review and spectroscopic follow-up.

Model	TP / FP	Precision	F1	F2
Original Top-100 (after Stage 1&2)	61 / 39	61%	0.3961	0.3273
GPT-5	65 / 35	65%	0.4221	0.3487
021SFM VL Thinking	66 / 34	66%	0.4286	0.3541

Outlook

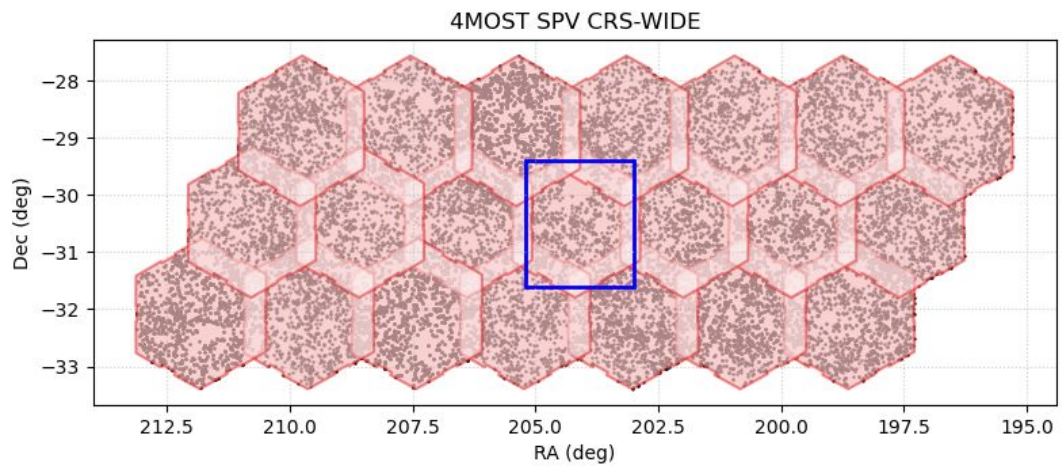
Application to 4MOST

Apply pre-trained models to Legacy Surveys DR11 to generate candidates for spectroscopic follow-up



4MOST Pilot Application

SPV CRS-WIDE pilot field



Selected region: RA 203.0°–205.2°, Dec −31.6°–−29.4°
Area: 4.17 deg², close to one 4MOST field of view

Potential applications:

- **Broad scientific applications**

Adjusting prompts and scoring criteria could extend this approach to other scientific applications.

- **Instrument-specific ranking**

Include instrument specifications and observing conditions in VLM prompts. Combine candidate images and catalog information to assess expected detectability with a given instrument under specified conditions, then prioritize candidates for observation.

Pilot application: 434,566 quality-controlled sources, an empirical XMPG fraction of 0.1%, guides the final Top-435 candidate list.
Output: Candidate CSV, images and model evidence provided for review and spectroscopic follow-up.

4MOST Pilot Application

SPV CRS-WIDE pilot field

- Yields **2 cross-matches** (d<2”) with 4MOST SPV data
- Both are from the **Bright Galaxies** sample

4MOST Pilot Application

SPV CRS-WIDE pilot field

- Yields **2 cross-matches** (d<2”) with 4MOST SPV data
- Both are from the **Bright Galaxies** sample

Conclusion

- We demonstrate a **scalable strategy** for identifying **high-priority XMPG candidates**
  - The first stage allows for efficient exclusion of non-candidates while maintaining high recall (up to 95%)
  - The 2nd and third stage then increase the precision, at the cost of recall

- First application to 4MOST data proves promising, candidates could be included as **spare targets** for spectroscopic surveys (eg. DESI Run 2, MUST, 4MOST)

- Next steps: confirm spectroscopic characteristics & refine selection efficiency for large areas

Thank you !