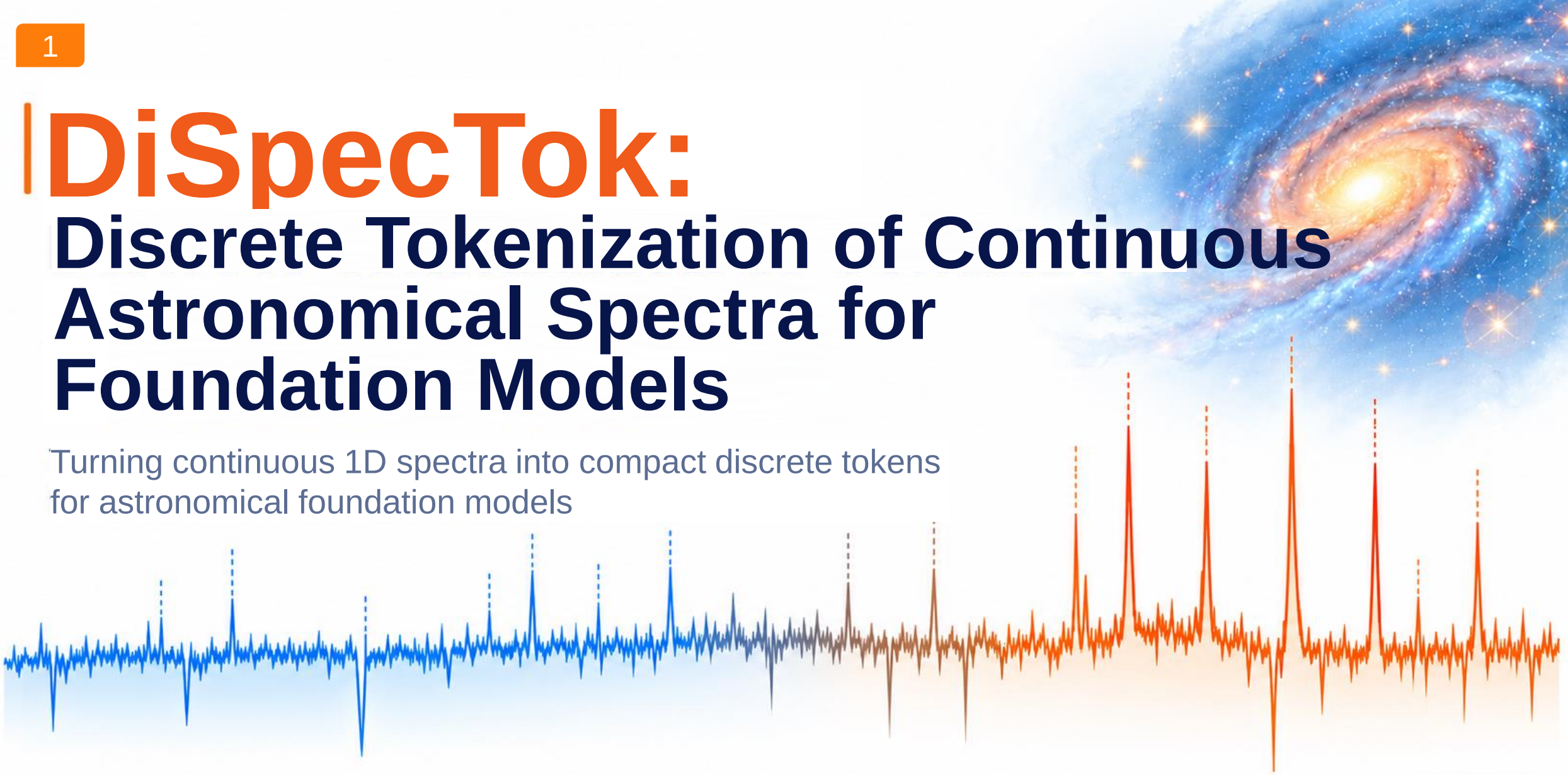
DiSpecTok:

Discrete Tokenization of Continuous Astronomical Spectra for Foundation Models

Turning continuous 1D spectra into compact discrete tokens for astronomical foundation models



Zhongyue Chen | Zhejiang Lab, Hangzhou, China |

The 3rd International Workshop on AI + Astronomy 2026

Talk Roadmap

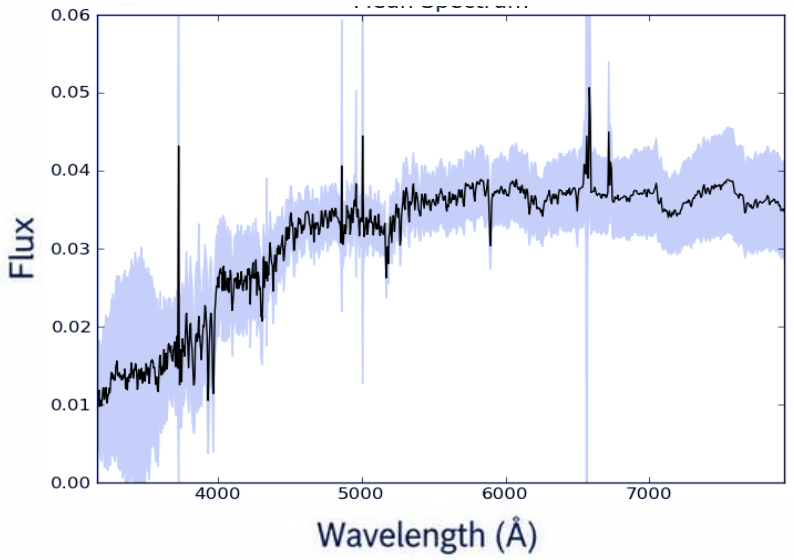
1. Introduction & Background
2. Scientific Problem and Challenges
3. Methodology: DiSpecTok
4. Pretraining Data and Setup
5. Reconstruction Results
6. Downstream Tasks: Redshift, Classification, Stellar & Galaxy Parameter Estimation
7. Key Takeaways & Conclusion



3 Research Background: Why Spectral Tokenization?

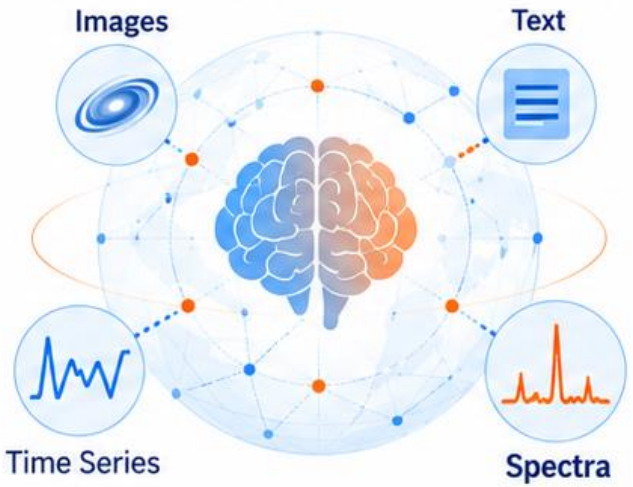
Astronomical Spectroscopy

- Encodes redshift, composition, stellar populations
- Large surveys: SDSS, DESI
- Continuous 1D flux over wavelength



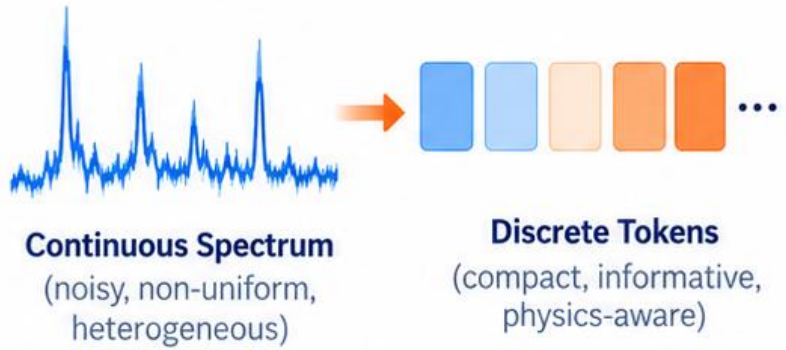
Foundation Models in Astronomy

- Unified representation across modalities
- Token-based Transformers power scalable pretraining
- Images / text / time series increasingly use discrete sequences



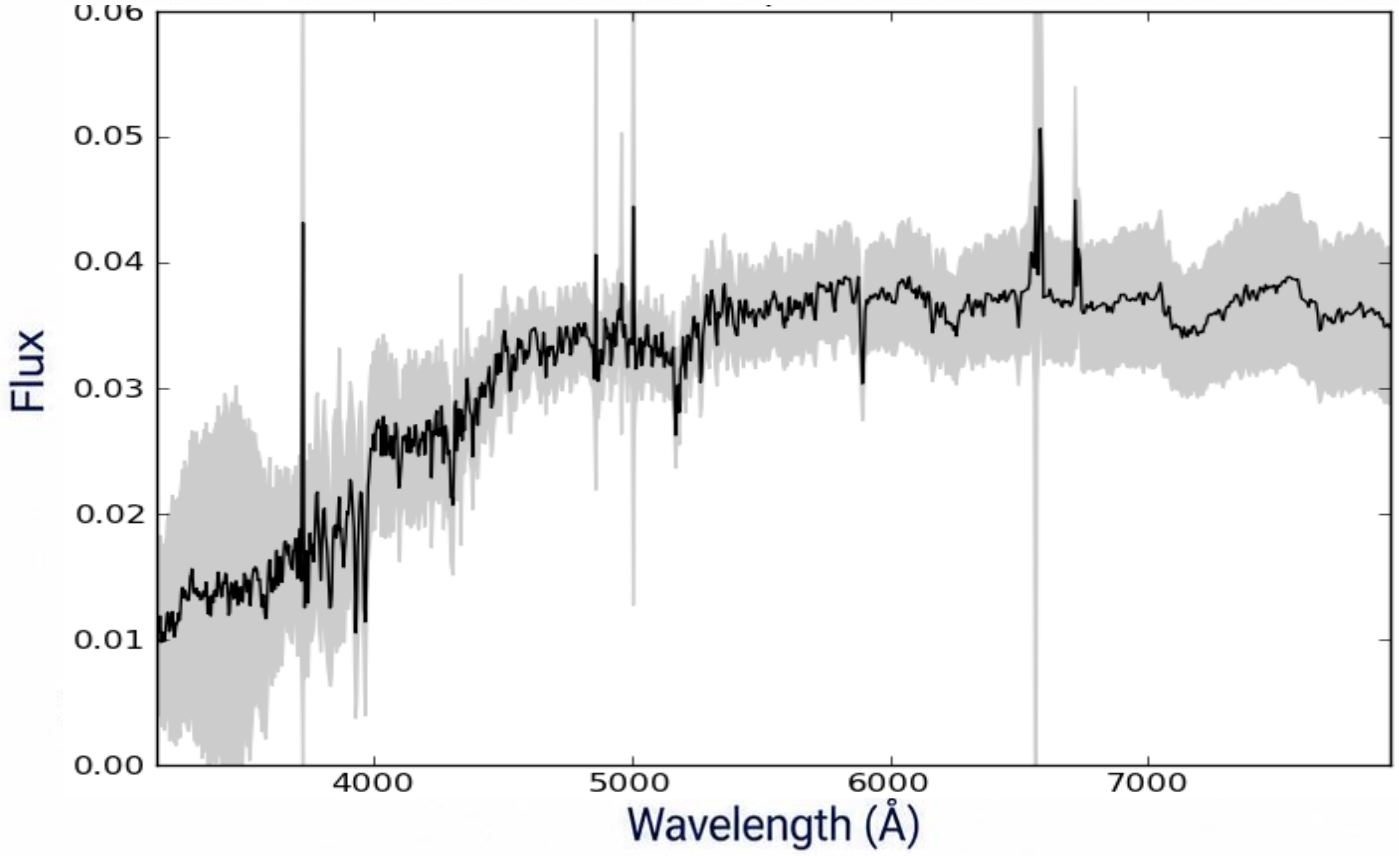
The Missing Interface

- Spectra are continuous, noisy, heterogeneous
- Discrete tokens enable autoregressive and masked-token modeling
- Need compact tokens that preserve physical information



4 Scientific Problem and Challenges

Observed 1D spectrum



Continuous, noisy spectra
instrumental noise, sky residuals, artifacts

Non-uniform wavelength grids
different coverage and resolution across surveys

Uncertainty & masks
per-pixel inverse variance and bitmasks are informative

Mismatch with token models
LLMs / multimodal Transformers expect discrete sequences

Goal: finite-vocabulary spectral tokens that remain useful for reconstruction and scientific inference

5 Key Idea and Contributions

Core Idea

Map each continuous, noisy, variable-length spectrum into a sequence of **compact discrete tokens** while preserving **physically meaningful spectral features.**



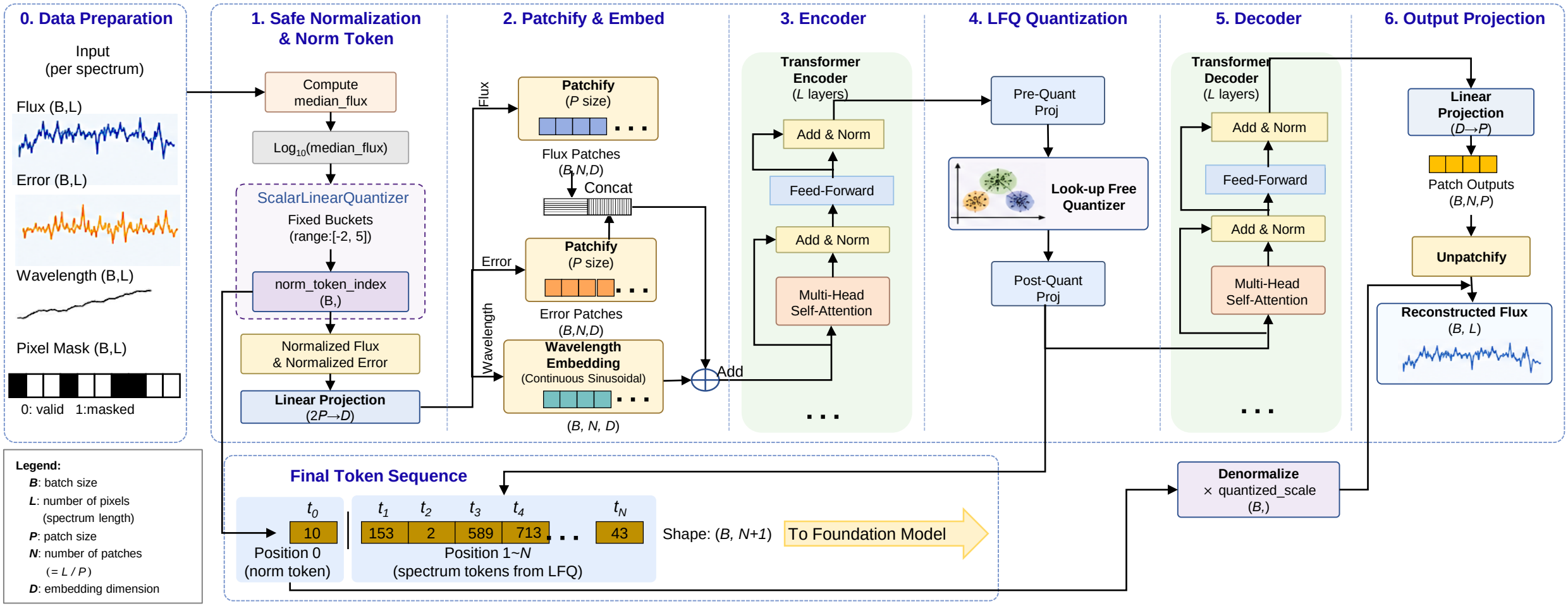
1. **Discrete Spectral Tokenizer (DiSpecTok)**
   finite-vocabulary interface for 1D spectra

2. **Wavelength- and uncertainty-aware architecture**
   native grid, flux, inverse variance, wavelength, mask

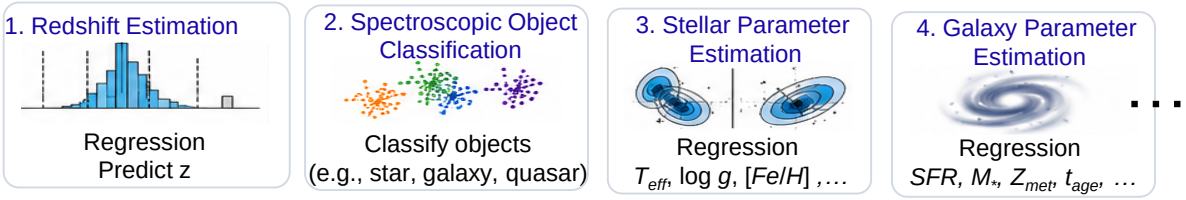
3. **Scale–shape decoupling + LFQ**
   scalar global flux token + lookup-free local shape tokens

4. **Broad downstream validation**
   redshift, object type, stellar and galaxy physical parameter estimation

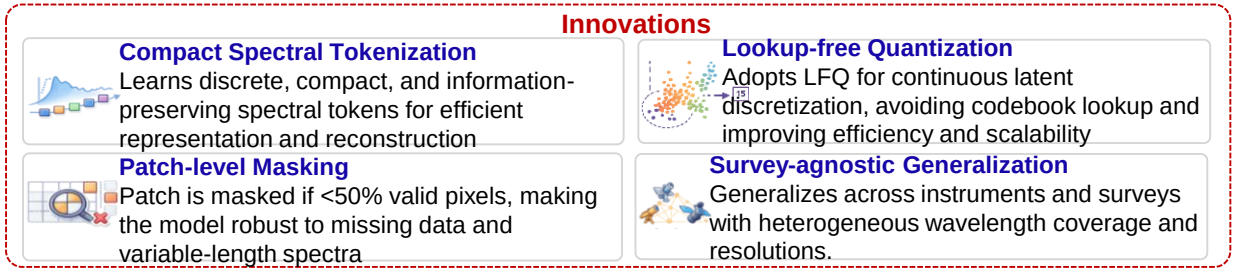
6. DiSpecTok Model Overview



Downstream Tasks

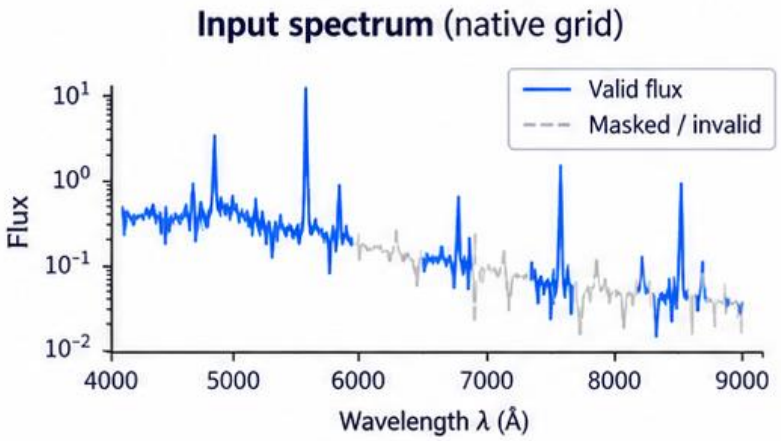


Innovations



7 Model Design I: Normalization, Patching and Embedding

1 Safe normalization + norm token

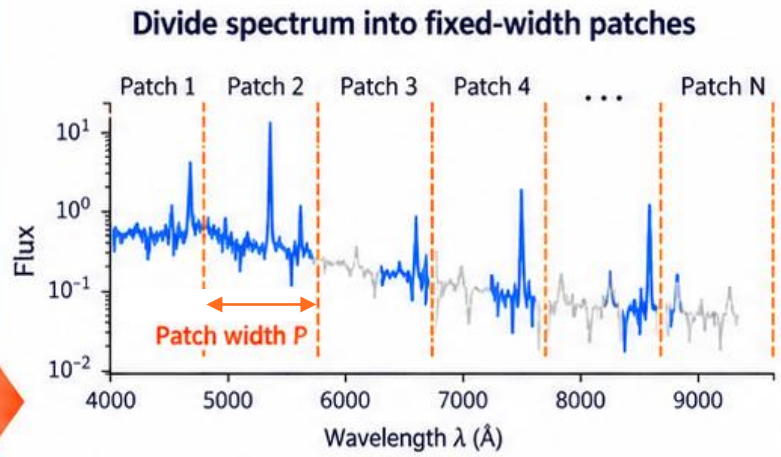


$$\sigma_{\text{norm}} = \log_{10}(\text{median}(\text{valid flux}))$$

$$t_0 = Q_{\text{scalar}}(\sigma_{\text{norm}})$$

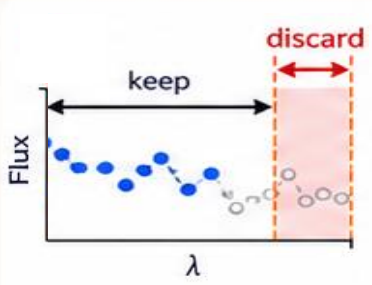
- decouples global flux scale
- prevents scale from dominating shape tokens

2 Bitmask-driven patching

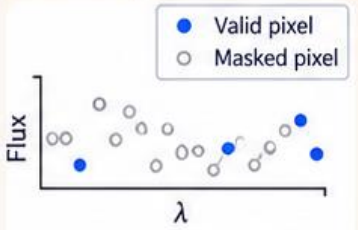


$$N = \text{ceil}(L \,/\, P)$$

Half-width discard criterion

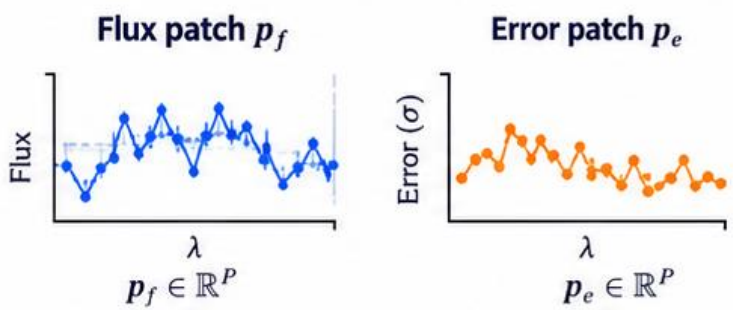


Quarantine low-coverage patches



quarantine patches with < 50% valid pixels

3 Dual-stream patch embedding



$$\boldsymbol{p}_{\text{joint}} = [\mathbf{p}_f \parallel \mathbf{p}_e] \in \mathbb{R}^{2P}$$

Linear projection: $2P \to D$

$$\boldsymbol{h}_j = \mathbf{W}_p\,\mathbf{p}_{\text{joint}} + \mathbf{b} \qquad \boldsymbol{h}_j \in \mathbb{R}^D$$

+ continuous sinusoidal wavelength encoding $\Psi(\bar{\lambda}_j)$



$$z_j = \boldsymbol{h}_j + \Psi(\bar{\lambda}_j) \qquad \in \mathbb{R}^D$$

The encoder receives **local spectral features, uncertainty information and physical wavelength anchors on the native grid.**

8 Model Mechanism II: Transformer Encoder and LFQ

DiSpecTok

1 Mask-aware Transformer encoder

Token embeddings (from patchify)

x_1 x_2 x_3 … x_N

$X \in \mathbb{R}^{N \times d}$
(patch embeddings with wavelength and uncertainty)

Attention mask (from validity mask)

m_1 m_2 m_3 … m_N

1 1 0 … 1

$m_j = 1$ (valid), 0 (invalid)

$M \in \{0,1\}^{N \times N}$
(pairwise attention mask)

Transformer encoder (uncertainty-aware)

× M layers

Add & Norm

Feed-Forward (MLP)

Add & Norm

Multi-Head Self-Attention (with attention bias)

Attention bias for invalid patches

$$A_{bias}(r,c) = \begin{cases} 0, & \text{if } m_r = 1 \text{ and } m_c = 1 \\ -\infty, & \text{otherwise} \end{cases}$$

corrupted patches cannot contaminate attention

Contextualized token representations

h_1 h_2 h_3 … h_N

$H \in \mathbb{R}^{N \times d}$
(contextualized latent vectors)

2 Lookup-Free Quantization (LFQ)

Latent vector h_j

$h_j \in \mathbb{R}^d$
(from Transformer)

Pre-quant projection $\hat{z}_j$

$\hat{z}_j = Wh_j$

$\hat{z}_j \in \mathbb{R}^B$
(linear projection)

Sign binarization b_j

+1 −1 … +1 −1

$b_j = \text{sign}(\hat{z}_j)$

$b_j \in \{-1, +1\}^B$

Binary token ID t_j

1 0 … 1 0

$t_j \in \{0, \cdots, 2^B - 1\}$
(B-bit binary code)

$t_j = \mathcal{B}(b_j)$
(binary-to-integer)

No embedding-table lookup

Directly discretize latent vectors via sign function.

STE enables end-to-end learning

Straight-Through Estimator for gradients through binarization.

Efficient finite vocabulary

Token ID $t_j \in \{0, \ldots, 2^B - 1\}$ with vocabulary size 2^B
(e.g., B=10 → 1024 tokens).

3 Final representation: scale token + local shape tokens = foundation-model-ready spectral sequence.

scale token (global information) t_0

local shape tokens (spectral details) t_1 t_2 t_3 t_4 … t_N

Final token sequence

$$T = [t_0, t_1, t_2, \cdots, t_N] \in \{0, \cdots, 2^B - 1\}^{N+1}$$

Foundation-model-ready spectral sequence

9 Pretraining Data and Setup

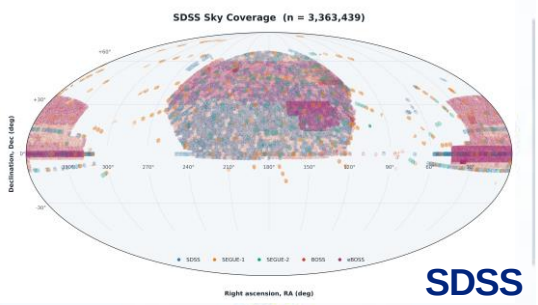
Dataset

SDSS:
3,363,439
spectra

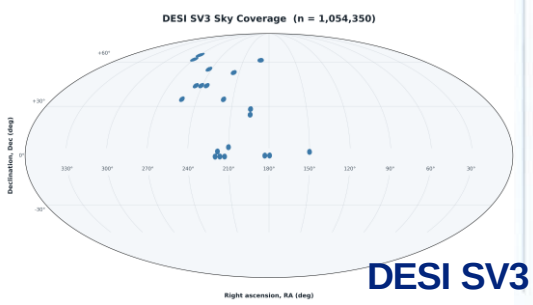
DESI SV3:
1,054,350
spectra

Total corpus:
4,417,789 spectra

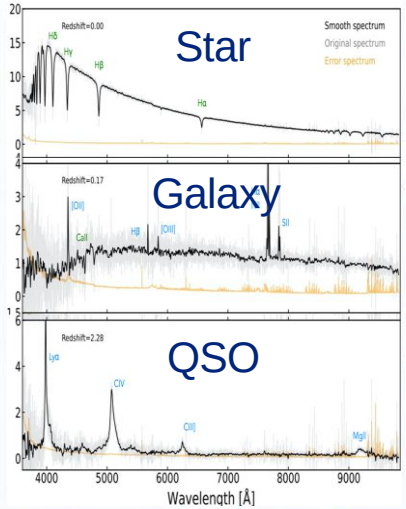
- stars, galaxies, quasars
- heterogeneous wavelength grids
- flux + inverse variance + wavelength + validity mask
- ~90/10 train-test split



SDSS



DESI SV3



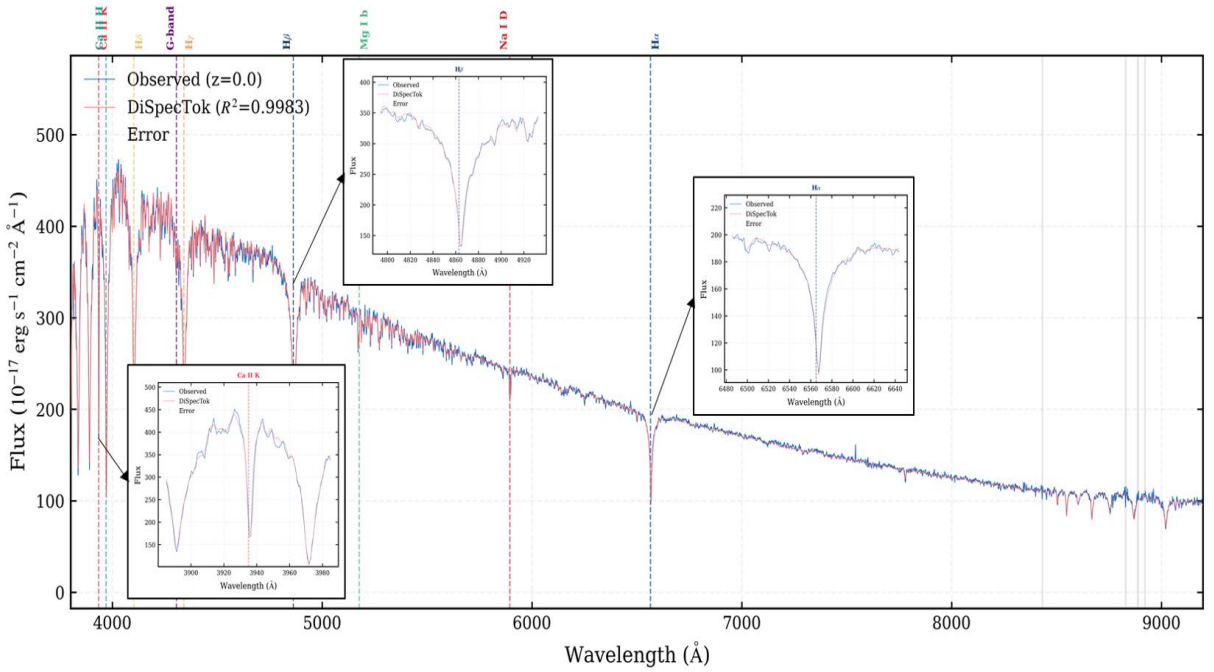
Architecture & Training

Patch / embed dim	24 / 512
Attention heads	8
Encoder / decoder layers	6 / 6
Codebook size / num.	1024 / 1
Entropy / commit weights	0.1 / 0.25
Optimizer	AdamW
Peak LR / weight decay	1e−4 / 0.01
Batch size	512
GPU	A100 × 4, 10 days

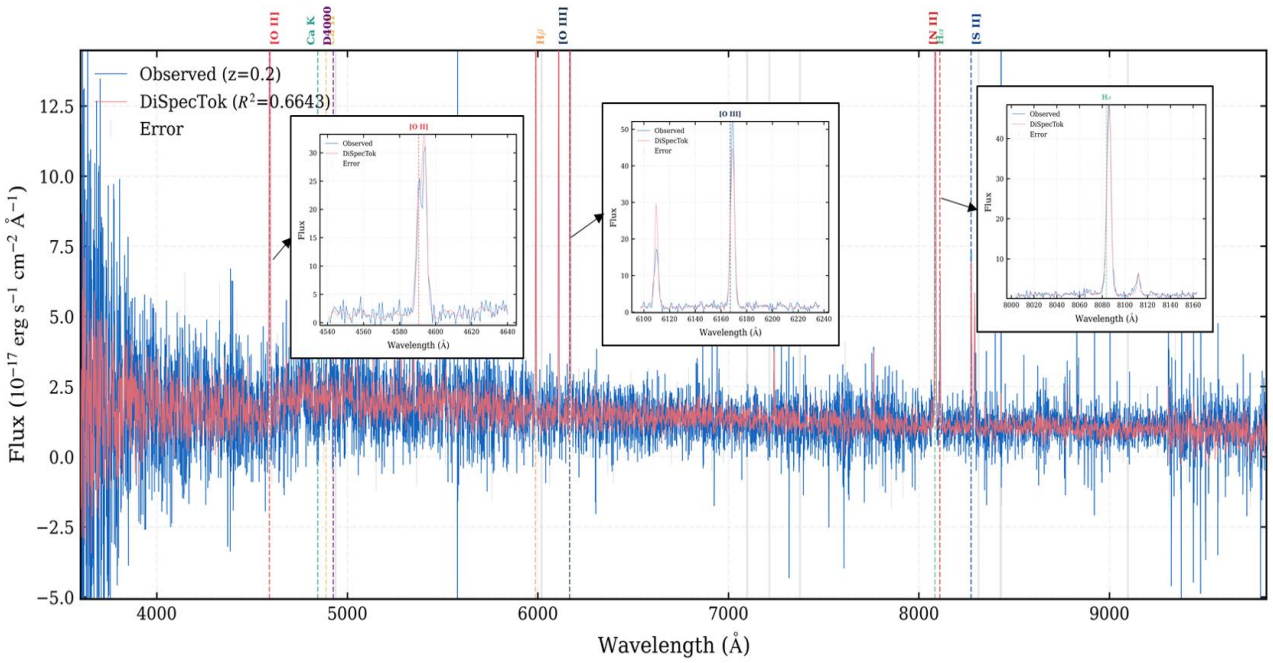
Self-supervised reconstruction objective; continuous variant DiSpecTok-C isolates quantization effects.

Reconstruction Evaluation: Compression vs. Fidelity

SEGUE stellar spectrum



DESI galaxy spectrum



Reconstruction on held-out test split

Method	Overall R^2 ↑	Overall wR^2 ↑
AION-Sp	0.579	0.725
DiSpecTok-C (Continuous ver)	0.9997	0.9999
DiSpecTok	**0.624**	**0.743**

Discrete bottleneck smooths pixel-level noise but preserves continuum and diagnostic line positions.

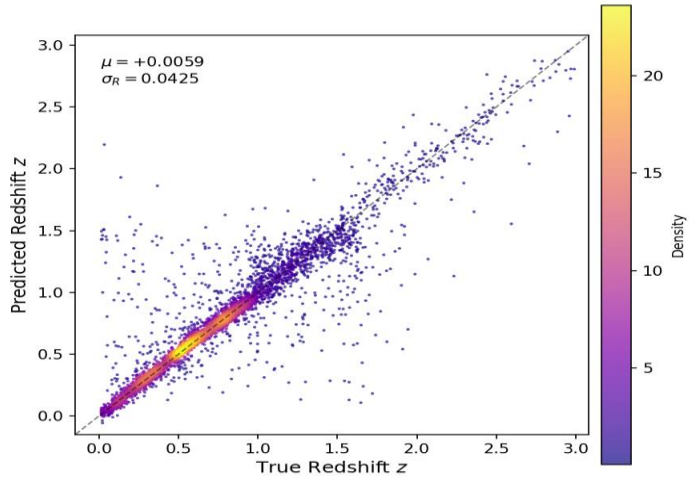
11 Downstream Task 1: Redshift Estimation

R^2
0.8659

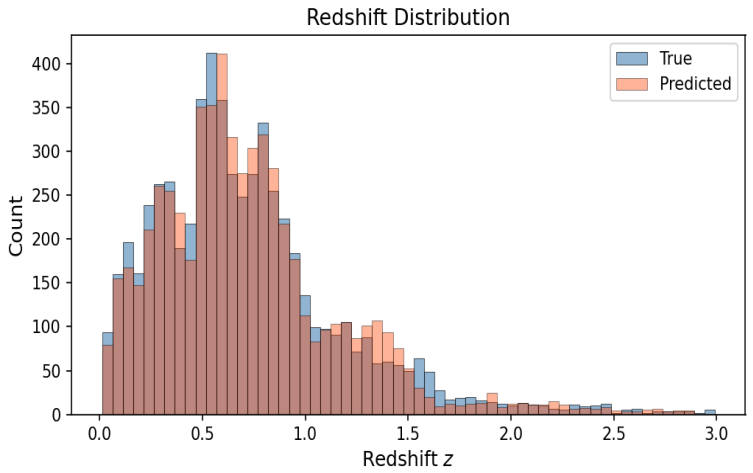
NMAD
0.0226

Outlier
4.66%

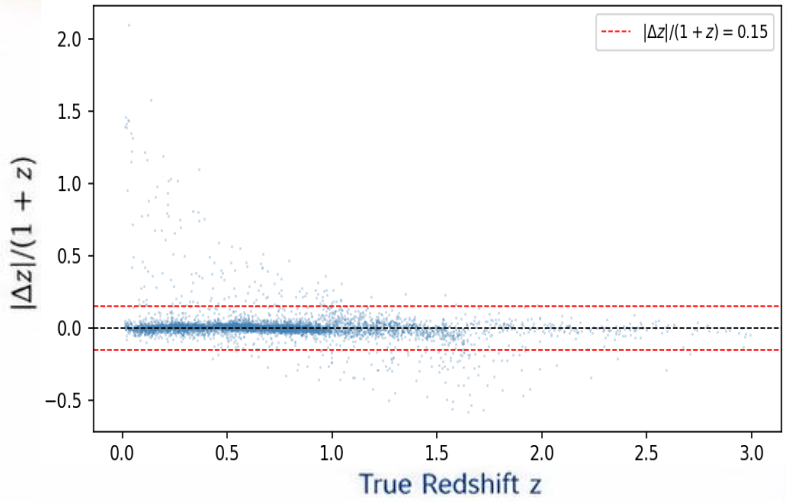
1 Predicted vs. True z



2 Redshift distribution



3 Normalized residuals



4 Test-set comparison

Method	R^2 ↑	MAE ↓	RMSE ↓	σR ↓	NMAD ↓	foutlier ↓
AION-Sp	0.7881	0.1277	0.2132	0.1266	0.0618	11.10%
DiSpecTok-C	0.8330	0.0872	0.1893	0.0529	0.0264	5.99%
DiSpecTok	**0.8659**	**0.0684**	**0.1696**	**0.0425**	**0.0226**	**4.66%**

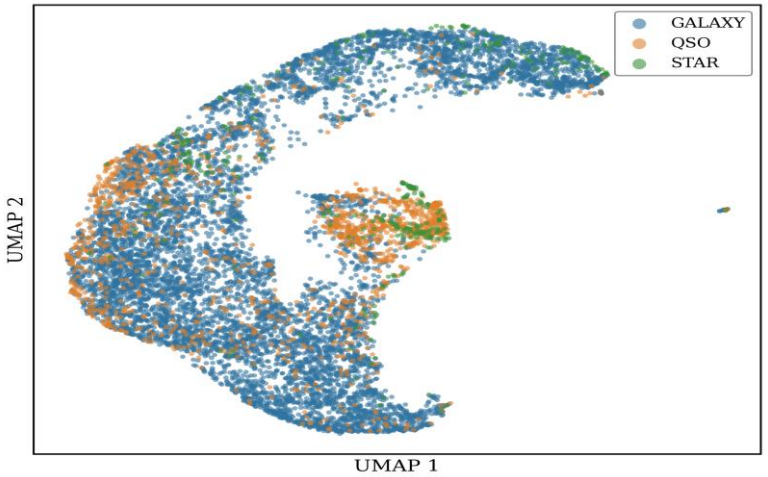
Downstream Task 2: Spectroscopic Object Classification

Accuracy = 0.9590

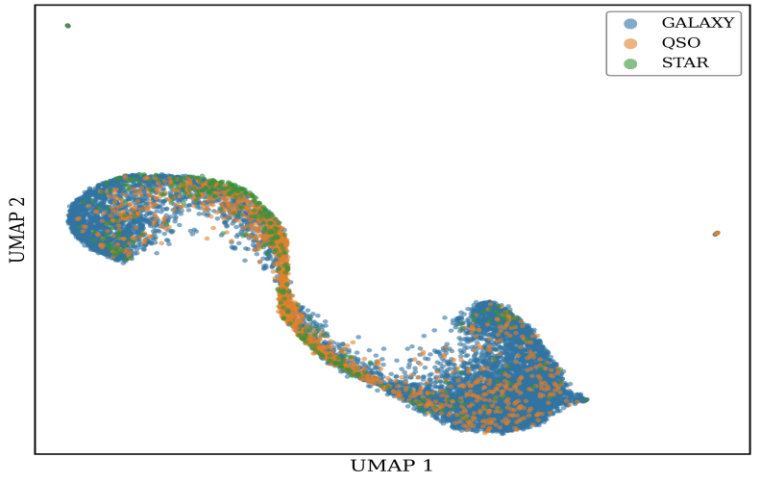
Balanced Acc. = 0.9145

Macro F1 = 0.9259

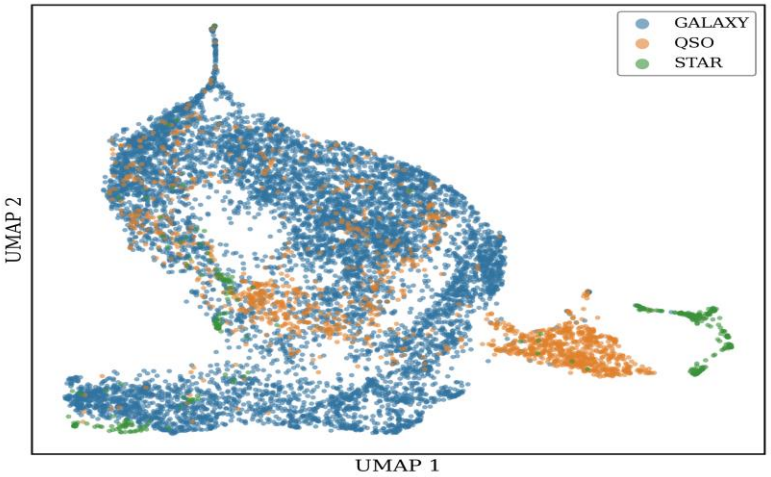
AION-Sp



DiSpecTok-C



DiSpecTok



DESI SV1 classification set

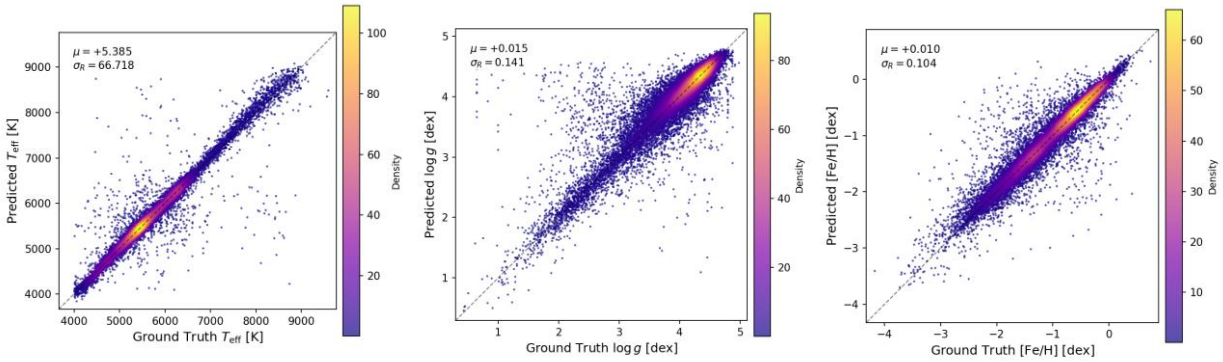
Method	Acc. ↑	Bal. Acc. ↑	Macro F1 ↑	Macro AUC ↑	MCC ↑
AION-Sp	0.9470	0.8744	0.8787	0.9680	0.8300
DiSpecTok-C	0.9464	0.8833	0.8995	0.9644	0.8261
DiSpecTok	**0.9590**	**0.9145**	**0.9259**	**0.9830**	**0.8680**

Takeaway: Quantization-trained representations improve class separability.

Downstream Tasks 3–4: Physical Parameter Estimation

Stellar parameters (SSPP)

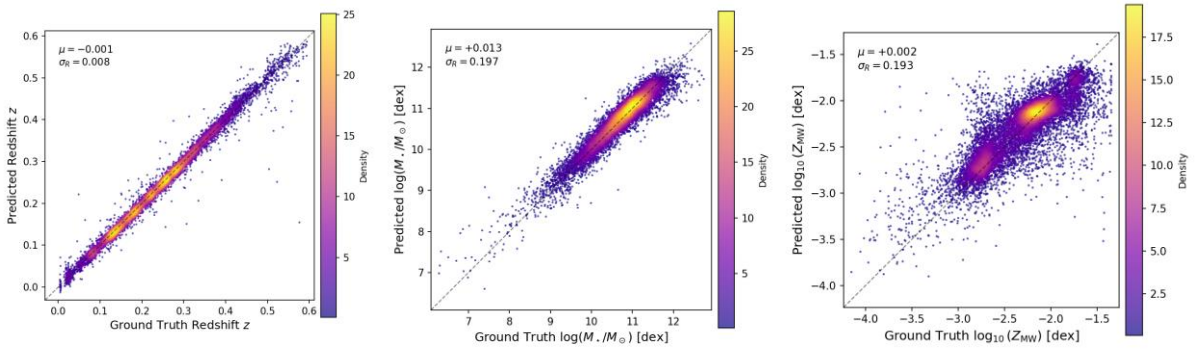
T_{eff} (K)	log g (dex)	[Fe/H] (dex)
R^2 = 0.944 MAE = 74.90 K	R^2 = 0.815 MAE = 0.142	R^2 = 0.909 MAE = 0.106



Parameter	R^2 ↑	MAE ↓	σ_R ↓	f_{outlier} ↓
T_{eff} (K)	**0.944**	**74.90 K**	**66.72**	**2.88%**
log g (dex)	**0.815**	**0.142**	**0.141**	**4.35%**
[Fe/H] (dex)	**0.909**	**0.106**	**0.104**	**6.54%**

Galaxy parameters (PROVABGS)

z	log M_*	log Z_{met}
R^2 = 0.986 MAE = 0.007	R^2 = 0.903 MAE = 0.161	R^2 = 0.652 MAE = 0.164



Parameter	R^2 ↑	MAE ↓	σ_R ↓	f_{outlier} ↓
z	0.986	0.007	0.008	0.09%
log M_*	0.903	0.161	0.197	13.18%
log Z_{met}	0.652	0.164	0.193	14.75%
t_{age} (Gyr)	0.521	0.720	0.714	59.10%
log sSFR	0.756	0.300	0.286	28.18%

Quantization-trained representations retain important stellar and galaxy physical information.

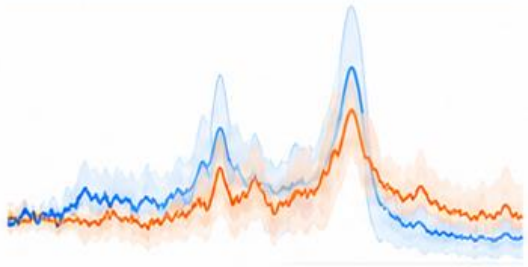
14 Innovations, Take-home Messages and Future Work

Key Innovations



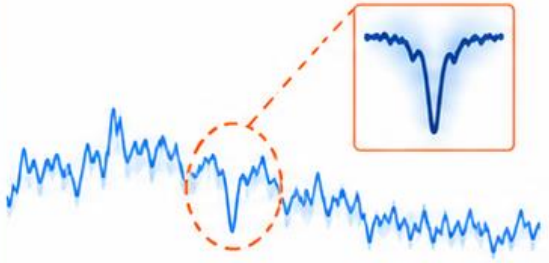
- Native-grid spectral tokenization
- Scale–shape decoupling
- Uncertainty-aware Transformer masking
- Lookup-Free Quantization for finite vocabulary

Main Findings



- 4.4M-spectrum pretraining corpus
- Competitive performance across multiple downstream
- Discrete bottleneck filters noise while retaining key spectral features.
- Tokens form a foundation-model-ready spectral interface

Limitations



- Fine-grained or rare spectral signatures may be smoothed
- Validated mainly on optical spectra
- Integrated LLM-scale reasoning remains future work

Future Work / Impact



- Extend to UV/IR and more surveys
- Enhance cross-modal understanding
- Evaluate generative and cross-modal reasoning
- Toward unified astronomical foundation models

DiSpecTok converts continuous astronomical spectra into compact, reusable tokens for scalable multimodal science.