Multi-Modal Masked Autoencoders for Learning Image-Spectrum Associations for Galaxy Evolution and Cosmology

Morgan Himes[1], Samiksha Krishnamurthy[2], Andrew Lizarraga[3], Srinath Saikrishnan[4], Vikram Seenivasan[1], Jonathan Soriano[1], Ying Nian Wu[3], Tuan Do[1]

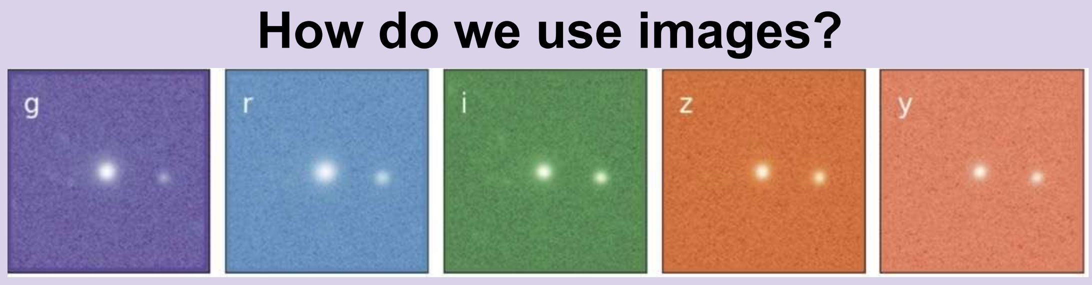
[1]Department of Physics and Astronomy, UCLA; [2]Department of Electrical and Computer Engineering, UCLA; [3]Department of Statistics and Data Science, UCLA; [4]Department of Computer Science, UCLA

NEURAL INFORMATION PROCESSING SYSTEMS

UCLA ASTROPHYSICS DATA LAB

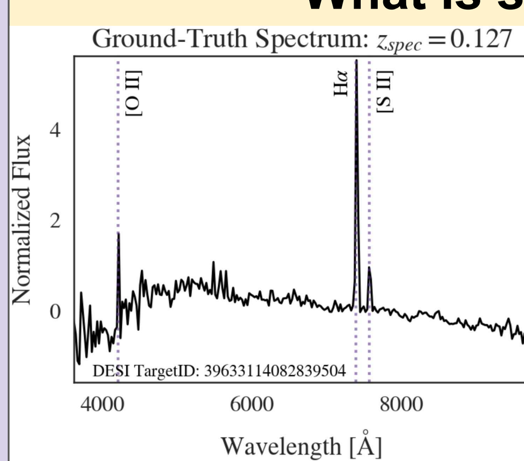
1. How Do Astronomical Observations Tell Us About Galaxies?

How do we use images?



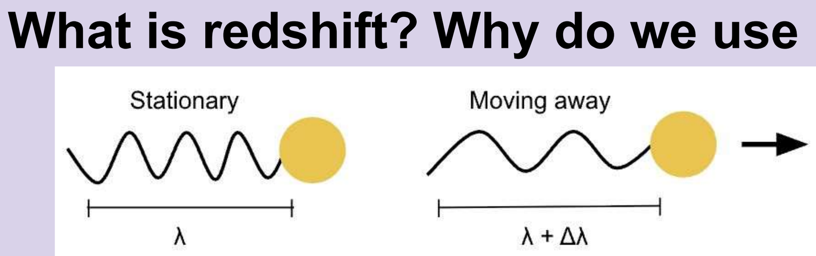
- Visualize morphology (shape) and color
- Galaxy brightness (photometry) measured from images at different wavelength bands and compared to galaxy templates. Redshift (z) of best-fit template is assigned
- Easier to collect for large surveys but less accurate in z

What is spectroscopy?



- Light from source is distributed across a spectrum
- Get redshift (z) by comparing known emission lines to their rest frame wavelengths
- Most accurate redshifts and encoded with physical information but time-consuming to obtain

What is redshift? Why do we use it?



- Redshift (z) is stretching of light to longer wavelengths as a source moves away from an observer
- When space expands, light from distant galaxies is stretched, allowing us to map out the universe
- Used to find speed, age, and distance of galaxies

2. A New Multi-Modal Dataset for ML: GalaxiesML-Spectra

- Assembled by cross-matching GalaxiesML and DESI DR1
- Hyper Suprime-Cam (HSC-PDR2) 5-band (g,r,i,z,y) images, 64x64 or 127x127
- DESI DR1 spectra and spectroscopic redshifts

Key Takeaway:
- New, publicly available, multimodal, ML-ready galaxy dataset

Dataset	Data Type	Source Count	z (90th pct.)	z (max)	i-mag (90th pct.)
DESI DR1	Spec, z	20,283,824	1.343	6.857	–
GalaxiesML	Img	286,401	1.155	4.000	22.171
GalaxiesML-Spectra	Spec, Img, z	134,533	1.581	4.119	20.635

Note: Disagreement between HSC and DESI spectroscopic redshifts causes a discrepancy in z (max).

3. Learning Shared Representations from Images and Spectra

Contribution: Multi-modal masked autoencoder for joint multi-modal reconstruction on images and spectra in astronomy, optimized for redshift prediction.

Architecture: Patch-based tokenization strategy, transformer encoders, cross-attention fusion, attention pooling, and three task-specific heads: see **Fig 1** and **Sec 4** for details.

Training: Randomly zero 75% of patch tokens (image and spectrum) per sample and reconstruct masked tokens. 50% of spectra are entirely zeroed during training. Uses AdamW optimization, gradient clipping, and a combined loss: weighted MSE reconstruction for masked image and spectral tokens and a custom redshift loss with tunable weights. Trained jointly for reconstruction and the auxiliary redshift task.

4. Multi-Modal Masked Autoencoder Architecture & Example Application

Images and spectra split into patches and tokenized:
- Images (64×64×5) → 8×8 patches → 256-dim embeddings
- Spectra (7783) → Downsampling (259) → 1D patches (8) → 256-dim embeddings

Masked and unmasked patches are fed into per-modality transformer encoders → cross-attention fusion layers (four cross-attention blocks where images query spectra and spectra query images) → attention pooling → global embeddings concatenated into a joint latent space → three heads operating on the latent: image decoder, spectrum decoder (both MLP decoders), and a redshift regression head.

Key Takeaway:
- Transformer-based model that reconstructs masked galaxy images/spectra and predicts redshift

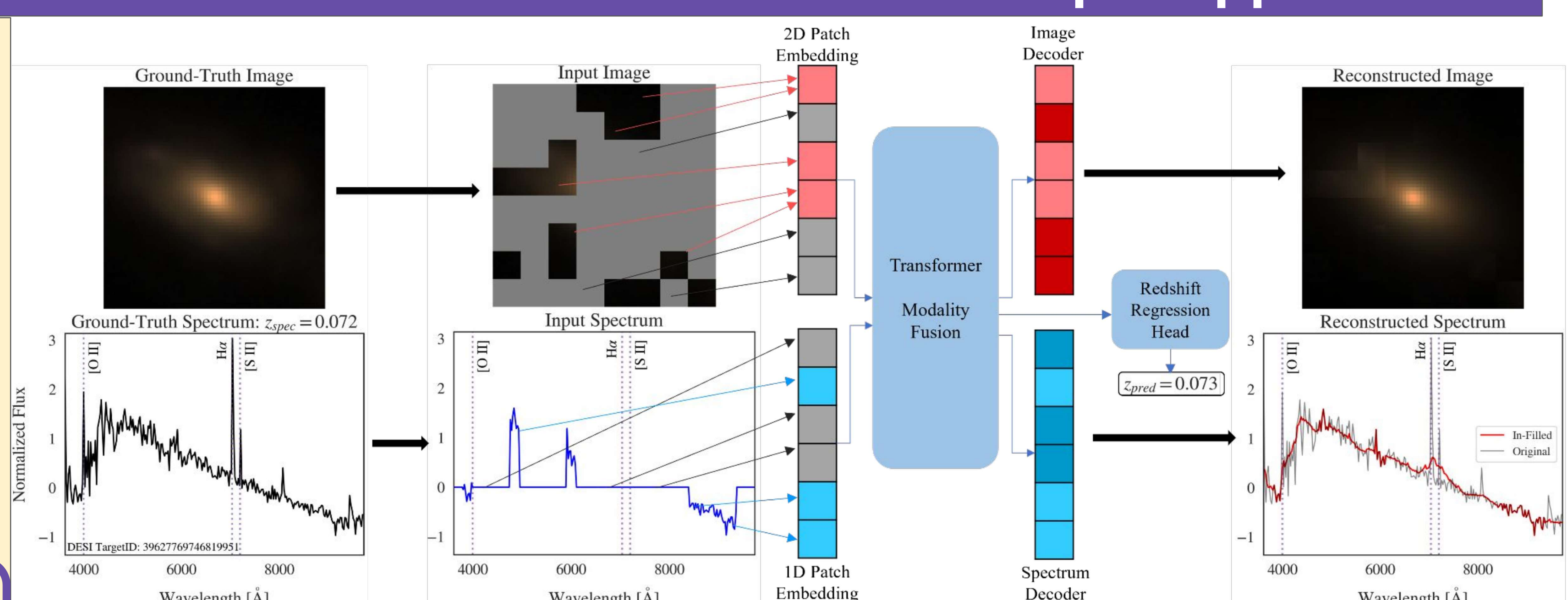


Fig 1: Architecture and reconstruction for a low redshift source with 75% masking (both modalities). Model reconstructs color, morphology, spectral continuum, and a common spectral line (H-α), and accurately predicts redshift. Limitations include spectral line width/height, less abundant spectral lines, and smooth integration of unmasked image patches with reconstructed patches.

5. Reconstructing Masked Data

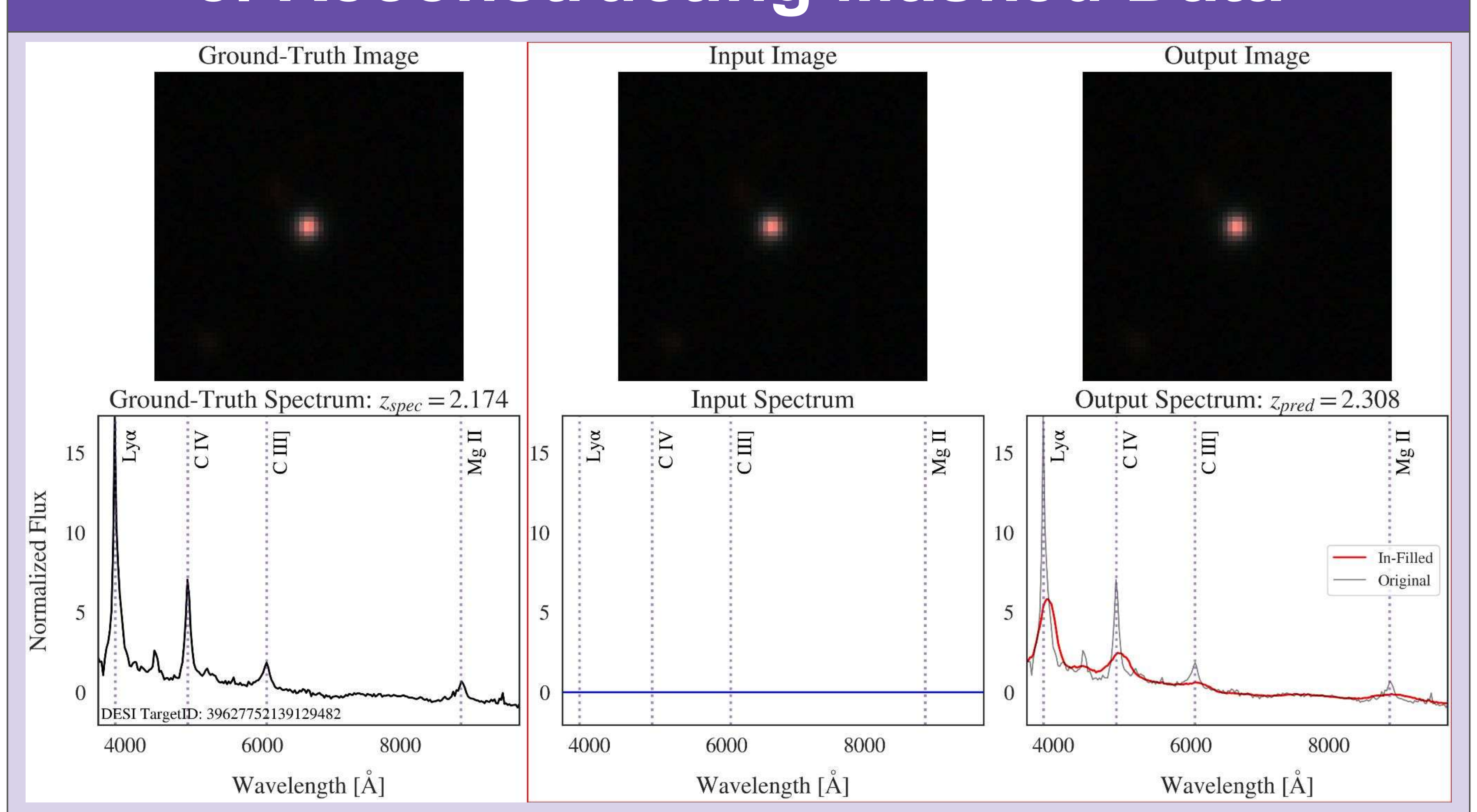


Fig 2: Reconstruction of a high redshift source with 100% spectrum masking and 0% image masking, demonstrating that the model has learned the Lyman-α and C IV emission lines, though underestimates their height and overestimates width.

Key Takeaways:
- *Images*: Reproduces shape/color but struggles with fine details and noise
- *Spectra*: Captures broad continuum, even in cases where entire spectrum is masked, but fails to reproduce random noise
- *Spectra*: Reproduces the locations of some common emission lines (H-α, Lyman-α, C IV), but line widths are systematically overestimated, heights underestimated

6. Predicting Galaxy Redshifts

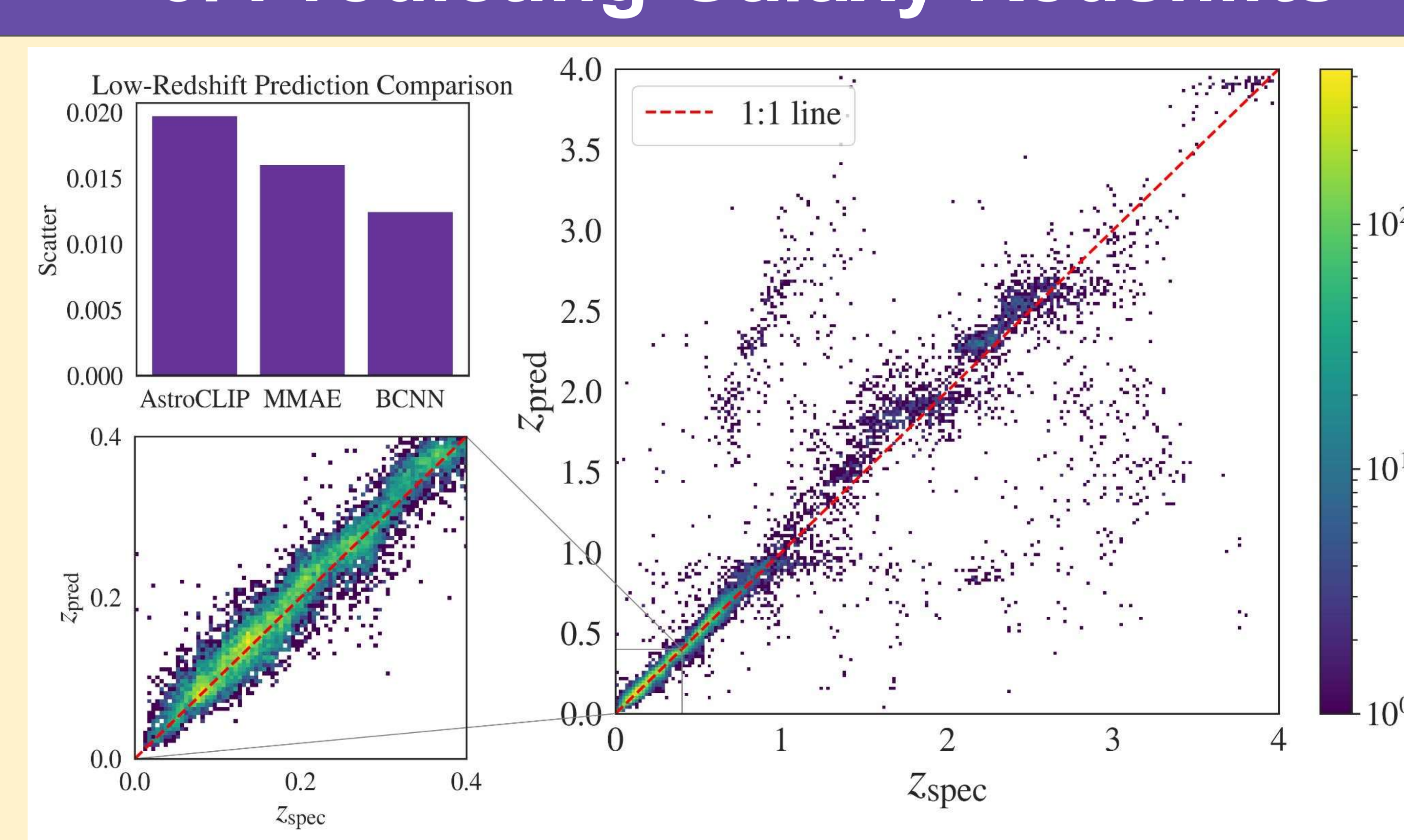


Fig 3: *Right*: Redshift regression results for the case of 25% image masking and 100% spectrum masking. *Bottom Left*: Low-redshift regime used for comparison to AstroCLIP. *Top Left*: MMAE scatter compared to AstroCLIP and a BCNN model for low-redshift.

Key Takeaways:
- Masking 25% of the image at test time produces better photo-z results than supplying the full image (masking acts as regularization to prevent overfitting)
- Performance degrades at high z where the model has less data or features may shift out of band
- MMAE achieves better or comparable results to other models in terms of scatter, but transformer-based architectures still underperform relative to inception-style convolutional models for redshift prediction