The Galaxy's Guide to the Tokenizer: A Benchmark for Scientific Foundation Models

Sogol Sanjaripour[1,*] **Michael J. Smith**[2,3,4] **Manuel Pérez-Carrasco**[2,3]
Juan Rafael Martínez-Galarza[2,3] **Bahram Mobasher**[1] **Gabriela Canalizo**[1]

[1]University of California, Riverside [2]AstroAI
[3]Center for Astrophysics | Harvard & Smithsonian [4]UniverseTBD

*sogol.sanjaripour@email.ucr.edu

Abstract

Tokenization is central to adapting scientific data for transformer-based foundation models, yet its impact on learned representations remains poorly understood. We compare four tokenization strategies—Affine, AIM, JetFormer, and VQ-VAE—within a unified transformer framework for astronomical imaging. Using 640 000 galaxy images from the DESI Legacy Survey and a shared AstroPT backbone, we evaluate each method on reconstruction fidelity and physical property prediction.

Our results reveal trade-offs across approaches. The flow-based JetFormer achieves higher reconstruction quality, while VQ-VAE yields strong probe performance for galaxy physical properties. Affine and AIM better preserve localized morphological information. We find that reconstruction and representation quality are decoupled, and no single method consistently performs best across the tasks considered here. By grounding our evaluation in independently measured physical quantities, we hope this study serves to highlight the potential of scientific data as a basis for constructing interpretable benchmarks for foundation models.

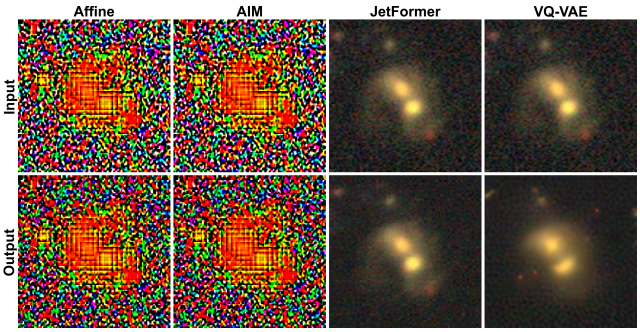

Figure 1: Comparison of image reconstruction quality using different tokenization strategies with a shared transformer backbone (AstroPT). The top row shows the input galaxy images, while the bottom row presents the reconstructions produced by each tokenizer. For the AIM and Affine models, both input and reconstructed images are shown in their patch-wise normalized representation.

1 From pixels to physics

The Big Data Era of astronomy has arrived. Surveys such as Euclid [1], LSST [9], and DESI [4] are producing imaging data at unprecedented scale, and transformer-based foundation models have emerged as a leading paradigm for extracting scientific insight [e.g. 18, 11, 10]. Yet, while architectural choices have been widely studied, far less attention has been paid to *tokenization*—the process by which observations are converted into sequences that transformers can process—and how it shapes learned representations. Scientific domains provide a natural setting for this question, as observations are governed by known physical processes and accompanied by independently measured quantities that serve as objective ground truth. In astronomy, these include continuous properties

Conference on Physics and AI at Stanford University (PAI 2026).

such as redshift, stellar mass, and star formation rate, enabling physically grounded evaluation of representations beyond task-specific metrics [17, 16, 8]. These questions also have practical relevance. Astronomical foundation models increasingly aim to unify multiple modalities within a shared framework [see e.g. 20, 7, 13], and tokenization strategies that perform well for one modality may not generalize to others. Moreover, representations optimized for reconstruction may discard information relevant for scientific inference, a concern also observed in machine learning [e.g., 2, 12].

We systematically compare four tokenization strategies within a controlled framework: Affine (linear) projection; MLP-based [AIM-like; 5]; flow-based [Jetformer-like; 21]; and VQ-VAE-based approaches [6, 25]. Our goal is to characterize how tokenization choices influence learned representations and their relation to scientific inference.

2 AstroPT backbone, tokenization strategies, and experimental setup

We adopt AstroPT [19] as our shared backbone: AstroPT is a decoder-only transformer [23, 14] trained autoregressively on sequences of astronomical data. We choose an autoregressive, decoder-only architecture because it is the dominant paradigm in current foundation models and imposes minimal inductive bias on the representation, making it a neutral testbed for comparing tokenization strategies. Our four tokenization strategies span a range of design choices: linear versus non-linear mappings, continuous versus discrete representations, and deterministic versus probabilistic decoding. We go into further detail about each strategy below.

The Affine tokenizer defines the simplest possible mapping; a linear projection from image patches to the AstroPT embedding space, with a second linear projection for reconstruction.

The AIM tokenizer replaces the linear projection with an MLP, following [5] and the original AstroPT framework. Following [5] and [19], we apply z-score normalization to each Affine and AIM tokenizer patch to stabilize training.

The JetFormer tokenizer, motivated by [21], produces continuous tokens through an invertible normalizing flow. Inputs are uniformly dequantized, then a two-dimensional flow produces a latent image preserving all input information.

The VQ-VAE tokenizer discretizes images into a learned codebook of visual tokens [22]. An encoder network maps image patches to a continuous latent space, where each latent vector is quantized to its nearest neighbor in a learned codebook. Following the literature, our VQ-VAE tokenizer is pre-trained on in-distribution galaxy imagery and then frozen during AstroPT training.

We must note here that our tokenization comparison is not entirely ablative, and that this is a genuine methodological tension. However, we believe it is endemic to any realistic comparison of tokenizers: each strategy requires its own training procedure to be evaluated competitively, and forcing an entirely uniform regime would produce unrealistic and unfavorable setting for some methods. We therefore instead follow the training methodologies described in the literature as faithfully as possible.

Data selection and pre-processing. We use a publicly available galaxy imaging dataset[1] derived from the DESI Legacy Surveys (DESI-LS) Data Release 8 [3, 24], comprising 8.6 million galaxies, from which we select a representative subset of $640\,000$ galaxies for training. For each galaxy we extract 256×256 pixel g, r, z–band postage stamps at a resolution of 0.262 arcsec pixel^{-1}.

Hyperparameters and pre-training routines. All experiments employ an 89M-parameter AstroPT backbone[2] with 12 transformer layers, model dimension 768, and 12 attention heads, with a context length of 1024 tokens, causal self-attention, zero dropout, and bias-free linear layers. Our pre-training code for all tokenizers is publicly available.[3]

Physical knowledge and reconstruction performance tests. We train linear and MLP probes on embeddings extracted from the center layer of each AstroPT model to predict physical properties of $167\,000$ held out galaxies. We select a range of properties covering photometry, morphology, and spectroscopy. Each property is evaluated with a k=10-fold cross validation. Reconstruction quality is assessed via SSIM and PSNR on $5\,000$ randomly sampled test galaxies.

[1] https://huggingface.co/datasets/Smith42/galaxies
[2] https://github.com/Smith42/astroPT
[3] https://github.com/Smith42/astroPT/tree/sogol_branch

3 Does tokenization affect latent physical knowledge?

We evaluate whether tokenization affects (a) how much physical information is retained in the learned representations, (b) how that information is organized (linearly vs. nonlinearly), and (c) whether reconstruction quality predicts either of these. Table 1 summarizes our main results across linear and MLP probes for physical property prediction, as well as SSIM and PSNR reconstruction metrics. The probed properties include DESI photometric magnitudes (M_g, M_r) and colors ($g-r$, $r-z$), photometric and spectroscopic redshifts (photo-z, spec-z), specific star formation rate (sSFR), stellar mass M_*, and morphological indicators (smoothness, disk fraction, artifact fraction, edge-on fraction, and spiral winding tightness).

Regression on galaxy physical parameters. VQ-VAE achieves the strongest linear and MLP probe performance for most physical properties. Notably, VQ-VAE's linear probe R^2 surpasses the MLP probe R^2 for most properties, which suggests that the discrete bottleneck organises physical information in a more linearly accessible manner compared to the other tokenizers.

JetFormer consistently outperforms AIM and Affine on linear probes for color and spectrally related quantities (magnitudes, colors, redshifts, star formation rate, stellar mass), while AIM shows stronger performance on morphology-driven parameters (smoothness, disk fraction, edge-on fraction). Interestingly, the linear and MLP probe performances for JetFormer are within the error spread from each other, suggesting that the JetFormer's normalizing flow produces embeddings that are somewhat linearly accessible, with little-to-no additional information accessible via a non-linear method. Affine and AIM perform near-identically across nearly all properties. This indicates that the MLP in the AIM tokenizer head adds negligible benefit over a simple linear projection when the transformer backbone is sufficiently expressive.

Across all tokenizers and probe types, $g-r$ color is predicted more accurately than $r-z$. This is initially surprising given that $g-r$ is more susceptible to dust obscuration, and the associated age-dust degeneracy (an old galaxy with no dust can have the same $g-r$ color as a young, star-forming galaxy that is heavily obscured). However, $g-r$ spans a larger dynamical range in our redshift regime and varies more strongly over the training set [15], making it an easier regression target. Physically, $g-r$ traces specific star formation rate (sSFR)via emission from young stars, whereas $r-z$ is a better predictor of older stellar populations that form the bulk of the stellar population, making it a better predictor of the total stellar mass, M_*. The greater dust sensitivity of $g-r$ explains why sSFR, is harder to predict than M_*, and why the MLP probe shows a larger improvement over the linear probe for sSFR; non-linear degeneracies between dust, age, and color are better captured by the more expressive model.

The observed differences in performance likely reflect architectural design. We hypothesize that JetFormer's multiple non-linear transformations emphasize global and spectrally informative features, but may smooth fine-scale morphological information. Patch-based methods (AIM, Affine), on the other hand, preserve localized spatial information, aiding encoding of morphological features. Finally, VQ-VAE further enforces a discrete information bottleneck through its finite codebook, compressing image patches into a small set of representative tokens. This hard tokenization discards nuisance variation while encouraging semantic clustering aligned with underlying physical properties, resulting in embeddings that are particularly well suited for regression tasks.

Table 1: Model evaluation results. Linear and MLP probe R^2 for physical properties. SSIM and R^2 values are multiplied by a factor of 100.

	Linear probe				MLP probe			
	Affine	AIM	Jet	VQ-VAE	Affine	AIM	Jet	VQ-VAE
M_g	72	72	74	76	77	77	74	80
M_z	68	68	70	71	73	73	71	75
$g-r$	71	71	74	82	78	79	75	86
$r-z$	64	64	66	74	70	71	68	79
photo-z	58	59	62	68	64	65	61	72
spec-z	77	77	80	85	81	82	79	87
sSFR	37	37	47	57	50	50	51	65
M_*	58	58	59	67	63	64	60	70
smooth	53	55	49	55	58	60	49	61
disc	46	48	42	49	53	55	42	56
artifact	64	64	66	64	69	70	66	74
edge-on	54	57	45	62	57	60	42	65
tight spiral	45	45	37	49	44	45	39	53

Reconstruction

	Jet	VQ-VAE
SSIM	77 ± 14	54 ± 6
PSNR	31 ± 4	23 ± 1

To investigate how each tokenizer allocates its representational capacity, we perform a second run of $k = 10$-fold linear probes to predict our embedding spaces given groups of physical properties. We find that JetFormer dedicates larger

shares of its representational capacity to directly observable image properties, with the 'apparent photometry' and 'structural parameter' group probes reaching $R^2 = 0.33 \pm 0.02$ and $R^2 = 0.29 \pm 0.02$, compared to VQ-VAE's $R^2 = 0.19 \pm 0.01$ and $R^2 = 0.16 \pm 0.01$. Conversely, VQ-VAE dedicates more capacity 'higher level' physical properties that require some abstraction beyond pixelwise analysis such as redshift ($R^2 = 0.058 \pm 0.003$ vs $R^2 = 0.033 \pm 0.006$), and absolute photometry ($R^2 = 0.090 \pm 0.004$ vs $R^2 = 0.05 \pm 0.01$). This suggests that the VQ-VAE tokenizer enables higher level reasoning which allows our AstroPT backbone to dedicate more space to complex physical properties.

Image reconstruction. JetFormer achieves the highest reconstruction fidelity, preserving fine spatial details and low surface brightness components more effectively than other approaches (Figure 1). On 5 000 held-out images JetFormer achieves a mean PSNR of $31.11\,\mathrm{dB}$ ($\sigma = 3.54$) and mean SSIM of 0.762 ($\sigma = 0.138$), confirming that it preserves structural morphologies (spiral arms, galactic cores, and diffuse background emission). For comparison, VQ-VAE attains a substantially lower mean PSNR of $23.57\,\mathrm{dB}$ ($\sigma = 0.93$) and SSIM of 0.544 ($\sigma = 0.054$), quantitatively supporting the reconstruction patterns observed in Figure 1. VQ-VAE produces visually plausible reconstructions but exhibits diffuse or 'cloudy' artifacts surrounding bright cores and spurious localized features (small red structures absent in the input data). It also underperforms on low-surface-brightness backgrounds, indicating limited recovery of faint, extended emission. In contrast, JetFormer generates consistently sharper reconstructions with improved contrast and structural coherence. For AIM and Affine, patch-level z-score normalization produces a block-wise appearance in reconstruction. This normalization enhances training stability and contributes to strong probe performance, but limits recovery of fine-grained intensity variations in the reconstructed images. Reconstructed images from these methods may visually differ from those of other methods despite sharing the same underlying target. Taking our probe and reconstruction results together, these findings highlight a trade-off between information preservation and abstraction: while soft, dense, information-complete representations favor high-fidelity reconstruction, hard and patch-based tokenization strategies promote more directly interpretable and easily accessible representations for physical inference.

4 Trade-offs, limitations, and the case for physics-grounded benchmarks

Our comparison of four tokenization strategies within a shared AstroPT backbone shows that tokenization is not a neutral preprocessing step, but shapes what information is learned and how it is organized within the representation. We find a decoupling between reconstruction fidelity and the physical information accessible in the representation. JetFormer achieves the best reconstruction, while VQ-VAE yields the most informative embeddings for physical property prediction. This suggests that compressive tokenization may retain physically meaningful structure while discarding pixel-level detail, whereas information-preserving approaches may distribute information in ways that make it less accessible to simple probes. Patch-based methods (AIM, Affine) occupy a middle ground, performing well on morphology-driven features and showing that the transformer backbone largely drives representation learning in this regime.

These trade-offs suggest that tokenization should be chosen based on performance on the eventual downstream task rather than reconstruction metrics alone. A model intended for photometric or spectroscopic inference may benefit from VQ-VAE's compressed representations, while one targeting image generation or reconstruction should prefer JetFormer's approach. For compute-constrained settings, the Affine tokenizer achieves performance comparable to AIM at lower complexity.

We note several limitations. Our evaluation relies on linear and MLP probes, which may not fully capture the information present in the embeddings. Differences in tokenizer training procedures mean the comparison is not purely ablative, and VQ-VAE's performance may partly reflect its pre-trained encoder. We also do not explore scaling with dataset size or model capacity, and patch normalization complicates direct reconstruction comparisons.

To summarize, astronomy provides a particularly powerful setting for tokenizer analysis: known physics and independently measured quantities offer objective benchmarks that avoid the circularity of evaluating representations solely on human-annotated tasks or toy datasets. We expect that similar evaluation frameworks would be valuable in other domains where physical ground truth is likewise available; future work will lay the groundwork for this by extending this study's framework towards realizing a comprehensive 'tokenization bench' built on verifiable astronomical ground-truth data.

Acknowledgments

We would like to thank Stella Biderman and EleutherAI for their generous donation of compute which made this project possible.

References

[1] J. Amiaux, R. Scaramella, Y. Mellier, B. Altieri, C. Burigana, Antonio Da Silva, P. Gomez, J. Hoar, R. Laureijs, E. Maiorano, D. Magalhães Oliveira, F. Renk, G. Saavedra Criado, I. Tereno, J. L. Auguères, J. Brinchmann, M. Cropper, L. Duvet, A. Ealet, P. Franzetti, B. Garilli, P. Gondoin, L. Guzzo, H. Hoekstra, R. Holmes, K. Jahnke, T. Kitching, M. Meneghetti, W. Percival, and S. Warren. Euclid mission: building of a reference survey. In Mark C. Clampin, Giovanni G. Fazio, Howard A. MacEwen, and Jr. Oschmann, Jacobus M., editors, *Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*, volume 8442 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 84420Z, 2012.

[2] Black Forest Labs. FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison, 2025.

[3] A. Dey, D. J. Schlegel, D. Lang, R. Blum, K. Burleigh, X. Fan, J. R. Findlay, D. Finkbeiner, D. Herrera, S. Juneau, et al. Overview of the DESI Legacy Imaging Surveys. *Astronomical Journal*, 157(5):168, 2019.

[4] A. Dey, D. J. Schlegel, D. Lang, R. Blum, K. Burleigh, X. Fan, J. R. Findlay, D. Finkbeiner, D. Herrera, S. Juneau, M. Landriau, M. Levi, I. McGreer, A. Meisner, A. D. Myers, J. Moustakas, P. Nugent, A. Patej, E. F. Schlafly, A. R. Walker, F. Valdes, B. A. Weaver, C. Yèche, H. Zou, X. Zhou, B. Abareshi, T. Abbott, B. Abolfathi, C. Aguilera, S. Alam, L. Allen, A. Alvarez, J. Annis, B. Ansarinejad, M. Aubert, J. Beechert, E. F. Bell, S. Y. BenZvi, F. Beutler, R. M. Bielby, A. S. Bolton, C. Briceño, E. J. Buckley-Geer, K. Butler, A. Calamida, R. G. Carlberg, P. Carter, R. Casas, F. J. Castander, Y. Choi, J. Comparat, E. Cukanovaite, T. Delubac, K. DeVries, S. Dey, G. Dhungana, M. Dickinson, Z. Ding, J. B. Donaldson, Y. Duan, C. J. Duckworth, S. Eftekharzadeh, D. J. Eisenstein, T. Etourneau, P. A. Fagrelius, J. Farihi, M. Fitzpatrick, A. Font-Ribera, L. Fulmer, B. T. Gänsicke, E. Gaztanaga, K. George, D. W. Gerdes, S. Gontcho A. Gontcho, C. Gorgoni, G. Green, J. Guy, D. Harmer, M. Hernandez, K. Honscheid, L. Wendy Huang, D. J. James, B. T. Jannuzi, L. Jiang, R. Joyce, A. Karcher, S. Karkar, R. Kehoe, J. Kneib, A. Kueter-Young, T. Lan, T. R. Lauer, L. Le Guillou, A. Le Van Suu, J. Hyeon Lee, M. Lesser, L. Perreault Levasseur, T. S. Li, J. L. Mann, R. Marshall, C. Martínez-Vázquez, P. Martini, H. du Mas des Bourboux, S. McManus, T. Gabriel Meier, B. Ménard, N. Metcalfe, A. Muñoz-Gutiérrez, J. Najita, K. Napier, G. Narayan, J. A. Newman, J. Nie, B. Nord, D. J. Norman, K. A. G. Olsen, A. Paat, N. Palanque-Delabrouille, X. Peng, C. L. Poppett, M. R. Poremba, A. Prakash, D. Rabinowitz, A. Raichoor, M. Rezaie, A. Robertson, N. A. Roe, A. J. Ross, N. P. Ross, G. Rudnick, S. Safonova, A. Saha, F. Javier Sánchez, E. Savary, H. Schweiker, A. Scott, H. Seo, H. Shan, D. R. Silva, Z. Slepian, C. Soto, D. Sprayberry, R. Staten, C. M. Stillman, R. J. Stupak, D. L. Summers, S. Sien Tie, H. Tirado, M. Vargas-Magaña, A. Katherina Vivas, R. H. Wechsler, D. Williams, J. Yang, Q. Yang, T. Yapici, D. Zaritsky, A. Zenteno, K. Zhang, T. Zhang, R. Zhou, and Z. Zhou. Overview of the desi legacy imaging surveys. *The Astronomical Journal*, 157(5):168, 2019.

[5] A. El-Nouby, M. Klein, S. Zhai, M. A. Bautista, A. Toshev, V. Shankar, J. M. Susskind, and A. Joulin. Scalable Pre-training of Large Autoregressive Image Models. *ArXiv e-prints*, 2024.

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[7] Euclid Collaboration, M. Siudek, M. Huertas-Company, M. Smith, G. Martinez-Solaeche, F. Lanusse, S. Ho, E. Angeloudi, P. A. C. Cunha, H. D. Sánchez, M. Dunn, Y. Fu, P. Iglesias-Navarro, J. Junais, J. H. Knapen, B. Laloux, M. Mezcua, W. Roster, G. Stevens, J. Vega-Ferrero, N. Aghanim, B. Altieri, A. Amara, S. Andreon, N. Auricchio, H. Aussel, C. Baccigalupi,

M. Baldi, S. Bardelli, P. Battaglia, A. Biviano, A. Bonchi, E. Branchini, M. Brescia, J. Brinchmann, S. Camera, G. Cañas-Herrera, V. Capobianco, C. Carbone, J. Carretero, S. Casas, F. J. Castander, M. Castellano, G. Castignani, S. Cavuoti, K. C. Chambers, A. Cimatti, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, F. Courbin, H. M. Courtois, M. Cropper, A. Da Silva, H. Degaudenzi, G. De Lucia, A. M. Di Giorgio, J. Dinis, C. Dolding, H. Dole, F. Dubath, C. A. J. Duncan, X. Dupac, S. Dusini, S. Escoffier, M. Farina, R. Farinelli, F. Faustini, S. Ferriol, F. Finelli, S. Fotopoulou, M. Frailis, E. Franceschi, S. Galeotta, K. George, B. Gillis, C. Giocoli, J. Gracia-Carpio, B. R. Granett, A. Grazian, F. Grupp, S. Gwyn, S. V. H. Haugan, W. Holmes, I. M. Hook, F. Hormuth, A. Hornstrup, K. Jahnke, M. Jhabvala, E. Keihänen, S. Kermiche, A. Kiessling, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, Q. L. Boulc'h, A. M. C. L. Brun, D. L. Mignant, S. Ligori, P. B. Lilje, V. Lindholm, I. Lloro, G. Mainetti, D. Maino, E. Maiorano, O. Mansutti, S. Marcin, O. Marggraf, M. Martinelli, N. Martinet, F. Marulli, R. Massey, S. Maurogordato, H. J. McCracken, E. Medinaceli, S. Mei, M. Melchior, Y. Mellier, M. Meneghetti, E. Merlin, G. Meylan, A. Mora, M. Moresco, L. Moscardini, R. Nakajima, C. Neissner, S.-M. Niemi, J. W. Nightingale, C. Padilla, S. Paltani, F. Pasian, K. Pedersen, W. J. Percival, V. Pettorino, S. Pires, G. Polenta, M. Poncet, L. A. Popa, L. Pozzetti, F. Raison, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, R. Saglia, Z. Sakr, A. G. Sánchez, D. Sapone, B. Sartoris, J. A. Schewtschenko, P. Schneider, T. Schrabback, M. Scodeggio, A. Secroun, G. Seidel, M. Seiffert, S. Serrano, P. Simon, C. Sirignano, G. Sirri, L. Stanco, J. Steinwagner, P. Tallada-Crespí, A. N. Taylor, I. Tereno, S. Toft, R. Toledo-Moreo, F. Torradeflot, I. Tutusaus, L. Valenziano, J. Valiviita, T. Vassallo, G. V. Kleijn, A. Veropalumbo, Y. Wang, J. Weller, A. Zacchei, G. Zamorani, F. M. Zerbi, I. A. Zinchenko, E. Zucca, V. Allevato, M. Ballardini, M. Bolzonella, E. Bozzo, C. Burigana, R. Cabanac, A. Cappi, D. Di Ferdinando, J. A. E. Vigo, L. Gabarra, J. Martín-Fleitas, S. Matthew, N. Mauri, R. B. Metcalf, A. Pezzotta, M. Pöntinen, C. Porciani, I. Risso, V. Scottez, M. Sereno, M. Tenti, M. Viel, M. Wiesmann, Y. Akrami, I. T. Andika, S. Anselmi, M. Archidiacono, F. Atrio-Barandela, C. Benoist, K. Benson, D. Bertacca, M. Bethermin, L. Bisigello, A. Blanchard, L. Blot, M. L. Brown, S. Bruton, A. Calabro, B. C. Quevedo, F. Caro, C. S. Carvalho, T. Castro, Y. Charles, F. Cogato, A. R. Cooray, O. Cucciati, S. Davini, F. De Paolis, G. Desprez, A. Díaz-Sánchez, J. J. Diaz, S. Di Domizio, J. M. Diego, P.-A. Duc, A. Enia, Y. Fang, A. G. Ferrari, P. G. Ferreira, A. Finoguenov, A. Fontana, A. Franco, K. Ganga, J. García-Bellido, T. Gasparetto, V. Gautard, E. Gaztanaga, F. Giacomini, F. Gianotti, G. Gozaliasl, M. Guidi, C. M. Gutierrez, A. Hall, W. G. Hartley, S. Hemmati, C. Hernández-Monteagudo, H. Hildebrandt, J. Hjorth, J. J. E. Kajava, Y. Kang, V. Kansal, D. Karagiannis, K. Kiiveri, C. C. Kirkpatrick, S. Kruk, J. L. Graet, L. Legrand, M. Lembo, F. Lepori, G. Leroy, G. F. Lesci, J. Lesgourgues, L. Leuzzi, T. I. Liaudat, A. Loureiro, J. Macias-Perez, G. Maggio, M. Magliocchetti, E. A. Magnier, F. Mannucci, R. Maoli, C. J. A. P. Martins, L. Maurin, M. Miluzio, P. Monaco, C. Moretti, G. Morgante, C. Murray, K. Naidoo, A. Navarro-Alsina, S. Nesseris, F. Passalacqua, K. Paterson, L. Patrizii, A. Pisani, D. Potter, S. Quai, M. Radovich, S. Sacquegna, M. Sahlén, D. B. Sanders, E. Sarpa, A. Schneider, D. Sciotti, D. Scognamiglio, E. Sellentin, L. C. Smith, K. Tanidis, G. Testera, R. Teyssier, S. Tosi, A. Troja, M. Tucci, C. Valieri, A. Venhola, D. Vergani, G. Verza, P. Vielzeuf, N. A. Walton, and J. G. Sorce. Euclid Quick Data Release (Q1) Exploring galaxy properties with a multi-modal foundation model. *ArXiv e-prints*, 2025.

[8] Shoubaneh Hemmati, Jessica Krick, Daniel Stern, Vandana Desai, Andreas Faisst, Lucas Martín-García, Varoujan Gorjian, Aryana Haghjoo, Farnik Nikakhtar, Troy Raen, Sogol Sanjaripour, Brigitta M. Sipőcz, and David Shupe. Reducing the dimensions of active galactic nuclei light-curve manifolds. *The Astrophysical Journal*, 998(1):130, feb 2026.

[9] Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew, et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873:111, 2019.

[10] N. Koblischke and J. Bovy. SpectraFM: Tuning into Stellar Foundation Models. *ArXiv e-prints*, 2024.

[11] H. W. Leung and J. Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *Monthly Notices of the Royal Astronomical Society*, 527(1):1494–1520, 2023.

[12] P. Mousavi, G. Maimon, A. Moumen, D. Petermann, J. Shi, H. Wu, H. Yang, A. Kuznetsova, A. Ploujnikov, R. Marxer, B. Ramabhadran, B. Elizalde, L. Lugosch, J. Li, C. Subakan, P. Woodland, M. Kim, H.-y. Lee, S. Watanabe, Y. Adi, and M. Ravanelli. Discrete Audio Tokens: More Than a Survey! *ArXiv e-prints*, 2025.

[13] L. Parker, F. Lanusse, J. Shen, O. Liu, T. Hehir, L. Sarra, L. Meyer, M. Bowles, S. Wagner-Carena, H. Qu, S. Golkar, A. Bietti, H. Bourfoune, N. Casserau, P. Cornette, K. Hirashima, G. Krawezik, R. Ohana, N. Lourie, M. McCabe, R. Morel, P. Mukhopadhyay, M. Pettee, B. R.-S. Blancard, K. Cho, M. Cranmer, and S. Ho. AION-1: Omnimodal Foundation Model for Astronomical Sciences. *ArXiv e-prints*, 2025.

[14] A. Radford, J. Wu, R. Child, D. Luan, Dario A., and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Whitepaper*, 2019.

[15] A. Raichoor, J. Moustakas, J. A. Newman, T. Karim, S. Ahlen, S. Alam, S. Bailey, D. Brooks, K. Dawson, A. de la Macorra, A. de Mattia, A. Dey, B. Dey, G. Dhungana, S. Eftekharzadeh, D. J. Eisenstein, K. Fanning, A. Font-Ribera, J. García-Bellido, E. Gaztañaga, S. G. A. Gontcho, J. Guy, K. Honscheid, M. Ishak, R. Kehoe, T. Kisner, A. Kremin, T.-W. Lan, M. Landriau, L. Le Guillou, M. E. Levi, C. Magneville, M. Manera, P. Martini, A. M. Meisner, A. D. Myers, J. Nie, N. Palanque-Delabrouille, W. J. Percival, C. Poppett, F. Prada, A. J. Ross, V. Ruhlmann-Kleider, C. G. Sabiu, E. F. Schlafly, D. Schlegel, G. Tarlé, B. A. Weaver, C. Yèche, R. Zhou, Z. Zhou, and H. Zou. Target Selection and Validation of DESI Emission Line Galaxies. *Astronomical Journal*, 165(3):126, 2023.

[16] Sogol Sanjaripour, Archana Aravindan, Gabriela Canalizo, Shoubaneh Hemmati, Bahram Mobasher, Alison L. Coil, and Barry C. Barish. Selection of dwarf galaxies hosting active galactic nuclei: A measure of bias and contamination using unsupervised machine learning techniques. *The Astrophysical Journal*, 992(1):138, oct 2025.

[17] Sogol Sanjaripour, Shoubaneh Hemmati, Bahram Mobasher, Gabriela Canalizo, Barry C. Barish, Irene Shivaei, Alison L. Coil, Nima Chartab, Marziye Jafariyazani, Naveen A. Reddy, and Mojegan Azadi. The application of manifold learning to a selection of different galaxy populations and scaling relation analysis. *The Astrophysical Journal*, 977(2):202, dec 2024.

[18] M. J. Smith and J. E. Geach. Astronomia ex machina: a history, primer and outlook on neural networks in astronomy. *R. Soc. Open Sci.*, 10(5):221454, 2023.

[19] M. J. Smith, R. J. Roberts, E. Angeloudi, and M. Huertas-Company. AstroPT: Scaling Large Observation Models for Astronomy. *ArXiv e-prints*, 2024.

[20] The Multimodal Universe Collaboration. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data. *Advances in Neural Information Processing Systems*, 37:57841–57913, 2024.

[21] M. Tschannen, A. S. Pinto, and A. Kolesnikov. JetFormer: An Autoregressive Generative Model of Raw Images and Text. *ArXiv e-prints*, 2024.

[22] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Neural Information Processing Systems*, 2017.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *arXiv*, 2017.

[24] M. Walmsley, T. Géron, S. Kruk, A. M. M. Scaife, C. Lintott, K. L. Masters, J. M. Dawson, H. Dickinson, L. Fortson, I. L. Garland, et al. Galaxy Zoo DESI: Detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys. *Monthly Notices of the Royal Astronomical Society*, 526(3):4768–4786, 2023.

[25] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.