

---

# Multi-modal Foundation Model for Cosmological Simulation Data

---

**Bin Xia\***

Center for Relativistic Astrophysics, School of Physics  
Georgia Institute of Technology  
Atlanta, GA 30332  
xiabin@gatech.edu

**Nesar Ramachandra†**

Computational Science Division  
Argonne National Laboratory  
Lemont, IL 60439  
nramachandra@anl.gov

**Azton I. Wells†**

Computational Science Division  
Argonne National Laboratory  
Lemont, IL 60439  
awells@anl.gov

**Salman Habib**

Computational Science Division  
Argonne National Laboratory  
Lemont, IL 60439  
habib@anl.gov

**John Wise**

Center for Relativistic Astrophysics, School of Physics  
Georgia Institute of Technology  
Atlanta, GA 30332, USA  
jwise@physics.gatech.edu

## Abstract

We present a multi-modal foundation model for astrophysical galaxy data, designed to map between simulation- and observation-based galactic features. Our encoder-only transformer flexibly ingests scalar quantities (e.g., redshifts, galaxy masses) and vectors (e.g., star formation histories, spectra), supporting multi-task training that includes within-modality reconstruction and cross-modality prediction. With a dynamic masking strategy, the model can query arbitrary galaxy properties from partial inputs—including predicting spectra from redshift and mass, or estimating photometric redshifts from broadband magnitudes—while also recovering missing segments within a modality. Trained on 185,000 simulated galaxies from a gigaparsec-scale Cosmology simulation, the model yields a 50% improvement in redshift estimation when combining LSST and SPHEREx photometry over LSST photometry alone, and a 63% improvement in stellar mass inference when combining late-time SFH with LSST photometry over early-time SFH with LSST photometry. The model demonstrates strong generalization across multi-modal tasks and lays the groundwork for future integration of higher-dimensional and structured data such as images, merger trees, and 3D fields. This approach provides a unified framework for connecting simulations and observations, advancing the development of generalizable astrophysical foundation models.

---

\*Corresponding author. Work completed during an internship at Argonne National Lab.

†These authors contributed equally.

# 1 Introduction

Understanding the connection between cosmological simulations and astronomical observations is a central challenge in modern astrophysics [1, 2, 3]. Large-scale simulations, such as the Last Journey project created using the Hardware/Hybrid Accelerated Cosmology Code (HACC) [4, 5], provide detailed predictions of dark matter structures and their evolution. Galaxies, their formation histories, and synthetic observables are subsequently derived from these simulation outputs through post-processing heuristic models of galaxy-halo connection and stellar population synthesis [6, 7]. Modern hydrodynamical simulations [8, 9] directly simulate galaxies using gas dynamics and subgrid models, in addition to gravity, and require fewer post-processing steps to obtain observable quantities. Meanwhile, wide-field surveys from Rubin Telescope [10] and SPHEREx [11] will deliver high-dimensional observational data across millions of galaxies. Bridging these two domains—simulation-derived physical quantities and observation-driven measurements—is essential for cosmological inference, yet remains difficult due to their heterogeneous representations and incomplete overlap.

Traditional machine learning approaches in cosmology often focus on narrow, single-modality tasks, such as photometric redshift estimation [12, 13] or classifying strong lenses [14]. These models rely on hand-crafted datasets and are not extendable across diverse data sources and representations. In contrast, the success of foundation models in other scientific and engineering domains suggests a new paradigm: transformer-based architectures trained on large [15, 16, 17], heterogeneous, multi-modal datasets can capture general-purpose representations that support flexible downstream tasks. For instance, Gloeckler et al. [18] show that transformer-based probabilistic diffusion models can perform flexible amortized Bayesian inference on simulation-based models, handling missing or unstructured data across diverse tasks. In astrophysics, efforts such as AstroCLIP [19] have demonstrated the potential of contrastive multi-modal pretraining for aligning images and catalogs. A foundation model for stars [20] has demonstrated that transformer-based approaches can effectively unify the modalities of stellar data. Existing efforts, such as OmniJet- $\alpha$  [21], target specific domains in particle physics, while large-scale multimodal datasets, like the Multi-Modal Universe (MMU) [22], underscore the need for truly general-purpose models. These advances indicate that astronomy, with its rich and heterogeneous data landscape, is well-positioned to benefit from such a foundation model approach.

Motivated by these developments, we present a multi-modal foundation model for cosmological simulation data. We adopt an encoder-only transformer, similar to BERT’s encoder [23], that jointly encodes scalar and vector modalities, including redshift, halo mass, stellar mass, star formation histories (SFHs), photometric magnitudes, and spectral energy distributions (SEDs). To enable cross-modality and within-modality reconstruction, we develop a novel dynamic masking strategy within this framework. By training on a subset of galaxies derived from the Last Journey simulation, we demonstrate that the model learns a coherent latent representation supporting flexible inference tasks such as photometric redshift estimation and missing data reconstruction.

## 2 Model Architecture and Methodology

We introduce **MOSAIC** (short for Multi-modal Observation-Simulation Integration for Cosmology), a model designed to enable translation and joint understanding of simulation-only quantities and synthetic observables. Our approach leverages an encoder-only transformer, inspired by BERT’s design [23], to learn shared representations across diverse scalar and vector modalities derived from the massive-volume HACC cosmological simulations [4, 5]. These modalities encompass both simulation-only quantities (e.g., dark matter halo masses) and have been post-processed to include synthetic observations (e.g., photometric magnitudes and SEDs, allowing the model to bridge the gap between simulation and observation through a unified latent space.

**Data Types and Normalization.** Our dataset comprises 185,247 training samples and 20,583 test samples from an extragalactic catalog generated from the Last Journey simulation [5]. While this gravity-only simulation includes only dark matter, we assign galaxies, their formation histories, associated spectra, and broad magnitudes onto the SMACC dark matter cores using a combination of galaxy-halo connection models, stellar population synthesis, and heuristic fits to observational data [5, 24]. Each galaxy sample in the resulting extragalactic catalog comprises a mixture of scalar and vector modalities, encompassing a broad range of galaxy types, environments, and formation histories. However, we note that the distributions of the galactic properties are not representative of any single survey. This extragalactic dataset is a randomly chosen subset used to train the foundation model

and evaluate the model’s ability to perform multi-modal reconstruction and prediction. Realistic extragalactic datasets like this have been successfully deployed on observational targets with pre-defined mapping, like photometric redshift estimations [25], or strong lensing parameter estimation [26], but have not been fully utilized in a ‘task-agnostic’ setting of a foundation model.

We categorize the synthetic catalog entries into scalar and vector modalities. Scalar inputs (zero-dimensional), including *redshift*, *halo mass*, and *stellar mass*, are standardized by subtracting the mean and dividing by the standard deviation computed over the entire training set for each scalar. Vector inputs (one-dimensional) include photometric magnitudes (AB-magnitudes in 6-band LSST (ugrizY) and 102 near-infrared colors from SPHEREx), star formation histories (117 cosmic-time bins, in log-scale units of  $\log_{10} M_{\odot}/\text{yr}$ ), and rest-frame spectral energy distributions (921 wavelength bins, in units of  $\log_{10} \text{Jy}$ ). Each vector modality is normalized by subtracting its global mean and then dividing by its standard deviation, which is computed over the entire training set.

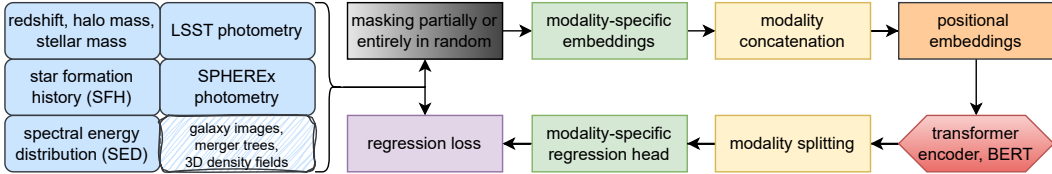


Figure 1: Schematic of MOSAIC architecture. Each modality is masked and projected into modality-specific embeddings, concatenated along the sequence dimension, combined with positional embeddings, and processed by a Transformer encoder. The sequence is split into modality-specific segments and fed into dedicated regression heads to produce predictions and compute regression loss.

**Model Architecture.** As illustrated in Fig. 1, MOSAIC handles scalar and vector modalities using an encoder-only Transformer backbone, similar to BERT [23]. To enable flexible inference from incomplete or heterogeneous data, each modality is first randomly masked, either fully or in contiguous blocks—simulating stretches of missing data—following the hierarchical scheme described in Sec. A. Physically, masking corresponds to simulating missing observations or unmeasured properties for a galaxy. In practice, masked entries are set to zero in our experiments.

The resulting masked and unmasked inputs are projected into modality-specific embeddings (see Sec. B), which map each physical quantity into a continuous latent space. Physically, these embeddings capture the relative magnitudes, patterns, or shapes of the underlying scalar and vector quantities, allowing the model to learn relationships between different galaxy properties. The embeddings from all modalities are then concatenated along the sequence dimension, combined with shared positional embeddings via element-wise addition, and processed by the Transformer encoder. After encoding, the sequence is split back into modality-specific segments, each of which is passed through its dedicated regression head. The model is trained to perform both in-modality reconstruction (predicting missing values within a modality, e.g., SED reconstruction when only visible spectra are available) and cross-modality prediction (predicting one modality from another, e.g., estimating photometric magnitudes from halo mass and SFH). The predictions are compared to the corresponding ground truth to compute regression loss, as detailed in Sec. C.

### 3 Results

We evaluate MOSAIC on five representative input configurations. For each configuration, the model receives partial observations and predicts all target astrophysical quantities. This setup illustrates MOSAIC’s flexibility in handling diverse combinations of available catalog inputs and performing both cross-modality prediction and partial reconstruction within a modality. According to Fig. 2, key prediction performance, measured by the mean absolute error (MAE), is summarized in Table 1, with smaller values indicating better performance. Some entries are omitted, as they correspond to input-target combinations that are less informative for this illustrative problem.

These results indicate several key insights. For stellar mass prediction, late SFH combined with LSST photometry provides the strongest constraint, whereas for redshift estimation, combined photometry (LSST + SPHEREx) or LSST with late SFH achieves the highest accuracy. LSST alone yields moderate performance, demonstrating that multi-modal inputs substantially improve predictive

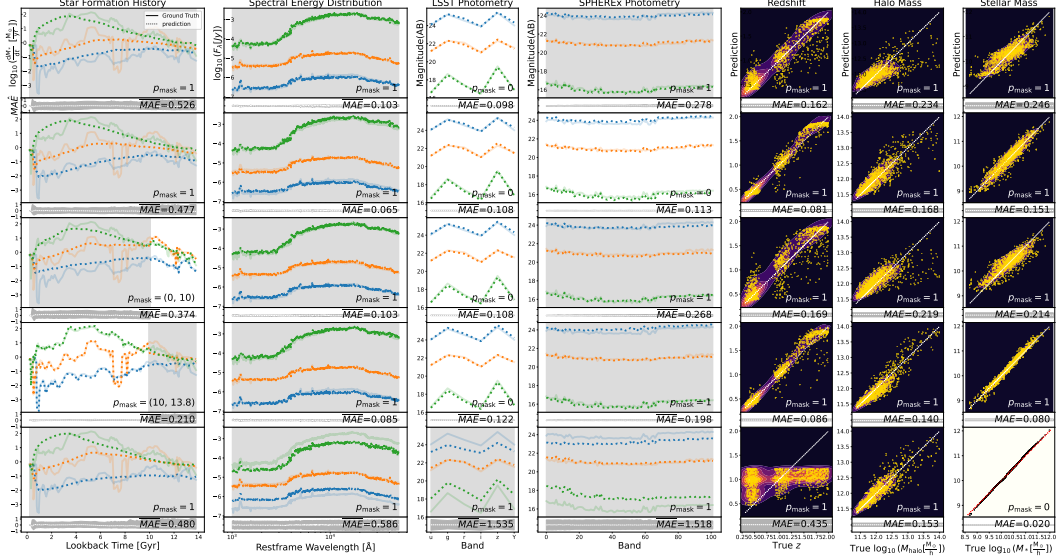


Figure 2: Illustration of multi-modal prediction tasks and input masking configurations. Each row corresponds to a different input combination: (1) LSST magnitudes only, (2) LSST magnitudes + SPHEREx colors, (3) LSST magnitudes + early SFH (0–10 Gyr masked), (4) LSST magnitudes + late SFH (10–13.8 Gyr masked), (5) stellar mass only. White areas indicate provided data, and shaded areas indicate masked portions. For vector modalities, each subplot shows three samples with light solid lines for ground truth and dark dotted lines for predictions. For scalar modalities, each subplot displays scatter points for 1000 samples, with points near the diagonal indicating accurate predictions. Shaded bands indicate the 16–84% range ( $\approx 1\sigma$ ) of the normalized mean absolute error (MAE) computed over 1000 samples; unnormalized mean absolute error ( $\overline{\text{MAE}}$  for vectors and MAE for scalars) is annotated for reference.

Table 1: Mean Absolute Error (MAE) for predictions of redshift, SFH (averaged over time), and stellar masses with different input modalities. Smaller is better.

| Input Modalities                 | Redshift | SFH   | Stellar mass |
|----------------------------------|----------|-------|--------------|
| LSST magnitudes only             | 0.162    | 0.526 | -            |
| LSST magnitudes + SPHEREx colors | 0.081    | 0.477 | -            |
| LSST magnitudes + early SFH      | 0.169    | -     | 0.214        |
| LSST magnitudes + late SFH       | 0.086    | -     | 0.080        |
| Stellar mass only                | 0.435    | 0.480 | -            |

capability. Interestingly, using stellar mass alone as input, while insufficient to accurately predict SED, photometry, or redshift, still allows the model to recover the overall trend of the SFH and provides reasonable estimates of halo mass. Certain tasks, such as SFH prediction from LSST only, remain challenging, reflecting the necessity of complementary modalities to fully constrain complex vector quantities. These experiments demonstrate MOSAIC’s ability to perform multi-task prediction from heterogeneous inputs while highlighting which input combinations are most informative for different astrophysical properties. Scatter plots show some clumping patterns, which arise from the non-uniform sampling of the training and test sets: oversampled parameter regions lead to denser clusters along the  $x$ -axis, while the model’s predictions in high-uncertainty regions concentrate around the most likely outcomes seen in the training set, producing vertical clustering.

To further investigate the physical insights captured by the model, we visualize the last hidden state embeddings using UMAP for the five input configurations in Fig. 3. Within each subplot, embeddings from the five modalities (SFH, SED, SPHEREx, LSST, Scalars) are shown for 10000 samples, enclosed by convex hulls that exclude extreme outliers (points outside the 0.1–99.9% range in  $x$  and  $y$ ). For the first four input configurations, embeddings of different modalities form well-separated clusters, indicating that the model can effectively distinguish and reconstruct distinct modalities. Furthermore, within each cluster, points are colored by their normalized ground-truth

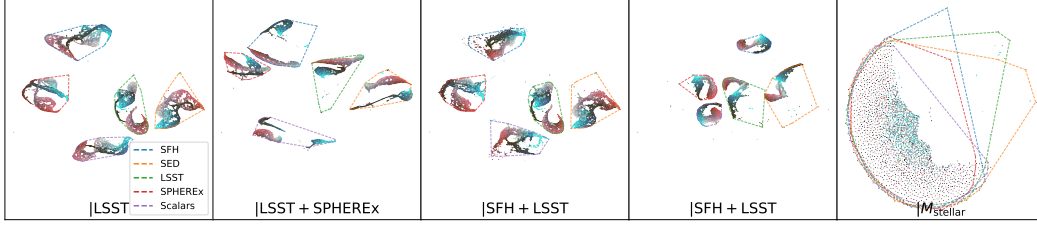


Figure 3: UMAP visualization of the last hidden state embeddings for five input configurations corresponding to Fig. 2. Each subplot shows embeddings of five modalities (SFH, SED, SPHEREx, LSST, Scalars) for 10000 galaxies. Convex hulls are drawn around points between the 0.1 and 99.9 percentiles to reduce outlier effects, with a distinct color for each modality. Point colors encode normalized scalar ground-truth values (redshift, halo mass, stellar mass). For the first four configurations, embeddings form well-separated clusters corresponding to each modality, and within each cluster, points with similar physical properties (similar colors) are located close to each other, forming smooth color gradients. This indicates that the latent space captures astrophysical correlations even when input scalars are masked. For the fifth configuration, where only stellar mass is provided, clusters overlap and color continuity is absent, consistent with weaker predictive performance. The axes correspond to the two UMAP embedding dimensions, which are abstract and unitless.

scalar properties (redshift, halo mass, stellar mass), and galaxies with similar physical properties lie closer together, producing smooth color gradients. This demonstrates that MOSAIC captures meaningful astrophysical correlations in its latent space, even when those scalar quantities are fully masked at the input stage. By contrast, when stellar mass alone is provided (fifth configuration), embeddings lack clear cluster separation, consistent with the weaker predictive performance shown in Fig 2 and Table 1.

## 4 Discussion and Conclusion

Multiwavelength astronomy has proven highly effective, as different surveys probe distinct physical processes [27, 28]. Likewise, simulations provide correlated quantities—that, if brought into a unified framework, offer significant opportunities for new physical insights. MOSAIC can flexibly predict galaxy properties from partial inputs: it can perform photometric redshift and stellar mass estimation, reconstruct SEDs and SFHs, and translate between photometric and spectroscopic measurements from different telescopes. For example, combining LSST and SPHEREx photometry yields an MAE of 0.081 for redshift estimation, while stellar mass can be inferred from late-time SFH combined with LSST photometry with an MAE of 0.080, substantially outperforming early-time SFH (MAE = 0.214). These results demonstrate the complementary roles of photometric and spectroscopic inputs, highlighting the model’s ability to integrate heterogeneous data for robust cross-modality prediction.

Predictive accuracy varies across properties, reflecting underlying astrophysical correlations. Quantities with direct observational imprints, such as redshift from photometric colors or stellar mass from late-time SFH, are more strongly constrained and yield higher accuracy. In contrast, reconstructing full star formation histories from photometry alone remains challenging, since cumulative SED features do not uniquely encode detailed temporal evolution. This explains why some tasks exhibit stronger performance than others and highlights the complementary role of multi-modal inputs.

Compared with existing foundation models in astronomy, which are trained solely on observational data [19, 20], MOSAIC is the first framework designed for large-scale cosmological simulation datasets, aiming to bridge simulations and observations. A key strength in our approach is the masking-based training strategy, which enables learning from incomplete and heterogeneous data. We note that, since our current training relies on simulated samples, predictive performance may partly reflect simulator-dependent biases; future work will incorporate observational datasets when scaling to larger data volumes to mitigate such biases. The clumping patterns observed in Fig. 2 exemplify such sampling-driven biases, highlighting the need for careful treatment of training distributions. Moreover, analysis of the latent space (Fig. 3) suggests that MOSAIC embeddings capture meaningful

astrophysical correlations, with similar galaxies mapped nearby even when input scalars are masked, providing additional interpretability beyond raw predictive accuracy.

While the current implementation focuses on point predictions, the learned multi-modal latent space could naturally support more complex generative tasks. For instance, adding a probabilistic decoder on top of the latent embeddings would allow joint generative modeling of multiple galaxy properties, enabling uncertainty quantification and conditional sampling. Approaches such as variational autoencoders, normalizing flows, or diffusion models could be integrated into the framework to produce full posterior distributions rather than single-point estimates, extending the model’s applicability to simulation-based inference, synthetic data generation, or probabilistic survey planning.

There remains scope for further enhancements: extending to higher-dimensional modalities such as galaxy images, 3D density fields, or merger trees, requiring modality-specific encoders (e.g., convolutional or graph-based) for spatial data before integration into the shared transformer latent space; handling domain shifts when applying to real surveys; exploring pretraining strategies; and evaluating downstream cosmological analyses such as weak lensing and galaxy clustering. In the near term, MOSAIC could be applied directly to observational datasets to infer missing galaxy properties or to simulations to generate synthetic observables, offering a versatile tool for survey planning, analysis, and astrophysical interpretation.

## Acknowledgments and Disclosure of Funding

This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DEAC02-06CH11357. The training is conducted on Swing, a GPU system at the Laboratory Computing Resource Center (LCRC) of Argonne National Laboratory, and utilizes the resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy User Facility, with an NERSC award, GenAI@NERSC.

## References

- [1] Mark Vogelsberger, Federico Marinacci, Paul Torrey, and Ewald Puchwein. Cosmological simulations of galaxy formation. *Nature Reviews Physics*, 2(1):42–66, 2020.
- [2] Thorsten Naab and Jeremiah P. Ostriker. Theoretical Challenges in Galaxy Formation. *Annual Review of Astronomy and Astrophysics*, 55(1):59–109, August 2017.
- [3] Rachel S. Somerville and Romeel Davé. Physical Models of Galaxy Formation in a Cosmological Framework. *Annual Review of Astronomy and Astrophysics*, 53:51–113, August 2015.
- [4] Salman Habib, Adrian Pope, Hal Finkel, Nicholas Frontiere, Katrin Heitmann, David Daniel, Patricia Fasel, Vitali Morozov, George Zagaris, Tom Peterka, et al. Hacc: Simulating sky surveys on state-of-the-art supercomputing architectures. *New Astronomy*, 42(C), 07 2015.
- [5] Katrin Heitmann, Nicholas Frontiere, Esteban Rangel, Patricia Larsen, Adrian Pope, Imran Sultan, Thomas Uram, Salman Habib, Hal Finkel, Danila Korytov, Eve Kovacs, Silvio Rizzi, Joe Insley, and Janet Y. K. Knowles. The Last Journey. I. An Extreme-scale Simulation on the Mira Supercomputer. *Astrophysical Journal Supplement Series*, 252(2):19, February 2021.
- [6] Danila Korytov, Andrew Hearin, Eve Kovacs, Patricia Larsen, Esteban Rangel, Joseph Hollowed, Andrew J Benson, Katrin Heitmann, Yao-Yuan Mao, Anita Bahmanyar, et al. Cosmodc2: A synthetic sky catalog for dark energy science with lsst. *The Astrophysical Journal Supplement Series*, 245(2):26, 2019.
- [7] Alex Alarcon, Andrew P. Hearin, Matthew R. Becker, and Jonás Chaves-Montero. Diffstar: a fully parametric physical model for galaxy assembly history. *MNRAS*, 518(1):562–584, January 2023.
- [8] Rüdiger Pakmor, Volker Springel, Jonathan P Coles, Thomas Guillet, Christoph Pfrommer, Sownak Bose, Monica Barrera, Ana Maria Delgado, Fulvio Ferlito, Carlos Frenk, et al. The millenniumtng project: the hydrodynamical full physics simulation and a first look at its galaxy clusters. *Monthly Notices of the Royal Astronomical Society*, 524(2):2539–2555, 2023.
- [9] Nicholas Frontiere, JD Emberson, Michael Buehlmann, Joseph Adamo, Salman Habib, Katrin Heitmann, and Claude-André Faucher-Giguère. Simulating hydrodynamics in cosmology with crk-hacc. *The Astrophysical Journal Supplement Series*, 264(2):34, 2023.

- [10] Quentin Le Boulc’h, Fabio Hernandez, and Gabriele Mainetti. The rubin observatory’s legacy survey of space and time dp0. 2 processing campaign at cc-in2p3. In *EPJ Web of Conferences*, volume 295, page 04049. EDP Sciences, 2024.
- [11] Brendan P. Crill, Michael Werner, Rachel Akeson, Matthew Ashby, Lindsey Bleem, James J. Bock, Sean Bryan, Jill Burnham, Joyce Byunh, Tzu-Ching Chang, Yi-Kuan Chiang, Walter Cook, Asantha Cooray, Andrew Davis, Olivier Doré, C. Darren Dowell, Gregory Dubois-Felsmann, Tim Eifler, Andreas Faisst, Salman Habib, Chen Heinrich, Katrin Heitmann, Grigory Heaton, Christopher Hirata, Viktor Hristov, Howard Hui, Woong-Seob Jeong, Jae Hwan Kang, Branislav Kecman, J. Davy Kirkpatrick, Phillip M. Korngut, Elisabeth Krause, Bomee Lee, Carey Lisse, Daniel Masters, Philip Mauskopf, Gary Melnick, Hiromasa Miyasaka, Hooshang Nayyeri, Hien Nguyen, Karin Öberg, Steve Padin, Roberta Paladini, Milad Pourrahmani, Jeonghyun Pyo, Roger Smith, Yong-Seong Song, Teresa Symons, Harry Teplitz, Volker Tolls, Steve Unwin, Rogier Windhorst, Yujin Yang, and Michael Zemcov. SPHEREx: NASA’s near-infrared spectrophotometric all-sky survey. In Makenzie Lystrup and Marshall D. Perrin, editors, *Space Telescopes and Instrumentation 2020: Optical, Infrared, and Millimeter Wave*, volume 11443 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 114430I, December 2020.
- [12] Jeffrey A Newman and Daniel Gruen. Photometric redshifts for next-generation surveys. *Annual Review of Astronomy and Astrophysics*, 60(1):363–414, 2022.
- [13] Antonio D’Isanto and Kai Lars Polsterer. Photometric redshift estimation via deep learning-generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111, 2018.
- [14] AJ Shajib, G Vernardos, TE Collett, V Motta, Dominique Sluse, LLR Williams, Prasenjit Saha, S Birrer, C Spiniello, and Tommaso Treu. Strong lensing by galaxies. *Space Science Reviews*, 220(8):87, 2024.
- [15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.
- [16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [17] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Saucedo Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.
- [18] Manuel Gloeckler, Michael Deistler, Christian Weillbach, Frank Wood, and Jakob H. Macke. All-in-one simulation-based inference. *arXiv e-prints*, page arXiv:2404.09636, April 2024.
- [19] Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, et al. Astroclip: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, 2024.
- [20] Henry W Leung and Jo Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *Monthly Notices of the Royal Astronomical Society*, 527(1):1494–1520, October 2023.
- [21] Joschka Birk, Anna Hallin, and Gregor Kasieczka. Omnijet- $\alpha$ : the first cross-task foundation model for particle physics. *Machine Learning: Science and Technology*, 5(3):035031, 2024.
- [22] The Multimodal Universe Collaboration, Jeroen Audenaert, Micah Bowles, Benjamin M. Boyd, David Chemaly, Brian Cherinka, Ioana Ciucă, Miles Cranmer, Aaron Do, Matthew Grayling, Erin E. Hayes, Tom Hehir, Shirley Ho, Marc Huertas-Company, Kartheik G. Iyer, Maja Jablonska, Francois Lanusse, Henry W. Leung, Kaisey Mandel, Juan Rafael Martínez-Galarza, Peter Melchior, Lucas Meyer, Liam H. Parker, Helen Qu, Jeff Shen, Michael J. Smith, Connor Stone, Mike Walmsley, and John F. Wu. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data. *arXiv e-prints*, page arXiv:2412.02527, December 2024.
- [23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

- [24] Imran Sultan, Nicholas Frontiere, Salman Habib, Katrin Heitmann, Eve Kovacs, Patricia Larsen, and Esteban Rangel. The Last Journey. II. SMACC—Subhalo Mass-loss Analysis Using Core Catalogs. *ApJ*, 913(2):109, June 2021.
- [25] Nesar Ramachandra, Jonás Chaves-Montero, Alex Alarcon, Arindam Fadikar, Salman Habib, and Katrin Heitmann. Machine learning synthetic spectra for probabilistic redshift estimation: Syth-z. *Monthly Notices of the Royal Astronomical Society*, 515(2):1927–1941, 2022.
- [26] François Lanusse, Quanbin Ma, Nan Li, Thomas E Collett, Chun-Liang Li, Siamak Ravanbakhsh, Rachel Mandelbaum, and Barnabás Póczos. Cmu deeplens: deep learning for automatic image-based galaxy–galaxy strong lens finding. *Monthly Notices of the Royal Astronomical Society*, 473(3):3895–3906, 2018.
- [27] Péter Mészáros, Derek B Fox, Chad Hanna, and Kohta Murase. Multi-messenger astrophysics. *Nature Reviews Physics*, 1(10):585–599, 2019.
- [28] Michael A Troxel, C Lin, A Park, C Hirata, Rachel Mandelbaum, M Jarvis, Ami Choi, J Givans, M Higgins, B Sanchez, et al. A joint roman space telescope and rubin observatory synthetic wide-field imaging survey. *Monthly Notices of the Royal Astronomical Society*, 522(2):2801–2820, 2023.

## A Masking Strategy

To support both in-modality reconstruction and cross-modality prediction, we employ a hierarchical masking scheme applied at data loading time. This design ensures balanced training across different modality combinations while also simulating realistic scenarios of incomplete data.

**Implementation Details of Masking.** Masked entries are set to a fixed value, `mask_token = 0`, corresponding to the mean of each modality after normalization. This value lies within the physical range and has been empirically found to yield the best performance. Both masked and unmasked inputs are passed through the model during training, allowing it to leverage available information while reconstructing missing values. This approach stabilizes training and improves generalization across diverse cross-modal and in-modality prediction tasks.

**Global Modality Masking.** Each modality, scalar or vector, is independently masked entirely with probability 0.5. Consequently, all  $2^N$  possible modality subsets (for  $N$  modalities) occur with equal probability  $2^{-N}$ , ensuring balanced exposure to every cross-modal configuration during training.

**Masking for Scalar Modalities.** All scalar inputs (e.g., redshift, halo mass, stellar mass) are concatenated into a single vector, denoted as *scalars*. Each element within this vector is independently masked with probability 0.5. Unlike vector modalities, scalars do not undergo partial token masking; masking corresponds simply to replacing the value with `mask_token`.

**Local Token Masking for Vector Modalities.** For vector modalities (e.g., SED or SFH) that are not fully masked, we apply *partial token masking*. A masking ratio  $p = 0.2$  defines the total number of tokens to mask:

$$N_{\text{mask}} = dp \quad de,$$

where  $d$  is the sequence length. Masking is performed iteratively by sampling spans of random lengths and positions until the budget  $N_{\text{mask}}$  is exhausted. By sampling spans of varying lengths, the masked tokens include both isolated points and contiguous segments, encouraging the model to learn local interpolation as well as global distributional patterns. Since the masking ratio is relatively small, overlaps between spans are allowed.

## B Input Embedding and Representation

All inputs are organized into modality-specific token sequences. In particular, all scalar quantities (e.g., redshift, halo mass, stellar mass) are concatenated into a single modality denoted as *scalars*, so that they can be treated uniformly as a sequence of tokens. Vector modalities (e.g., SED, SFH) naturally form token sequences according to their sampled dimensions.

Each token, whether from *scalars* or a vector modality, is mapped into the hidden space via a modality-specific linear projection. For a given modality  $M$ , let  $x_{b,\ell}^{(M)}$  denote the value of token  $\ell$  in batch  $b$ , and let  $L_M$  be the sequence length of modality  $M$ . Each token is mapped into the hidden space via a modality-specific linear projection:

$$\mathbf{h}_{b,\ell}^{(M)} = W_M x_{b,\ell}^{(M)} + b_M, \quad W_M \in \mathbb{R}^{d_{\text{hidden}} \times 1}, \quad b_M \in \mathbb{R}^{d_{\text{hidden}}},$$

yielding the embedded representation  $\mathbf{H}^{(M)} \in \mathbb{R}^{B \times L_M \times d_{\text{hidden}}}$  for modality  $M$ .

This design avoids discrete tokenization, which is natural for text but suboptimal for continuous scientific data, while preserving numerical continuity and allowing each modality to learn a flexible embedding in a shared representation space. All embedded modalities are then concatenated, augmented with positional encodings, and processed by a shared Transformer encoder.

## C Training Strategy and Loss Function

We adopt masked regression as the primary training objective. For each modality  $M$  with sequence length  $L_M$ , let  $m_{b,\ell}^{(M)} \in \{0, 1\}$  denote the mask indicator (1 if masked, 0 otherwise). The model receives

$$\mathbf{x}_{b,\ell}^{(M)} = \begin{cases} \text{mask\_token}, & m_{b,\ell}^{(M)} = 1, \\ x_{b,\ell}^{(M)}, & m_{b,\ell}^{(M)} = 0, \end{cases}$$

and produces predictions  $\hat{x}_{b,\ell}^{(M)}$ . In practice, we set `mask_token` = 0, which corresponds to the global mean of normalized data, as we found this choice performs best across modalities. The per-token loss is

$$L_{b,\ell}^{(M)} = (\hat{x}_{b,\ell}^{(M)} - x_{b,\ell}^{(M)})^2.$$

The loss for modality  $M$  is the mean over all tokens in the batch:

$$L^{(M)} = \frac{1}{BL_M} \sum_{b=1}^B \sum_{\ell=1}^{L_M} L_{b,\ell}^{(M)}.$$

Finally, the overall training objective averages equally over all  $K$  modalities:

$$L = \frac{1}{K} \sum_{M=1}^K L^{(M)}.$$

**Why include unmasked tokens?** While the loss on masked tokens drives imputation and cross-modal reasoning, we also include reconstruction loss on unmasked tokens. This acts as a form of curriculum: initially, the model can reliably learn the trivial identity mapping, which stabilizes optimization and anchors the representation space. Once this foundation is established, the model progressively learns to leverage contextual and cross-modal information to recover masked values. In this way, unmasked-token supervision accelerates convergence, regularizes training, and improves overall robustness.

Moreover, all tokens within a modality are treated with equal weight, regardless of whether they are involved in identity reconstruction, imputation, or cross-modality prediction. Similarly, losses from different modalities are averaged with equal weight, independent of their token counts. This design avoids bias toward particular tasks or modalities, ensuring a balanced and unbiased training signal that promotes generalizable representations.