Self-supervised Multimodal Model for Astronomy

Mariia Rizhko
Department of Astronomy
University of California, Berkeley
Berkeley, CA, USA
mariia.rizhko@berkeley.edu

Joshua S. Bloom
Department of Astronomy
University of California, Berkeley
Berkeley, CA, USA
joshbloom@berkeley.edu

Abstract

While machine-learned models are now routinely employed to facilitate astronomical inquiry, model inputs tend to be limited to a primary data source (namely images or time series) and, in the more advanced approaches, some metadata. Yet with the growing use of wide-field, multiplexed observational resources, individual sources of interest often have a broad range of observational modes available. Here we construct an astronomical multimodal dataset and propose a self-supervised pre-training approach that enables a model to learn from multiple modalities simultaneously. Specifically, we extend the CLIP (Contrastive Language-Image Pretraining) model to a trimodal setting, allowing the integration of time-series photometry data, spectra, and astrophysical metadata. In a fine-tuning supervised setting, our results demonstrate that CLIP pre-training improves classification performance for time-series photometry, where accuracy increases from 84.6% to 91.5%. Furthermore, CLIP boosts classification accuracy by up to 12.6% when the availability of labeled data is limited, showing the effectiveness of leveraging larger corpora of unlabeled data. To our knowledge this is the first construction of an $n > 2$ mode model in astronomy. Extensions to $n > 3$ modes is naturally anticipated with this approach.

1 Introduction

Despite the vast volumes of publicly available raw astronomical data, with a few notable subfield exceptions, the application of machine learning to discovery and inference has yet to broadly permeate the field. One impediment stems from the challenge of fusing data across heterogeneous modes of collection. Off-the-shelf architectures do not easily accommodate an admixture of irregularly sampled multi-spectral multi-scale heteroskedatic time-series data, images, spectra, and metadata. Another issue, arising in the classification context, is that very few ground-truth labels exist in a given context. This "small label" problem arose, for example, in Richards et al. (2012) who sought to probabilistically classify 50,124 variable stars using only 810 labels over 28 classes. Last, models learned on a dataset from one survey do not easily transfer to other data collected on the same objects from different surveys (e.g., Long et al. 2012; Kim et al. 2021). Our self-supervised multimodal architecture addresses the first two challenges, establishing methods and milestones for a more generalized foundation model applicable to inference tasks on unseen survey data.

Our work builds upon the Contrastive Language-Image Pretraining (CLIP) framework, originally introduced by Radford et al. (2021); CLIP demonstrated the power of contrastive learning on large-scale image and text datasets to learn joint representations. Since its introduction, CLIP has been extensively researched and improved in various ways. For example, Li et al. (2021) enhanced data efficiency through supervision, while Yao et al. (2021) focused on improving semantic alignment. Cherti et al. (2023) introduced scaling laws, and Sun et al. (2023) optimized the model for faster

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

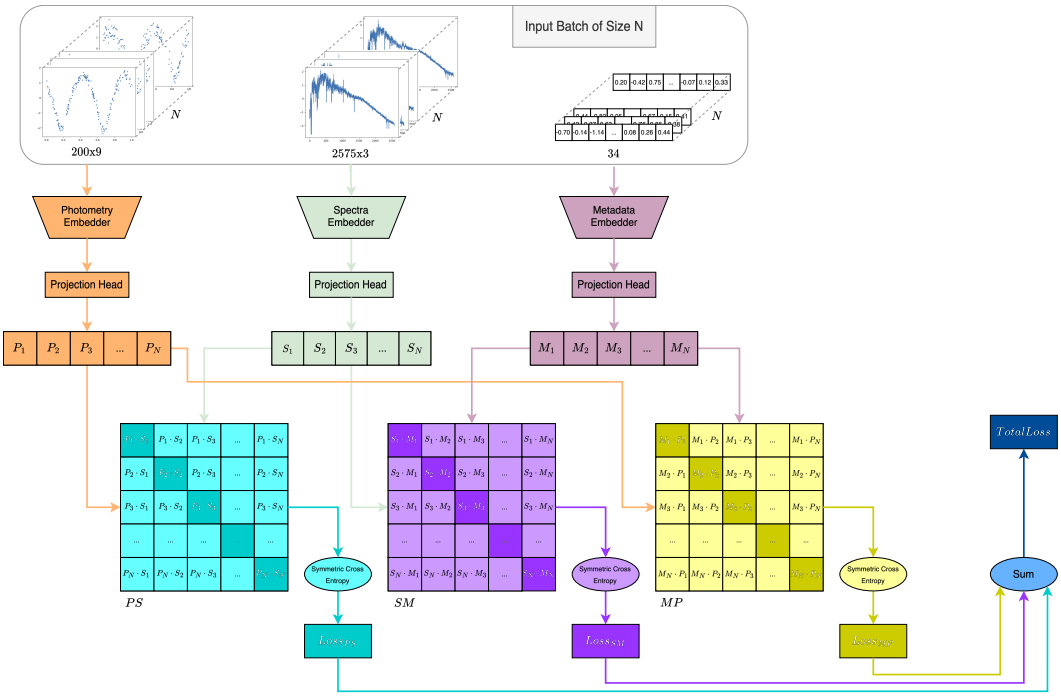

Figure 1: Overview of the multimodal CLIP framework extended to three modalities: photometry, spectra, and metadata. Each modality is processed by its respective encoder to produce embeddings, which are then aligned in a shared embedding space using a symmetric cross-entropy loss over pairwise similarity matrices.

training. Additionally, CLIP has been combined with other pretraining objectives: Mu et al. (2022) incorporated image self-supervision, and Singh et al. (2022) along with Li et al. (2022) added masked multimodal, image, and language modeling. Furthermore, CLIP has been extended to other modalities: audio-text (Wu et al., 2023), video-text (Luo et al., 2021; Xu et al., 2021; Ma et al., 2022), and point cloud-text (Zhang et al., 2022). In the astronomical context, Parker et al. (2024) used dual-mode CLIP on static-sky galaxy images and spectra. Closest to the approach of our work outside of astronomy, Guzhov et al. (2022) adapted CLIP for use with three modalities: audio, image, and text. Given the proven versatility and success of CLIP, we build upon this approach herein. We extend CLIP to work on three modalities: time-series photometry, spectra, and metadata (see Figure 1). Our work and a recent preprint from Zhang et al. (2024) are the first efforts to incorporate time-series data with CLIP and ours is the only three-mode model in astronomy, a critical step towards foundational multimodal model for time-domain astronomy.

2 Related Work

Early classification-focused research used hand-crafted features of time-series photometry and metadata with decision forests in a supervised context (Debosscher et al., 2007; Richards et al., 2011; Dubath et al., 2011; Palaversa et al., 2013). Neural network approaches to learn representations of time-series photometry (both in supervised and self-supervised contexts) then achieved state of the art, first with flavors of RNNs (e.g., LSTMs: Naul et al. 2018, GRUs: Muthukrishna et al. 2019; Becker et al. 2020) and more recently with convolution (Jamal & Bloom, 2020; Boone, 2021) and Transformers (Donoso-Oliva et al., 2023; Leung & Bovy, 2024). CNNs have been used to achieve state of the art classification on galaxy spectra (e.g., GalSpecNet: Wu et al. 2024a). Hayat et al. (2021) use CNN autoencoders with contrastive learning for self-supervised embedding of galaxy images.

AstroCLIP (Parker et al., 2024) fused pre-trained embeddings of galaxy spectra and images with constrastive learning and showed the trained model to be competitive with purpose-built classification models. Our work differs from AstroCLIP in that 1) our primary objects are individual sources that vary in time (ie. not static like galaxies), 2) we explicitly build embeddings for three different modes of data, 3) our approach does not rely upon pretraining of embeddings for the different modes, but instead learns all embeddings simultaneously, and 4) we examine the efficacy of the model with missing modes at test time. Like with AstroCLIP we find our model outperforms purpose-built supervised models for downstream tasks. To our knowledge, MAVEN (Zhang et al., 2024) is the only other CLIP-centric model applied in the astronomical time domain. It is a dual-mode model built for "one off" explosive supernovae events, whereas ours is focused on persistently variable sources. MAVEN first learns spectroscopic and photometric embeddings from synthetic data and then requires a fine-tuning step on real survey data. Our work model is trained directly on real observational data.

3 Dataset Assembly

The basis of our observational dataset is the variable star catalog observed and curated (Jayasinghe et al., 2019) by the All-Sky Automated Survey for SuperNovae (ASAS-SN) project (Shappee et al., 2014). We downloaded the lightcurve data from the 2021 assembly of the 687,695 v-band variables and the 2022 assembly of the 378,861 g-band variables, along with the associated metadata catalogs. These catalogs contain cross-matched photometry information for each source from WISE (Wright et al., 2010), GALEX (Morrissey et al., 2007), 2MASS (Skrutskie et al., 2006) and Gaia EDR3 (Gaia Collaboration et al., 2021), variability statistics derived from the lightcurves in each bandpass (such as period and peak-to-peak amplitude), astrometric information from Gaia (such as parallax and proper motion), and a machine-learned classification from the ASAS-SN group (Jayasinghe et al., 2019). We deduplicated and merged these data using the cross-matched `source_id` from Gaia EDR3, with the merged catalog serving as the basis of the `metadata` mode.

To facilitate the use of positional information in the models, we transformed the galactic latitude to $b \rightarrow \sin(b)$ and galactic longitude to $l \rightarrow \cos(l)$. We also transformed all catalog apparent photometry m to absolute magnitude using the Gaia EDR3 parallax π (units of milliarcseconds) using $M = m + 5\log_{10}\pi - 10$. We did not deredderen any values. To cleanly delineate the `time-series` mode from the `metadata` mode, we removed features derived from photometric time-series data from the `metadata` catalog (and later used such features as auxiliary inputs in the `time-series` channel, see 4.1 below). We also removed any columns from the `metadata` catalog related to indices (such as source names). Last, we removed the assigned classification of each source (later used to test downstream tasks; see 5).

To build the `spectral` mode, we cross-matched the sources with the v2.0 DR9 Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012) public catalog using the Gaia EDR3 ID. We downloaded the 41,204 1D spectra identified in the the cross match and constructed a lookup table matching specific variable sources to LAMOST spectra. Most variable sources had zero associated spectra but a small subset had multiple spectra of the same source obtained over multiple epochs.

We filtered the dataset based on the following criteria: (1) each object must have data available for all three modalities—time-series photometry, spectra, and metadata; (2) the metadata cannot have any missing values to ensure a complete dataset for training; and (3) the object must belong to one of the top 10 classes to ensure there are sufficient samples for effective CLIP training (Xu et al., 2023; Alabdulmohsin et al., 2024). The selected classes and the corresponding number of objects are listed in Table 1.

4 Method

Our objective is to develop a self-supervised multimodal model that can learn from astronomical data across three distinct modalities: time-series photometry, spectra, and astrophysical metadata. To achieve this, we extend the Contrastive Language-Image Pretraining (CLIP) framework (Radford et al., 2021) to a trimodal setting, enabling simultaneous learning from multiple data types. In this section, we describe the models used for each modality and how they are integrated into our multimodal CLIP framework.

Class	Description	Total Objects
EW	W Ursae Majoris type binaries	6168
SR	Semi-regular variables	4590
EA	Detached Algol-type binaries	2916
RRAB	Fundamental Mode RR Lyrae variables	2351
EB	β Lyrae-type binaries	1976
ROT	Spotted Variables with rotational modulation	1839
RRC	First Overtone RR Lyrae variables	796
HADS	High amplitude δ Scuti type variables	281
M	Mira variables	268
DSCT	δ Scuti type variables	255

Table 1: Summary of variable star classes, including abbreviations, descriptions, and total object counts for each class used in the dataset.

4.1 Photometric Time-Series Model

Photometric time-series data are flux measurements of astronomical objects over time. To effectively capture the temporal dependencies and handle sequences of varying lengths, we employ the Encoder component from the Informer model (Zhou et al., 2021).

Model Architecture. The photometric time-series encoder consists of:

- Input Embedding Layer: Projects the input features to a higher-dimensional space.
- Informer Encoder Layers: Eight encoder layers with a hidden dimension of 128, four attention heads, and a feedforward dimension of 512.
- Output Layer: Produces a fixed-length embedding representing the input time-series data.

Data Preprocessing. Each light curve is a sequence of flux measurements $f = \{f_1, f_2, \ldots, f_T\}$ and flux errors $\sigma_f = \{\sigma_{f_1}, \sigma_{f_2}, \ldots, \sigma_{f_T}\}$ at corresponding times $t = \{t_1, t_2, \ldots, t_T\}$. We normalize the flux by subtracting the mean μ_f and dividing by the median absolute deviation MAD_f: $\tilde{f}_i = \frac{f_i - \mu_f}{\mathrm{MAD}_f}$. Flux errors are normalized by the flux median absolute deviation division: $\tilde{\sigma}_{f_i} = \frac{\sigma_{f_i}}{\mathrm{MAD}_f}$. Time is scaled between 0 and 1 for each light curve: $\delta_t = t_{\max} - t_{\min}$; $\tilde{t}_i = \frac{t_i - t_{\min}}{\delta_t}$. Auxiliary features such as amplitude, period, Lafler-Kinmann string length statistic (Lafler & Kinman, 1965), peak-to-peak variability, delta time $\frac{\delta_t}{365}$ and logarithm of median absolute deviation $\log \mathrm{MAD}_f$ are included as additional inputs.

Handling Variable Sequence Lengths. We set a maximum sequence length of $L = 200$. Sequences longer than this are randomly cropped during training and center-cropped during validation and testing. Shorter sequences are padded with zeros, and an attention mask is used to differentiate between valid data and padding.

4.2 Spectra Model

Spectral data provides detailed information about the composition and physical properties of astronomical objects. We adapt the GalSpecNet architecture (Wu et al., 2024b), which is specifically designed for processing one-dimensional astronomical spectra.

Model Architecture. The spectra encoder consists of:

- Convolutional Layers: Four layers (64, 64, 32, 32 channels) followed by ReLU activations.
- Pooling Layers: Max-pooling layers after each convolutional layer except for the last one.
- Dropout Layer: Applied after the last convolutional layer for regularization.
- Output Layer: Generates a fixed-length embedding of the spectral data.

Modifications. We replace the original fully connected layers with a single fully connected layer for classification tasks or omit it entirely when the model serves as a feature extractor. We also add additional input channels for spectra errors and auxiliary data.

Data Preprocessing. Spectra are limited to the wavelength range of 3850–9000 Å and resampled at regular intervals of 2Å using linear interpolation. Each spectrum $s = \{s_1, s_2, \ldots, s_W\}$ and its uncertainties $\sigma_s = \{\sigma_{s_1}, \sigma_{s_2}, \ldots, \sigma_{s_W}\}$ at corresponding wavelengths $w = \{w_1, w_2, \ldots, w_W\}$ are normalized in a similar way as photometry data: values are normalized by subtracting the mean μ_s and dividing by the median absolute deviation MAD_s: $\tilde{s}_i = \frac{s_i - \mu_s}{\mathrm{MAD}_s}$, while uncertainties are divided by MAD_s: $\tilde{\sigma}_{s_i} = \frac{\sigma_{s_i}}{\mathrm{MAD}_s}$. The logarithm of the median absolute deviation $\log \mathrm{MAD}_s$ is included as an auxiliary feature.

4.3 Metadata Model

The metadata modality consists of astrophysical parameters and observational data not included in the other two modalities. This includes features like absolute magnitudes in various bands, astrometric information, and other cross-matched catalog data. A full list of features and their descriptions is provided in Table 5.

Model Architecture. The metadata encoder is a Multilayer Perceptron consisting of:

- Input Layer: Accepts the 34 preprocessed features.
- Hidden Layers: Two hidden layers with 512 units each followed by ReLU activations.
- Dropout Layers: Applied after hidden layers for regularization.
- Output Layer: Provides a fixed-length metadata embedding.

Data Preprocessing. Except for the steps already mentioned during the dataset assembly (see 3), we apply logarithm to period and then standardize each feature to have zero mean and unit variance.

4.4 Multi-modal CLIP Model

To integrate the three modalities we extend the CLIP model to a trimodal setting. The goal is to learn a shared embedding space where representations from different modalities corresponding to the same astronomical object are close together (see Figure 1).

Projection Heads. Each modality has its own architecture, producing embeddings of different sizes. To bring these embeddings into a shared space, we apply a projection head to each modality. The projection head is a fully connected layer that maps the embeddings to a fixed size of 512. Let the original embeddings of photometry, spectra, and metadata be denoted as $\tilde{P}_i$, $\tilde{S}_i$, and $\tilde{M}_i$, where i denotes the i-th sample in a batch of size N. The projection heads transform these original embeddings as follows:

$$P_i = W_P \tilde{P}_i + b_P \tag{1}$$

$$S_i = W_S \tilde{S}_i + b_S \tag{2}$$

$$M_i = W_M \tilde{M}_i + b_M, \tag{3}$$

where W_P, W_S, and W_M are the weight matrices, and b_P, b_S, and b_M are the bias terms for the projection head of each modality. After applying these transformations, the projected embeddings P_i, S_i, and M_i all have a fixed size of 512, making them suitable for comparison in the shared embedding space.

Pairwise Similarity Matrices. For each pair of modalities (photometry-spectra, spectra-metadata, metadata-photometry) we compute similarity matrices using cosine similarity:

$$PS_{ij} = \frac{P_i \cdot S_j}{\|P_i\|\|S_j\|} \tag{4}$$

$$SM_{ij} = \frac{S_i \cdot M_j}{\|S_i\|\|M_j\|} \tag{5}$$

$$MP_{ij} = \frac{M_i \cdot P_j}{\|M_i\|\|P_j\|} \tag{6}$$

Contrastive Loss. We use a symmetric cross-entropy loss to align the embeddings:

$$\mathcal{L}^{PS} = \mathcal{L}_{\text{CE}}(PS, Y) + \mathcal{L}_{\text{CE}}(PS^\top, Y) \tag{7}$$

$$\mathcal{L}^{SM} = \mathcal{L}_{\text{CE}}(SM, Y) + \mathcal{L}_{\text{CE}}(SM^\top, Y) \tag{8}$$

$$\mathcal{L}^{MP} = \mathcal{L}_{\text{CE}}(MP, Y) + \mathcal{L}_{\text{CE}}(MP^\top, Y) \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss and Y is the label matrix defined as:

$$Y_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Total Loss. The overall loss is the sum of the individual pairwise losses:

$$\mathcal{L} = \mathcal{L}^{PS} + \mathcal{L}^{SM} + \mathcal{L}^{MP} \tag{11}$$

By minimizing this loss, the model learns to align the embeddings across all three modalities, bringing representations of the same object closer together in the embedding space while pushing apart those of different objects.

5 Results

We evaluated the models on downstream classification across four modes: photometry only, spectra only, metadata only, and all modalities combined. For single modalities, we added a fully connected layer on top of the respective encoders for classification. In the multimodal setting, we averaged the embeddings from all three modalities and then applied a fully connected layer for classification. Each model was trained both with and without CLIP pre-training. "With CLIP pre-training" indicates that the model was initially trained using the CLIP framework, followed by fine-tuning the encoders for the downstream task. "Without CLIP pre-training" refers to models trained directly on the downstream task with randomly initialized weights. The training setup and hyperparameter search process are detailed in Appendix A. All models were cross-validated using 5 random seeds and data splits to ensure robust evaluation.

5.1 CLIP Evaluation

The results in Table 2 show that while there is no statistically significant difference between using CLIP and not using CLIP for spectra, metadata and combined modalities, CLIP has a strong impact on photometry classification. It increased the average accuracy **from 84.64% to 91.47%** and significantly reduced the standard deviation (from 6.32 to 0.45), indicating better model stability. With or without CLIP, we also show that *by using all three modalities at the same time, we achieve better accuracy than by using any single modality alone*.

5.2 Limited Labeled Data

To evaluate the effectiveness of CLIP pre-training when the availability of labeled data is limited, we conducted experiments on smaller subsets of the original dataset. Specifically, we created subsets containing 10%, 25%, and 50% of the data by downsampling the most common classes, ensuring a balanced class distribution. Table 3 provides details on the class distribution across these subsets.

Data Type	No CLIP	CLIP
Spectra	76.278 ± 0.931	77.396 ± 0.614
Metadata	85.623 ± 0.628	85.855 ± 0.856
Photometry	84.642 ± 6.317	$\mathbf{91.468 \pm 0.446}$
All	94.065 ± 0.390	94.153 ± 0.577

Table 2: Comparison of accuracy between models with and without CLIP. Statistically important results are in bold.

Class	Train				Val				Test			
	Full	**50%**	**25%**	**10%**	**Full**	**50%**	**25%**	**10%**	**Full**	**50%**	**25%**	**10%**
EW	4890	1209	516	166	597	149	64	21	681	160	69	22
SR	3647	1209	516	166	479	149	64	21	464	160	69	22
EA	2343	1209	516	166	272	149	64	21	301	160	69	22
RRAB	1886	1209	516	166	231	149	64	21	234	160	69	22
EB	1571	1209	516	166	207	149	64	21	198	160	69	22
ROT	1454	1209	516	166	189	149	64	21	196	160	69	22
RRC	624	624	516	166	93	93	64	21	79	79	69	22
HADS	226	226	226	166	29	29	29	21	26	26	26	22
M	216	216	216	166	30	30	30	21	22	22	22	22
DSCT	206	206	206	166	25	25	25	21	24	24	24	22

Table 3: Class distribution across different dataset splits (Full, 50%, 25%, 10%) for training, validation, and test sets.

Note that we choose to downsample the overrepresented sources at random. An interesting alternative to this, to approximate the ways in which brighter sources preferentially are easier to label on new survey data, would be to select only the brightest (or highest signal-to-noise) sources to include in the training data.

Models. For each subset, we retrained all models, with and without CLIP pre-training, using the same optimization settings and hyperparameter search as previously applied. It is important to note that the CLIP model used for these experiments was the same as before: pre-trained on the full dataset without using any labels. This setup is designed (for future applications) to leverage large amounts of unlabeled data for pre-training and then fine-tuning the model on smaller labeled datasets.

Results. The results in Table 4 demonstrate that CLIP pre-training improves model performance when labeled data is limited. For example, at the 25% data split, CLIP increased the accuracy of the spectra model by **4.14%** (from 63.73% to 67.87%), and by **12.56%** at the 10% data split (from 46.68% to 59.24%). Photometry shows a similar trend, with accuracy increasing by **5.21%** at the 25% data split (from 83.22% to 88.43%), and by **7.65%** at the 10% split (from 83.07% to 90.72%). For metadata and all modalities combined, although the difference in accuracy between models with and without CLIP pre-training was not statistically significant, CLIP models generally performed better. These findings suggest that CLIP is beneficial, especially when labeled training data is limited, making it an effective approach for leveraging large unlabeled datasets in future work.

6 Conclusion

We present the curation of a large labeled dataset suitable for building and testing next-generation multi-modal self-supervised models. This includes 21,440 objects with time-series photometry, spectra, and metadata. We also introduce self-supervised pre-training framework that leverages all three data modalities. By extending the Contrastive Language-Image Pretraining model to handle a trimodal setting, our approach effectively learns joint representations across diverse astronomical data types, enhances classification accuracy, and leverages unlabeled data to improve performance when labeled data is limited.

Data Type	Pre-train	50%	25%	10%
Spectra	No CLIP	68.072 ± 1.759	63.729 ± 1.637	46.677 ± 3.486
	CLIP	71.609 ± 1.814	$\mathbf{67.869 \pm 1.303}$	$\mathbf{59.235 \pm 1.399}$
Photometry	No CLIP	89.177 ± 0.518	83.218 ± 2.709	83.073 ± 1.762
	CLIP	90.272 ± 0.695	$\mathbf{88.434 \pm 0.781}$	$\mathbf{90.720 \pm 1.359}$
Metadata	No CLIP	82.035 ± 1.452	79.649 ± 1.148	76.524 ± 1.309
	CLIP	83.830 ± 1.083	81.953 ± 1.492	79.073 ± 1.711
All	No CLIP	91.870 ± 0.470	90.741 ± 1.053	88.264 ± 2.188
	CLIP	91.978 ± 0.746	92.073 ± 1.066	90.628 ± 1.509

Table 4: Accuracy comparison across data splits (50%, 25%, 10%) with and without CLIP pre-training for different data types (Spectra, Photometry, Metadata, All). Statistically significant improvements in bold.

Future Work. Given the abundance of photometry and metadata compared to spectra, one key area is to develop an algorithm capable of handling missing modalities *during training*, allowing us to leverage all available photometry and metadata. Additional directions include expanding the framework to integrate even more modalities, such as photometry from other bands and human comments on sources; learning to manage varying and missing metadata; and incorporating new classes, including non-periodic ones. Building a larger, more diverse dataset and applying the models to tasks like prediction and anomaly detection are essential next steps toward creating a truly foundational multimodal model for astronomy.

References

Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp 2623–2631

Alabdulmohsin I., Wang X., Steiner A., Goyal P., D'Amour A., Zhai X., 2024, arXiv preprint arXiv:2403.04547

Becker I., Pichara K., Catelan M., Protopapas P., Aguirre C., Nikzat F., 2020, MNRAS, 493, 2981

Boone K., 2021, AJ, 162, 275

Cherti M., et al., 2023, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 2818–2829

Cui X.-Q., et al., 2012, Research in Astronomy and Astrophysics, 12, 1197

Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, AAP, 475, 1159

Donoso-Oliva C., Becker I., Protopapas P., Cabrera-Vives G., Vishnu M., Vardhan H., 2023, AAP, 670, A54

Dubath P., et al., 2011, MNRAS, 414, 2602

Gaia Collaboration Brown A. G. A., et al., 2021, AAP, 649, A1

Guzhov A., Raue F., Hees J., Dengel A., 2022, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 976–980

Hayat M. A., Stein G., Harrington P., Lukić Z., Mustafa M., 2021, ApJ Letters, 911, L33

Jamal S., Bloom J. S., 2020, ApJ Supp., 250, 30

Jayasinghe T., et al., 2019, MNRAS, 486, 1907

Kim D.-W., Yeo D., Bailer-Jones C. A. L., Lee G., 2021, AAP, 653, A22

Lafler J., Kinman T. D., 1965, ApJS, 11, 216

Leung H. W., Bovy J., 2024, MNRAS, 527, 1494

Li Y., Liang F., Zhao L., Cui Y., Ouyang W., Shao J., Yu F., Yan J., 2021, arXiv preprint arXiv:2110.05208

Li J., Li D., Xiong C., Hoi S., 2022, in International conference on machine learning. pp 12888–12900

Long J. P., El Karoui N., Rice J. A., Richards J. W., Bloom J. S., 2012, PASP, 124, 280

Luo H., Ji L., Zhong M., Chen Y., Lei W., Duan N., Li T., 2021, arXiv preprint arXiv:2104.08860

Ma Y., Xu G., Sun X., Yan M., Zhang J., Ji R., 2022, in Proceedings of the 30th ACM International Conference on Multimedia. pp 638–647

Morrissey P., et al., 2007, ApJS, 173, 682

Mu N., Kirillov A., Wagner D., Xie S., 2022, in European conference on computer vision. pp 529–544

Muthukrishna D., Narayan G., Mandel K. S., Biswas R., Hložek R., 2019, PASP, 131, 118002

Naul B., Bloom J. S., Pérez F., van der Walt S., 2018, Nature Astronomy, 2, 151

Palaversa L., et al., 2013, AJ, 146, 101

Parker L., et al., 2024, Monthly Notices of the Royal Astronomical Society, 531, 4990–5011

Radford A., et al., 2021, in International conference on machine learning. pp 8748–8763

Richards J. W., et al., 2011, ApJ, 733, 10

Richards J. W., Starr D. L., Miller A. A., Bloom J. S., Butler N. R., Brink H., Crellin-Quick A., 2012, ApJ Supp., 203, 32

Shappee B. J., et al., 2014, ApJ, 788, 48

Singh A., Hu R., Goswami V., Couairon G., Galuba W., Rohrbach M., Kiela D., 2022, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 15638–15650

Skrutskie M. F., et al., 2006, AJ, 131, 1163

Sun Q., Fang Y., Wu L., Wang X., Cao Y., 2023, arXiv preprint arXiv:2303.15389

Wright E. L., et al., 2010, AJ, 140, 1868

Wu Y., Chen K., Zhang T., Hui Y., Berg-Kirkpatrick T., Dubnov S., 2023, in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 1–5

Wu Y., Tao Y., Fan D., Cui C., Zhang Y., 2024a, MNRAS, 527, 1163

Wu Y., Tao Y., Fan D., Cui C., Zhang Y., 2024b, Monthly Notices of the Royal Astronomical Society, 527, 1163

Xu H., Ghosh G., Huang P.-Y., Okhonko D., Aghajanyan A., Metze F., Zettlemoyer L., Feichtenhofer C., 2021, arXiv preprint arXiv:2109.14084

Xu H., et al., 2023, arXiv preprint arXiv:2309.16671

Yao L., et al., 2021, arXiv preprint arXiv:2111.07783

Zhang R., et al., 2022, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 8552–8562

Zhang G., Helfer T., Gagliano A. T., Mishra-Sharma S., Villar V. A., 2024, arXiv e-prints, p. arXiv:2408.16829

Zhou H., Zhang S., Peng J., Zhang S., Li J., Xiong H., Zhang W., 2021, in Proceedings of the AAAI conference on artificial intelligence. pp 11106–11115

A Training Setup and Hyperparameters

In this work, we used Optuna (Akiba et al., 2019) to perform hyperparameter optimization for our models. Our goal was to minimize the validation loss across multiple architectures and pre-training strategies. We tuned CLIP itself, as well as models for photometry, spectra, metadata, and multimodal data, with two initialization options: random initialization or pre-trained CLIP weights.

For each model type, the hyperparameters we explored included:

- Learning rate (`lr`): Sampled from a logarithmic scale between 1×10^{-5} and 1×10^{-2}
- Dropout rates for photometry (`p_dropout`), spectra (`s_dropout`) and metadata (`m_dropout`): All sampled from a uniform distribution between 0.0 and 0.4.
- Adam optimizer parameters:
  - Beta1 (`beta1`): Sampled from a uniform distribution between 0.7 and 0.99.
  - Weight decay (`weight_decay`): Sampled from a logarithmic scale between 1×10^{-5} and 1×10^{-1}.
- Learning rate scheduler factor (`factor`): Sampled from a uniform distribution between 0.1 and 1.0 for the `ReduceLROnPlateau` scheduler.

Training Setup. For each trial, additional techniques were applied to ensure model stability and improve convergence:

- Gradient clipping was applied to stabilize training. For CLIP, a clipping value of 45 was used, while for the photometry and spectra models, the clipping value was set to 5.
- Training duration: The models were trained for a fixed number of epochs: 100 epochs for CLIP and 50 epoch for others
- A warmup scheduler was employed to gradually increase the learning rate from a very low value to the target learning rate over the first 10 epochs.
- Early stopping based on validation loss was used with a patience of 6 epochs.

Feature	Description
mean_vmag	Mean magnitude in the visible band
phot_g_mean_mag	Gaia G-band mean magnitude
e_phot_g_mean_mag	Uncertainty in Gaia G-band mean magnitude
phot_bp_mean_mag	Gaia BP band mean magnitude
e_phot_bp_mean_mag	Uncertainty in Gaia BP band mean magnitude
phot_rp_mean_mag	Gaia RP band mean magnitude
e_phot_rp_mean_mag	Uncertainty in Gaia RP band mean magnitude
bp_rp	BP mean magnitude minus RP mean magnitude
parallax	Gaia DR3 Parallax measurement
parallax_error	Uncertainty in parallax measurement
parallax_over_error	Signal-to-noise ratio for parallax measurement
pmra	Proper motion in the Right Ascension direction
pmra_error	Uncertainty in pmra
pmdec	Proper motion in the Declination direction
pmdec_error	Uncertainty in pmdec
j_mag	2MASS J-band magnitude
e_j_mag	Uncertainty in 2MASS J-band magnitude
h_mag	2MASS H-band magnitude
e_h_mag	Uncertainty in 2MASS H-band magnitude
k_mag	2MASS K-band magnitude
e_k_mag	Uncertainty in 2MASS K-band magnitude
w1_mag	WISE W1 band magnitude
e_w1_mag	Uncertainty in WISE W1 band magnitude
w2_mag	WISE W2 band magnitude
e_w2_mag	Uncertainty in WISE W2 band magnitude
w3_mag	WISE W3 band magnitude
w4_mag	WISE W4 band magnitude
j_k	J-band minus K-band magnitude
w1_w2	W1 band minus W2 band magnitude
w3_w4	W3 band minus W4 band magnitude
pm	Total proper motion
ruwe	Renormalized unit weight error
l	Galactic longitude
b	Galactic latitude

Table 5: Descriptions of metadata features used in the dataset.