HubbleCLIP: Associating Astronomical Observations and Natural Language with Multi-Modal Models

Siddharth Mishra-Sharma *smsharma@mit.edu*
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Department of Physics, Harvard University, Cambridge, MA 02138, USA

Yiding Song *ydsong@mit.edu*
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Jesse Thaler *jthaler@mit.edu*
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

We present a multi-modal model which associates astronomical observations imaged by the *Hubble* Space Telescope (HST) with natural language. The model is fine-tuned from a base CLIP model using successful proposal abstracts corresponding to HST observations, summarized via guided large language (LLM) model generation. We show that the model embodies a meaningful joint representation between observations and text through experiments targeting observation retrieval (i.e., retrieving the most relevant observations from a set using natural language queries) and description retrieval (i.e., querying the astrophysical object classes and science use cases most relevant to a given observation). The model demonstrates the potential for using generalist rather than task-specific models for astrophysics research, in particular by leveraging text as an interface.

1 Introduction

Machine learning (ML) is starting to have a significant impact on the sciences, with astrophysics being no exception. Machine learning methods have demonstrated promise when applied to every part of the astrophysics pipeline, from instrument design, to data acquisition, to its analysis. Until recently, most applications of ML within astrophysics have focused on replacing or augmenting traditional techniques with ML counterparts in order to improve performance on specific tasks.

The *Foundation Model* paradigm, in contrast, seeks to develop generalist models which can be deployed on a wide range of tasks. The paradigm has been highly successful in domains like computer vision and natural language processing, as demonstrated by the widespread adoption of tools like CLIP, ChatGPT, Dall-E, and Stable Diffusion. These models are typically pre-trained on massive amounts of unlabeled data using self-supervised learning techniques, enabling them to learn powerful representations which can be optionally fine-tuned to address domain-specific tasks. At the heart of the paradigm lies the triumph of scale – scaling up model size, dataset size, and compute. However, foundation models often benefit from fine tuning using a relatively small amounts of domain-specific data, increasing their usefulness when applied to those specific domains.

There is considerable interest in developing custom foundation models for the sciences, with astrophysics being ripe for such an effort due to several reasons. The first is the availability of large amounts of publicly-available data as a contingency of publicy-funded data-taking efforts. The second is the multi-modality

inherent astrophysical observations, with different types of data (e.g., images, spectra, light curves, textual descriptions) often available for each observation. This multi-modality was recently exploited to train AstroCLIP (Lanusse et al., 2023) – a joint representation between multi-band images and optical spectra from the Dark Energy Spectroscopic Instrument (DESI). AstroLLaMA (Nguyen et al., 2023) is another recent effort to fine-tune a publicly-available model (Llama-2) on astrophysics-specific textual data from the arXiv.

The CLIP (Radford et al., 2021) family of models has shown strong performance on a variety of downstream tasks including zero-shot classification and image retrieval.

GeoCLIP (Cepeda et al., 2023).

The success of the foundation model paradigm partly relies on the ability to flexibly leverage text as a *universal interface*. In this work, we take this outlook and train a joint representation between observations taken by the *Hubble* Space Telescope (HST) and natural language. We do so by using the associations between observation proposals and corresponding downstream observations. We show that fine-tuning a CLIP (Contrastive Language-Image Pre-training) model on this data enables learning meaningful joint representations.

The paper is organized as follows. In Sec. 2, we describe the dataset used in this work, including its curation and processing. In Sec. 3, we describe the methodology used to train and evaluate the model. In Sec. 4, we present the results of our experiments on image and text retrieval tasks. We discuss future prospects and conclude in Sec. 5.

2 Dataset Construction

We curate a dataset of images of *Hubble* observations and corresponding text descriptions. We rely on summarized versions of proposal abstracts from the Proposal Abstracts Catalog[1] – a catalog of all accepted *Hubble* proposals – to derive captions for the observations. The HST has been operational for 33 years, having been launched on April 24, 1990. We use available proposals and observations up to Cycle 30, which commenced data-taking in 2022.

Examples of images and corresponding captions are shown in Tab. 1. It can be seen that these images have specific characteristics as well as artifacts particular to the nature of data-taking which distinguish them from the distribution of natural images typically used for large-scale pre-training of foundation models. This further motivates the need for fine-tuning on domain-specific data.

2.1 Data Selection and Pre-processing

Observations corresponding to individual proposal IDs are queried through the Mikulski Archive for Space Telescopes (MAST) via the `astroquery` interface. Products of type `PREVIEW` are filtered in, corresponding to preview postcard images. Note that these are not science-grade observations, but rather low-resolution images used for quick-look purposes; given the nature of associations we aim to learn, we deem this adequate for our current purposes. A maximum of 20 images are downloaded per proposal ID, selected at random, in order to avoid biasing the model towards proposals with a larger number of observations. Images are centered and resized to a resolution per side of 512 pixels before saving. Color previews (i.e., observations taken with multiple wavelength filters assigned to individual RGB channels) are manually excluded via a filename filter in order to maintain consistency across the dataset; models trained on datasets with color images included were observed to show worse performance on generalization metrics. If no appropriate images corresponding to an abstract are found, the abstract is excluded from the dataset.

In total 31,859 images corresponding to 4,438 abstracts are included in the fine-tuning dataset. 3,194 images are held out for validation, with no abstract being common between training and validation sets in order to ensure an independent set of text-image pairs for testing.

[1] https://archive.stsci.edu/hst/proposal_abstracts.html

2.2 Summarization via Guided Generation

Raw abstracts summarize the corresponding successful HST observing proposals, which intend to make the case for allocating *Hubble* telescope time towards a particular set of observations. These abstracts are written in a diversity of styles, formats, and lengths, being highly variable in the nature of content as well. Although the abstracts can be used as-is as image captions, we explore the use of summarization via guided LLM generation to standardize the captions used for fine-tuning the CLIP model. The goal is to summarize the objects and phenomena, as well as potential downstream science use cases corresponding to the HST observations in order to increase the signal between text and images.

The method from Willard & Louf (2023) is used to produce an LLM-generated summary of the abstract conforming to a particular schema, specified in JSON format. The schema is designed to represent a list of the objects (e.g., 'Type Ia supernova') and phenomena (e.g., 'gravitational lensing'), as well as potential downstream science uses cases (e.g., 'set constraints on supernova explosion models') that could correspond to the eventual imaged observation given the abstract text.

The procedure guides the generation of LLM outputs while ensuring that the schema is respected at every point in the generation by masking out tokens that would violate the intended format. By framing the problem in terms of transitions between a set of finite-state machines, Willard & Louf (2023) showed that guided generation can be performed with negligible overhead compared to unconstrained generation. This ensures that the output of the LLM strictly conforms to the format of the following example:

```
{
  'objects_and_phenomena': ['star forming galaxy', 'lensed galaxy', ...],
  'science_use_cases': ['measure lensing magnification', 'probe spectral energy
    distributions', ...]
}
```

which is then used to construct the summarized caption by combining the two key elements. Examples of raw abstracts and corresponding LLM-generated summaries are shown in Tab. 2. Further details on the summarization procedure, including the prompts used and a more detailed description of guided generation, are provided in App. A.

The open-weights, instruction-tuned model Mixtral-8x7B[2] is used to generate the summaries, with guided generation performed using the `outlines`[3] package.

The goal if summarization-via-guided-generation is to increase the signal between text and images by standardizing the captions used for fine-tuning the CLIP model.

[2] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[3] https://github.com/outlines-dev/outlines

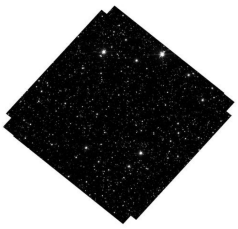
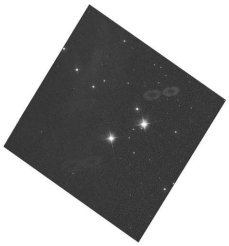
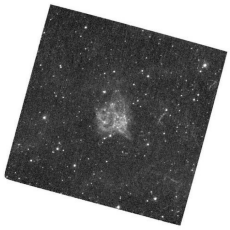
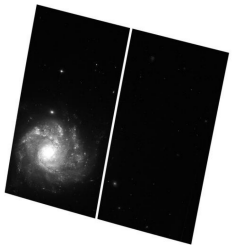
Image	Obs. cycle	Prop. ID	LLM-extracted summary
	26	15513	isolated black holes, background stars, Galactic bulge; constrain mass of isolated black holes, distinguish between black hole scenarios, analyze relative proper motions of stars
	19	12577	Cas A supernova, light echoes, interstellar dust, supernova outburst, shock breakout; Estimate radius of Cas A progenitor star, connect progenitor star to explosion to supernova to supernova remnant (SNR), analyze evolution of Cas A's spectrum over time, determine maximum-light characteristics of the supernova, probe properties of cooling envelope after shock breakout
	7	7340	young oxygen-rich supernova remnants SN0540–69.3, LMC, SMC, supernova debris, active pulsar, synchrotron nebula; characterize ionization structure and distribution of chemically peculiar debris in SN0540–69.3, determine ionization structure in the SN debris of E0102.2–7219, provide benchmarks for models of nucleosynthesis in massive stars, excitation mechanisms in extremely metal-rich plasmas, and supernova explosion dynamics, study the pulsar and synchrotron nebula in SN0540–69.3, investigate SN0540–69.3's proximity to SN 1987A in both space and time, and relation to the same extended complex of young stars
	22	13757	type Iax supernovae, white dwarfs, possible companion stars, accretion disks, luminous blue stars; constrain progenitor systems of type Iax supernovae, distinguish between explosion mechanisms, investigate mass transfer processes in accretion disks, determine if type Iax supernovae originate from massive stars

Table 1: Examples of images and corresponding captions, constructed using the LLM-extracted summaries. The CLIP model is fine-tuned on these text-image associations.

Prop. ID	Proposal abstract	LLM-extracted summary	
		Objects and phenomena	Science use cases
15513	Category: Stellar Physics. A significant fraction of the mass of an old stellar population should be in the form of isolated black holes (BHs). Yet there has never been an unambiguous detection of a solitary BH. The only technique available to detect isolated BHs is astrometric microlensing–relativistic deflection of light from background stars. We have...	isolated black holes, background stars, Galactic bulge	constrain mass of isolated black holes, distinguish between black hole scenarios, analyze relative proper motions of stars
12577	Category: ISM AND CIRCUMSTELLAR MATTER. We propose to obtain time-resolved spectroscopy of the outburst of the enigmatic historical supernova Cas A using STIS spectroscopy of light scattered by a narrow filament of interstellar dust. Our group has identified recent, high-surface brightness filaments that are likely to provide high signal-to-noise reproduction of the evolving spectrum of...	Cas A supernova, light echoes, interstellar dust, supernova outburst, shock breakout	Estimate radius of Cas A progenitor star, connect progenitor star to explosion to supernova to supernova remnant (SNR), analyze evolution of Cas A's spectrum over time, determine maximum-light characteristics of the supernova, probe properties of cooling envelope after shock breakout
7340	Category: STELLAR EJECTA. We propose to use the WFPC2 and STIS CCD to obtain maximum spatial resolution emission-line images of the young, oxygen- rich supernova remnants SN0540–69.3 in the LMC and E0102.2– 7219 in the SMC. O IIILambda5007, S IILambdaLambda6724 and O IILambdaLambda3727 images of SN0540–69.3 will be used to characterize the ionization structure and...	young oxygen-rich supernova remnants SN0540–69.3, LMC, SMC, supernova debris, active pulsar, synchrotron nebula	characterize ionization structure and distribution of chemically peculiar debris in SN0540–69.3, determine ionization structure in the SN debris of E0102.2–7219, provide benchmarks for models of nucleosynthesis in massive stars, excitation mechanisms in extremely metal-rich plasmas, and supernova explosion dynamics, study the pulsar and synchrotron nebula in SN0540–69.3, investigate SN0540–69.3's proximity to SN 1987A in both space and time, and relation to the same extended complex of young stars
13757	Category: HOT STARS. Type Ia supernovae (SN Ia) have enormous importance to cosmology and astrophysics, but their progenitors and explosion mechanisms are not known in detail. Recently, observations and theoretical models have suggested that not all thermonuclear white-dwarf supernova explosions are normal SN Ia. In particular, type Iax supernovae (peculiar cousins to SN Ia), are...	type Iax supernovae, white dwarfs, possible companion stars, accretion disks, luminous blue stars	constrain progenitor systems of type Iax supernovae, distinguish between explosion mechanisms, investigate mass transfer processes in accretion disks, determine if type Iax supernovae originate from massive stars

Table 2: Examples of the initial parts of raw proposal abstracts (second column) and LLM (Mixtral-8x7B)-extracted summaries (rightmost two columns), separately extracting objects and phenomena as well as potential downstream science use cases. The LLM-extracted summaries are used for associating text with observations.

3 Methodology

Our goal is to learn a semantically meaningful joint representation of HST image observations and natural language, with the intention of using it for a variety of downstream tasks. We leverage the strong generalization capabilities demonstrated by CLIP (Contrastive Language-Image Pretraining) and adapt these to work with domain-specific *Hubble* data via fine-tuning; we describe these below.

3.1 Language-Image Pre-training

CLIP (Contrastive Language-Image Pretraining; Radford et al., 2021) is a multi-modal model pre-trained on a large corpus of image-text pairs via weak supervision using a contrastive loss. Given a minibatch $\mathcal{B}$ of $|\mathcal{B}|$ image-text pairs $\{(I_i, T_i)\}$, the goal is to align the learned representations of corresponding (positive) pairs (I_i, T_i) while repelling the representations of unaligned (negative) pairs $(I_i, T_{j\neq i})$. Image and text encoders $f(\cdot)$ and $g(\cdot)$ are used to map images and text to a common embedding space. The standard softmax-based bidirectional variation of the InfoNCE (Oord et al., 2018) contrastive loss function, as used by CLIP, is particularly effective for multimodal learning. This bidirectionality is crucial as it ensures the model learns to map both images to text and text to images with equal importance. This symmetry in learning is essential for tasks that require a mutual understanding and interchangeability between visual and textual representations, such as image captioning, text-to-image synthesis, and cross-modal retrieval. The bidirectional loss is given by (Radford et al., 2021)

$$\mathcal{L}(\mathcal{B}) = -\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{x_i\cdot y_i/\tau}}{\sum_{j=1}^{|\mathcal{B}|}e^{x_i\cdot y_j/\tau}}}^{\text{image}\,\rightarrow\,\text{text softmax}} + \overbrace{\log\frac{e^{x_i\cdot y_i/\tau}}{\sum_{j=1}^{|\mathcal{B}|}e^{x_j\cdot y_i/\tau}}}^{\text{text}\,\rightarrow\,\text{image softmax}}\right) \tag{1}$$

where $x_i = \frac{f(I_i)}{\|f(I_i)\|_2}$ and $y_i = \frac{g(T_i)}{\|g(T_i)\|_2}$ are the normalized representations of the i-th image and text, respectively, and τ is a (typically learnable) temperature hyperparameter. This bidirectional approach not only enhances the model's ability to create robust and versatile embeddings but also aligns with the intuitive understanding of how humans interact with and interpret multimodal information, further bridging the gap between artificial and natural intelligence in multimodal contexts.

As the base model, we use the CLIP-ViT-B/16 (Radford et al., 2021) variant trained by OpenAI. This model uses a (12-layer, 12-head, 768-embedding dimension) vision transformer as the image encoder and a (12-layer, 8-head, 512-embedding dimension) sequence transformer as the text backbone. The text encoder has a maximum length of 77-tokens and the image encoder a native resolution of 224×224. Linear projection layers map the outputs of the image and text encoders to a common embedding space of dimension 512. In total, the model has 149,620,737 trainable parameters. The model was trained on 400 million image-text pairs from the internet.

3.2 Fine-tuning Procedure and Evaluation

The base CLIP model is fine-tuned using the dataset described in Sec. 2, using either the LLM-summarized data or the raw proposal abstracts. When using raw proposal abstracts, random chunks of the text delimited by periods are selected on the fly to fit within the maximum token length of the text encoder. Images are randomly cropped to the native resolution of the image encoder and randomly rotated at each training step. Given the relatively modest size of the fine-tuning set, a batch size $|\mathcal{B}| = 32$ is used throughout. We note that the positive and negative image-text association is noisy and imperfect, since multiple images can be associated with the same abstract.

We explore three different methods of training the model on our domain dataset: *(1)* Fine-tuning the entire network, starting from the pre-trained base model; *(2)* Freezing the base image and text encoders, and training a small projection head on top of these; and *(3)* Training the entire model from scratch.

The model is trained over 50,000 steps with 5000 linear warmup steps and cosine decay using the AdamW optimizer (Loshchilov & Hutter, 2019; Kingma & Ba, 2015) with peak learning rate of either 10^{-5} or 10^{-6} and weight decay 10^{-3}. Training takes approximately 6 hours on 4 Nvidia A100 GPUs.

4 Results and Discussion

4.1 Fine-tuned Retrieval Accuracy

4.2 'Zero-shot' Hypothesis and Object Retrieval

4.3 Text-to-Image Retrieval

5 Outlook and Conclusions

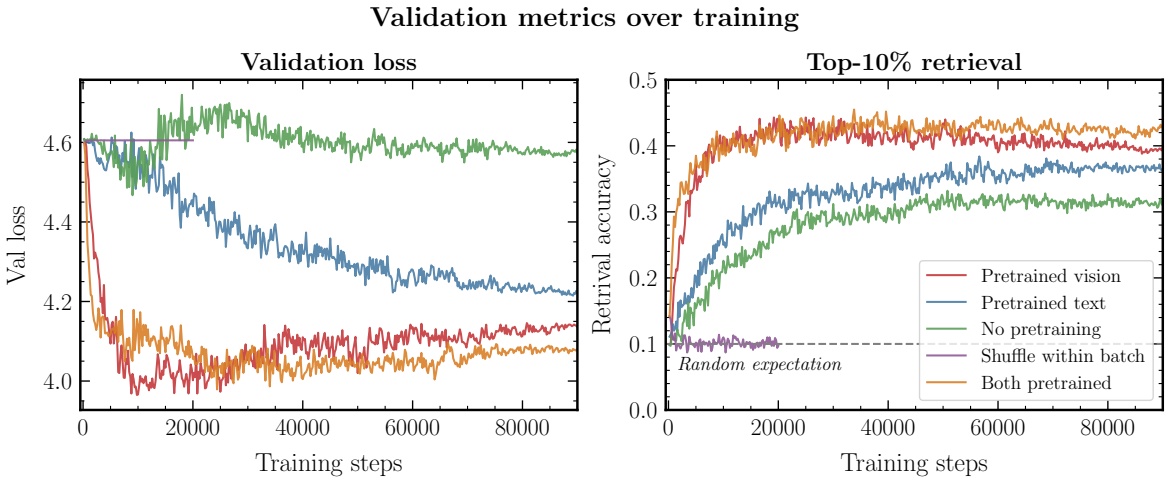


Figure 1: Retrieval accuracy

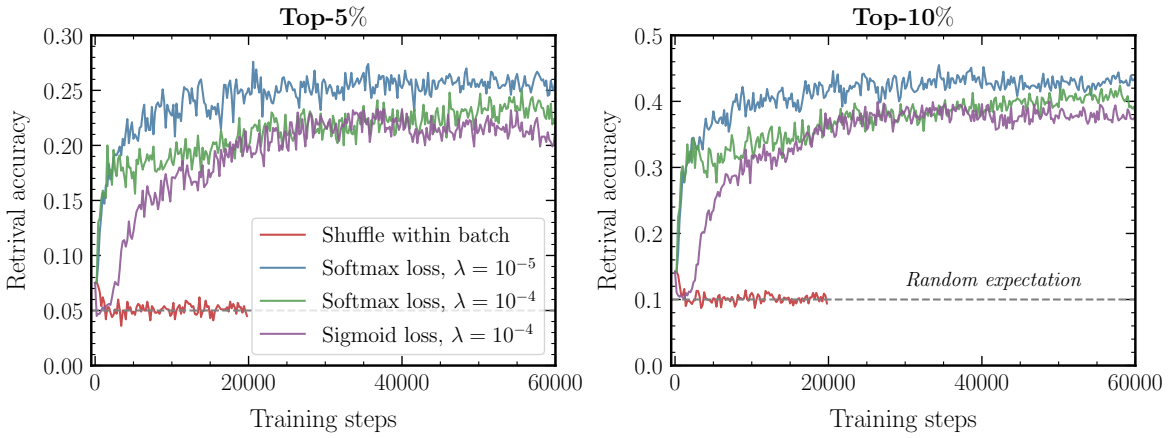


Figure 2: Retrieval accuracy

Image-to-Text Retrieval

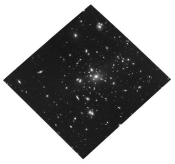
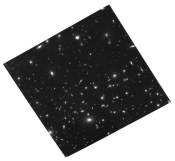
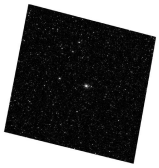
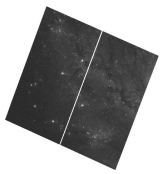
Image	Top classes (fine-tuned)	Top classes (base)	Abstract
	1. dark matter 2. Einstein rings 3. galaxy mergers 4. gravitational lensing 5. galaxy clusters	1. galaxy clusters 2. ultra diffuse galaxies 3. dwarf galaxies 4. gravitational lensing 5. crowded stellar field	Category: COSMOLOGY. We propose to study the physical nature of dark matter by using massive, merging clusters of galaxies. As shown with the Bullet Cluster (1E0657-56), such massive well-measured systems are critical for our understanding of dark matter. By more than doubling the number of clusters in the sample and obtaining systems at different observation angles, impact parameters, geometrical arrangements, and merger velocities, the systematic uncertainties in the dark matter cross section calculations can be improved substantially, allowing us to move from rough order of magnitude estimates to measurements with quantifiable uncertainties that can be compared usefully with the predictions from numerical simulations, and the constraints on alternate gravity models become unambiguous. Our proposed targets are three extraordinary, merging galaxy clusters with X-ray and optical offsets that are placed at ideal redshifts for such a study: A520, A1758N, and A2163. To pin down the position of the dark matter component we require high resolution, absolutely calibrated mass maps. High resolution gravitational lensing data is needed to attain this goal, which can only be achieved with the excellent resolving power of the HST.
	1. dark matter 2. galaxy mergers 3. Einstein rings 4. gravitational lensing 5. dark energy	1. ultra diffuse galaxies 2. galaxy clusters 3. gravitational lensing 4. high-redshift quasars 5. dwarf galaxies	Category: COSMOLOGY. We propose to study the physical nature of dark matter by using massive, merging clusters of galaxies. As shown with the Bullet Cluster (1E0657-56), such massive well-measured systems are critical for our understanding of dark matter. By more than doubling the number of clusters in the sample and obtaining systems at different observation angles, impact parameters, geometrical arrangements, and merger velocities, the systematic uncertainties in the dark matter cross section calculations can be improved substantially, allowing us to move from rough order of magnitude estimates to measurements with quantifiable uncertainties that can be compared usefully with the predictions from numerical simulations, and the constraints on alternate gravity models become unambiguous. Our proposed targets are three extraordinary, merging galaxy clusters with X-ray and optical offsets that are placed at ideal redshifts for such a study: A520, A1758N, and A2163. To pin down the position of the dark matter component we require high resolution, absolutely calibrated mass maps. High resolution gravitational lensing data is needed to attain this goal, which can only be achieved with the excellent resolving power of the HST.
	1. crowded stellar field 2. supernova remnants 3. compact stellar remnants 4. primordial black holes 5. pre-main sequence stars	1. stellar abundances 2. stellar populations 3. interstellar medium 4. pre-main sequence stars 5. Cepheid variables	Category: RESOLVED STELLAR POPULATIONS. Exploiting the full power of the Wide Field Camera 3 {WFC3}, we propose deep panchromatic imaging of four fields in the Galactic bulge. These data will enable a sensitive dissection of its stellar populations, using a new set of reddening-free photometric indices we have constructed from broad-band filters across UV, optical, and near-IR wavelengths. These indices will provide accurate temperatures and metallicities for hundreds of thousands of individual bulge stars. Proper motions of these stars derived from multi-epoch observations will allow separation of pure bulge samples from foreground disk contamination. Our catalogs of proper motions and panchromatic photometry will support a wide range of bulge studies. Using these photometric and astrometric tools, we will reconstruct the detailed star-formation history as a function of positon within the bulge, and thus differentiate between rapid- and extended-formation scenarios. We will also measure the dependence of the stellar mass function on metallicity, revealing how the characteristic mass of star formation varies with chemistry. Our sample of bulge stars with accurate metallicities will include 12 candidate hosts of extrasolar planets. Planet frequency is correlated with metallicity in the solar neighborhood; our measurements will extend this knowledge to a remote environment with a very distinct chemistry. Our proposal also includes observations of six well-studied globular and open star clusters; these observations will serve to calibrate our photometric indices, provide empirical population templates, and transform the theoretical isochrone libraries into the WFC3 filter system. Besides enabling our own program, these products will provide powerful new tools for a host of other stellar-population investigations with HST/WFC3. We will deliver all of the products from this Treasury Program to the community in a timely fashion.
	1. low surface brightness galaxies 2. star formation histories 3. galaxy formation 4. ultra diffuse galaxies 5. circumgalactic medium	1. gravitational lensing 2. high-redshift quasars 3. brown dwarfs 4. trans-Neptunian objects 5. Kuiper Belt objects	Category: Stellar Populations and the Interstellar Medium. Observations of the ultra-faint dwarfs (UFDs), as relics of the epoch of reionization, allow us to probe the earliest epochs of star formation (SF). In particular, the UFDs in low density environments for most of their lifetimes provide unique tools to probe the effects of early environmental conditions on the SF histories (SFHs) of the UFDs and their sub-solar initial mass function (IMF) because they likely maintain the best \' fossil' record of early local environments. We propose to obtain deep ACS and UVIS imaging in F606W and F814W for 2 LMC satellite UFDs that are on their first approach to our Galaxy, and thus resided in the outskirts of the Local Group at high redshift. This program is designed to identify systematic differences in the stellar populations of recently captured UFDs vs. long-term MW satellites (data available from previous programs) by using high-fidelity color-magnitude diagrams constructed from deep HST imaging as well as spectroscopically measured metallicity distribution functions. We will: (1) Establish whether SF is quenched at different times with different rate in UFDs in low density environment at early times, probing the patchiness of reionization by directly comparing with theoretical predictions; (2) Identify variations in the sub-Solar IMF across UFDs born in different environments; (3) Pave the way for a more accurate constraint on the MW halo mass.
	1. supernovae 2. Cepheid variables 3. star clusters 4. star forming galaxies 5. dust	1. high-redshift quasars 2. gravitational lensing 3. Kuiper Belt objects 4. compact stellar remnants 5. ultra diffuse galaxies	Category: RESOLVED STELLAR POPULATIONS. We propose to test two of the clearest predictions of the theory of evolution of massive-star evolution: 1) The formation of Wolf-Rayet stars depends strongly on these stars' metallicity {Z}, with relatively fewer WR stars forming at lower Z, and 2) Wolf-Rayet stars die as Type Ib or Ic supernovae. To carry out these tests we propose a deep, narrowband imaging survey of the massive star populations in the Scl spiral galaxy M101. Just as important, we will test the hypothesis that Superclusters like 30 Doradus are always richly populated with WR stars, and by implication that these complexes are responsible for the spectral signatures of starburst galaxies. Our previous HST survey of the HII regions in the SclII galaxy NGC 2403 suggested that the distribution of WR stars and RSG is a sensitive diagnostic of the recent star-forming history of these large complexes: young cores of O and WR stars are surrounded by older halos containing RSG. Theory predicts that this must change with metallicity; relatively fewer WR stars form at lower Z. A key goal of our proposal is to directly test this paradigm in a single galaxy, M101 being the ideal target. The abundance gradient across M101 (a factor of 20) suggests that relatively many more WR will be found in the inner parts of this galaxy than in the outer "suburbs". Second, we note that WR stars are predicted to end their lives as core-collapse or pair-instability supernovae. The WR population in M101 may be abundant enough for one to erupt as a Type Ib or Ic supernova within a generation. The clear a priori identification of a WR progenitor would be a major legacy of HST. Third, we will also determine if "superclusters", heavily populated by WR stars, are common in M101. It is widely claimed that such Superclusters produce the integrated spectral signatures of Starburst galaxies. We will be able to directly measure the numbers and emission-line luminosities of thousands of Wolf Rayet stars located in hundreds of M101 Superclusters, and correlate those numbers against the Supercluster sizes and luminosities. It is likely (but far from certain) that Supercluster sizes and emission-line luminosities are driven by their Wolf-Rayet star content. Our sample will be the largest and best-ever Supercluster/Wolf Rayet sample, an excellent local proxy for characterizing starburst galaxies' Superclusters.

Figure 3: Retrieval accuracy

Code

Acknowledgments

This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under grant Contract Number DE-SC0012567. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

This research is based on observations made with the NASA/ESA Hubble Space Telescope obtained from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555.

Based on observations made with the NASA/ESA Hubble Space Telescope, and obtained from the Hubble Legacy Archive, which is a collaboration between the Space Telescope Science Institute (STScI/NASA), the

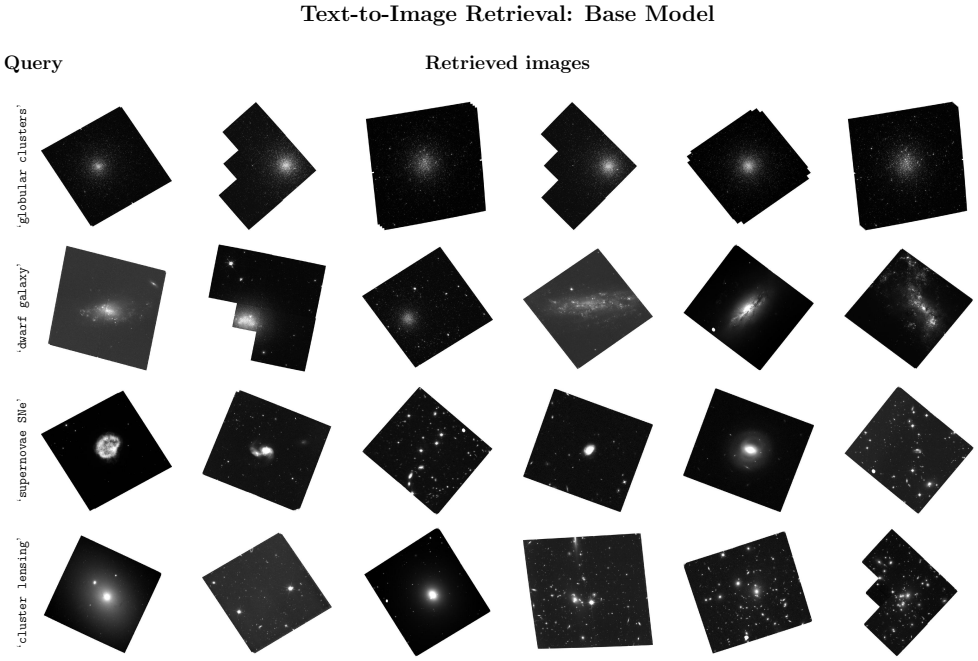


Figure 4: Retrieval accuracy

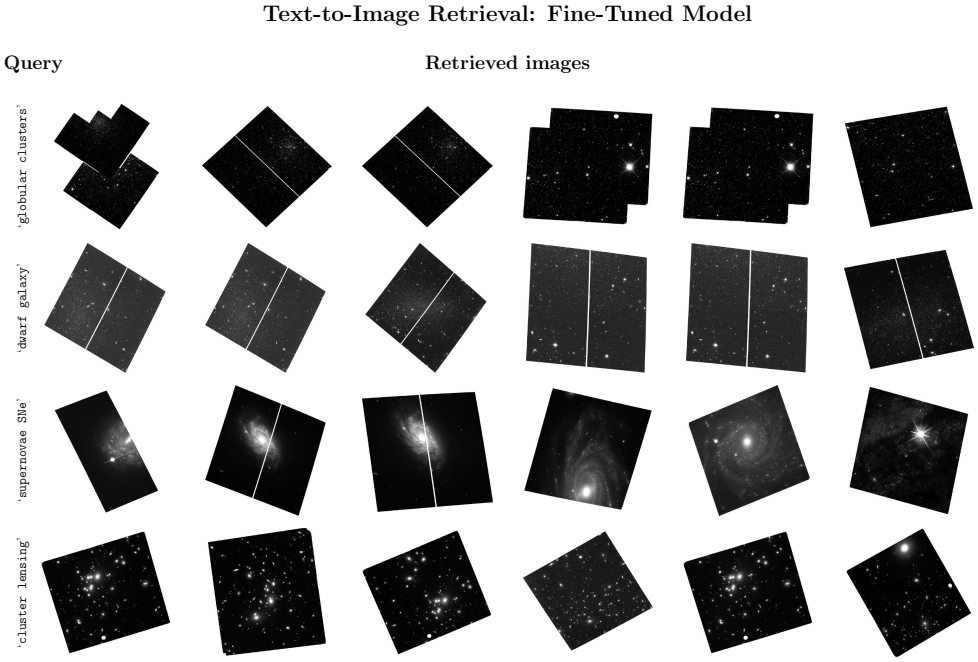


Figure 5: Retrieval accuracy

Space Telescope European Coordinating Facility (ST-ECF/ESAC/ESA) and the Canadian Astronomy Data Centre (CADC/NRC/CSA).

References

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv preprint arXiv:2309.16020*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Francois Lanusse, Liam Parker, Siavash Golkar, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Ruben Ohana, Mariel Pettee, et al. Astroclip: Cross-modal pre-training for astronomical foundation models. *arXiv preprint arXiv:2310.03024*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, et al. Astrollama: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2309.06126*, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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*, pp. 8748–8763. PMLR, 2021.

Brandon T Willard and Rémi Louf. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*, 2023.

A Summarization via regex-guided generation

The following prompt is used to summarize the abstracts.

```
@outlines.prompt
def prompt_fn(abstract):

      """<s>[INST] You are an expert astrophysicist, with broad expertise across
    observational and theoretical astrophysics. You are able to extract core information
    from astrophysical texts.

Abstract: "{{abstract}}"

Based on the above observational proposal abstract, your task is to summarize the nature of
    the eventual observations. You will identify the astrophysical objects and phenomena, as
     well as the potential science use cases described in the abstract.

Follow these instructions exactly:
- Mention up to 5 items for both categories; do not mention more than 5 items in either
    category.
- Choose the most relevant ones if there are more than 5 items in a category.
- Never mention the Hubble Space Telescope, HST, or the HST archive.
- Mention the class (e.g., barred spiral galaxy) and not just the specific instance (e.g.,
    Andromeda).
```

```
- Name the objects in the science use cases, if appropriate.
- Write out full names of objects in addition to acronyms.
- Do not list irrelevant objects which do not describe the eventual observation, such as
    units or proposal Cycle numbers. List fewer but more relevant objects, if in doubt.
- Each science case listed must be self-contained but succinct.
- Only write in English.
- Do not list items that are too generic (e.g., galaxy, faint object, kinematics)
- The total length of text should not exceed 80 words.
- Present your lists in a comma-separated format; no dashed or numbered lists.

Example output: {'objects_and_phenomena':'spiral galaxies, galaxy clusters, supernova
    remnants', 'science_use_cases':'model galactic structure and evolution, characterize
    dark matter distribution in clusters, analyze expansion rates of supernova remnants'}

Answer in JSON format. The JSON should be a dictionary with keys "objects_and_phenomena" and
     "science_use_cases".

[/INST]
"""
```