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 Contrastive Language–Image Pre-training (CLIP) model using successful proposal abstracts corresponding to HST observations, summarized via guided large language model (LLM) 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 lower-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-Instruct[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

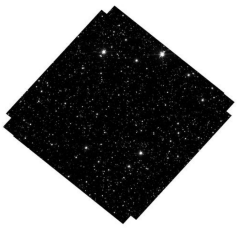
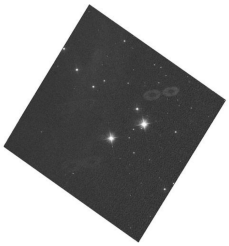
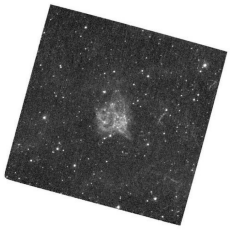
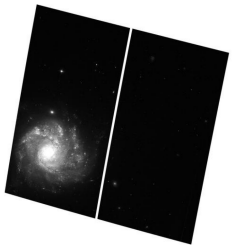
Image	Obs. cycle	Prop. ID	LLM-extracted summary
	26	15513	isolated black holes, background stars, Galactic bulge; constrain mass of isolated black holes, distinguish between black hole scenarios, analyze relative proper motions of stars
	19	12577	Cas A supernova, light echoes, interstellar dust, supernova 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)

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

where $x_i = f(I_i)/\|f(I_i)\|$ and $y_i = g(T_i)/\|g(T_i)\|$ are the normalized representations of the i-th image and text, respectively, and τ is a learnable temperature hyperparameter.

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; larger batch sizes were seen to susceptible to overfitting. 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. For *(2)*, we use a 2-layer MLP with 1024 units and a GELU activation layer, projecting to the 512-dimensional embedding space.

Models were instantiated and trained using the *Jax* (Bradbury et al., 2018) ecosystem. The model is trained over 20,000 steps with 2000 linear warmup steps using the AdamW optimizer (Loshchilov & Hutter, 2019;

Kingma & Ba, 2015) with peak learning rate of 10^{-5} and weight decay 10^{-3}. Training takes approximately 3 hours on 4 Nvidia A100 GPUs.

The model is evaluated by tracking the loss in Eq. 1 as well as the top-k% retrieval accuracy on the held out validation set over the course of training. The retrieval accuracy is defined as the fraction of associated captions which fall within the top k% of captions by cosine similarity of the (normalized) embeddings $x_i \cdot y_j$, averaged over the images in the validation set.

4 Results and Discussion

Validation metrics Fig. 1 shows the contrastive loss (left) and the top-10% retrieval accuracy on the held out validation set over the course of training, for different variations considered. The red lines show the metrics evaluated when training with batches where the image-text associations are randomly shuffled, serving as a baseline. This baseline is seen to do on par with random expectation, unlike the others, validating the presence of a positive image-text association signal in the dataset. Interestingly, the base model performs better than random expectation, with a top-10% retrieval accuracy of $\sim$ 15%. We will therefore compare the qualitative performance of the base model with the fine-tuned model on downstream retrieval tasks.

The fiducial model with LLM-guided summarization (orange lines) is seen to perform significantly better than the model using raw abstracts as captions (purple line), validating the stronger association signal in the summarized dataset we curate. Fine-tuning a small MLP head over frozen vision and text backbones (green lines) and training from scratch (blue lines) show a non-trivial improvement in the retrieval metrics compared to the random baseline as well as base model, but with deteriorated performance compared to the fiducial set-up.

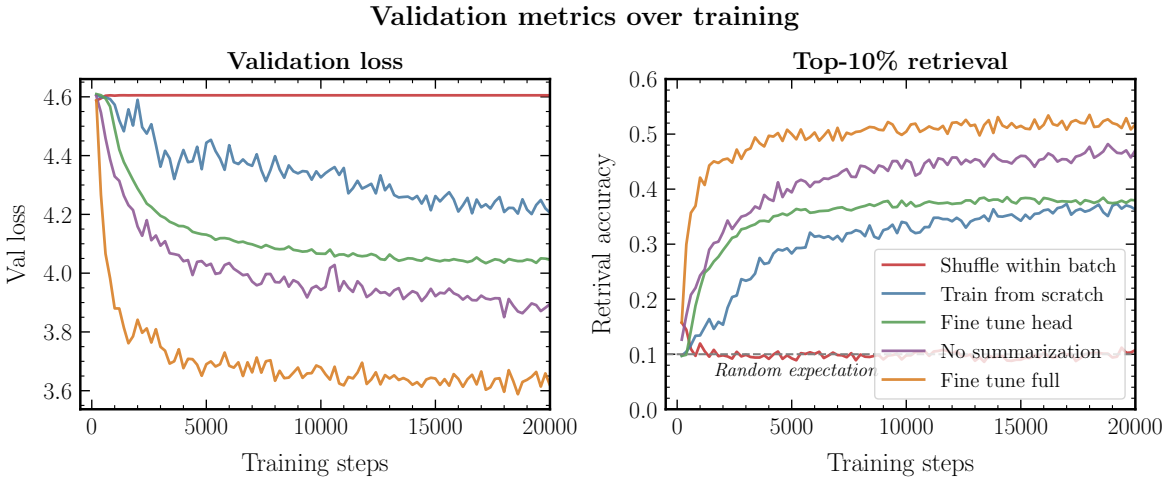


Figure 1: The CLIP-contrastive loss from Eq. 1 and the top-10% retrieval accuracy computed on the validation set over the course of training. Shown for the fiducial set-up (orange), dataset using raw proposal abstracts (purple), only fine-tuning a small MLP head (green), training from scratch (blue), and a baseline trained with shuffled image-text pairs.

Image retrieval Having aligned the image and text representations, we can query the validation set using natural language and show the most "relevant" images when ranked by cosine similarity. We show this in Figs. 2 and 3 for the base model and fine-tuned model respectively using four simple curated queries: 'globular clusters', 'dwarf galaxy', 'SN1987A', and 'cluster lensing'.

The base model shows signs of meaningful retrieval – it returns observations that clearly visually resemble globular clusters (first row of Fig. 2), for example. Beyond this, the 'dwarf galaxy' query returns a mix of

galaxies and globular clusters, and it is challenging to discern meaningful associations between the other retrieved images and corresponding query.

The fine-tuned model (Fig. 3) shows strikingly different behavior. For example, it is able to return images with processing and assembly artifacts particular to the telescope (the lines through the middle), which typically receive low similarity scores with the base model. The 'dwarf galaxy' images correspond to proposals aiming to measure the kinematics of the stellar cores of dwarf galaxies. Supernova SN1987 itself can be seen as the most relevant image for the 'SN1987A' query. Cluster-scale gravitational lenses are returned by the 'cluster lensing' query, with lensing patterns visible in the images.

Text retrieval Finally, we can use images from the validation set as queries and retrieve the most relevant text chunks (e.g., contained objects and use cases) from a curated list. We show the result of image-to-text retrieval in Fig. 4, for the base as well as fine-tuned models, using the same genre of observations as for the text-to-image retrieval examples. We curate a list of possible text associations by querying the Claude large language model for such a list, which we show in App. B.

The top 3 text associations are shown for each image query. The 'ground truth' summarized abstract is shown in the right column. The base model is seen to return a mix of relevant and less-relevant associations. While it can often return the nature of objects imaged, we observe it to seldom return scientific phenomena (e.g., 'dark matter' as successfully done by the fine-tuned model in the second row). The third row (supernova 1987A) highlights interesting behavior – the base model erroneously attributes the object at the center of the image to a gravitational lens or protoplanetary disk, while the fine-tuned model correctly identifies it as a supernova remnant (which play a crucial role for interstellar chemistry – another returned snippet).

5 Outlook and Conclusions

We present HubbleCLIP, a multi-modal fountation model that associates observations imaged by the *Hubble* Space Telescope with natural language in a common embedding space. The model is fine-tuned from a pre-trained CLIP model on LLM-summarized versions of *Hubble* proposal abstracts, leveraging a noisy signal associating text and images. We show that Hubble CLIP significantly outperforms the base CLIP model in quantitative metrics, such as retrieval accuracy, as well as quality of text-to-image and image-to-text retrieval. The procedure demonstrates the efficacy on fine-tuning generalist pre-trained models on small amounts of domain-specific data, in particular astronomical datasets.

Code and data availability The code, dataset, and models used in this work is available at `https://www.github.com/smsharma/HubbleCLIP`.

Software This work relied on the *Astroquery* (Ginsburg et al., 2019), *BitsAndBytes* (Dettmers et al., 2022), *Flax* (Heek et al., 2023), *Jax* (Bradbury et al., 2018), *Jupyter* (Kluyver et al., 2016), *Matplotlib* (Hunter, 2007), *Numpy* (Harris et al., 2020), *Optax* (Babuschkin et al., 2020), *Outlines*, *Pandas* (Virtanen et al., 2020), *Pydantic*, *PyTorch* (Paszke et al., 2019), *SciPy* (Virtanen et al., 2020), *Transformers* (Wolf et al., 2019), and *Wandb* (Biewald, 2020) software packages.

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. YS was supported by the Research Science Institute (RSI) program at MIT. 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.

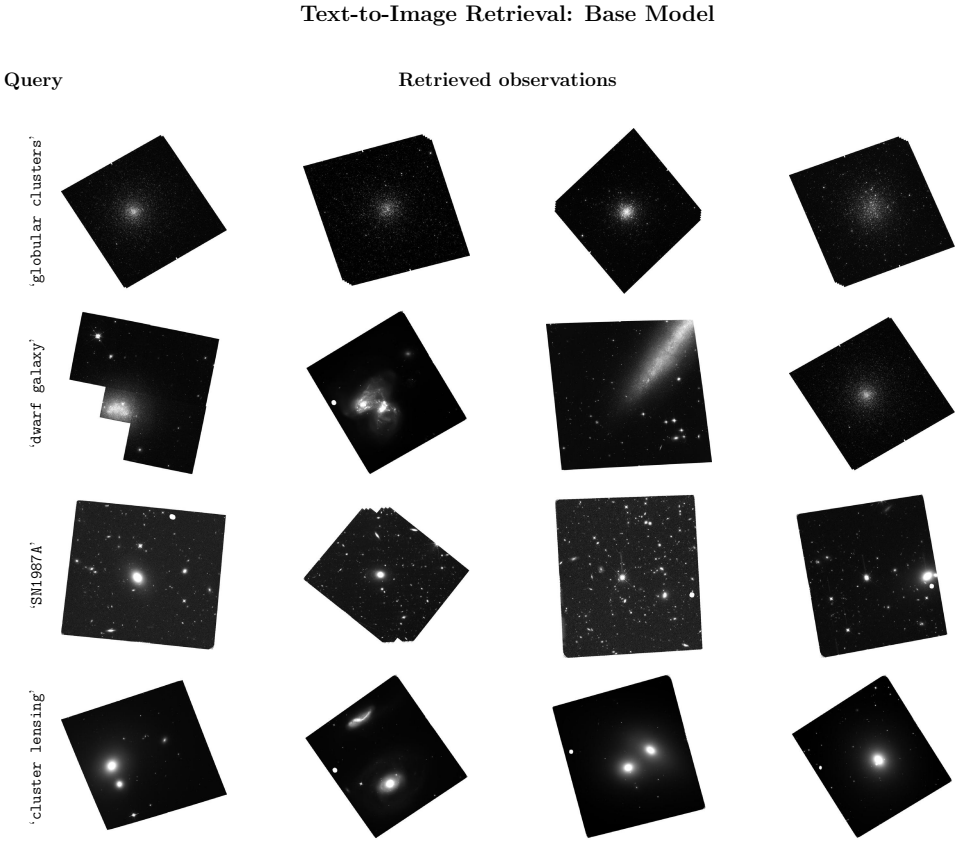


Figure 2: Image retrieval using the base CLIP model on four curated queries.

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

References

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/deepmind.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

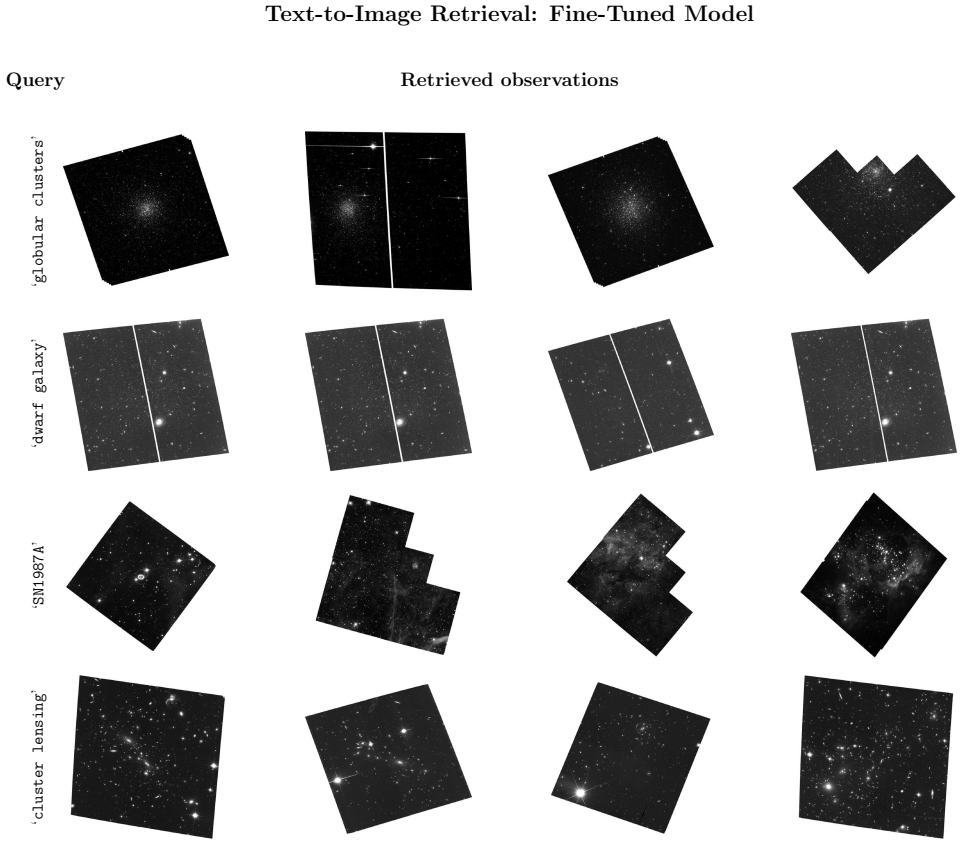


Figure 3: Image retrieval using the fine-tuned CLIP model on four curated queries.

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.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

A. Ginsburg, B. M. Sipőcz, C. E. Brasseur, P. S. Cowperthwaite, M. W. Craig, C. Deil, J. Guillochon, G. Guzman, S. Liedtke, P. Lian Lim, K. E. Lockhart, M. Mommert, B. M. Morris, H. Norman, M. Parikh, M. V. Persson, T. P. Robitaille, J.-C. Segovia, L. P. Singer, E. J. Tollerud, M. de Val-Borro, I. Valtchanov, J. Woillez, The Astroquery collaboration, and a subset of the astropy collaboration. astroquery: An Astronomical Web-querying Package in Python. *Astrophysical Journal*, 157:98, March 2019. doi: 10.3847/1538-3881/aafc33.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

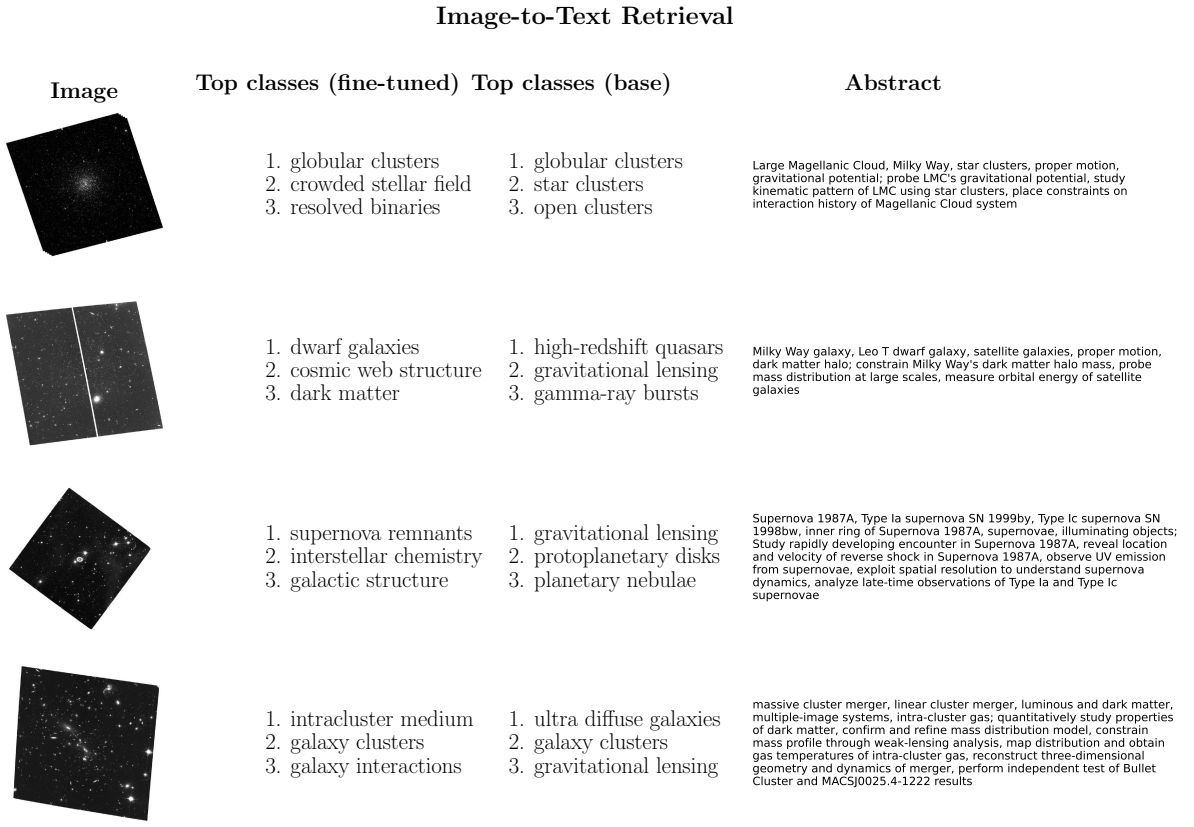


Figure 4: Text associations from a curated list most closely matching a given image query, for both the fine-tuned and base models. The 'ground truth' LLM-summarized abstract is shown in the right column.

Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2023. URL http://github.com/google/flax.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

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.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87 – 90. IOS Press, 2016.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

A Summarization via Regex-Guided Generation

The following prompt is used to summarize the abstracts using the *Outlines* package interfacing with Mixtral-8x7B-Instruct.

```
import outlines

@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]
"""
```

B List of Categories

```
["star forming galaxies", "lyman alpha", "dust", "crowded stellar field", "core-collapse supernova", "cosmology", "gravitational lensing", "supernovae", "diffuse galaxies", "globular clusters", "stellar populations", "interstellar medium", "black holes", "dark matter", "galaxy clusters", "galaxy evolution", "galaxy formation", "quasars", "circumstellar disks", "exoplanets", "Kuiper Belt objects", "solar system objects", "cosmic web structure", "distant galaxies", "galaxy mergers", "galaxy interactions", "star formation", "stellar winds", "brown dwarfs", "white dwarfs", "nebulae", "star clusters", "galaxy archeology", "galactic structure", "active galactic nuclei", "gamma-ray bursts", "stellar nurseries", "intergalactic medium", "dark energy", "dwarf galaxies", "barred spiral galaxies", "irregular galaxies", "starburst galaxies", "low surface brightness galaxies", "ultra diffuse galaxies", "circumgalactic medium", "intracluster medium", "cosmic dust", "interstellar chemistry", "star formation histories", "initial mass function", "stellar proper motions", "binary star systems", "open clusters", "pre-main sequence stars", "protostars", "protoplanetary disks", "jets and outflows", "interstellar shocks", "planetary nebulae", "supernova remnants", "red giants", "Cepheid variables", "RR Lyrae variables", "stellar abundances", "stellar dynamics", "compact stellar remnants", "Einstein rings", "trans-Neptunian objects", "cosmic microwave background", "reionization epoch", "first stars", "first galaxies", "high-redshift quasars", "primordial black holes", "resolved binaries", "binary stars"]
```

C Additional Evaluation Metrics and Ablations

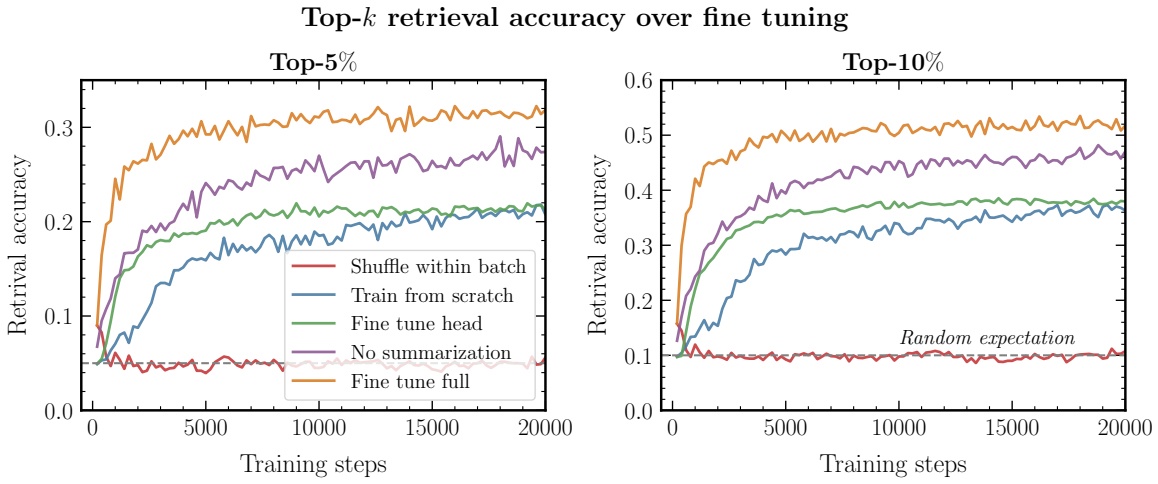


Figure 5: Retrieval accuracy