PAPERCLIP: 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 PAPERCLIP (Proposal Abstracts Provide an Effective Representation for Contrastive Language-Image Pre-training), a method which associates astronomical observations imaged by surveys and telescopes with natural language using a neural network model. The model is fine-tuned from a pre-trained Contrastive Language–Image Pre-training (CLIP) model using successful observing proposal abstracts, optionally summarized via guided generation using large language models (LLMs), and corresponding downstream observations. Using observations from the *Hubble* Space Telescope (HST) as an example, we show that the fine-tuned model embodies a meaningful joint representation between observations and text through tests targeting image retrieval (i.e., finding the most relevant observations using natural language queries) and description retrieval (i.e., querying for astrophysical object classes and science use cases most relevant to a given observation). Our study demonstrates the potential for using generalist foundation models rather than task-specific models for interacting with astronomical data by leveraging text as an interface.

Contents

1 Introduction

Machine learning (ML) is starting to have a significant impact in the sciences, with astrophysics being no exception. ML methods have demonstrated promise at every stage of the research pipeline, from instrument design, to data acquisition, to its analysis (Huertas-Company & Lanusse, 2022). Until recently, most applications of ML within astrophysics have focused on augmenting traditional techniques in order to improve performance on specific tasks. The Foundation Model paradigm, in contrast, seeks to develop generalist models which can be deployed to simultaneously tackle a wide range of tasks (Bommasani et al., 2021). These models are typically pre-trained on massive amounts of unlabeled data using self-supervised or weakly-supervised learning techniques, enabling them to learn powerful representations which can then be used downstream in different ways. Foundation models can often benefit from additional training (fine tuning) using a relatively small amounts of domain-specific data in order to increase their usefulness when applied to specialized domains.

There is considerable interest in developing custom foundation models for the sciences (e.g., Batatia et al., 2023; Subramanian et al., 2023), with astrophysics being ripe for such an effort given the large amounts of publicly-available data and diverse modes of interacting with it. The multi-modality inherent to astrophysical observations, with different types of data (e.g., images, spectra, light curves, textual descriptions) often available for a given target object, presents a unique opportunity. This multi-modality was recently exploited in AstroCLIP (Lanusse et al., 2023) to construct a joint embedding space between multi-band images and optical spectra from the Dark Energy Spectroscopic Instrument (DESI). AstroLLaMA (Nguyen et al., 2023; Perkowski et al., 2024) is another recent effort to fine-tune a publicly-available model (Llama-2) on astrophysics-specific textual data from the arXiv.

In this paper, we describe PAPERCLIP (Proposal Abstracts Provide an Effective Representation for Contrastive Language-Image Pre-training[1]), a method that connects astronomical image observations with natural language by leveraging the association between abstracts of successful observing proposals and images corresponding to downstream observations. Concretely, here we showcase the method using observations imaged by the *Hubble* Space Telescope (HST). We show that fine-tuning a pre-trained CLIP (Contrastive Language-Image Pre-training; Radford et al., 2021) image-text model on observation-abstract pairs results in meaningful joint representations through quantitative and qualitative evaluation tests. Our method

[1]Technically, we fine tune rather than pre train, but "PAPERCLIFT" was rejected by the senior author of this paper.

opens up the possibility of interacting with astronomical survey data using free-form natural language as an interface, which is a cornerstone of the success of the modern foundation model paradigm.

The CLIP family of foundation models, which in their original form embed images and associated captions into a common representation space via contrastive learning, have shown strong performance and generalization capabilities on a variety of downstream tasks including zero-shot classification and image retrieval. The concept of associating diverse modalities via contrastive training has been employed in other scientific domains (e.g., Liu et al., 2023; Sanchez-Fernandez et al., 2023; Lanusse et al., 2023; Cepeda et al., 2023), and has been shown to be effective in learning semantically meaningful joint representations. Here, we present for the first time an application associating astronomical data with the text modality.

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

2 Dataset Construction

We curate a dataset of *Hubble* Space Telescope (HST) image observations and corresponding text descriptions from publicly available sources. We rely on proposal abstracts from the Proposal Abstracts Catalog[2] – a catalog of successful HST proposals – to generate captions for the observations, optionally summarizing them via guided generation using LLMs (described in Sec. 2.2 below). The HST has been operational since its launch on April 24, 1990, and we use available proposals and observations up to the Cycle 30 science program, which commenced data-taking in 2022.

Table 1 shows examples of images and their corresponding (clipped) proposal abstracts. It can be seen that the images in this dataset exhibit specific characteristics as well as artifacts particular to HST data-taking and processing which distinguishes them from the distribution of natural images typically used for large-scale pre-training of foundation models (Deng et al., 2009). This further motivates the need for fine tuning on domain-specific data.

2.1 *Hubble* Data Selection and Pre-processing

Observations corresponding to individual proposal IDs are queried through the Mikulski Archive for Space Telescopes (MAST)[3] via the *Astroquery* (Ginsburg et al., 2019) API. Products of type `PREVIEW` are filtered in, corresponding to preview postcard images. We note that these are not science-grade observations, but rather lower-resolution images useful for diagnostic 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 and survey-style campaigns. Images are centered and resized to a resolution-per-side of 512 pixels. 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 evaluation metrics. If no appropriate images corresponding to an abstract are found, it 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 image-text pairs for evaluation. The held out images correspond to 438 unique abstracts.

We note that some fraction of the image-caption pairs in the constructed dataset will primarily concern instrumentation and/or calibration rather than scientific content. We choose to not filter out these pairs from our dataset, in order to have a larger sample of HST observations that the model can leverage to adapt to the distinctive characteristics of *Hubble* images.

[2] https://archive.stsci.edu/hst/proposal_abstracts.html
[3] https://mast.stsci.edu/

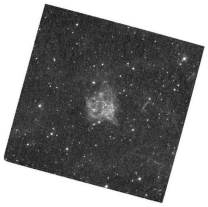
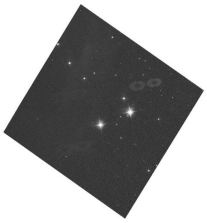
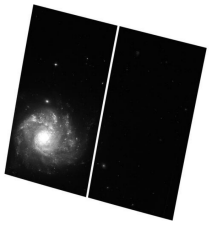
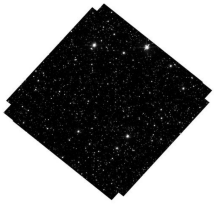
Hubble image	Obs. cycle (Year)	Prop. ID	Proposal abstract (clipped)
	7 (1999)	7340	Category: STELLAR EJECTA. We propose to use the WFPC2 and STIS CCD to obtain maximum spatial resolution emission-line images of the young, oxygen- rich supernova remnants SN0540–69.3 in the LMC and E0102.2– 7219 in the SMC. O IIILambda5007, S IILambdaLambda6724 and O IILambdaLambda3727 images of SN0540–69.3 will be used to characterize the ionization structure and...
	19 (2013)	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...
	22 (2016)	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...
	26 (2019)	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...

Table 1: Examples of *Hubble* images (left-most column) and corresponding clipped proposal abstracts (right-most column). The observation cycle and corresponding year, as well as proposal ID, are shown in the second and third columns, respectively. The proposal IDs link to the Mikulski Archive for Space Telescopes (MAST) page corresponding to the proposal.

2.2 Abstract Summarization via Guided Generation

Raw proposal 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 while also being highly variable in their content. Although the abstracts can be used as-is as image captions, we experiment with summarizing them via guided large language model (LLM) generation to standardize the captions used for fine-tuning the CLIP model. Captions are summarized by extracting a list of objects and phenomena, as well as potential downstream science use cases, corresponding to the eventual imaged observation. The intended goal of the summarization process is to increase the strength of the association 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

Prop. ID	LLM-extracted summary	
	Objects and phenomena	**Science use cases**
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
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
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
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

Table 2: For the *Hubble* proposal abstracts shown in Tab. 1, the LLM (Mixtral-8x7B)-extracted summaries showing objects and phenomena (middle column) as well as potential downstream science use cases (last column) separately.

to the eventual imaged observation given the abstract text, with a minimum of 1 and a maximum of 5 elements per list.

The procedure guides the generation of LLM outputs while ensuring that the schema is respected at every step in the generation process by masking out tokens that would violate the intended format. By framing the problem in terms of transitions between a set of finite states (a finite-state machine), Willard & Louf (2023) showed that guided generation can be performed with negligible overhead compared to unconstrained generation. See App. A.1 for a more detailed description of the guidance generation method used here.

While the schema-guided generation ensures the *format* of the output, the prompt and choice of LLM will dictate the *content* of the generated summaries. We use the open-weights, instruction-tuned model Mixtral-8x7B-Instruct (Jiang et al., 2024) to generate the summaries, with guided generation performed using the *Outlines*[4] package. Further details on the summarization procedure, including the prompts and schema used, are provided in App. A.2.

[4] https://github.com/outlines-dev/outlines

The guided generation process ensures that, in this case, 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']
}
```

which is then used to construct the summarized caption by combining the two key elements. Examples of raw abstract snippets and corresponding LLM-generated summaries are shown in Tab. 2. We train separate models using the raw abstracts and the LLM-generated summaries, and compare their performance on downstream tasks in Sec. 4. We note that, even after summarization, the association signal is expected to be noisy, since parts of the summarized caption may not be directly descriptive of the observed images. The goal of the fine-tuning process is to leverage the signal contained in this noisy association.

3 Methodology

Our goal is to learn a semantically meaningful joint representation between images corresponding to HST observation and natural (English) language. With PAPERCLIP, we leverage the strong generalization capabilities demonstrated by pre-trained CLIP models and adapt these to work with domain-specific *Hubble* data via fine tuning.

3.1 Contrastive Language-Image Pre-training (CLIP)

CLIP (Contrastive Language-Image Pre-training; Radford et al., 2021) is a multi-modal neural network 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 $f : I \to \mathbb{R}^{n_{\rm emb}}$ and $g : T \to \mathbb{R}^{n_{\rm emb}}$ are used to map images and text to a common embedding space of dimension $n_{\rm emb}$. We use the standard softmax-based bidirectional variant of the InfoNCE (Oord et al., 2018) contrastive loss function introduced for training CLIP-style architectures (Radford et al., 2021)

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

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. Note that this loss treats the image and text representations symmetrically, ensuring that the two modalities are considered on the same footing.

We use the `CLIP-ViT-B/16` (Radford et al., 2021) variant as the base pre-trained CLIP model. 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 text 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 pixels. Linear projection layers map the outputs of the image and text encoders to a common embedding space of dimension $n_{\rm emb} = 512$. In total, the model contains ~ 149 million trainable parameters. This model was originally pre-trained on 400 million image-text pairs from internet data (Radford et al., 2021).

3.2 Fine-tuning Procedure

The base CLIP model is fine-tuned using the dataset described in Sec. 2, using either the LLM-summarized abstracts or raw proposal abstracts paired with observations. 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 augmented via random four-fold rotations (increments of 90°) and randomly

cropped to the native resolution of the image encoder, maintaining $\sim 20\%$ of the area of the original image, at each training step. Given the relatively modest size of the fine-tuning dataset, a batch size $|\mathcal{B}| = 32$ is used throughout; larger batch sizes were observed to be susceptible to overfitting. The temperature hyperparameter τ was initialized to its pre-trained value. We emphasize that the positive and negative image-text association is noisy and imperfect, since multiple images can be associated with the same abstract, and the goal of the fine-tuning process is to leverage the signal contained in this noisy association.

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/text encoders and training a small projection head; and *(3)* Training the entire model from scratch. For *(2)*, we use a 2-layer MLP with 1024 hidden units and a GELU activation layer, projecting onto the 512-dimensional common embedding space.

All models were trained over 20,000 steps with 2000 linear warmup steps using the AdamW optimizer (Loshchilov & Hutter, 2019; Kingma & Ba, 2015) with learning rate 10^{-5} and weight decay 10^{-3}. Training takes approximately 3 hours on 4 Nvidia A100 GPUs. Models were instantiated using the *Transformers* (Wolf et al., 2019) library and trained using packages from the *Jax* (Bradbury et al., 2018) ecosystem.

3.3 Evaluation Metrics

The model is evaluated by tracking the contrastive 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 image and caption embeddings, averaged over the images in the validation set:

$$\text{Retrieval accuracy}_k = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} \mathbb{1}\left[\text{rank}\left(x_i \cdot y_i; \{x_i \cdot y_j\}_{j=1}^{|\mathcal{V}|}\right) \leq \left\lfloor \frac{k}{100}|\mathcal{V}| \right\rfloor\right] \tag{2}$$

where $|\mathcal{V}|$ is the total number of images in the validation set, $\mathbb{1}[\cdot]$ is the indicator function that returns 1 if the condition inside the brackets is true and 0 otherwise, $\text{rank}\left(x_i \cdot y_i; \{x_i \cdot y_j\}_{j=1}^{|\mathcal{V}|}\right)$ is a function that returns the rank of the cosine similarity between x_i and y_i among the cosine similarities between x_i and all captions y_j in the validation set, and k is the percentage of top captions considered for the retrieval accuracy. Note that this metric is symmetric in the image and text modalities.

We also qualitatively evaluate the learned embeddings through image retrieval (i.e., retrieving the most relevant images from the validation 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, akin to zero-shot classification) experiments. For the description/text retrieval evaluation, we define a list of possible text associations (i.e., classes), which we show in App. B, by querying the Claude 2[5] large language model along with manual curation.

4 Results and Discussion

4.1 Quantitative Evaluation

Validation metrics during training Figure 1 shows the contrastive loss (left) and the top-10% retrieval accuracy (right) evaluated on the held out validation set over the course of training, for different training configurations considered. The dashed orange lines show the metrics evaluated when training with batches where the image-text associations are randomly shuffled. This randomized baseline is seen to do on par with random expectation (i.e., a 10% retrieval accuracy), unlike the others, validating the presence of a significant association signal between images and text in the dataset. Interestingly, the base pre-trained model performs better than random expectation, with a top-10% retrieval accuracy of $\sim 15\%$. We therefore also compare the qualitative performance of the base model with the fine-tuned models on downstream retrieval tasks.

[5] https://claude.ai/

The model trained using LLM-summarized abstracts (red lines) is seen to perform slightly worse than the model using raw abstracts as captions (blue lines), despite the curation of the summarized-abstract dataset intended to provide a stronger image-text association signal. Fine-tuning a small MLP head over frozen vision and text backbones (dotted green lines) and training from scratch with summarized abstracts as captions (yellow lines) show a non-trivial improvement compared to the base model, although with deteriorated performance compared to fine-tuning with either summarized or raw abstracts.

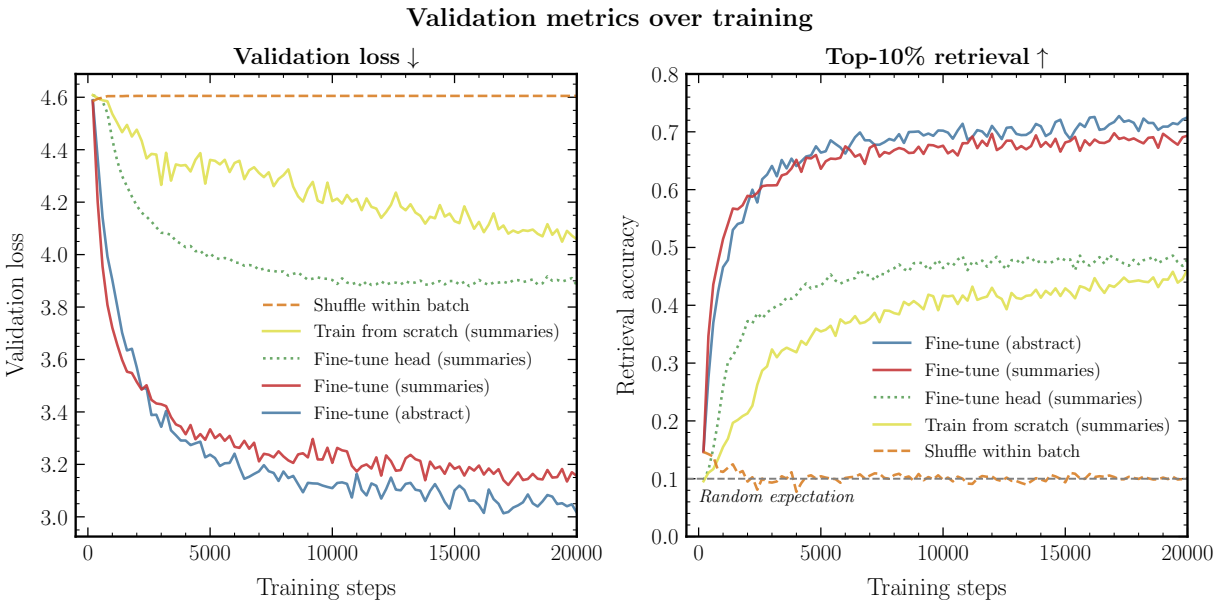


Figure 1: The CLIP contrastive loss from Eq. (1) (left) and the top-10% retrieval accuracy from Eq. (2) (right) computed on the validation set over the course of training. Shown for the dataset with summarized abstracts as captions (red), dataset using raw proposal abstracts as captions (blue), only fine-tuning a small MLP head (dotted green), training from scratch with summarized abstracts as captions (yellow), and trained with shuffled image-text pairs (dashed orange).

Distribution of text-image cosine similarities Figure 2 (left) shows the distribution of cosine similarities between corresponding image and text embeddings, x_i and y_i, for the base CLIP model (purple line), and for the LLM-summarized abstracts using the fine-tuned CLIP model (red line). Distributions evaluated for a shuffled order of text embeddings – therefore randomizing the image-text correspondence during evaluation – are shown as dashed lines. We note that the shuffling here is performed at the evaluation stage, and not the training stage. The distributions for the base model is seen to be sharply peaked at a specific value, showing little diversity and being very similar between the shuffled (dashed purple) and non-shuffled (solid purple) versions. Distributions for the fine-tuned model, on the other hand, show a clear separation when evaluated on shuffled (dashed red) and corresponding (solid red) text-image pairs.

Retrieval accuracy Figure 2 (right) shows the retrieval accuracy, as defined in Eq. (2), as a function of the retrieval fraction k%. In this case, we evaluate all four models (fine tuned on raw abstracts (blue), fine-tuned on LLM-summarized abstracts (red), trained on LLM-summarized abstracts from scratch (yellow), and the base model (purple)) on the same captions dataset – the summarized abstracts – for a direct comparison. Remarkably, the model trained on raw abstracts shows very similar performance when evaluated on the summarized abstracts compared to that trained on the summarized abstracts themselves, indicating that *(1)* the image-text association signal is preserved in the summarization process, and *(2)* the model is able to effectively leverage meaningful concepts in the noisy raw abstracts through weak supervision. The significantly worse performance of the model trained from scratch, compared to the fine-tuned models, highlights the crucial role of the inductive bias inherent in the base pretrained model, which effectively captures the rich associations between images and language.

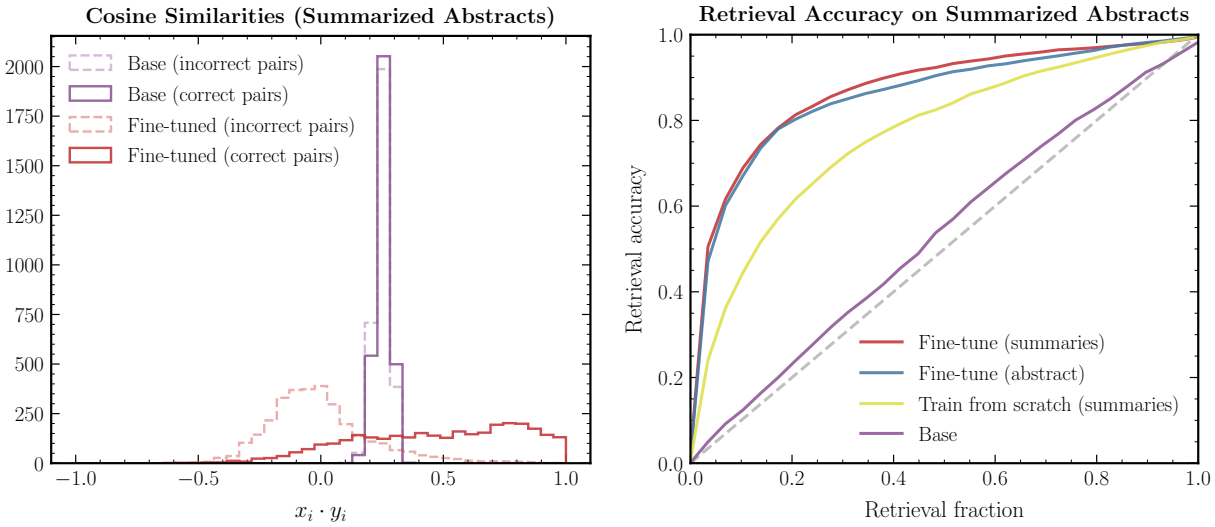


Figure 2: (Left) Distribution of cosine similarities between corresponding image and text embeddings, x_i and y_i, shown when using the base CLIP model (purple lines), and the summary fine-tuned CLIP model (red line). Dashed lines correspond to models evaluated on image-text pairs with associations shuffled. (Right) Retrieval accuracy as a function of the retrieval fraction k for the fine-tuned model on the summarized abstracts (red), fine-tuned on raw abstracts (blue), trained on summarized abstracts from scratch (yellow), and the base model (purple).

We show retrieval accuracy performance for additional variations on the model and training configuration in App. D.

4.2 Image Retrieval

Having aligned the image and text representations, we can embed a natural language query using the model and show the closest images by embedding from the validation set when ranked by cosine similarity. We show these in Tabs. 3 and 4 for the base and fine-tuned models respectively using four simple curated queries: `dwarf galaxy` (small galaxies that typically orbit larger galaxies like the Milky Way), `Jupiter`, `SN1987A` (a specific, prominent supernova), and `strong lensing` (the phenomenon of bending of light due to the gravitational influence of a foreground distribution of matter). The proposal ID corresponding to the retrieved images is shown below each image, and contains a hyperlink to the MAST page corresponding to the proposal for further details.

While the base model shows some signs of meaningful retrieval (e.g., the image of Jupiter in the second row of Tab. 3, and images of galaxies in first row), it is challenging to discern meaningful, strong associations between the retrieved images and corresponding query.

The model fine-tuned with summarized abstracts, meanwhile, shows strikingly different behavior (Tab. 4). The `dwarf galaxy`-queried images correspond to proposals aiming to measure the kinematics of the stellar cores of dwarf galaxies. Images of Jupiter are returned for the `Jupiter` query, with the planet clearly visible in all four images. Supernova SN1987 itself can be seen in the three closest images for the `SN1987A` query with the fourth image being a supernova remnant. Cluster-scale as well as galaxy-scale gravitational lenses are returned by the `strong lensing` query, with lensing patterns visible in the images.

4.3 Text Retrieval

We can use images from the validation set as queries and retrieve the most relevant text chunks (e.g., objects and use cases) from a curated list as described in Sec. 3.3. We show the result of image-to-text retrieval

in Tab. 5, for the base (second column) as well as summary fine-tuned (third column) models, using four observations (left-most column) from the validation set.

The top four text associations are shown for each image query. The 'ground truth' summarized abstract is shown in the right column. The base as well as fine-tuned models are seen to return a mix of relevant and less-relevant associations, although showing different qualitative behavior. Purely qualitatively, the fine-tuned model is seen to consistently return more relevant associations compared to the base model.

The second row (an image of supernova 1987A) highlights an interesting pattern – the base model erroneously attributes the object at the center of the image to a gravitational lens, while the fine-tuned model correctly identifies it as a supernova remnant. This kind of reasonable misattribution is common when querying the base model, and largely absent in the fine-tuned model.

Note that we chose to illustrate qualitative performance on text and image retrieval using the model fine-tuned on summarized abstracts, rather than raw abstracts. We show analogous results for the model fine-tuned on raw abstracts in App. C. Although the two models show very similar quantitative performance on retrieval metrics (as shown in Fig. 2), they exhibit characteristically different behaviors in terms of objects (images/text) retrieved. We emphasize that for scientific usefulness, the goal is not necessarily to correctly retrieve the most "relevant" objects, but rather to identify a diverse set of interesting candidates for manual follow-up and further analysis; both models are seen to perform sensibly, even if differently, in this regard.

5 Outlook and Conclusions

In this paper, we present PAPERCLIP, a method for training domain-specific multi-modal models for astrophysics that associates observations imaged by telescopes with natural language in a common, semantically-meaningful embedding space. We showcase an application to *Hubble* Space Telescope (HST) observations, where the model is fine-tuned from a pre-trained CLIP model using abstracts of successful *Hubble* proposals, optionally summarized, leveraging a noisy association signal between text and images. We show that PAPERCLIP 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. We also introduce a novel LLM summarization process which leverages guided generation to distill the content of proposal abstracts while preserving salient information. Overall, the procedure demonstrates the efficacy of fine-tuning generalist pre-trained models on small amounts of domain-specific data, in particular astronomical datasets, and leveraging text as an interface.

Although the model explored here is fine-tuned using postage stamp images (i.e., preview-quality and not science-grade data), we highlight potential immediate as well as downstream use cases. A model trained using weakly-supervised image-text pairs can be used to query survey data e.g., PHANGS (Lee et al., 2022), COSMOS (Scoville et al., 2007) using natural language, as well as to efficiently find patterns in such data that may not be apparent using specialized models or manual inspection. The learned representations, having shown to correlate with physical characteristics of imaged objects, can also be fine-tuned via transfer learning to adapt to either specific tasks e.g., classification (Wei et al., 2020) or segmentation (Hausen & Robertson, 2020), or observations imaged by other telescopes.

Finally, while the CLIP model is restricted to retrieving nearest-neighbour associations within and across text/image modalities, the learned embeddings can be used as a starting point for training or fine-tuning multi-modal large-language models for interacting with survey data and receiving responses in natural language form, as well as grounding the responses based on an existing set of observations.

Code and Data Availability

The code, dataset, and fine-tuned models used in this work are available at https://www.github.com/smsharma/HubbleCLIP.

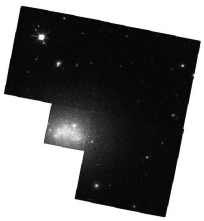
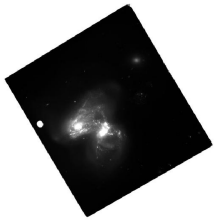
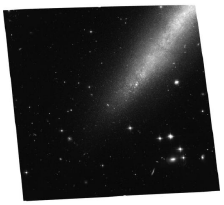
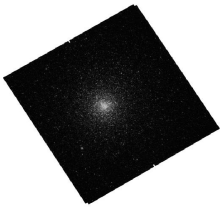
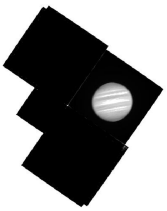
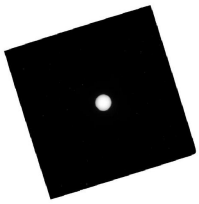
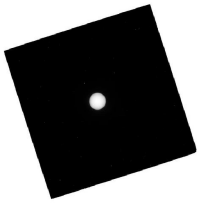
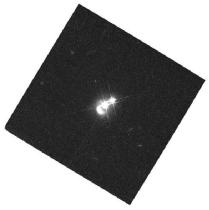
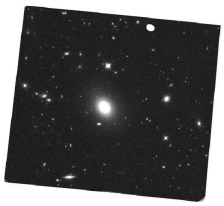
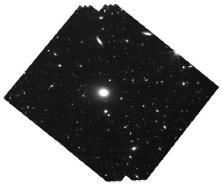
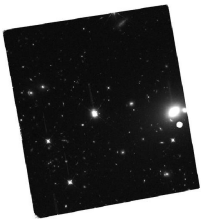
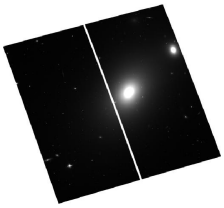
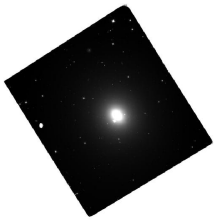
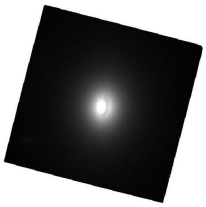
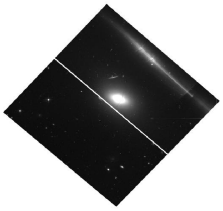
Query	Top-4 most similar images using base CLIP model			
dwarf galaxy	 8122	 15649	 12196	 12605
Jupiter	 6028	 10170	 10170	 6303
SN1987A	 13830	 15475	 14594	 14594
strong lensing	 10787	 14654	 9106	 16025

Table 3: For four text queries (left-most column), the four most similar images from the validation dataset by cosine similarity when using the **base CLIP model** (CLIP-ViT-B/16). The proposal ID associated with each image is given below the image and contains a hyperlink to the MAST page corresponding to the proposal.

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),

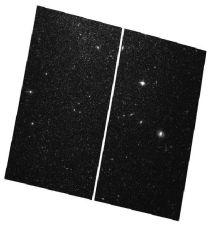
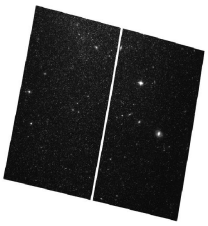
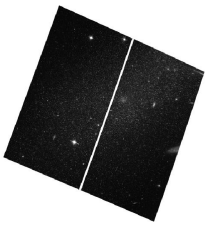
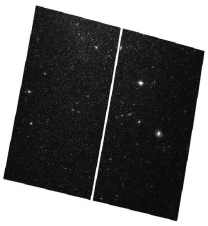
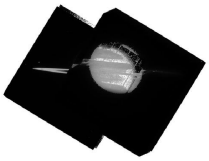
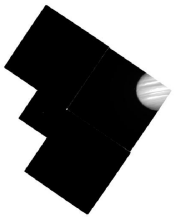
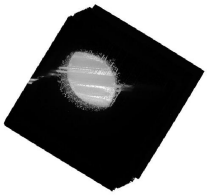
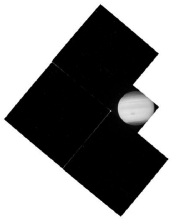
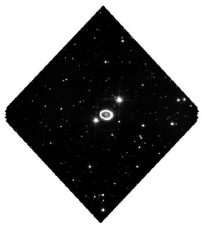
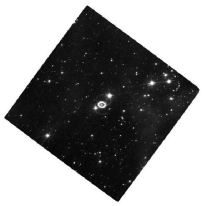
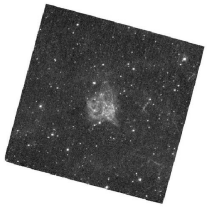
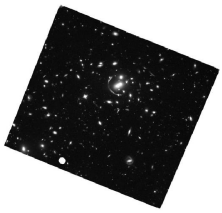
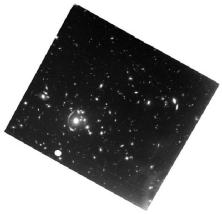
Query	Top-4 most similar images using summary fine-tuned CLIP model			
`dwarf galaxy`	 13768	 13768	 13768	 13768
`Jupiter`	 11956	 6028	 11956	 11096
`SN1987A`	 11653	 11653	 8648	 7340
`strong lensing`	 14098	 11602	 11602	 12068

Table 4: Same as Tab. 3, but using the **summary fine-tuned CLIP model**.

Pydantic, *PyTorch* (Paszke et al., 2019), *SciPy* (Virtanen et al., 2020), *Transformers* (Wolf et al., 2019), and *Wandb* (Biewald, 2020) software packages.

Broader Impact

This work relies on using abstracts from successful *Hubble* Space Telescope observing proposals as part of a dataset for training and evaluating machine learning models. While these abstracts are publicly available, the authors likely did not anticipate their text being used in this manner, raising questions around consent, attribution, and appropriate use of data. Since this research intends to develop methods to aid astronomical

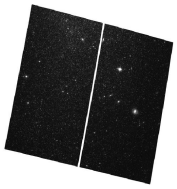
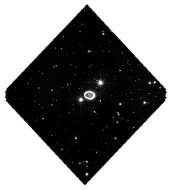
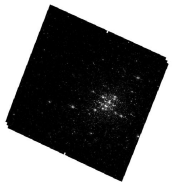
Hubble image	Top-4 text (base)	Top-4 text (summary fine-tuned)	Summarized abstract (objects; 'ground truth')
	1. high-redshift quasars 2. gravitational lensing 3. white dwarfs 4. dwarf galaxies	1. dwarf galaxies 2. RR Lyrae variables 3. red giants 4. trans-Neptunian objects	isolated dwarf galaxies, WLM, Pegasus Dwarf Irregular Galaxy, stellar mass, main sequence stars
	1. gravitational lensing 2. supernovae 3. binary star systems 4. circumstellar disks	1. supernova remnants 2. protostars 3. galactic structure 4. core-collapse supernova	supernova SN 1987A, circumstellar ring, supernova remnant, shocked ring, radioactive isotopes
	1. gravitational lensing 2. high-redshift quasars 3. ultra diffuse galaxies 4. galaxy clusters	1. galaxy clusters 2. lyman alpha 3. intracluster medium 4. dark energy	X-ray luminous galaxy clusters, eMACS clusters, Balmer Break Galaxies, Lyman-break galaxies, gravitational telescopes
	1. star clusters 2. globular clusters 3. open clusters 4. stellar populations	1. globular clusters 2. star clusters 3. galactic structure 4. crowded stellar field	pre-main sequence stars, Large Magellanic Cloud, young clusters, color-magnitude diagrams, main-sequence turn offs

Table 5: Text snippets from a curated list most closely matching a given image query (left-most column) by cosine similarity of respective embeddings, shown for the base (CLIP-ViT-B/16) and summary fine-tuned models. The 'ground truth' LLM-summarized abstract (only objects/phenomena) is shown in the right-most column.

research and does not use sensitive personal information or target commercial gain, we believe that the scientific benefits outweigh the potential concerns in this case while acknowledging good-faith arguments to the contrary. As the use of foundation models in the sciences increases, it will be important for the community to consider norms and guidelines around the appropriate use and attribution of various data sources for model training and evaluation, including qualitative textual data, to ensure transparency and maintain trust.

Acknowledgments

We thank Michael Brenner for helpful conversations. 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. This research was supported by an award from Google, "Interpretation of Multimodal Images from Astronomy". This research was supported by the Munich Institute for Astro-, Particle and BioPhysics (MIAPbP), which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311. 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 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.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023.

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

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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.

Ryan Hausen and Brant E Robertson. Morpheus: A deep learning framework for the pixel-level analysis of astronomical image data. *The Astrophysical Journal Supplement Series*, 248(1):20, 2020.

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.

Marc Huertas-Company and François Lanusse. The dawes review 10: The impact of deep learning for the analysis of galaxy surveys. *arXiv preprint arXiv:2210.01813*, 2022.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

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.

Janice C Lee, Bradley C Whitmore, David A Thilker, Sinan Deger, Kirsten L Larson, Leonardo Ubeda, Gagandeep S Anand, Médéric Boquien, Rupali Chandar, Daniel A Dale, et al. The phangs-hst survey: Physics at high angular resolution in nearby galaxies with the hubble space telescope. *The Astrophysical Journal Supplement Series*, 258(1):10, 2022.

Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 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.

Ernest Perkowski, Rui Pan, Tuan Dung Nguyen, Yuan-Sen Ting, Sandor Kruk, Tong Zhang, Charlie O'Neill, Maja Jablonska, Zechang Sun, Michael J Smith, et al. Astrollama-chat: Scaling astrollama with conversational and diverse datasets. *Research Notes of the AAS*, 8(1):7, 2024.

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.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *bioRxiv*, 2023. doi: 10.1101/2022.11.17.516915. URL https://www.biorxiv.org/content/early/2023/06/01/2022.11.17.516915.

Nick Scoville, H Aussel, Marcella Brusa, Peter Capak, C Marcella Carollo, M Elvis, M Giavalisco, L Guzzo, G Hasinger, C Impey, et al. The cosmic evolution survey (cosmos): overview. *The Astrophysical Journal Supplement Series*, 172(1):1, 2007.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. *arXiv preprint arXiv:2306.00258*, 2023.

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.

Wei Wei, EA Huerta, Bradley C Whitmore, Janice C Lee, Stephen Hannon, Rupali Chandar, Daniel A Dale, Kirsten L Larson, David A Thilker, Leonardo Ubeda, et al. Deep transfer learning for star cluster classification: I. application to the phangs–hst survey. *Monthly Notices of the Royal Astronomical Society*, 493(3):3178–3193, 2020.

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 Details on the Abstract Summarization Procedure

A.1 Guided LLM Generation with *Outlines*

As mention in Sec. 2.2, we employ the guided generation method introduced by Willard & Louf (2023) and implemented in *Outlines* to ensure that the LLM summarization of the raw proposal abstracts adheres to specific pattern, specified in JSON format (Sec. A.2 below), which we briefly describe here. This approach represents the desired output format as a finite-state machine (FSM) that encodes the JSON schema as a regular expression. The JSON schema constraint is therefore first converted into a regular expression.

The key idea then is to pre-compute an index that maps each state of the FSM to the subset of tokens from the LLM's vocabulary that can be generated from that state while still allowing for a valid completion of the pattern. By doing so, we can efficiently determine the valid next tokens at each step of the generation process without having to check the entire vocabulary.

Formally, let $\mathcal{M} = (Q, \Sigma, \delta, q_0, F)$ be the FSM representing the regular expression, where Q is the set of states, Σ is the alphabet of the regular expression, $\delta : Q \times \Sigma \to Q$ is the transition function between states, q_0 is the start state, and F is the set of accept states which terminate the generation. An index $\sigma : Q \to \mathcal{P}(V)$

is first constructed, where V is the LLM's token vocabulary and $\mathcal{P}(V)$ denotes the power set of V. For each state $q \in Q$, $\sigma(q)$ contains the allowed tokens that can be generated from state q while maintaining the possibility of reaching an accept state. The construction of σ involves finding all token sequences that, when processed by the FSM starting from each state q, lead to an accept state.

During the token-by-token generation process, we keep track of the current FSM state q_t after sampling each token v_t. At each step t, we mask the LLM's output logits based on the valid next tokens $\sigma(q_t)$, setting the logits of invalid tokens to $-\infty$. The next token is then sampled from the categorical distribution defined by the unmasked logits, and the FSM transitions to the next state $q_{t+1} = \delta(q_t, v_{t+1})$, where $v_{t+1} \in \Sigma$ is the token in the regular expression alphabet corresponding to the sampled token. This process continues until an accept state with no outgoing transitions is reached, indicating a valid completion of the pattern.

A.2 Prompts and Schema Used for Summarization

We list here the prompts and schema (i.e., desired output formats) used for guided text generation via *Outlines* package interfacing with the MIXTRAL-8X7B-INSTRUCT open-weights LLM.

The following schema is used to guide the generation of the summaries, intended to produce between one and five objects and hypotheses, as well as science use cases.

```
from pydantic import BaseModel, conlist

class ConstrainedResponseHST(BaseModel):
      objects_and_phenomena: conlist(str, min_length=1, max_length=5)
      science_use_cases: conlist(str, min_length=1, max_length=5)
```

The following prompt function is used to produce a list of possible objects and phenomena shown in HST observations downstream of a proposal abstract, as well as one to five possible science use cases.

```
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 for Text Retrieval Task

The following curated categories are used in the text retrieval experiment in Sec. 4. These are derived by initially prompting CLAUDE 2, having attached a subsample of 30 proposal abstracts in the online interface to be used as context, to produce a list of categories corresponding to typical HST observations. The list is then manually curated to remove similar entries and ensure a representative sample 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"]
```

The following prompt is used to generate the initial list before manual curation: *"Here is a list of Hubble proposals. Base on this, please provide a list of about 100 strings, each describing a science target or use case for observations imaged by the Hubble Space Telescope. You may use these proposals and also rely on your general knowledge. For example, ["gravitational lensing", "supernovae", "diffuse galaxies", ...]"*

C Evaluation of Model Trained on Raw Abstracts

In the main text, we illustrated qualitative evaluation (image and text retrieval) for the model fine-tuned on summarized abstracts. Here, we show the same for the model fine-tuned on raw proposal abstracts. Table 6 shows the top-4 most similar images for the abstract fine-tuned CLIP model on the same curated queries as in Tab. 4 for the summary fine-tuned model. Table 7 shows text associations from the curated list most closely matching the image queries, for the base and abstract fine-tuned models, as well as the summary fine-tuned model, for comparison. Although qualitatively different behavior is observed for both tasks, the objects retrieved are seen to, in most cases, meaningfully correspond to the given image/text queries.

D Variations on Model and Training

Figure 3 shows the retrieval accuracy as defined in Eq. (2) as a function of the retrieval fraction for further variations of the model or training, evaluated and trained on summarized abstracts. The red line corresponds to the model trained on summarized abstract described in the main text (fine-tuned on `CLIP-ViT-B/16` with constant learning rate LR $= 10^{-5}$ after linear warmup). The purple line corresponds to the base `CLIP-ViT-B/16` model.

Curves for the model fine-tuned on the larger base CLIP model `CLIP-ViT-L/14` (dotted red), with a smaller learning rate LR $= 10^{-6}$ (dashed green), and with a cosine learning rate schedule (green) are also shown.

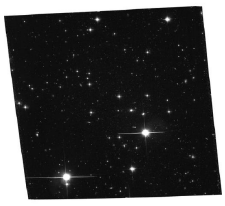
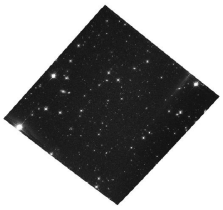
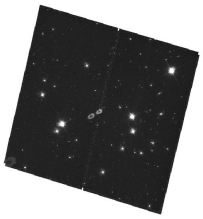
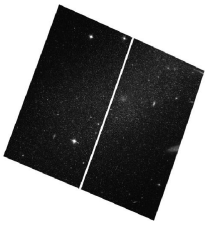
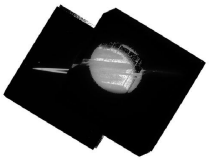
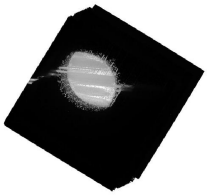
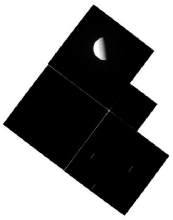
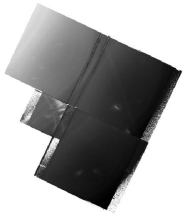
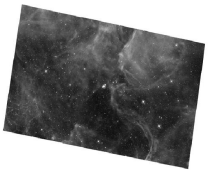
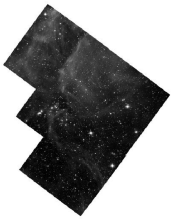
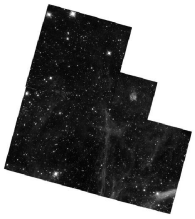
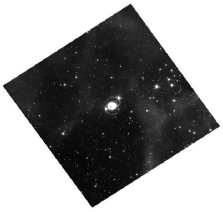
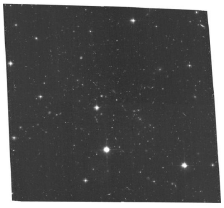
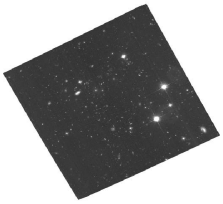
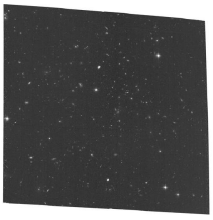
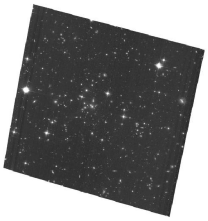
Query	Top-4 most similar images using abstract fine-tuned CLIP model			
dwarf galaxy				
	14259	14259	16293	13768
Jupiter				
	11956	11956	5783	5662
SN1987A				
	14904	8648	7340	16265
strong lensing				
	13412	13412	14098	14098

Table 6: Same as Tabs. 3 and 4, but using the **abstract fine-tuned CLIP model**.

All these models are seen to perform similarly, with the exception of the model trained with smaller learning rate showing degraded performance. Given the similar performance between `CLIP-ViT-L/14` ($\sim$ 428 million parameters) and `CLIP-ViT-B/16` ($\sim$ 149 million parameters), we chose the latter as the base model in the main text for computational efficiency.

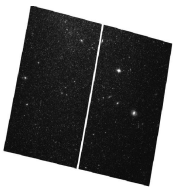
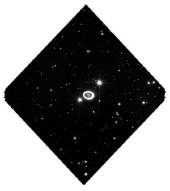
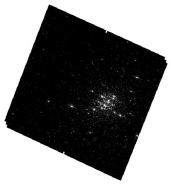
Hubble image	Top-4 text (base)	Top-4 text (abstract fine-tuned)	Top-4 text (summary fine-tuned)
	1. high-redshift quasars 2. gravitational lensing 3. white dwarfs 4. dwarf galaxies	1. dwarf galaxies 2. RR Lyrae variables 3. stellar populations 4. primordial black holes	1. dwarf galaxies 2. RR Lyrae variables 3. red giants 4. trans-Neptunian objects
	1. gravitational lensing 2. supernovae 3. binary star systems 4. circumstellar disks	1. supernova remnants 2. pre-main sequence stars 3. crowded stellar field 4. planetary nebulae	1. supernova remnants 2. protostars 3. galactic structure 4. core-collapse supernova
	1. gravitational lensing 2. high-redshift quasars 3. ultra diffuse galaxies 4. galaxy clusters	1. galaxy clusters 2. cosmic web structure 3. intracluster medium 4. dark matter	1. galaxy clusters 2. lyman alpha 3. intracluster medium 4. dark energy
	1. star clusters 2. globular clusters 3. open clusters 4. stellar populations	1. star clusters 2. stellar populations 3. primordial black holes 4. globular clusters	1. globular clusters 2. star clusters 3. galactic structure 4. crowded stellar field

Table 7: Text associations from a curated list most closely matching four image queries (first column, the same as in Tab. 5), for the base (CLIP-ViT-B/16), abstract fine-tuned, and summary fine-tuned models.

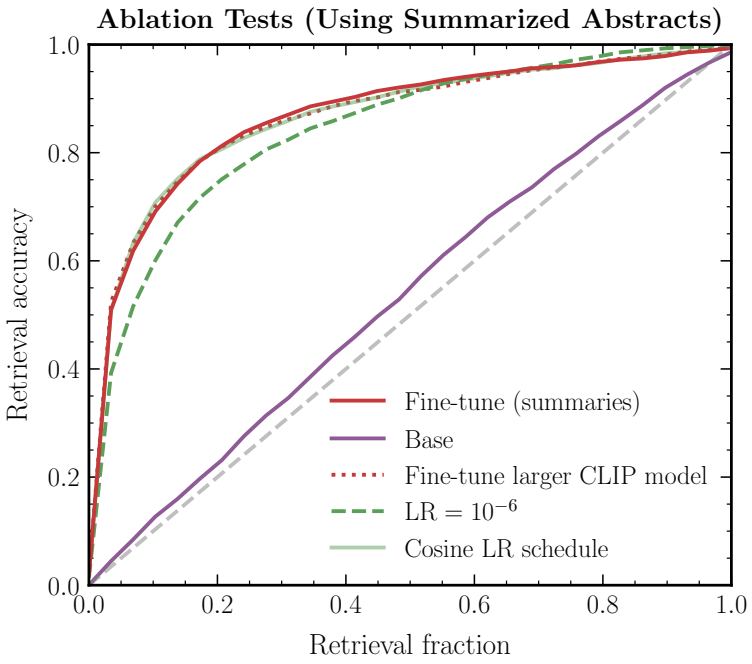


Figure 3: Same as Fig. 2 (right) – retrieval accuracy as a function of the retrieval fraction – for further variations on the model or training. The red and purple lines correspond to the model trained on summarized abstract, described in the main text, and the base `CLIP-ViT-B/16` model, respectively. Curves for the model fine-tuned on the larger base CLIP model `CLIP-ViT-L/14` (dotted red), with a smaller learning rate LR $= 10^{-6}$ (dashed green), and with a cosine learning rate schedule (green) are also shown.