Semantic search for 100M+ galaxy images using AI-generated captions

Nolan Koblischke[*1,2], **Liam Parker**[3,2,4,5], **Francois Lanusse**[6,5],
Irina Espejo Morales[2], **Jo Bovy**[1], **Shirley Ho**[2,5,7]

[1]University of Toronto, [2]New York University, [3]University of California, Berkeley, [4]Lawrence Berkeley National Laboratory, [5]Flatiron Institute, [6]Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, [7]Princeton University

Abstract

Finding scientifically interesting phenomena through slow, manual labeling campaigns severely limits our ability to explore the billions of galaxy images produced by telescopes. In this work, we develop a pipeline to create a *semantic* search engine from completely unlabeled image data. Our method leverages Vision-Language Models (VLMs) to generate descriptions for galaxy images, then contrastively aligns a pre-trained multimodal astronomy foundation model with these embedded descriptions to produce searchable embeddings at scale. We find that current VLMs provide descriptions that are sufficiently informative to train a semantic search model that outperforms direct image similarity search. Our model, AION-Search, achieves state-of-the-art zero-shot performance on finding rare phenomena despite training on randomly selected images with no deliberate curation for rare cases. Furthermore, we introduce a VLM-based re-ranking method that nearly doubles the recall for our most challenging targets in the top-100 results. For the first time, AION-Search enables flexible semantic search scalable to 140 million galaxy images, enabling discovery from previously infeasible searches. More broadly, our work provides an approach for making large, unlabeled scientific image archives semantically searchable, expanding data exploration capabilities in fields from Earth observation to microscopy. The code, data, and app are publicly available at https://github.com/NolanKoblischke/AION-Search.

1 Introduction

Recent advances in AI capabilities have prompted visions of "a country of geniuses in a datacenter", highly capable AI agents available on demand to accelerate scientific discovery [Amodei, 2024]. In astrophysics, where telescopes will generate billions of galaxy images that far exceed human capacity for manual inspection, deploying such AI agents at scale presents an opportunity to analyze vast astronomical datasets. Here, we take a step towards this vision by using Vision-Language Models (VLMs) as image annotators. We investigate whether VLMs can generate informative descriptions of galaxy images, then use this capability to train AION-Search, a CLIP-based search engine on these generated captions. PAPERCLIP pioneered astronomical text-image retrieval using ~4,000 abstracts and associated observations for Hubble Space Telescope observing proposals, though the approach faced inherent limitations from misaligned text-image pairs, where one text can map to multiple images or contain irrelevant information [Mishra-Sharma et al., 2024]. With VLM annotators, we can overcome these constraints by generating specific descriptions for each image, enabling semantic search for any telescope at scale. The Legacy Survey [Dey et al., 2019] and Hyper Suprime-Cam (HSC) [Aihara et al., 2018] are wide-field imaging surveys that have collectively imaged hundreds of millions of galaxies. While these datasets are a valuable resource for studying galaxy evolution and

---

cosmology, finding specific phenomena of interest, such as gravitational lenses or specific galaxy morphologies, remains challenging due to the sheer data volume.

Currently, scientists rely on volunteers to answer fixed-choice questions, which are then used to train classifiers for finding specific types of galaxies [Walmsley et al., 2018, Gonzalez et al., 2025]. These sorts of labeling campaigns, like Galaxy Zoo, are costly and slow, often requiring months of effort [Walmsley et al., 2022, Walmsley et al., 2025, Lintott et al., 2008]. For Galaxy Zoo, volunteers answer hierarchical fixed-choice questions such as "Is there any sign of a spiral arm pattern?" followed by "How many spiral arms are there?" With upcoming surveys like Euclid (1.5 billion galaxies by 2031 [Laureijs et al., 2011]) and LSST (20 billion by 2035 [Ivezić et al., 2019]), these manual approaches will inevitably miss discoveries. These labeling campaigns have been used to train supervised deep-learning models [Walmsley et al., 2023, Walmsley et al., 2022]. However, such models inherit the biases of the volunteer-labeled training set and are constrained to predicting answers to fixed question sets (the models explicitly learn the fraction of volunteers selecting each answer), and therefore do not support free-form, open-vocabulary queries.

While self-supervised methods offer similarity search [Stein et al., 2021, Parker et al., 2024, Hayat et al., 2021], they retrieve the most visually similar images to the query image, rather than all images containing a desired feature, making them inadequate for targeted searches. An effective semantic search engine would require informative captions that capture the physical phenomena present in galaxy images. VLMs demonstrate promising capabilities for this task, having been pretrained on extensive corpora that include research papers, observational annotations, and online discourse about galaxy images. This allows them to describe a broad range of observable phenomena rather than the predefined categories derived from volunteer labeling campaigns typically used for supervised training. Our objective is to evaluate whether VLM-generated descriptions possess sufficient accuracy and information content to serve as training data for a contrastive learning framework. To enable efficient semantic search at scale, we train a CLIP-based model that aligns image representations from AION [Parker et al., 2025], an astronomy multimodal foundation model, with these VLM-generated descriptions. With the flexibility provided by AION and VLM descriptions, this approach provides scalable semantic search extensible to any future telescopes.

1.1 Related Work

In recent years, approaches to improve text-visual understanding, such as CLIP [Radford et al., 2021], have achieved remarkable performance on a variety of downstream tasks, including retrieval, by training on pairs of image-captions extracted from the web. However, because high-quality image-text pairs are finite, recent work has explored using synthetic captions to augment training data [Zheng et al., 2024, Zhang et al., 2025]. For CLIP-like models, SynthCLIP [Hammoud et al., 2024] demonstrates that while performance with fully synthetic captions does not match performance with real captions, it exhibits good scalability trends and has competitive performance in some of the downstream tasks. For retrieval, CLIPS [Liu et al., 2024b] proposes to learn both with web and synthetically augmented captions, achieving state-of-the-art in cross-modal retrieval tasks.

In the context of scientific images, the scarcity of text-image pairs is even more pronounced because it relies on domain expertise and standardized benchmarks are rare. A common workaround is to convert categorical labels into natural-language prompts with templates and train CLIP-style models on these synthetic pairs [Liu et al., 2024a, Shi and Zou, 2017, Khattak et al., 2024]. In particular, INQUIRE [Vendrow et al., 2024] produces a benchmark for text-to-image retrieval in ecology and additionally evaluates reranking with multimodal models. More recently, a few efforts experiment with generating captions directly from images using pre-trained VLMs [Chen et al., 2025, He et al., 2025, Yuan et al., 2025]. VLMs have also been used in astronomy for image classification [Tanoglidis and Jain, 2024, Zaman et al., 2025]. To our knowledge, this work is the first example in astrophysics of using VLM-generated captions to generate training sets for semantic retrieval.

1.2 Paper Outline

We begin in §2.1 by benchmarking whether VLMs can generate astronomy-relevant image descriptions by comparing them to human annotations on a set of galaxy images. Next, in §2.2, we use the best prompt-model pair to generate free-form captions for a large collection of telescope images and embed the text. To generate semantically searchable embeddings directly from images, we learn
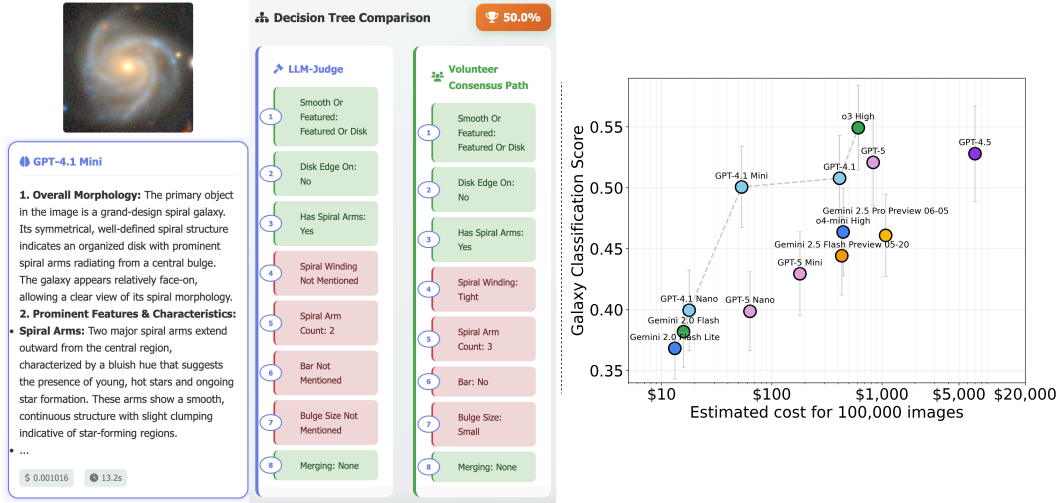
Figure 1: **Evaluating VLM descriptions and the accuracy-cost trade-off.** *Left:* GPT-4.1-mini generates a free-form description for a galaxy image. An LLM-judge converts the text into Galaxy Zoo answers and we compare them to the volunteer consensus answers. *Right:* Mean performance of OpenAI and Google VLMs on the benchmark versus the estimated batch-API cost to caption 100,000 images. Points show the mean over three runs (GPT-4.5: one run), vertical bars denote standard errors on the mean. The dashed line highlights the Pareto frontier. GPT-4.1-mini offers a strong accuracy per dollar, while larger models (e.g., o3) are more accurate but more costly.

in §2.3 a shared text-image space by contrastively aligning a frozen, pre-trained astronomy image encoder to the VLM text embeddings using shallow projection heads. We evaluate in §2.4 zero-shot retrieval with natural-language queries targeting three scientifically interesting galaxy-image phenomena: *spiral galaxies* (galaxies with winding arms), *galaxy mergers* (two galaxies colliding or recently merged, sometimes leaving distorted shapes), and *strong gravitational lenses* (where a foreground galaxy bends the light from a more distant object so it appears as arcs, rings, or multiple images), and we test in §2.5 a VLM-based re-ranking stage. This makes large datasets of unlabeled astronomical images semantically searchable.

2 Methodology

2.1 Evaluating VLM image descriptions

We evaluate the capacity of Vision-Language Models to generate scientifically accurate descriptions of galaxy images using the Galaxy Zoo-DECaLS catalog [Walmsley et al., 2022], which provides crowd-sourced annotations for Legacy Survey images, where human volunteers answer a series of hierarchical questions about each galaxy's visual properties. We use this dataset as a testbed to evaluate relative performance across models in capturing key features of galaxies, rather than as a comprehensive benchmark. For cost-effective evaluation, we curate a test set of 64 images meeting the criteria: (i) each question having strong consensus ($>70\%$ agreement) among sufficient annotators (>10), and (ii) diversity, enforced by limiting any identical classification to at most five examples. This size enables multiple evaluation runs across 13 different VLMs while keeping API costs manageable (with GPT-4.5, the most expensive model, costing $9 per evaluation) and provides sufficient statistical power to identify performance tiers.

We conduct two forms of evaluation. First, we directly prompt VLMs to answer the same questions that human annotators answered. Second, we prompt VLMs to generate free-form image descriptions, since these descriptions will serve as training data for our contrastive model, which we subsequently evaluate using an LLM-judge to extract answers to the annotation questions from these descriptions. Specifically, we use Gemini-2.5-Flash [`gemini-2.5-flash-preview-05-20`] [Comanici et al., 2025] to extract decision tree answers from the generated descriptions, allowing responses of "not stated" when information is absent which is scored as incorrect. As seen in Figure 1, accuracy

generally trades off with cost, forming a Pareto frontier. We choose GPT-4.1-mini for mass captioning since it offers the strongest per-question accuracy ($50.1 \pm 3.3\%$) within our budget.

2.2 Caption generation for scientific images

To construct our training dataset, we sample 300,000 galaxy images brighter than 19 magnitudes in r-band. We keep the selection simple, a single brightness cut to remove faint sources, and we do not build the training set from curated rare-object catalogs (e.g., Galaxy Zoo, strong lens lists), so the model stays general rather than biased toward any one target class. We split our sampling across telescopes to demonstrate that our search can work across different imaging modalities, 120,000 from HSC (PDR3 Wide [Aihara et al., 2022]) and 180,000 from Legacy Survey (DR10 South [Dey et al., 2019]). This data is sourced from the Multi Modal Universe (MMU) dataset, which has been pre-processed to remove non-galaxy sources and low quality images [Angeloudi et al., 2024]. The final dataset consists of 160×160 pixel cutouts with 5 channels for HSC and 4 channels for Legacy Survey. We turn each multi-band cutout into a color image for the VLMs by putting the bands on the same scale, mapping (z,r,g) bands to (R,G,B), and applying a standard transformation for galaxy images from Lupton et al. [2004] so faint features show up without blowing out bright features. Using the optimal prompt–model configuration identified through our evaluation framework (i.e., `gpt-4.1-mini` with the prompt in Appendix Fig. 4), we generate descriptions for each image. These textual descriptions are subsequently embedded using OpenAI's `text-embedding-3-large` [OpenAI, 2024b]. After excluding overlap with downstream benchmarks, 255,948 galaxies receive embedded descriptions and are used for training, with generation and embedding costs totaling ~$150.

2.3 Learning semantic embeddings through contrastive alignment

These generated embeddings already enable semantic search, however writing descriptions for all 140 million images in the Legacy Survey and HSC MMU catalogs would be prohibitively expensive. To address this, we predict the semantic embeddings from images directly. We use the pre-trained AION as our image encoder, an astronomy foundation model pretrained via masked modeling on over 200 million observations that achieves state-of-the-art performance on galaxy image similarity search [Parker et al., 2025]. The AION Transformer-based encoder-decoder architecture accommodates varying input formats, capable of handling both HSC images (5 bands) and Legacy Survey images (4 bands), and can be used for future telescope surveys. We use AION-1-Base, the 300M parameter variant with 768-dimensional embeddings, contrastively aligning its image encoder outputs with our text embeddings to create a shared semantic space. The encoder weights of AION-1-Base and `text-embedding-3-large` are not updated during training.

Our alignment architecture uses four-layer residual MLPs as projection heads for both image and text, learning a shared 1024-dimensional embedding space through a contrastive loss. Motivated by prior work suggesting that shorter captions may lead to higher performance in contrastive learning [Li et al., 2023], we compare training on the original multi-paragraph descriptions against those using single-sentence summaries generated by `gpt-4.1-nano` (see Appendix Fig. 5 for prompt). We train on a single NVIDIA A100 GPU using AdamW optimizer [Loshchilov and Hutter, 2019] with learning rate `1e-4`, weight decay 0.05, and cosine annealing. We leverage AION-1-Base encoder embeddings averaged over token-space (768-dimensional) for images and `text-embedding-3-large` (3072-dimensional) for text. This average pooling follows the AION embedding protocol for similarity search and morphology classification [Parker et al., 2025]. We use symmetric cross-entropy (InfoNCE [van den Oord et al., 2018]) as the contrastive loss.

2.4 Evaluating search capabilities

We adopt the retrieval evaluation protocol established for AION, assessing performance on three categories with varying rarity in the AION benchmark datasets of Legacy Survey images: spiral galaxies (24 622 in GZ-DECaLS dataset, making up 26% of the dataset being searched), mergers (726 in GZ-DECaLS dataset, 2%), and gravitational lenses (758 in HSC strong lens catalog cross-matched with Legacy Survey images, 0.1%). Full details about dataset construction is described in the Appendix A. Each image has relevance scores ranging from 0.0 to 1.0: for spirals and mergers, the score represents the fraction of human volunteers who identified each category; for lenses, the images

receive 1.0 if present in expert-curated lens catalogs and 0.0 otherwise. The rarity of gravitational lenses, with only a few thousand confident examples in the literature [Vujeva et al., 2025], coupled with their usefulness including constraining cosmology and measuring the mass and shape of dark matter halos makes them a challenging and scientifically valuable target.

To measure the search performance, we use the Discounted Cumulative Gain (DCG) to evaluate ranking quality, capturing both precision and the relative ordering of retrieved results. Normalizing DCG (nDCG) by the DCG with ideal ordering (IDCG). If r_i is the relevance label of a candidate ranked at position i, the discounted and normalized cumulative gain of the top-10 images are

$$\text{DCG@10} = \sum_{i=1}^{10} \frac{2^{r_i} - 1}{\log_2(i+1)}; \quad \text{nDCG@10} = \frac{\text{DCG@10}}{\text{IDCG@10}}$$

We evaluate against unsupervised baseline models from astronomy (AION-1 variants, Astro-CLIP [Parker et al., 2024], and Stein et al. [2021]), and computer vision (DINOv2 [Oquab et al., 2023]) that were trained via self-supervised learning and perform retrieval by computing cosine similarity between query image embeddings and candidate embeddings, where queries are selected as high-confidence examples (>90% volunteer agreement for morphologies, or presence in lens catalogs), with performance measured by averaging nDCG@10 across queries. For our model, we query with the text embeddings for: "visible spiral arms", "merging", and "gravitational lens".

2.5 Post-search VLM re-ranking

To improve the accuracy of a search, we experiment with a re-ranking step where a VLM examines the top-k retrieved images and assigns a score (1-10) based on the search query (e.g., "Does this galaxy image display signs of gravitational lensing? Rank 1-10.'), then reorder based on these scores. Specifically, for each retrieval task (spirals, mergers, and lenses) conducted in Section 2.4, we apply GPT-4.1 to score each of the top-1000 retrieved images and reorder them based on these scores. This automated verification step is inspired by current discovery pipelines, where classifiers produce thousands of false positives that require manual inspection by expert teams (e.g. Schuldt et al. [2025] for lenses).

To push this approach further, we investigate whether scaling test-time compute can improve retrieval results, analogous to how OpenAI scaled test-time compute to enhance logical reasoning capabilities [OpenAI, 2024a]. To evaluate this hypothesis, we construct a controlled test dataset using higher-resolution HSC images, the same data used for initial human identification, rather than lower-resolution Legacy Survey images, comprising 20,000 non-lenses and 200 confirmed gravitational lenses from published HSC strong-lens catalogs (see Appendix A).

We perform semantic search using AION-Search with the query "gravitational lens" and measure performance by counting confirmed lenses in the top-100 results after re-ranking the initial top-1000 retrievals, comparing against two baselines: the original AION-Search ranking and AION-1-B similarity search. Our experimental design investigates two dimensions of computational scaling: (1) model capacity, comparing GPT-4.1-nano, GPT-4.1-mini, and GPT-4.1 in ascending order of size and cost; and (2) inference-time strategy through n-sample averaging, where we generate n independent scores per image and average them for the final ranking.

3 Results

3.1 VLM image descriptions

We first evaluate whether current VLMs from OpenAI [OpenAI, 2025a,b] and Google [Comanici et al., 2025] can accurately describe the visual properties of galaxy images using our curated Galaxy Zoo test set (Section 2.1). We find that having o3 fill out the GalaxyZoo questions directly results in $23.6 \pm 1.0\%$ of images (over three runs) where the full decision tree paths matching that of the majority human consensus. While this is lower than the average volunteer (51.9 ± 2.4 %), it significantly exceeds random chance ($4.1 \pm 1.1\%$), demonstrating that VLMs can provide a noisy but informative signal for human labeling tasks. This average volunteer performance represents the mean probability across galaxies that any individual volunteer would match the majority, computed as the product of the vote fractions for each question and assuming the majority vote as ground truth.
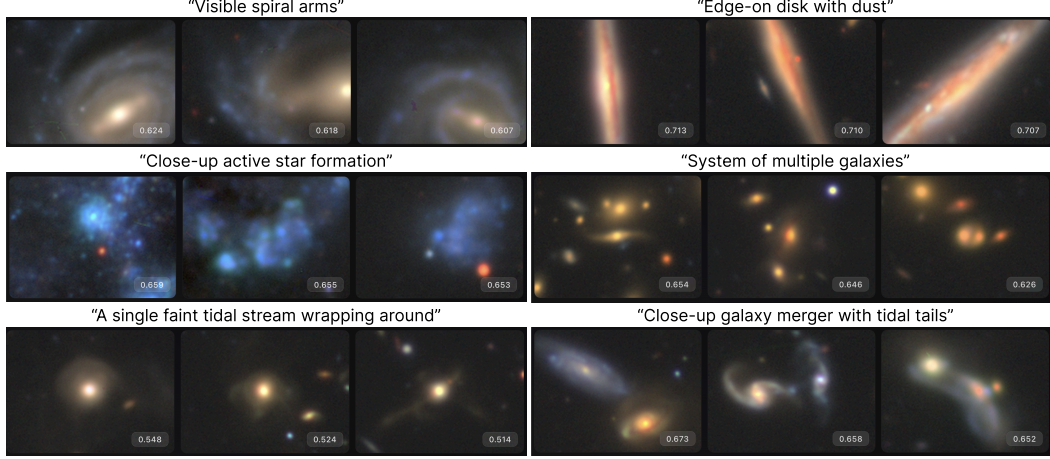
Figure 2: **AION-Search enables semantic retrieval of galaxies matching free-form natural language queries.** Top three retrieved images from HSC survey for each query demonstrate the system's ability to identify specific astronomical phenomena that would traditionally require volunteer labeling to catalog and train a supervised classifier. Cosine similarity scores are shown in the bottom right of each image

Testing our free-form description benchmark on 13 different models from Google and OpenAI with three runs each (with one run for GPT-4.5), we evaluated performance across different price points (Figure 1). We found GPT-4.1-mini to be optimal for our budget constraints, while models like o3 offer stronger performance but higher cost. We then use the benchmark to optimize a general prompt using AIDE [Jiang et al., 2025] which led to $50.1 \pm 3.3\%$ with GPT-4.1-mini, ensuring that the prompt is sufficiently general to capture galaxy features beyond Galaxy Zoo's fixed question set (see Appendix Fig. 4 for prompt). We use this setup for the CLIP training set caption generation. When the LLM-judge infers decision-tree answers from free-form captions generated by o3, only 12.5% of images have all the answers exactly matching the volunteer majority-vote, substantially lower than when o3 is directly provided the Galaxy Zoo questions.

3.2 Search performance

We demonstrate our system retrieving galaxies for complex queries like 'close-up active star formation' and 'a single faint tidal stream' in Figure 2, which are free-form searches that traditional supervised methods cannot perform and would typically require months-long volunteer labeling campaigns to enable.

Table 1 reports nDCG@10 scores for retrieval for spiral galaxies, mergers, and gravitational lenses. Our AION-Search model, trained on the summarized VLM-generated captions, consistently outperforms similarity-based baselines across all categories. For spirals, we observe an nDCG@10 of 0.941 compared to 0.643 for the strongest baseline (AION-1-L). For mergers, the model achieves 0.554 versus 0.384 (AION-1-XL), and for gravitational lenses 0.180 compared to 0.015 (AION-1-XL).

Consistent with prior work [Li et al., 2023], training on single-sentence summaries improves nDCG@10 by 0.143 (spirals), 0.084 (mergers), and 0.180 (lenses) relative to multi-paragraph descriptions, likely because shorter captions reduce noise and overfitting to caption-specific details. Baseline methods using visual similarity to find lenses return fewer than a single lens on average in their top-10 results when averaged across multiple queries. In contrast, our semantic approach retrieves two lenses in the top-10.

3.3 Re-ranking

On our dataset of 200 confirmed gravitational lenses hidden among 20,000 non-lenses, AION-Search (zero-shot) successfully retrieves 38 lenses within its top-1000 retrievals, demonstrating strong initial recall. While the top-100 precision matched AION-1-B one-shot similarity search (7 lenses), we

6

	Spirals	Mergers	Lenses
AION-1-B	0.632	0.281	0.012
AION-1-L	0.643	0.303	0.011
AION-1-XL	0.621	0.384	0.015
Parker et al. [2024]	0.602	0.248	0.006
Stein et al. [2021]	0.590	0.340	0.007
Oquab et al. [2023]	0.477	0.060	0.003
Random	0.263	0.037	0.000
AION-Search	0.941	0.554	0.180
AION-Search (re-rank)	0.992	0.678	0.290

Table 1: **Semantic search performance on astronomical images demonstrates superiority over similarity-based methods.** Quantitative comparison of retrieval performance (nDCG@10) across three astronomical phenomena of varying rarity. AION-Search, trained on VLM-generated captions, achieves strong performance using text queries ("visible spiral arms", "merging", "gravitational lens") compared to similarity-based baselines using image queries. Performance gains are most pronounced for rare phenomena. GPT-4.1 re-ranking of top-1000 results further improves performance across all categories. Baseline methods include various AION-1 model sizes (B, L, XL), AstroCLIP [Parker et al., 2024], astronomical self-supervised models [Stein et al., 2021], and DINOv2 [Oquab et al., 2023]. Retrieval scores for randomly shuffled data are shown for reference.

find that VLM re-ranking can greatly improve performance. By applying GPT-4.1 with 5-sample averaging to re-rank the initial 1000 retrievals, we nearly doubled the number of gravitational lenses in the top-100, increasing from 7 to 13.

Furthermore, our results reveal a relationship between test-time compute in re-ranking and retrieval performance for rare astronomical phenomena (Figure 3). As we scale from GPT-4.1-nano through GPT-4.1, and from single evaluations to 5-sample averaging, we increase the number of retrieved lenses.

4 Discussion & Conclusion

Our results demonstrate that VLMs can understand galaxy images well enough to allow for 1) generating descriptions for semantic search and 2) scoring an image to enable effective reranking of search results. Although VLM descriptions contain hallucinations (e.g. saying a galaxy has two spiral arms instead of three (Fig. 1)), these imperfect annotations nonetheless provide an effective training signal for contrastive learning. Our resulting semantic search model achieves strong zero-shot performance with text queries for spiral galaxies, mergers, and gravitational lenses, substantially outperforming similarity-based retrieval methods that rely on example images.

VLM-based re-ranking consistently improves retrieval performance across all tasks (Table 1): spirals improve from 0.941 to 0.992 nDCG@10 and mergers from 0.554 to 0.678, improvements even when AION-Search already achieves strong baseline performance. The most substantial gains occur for gravitational lenses, where re-ranking nearly doubles the number found in the top-100 (from 7 to 13), with performance scaling with test-time compute. This pattern makes re-ranking particularly valuable for discovering rare astronomical phenomena at scale, offering an alternative to resource-intensive manual labeling campaigns that typically require months to years of volunteer effort [Walmsley et al., 2023, 2025, Lintott et al., 2008].

Future work will focus on using AION-Search for astronomical discovery. Beyond gravitational lenses, the system could be applied to other scientifically valuable but under-explored phenomena, yet remain poorly cataloged due to their rarity with no dedicated classifiers or labeling campaigns. For example, finding galaxies with two tidal streams which can provide stronger constraints on the shape of dark matter halos [Nibauer et al., 2023].

Re-ranking could be extended through agentic approaches that provide models with image manipulation capabilities and domain-specific computational tools, such as astrophysical modeling packages (e.g. lens modeling with the `lenstronomy` package [Birrer et al., 2021]). Recent work demonstrates that tool-augmented multimodal reasoning substantially improves visual understanding tasks [Shao
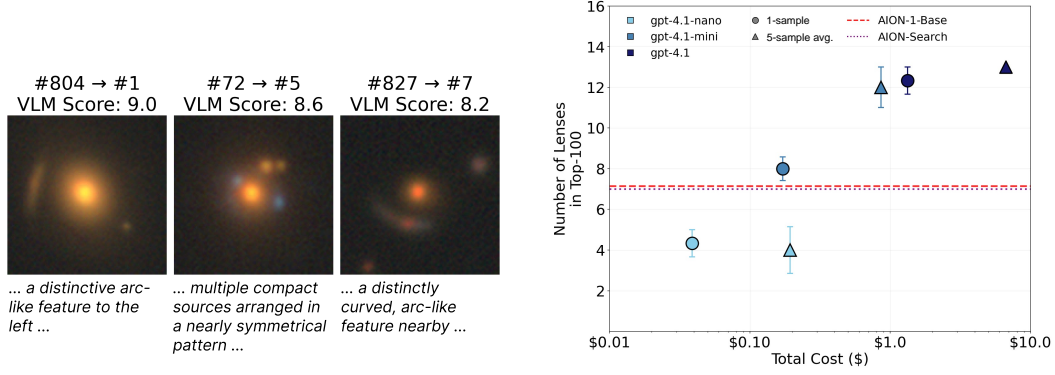
Figure 3: **VLM re-ranking improves gravitational lens discovery and performance scales with test-time compute budget.** *Left:* First three gravitational lenses identified by GPT-4.1 with 5-sample averaging for re-ranking, showing rank improvements from initial AION-Search results along with excerpts from the GPT-4.1 explanations (see Appendix Fig. 6 for prompt). The distinctive lensing features detected would traditionally require expert astronomers to visually inspect thousands of candidates [Schuldt et al., 2025]. *Right:* Performance scales with test-time compute: number of confirmed lenses in top-100 results after re-ranking the top-1000 AION-Search results. Baseline methods (AION similarity search and no re-ranking) finds only seven lenses. VLM re-ranking performance improves with both model size (`gpt-4.1-nano` to `gpt-4.1`) and inference-time compute via N-sample averaging, reaching 13 lenses with GPT-4.1 5-sample averaging. This demonstrates that spending more compute at search-time improves discovery of rare astronomical phenomena.

et al., 2024, Hu et al., 2024, Yang et al., 2023]. If a discovery is made, these analysis pipelines produced by the agent can be examined and verified by a human expert. Such methods could provide an effective way to scale test-time compute, beyond the N-sample averaging explored in this work.

The minimal training required for AION-Search demonstrates the quality of the AION representations. With LSST expected to image 20 billion galaxies by 2035 [Ivezić et al., 2019], full manual inspection becomes infeasible, making automated semantic search necessary. The AION architecture addresses this need through its unified latent space that handles multiple modalities (4-band Legacy Survey and 5-band HSC images as demonstrated here, plus spectroscopic measurements of stars and galaxies), allowing the same model to process new telescope data. Future implementations of AION-Search could incorporate time-series photometry or spectroscopic observations of stars and other astronomical sources, providing a single general-purpose semantic search engine for astronomy.

Several limitations constrain our approach. VLMs fail to detect subtle features, and our model inherits biases from GPT-4.1-mini. Our VLM benchmark captures limited visual information through its fixed question structure, providing insufficient assessment of description completeness. To get past this, one could evaluate VLM responses against expert-curated free-form descriptions containing all observable physical properties of an image. Furthermore, the astronomy performance of these VLMs likely benefits from pretraining on online image-text discourse and annotations (e.g., Galaxy Zoo), and VLM descriptions may be weaker in scientific domains without comparable corpora.

This approach could generalize to other scientific domains with large unlabeled image archives such as Earth observation, microscopy, and materials science, particularly where rare phenomena are valuable but costly to identify manually. As VLM capabilities generally improve over time, the gap between AI and human annotation quality will likely narrow.

We demonstrate that synthetic VLM descriptions provide a sufficient signal for semantic retrieval, significantly improving on self-supervised baselines and enabling open-vocabulary search. By training cross-modal encoders on ~250,000 synthetic captions, we enable natural language search over 140 million astronomical images without human annotation. A VLM-based re-ranking step further improves results, and its performance increases with test-time compute. We establish that large-scale scientific image archives can be made semantically searchable without human supervision, offering a generalizable framework for scientific image retrieval.

Acknowledgments and Disclosure of Funding

References

Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, et al. The Hyper Suprime-Cam SSP Survey: Overview and survey design. *Publications of the Astronomical Society of Japan*, 70:S4, January 2018. doi: 10.1093/pasj/psx066.

Hiroaki Aihara, Yusra AlSayyad, Makoto Ando, et al. Third data release of the hyper suprime-cam subaru strategic program. *Publications of the Astronomical Society of Japan*, 74(2):247–272, February 2022. ISSN 2053-051X. doi: 10.1093/pasj/psab122. URL http://dx.doi.org/10.1093/pasj/psab122.

Dario Amodei. Machines of loving grace: How ai could transform the world for the better, October 2024. URL https://www.darioamodei.com/essay/machines-of-loving-grace.

Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, Benjamin M. Boyd, David Chemaly, Brian Cherinka, Ioana Ciuca, Miles Cranmer, Aaron Do, Matthew Grayling, Erin Elizabeth Hayes, Tom Hehir, Shirley Ho, Marc Huertas-Company, Kartheik G. Iyer, Maja Jablonska, Francois Lanusse, Henry W. Leung, Kaisey Mandel, Juan Rafael Martínez-Galarza, Peter Melchior, Lucas Thibaut Meyer, Liam Holden Parker, Helen Qu, Jeff Shen, Michael J. Smith, Connor Stone, Mike Walmsley, and John F Wu. The multimodal universe: Enabling large-scale machine learning with 100TB of astronomical scientific data. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=EWm9zR5Qy1.

Simon Birrer, Anowar Shajib, Daniel Gilman, Aymeric Galan, Jelle Aalbers, Martin Millon, Robert Morgan, Giulia Pagano, Ji Park, Luca Teodori, Nicolas Tessore, Madison Ueland, Lyne Van de Vyvere, Sebastian Wagner-Carena, et al. lenstronomy II: A gravitational lensing software ecosystem. *The Journal of Open Source Software*, 6(62):3283, June 2021. doi: 10.21105/joss.03283.

Weizhi Chen, Jingbo Chen, Yupeng Deng, Jiansheng Chen, Yuman Feng, Zhihao Xi, Diyou Liu, Kai Li, and Yu Meng. Lrsclip: A vision-language foundation model for aligning remote sensing image with longer text, 2025. URL https://arxiv.org/abs/2503.19311.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Arjun Dey, David J. Schlegel, Dustin Lang, et al. Overview of the DESI Legacy Imaging Surveys. *The Astronomical Journal*, 157(5):168, May 2019. doi: 10.3847/1538-3881/ab089d.

J. Gonzalez, P. Holloway, T. Collett, A. Verma, K. Bechtol, P. Marshall, A. More, J. Acevedo Barroso, G. Cartwright, M. Martinez, T. Li, K. Rojas, S. Schuldt, S. Birrer, H. T. Diehl, R. Morgan, A. Drlica-Wagner, J. H. O'Donnell, E. Zaborowski, B. Nord, E. M. Baeten, L. C. Johnson, C. Macmillan, A. Roodman, A. Pieres, A. R. Walker, A. A. Plazas Malagón, A. Carnero Rosell, B. Santiago, B. Flaugher, D. Gruen, et al. Discovering strong gravitational lenses in the dark energy survey with interactive machine learning and crowd-sourced inspection with space warps, 2025. URL https://arxiv.org/abs/2501.15679.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training?, 2024. URL https://arxiv.org/abs/2402.01832.

Md Abul Hayat, George Stein, Peter Harrington, Zarija Lukić, and Mustafa Mustafa. Self-supervised Representation Learning for Astronomical Images. *The Astrophysical Journal Letters*, 911(2):L33, April 2021. doi: 10.3847/2041-8213/abf2c7.

Yiguo He, Junjie Zhu, Yiying Li, Xiaoyu Zhang, Chunping Qiu, Jun Wang, Qiangjuan Huang, and Ke Yang. Enhancing remote sensing vision-language models through mllm and llm-based high-quality image-text dataset generation, 2025. URL `https://arxiv.org/abs/2507.16716`.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=GNSMl1P5VR`.

Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

Anton T. Jaelani, Anupreeta More, Kenneth C. Wong, Kaiki T. Inoue, Dani C. Y. Chao, Premana W. Premadi, and Raoul Cañameras. Survey of gravitationally lensed objects in HSC imaging (SuGOHI) - X. Strong lens finding in the HSC-SSP using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 535(2):1625–1639, December 2024. doi: 10.1093/mnras/stae2442.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. Aide: Ai-driven exploration in the space of code, 2025. URL `https://arxiv.org/abs/2502.13138`.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities, 2024. URL `https://arxiv.org/abs/2412.10372`.

R. Laureijs, J. Amiaux, S. Arduini, J. L. Auguères, et al. Euclid Definition Study Report. *arXiv e-prints*, art. arXiv:1110.3193, October 2011. doi: 10.48550/arXiv.1110.3193.

Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for CLIP training. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=LMU2RNwdh2`.

Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, September 2008. doi: 10.1111/j.1365-2966.2008.13689.x.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing, 2024a. URL `https://arxiv.org/abs/2306.11029`.

Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. Clips: An enhanced clip framework for learning with synthetic captions, 2024b. URL `https://arxiv.org/abs/2411.16828`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Robert Lupton, Michael R. Blanton, George Fekete, David W. Hogg, Wil O'Mullane, Alex Szalay, and Nicholas Wherry. Preparing red-green-blue images from ccd data. *Publications of the Astronomical Society of the Pacific*, 116(816):133, feb 2004. doi: 10.1086/382245. URL `https://dx.doi.org/10.1086/382245`.

Siddharth Mishra-Sharma, YIDING SONG, and Jesse Thaler. PAPERCLIP: Associating astronomical observations and natural language with multi-modal models. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=8TdcXwfNRB`.

Jacob Nibauer, Ana Bonaca, and Kathryn V. Johnston. Constraining the gravitational potential from the projected morphology of extragalactic tidal streams. *The Astrophysical Journal*, 954(2):195, sep 2023. doi: 10.3847/1538-4357/ace9bc. URL `https://dx.doi.org/10.3847/1538-4357/ace9bc`.

OpenAI. Learning to reason with llms. `https://openai.com/index/learning-to-reason-with-llms`, 2024a. Accessed: 2025-08-12.

OpenAI. New embedding models and api updates, January 2024b. URL `https://openai.com/index/new-embedding-models-and-api-updates/`. New embedding models and API updates.

OpenAI. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`, 2025a. Accessed Aug 26, 2025.

OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, 2025b. URL `https://openai.com/index/o3-o4-mini-system-card/`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv e-prints*, art. arXiv:2304.07193, April 2023. doi: 10.48550/arXiv.2304.07193.

Liam Parker, Francois Lanusse, Siavash Golkar, et al. AstroCLIP: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, July 2024. doi: 10.1093/mnras/stae1450.

Liam Holden Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Thibaut Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Pierre Cornette, Keiya Hirashima, Geraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay, Mariel Pettee, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. AION-1: Omnimodal foundation model for astronomical sciences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=6gJ2ZykQ5W`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

S. Schuldt, R. Cañameras, Y. Shu, I. T. Andika, S. Bag, C. Grillo, A. Melo, S. H. Suyu, and S. Taubenberger. HOLISMOKES: XVI. Lens search in HSC-PDR3 with a neural network committee and post-processing for false-positive removal. *Astronomy & Astrophysics*, 699:A350, July 2025. doi: 10.1051/0004-6361/202554425.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=aXeiCbMFFJ`.

Zhenwei Shi and Zhengxia Zou. Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3623–3634, 2017. doi: 10.1109/TGRS.2017.2677464.

George Stein, Peter Harrington, Jacqueline Blaum, Tomislav Medan, and Zarija Lukic. Self-supervised similarity search for large scientific datasets. *arXiv e-prints*, art. arXiv:2110.13151, October 2021. doi: 10.48550/arXiv.2110.13151.

Dimitrios Tanoglidis and Bhuvnesh Jain. At first sight! zero-shot classification of astronomical images with large multimodal models. *Research Notes of the AAS*, 8(10):265, oct 2024. doi: 10.3847/2515-5172/ad887a. URL `https://doi.org/10.3847/2515-5172/ad887a`.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv e-prints*, art. arXiv:1807.03748, July 2018. doi: 10.48550/arXiv.1807.03748.

Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate E. Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 126500–126514. Curran Associates, Inc., 2024. doi: 10.52202/079017-4018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e4ad9c75f0d60ed75700f020adb3f705-Paper-Datasets_and_Benchmarks_Track.pdf.

L. Vujeva, R. K. L. Lo, J. M. Ezquiaga, and J. C. L. Chan. lenscat: a public and community-contributed catalogue of known strong gravitational lenses. *Philosophical Transactions of the Royal Society of London Series A*, 383(2294):20240168, April 2025. doi: 10.1098/rsta.2024.0168.

M. Walmsley, P. Holloway, N. E. P. Lines, K. Rojas, T. E. Collett, A. Verma, T. Li, J. W. Nightingale, G. Despali, S. Schuldt, R. Gavazzi, et al. Euclid quick data release (q1): The strong lensing discovery engine a – system overview and lens catalogue, 2025. URL https://arxiv.org/abs/2503.15324.

Mike Walmsley, Annette M N Ferguson, Robert G Mann, and Chris J Lintott. Identification of low surface brightness tidal features in galaxies using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 483(3):2968–2982, November 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty3232.

Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W. Willett, Steven Bamford, Lee S. Kelvin, Lucy Fortson, Yarin Gal, William Keel, Karen L. Masters, Vihang Mehta, Brooke D. Simmons, Rebecca Smethurst, Lewis Smith, Elisabeth M. Baeten, and Christine Macmillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3): 3966–3988, January 2022. doi: 10.1093/mnras/stab2093.

Mike Walmsley, Tobias Géron, Sandor Kruk, Anna M M Scaife, Chris Lintott, Karen L Masters, James M Dawson, Hugh Dickinson, Lucy Fortson, Izzy L Garland, Kameswara Mantha, David O'Ryan, Jürgen Popp, Brooke Simmons, Elisabeth M Baeten, and Christine Macmillan. Galaxy zoo desi: Detailed morphology measurements for 8.7m galaxies in the desi legacy imaging surveys. *Monthly Notices of the Royal Astronomical Society*, 526(3):4768–4786, 09 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad2919. URL https://doi.org/10.1093/mnras/stad2919.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023. URL https://arxiv.org/abs/2303.11381.

Z. Yuan, Z. Xiong, L. Mou, and X. X. Zhu. Chatearthnet: a global-scale image–text dataset empowering vision–language geo-foundation models. *Earth System Science Data*, 17(3):1245–1263, 2025. doi: 10.5194/essd-17-1245-2025. URL https://essd.copernicus.org/articles/17/1245/2025/.

Sharaf Zaman, Michael J. Smith, Pranav Khetarpal, Rishabh Chakrabarty, Michele Ginolfi, Marc Huertas-Company, Maja Jabłońska, Sandor Kruk, Matthieu Le Lain, Sergio José Rodríguez Méndez, and Dimitrios Tanoglidis. Astrollava: towards the unification of astronomical data and natural language, 2025. URL https://arxiv.org/abs/2504.08583.

Xinsong Zhang, Yarong Zeng, Xinting Huang, Hu Hu, Runquan Xie, Han Hu, and Zhanhui Kang. Low-hallucination synthetic captions for large-scale vision-language model pre-training, 2025. URL https://arxiv.org/abs/2504.13123.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions, 2024. URL https://arxiv.org/abs/2403.17007.

13

A Retrieval benchmark details

All retrieval experiments use Legacy Survey images [Dey et al., 2019] following the AION retrieval protocol in Section 7.3 in Parker et al. [2025], restated here for comparability.

Galaxy Zoo-DECaLS is a citizen-science project where volunteers label galaxy morphology [Walmsley et al., 2022]. We keep galaxies with at least three volunteer votes and cross-match with Legacy Survey South resulting in 171,000 galaxies. For evaluation, each candidate image gets a relevance score equal to the fraction of volunteers who chose that class ($r \in [0, 1]$).

We reproduce the AION parent sample for strong gravitational lenses by first cross-matching the Legacy Survey and HSC datasets to approximately reproduce the sample used in the HSC strong lensing searches [Jaelani et al., 2024]. Then we require: $0.2 \leq z_{\mathrm{phot}} \leq 1.2$ (photometric redshift, i.e. an estimated distance from broadband colors), $M_\star > 5 \times 10^{10} \, M_\odot$ (stellar mass, i.e. the total mass in stars), and star formation rate per unit stellar mass less than $10^{-10} \, \mathrm{yr}^{-1}$. We then cross-match to published strong-lens catalogs as listed in Parker et al. [2025] Section 7.3 and assign relevance $r = 1.0$ to cataloged lenses and $r = 0.0$ otherwise, yielding 758 lenses.

Since lenses are extremely rare in this dataset, nDCG@10 is inherently noisy. Unlike the similarity-based baselines in Table 1, which can vary image queries per task, our lens experiment use a single text query, "gravitational lens", so we cannot vary queries to estimate variability. Instead, we randomly partition the lens evaluation set into 10 disjoint subsets and recompute nDCG@10 on each. Doing so results in nDCG@10 scores ranging from 0.00 to 0.29, with a mean of 0.123 ± 0.114 (standard deviation), consistent with the overall score of 0.180 but highlighting the high variance induced by the small number of positives. For comparison, performing the same 10-fold evaluation for spirals and mergers yields mean nDCG@10 values of 0.973 ± 0.033 (range $[0.889, 1.000]$) and 0.456 ± 0.077 (range $[0.310, 0.620]$), respectively. To remain comparable with the original AION retrieval benchmark [Parker et al., 2025], Table 1 reports nDCG@10 values computed on the full evaluation set for all methods.

B Prompts

Training set generation prompt

Analyze this astronomical image and provide a detailed description, assuming the reader has domain expertise and the description will be reviewed by astrophysicists.

1. Morphology: Describe a detailed morphological classification of the main object(s) (e.g., spiral, elliptical, irregular, merging system, etc.).

2. Prominent Features & Characteristics: Detail specific astronomical features observed, for example: spiral arms, bars, dust lanes, star-forming regions, tidal features, merger remnants, foreground stars, lensing effects, etc. Mention these features if and only if they are present in the image. If they are not present, do not mention them. The features mentioned here were just examples, state any feature you believe astrophysicists would be interested in. For each feature, describe its key characteristics (e.g., morphology, extent, brightness, orientation, interaction with other components, physical phenomena occurring, etc.).

3. Additional Context & Interpretation: Include any further observational details or astrophysical interpretations.

Only focus on the center object(s) in the image. Keep your response under 300 words.

Figure 4: Prompt for generating free-form astronomical image descriptions with VLMs, designed to generate comprehensive general descriptions about the physical phenomena in the image rather than answers to specific questions (as in Galaxy Zoo's decision tree).

Please summarize the following description into a single sentence CLIP query:

```
<original_description>
{{original_description}}
</original_description>
```

Only output the summary without any additional text.
Do not use any of the following words: 'no', 'not', 'without', 'absence', 'lack of', 'no obvious', 'no signs', 'absence of' or any other negation words or phrases.
Ignore any phrases containing negation from the original description containing these words.

Figure 5: Prompt for generating single-sentence summaries from multi-paragraph VLM descriptions using `gpt-4.1-nano`. The prompt leads to shorter descriptions that exclude negations to improve contrastive learning performance [Li et al., 2023].

Reranking prompt

Does this galaxy image display signs of gravitational lensing? Rank 1-10 where 10 means you are entirely sure there are signs of gravitational lensing and 1 being you are entirely sure there are no signs of gravitational lensing.

Figure 6: Prompt used in the VLM re-ranking stage for the experiment in Section 2.5: for each retrieved galaxy image, the model assigns a 1–10 score indicating confidence in gravitational lensing.