AstroCLIMB: Astronomy Citation Linking from Illustrations - a Multimodal Benchmark and Overview of the WASP 2026 Shared Task

Felix Grezes[1], **Atilla Kaan Alkan**[1], **Kelly Lockhart**[1], **Alberto Accomazzi**[1], **Ethan Seefried**[2] **and Tirthankar Ghosal**[2]
[1]Center for Astrophysics, Harvard & Smithsonian, USA
[2]Oak Ridge National Laboratory, USA
[1]{felix.grezes,atilla.alkan, kelly.lockhart,aaccomazzi}@sao.si.edu
[2]{seefriedej, ghosalt}@ornl.gov

Abstract

Scientists rely on figures to share their discoveries, but this makes the information contained in the figures difficult to parse, archive, and search. Multimodal neural-network models, designed to handle multiple types of data, promise to extract this information, but have not yet been widely tested and adopted by the astronomy community. In partnership with astroexplorer.org, we propose a novel dataset with over 100K figures and captions, and an associated task as a benchmark to evaluate a model's multimodal capabilities. The task is to reconstruct the citation graph of astronomy papers solely from their figures and captions. AstroCLIMB was the shared task for WASP 2026. It attracted 15 participants on the open competition platform Kaggle, of which 10 submitted system papers. Results show that zero-shot methods and general purpose metrics underperform compared to finetuned methods, indicating that scientific figures are still hard to understand for currently available multimodal models. The dataset, benchmark task, and associated code are publicly available online.

1 Introduction

The Science Explorer (SciX) (Accomazzi and SciX Team, 2026), formerly the Astrophysics Data System (Kurtz et al., 2000), is a digital library portal for researchers in astronomy, Earth science, heliophysics, physics, and planetary science, with the mission of facilitating and increasing research efficiency. Currently, SciX processes the textual information present in the papers it ingests. Author names, keywords, title and abstract, references, and many more are harvested, cleaned, categorized, and finally formatted into a rich data schema, allowing SciX users to search and filter for papers most useful to them.

However, missing from that harvest is the information present in the various figures in the papers. Unlike text which is linear in nature, these figures allow researchers to parse and absorb new scientific findings holistically. In astronomy, the image itself is often the valuable information, and any text attempting to describe it would be incomplete; for example a notable star formation in a specific location in the sky. Ideally, scientists could use natural language to describe figures to search and find papers of interest, even if the query doesn't strictly relate to the textual information in the caption or body of the papers. An example of query that is not currently possible is:

```
Find JWST papers with a plot showing a red to blue
heatmap and black contour lines.
```

Harvesting the information from figures is non-trivial: classical OCR systems can fail when the text is rotated at different angles over different color shades in the same figure; multiple plots are often combined into one figure; relevant information might be missing from the direct caption. Image 1 shows an example of these difficulties.

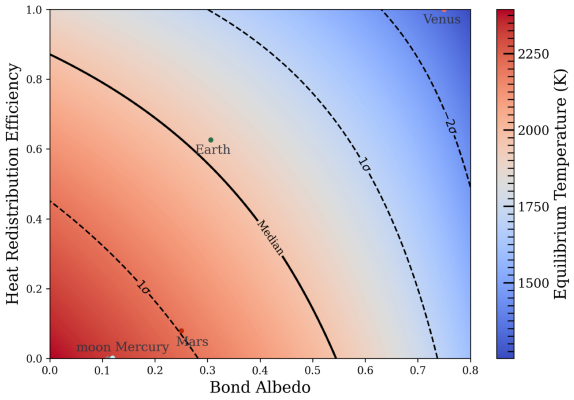


Figure 1: Sample astronomy image. To extract relevant information, a system needs to be able to read text at 5 different angles over continuous shades of blue and red, translate the XY and color coordinates of the Earth position to numbers, and understand that this is likely a JWST paper based on the context of the image and caption. Image courtesy of astroexplorer.org, original by Smith et al. (2026)

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 10–16
November 10, 2026 ©2026 Association for Computational Linguistics

Since the breakout of CLIP (Radford et al., 2021), neural network models that harness the technology of transformers to process multiple modes of information (text, image, audio, etc...) have proliferated. Nearly all frontier models offer multimodal variants, both closed (Gemini Team, 2025; Anthropic Team, 2026; OpenAI Team, 2025) and open source (Qwen Team, 2026; DeepSeek Team, 2026; Gemma Team, 2026) just to name a few. While these models are capable of embedding images and text in a shared vector space, and showcase that these embeddings preserve some of the shared meaning between images and associated text, they have not been rigorously tested on astronomy scientific figures.

An potential approximation for semantic understanding comes from the scientific citation graph, in which authors connect papers by their references. Similarly to Word2Vec (Mikolov et al., 2013) or more modern semantic preserving embedding methods, a thesis of this paper is that a model that can approximate the citation graph of the astronomy papers citation graph from its figures is a model that has some understanding of these.

To test this thesis, we contribute a novel dataset and benchmark: the Astronomy Citation Linking from Illustrations - a Multimodal Benchmark, or AstroCLIMB.

Contributions

- An open dataset consisting of over 100K figures from over 10K recent open access astronomy papers, along with their associated caption, links to citing and referring papers, and metadata, hosted publicly on Huggingface[1].
- A shared task platform to benchmark multimodal models, hosted publicly on Kaggle[2].
- A baseline analysis, evaluating off-the-shelf multimodal LLMs on the task.
- A description of the systems developed by participants to WASP@IJCNLP-AACL 2026[3]

2 Dataset Description

Since its launch in 2013, the Astronomy Image Explorer (AIE, astroexplorer.org) has archived hundreds of thousands of high-resolution images published in the open-access journals of the American Astronomical Society. Alongside the image, metadata containing information about author, journal, citation, format, and keywords is also stored. In partnership with the AIE, we harvested images and associated information for years ranging from 2022 to 2025. After filtering for images that may not be open-access, the final full dataset size is 102629 pairs of image and captions, coming from 12671 unique papers, of which 10843 have citation or reference links to other papers in the dataset. Figure 2 shows the distributions of reference and citation count per paper.

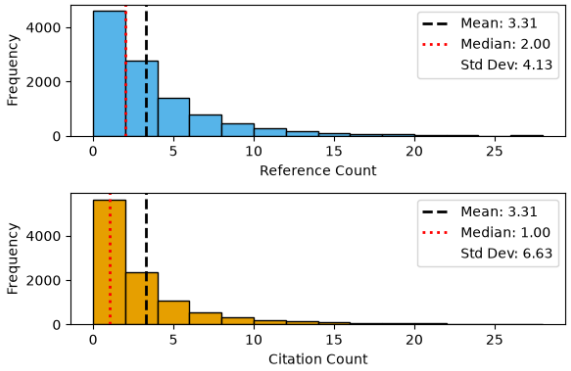


Figure 2: Distributions of reference and citation count for the papers within AstroCLIMB.

The dataset was split into 94233 rows for training and 8396 for validation. It should be noted that because AstroCLIMB is a benchmark rather than an inherently useful task, the 'training' mentioned here is about simple statistics computations rather than fully finetuning model weights.

2.1 Technical Details

The full dataset splits are Huggingface `Dataset` objects. The training split look as follows:

```
Dataset({
    features: ['image',
               'UUID',
               'Image ID',
               'Paper DOI',
               'Paper Title',
               'Image Caption',
               'Image Authors',
               'References DOIs',
               'Citing DOIs'],
    num_rows: 94233
})
```

The first row provides an example:

```
train_dataset[0] == {
'image':    <PIL.PngImagePlugin.PngImageFile
            image mode=RGB size=1000x1066>,
'UUID':     '10.3847.1538-4357.ac983d_apj[...]',
'Image ID': 'apjac983df5',
'Paper DOI': '10.3847/1538-4357/ac983d',
```

[1] huggingface.co/datasets/adsabs/AstroCLIMB
[2] kaggle.com/competitions/astroclimb
[3] ui.adsabs.harvard.edu/WIESP/2026/

```
'Paper Title': 'The Radius of PSR J0740+6[...]',
'Image Caption': 'Effect of the SW backgr[...]',
'Image Authors': 'Authors: Tuomo Salmi, S[...]',
'References DOIs': ['10.3847/1538-3881/ac4ae6'],
'Citing DOIs': ['10.3847/2041-8213/ace5b2',
    '10.3847/2041-8213/ad5beb',
    '10.3847/1538-4357/ae145d',
    '10.3847/1538-4357/ad81d2',
    '10.3847/1538-4357/acf9a0',
    '10.3847/1538-4357/ad7255',
    '10.3847/2041-8213/ad9f3c',
    '10.3847/2041-8213/ad5f02',
    '10.3847/1538-4357/acf49d',
    '10.3847/2041-8213/ad5a6f']}
```


Going through each of these:

- `'image': <PIL.PngImagePlugin.PngImageFile image mode=RGB size=1000x1066` is the image data for this figure.
- `Image Caption` is the other important feature of this row. It is the text caption associated with this figure.
- `References DOIs` and `Citing DOIs` are lists that link to other DOIs which are either referenced by or cite the current row's DOI. Note that are only DOIs listed elsewhere in the dataset, not every reference and citation in the full paper.
- `'UUID': '10.3847.1538-4357.ac983d_apjac983df5_8'` serves as a unique identifier for the row.
- `'Paper DOI': '10.3847/1538-4357/ac983d'` is the Digital Object Identifier (DOI) of the paper from which the figure was extracted.
- `'Image ID': 'apjac983df5'` is an identifier to distinguish from other figures in the same paper.
- `Paper`, `Title`, and `Image Authors` provide potentially useful metadata.

3 Shared Task Description

The goal of the shared task for WASP 2026 is multi-class, single-label classification. The inputs are pairs of objects. It could be two figures, a figure and a caption, or two captions. The possible outputs are:

- `same_paper` if the two objects originated from the same scientific paper (same DOI).
- `related_papers` if the two objects originated from related papers (one object's DOI cites the DOIs of the second object).
- `same_figure` if the two objects are actually from the same scientific figure (only for figure+caption pairs).
- `unrelated_papers` if the two objects originated from unrelated papers (none of the previous cases).

Input pairs only have one label and the relations between objects is symmetrical. This dataset is intended to be more of a benchmark for multimodal models, rather than a meaningful task in itself. As such, participants were not necessarily expected to fully train large models solely for AstroCLIMB. Basic statistics or incorporation into a larger multimodal training process seemed more appropriate.

From the `References DOIs` and `Citing DOIs` in the Huggingface dataset columns we can build the version of the dataset used on Kaggle for automatic benchmarking. However with over 100K rows in the full dataset, there are over 5B possible pairs of objects. The version of the dataset on Kaggle contains a small, balanced sample of those pairs. Both `train.csv` and `test.csv` contain 10K rows with 1K sample for each combination possible of image+image, image+caption, caption+caption input features with `same_figure`, `same_paper`, `related_papers` and `unrelated_papers` output labels. Submissions were evaluated using macro-F1 score over the four target classes.

4 Baseline Experiments

To establish reference performance for AstroCLIMB, we evaluated two open-weight vision-language models (VLMs): Llama-3.2-11B-Vision-Instruct (11B parameters) (Grattafiori et al., 2024) and Gemma-4-31B-it (31B parameters) (Gemma Team, 2026). Both models were evaluated zero-shot, without task-specific fine-tuning or in-context examples. These experiments provide a reference for assessing how general-purpose multimodal models perform on scientific figure and citation relationship classification.

4.1 Experimental Setup

Each pair was presented through a chat-formatted prompt defining the four relationship classes: `same_figure`, `same_paper`, `related_papers`, and `unrelated_papers`. Objects were provided in their original modalities, either as images or text,

Metric	Llama 3.2 11B	Gemma 4 31B
Overall macro-F1	0.2145	**0.4890**
Per-class F1		
Same figure	0.11	**0.87**
Same paper	0.42	**0.63**
Related papers	0.32	**0.46**
Unrelated papers	0.01	0.00
Modality-specific macro-F1		
Text + image	0.1662	**0.4212**
Image + image	0.1463	**0.3765**
Text + text	0.3426	**0.3891**

Table 1: Zero-shot VLM performance on AstroCLIMB. Overall macro-F1 is reported from the official Kaggle public leaderboard. Modality-specific macro-F1 is computed over structurally valid classes for each pair type: four classes for text-image pairs and three for image-image and text-text pairs.

followed by a `Label:` cue. The prompt explicitly specified that `same_figure` is only applicable to an image and its corresponding caption.

Both models were loaded using Hugging Face Transformers and evaluated in bfloat16 precision on NVIDIA A100 80GB GPUs. Inference was performed independently for each pair using greedy decoding with a maximum of 10 generated tokens. The first recognized class label was taken as the prediction, defaulting to `unrelated_papers` if no valid label was generated. No calibration, class balancing, or task-specific post-processing was applied.

We report macro-averaged F1 using the official Kaggle public leaderboard, along with locally computed class-level and modality-specific metrics.

4.2 Baseline Performance

Table 1 presents the results of both zero-shot baselines. Gemma substantially outperformed Llama, achieving a macro-F1 of 0.4890 compared with 0.2145. However, performance varied considerably across relationship categories and input modalities.

Gemma performed particularly well on direct figure-caption correspondence, achieving an F1 score of 0.87 for `same_figure`, with recall of 1.00 and precision of 0.77. In contrast, Llama achieved an F1 of only 0.11 for this category. Both models performed less effectively when distinguishing relationships between different scientific papers.

Modality-specific results reveal additional differences. Llama performed substantially better on text-text pairs (macro-F1 0.3426) than image-image pairs (0.1463). Gemma exhibited a smaller difference between these configurations, achieving 0.3891 and 0.3765, respectively. These results suggest that the models differ not only in overall classification performance but also in their handling of visual inputs.

Model	Free generation	Forced choice
Llama 3.2 11B	0.5%	100.0%
Gemma 4 31B	0.0%	68.9%

Table 2: Percentage of predictions assigned to `unrelated_papers` under greedy generation (full test set, $n = 10{,}000$) and forced-choice label scoring (stratified sample, $n = 450$). The two columns are descriptive and are not directly matched-sample comparisons.

We additionally examined predictions that violated the task's structural constraints. Although `same_figure` is only valid for image-caption pairs, Llama assigned this label to 1,064 of 3,000 image-image pairs (35.5%) and 151 of 3,000 text-text pairs (5.0%). Gemma made the same error on only 2 image-image pairs (0.07%) and 10 text-text pairs (0.3%). This substantial difference indicates that Gemma more reliably follows the modality-dependent definition of direct figure-caption correspondence.

4.3 Analysis of Prediction Behavior

Despite their differences in overall performance, both models exhibited a striking tendency to avoid predicting `unrelated_papers` during greedy generation. Although this category accounts for 30% of the evaluation set, Llama predicted it only 48 times out of 10,000 pairs, while Gemma never predicted it. Instead, Llama predominantly predicted `same_paper` (5,971 predictions), whereas Gemma favored `related_papers` (5,879 predictions).

To investigate whether this behavior was specific to the generation procedure, we conducted an additional forced-choice diagnostic on a stratified sample of 450 test pairs, containing 150 examples of each input configuration. Rather than generating labels freely, we computed the teacher-forced, length-normalized log-likelihood of each candidate label as a continuation of the original prompt. The label with the highest average token log-likelihood was selected.

The forced-choice results differ dramatically from free generation. Llama selected `unrelated_papers` for all 450 sampled pairs, while Gemma selected it for 310 pairs (68.9%). Neither model selected `related_papers` under

forced-choice scoring in this sample, despite its frequent occurrence in free-generation predictions.

This reversal indicates that the observed absence of `unrelated_papers` predictions is strongly dependent on the output selection procedure, rather than demonstrating that the models cannot assign likelihood to this relationship. However, forced-choice scoring introduces its own biases, and a high likelihood for a particular label does not necessarily indicate correct relational reasoning. Further experiments would be needed to determine whether calibrated or constrained prediction improves classification accuracy.

Overall, these results highlight two challenges for general-purpose VLMs on AstroCLIMB. First, the models differ substantially in their ability to respect modality-dependent relationship definitions. Second, their predicted relationship distributions are highly sensitive to how candidate labels are selected, suggesting that evaluation protocols and output calibration are important considerations when using generative VLMs for scientific relationship classification.

5 Participant Systems

The AstroCLIMB shared task attracted 15 participating teams on Kaggle, of which 10 submitted system papers to WASP 2026.

- Zhu and Tan (2026) test a hosted multimodal judge under a fixed, auditable multi-view protocol. The experiments show that this constrained approach substantially outperforms a content-only lexical and image-statistics baseline on local validation.

- Bang et al. (2026) combine text-centered OCR representations, citation-graph pretraining, and fine-tuned large language models. Their approach demonstrates that ensembling complementary architectures overcomes fine-grained confusions between same-paper and citation-linked pairs.

- Fedosov (2026) evaluates a factorized gradient-boosted classifier that merges frozen semantic embeddings with explicit visual and textual features. The results reveal that rich symmetric pair features provide significant performance gains over raw cosine similarities, whereas nested model selection yields no aggregate benefit.

- Hoque (2026) applies parameter-efficient adapters and probability ensembling to a vision-language model. The author finds that model averaging and validity constraints improve test performance, while broader visual adaptation and auxiliary training provide no further improvements.

- Sourav et al. (2026) evaluate modality-conditioned ensembles and zero-shot cross-encoder reranking. The experiments show that representation quality serves as the primary performance driver, whereas extensive feature engineering and distant supervision yield minimal gains.

- Quezada Ramírez (2026) evaluates public scientific-figure adaptation for a vision-language model. The author shows that pretraining on unlabeled public pairs improves downstream validation performance, whereas alternative generative and reinforcement-learning setups yield no additional benefit.

- Ghimire et al. (2026) evaluate parameter-efficient multimodal fine-tuning for vision-language models. The experiments show that low-rank adaptation substantially improves performance on challenging citation-relation classes, while demonstrating that unadapted few-shot reasoning yields no additional gains.

- Ramalakshmi et al. (2026) test parameter-efficient vision-language models across diverse modalities. Results show that fine-tuned generative backbone models achieve strong performance, while highlighting that textual semantics provide more reliable citation signals than visual features alone.

- Mane et al. (2026) build a multimodal feature-based pipeline combining textual, visual, semantic, and domain-oriented signals with a gradient-boosted classifier. The experiments show that complementary representations provide useful predictive features, while demonstrating that distinguishing related papers from other relationship types remains a persistent challenge.

- Maiti and Chowdhury (2026) evaluate a modality-specialized and relation-aware pipeline combining custom encoders with decomposed correspondence gates. They show

Team / Model	Macro-F1
Quezada Ramírez (2026)	0.75
Bang et al. (2026)	0.74
Ramalakshmi et al. (2026)	0.74
Hoque (2026)	0.74
Sourav et al. (2026)	0.69
Ghimire et al. (2026)	0.64
Maiti and Chowdhury (2026)	0.61
Gemma-4-31B-it (Zero-shot)	0.49
Fedosov (2026)	0.48
Mane et al. (2026)	0.48
Zhu and Tan (2026)	0.40
Random Baseline	0.25
Llama-3.2-11B-Vision (Zero-shot)	0.21

Table 3: AstroCLIMB shared-task leaderboard results and zero-shot VLM baselines, reported using macro-F1. Scores are rounded to two decimal places.

that separating exact match prediction from residual paper relations substantially improves performance, while identifying cross-paper citation links remains a persistent challenge.

We report the results of the participating teams in table 3.

6 Conclusion and Future Directions

In this work, we introduced AstroCLIMB, a novel multimodal dataset and benchmark task designed to evaluate the capability of neural network models to parse and understand scientific figures within astronomy literature. By framing the task around reconstructing citation and reference relationships from figures and captions, we provided a rigorous proxy for semantic comprehension of visual data in digital library portals like the Science Explorer (SciX).

Our baseline evaluations with off-the-shelf vision-language models—such as Llama-3.2-11B-Vision-Instruct and Gemma-4-31B-it—demonstrated that zero-shot general-purpose models struggle significantly with scientific illustrations, exhibiting sensitivity to output selection procedures and limitations in respecting modality-dependent relationship definitions. However, the response of the research community through the WASP 2026 shared task highlights the viability of addressing these challenges. With 15 participating teams on Kaggle and 10 submitted system papers, participants showed that strategies such as domain-adapted fine-tuning, parameter-efficient adapters (e.g., QLoRA), modality-conditioned ensembling, and factorized gradient-boosted classifiers can substantially boost performance, raising top macro-F1 scores to 0.75.

Despite these advances, several challenges remain. Participant systems and baseline analyses consistently identified that distinguishing cross-paper citation links and fine-grained relational categories (such as related papers versus same paper) remains a persistent hurdle. We hope that future works will use AstroCLIMB to build systems that enable natural language visual search and intelligent artifact discovery for digital libraries.

7 Ethics Statement

The authors of this paper follow principles of transparency and reproducibility. The dataset and code described are publicly available and open source, ensuring accessibility for verification and future research. Large language models were employed solely as baseline comparisons in our experiments, in a non-generative mode only, as classifiers. We acknowledge that LLMs may carry inherent biases present in their training data, and we have taken care to document these limitations in our analysis. The use of LLMs as baselines does not constitute endorsement of their outputs, but rather provides a standardized benchmark for evaluating our proposed methods. We are committed to responsible research practices and have considered the potential societal impacts of this work throughout the research process.

References

Alberto Accomazzi and SciX Team. 2026. The Science Explorer: An Open Science Discovery Platform in the Earth and Space Sciences. In *46th COSPAR Scientific Assembly. Held 1-9 August*, volume 46, page 3448.

Anthropic Team. 2026. The claude model family: Opus, sonnet, haiku. anthropic.com/system-cards. Model Card and Technical Specifications, archived URL: archive.ph/sVkGb.

My-Linh Bang, Phu-Quy Nguyen-Lam, Duy-Minh Dao-Sy, Trung-Kiet Huynh, Phu-Hoa Pham, and Chi-Nguyen Tran. 2026. HCMUS_beyonders at AstroCLIMB: Same paper or citing paper? linking astronomy figures with OCR, citation graphs, and LLMs. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

DeepSeek Team. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv e-prints*, arXiv:2606.19348. huggingface.co/deepseek-ai.

Egor Fedosov. 2026. GRAIL: Compact multimodal similarities for astronomy citation linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv e-prints*, arXiv:2507.06261.

Gemma Team. 2026. Gemma 4 technical report. *arXiv e-prints*, arXiv:2607.02770. huggingface.co/google.

Prajwal Ghimire, Shiv Ram Saud, Aashish Mahato, Sundeep Dawadi, Sunil Regmi, and Manoj Shakya. 2026. PSGS@2026: From figures to citation graphs: QLoRA based multimodal linking for AstroCLIMB. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Syed Mohaiminul Hoque. 2026. Parameter-efficient adaptation of Qwen3-VL for astronomy citation linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and Joyce M. Watson. 2000. The NASA Astrophysics Data System: Overview. , 143:41–59.

Bodhisatta Maiti and Debshree Chowdhury. 2026. Modality-specific and relation-aware modeling for scientific figure linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Raj Mane, Ananya Kalia, Aashna Shetty, Rutal Mahajan, and Preksha Pareek. 2026. Multimodal relationship classification of astronomy figures and captions. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, arXiv:1301.3781.

OpenAI Team. 2025. OpenAI GPT-5 System Card. *arXiv e-prints*, arXiv:2601.03267.

Irving Ernesto Quezada Ramírez. 2026. Public scientific-figure adaptation for AstroCLIMB with a 4B vision-language model. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Qwen Team. 2026. Qwen-image-2.0-rl technical report. *arXiv e-prints*, arXiv:2606.27608. huggingface.co/Qwen.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv e-prints*, arXiv:2103.00020.

K S Ramalakshmi, Shubham Sharma, Ashmani Kumar, and Ankush Chopra. 2026. Beyond visual similarity: Parameter-efficient vision–language models for scientific citation linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Cole Smith, Brandon Park Coy, Megan Weiner Mansfield, Rafael Luque, Anjali A. A. Piette, Qiao Xue, Michael Zhang, Jacob L. Bean, Madison Brady, Peter Gao, and et al. 2026. JWST MIRI reveals a potential atmosphere on the ultra-hot rocky planet TOI-431b. *arXiv e-prints*, arXiv:2609.03029.

Tonmoy Sarker Sourav, M. Tahsan Zahid, and Most. Atkia Farzana. 2026. Modality-conditioned ensembles and zero-shot cross-encoder reranking for relation classification between astronomy-paper objects. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Kewen Zhu and Junyuan Tan. 2026. Auditable multi-view hosted multimodal judging for AstroCLIMB. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.