Multimodal Feature Integration for Pairwise Relationship Classification in Astronomy Literatures

Raj Mane[1], Ananya Kalia[1], Aashna Shetty[1], Rutal Mahajan[2], Preksha Pareek[1]

[1]Department of Computer Engineering,
Thakur College of Engineering and Technology, Mumbai, India
[2]Department of Computer Science,
Utica University, USA

Abstract

In scientific articles, figures often convey experiments and observations, as well as physical intuition, to the readers. Yet, surface metadata alone can be insufficient to unpack how two scientific objects relate to each other. We explore this with AstroCLIMB, which infers relationships between pairs of figures and metafigures based on four classes: same figure, same paper, related papers and unrelated papers. We design a multimodal pairwise pipeline that consolidates the textual visual semantic, and domain-specific signal sources from object pairs. TF-IDF and SBERT text and image-based text-image feature pairings, combined with lexical overlap, entity, and instrument overlap, numerical overlap, and interaction features conditioned on methods, comprise the feature set fed into a weighted ensemble of LightGBM, XGBoost, and CatBoost trained with 313 engineered features on five-fold stratified cross-validation, optimizing for class-specific thresholds to produce a Macro-F1 of 0.5016 on out-of-fold data. Results show that yet stronger textual and multimodal counterpart representations exhibit complementary yet useful pairwise signals, although establishing related papers still remains a significant challenge.

1 Introduction

Scientific publications are multimodal documents, which include prose, figures, tables, and captions that communicate significant information (Kashyap et al., 2023). Figures communicate observations, model results, measurements, and comparisons and can be central to any scientific work.

Recently, there have been some work on scientific figure captioning, figure extraction, multimodal scientific understanding, scientific information retrieval, and citation prediction. AstroCLIMB considers a different task: instead of generating a caption, retrieving a document, or answering a question on a single figure, it classifies the type of relationship between a pair of scientific objects. AstroCLIMB offers four symmetric relationship types namely same figure, same paper, related papers, and unrelated papers, and attempts to recover, in a limited fashion, the citation graph of astronomy publications using figures and captions rather than relying solely on standard metadata. The method uses several visual textual semantic, and domain-specific signals in a single feature-based classification method.

It then generates pairwise features based on overlap in lexical units, TF-IDF and SBERT similarity, CLIP-based text and image representations, overlap in astronomical entity and instrument concepts, numerical similarity, length of objects, sentence-based features, and modals. The trained classification model is a weighted meta ensemble of LightGBM, XGBoost, and CatBoost classifiers, exploiting 313 engineered features among the evidence sources. The model is assessed by five-fold stratified cross-validation, with an out-of-fold Macro-F1 of 0.5016 after improvement of the decision-thresholds for each class.

2 Related work

2.1 Scientific Figure Understanding

Figure captioning for scientific figures has demonstrated that there is information in figures that might not be captured in author captions. SciCap established a large-scale benchmark for scientific figure captioning and demonstrated both the potential and the challenges of generating an informative scientific caption (Hsu et al., 2021). SciCap+ demonstrated that figure-mention paragraphs and OCR information can be used for captioning, and so the importance of surrounding text (Yang et al., 2024). Summarization-based captioning in the same way uses the paragraphs that mention figures (Huang et al., 2023). Scientific figure captioning with large multimodal models was then

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 216–223
November 10, 2026 ©2026 Association for Computational Linguistics

explored and shown to be a challenging problem (Hsu et al., 2026).

Another task that we will need is figure extraction from scientific documents. Figure extraction is needed for caption generation. FigEx has investigated extraction and alignment of subfigures and subcaptions (Song et al., 2025) while SciFig offers large-scale data for scientific figure classification and understanding (Anonymous, 2021). These are related to AstroCLIMB as figure extraction and multimodal understanding is required for relation discovery, but in fact their main focus is on extraction, captioning, or classification of figures, rather than predicting the relationship between two scientific figures.

2.2 Scientific Multimodal Learning

Multimodal ArXiv offers scientific large-scale image-text data and demonstrates how domain-adaptive training can cause scientific understanding, but there are still differences in capturing subtle scientific semantics (Li et al., 2024). SciMMIR benchmarks sci-multimodal scientific retrieval and compare vision language models under the zero-shot and fine-tuned settings (Wu et al., 2024). SciVQA tests reasoning on scientific figures, the results show that there is still room to improve on visual reasoning tasks, as the current strong multimodal systems do not yet outperform humans on this task (Borisova et al., 2025). The experiments above emphasize the importance of representations that are adapted to scientific content, rather than solely relying on general web training of vision-language models.

2.3 Scientific Similarity and Citation Analysis

SPECTER Learning document representations using citation information and handling tasks such as citation prediction and classification (Cohan et al., 2020). Scientific RepEval: a system for evaluating representations of scientific documents over multiple tasks (Singh et al., 2023). Multi-vector methods learn fine-grained document similarities and aspect-based similarity learning directly from citation sections (Mysore et al., 2022; Ostendorff et al., 2020). Citation recommendation methods use document contents or citation context to recommend relevant scientific papers (Bhagavatula et al., 2018; Duma et al., 2016). These techniques demonstrate the importance of scientific relations and domain-specific knowledge. Yet, most of these methods are based on text and never deal with classifying relationships between scientific figures directly by integrating both visual and text methods.

2.4 Vision-Language Models

Recently, a lot of state-of-the-art approaches have been proposed that seamlessly connect visual and textual ways. CLIP learns text and image representations aligned with large scale image-text data (Radford et al., 2021). SigLIP learns image and text using a sigmoid loss for the image-text relation (Zhai et al., 2023). BLIP-2 connects a frozen visual encoder with a language model through a Querying Transformer (Li et al., 2023), LLaVA combines a vision encoder with a large language model to follow instructions from visual input (Liu et al., 2023). ColPali learns visual document representation and late interaction for document retrieval (Faysse et al., 2024). These approaches gave us some clues on multimodal learning but ultimately AstroCLIMB is task-specific, which is to classify the relation types between astronomy figure-text pairs.

3 Methodology

The proposed system performs this multimodal relationship classification of between pairs of astronomy figures and their respective captions. The system flow is illustrated in Figure 1.

3.1 Extraction and Preprocessing

The data is a collection of astronomy figures subpanels captions, and text data from the NASA ADS and AstroCLIMB collection (Grezes et al., 2026). Textual data is stored as UTF-8 string and subject to simple normalization as well as whitespace handling before text feature extraction.

For each object, we obtain characters words sentences, references to figures, tables, and equations. Astronomy-specific features are obtained from catalogue and object identifier regular-expression patterns, which include NGC IC M, HD HIP Kepler, Wasp, and other common catalogue and object identification conventions, and for instruments and surveys, by matching text to a list of instrument and survey names, including JWST, Hubble, Chandra, ALMA, TESS, Gaia, and SDSS. Identifiers and instrument names are normalized for consistent matching. For each pair, we then extract shared identifiers and shared instrument labels as well as their Jaccard similarities, and the number of numerical entries in common. These features work as explicit features for the presences of common objects

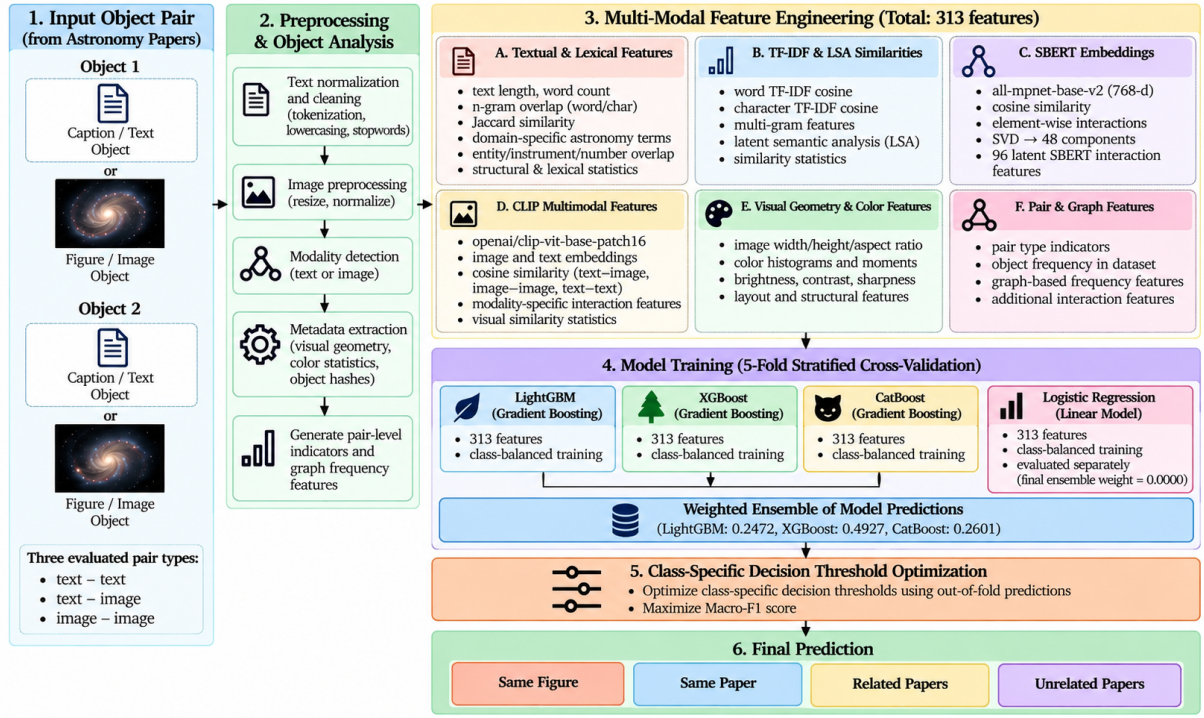


Figure 1: Implemented AstroCLIMB pipeline showing pairwise feature extraction, modality-aware interactions, weighted ensemble classification using LightGBM, XGBoost, and CatBoost, and class-specific threshold optimization.

and common observational instrumentation. We also extract object-hash frequency features from the available train and test pairs. For each pair, we extract the minimum and maximum of the two object-hash frequencies to provide an idea of the frequency of the two objects being paired.

We augment the lexical features with pretrained semantic encoders. For text, SBERT gives us 768-dimension text embeddings with the `all-mpnet-base-v2` checkpoint. For visual and cross-modal encodings, we use CLIP with the `openai/clip-vit-base-patch16` checkpoint, which gives us 512-dimension projected embeddings. The TF-IDF features encode the lexical similarity, while the SBERT embeddings measure semantic similarity. The TF-IDF vocabulary is fitted on the concatenated train and test corpus without the class labels.

Before CLIP text encoding each string is limited to 350 characters. The CLIP text encoder defaults to truncating the input to 77 tokens (including any special tokens required by the encoder), due to the context-length restriction of the encoder. As a result, information provided later in longer captions may not be encoded by CLIP. Both CLIP encoders are used in their original, pretrained state (i.e. without fine-tuning on the AstroCLIMB dataset). In the feature-extraction run reported here, CLIP text embeddings were computed for 14,581 text instances, while CLIP image embeddings were computed for 14,598 images. All images were successfully decoded in this run.

For images, we use CLIP image features and compare them with the paired object representations. This is image-image or cross-modality text-image matching.

Lastly, we explicitly encode the modality (text or image) of each object, which means that the pairwise feature construction stage can identify whether a pair is text-text, text-image, or image-image and include this information when constructing interaction features for the final classification.

3.2 Pairwise Feature Construction

For each pair of objects, the system generates a 313-dimensional feature vector of combined textual, semantic, visual, structural, and modality-aware features. The set of features contains the TF-IDF and SBERT similarities, CLIP-based text and image features, lexical overlap, features specific to astronomy, and the interactions of the compressed embeddings, aiming to combine complementary signals between the paired objects.

Object-level statistics object length, number of

words, number of sentences; Structural markers references to figure, table, and equation; Domain-oriented features number of astronomical objects and instruments in each object, and shared number of astronomical objects, shared astronomical instruments, and shared numerical values between the two objects.

Lexical overlap is measured with word-level Jaccard similarity, word overlap count, character 3-gram Jaccard similarity, and the fraction of common words compared to the shorter object. Entity, instrument, and numerical overlaps are captured using both shared-count and Jaccard-based features.

The pipeline explicitly encodes whether each object is an image and whether the pair is a mixed-modality pair. These semantic similarities are combined with modality indicators to create interaction features: TF-IDF, SBERT, and CLIP similarities are multiplied by indicators for each of the four possible modality configurations (text-text, text-image, image-text, image-image). The evaluated dataset contains text-text, text-image, and image-image pairs; no image-text pairs are present in the evaluated instances. Additional difference features, including SBERT minus TF-IDF, CLIP minus SBERT, and CLIP minus TF-IDF, are included to capture disagreement between the different representation spaces. The average and maximum text similarity signals are included as aggregate semantic features.

3.3 Ensemble Classification

The 313-dimensional pairwise feature vectors act as input to an ensemble of LightGBM, XGBoost, and CatBoost classifiers. A linear classifier is also tested, but is assigned zero weight when forming the ensemble. The training set includes 10,000 labeled instances sorted into four classes of relationship: 1,000 same-figure 3 000 same-paper 3 000 related-paper, and 3,000 unrelated-paper.

The optimal parameters used for the LightGBM classifier are a gradient-boosted decision tree with a learning rate of 0.03, 127 leaves, maximum depth of 9, row subsample of 0.8, feature subsample of 0.7, and up to 3,500 estimators. For XGBoost the parameters are a learning rate of 0.03, maximum depth of 8, minimum child weight of 15, row subsample of 0.8, feature subsample of 0.7, and up to 3,500 estimators. For CatBoost these are a learning rate of 0.03, depth 7 3 500 maximum iterations, and an L2 leaf reg regularization coefficient of 4.0. Early stopping with a patience of 150 rounds is set for all three tree-based classifiers. Weights for class imbalance are used in CatBoost, while sample weights are applied for LightGBM and XGBoost.

Table 1: Macro-F1 scores of individual classifiers and the weighted ensemble before and after threshold optimization.

Model / Configuration	**Macro-F1**
LightGBM (mean fold score)	0.48358
XGBoost (mean fold score)	0.49150
CatBoost (mean fold score)	0.47694
Linear classifier (mean fold score)	0.48082
Weighted ensemble (raw OOF)	0.48846
Weighted ensemble (threshold-optimized OOF)	**0.50163**

For scoring, 5-fold stratified cross validation is used to estimate out of fold performance with random seed 42. The predictions of the three models used contribute to the ensemble and are combined with weights chosen from out of fold predictions. Class specific decision multipliers are then multiplied into the ensemble scores to optimize about Macro-F1. In Table 1 ,the final macro-F1 after threshold improvement is presented as 0.50163 versus 0.48846 pre-tuning.

The final level evaluation is based on the out-of-fold predictions from all 5 folds. This way, each example in the train set has a prediction from a model that did not see that example during training.

4 Experiments and Results

The updated system is evaluated using five-fold stratified cross-validation on 10,000 labeled training instances. The dataset contains four relationship classes: same figure (1,000 instances), same paper (3,000 instances), related papers (3,000 instances), and unrelated papers (3,000 instances). Each fold evaluates the models on held-out instances that are excluded from the corresponding training phase.

Four classifiers are evaluated: LightGBM, XGBoost, CatBoost, and a linear classifier. Their out-of-fold predictions are used to determine the ensemble weights. The final ensemble assigns weights of 0.2472 to LightGBM, 0.4927 to XGBoost, and 0.2601 to CatBoost. The linear classifier receives zero weight in the final ensemble. Thus, the final predictions combine the outputs of the three contributing tree-based models.

The individual-classifier results show the model level comparison when using the same feature space. In Table 2, XGBoost scored the highest

Table 2: Per-class performance of the threshold-optimized ensemble on out-of-fold predictions.

Class	Prec.	Recall	F1
Same Figure	0.4612	0.7310	0.5656
Same Paper	0.6123	0.4170	0.4961
Related Papers	0.4071	0.4153	0.4112
Unrelated Papers	0.5086	0.5613	0.5337
Macro Avg.	0.4973	0.5312	0.5016
Weighted Avg.	0.5045	0.4912	0.4888

mean five-fold validation Macro-F1 (0.49150), followed by LightGBM (0.48358), the linear classifier (0.48082), and CatBoost (0.47694). The weighted ensemble resulted in an OOF Macro-F1 of 0.48846, which increased to 0.50163 following class-specific threshold tuning. This class-specific threshold improvement results helps spell out the effect of classifier choice and ensemble calibration; they do not yet show the effects of the individual feature ways.

The experiments above report performance for the ensemble with five-fold out-of-fold predictions. The raw weighted ensemble obtains a Macro-F1 of 0.48846, rising to 0.50163 after class-specific decision-multiplier tuning. These results provide performance benchmarks for the specified system; other than the projected TF-IDF-only and CLIP-only baselines, we cannot compare majority-class predictions under the same validation regime to measure the additive contribution of the whole multimodal feature set. For this reason we consider these results an experimental validation of the specified pipeline and not as a controlled experiment demonstrating the individual benefit of each modality.

Class-specific threshold tuning results before class-specific decision thresholds tuning, the ensembleout-of-fold Macro-F1 is 0.48846. When using class-specific decision multipliers, the threshold-optimized Macro-F1 is 0.50163. The absolute gain is 0.01317. The results show that class-specific decision thresholds improvement can improve the class balance of the ensemble for the four classes.

The threshold-optimized results show that the same-figure class obtains the highest F1 score at 0.5656, followed by unrelated papers at 0.5337 and same paper at 0.4961. Related papers remain challenging, with an F1 score of 0.4112. The same-figure class has the highest recall (0.7310), whereas the same-paper class has the highest precision (0.6123). This illustrates the different precision-recall trade-offs across the relationship categories.

The threshold-optimized classification report gives an overall accuracy of 0.4912. The Macro-F1 score is 0.50163, while the weighted-average F1 score is 0.4888. Macro-F1 is used as the primary metric because it gives equal importance to each relationship class despite the difference in class support.

The final submission contains 10,000 test predictions, with one predicted class assigned to each instance. The predicted distribution is 1,309 same-figure instances (13.09%), 1,657 same-paper instances (16.57%), 3,474 related-paper instances (34.74%), and 3,560 unrelated-paper instances (35.60%). The submission file and the corresponding soft predictions were generated by the final pipeline.

In addition to cross-validation evaluation, the proposed system was evaluated in the AstroCLIMB Kaggle competition. At the time of recording the leaderboard results, the submission ranked 10th with a score of 0.48121. This competition result is reported separately from the out-of-fold Macro-F1 scores because the two results arise from different evaluation settings. The competition leaderboard is available at https://www.kaggle.com/competitions/astroclimb/leaderboard.

5 Error Analysis and Discussion

Class-by-class results reveal that the four relation classes display varying precision-recall behavior. The same-figure class has the highest F1 of 0.5656, but it also has a recall of 0.7310. This shows the threshold-optimized model is capturing many same-figure pairs, but with a precision of 0.4612, some of the pixels predicted to belong to this class are actually false positives.

The most precise class is the same-paper class at 0.6123, but the class recall is low at 0.4170. This shows that predictions made for this class are relatively stringent, although we are assigning a fair number of actual same-paper relations to other classes. It may be more difficult to differentiate same-paper versus other links without specific evidence of shared source documents.

Related papers is also an ambiguous class with an F1 score of 0.4112. Truly related papers can mention similar scientific terms, astronomical objects instruments observations, and research goals without being the same paper (or cited by the same paper). Because of this, using lexical or seman-

tic similarity on its own may not be sufficient to tell related-paper relationships apart from same- or unrelated-paper relationships.

F1 for the unrelated-papers class is 0.5337, with precision 0.5086 and recall 0.5613. Of course, two papers from different sources may still talk about the same objects, instruments, measurements, and scientific things. So high semantic similarity does not mean that two objects must be related to each other.

The final ensemble combines LightGBM, XGBoost, and CatBoost, with XGBoost receiving the largest weight. By combining their predictions, the system uses the complementary decision patterns of the three classifiers. Class-specific decision multipliers further improve the out-of-fold Macro-F1 from 0.48846 to 0.50163. Yet, since these multipliers are optimized using out-of-fold predictions, the threshold-optimized score should be interpreted in this improvement procedure, rather than as an independent estimate of generalization performance.

Overall, the findings indicate that integrating lexical semantic visual, structural, and astronomy-specific features can aid in multimodal relationship classification. Yet, the observed differences in precision and recall across classes show that accurately identifying the correct relationship involves more than simply assessing similarity between paired objects.

The system uses lexical, semantic, visual, structural, and astronomy-oriented features. These groups of features complement each other though we do not attempt to measure the impact of each group of features in isolation by 'ablating' features in a controlled manner. As such, these results should not be used so that each group of features contributes to the improvements in Macro-F1.

Object length, number of words, and number of sentences are cheap content-based indicators that could signal whether an object has lots of text and might this way help differentiate between text-rich objects and short captions, images, or other objects. But these statistics could also be a function of object modality or formatting conventions, or properties of dataset construction rather than the relationship of interest. We can't claim that they encode the semantics of the relationship without a rigorous experimental test.

6 Conclusion

This work investigates multimodal relationship classification for astronomy figures and text using the AstroCLIMB benchmark. The proposed system combines lexical, semantic, visual, structural, and domain-oriented pairwise features, including TF-IDF and SBERT similarities, CLIP-based representations, modality-aware interaction features, astronomical entity and instrument overlap, numerical overlap, and object-level statistics.

The final system uses 313 engineered features and combines LightGBM, XGBoost, and CatBoost predictions through a weighted ensemble. Under five-fold stratified cross-validation, the ensemble achieves an out-of-fold Macro-F1 of 0.48846 before class-specific threshold optimization and 0.50163 after optimization. The per-class results show that same-figure relationships are identified with the highest F1 score, while related-paper relationships remain more difficult to classify.

These findings demonstrate the potential of combining complementary textual and multimodal signals for relationship classification. Future work can investigate OCR-enhanced figure representations, richer document structure, improved figure-caption alignment, and end-to-end vision-language models to better distinguish semantic similarity from direct relationships between scientific objects.

7 Limitations

Our system has several limitations. First, distinguishing related-paper relationships remains challenging, with an F1 score of 0.4112 for this class. Scientific papers from the same research area may share terminology, astronomical entities, instruments, observations, and semantic content even when their relationship is indirect.

Second, visual information is primarily represented through pretrained CLIP embeddings rather than explicit interpretation of plot components. The current pipeline does not explicitly perform OCR over axis labels, legends, tick values, or annotations. Consequently, visually similar figures with different numerical or textual content may remain difficult to distinguish.

Third, the system relies on engineered pairwise features and a weighted ensemble of LightGBM, XGBoost, and CatBoost. Although this approach combines several classifiers and different sources of information, it may not capture all fine-grained

relationships between figure structure and surrounding scientific text.

Finally, the reported threshold-optimized Macro-F1 is calculated from out-of-fold predictions after class-specific decision multipliers have been optimized. This optimization may make the reported score optimistic as an estimate of generalization performance. Independent evaluation of the selected thresholds would provide a more reliable assessment. The pipeline also depends on the quality and completeness of the available textual and visual data; missing, incomplete, or weakly informative captions can reduce the usefulness of lexical and semantic features.

Future work can investigate OCR-enhanced visual representations, richer document structure, improved figure-caption alignment, and end-to-end vision-language fine-tuning.

8 Acknowledgment

This paper was prepared for the AstroCLIMB shared task at WASP 2026. We thank the organizer and reviewers. The implementation and benchmark results described here are based on the project model and evaluation information supplied for the research study.

References

Anonymous. 2021. Scifig: A scientific figure dataset for figure understanding. OpenReview submission.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–251.

Elena Borisova, Nina Rauscher, and Georg Rehm. 2025. Scivqa 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing*, pages 182–210.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Daniel Duma, Maria Liakata, Amanda Clare, James Ravenscroft, and Ewan Klein. 2016. Applying core scientific concepts to context-based citation recommendation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1737–1742.

Mickael Faysse, Hugues Sibille, Tong Wu, Behrooz Omrani, Gregory Viaud, Celine Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. Astroclimb: Astronomy citation linking from illustrations - a multimodal benchmark (overview of the wasp 2026 shared task). In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Ting-Yun Hsu, C. Lee Giles, and Ting-Hao Huang. 2021. Scicap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264.

Ting-Yun Edward Hsu, Yu-Lin Hsu, Saksham Rohatgi, Chia-Yu Huang, H. Y. S. Ng, Ryan Rossi, Seung Kim, Tong Yu, Lun-Wei Ku, C. Lee Giles, and Ting-Hao Kenneth Huang. 2026. Do large multimodal models solve caption generation for scientific figures? lessons learned from scicap challenge 2023. *Transactions of the Association for Computational Linguistics*, 14:233–252.

Chieh-Yang Huang, Ting-Yao Hsu, Ryan Rossi, Ani Nenkova, Sungchul Kim, Gromit Yeuk-Yin Chan, Eunyee Koh, Clyde Lee Giles, and Ting-Hao Huang. 2023. Summaries as captions: Generating figure captions for scientific documents with automated text summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 80–92.

Abhishek Kashyap, Zhen Yang, and Min-Yen Kan. 2023. Scientific document processing: Challenges for modern learning methods. *International Journal on Digital Libraries*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lei Li, Yufei Wang, Ruochen Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14369–14387.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470.

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based document similarity for research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6194–6206.

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. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566.

Jinyu Song, Abhishek Das, Guanhua Cui, and Yufeng Huang. 2025. Figex: Aligned extraction of scientific figures and captions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16558–16571.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. Scimmir: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574.

Zhen Yang, Raj Dabre, Hiroki Tanaka, and Naoaki Okazaki. 2024. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing*, 31(3):1140–1165.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.