PSGS@2026: From Figures to Citation Graphs: QLoRA Based Multimodal Linking for AstroCLIMB

Prajwal Ghimire[1,2] **and Shiv Ram Saud**[1,2] **and Aashish Mahato**[1,2] **and Sandeep Dawadi**[1,2]
Ramesh Saud and **Sunil Regmi**[1,2] and **Manoj Shakya**[1,2]

[1]Department of Artificial Intelligence, Kathmandu University
[2]Data Science & Artificial Intelligence Lab, Kathmandu University
{prajwalghimire22,saudshivram373,aashishmtho24,sundeepdwd,rajusaud245}@gmail.com
{sunil.regmi,manoj}@ku.edu.np

Abstract

We describe our system for the AstroCLIMB shared task (WASP 2026), which reconstructs the citation graph of astronomy papers from scientific figures and captions alone. The task requires identifying the relationship between pairs of figures and captions as `same_figure`, `same_paper`, `related_papers`, or `unrelated_papers`. Our system finetunes `google/gemma-4-E4B-it` using QLoRA on 8,200 competition training rows with a shared descriptive prompt and greedy decoding. A zero shot baseline achieves 0.4408 macro F1 on a balanced 1,800 pair validation split held out from the competition training file, and the finetuned system substantially improves this citation-relation classification for the challenging `related_papers` class in particular. Our submitted system achieves 0.64282 F1 on all 10,000 official test pairs. These results show that multimodal finetuning can improve citation-relation inference from scientific figures and captions when document metadata is unavailable.

1 Introduction

AstroCLIMB (Astronomy Citation Linking from Illustrations) is a shared task of the 4th Workshop on Artificial Intelligence for Scientific Publications (WASP) at AACL-IJCNLP 2026 (Grezes et al., 2026). Systems predict exactly one of four relations between a pair of objects (two figures, a figure and a caption, or two captions): `same_figure` (both objects from the same figure; figure–caption pairs only), `same_paper` (same DOI, different figures), `related_papers` (one DOI cites the other), or `unrelated_papers` (none of the above). The relation is symmetric; each pair carries a single label, and test pairs carry no DOI or metadata. Therefore, classification must rely on visual–textual content alone.

Our early experiments showed that our strongest zero shot vision language model already performs well on the `same_figure` class while, like the other baselines, largely failing on the citation graph classes, with `related_papers` near zero across every baseline (Section 4). We therefore finetune **Gemma-4-E4B-it** with QLoRA (Dettmers et al., 2023) on 8,200 labeled pairs, optimized with AdamW (Loshchilov and Hutter, 2019), using the exact prompt format also used at inference, with a fixed evaluation protocol (greedy decoding, label only output, macro F1 via scikit-learn (Pedregosa et al., 2018)). The resulting system (Gemma-4-E4B plus the `adapter_best` checkpoint) is our official submission, scoring 0.64282 on all 10,000 test rows.

2 Related Work

The scientific figure–caption literature provides the visual–textual grounding for our task. SciCap introduces scientific-figure caption generation with baselines on graph plots (Hsu et al., 2021), while Multimodal ArXiv introduces ArXivCap and ArXivQA and benchmarks vision-to-text tasks, showing that in-domain training substantially improves scientific-figure comprehension (Li et al., 2024). SciMMIR benchmarks text–figure/table retrieval over 530K pairs (Wu et al., 2024), whereas SciVer evaluates multimodal scientific claim verification on 3,000 expert-annotated examples from 1,113 papers across four reasoning types (Wang et al., 2025). Neither task infers citation relations from scientific figures. For scholarly document relations, SPECTER learns document representations from titles and abstracts using a citation graph triplet objective (Cohan et al., 2020). Citation intent is studied separately by Structural Scaffolds, which incorporates structural information through multitask learning (Cohan et al., 2019), and GraphCite, which combines citation-graph embeddings with

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 192–198
November 10, 2026 ©2026 Association for Computational Linguistics

citation phrases (Berrebbi et al., 2022). Content Based Citation Recommendation instead ranks citation candidates for a query document from textual content without requiring metadata (Bhagavatula et al., 2018). Text-only astronomy models (astroBERT (Grezes et al., 2021); AstroLLaMA (Nguyen et al., 2023)) motivate the domain focus but use no figures. Thus, prior work has separately studied scientific figure–caption modeling, scientific multimodal retrieval and reasoning, and scholarly citation/document relationships. AstroCLIMB combines these dimensions by requiring relation inference directly from scientific figures and captions without document metadata at test time.

3 Methodology

Task. Each (Object A, Object B) pair is classified into one of four classes C. An object is a figure image or a caption text, giving three pair types: TXT-IMG (figure–caption), IMG-IMG (figure–figure), and TXT-TXT (caption–caption). `same_figure` occurs only for TXT-IMG pairs by construction.

Competition data and distribution. The Kaggle competition release (Grezes, 2026) provides 10,000 labeled training pairs and 10,000 test pairs, with object cells containing base64-encoded image bytes or caption text. A separate **sampled development release** contains 1,800 balanced rows with figure images stored as UUID-keyed PNGs `https://www.kaggle.com/datasets/jcxiv42/astroclimb` and is used only for the four few shot demonstrations but never for training or scoring. The two releases contain different rows and are never mixed. The full training file is imbalanced with 10.0% `same_figure` and 30.0% for each other class, while our validation split contains 450 examples per class (seed 42) that leaves 8,200 training rows with 6.7% `same_figure` and 31.1% for each other class. The split is by pair row index with zero index overlap; figures, captions, or source papers may therefore recur across splits, so validation measures in-distribution pair generalization rather than generalization to unseen papers. The primary evaluation metric is macro F1 which is the unweighted mean of per class F1 (Pedregosa et al., 2018).

Multimodal input. Each example pairs the two objects of one row as $x_i = (I_i, T_i)$, where I_i is the ordered list of figure images (two for IMG-IMG, one for TXT-IMG, empty for TXT-TXT) and T_i is the shared descriptive system prompt plus per object caption blocks (truncated to 2,000 characters) and the closing instruction "Answer with one label:". Images appear before their corresponding text and are resized to 448×448. The true label $y_i \in \mathcal{C}$ is appended after the prompt for supervised fine tuning (SFT).

Model architecture. The backbone is **google/gemma-4-E4B-it** (Team et al., 2026a), a dense model with 4.5B effective parameters and a 150M vision encoder, loaded as `AutoModelForMultimodalLM` (falling back to `AutoModelForImageTextToText`). We train LoRA (Hu et al., 2021) adapters on all attention and MLP projections with rank r=64, α=128, dropout 0.05 ($\Delta W = \frac{\alpha}{r} BA$, effective scaling 2) that minimizes causal LM cross entropy with AdamW (Loshchilov and Hutter, 2019). Both training and inference quantize the frozen backbone to 4-bit NF4 with double quantization and bf16 compute (Dettmers et al., 2023).

Training. The model was trained on remaining residual 8,200 rows after removing the balanced 1,800 row validation split (zero index overlap). Training uses the HF `Trainer`, learning rate 10^{-4}, cosine schedule with 5% warmup (204 steps), per device batch 2 with gradient accumulation 8 (effective batch 16), max sequence length 1,536, and a cap of 2,560 steps with early stopping (patience 3, threshold 0.001, evaluated every 200 steps) on a single NVIDIA A100 (80GB). The best validation loss (0.3464) is observed at step 1,000 and the corresponding checkpoint is saved as `adapter_best`. Early stopping terminates the run at step 1,600. Decoding is greedy (`max_new_tokens=32`, `do_sample=False`). The raw generation is mapped to a class by case insensitive regex over the four labels with falling back to `unrelated_papers` when nothing matches. Figure 1 summarizes the training pipeline of the submitted system.

Inference and submission. Testing uses the official 10,000 row `test.csv`. The base model at test time is loaded in 4-bit on 2×T4 GPUs (Kaggle) with `adapter_best` attached unmerged and the predictions are written one-hot in the required column order. The local evaluations report macro F1 and accuracy via scikit-learn (Pedregosa et al., 2018). Figure 2 summarizes the test pipeline.

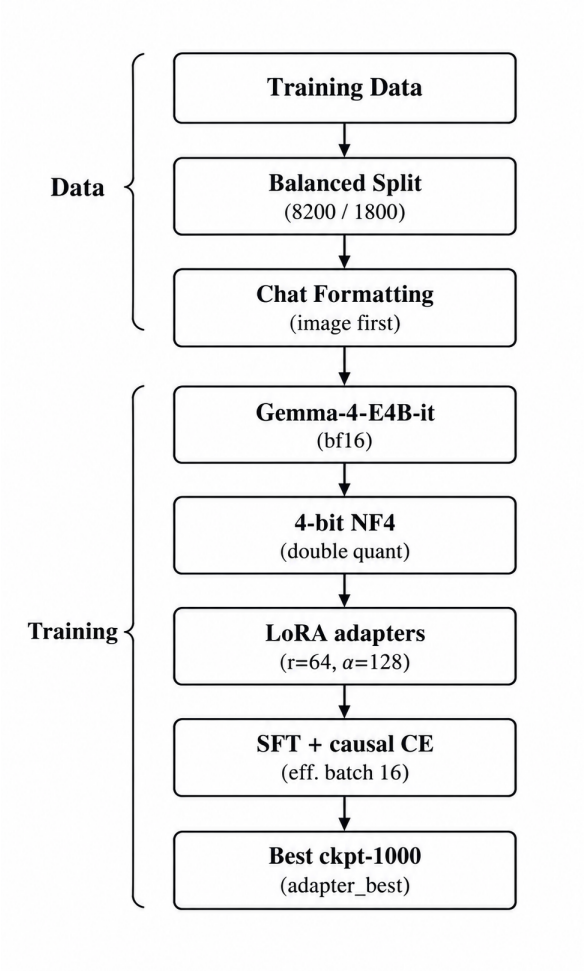


Figure 1: Training pipeline of the submitted system.

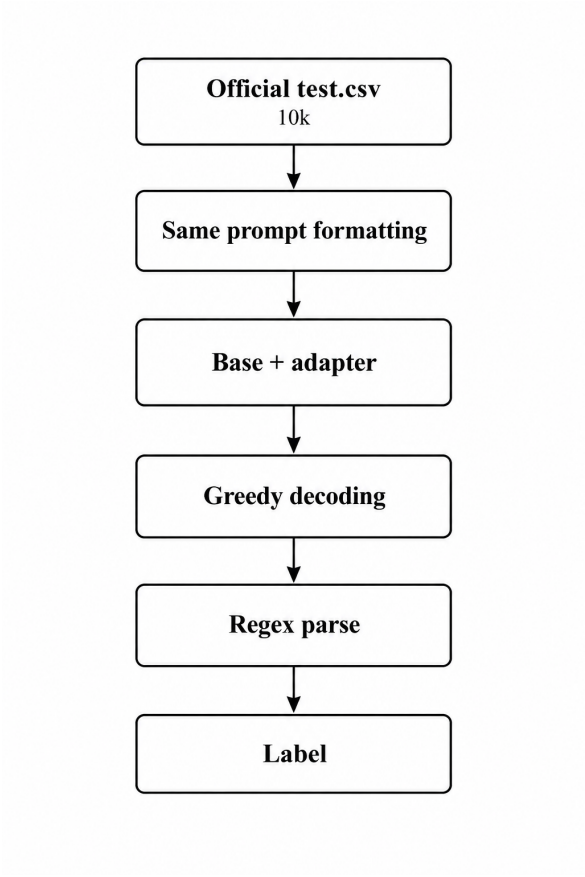


Figure 2: Test/submission pipeline.

4 Results

We evaluated zero shot and few shot vision language model baselines using macro F1 and prediction coverage, and examined the effects of few shot demonstrations and native thinking capabilities, and finetuning on performance.

Zero/few shot baselines. All baselines share one descriptive system prompt and greedy decoding with only the backbone and its native thinking mode varying. Gemma-4-E4B (Team et al., 2026a) leads zero shot at 0.4408 macro F1 (Table 1) whereas Qwen3-VL-8B-Thinking (Bai et al., 2025) collapses to `unrelated_papers` (0.1068), and SmolVLM-500M (Marafioti et al., 2025) scores 0.1723.

The few shot runs include 4 demonstrations (one per class, seed 42, figure–caption pairs preferred) and enable native thinking where the backbone supports it. The decoding is otherwise greedy as in zero shot except that thinking runs use `max_new_tokens=512` instead of 32. The smaller Qwen3-VL-4B and GLM-4.1V-9B variants are used here in place of the Qwen3-VL-8B and GLM-4.6V-Flash variants evaluated zero shot (Team et al., 2026b), so the two tables are not paired backbone-for-backbone. Table 2 compares configurations and results. For Gemma-4-E4B, adding few shot thinking (0.3925) does *not* beat its zero shot score (0.4408) on the identical rows. The Qwen3-VL-4B model covers 397 pairs (0.2497), and GLM-4.1V-9B covers 445 pairs (0.4200). The four demonstrations are drawn from the separate sampled release (disjoint rows); all scores are computed on the validation split.

Model	Think.	Rows	Macro F1	Acc.
Gemma-4-E4B	no	1800	0.4408	0.5206
Qwen3-VL-8B	yes	1800	0.1068	0.2461
SmolVLM-500M	no	1800	0.1723	0.2511
GLM-4.6V-Flash	no	610	0.1020	0.2557

Table 1: Zero shot results on the 1,800-row validation split held out from the training file. Rows list evaluated pairs.

Finetuned system. Our submitted adapter reaches a best validation loss of 0.3464 at check-

Model	Demos	Rows	Macro F1	Acc.
Gemma-4-E4B	4	1800	0.3925	0.505
Qwen3-VL-4B	4	397	0.250	0.330
GLM-4.1V-9B	4	445	0.420	0.524

Table 2: Few shot results with native thinking enabled (demonstrations from the sampled release; evaluation on the same validation split as Table 1). Rows list evaluated pairs and partial row scores are not comparable to full split scores.

point 1,000 on the balanced 1,800 pair split and achieves a macro F1 of 0.5347 (accuracy 0.5483 & per class F1: 0.5705/0.5311/0.3508/0.6864). It substantially improves `related_papers` classification, from near zero F1 in the baselines to 0.3508, while `same_figure` decreases from 0.7760 to 0.5705 with a precision of 0.9945. A comparator SmolVLM-500M LoRA-32 adapter reaches a *lower* validation loss (0.3029) but predicts no `unrelated_papers` on the official test predictions and achieves only 0.23675 on the official leaderboard, compared with the official test F1 of 0.64282 for our system (Table 3).

Model	Val. Loss	Official Test F1
Gemma-4-E4B, LoRA-64	0.3464	**0.64282**
SmolVLM-500M, LoRA-32	0.3029	0.23675

Table 3: Fine-tuned model comparison.

4.1 Per class breakdown & Confusion Analysis

Table 4 reports per class F1 for the leading run per mode (full run list in Tables 1 and 2).

Run	Mode	sf	sp	rp	up
Gemma-4-E4B	ZS	0.7760	0.4108	0.0172	0.5593
Gemma-4-E4B	FS	0.7228	0.2047	0.0505	0.5918
SmolVLM-500M	ZS	0.0128	0.0088	0.3333	0.3341
Gemma-4-E4B	FT	0.5705	0.5311	0.3508	0.6864
GLM-4.1V-9B (445)	FS	0.8653	0.2571	0.0	0.5575
Qwen3-VL-4B (397)	FS	0.4709	0.2381	0.0612	0.2286

Table 4: Per class F1 on the 1,800 row validation split (truncated runs: evaluated subsets, pair counts in parentheses). Leading runs only (sf = `same_figure`, sp = `same_paper`, rp = `related_papers`, up = `unrelated_papers`). ZS = zero shot, FS = few shot, FT = finetuned. Truncated runs are not comparable with full split runs.

The submitted adapter shows strongly asymmetric errors (Figure 3) on the validation split. `same_figure` is highly precise (0.9945) but has low recall (0.4000), whereas `same_paper` has higher recall (0.7600) but lower precision (0.4081). `related_papers` remains weakest (precision 0.5392, recall 0.2600) and `unrelated_papers` is more balanced (precision 0.6170, recall 0.7733).

The primary confusion on validation is `same_figure` → `same_paper`: 256 of 270 missed `same_figure` pairs receive this label. Validation `related_papers` errors concentrate in figure–figure pairs, with recall of 0.0867 and 104 of 150 cases predicted as `same_paper`. The same pattern appears on the official test predictions but is less severe: `same_figure` precision is 0.9386 with recall 0.6420, and 333 of 358 missed `same_figure` cases are classified as `same_paper`. Test precision and recall for the remaining classes are 0.6453 and 0.6463 (`same_paper`) and 0.5573 and 0.3823 (`related_papers`), while figure–figure pairs again show the weakest `related_papers` F1 (0.3166) (Figure 4).

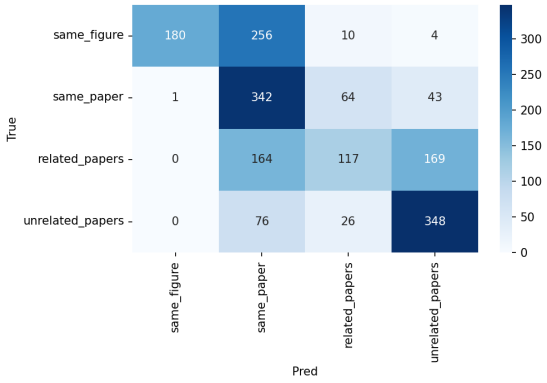


Figure 3: Confusion matrix of the submitted adapter on the balanced 1,800 pair split.

5 Competition Submission Result

Our submission scores all 10,000 test rows and achieves **0.64282**, with predicted class counts 684/3,005/2,058/4,253 (`same_figure`/`same_paper`/`related_papers`/`unrelated_papers`), which shows the use of all four labels rather than collapse onto a single class, and the expected rarity of direct figure–caption matches. No test labels were used in training, model selection, or prompt design. The test set analyses use the organizers' released solution file and were performed after the leaderboard submission. Errors on the official test predictions are asymmetric: `same_figure`

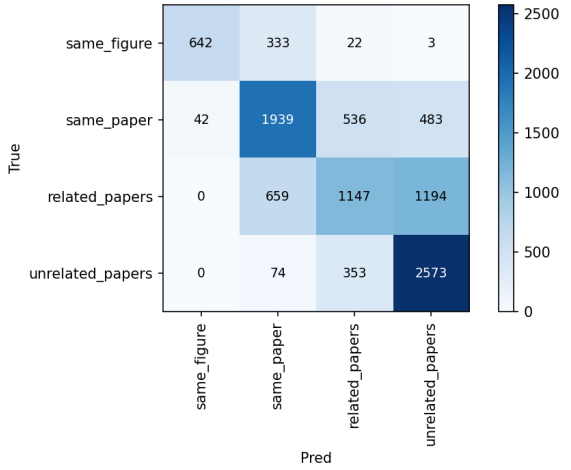


Figure 4: Confusion matrix of the official 10,000 test predictions.

has precision 0.9386 and recall 0.6420, with 333/358 missed pairs predicted as `same_paper`. The `related_papers` class remains weakest overall on test too (precision 0.5573, recall 0.3823, F1 0.4535, up from F1 0.3508 on validation), with figure–figure pairs again the hardest subset.

6 Conclusion

The QLoRA finetuning of Gemma-4-E4B-it on 8,200 competition training rows achieved a best validation loss of 0.3464 and a macro F1 of 0.5347 on the balanced 1,800 row validation split. The resulting submitted system achieved 0.64282 F1 on the official 10,000 row test set. Few shot thinking performed below zero shot for the same backbone (0.3925 vs. 0.4408). The larger thinking models evaluated also score below Gemma-4-E4B's zero shot baseline on their evaluated pairs. Several models cover only partial row subsets and are not directly comparable to full split results. We do not run caption-only or image-only ablations, so we cannot separate multimodal gains from textual topic similarity. The `adapter_best` weights (best validation loss 0.3464) and the official 10,000-row submission are public at https://huggingface.co/ShivRamSaud/gemma4-e4b-qlora-astroclimb-v2-fresh; together with the fixed protocol above and the seed 42 row index split of the public competition file, they document the run (see Appendix A). The training script, preprocessing code, and split row indices are not released; exact reproduction requires reimplementing them from the protocol, prompt, and seed 42 split specification given here. Future work should fine-tune the thinking capable backbones evaluated here as baselines and explore pair type aware training objectives for the hard `related_papers` class.

7 Acknowledgements

We thank the organizers of the AstroCLIMB shared task at WASP 2026 for the competition release and leaderboard evaluation (Grezes, 2026).

A Reproducibility artifacts

Released weights and predictions. The submitted `adapter_best` LoRA adapter (best validation loss 0.3464 at step 1,000) and the official 10,000-row test submission are public at https://huggingface.co/ShivRamSaud/gemma4-e4b-qlora-astroclimb-v2-fresh. The `adapter_best/` folder contains `adapter_model.safetensors`, `adapter_config.json`, the tokenizer and processor configs, and the training arguments. `submission/` holds the scored test predictions.

Not in the repository. The repository contains no training or preprocessing scripts, no prompt file, and no split-index file, and its README is the unfilled PEFT template (no model card). These are instead fixed by the specification in the paper: preprocessing is 2,000-character caption truncation with 448×448 image resizing, the split is seed 42 row indices of the public competition training file (8,200 train / 1,800 validation), decoding is greedy with regex label mapping falling back to `unrelated_papers`, and the exact inference prompt is reproduced below.

A.1 Prompt format

The training, validation, and test use a single fixed prompt contract which is a one descriptive system prompt followed by per object content blocks, with images ordered before text, closed by an explicit answer instruction. The processor renders the turns through its chat template with the generation prompt appended, so training and inference share identical formatting. The exact system prompt is:

```
You are an expert in
astrophysics figures and
captions.  Given Object A
and Object B (each is either
a figure image or a caption
text), classify their relation
into exactly ONE label based
ONLY on what you see/read - no
DOI or metadata is provided.
Classes:  - same_figure:  The
```

```
caption directly describes
the figure in front of you.
Visual elements (axes, labels,
numbers, morphology) are
mentioned verbatim in the
text, or the text reads like
"Figure X shows..." matching
the image. - same_paper:
Same study, different figures.
Similar writing style, same
instruments/datasets/authors
hinted in text, or visual
style (fonts, colors, layout)
is consistent, but NOT a
direct caption-figure match.
- related_papers: Different
papers where one builds on
the other. Overlapping
methods, shared datasets, or
a figure/caption that looks
like a cited prior result,
but style/authors differ. -
unrelated_papers: No clear link.
Different topics, instruments,
scales, or writing/visual style
with no overlap. Base your
decision only on visual and
textual content. Do not assume
same_figure is impossible
for any pair type - judge
from alignment. Output ONLY
the lowercase label (e.g.,
related_papers), no explanation,
no punctuation. Choose
exactly one of: same_figure,
same_paper, related_papers,
unrelated_papers.
```

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. Graphcite: Citation intent classification in scientific publications via graph embeddings. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 779–783, New York, NY, USA. Association for Computing Machinery.

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, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Felix Grezes. 2026. Astroclimb: Astronomy citation linking from illustrations. https://www.kaggle.com/competitions/astroclimb. Kaggle.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astrobert, a language model for astronomy & astrophysics. *Preprint*, arXiv:2112.00590.

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-Yao 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Lei Li, Yuqi Wang, Runxin 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 (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *Preprint*, arXiv:2504.05299.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, and 5 others. 2023. Astrollama: Towards specialized foundation models in astronomy. *Preprint*, arXiv:2309.06126.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python. *Preprint*, arXiv:1201.0490.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026a. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. 2025. SciVer: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8562–8579, Vienna, Austria. Association for Computational Linguistics.

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, Bangkok, Thailand. Association for Computational Linguistics.