Beyond Visual Similarity: Parameter-Efficient Vision–Language Models for Scientific Citation Linking on AstroCLIMB

Ashmani Kumar **Shubham Sharma** **K S Ramalakshmi** **Ankush Chopra**
AI COE / Tredence Analytics
Bengaluru, Karnataka, India
{ashmani.kumar, shubham.sharma, ramalakshmi.ks, ankush.chopra}@tredence.com

Abstract

This paper describes our submission to the AstroCLIMB shared task, which evaluates whether multimodal models can identify citation relationships among astronomy papers from pairs of figures and captions. We explore an empirical progression spanning frozen embedding baselines, dual encoders, cross-encoders, and generative vision–language models (VLMs) adapted via parameter-efficient fine-tuning (LoRA with 4-bit quantization). Among our evaluated configurations, a fine-tuned Qwen3.8-27B achieves a test Macro-F1 of approximately 0.74. An auxiliary specialist adapter trained to re-rank the frequently confused *same paper* and *related papers* classes yields comparable performance within expected run-to-run variance. Across input modalities, Text–Text pairs achieve the highest classification accuracy, followed by Text–Figure pairs, while Figure–Figure pairs prove the most challenging. These findings demonstrate that visual similarity alone remains an insufficient proxy for semantic citation linkage in specialized scientific literature.

Keywords: Multimodal Learning, VLM, Citation Linking, PEFT.

1 Introduction

Scientific figures and captions contain valuable insights, but they are often overlooked by automated research discovery tools. The AstroCLIMB(Grezes et al., 2026) shared task at WASP, IJCNLP–AACL 2026 evaluates whether multimodal models can partially infer citation linkages among astronomy publications directly from isolated pairs of figures and captions. The benchmark tests three configurations of the modality: Figure–Figure (`ii`), Text–Figure (`ti`) and Text–Text (`tt`). Each pair is assigned to one of four relational proximity tiers: *same figure*, *same paper*, *related papers*, or *unrelated papers*.

This task is challenging because astronomy figures from different studies often look similar, while distinguishing figures from the same paper and those from closely related papers requires deeper scientific understanding. We evaluate several models, from AstroBERT-based dual encoders to efficient vision-language models. Our best-performing configurations achieve approximately 0.74 test Macro-F1. A fine-grained analysis across modalities reveals a consistent performance hierarchy (`tt` $>$ `ti` $>$ `ii`), confirming that textual semantics currently provide more reliable citation signals than visual features alone.

2 Related Work

Scientific Document Understanding (SDU): While citation analysis historically relies on text (Beltagy et al., 2019; Devlin et al., 2019), SDU increasingly targets non-textual evidence like charts and plots, which pose unique challenges due to technical axes and visual conventions (Masry et al., 2022). In astronomy, domain-specific models like ASTROBERT (Grezes et al., 2021, 2022) parse literature from the NASA ADS, yet multimodal citation linking directly from isolated artifact pairs remains underexplored, motivating the AstroCLIMB benchmark.

Multimodal Modeling and PEFT: Architectures broadly follow two paradigms: dual encoders mapping features into a shared space (e.g., ViT (Dosovitskiy et al., 2021), SigLIP 1/2 (Zhai et al., 2023; Tschannen et al., 2025)) and generative VLMs combining patch encoders with causal decoders (e.g., PaliGemma (Beyer et al., 2024), Qwen (Wang et al., 2024)). To mitigate computational costs, Parameter-Efficient Fine-Tuning (PEFT) via LoRA (Hu et al., 2021) and 4-bit quantization (Dettmers et al., 2024) allows adapting large backbones on a single workstation GPU. We evaluate whether unified generative VLMs outperform mod-

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 210–215
November 10, 2026 ©2026 Association for Computational Linguistics

ular dual encoders on specialized scientific figures.

3 Dataset and Problem Statement

The AstroCLIMB shared task targets semantic citation relationships between multimodal artifact pairs from NASA's Astrophysics Data System (ADS). Formally, each instance is denoted as (x_1, x_2, τ, y), where x_1 and x_2 are scientific artifacts (figures or text captions), $\tau \in \{\texttt{ii}, \texttt{ti}, \texttt{tt}\}$ specifies the modality configuration (Figure–Figure, Text–Figure, or Text–Text), and $y \in \mathcal{Y}$ is the relationship label. The objective is to infer:

$$\hat{y} = \arg\max_{c \in \mathcal{Y}} P\big(y = c \mid x_1, x_2, \tau\big) \qquad (1)$$

The label space $\mathcal{Y}$ defines four hierarchical proximity tiers: (1) **Same Figure** (composite panels or true figure–caption pairs), (2) **Same Paper** (distinct artifacts from the same paper), (3) **Related Papers** (artifacts from papers connected via citations or topical overlap), and (4) **Unrelated Papers** (independent publications).

Dataset Breakdown: The benchmark comprises labeled training pairs and an unlabelled test set evaluated blindly via Kaggle. We partitioned the labeled data into training and validation sets using stratified sampling to strictly preserve both the class distribution and the proportions of each input modality. The complete breakdown across splits, classes, and modality configurations is detailed in Table 1.

4 Methodology

We evaluate model architectures across four progressive stages on all pair modalities (`ii`, `ti`, `tt`).

4.1 Architectural Progression

The benchmarked systems are organized into four paradigms:

- **Baselines (E0):** Zero-shot cosine similarity over frozen OpenAI text and SigLIP multimodal embeddings.
- **Biencoders and Cross-Encoders (E1):** Dual-tower metric models pairing text (ASTROBERT) and vision (SigLIP-1/2), alongside cross-encoders (PaliGemma).
- **Generative VLMs (E2):** Unified autoregressive decoders like Qwen3-VL-8B, Qwen3-VL-32B, and Qwen3.8-27B (checkpoint: `unsloth/Qwen3.8-27B-unsloth-bnb-4bit`) (Wang et al., 2024)—adapted via 4-bit NF4 and 16-bit LoRA.
- **Specialist and Cascade (E3):** A targeted 2-class adapter for the `same_paper` vs. `related_papers` boundary, alongside a two-level hierarchical cascade.

Table 2 summarizes the evaluated model configurations.

4.2 Training and Adaptation

Biencoder Fusion and Unfreezing: For dual encoders, visual (e_v) and textual (e_t) projections are combined via difference concatenation: $e_{\text{fused}} = [e_v \parallel e_t \parallel |e_v - e_t|]$. While base layers remain frozen, the top four transformer blocks of SigLIP-2 are unfrozen during training.

Parameter-Efficient VLM Adaptation: VLMs were fine-tuned via UNSLOTH (Unsloth, 2024) with LoRA (W_q, W_k, W_v, W_o) on two NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB VRAM each). The training of our best model (Qwen3.8-27B) took approximately 9.5 GPU-hours for 10 epochs, updating 124.4 million out of 27.4 billion parameters (0.45%). We benchmarked 4-bit NF4 quantization (Dettmers et al., 2024) against unquantized 16-bit LoRA and swept adapter ranks ($r \in \{32, 64\}$). We used the AdamW 8-bit optimizer with a learning rate of 1×10^{-4}, $\alpha = 16$, and a batch size of 4 with 8 gradient accumulation steps. At test time, inference validation took approximately 1 hour for 10,000 samples.

Data Augmentation: Training incorporated cross-modal Text–Figure (`ti`) sample upsampling to rebalance pair modalities. Although `ti` pairs constitute 40% of the dataset, initial cross-encoder validation showed it was the weakest modality (Macro-F1 of 0.3604). Therefore, `ti` pairs were deliberately upsampled to improve performance. We also used pair-order permutation ($A \leftrightarrow B$) to enforce symmetric invariance.

4.3 Inference and Post-Processing

Next-Token Logit Classification: To eliminate autoregressive generation latency, classification probabilities for $y \in \{0, 1, 2, 3\}$ were extracted directly via a 4-way softmax over the first-step vocabulary logits of the corresponding label tokens.

Test-Time Augmentation and Calibration: At test time, predictions were symmetrized via pair-order averaging: $P_{\text{sym}} = \frac{1}{2}(P(y \mid A, B) + P(y \mid$

	Class Distribution (Labeled)				Modality Distribution			
Split	**Same Fig.**	**Same Paper**	**Related**	**Unrelated**	**Fig–Fig (`ii`)**	**Text–Fig (`ti`)**	**Text–Text (`tt`)**	**Total**
Train	850	2,550	2,550	2,550	2,550 (30%)	3,400 (40%)	2,550 (30%)	8,500
Validation	150	450	450	450	450 (30%)	600 (40%)	450 (30%)	1,500
Test	—	—	—	—	3,000 (30%)	4,000 (40%)	3,000 (30%)	10,000
Total	**1,000**	**3,000**	**3,000**	**3,000**	**6,000 (30%)**	**8,000 (40%)**	**6,000 (30%)**	**20,000**

Table 1: AstroCLIMB dataset breakdown across target classes and input modalities. Ground-truth test labels were withheld for blind evaluation on the Kaggle leaderboard.

Model Family / Setup	**Paradigm**	**Precision**	**Adaptation Technique**	**Modalities**
OpenAI / SigLIP Embeddings	Metric Baseline	FP16 / FP32	Zero-Shot Cosine Similarity	`tt`, `ti`, `ii`
BERT / AstroBERT + ViT	Biencoder / CE	FP16	Supervised Metric Head	`ti`, `ii`
AstroBERT + SigLIP-1 / 2	Biencoder	FP16	Unfreeze Last 4 Blocks + Aug	`ti`, `ii`
PaliGemma	Cross-Encoder	FP16	End-to-End Supervised Tuning	`tt`, `ti`, `ii`
Qwen3-VL-8B / Qwen3-VL-32B	Generative VLM	4-bit (NF4)	Unsloth LoRA Adapter	`tt`, `ti`, `ii`
Qwen3.8-27B	Generative VLM	4-bit / 16-bit	Unsloth LoRA (Rank Sweeps)	`tt`, `ti`, `ii`
Specialist Adapter	Targeted VLM	4-bit (NF4)	Focused 2-Class Boundary SFT	`tt`, `ti`, `ii`

Table 2: Systematic overview of benchmarked architectures across evaluation stages (E0–E3).

$B, A)$). Finally, validation-fitted additive logit offsets were applied prior to argmax selection to calibrate against class skew.

Threshold Calibration for Embedding Baselines (E0): For the frozen embedding baselines (E0a and E0b), predictions are derived by partitioning continuous cosine similarities $s \in [-1, 1]$ across three strictly ordered decision thresholds $\tau_1 < \tau_2 < \tau_3$:

$$\hat{y}(s) = \begin{cases} \texttt{unrelated_papers} & \text{if } s < \tau_1, \\ \texttt{related_papers} & \text{if } \tau_1 \le s < \tau_2, \\ \texttt{same_paper} & \text{if } \tau_2 \le s < \tau_3, \\ \texttt{same_figure} & \text{if } s \ge \tau_3. \end{cases}$$

Because these frozen metric models lack parameterized classification heads, the cut points are calibrated on the training split to maximize Macro-F1 under this decision rule. We compare two threshold search procedures: (i) a uniform grid search over all ordered candidate triples, and (ii) Nelder–Mead optimization using an unconstrained, order-preserving reparameterization:

$$\tau_1 = x_1, \qquad \tau_2 = \tau_1 + e^{x_2}, \qquad \tau_3 = \tau_2 + e^{x_3},$$

initialized from midpoints between adjacent class-conditional similarity quantiles (the 75th percentile of class k and the 25th percentile of class $k+1$). The optimal thresholds are fixed and evaluated directly on validation and test splits. Subsequent supervised models (E1–E3) bypass this procedure, relying instead on direct softmax logit classification.

5 Experiments

We evaluate our modeling progression across all three pair modalities (`ii`, `ti`, `tt`) on the AstroCLIMB benchmark. Table 3 summarizes the official test set Macro-F1 scores across all development phases. Generative VLM experiments were evaluated using seed 3407, whereas bi-encoder and cross-encoder experiments used seed 42.

5.1 Experimental Setup and Protocol

Models are evaluated on test Macro-F1 across all four classes, alongside modality-stratified breakdowns. Training and evaluation were conducted on NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB VRAM) using the stratified validation split. Deterministic first-token logit extraction and test-time augmentation follow the protocol in Section 4.3.

5.2 Baselines and Discriminative Encoders (E0–E1)

Zero-shot cosine similarity over frozen embeddings (E0) yields the lowest overall performance, confirming that general-purpose metric spaces fail to capture hierarchical citation proximity (Table 3). Supervised bi-encoders and cross-encoders (E1) improve systematically over these baselines as visual towers advance from ViT to SigLIP-2, with further gains from unfreezing upper visual blocks and cross-modal Text–Figure augmentation. However, modular architectures ultimately plateau well below unified VLMs due to representation bottlenecks in late fusion.

ID	Stage	Experiment / Configuration	Test mF1
E0a	Embedding Baseline	OpenAI embedding cosine similarity	0.26608
E0b	Embedding Baseline	SigLIP embedding cosine similarity	0.26619
E1a	Biencoder / CE	AstroBERT + ViT biencoder	0.38688
E1b	Biencoder / CE	AstroBERT + SigLIP-1 biencoder	0.40393
E1c	Biencoder / CE	PaliGemma cross-encoder	0.38836
E1d	Biencoder / CE	AstroBERT + SigLIP-2 biencoder	0.44641
E1e	Biencoder / CE	AstroBERT + SigLIP-2 (unfreeze last 4 blocks)	0.45818
E1f	Biencoder / CE	AstroBERT + SigLIP-2 + Text–Figure sample augmentation	0.46935
E2a	VLM SFT	Qwen3-VL-8B	0.70171
E2b	VLM SFT	Qwen3-VL-32B	0.72534
E2c	VLM SFT	Qwen3.8-27B (4-bit LoRA, $r = 32$)	**0.73925**
E2d	VLM SFT	Qwen3.8-27B (increased LoRA rank, $r = 64$)	0.71072
E2e	VLM SFT	Qwen3.8-27B + data augmentation	0.73889
E2f	VLM SFT	Qwen3.8-27B (16-bit LoRA, $r = 32$)	0.73300
E3a	Specialist	Specialist model (focused 2-class adapter)	**0.73929**
E3b	Cascade	Two-level uncalibrated cascade	0.37401

Table 3: Catalogue of experiments with official test macro-F1 across all development phases.

5.3 Vision–Language Fine-Tuning (E2a–E2f)

Unified autoregressive VLMs substantially outperform modular encoders (Stage E2). Performance scales with model generation and capacity: the newer Qwen3.8-27B backbone consistently outperforms both Qwen3-VL-8B and the nominally larger Qwen3-VL-32B, reflecting stronger foundational pretraining. Across adaptation settings, 4-bit LoRA provides an effective, parameter-efficient baseline. Unquantized 16-bit training, higher adapter rank, and symmetric data augmentation all perform within a narrow score band; given our limited dataset size, these minor variations reflect optimization sensitivity and sample variance rather than fundamental scaling limits.

5.4 Specialist Modeling and Cascading (E3a–E3b)

To address the primary confusion between `same_paper` and `related_papers`, we trained an auxiliary 2-class specialist adapter (E3a). However, introducing this specialist did not meaningfully improve performance over the monolithic generalist baseline, yielding virtually identical results within expected run-to-run variance. This suggests that isolated artifact pairs may lack sufficient document-level context to resolve this subtle boundary, regardless of focused re-ranking. In contrast, an uncalibrated two-level cascade (E3b) degrades performance severely, falling to a macro-F1 of 0.374. This cascading failure happens because the first stage (differentiating `related_papers` vs `unrelated_papers`) achieved a weak macro-F1 of 0.55, while the second stage achieved ~ 0.75. The poor initial routing meant candidate pairs were incorrectly discarded early, causing these early-stage routing errors to propagate irreversibly.

6 Results and Analysis

6.1 Architectural Comparisons and Adaptation Ablations

Across the evaluated paradigms (Table 3), zero-shot embedding similarity performs worst, while supervised bi-encoders and cross-encoders markedly improve results. However, neither matches the performance of generative VLMs. Within the Qwen family, the newer Qwen3.8-27B consistently outperforms earlier backbones, including both the 8B model and the nominally larger Qwen3-VL-32B, benefiting from stronger foundational pretraining.

When adapting Qwen3.8-27B, ablations over LoRA rank and precision (4-bit vs. 16-bit) remained within a narrow score band. Given the limited size of the dataset, we cannot conclusively determine whether scaling rank or precision systematically improves performance or simply reflects optimization noise. Under these data-constrained conditions, standard 4-bit LoRA provides an effective, memory-efficient baseline.

6.2 Modality-Level Breakdown

Validation results follow a strict hierarchy: `tt` $>$ `ti` $>$ `ii` (Table 4). Text–Text (`tt`) pairs benefit from explicit domain terminology (e.g., coordinates and instruments), while Text–Figure (`ti`) pairs effectively ground text into visual elements.

Modality	Val mF1	Val Acc.	n
Overall	0.7439	0.7167	1,500
Text–Text (`tt`)	0.7798	0.7844	450
Text–Figure (`ti`)	0.7014	0.7067	600
Figure–Figure (`ii`)	0.6635	0.6622	450

Table 4: Modality-stratified validation metrics for Qwen3.8-27B.

Conversely, Figure–Figure (`ii`) pairs are the most difficult because independent studies often share standardized visualization conventions (e.g., contours, spectra, and color-magnitude plots) that create spurious visual similarity.

6.3 Class-Level Performance and Error Analysis

Validation confusion matrices show that errors are heavily concentrated along specific boundaries. While `same_figure` (via layout continuity) and `unrelated_papers` (via lexical divergence) are classified reliably, distinguishing `same_paper` from `related_papers` causes the vast majority of misclassifications. Papers that cite each other or target identical astronomical objects share nearly identical vocabulary and plotting styles, making isolated artifact pairs fundamentally ambiguous without broader document context.

7 Conclusion and Future Work

We investigated an empirical progression of multimodal architectures for scientific citation linking on AstroCLIMB. While metric dual encoders are constrained by cross-modal representation gaps, parameter-efficient fine-tuning of generative vision–language models achieves competitive performance, reaching approximately 0.74 test Macro-F1. Our experiments demonstrate that model performance is not primarily limited by architectural capacity, but rather by visual ambiguity in Figure–Figure pairs and the subtle semantic boundary separating artifacts from the same paper versus related papers.

Future work includes leveraging full-paper metadata, domain-specific vision encoder pretraining, and exploring dedicated decision models (e.g., KEV/JEV) for routing, enabling more reliable model selection and reducing early routing errors.

8 Limitations

Lack of Document Context: Looking at artifact pairs alone is not enough. Without surrounding text, methodology, or author metadata, telling apart *same paper* from *related papers* is very difficult—especially when two papers study the same celestial object or use the same data pipeline.

Visual Domain Gap on Figure–Figure (`ii`) Pairs: Standard vision models are not trained on astronomy plots. On `ii` pairs, the vision encoder often focuses on shared plot styles (such as colorbars, axes, and line layouts) instead of the actual science, causing false matches between unrelated papers.

Simple Decision Rule and Cascade Failures: Our system uses a simple argmax rule over logits rather than an adaptive decision model. As our cascade test (E3b) showed, hard multi-stage decisions fail easily: errors in the first step permanently discard correct pairs. A better decision model must adjust confidence thresholds for each modality, since performance varies widely ($tt > ti > ii$).

Inference Speed and Scalability: Although 4-bit LoRA allows training on a single GPU, running a 27B model is still slow and resource-heavy. This makes it difficult to scale the system across millions of papers in the NASA ADS.

Lack of Explanations: Taking predictions directly from the first token makes inference fast, but the model cannot output reasoning steps or explain why two items are linked.

Domain Focus: Our findings are specific to astronomy publications. They may not directly apply to fields like biomedicine or computer science, which use different types of figures and citation styles.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b vlm for transfer.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Felix Grezes, Thomas Allen, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Shinyi Chen, Jennifer Koch, Taylor Jacovich, and Pavlos Protopapas. 2022. Improving astrobert using semantic textual similarity.

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.

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.

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.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.

Unsloth. 2024. Unsloth: Fast finetuning for large language models. `https://github.com/unslothai/unsloth`.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.

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 (ICCV)*, pages 11975–11985.