Modality-Specific and Relation-Aware Modeling for Scientific Figure Linking

Bodhisatta Maiti
Independent Researcher
Durham, North Carolina, USA

Debshree Chowdhury
Independent Researcher
Durham, North Carolina, USA

Abstract

We present a modality-specific system for the AstroCLIMB shared task, which predicts relations among astronomical figures and captions. Caption–caption, image–image, and image–caption pairs use different combinations of frozen pretrained representations and supervised classifiers. Image–caption prediction first applies a Qwen3-VL 4B/8B exact-correspondence gate before classifying the remaining paper relations. Subsequent improvements add astronomy-domain sentence embeddings to the caption branch and a learned visual-projection specialist to the image branch. Across eight Kaggle submissions, macro-F1 increases from 0.3641 to 0.6096. Adding astronomy-specific caption features improves five-fold pooled out-of-fold (OOF) macro-F1 from 0.6071 to 0.6388; the image specialist offers a smaller incremental test-set gain. The direct citation relation, `related_papers`, remains the most difficult. Results support selecting representations and decision mechanisms according to modality, while highlighting the limitations of recovering citation links from isolated content.

1 Introduction

Figures are central to scientific communication. They condense experimental results, comparisons, trends, and domain-specific evidence into a form that is often easier for a reader to interpret than the surrounding prose. At the same time, much of this information remains difficult to exploit computationally in scholarly search and literature analysis. The AstroCLIMB shared task addresses this gap through an unusual multimodal formulation: given a pair of scientific objects—two figures, two captions, or a figure and a caption—a system must determine whether they correspond to the same figure, originate from the same paper, come from citation-linked papers, or are unrelated (Grezes et al., 2026). The task therefore asks models to recover part of the structure of the scholarly citation graph using only the content of scientific figures and captions.

Motivated by these differences, we use separate modeling strategies for each modality and relation type rather than a single end-to-end classifier. For caption-caption pairs, we combine scientific text representations with vision-language embeddings and lightweight lexical features. For image-image pairs, we use a higher-capacity, higher-resolution SigLIP2 visual encoder, motivated by the improved semantic and visual representation capabilities of the SigLIP2 family (Tschannen et al., 2025). Image-caption pairs are handled differently. We first use Qwen3-VL (Bai et al., 2025) as a dedicated exact-correspondence detector for the `same_figure` relation. Pairs not identified as exact matches are then passed to a separate classifier for the remaining scholarly relations. In the final system, the exact-match decision combines scores from Qwen3-VL 4B and 8B models, while the residual classifier combines complementary scientific-text and multimodal signals. We deliberately avoid DOI, citation-graph, or other source-metadata lookup at inference time so that predictions are based on the figure-caption content provided by the benchmark.

This specialization substantially outperformed a uniform representation. Our initial frozen SigLIP2 baseline obtained 0.3641 Kaggle macro-F1; adding scientific-text features increased this to 0.4472, and introducing the dedicated image–caption correspondence gate raised it to 0.5660. Stronger visual representations and score-level combinations produced a final score of 0.6096 after astronomy-specific caption features and a complementary visual specialist. Across experiments, `related_papers` remained the most difficult relation, highlighting the gap between semantic relatedness and direct citation linkage.

Our work makes three main contributions. First, we develop a modality-specialized architecture

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 262–271
November 10, 2026 ©2026 Association for Computational Linguistics

for AstroCLIMB that assigns different representations and decision mechanisms to caption-caption, image-image, and image-caption pairs instead of forcing all three modalities through a shared classifier. Second, we show that vision-language models are particularly effective when their role is narrowed to exact figure-caption correspondence; direct use of the same models for broader paper-provenance classification is considerably less reliable. Third, through controlled ablations, cross-fitted predictions, and a separately evaluated Qwen gate, we identify direct citation prediction as the main remaining bottleneck and show that stronger generic visual representations improve same-paper and unrelated-paper discrimination much more consistently than citation-linked-paper recognition. Together, these results support a view of scientific multimodal relation prediction as a collection of distinct inference problems rather than a single similarity task.

2 Related Work

Scientific-document representations have used citation relationships during pretraining. SPECTER learns document embeddings from citation-informed supervision (Cohan et al., 2020); SciRepEval studies scientific-text representations across multiple tasks and formats (Singh et al., 2023). In contrast, our inference setting supplies only a figure or caption pair, not full papers or their citation neighborhoods. We apply pretrained representations directly to captions, without access to the underlying citation graph.

For visual and cross-modal evidence, SigLIP2 provides pretrained image–text embeddings (Tschannen et al., 2025), and Qwen3-VL supports multimodal question answering (Bai et al., 2025). We use these models for distinct purposes: SigLIP2 supplies pairwise features, while Qwen3-VL provides continuous evidence of exact figure–caption correspondence. Our experiments also compare an astronomy-specialized sentence encoder (NeuML, 2026) with more general scientific-text embeddings. The AstroCLIMB benchmark defines provenance by paper identity or direct citation (Grezes et al., 2026); such relationships cannot in general be identified from visual similarity alone.

3 Task and Dataset

3.1 Task Definition

AstroCLIMB is a multiclass, single-label classification task designed to recover relationships between scientific objects drawn from astronomy papers (Grezes et al., 2026). Each example contains a pair of objects, where an object is either a scientific figure or its textual caption. This results in three input configurations: caption–caption (CC), image–image (II), and image–caption (IC).

Each pair is assigned exactly one of four labels:

- `same_figure`: the image and caption correspond to the same scientific figure. This relation is defined only for IC pairs.
- `same_paper`: the two objects originate from the same source paper.
- `related_papers`: the objects originate from different papers for which one paper directly cites the other.
- `unrelated_papers`: none of the preceding relationships holds.

The relations are symmetric with respect to the order of the two input objects (Grezes et al., 2026). The distinction between `related_papers` and the other classes is particularly important for our study. Unlike `same_figure` or `same_paper`, a direct citation edge is not necessarily recoverable from semantic similarity alone: two papers may investigate the same object or use similar methods without directly citing one another.

3.2 Released Data

The Kaggle release[1] contains 10,000 labeled training pairs and 10,000 test pairs (Grezes et al., 2026). Captions are provided as English text, whereas figures are represented as encoded image strings that can be converted back to image objects. We use the Kaggle pair dataset as the sole source of task-specific supervision.

Inspection of the released training split shows a deliberately structured distribution across modalities. The CC and II subsets each contain 3,000 examples, with 1,000 examples from each of `same_paper`, `related_papers`, and `unrelated_papers`. The IC subset contains 4,000

[1] https://www.kaggle.com/competitions/astroclimb

Pair	Same fig.	Same paper	Related	Unrel.
CC	0	1000	1000	1000
II	0	1000	1000	1000
IC	1000	1000	1000	1000
Total	1000	3000	3000	3000

Table 1: Class distribution in the 10,000-pair AstroCLIMB training split. CC, II, and IC denote caption–caption, image–image, and image–caption pairs, respectively.

examples, with 1,000 examples from each of the four classes. The test split follows the same modality counts—3,000 CC, 3,000 II, and 4,000 IC pairs—although its class labels are hidden.

The broader AstroCLIMB resource[2] also provides figure-caption records and bibliographic metadata from the underlying astronomy corpus (Grezes et al., 2026). We intentionally do not use source DOIs, citation adjacency information, paper titles, author metadata, or other corpus-level identifiers as task supervision or as lookup information at inference time. In particular, we do not recover test-pair relationships by matching test objects back to the underlying citation graph. This restriction keeps the system focused on the information available in the benchmark pair itself and avoids introducing paper-level graph information that could trivially reveal some target relations.

3.3 Evaluation Protocol

The official metric is macro-F1 over the four target classes (Grezes et al., 2026). We use stratified five-fold out-of-fold (OOF) evaluation for lightweight classifiers and separate development and validation subsets for the IC exact-correspondence component. Submitted labels are determined from model scores without imposing fixed class quotas. CV splits are stratified at the *pair* level, not guaranteed to be disjoint at figure, caption, or source-paper level. Thus OOF estimates support within-benchmark ablation comparisons but do not establish unseen-paper generalization. We report pooled OOF macro-F1 separately from the mean and sample standard deviation of the five fold scores; hidden-test Kaggle values are reported separately.

4 System Architecture

A single classifier is unlikely to receive equally useful evidence from all three AstroCLIMB input configurations. A caption–caption pair is primarily a scientific-language problem, an image–image pair depends on visual and stylistic evidence, and an image–caption pair additionally permits direct cross-modal correspondence. We therefore use a modality-specialized architecture in which CC, II, and IC pairs follow separate prediction paths. All pretrained models are kept frozen; only lightweight downstream classifiers and decision rules are fit using the AstroCLIMB training pairs.

4.1 Pair Representation

For two normalized embeddings $\mathbf{e}_1$ and $\mathbf{e}_2$ from a shared representation space, we construct a symmetric pair representation

$$\phi(\mathbf{e}_1, \mathbf{e}_2) = [|\mathbf{e}_1 - \mathbf{e}_2|; \mathbf{e}_1 \odot \mathbf{e}_2; \cos(\mathbf{e}_1, \mathbf{e}_2)], \quad (1)$$

where $\odot$ denotes element-wise multiplication. The absolute difference and element-wise product preserve dimension-wise agreement between the two objects, while cosine similarity provides a compact global similarity signal. Because the benchmark relations are symmetric, the representation does not depend on the order of the two inputs.

The resulting features are classified using XGBoost (Chen and Guestrin, 2016). In the CC configurations we evaluated, the XGBoost head improved over the simpler classifier when combined pretrained representations were available, while remaining inexpensive to train.

4.2 Caption–Caption Specialist

For CC pairs, both inputs contain scientific language, so we combine a general vision-language text representation with a scientific-document representation. Each caption is first embedded using the text encoder of SigLIP2 (Tschannen et al., 2025). In parallel, we encode captions using SPECTER2, the scientific-document representation family introduced with the SciRepEval benchmark (Singh et al., 2023).

We construct separate pair features using Eq. 1 for the SigLIP2 and SPECTER2 representations and concatenate them with lightweight lexical overlap features. The latter capture surface-level evidence such as shared tokens and relative caption length. The combined representation is passed to a three-class XGBoost classifier that predicts `same_paper`, `related_papers`, or `unrelated_papers`. The `same_figure` relation is not defined for CC pairs.

[2] https://huggingface.co/datasets/adsabs/AstroCLIMB

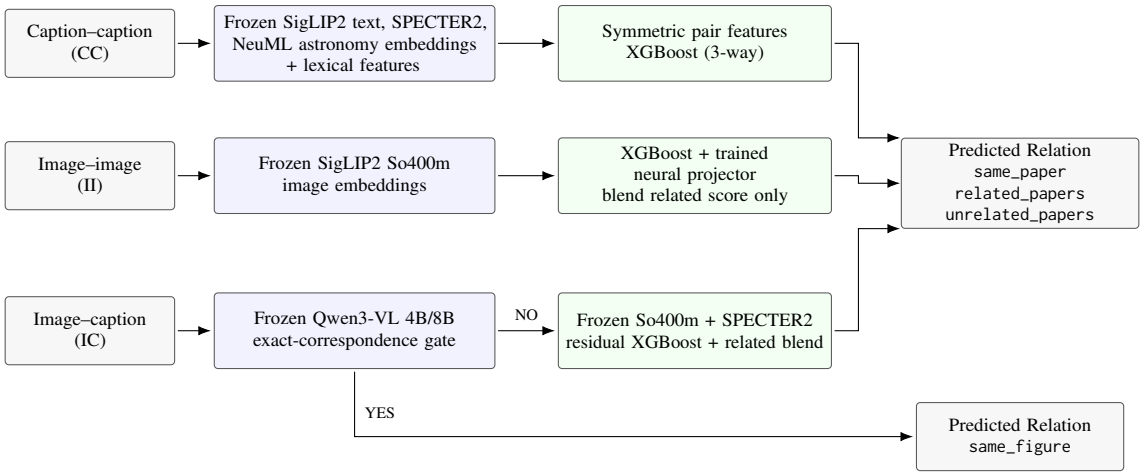


Figure 1: Final V8 modality-specific system. Pretrained encoders are frozen. The CC and II paper-relation classifiers and the II projection specialist learn from Kaggle training pairs. The IC gate routes exact matches directly to `same_figure`; remaining IC pairs use the unchanged V6 residual classifier to predict `same_paper`, `related_papers`, or `unrelated_papers`.

In V7, we additionally encode each caption with the frozen astronomy-specialized `NeuML/astrobert-small-embeddings` sentence model (NeuML, 2026). Its native sentence representation supplies a third symmetric pair-feature block alongside SigLIP2 and SPECTER2, while the original lexical features are retained. We compare replacing and augmenting SPECTER2; adding the astronomy-specific block performs best on OOF evaluation.

This branch is intended to separate topical similarity from stronger paper-level correspondence. In particular, the scientific-text representation provides information that is not well captured by the shorter text context of the vision-language encoder alone.

4.3 Image–Image Specialist

For II pairs, we use the image encoder from SigLIP2 So400m with 384×384 input resolution (Tschannen et al., 2025). Each image is encoded independently, and the two visual embeddings are converted into the symmetric representation in Eq. 1. An XGBoost classifier then predicts one of `same_paper`, `related_papers`, or `unrelated_papers`.

The evaluated So400m configuration uses both a larger encoder and higher image resolution; our experiments do not isolate the contribution of either change. Scientific figures contain small labels, panel structure, plotting conventions, and other visual cues that may be relevant to provenance. We use the So400m representation directly rather than concatenating it with the smaller SigLIP2 representation; our ablations show that naive feature-level fusion does not improve II performance.

For V8, we add a neural specialist on cached, frozen So400m embeddings. A shared two-layer projection maps each image to a normalized task-adapted space; the symmetric representation in Eq. 1 feeds a three-class multilayer perceptron. We average five trained models' scores at test time and blend only the `related_papers` score with the five-fold XGBoost ensemble (neural weight 0.90); the other two class scores remain those of XGBoost. The weight was chosen by searching OOF predictions, so the selected OOF blend score is not an independent estimate of generalization.

4.4 Image–Caption Routing

IC pairs require a different treatment because they are the only modality for which `same_figure` is possible. A flat four-class classifier forces one model to solve two qualitatively different problems: exact visual-text correspondence and paper-level provenance. We instead separate these decisions.

Exact correspondence gate. We first determine whether the caption describes the exact input figure. Qwen3-VL (Bai et al., 2025) is prompted with the figure and caption and asked a binary question: whether the caption corresponds to that exact figure. Rather than relying on a generated class label, we run deterministic generation (`max_new_tokens=1`, no sampling) immediately after the assistant-generation prefix. Both tokenizers encode uppercase `YES` and `NO` as single tokens

(IDs 14004 and 8996). From their first-step logits, ℓ_Y and ℓ_N, we compute the two-token normalized score $p = \exp(\ell_Y)/(\exp(\ell_Y) + \exp(\ell_N))$. This is a decision score, not a separately calibrated probability.

Our final system combines scores from the 4B and 8B Qwen3-VL variants. Let p_4 and p_8 denote their respective two-token normalized scores. Before the logit transformation, scores are clipped to $[10^{-8}, 1 - 10^{-8}]$ to avoid infinities. We combine them in log-odds space,

$$s_{\mathrm{IC}} = \sigma\left((1-\alpha)\mathrm{logit}(p_4) + \alpha\mathrm{logit}(p_8)\right), \quad (2)$$

where σ is the logistic function and α is selected on development data. If s_{IC} meets or exceeds a tuned threshold, the pair is assigned `same_figure`. Using the continuous model scores proved substantially more reliable than treating the generated binary answer as the final decision.

Residual paper-relation classifier. Pairs that do not pass the exact-correspondence gate are classified into `same_paper`, `related_papers`, or `unrelated_papers`. The image and caption are embedded in the shared SigLIP2 So400m space, and their cross-modal pair representation is constructed using Eq. 1. We additionally append the SPECTER2 representation of the caption, providing a longer-context scientific-language signal alongside the visual-text similarity features. An XGBoost classifier produces the three paper-relation scores.

We retain a small amount of complementary information from the earlier, lower-capacity IC classifier for the `related_papers` score. This probability-level combination is performed only after classification rather than by concatenating the underlying embedding spaces. The resulting three-way scores are used only when the Qwen correspondence gate rejects `same_figure`.

5 Experimental Setup

5.1 Pretrained Encoders

We use pretrained encoders as frozen feature extractors throughout the system. The initial SigLIP2 representation is obtained from `google/siglip2-base-patch16-224`, which produces 768-dimensional embeddings from 224×224 inputs. For the stronger visual and cross-modal branches, we use `google/siglip2-so400m-patch14-384`, with 1152-dimensional embeddings and 384×384 image inputs (Tschannen et al., 2025).

Scientific-text representations include `allenai/specter2_base` and, from V7 onward, `NeuML/astrobert-small-embeddings`. We use the base SPECTER2 encoder directly, without task-specific adapters, and represent each caption using the hidden state of the first token. SPECTER2 was introduced as part of the SciRepEval work on transferable scientific-document representations (Singh et al., 2023).

For IC exact-correspondence detection, we use `Qwen/Qwen3-VL-4B-Instruct` and `Qwen/Qwen3-VL-8B-Instruct` (Bai et al., 2025). The 8B model is loaded with 4-bit NF4 quantization and bfloat16 computation to fit within a single-GPU Colab environment. These models are used only for inference; no Qwen3-VL parameters are fine-tuned on AstroCLIMB.

All pretrained encoders remain frozen and embeddings are cached; the V8 projection and downstream classifiers are trained on Kaggle pairs in a single-GPU Google Colab environment.

5.2 Supervised Classifiers

The CC and residual IC branches use XGBoost classifiers; the II branch combines XGBoost and a neural specialist (Chen and Guestrin, 2016). Training examples are represented using the symmetric pair features described in Section 4, together with the additional modality-specific representations used by each branch.

For CC and II, the classifier predicts the three paper-level relations: `same_paper`, `related_papers`, and `unrelated_papers`. For IC Stage 2, training is restricted to the 3,000 non-`same_figure` IC examples, since the first-stage VLM gate is responsible for detecting exact figure–caption correspondence.

We use five-fold stratified cross-validation to generate out-of-fold predictions for the supervised classifiers. These predictions are used for model comparison and for constructing the probability-level combination in the IC residual classifier. The II related-score blend weight was also selected from OOF predictions; it is not included as an independently cross-validated result. Test predictions are generated by the corresponding trained branch and are not constrained to match the class proportions observed in the training set.

5.3 Image–Caption Development Protocol

The IC branch requires additional calibration because the output of the exact-correspondence detector is a continuous score rather than a direct four-class prediction. We therefore separate IC development from the out-of-fold evaluation used for the lightweight classifiers.

We construct a balanced 200-example IC development set containing 50 examples from each of the four target classes. This subset is used for selecting the exact-match threshold and, for the V6 and later systems, the relative contribution of the 4B and 8B Qwen3-VL models. We additionally maintain a separate balanced set of 400 IC examples, with 100 examples per class, for evaluating the exact-correspondence gate. This subset is not used when selecting the Qwen3-VL ensemble weight or correspondence threshold. For the hybrid evaluations on these rows, non-`same_figure` Stage-2 outputs are retrieved by row ID from the five-fold OOF predictions: the fitted classifier producing each such score excluded that row. Nevertheless, the residual-score blend coefficient was selected from pooled OOF predictions that included these rows; therefore, the hybrid results are cross-fitted diagnostics, not a fully independent validation of all model-selection choices.

For a figure–caption pair, each Qwen3-VL model receives the figure together with the caption and is asked whether the caption describes that exact figure. We obtain the two-token normalized correspondence decision score from the next-token `YES/NO` logits as defined in Section 4. For the final ensemble, the two model scores are combined in log-odds space as defined in Eq. 2. Development-set search selected an 8B ensemble weight of

$$\alpha = 0.65 \tag{3}$$

and a correspondence threshold of

$$\tau = 0.01778. \tag{4}$$

Pairs with $s_{\mathrm{IC}} \geq \tau$ are labeled `same_figure`; the remaining pairs are passed to the three-way IC Stage-2 classifier.

This protocol is important because several apparently stronger configurations on the 200-example development subset did not transfer to the held-out 400-example set. We therefore use the latter only as a validation check and do not retune the final system on it.

Version	Main change	Macro-F1
V1	Frozen SigLIP2 baseline	0.3641
V2	Scientific-text features	0.4472
V3	Qwen3-VL IC routing	0.5660
V4	So400m visual features	0.5843
V5	IC blend and gate calibration	0.5919
V6	Qwen 4B+8B ensemble	0.5982
V7	Astronomy-specific CC features	0.6081
V8	II neural specialist + V7 CC	**0.6096**

Table 2: Kaggle hidden-test leaderboard progression. Scores are rounded to four decimal places from submission records; the timing-only sample submission is omitted.

5.4 Residual IC Score Combination

The final IC Stage-2 classifier primarily uses the SigLIP2 So400m representation together with the SPECTER2 caption representation. An earlier IC classifier based on the lower-capacity representation retained complementary signal for `related_papers`. We therefore combine only the `related_papers` probability from the two classifiers:

$$p_{\mathrm{rel}} = \lambda p_{\mathrm{rel}}^{\mathrm{old}} + (1 - \lambda) p_{\mathrm{rel}}^{\mathrm{new}}, \tag{5}$$

with

$$\lambda = 0.35. \tag{6}$$

The `same_paper` and `unrelated_papers` probabilities are taken directly from the stronger So400m-based classifier. The final Stage-2 label is the class with the highest resulting score.

6 Results and Ablations

We report two types of results. Kaggle leaderboard scores measure end-to-end performance on the unlabeled test pairs, whereas local results are obtained either through five-fold out-of-fold (OOF) predictions or through the IC validation subsets described in Section 5. Since these evaluation sets differ, their absolute scores should not be compared directly.

6.1 End-to-End System Progression

Table 2 reports the eight completed Kaggle submissions (excluding a timing-only sample submission). The largest early gain comes from introducing the exact figure–caption gate (V3). V7 adds astronomy-specific sentence features only to CC, raising the hidden-test score from 0.5982 to 0.6081. V8 adds an II neural relatedness specialist while keeping the IC branch unchanged; its further gain is small (0.6096). These leaderboard differences are not paired statistical tests, since test labels are not released.

Branch	Configuration	$\mu \pm \sigma$	OOF
CC	SigLIP2 + linear	$.4059 \pm .0127$	.4058
CC	+ SPECTER2 + lexical	$.5966 \pm .0164$	.5970
CC	+ XGBoost (V6)	$.6070 \pm .0105$	.6071
CC	+ NeuML add (V7)	$\mathbf{.6386 \pm .0079}$	**.6388**
II	SigLIP2 base	$.4377 \pm .0141$	.4377
II	So400m XGBoost	$.4992 \pm .0172$	.4994
II	Base + So400m	$.4913 \pm .0179$	.4914
II	Neural projection alone	$.4887 \pm .0180$	.4917
IC	Flat classifier	$.3527 \pm .0101$	.3526
IC	Residual Stage 2 (V6)	$.4264 \pm .0172$	.4266

Table 3: Local five-fold stratified CV ablations. Mean $\pm$ sample SD uses fold macro-F1 values, while OOF reports pooled predictions. II's weight-selected V8 blend is excluded from this independent-fold table.

From V1 to V8, the absolute macro-F1 gain is 0.2455 (about 67.4% relative to V1). The central architectural change was the early separation of exact correspondence from provenance; later revisions add targeted representation and score-level improvements.

6.2 Modality-Specific Ablations

Table 3 reports pooled five-fold OOF macro-F1 and the mean $\pm$ sample SD of individual fold macro-F1. These are component-level experiments, not estimates of each complete V1–V8 pipeline. In particular, the IC hybrid experiments below use a separate 400-example Qwen evaluation with cross-fitted Stage-2 predictions, and appear in Table 5.

The clearest Phase-2 component gain is in CC: supplementing SigLIP2, SPECTER2, and lexical features with NeuML embeddings raises pooled OOF macro-F1 from 0.6071 to 0.6388. The CC improvement occurs on each of the five folds, although choosing the best observed configuration among multiple OOF ablations can introduce model-selection optimism. Alternative replacements and an `adsabs/astroBERT` variant (Grezes et al., 2021) did not exceed the NeuML-add configuration (Appendix A).

The original II So400m encoder raises OOF macro-F1 over the base image model, but naive feature concatenation is worse. The standalone neural projection reaches 0.4917, below the XGBoost 0.4994 baseline, yet improves the `related_papers` class (Table 6). The *OOF-selected* related-score-only blend reaches 0.5254, with weight chosen using those same predictions. It is reported as model-selection evidence only; the final Kaggle submission gains 0.0015 beyond V7.

Gate	Precision	Recall	F1
Qwen3-VL 4B	0.8824	0.9000	0.8911
Qwen3-VL 8B	0.8246	0.9400	0.8785
4B + 8B	**0.9029**	**0.9300**	**0.9163**

Table 4: Exact `same_figure` detection on the held-out 400-example IC validation set. The 4B row uses the calibrated gate employed before the final ensemble.

6.3 Exact Figure–Caption Matching

The behavior of Qwen3-VL further supports the decision to use the VLM as a specialist rather than as a general relation classifier. In an early balanced 60-example experiment, direct four-way Qwen3-VL classification obtained only 0.3416 macro-F1. However, the `same_figure` class alone achieved an F1 of 0.8750. We therefore reformulated the VLM component as binary exact-correspondence detection.

On the 200-example development set, Qwen3-VL-4B achieved an ROC-AUC of 0.9885 and a PR-AUC of 0.9643 for this binary task. The 8B model achieved still higher development values of 0.9943 ROC-AUC and 0.9863 PR-AUC. The larger model did not, however, translate into better held-out F1 when used alone.

As shown in Table 4, Qwen3-VL-8B alone favors recall at the cost of substantially lower precision. Combining the 4B and 8B scores produces a better balance and raises held-out exact-match F1 to 0.9163. This ensemble becomes the correspondence gate from V6 onward. The gate comparison is evaluated separately from the cross-fitted hybrid diagnostics in Table 5. A class-stratified bootstrap of the ensemble's retained 400-example confusion counts (93 true positives, 7 false negatives, 10 false positives) gives an approximate 95% interval of 0.877–0.952 for its F1. This interval characterizes the ensemble alone, not the paired improvement over 4B: aligned per-example predictions for the compared systems were not retained.

Importantly, the same models were much less reliable when asked to infer paper provenance directly. For example, an image–image same-paper detector based on Qwen3-VL achieved only 0.7494 ROC-AUC, and a small few-shot three-way II experiment obtained 0.2913 macro-F1. These results suggest that strong visual-language correspondence does not automatically translate into reliable citation or provenance reasoning.

IC hybrid variant	Diagnostic macro-F1
Qwen gate + old Stage 2	0.4863
Qwen gate + So400m Stage 2	0.5414
+ Stage-2 blend / calibration	0.5491

Table 5: IC hybrid diagnostics on 400 examples using the earlier Qwen3-VL 4B correspondence gate. Gate scores were evaluated on a separate validation set, while residual predictions for non-`same_figure` examples were generated out of fold. Because the related-score blending weight was selected using pooled OOF predictions that included these examples, the reported scores are not fully independent held-out estimates.

Configuration	Same	Rel.	Unrel.	F1
CC V6	.6813	.4498	.6903	.6071
CC V7 NeuML	.6803	.4936	.7424	.6388
II So400m	.5753	.3454	.5774	.4994
II neural alone	.5265	.4405	.5079	.4917
II blend*	.5667	.4378	.5717	.5254
IC residual Stage 2	.4540	.3425	.4833	.4266

Table 6: Per-class OOF F1. Labels abbreviate `same_paper`, `related_papers`, and `unrelated_papers`. *OOF-selected score-blend weight: descriptive model-selection result, not independent validation.

6.4 Relation-Level Behavior

Table 6 exposes which provenance relations change under V7/V8 ablations. The NeuML CC features particularly improve `related_papers` and `unrelated_papers`; the II projection raises related-paper F1 but reduces the other classes. The tuned II blend partly restores that balance. All entries are three-class OOF diagnostics, *not* end-to-end IC four-class results.

In the So400m II out-of-fold results, 415 of 1,000 true `related_papers` pairs were classified as `unrelated_papers` and 275 as `same_paper` (Appendix Table 8). The residual IC Stage-2 classifier exhibited a similar pattern, with 360 and 317 examples, respectively (Appendix Table 10). These component-level counts reveal confusion with both neighboring relations but do not establish the causes of individual errors.

6.5 Negative Ablations

Several alternatives were tested and discarded because their gains did not generalize or because they degraded local validation. Naively concatenating the original and So400m II features reduced macro-F1 from 0.4994 to 0.4914. Few-shot Qwen3-VL classification of II relations collapsed largely toward `unrelated_papers`, obtaining 0.2913 macro-F1. We also tested class-bias calibration for the residual IC classifier. Although the bias parameters were selected on separate OOF examples, held-out Stage-2 macro-F1 decreased from 0.4544 to 0.4500. We therefore retained the uncalibrated decision rule.

In Phase 2, substituting NeuML in the residual IC classifier also reduced pooled OOF macro-F1 from 0.4266 to 0.4226 (replacement) and 0.4131 (addition). Single-fold LoRA and neural-adapter pilots were not promoted to the final system and are not presented as five-fold results.

These negative results reinforce two design choices in the final system: representation improvements are most useful when matched to the appropriate modality, and the VLM is most effective when assigned the narrowly defined exact-correspondence decision rather than the full citation-relation problem.

7 Discussion

The main remaining difficulty is `related_papers`. Unlike exact correspondence or same-paper prediction, a direct citation edge is only indirectly observable from an isolated figure or caption. Two papers may contain highly similar scientific content without citing one another, while a citation may concern a method, catalog, instrument, or background result not visible in the sampled figure. This helps explain why `related_papers` remains the weakest paper-level class across CC, II, and IC.

The Qwen3-VL experiments further show that model capability depends strongly on how the model is used. Exact figure–caption correspondence provides concrete cross-modal evidence such as axes, annotations, plotted quantities, object names, and panel structure. Qwen3-VL performs well on this narrow decision but is considerably less reliable for broader provenance prediction. Similarly, the 8B model achieves stronger development AUC but lower standalone validation F1 than the 4B model; combining their partially complementary scores gives the best correspondence gate.

Our ablations also show that additional capacity or features do not automatically improve performance. Combining the base and So400m II representations reduced OOF macro-F1, and aggressive post-hoc calibration did not improve the residual IC classifier. The strongest gains instead came from matching representations and decision mechanisms to the evidence available in each modality.

Finally, our system intentionally excludes

paper identifiers, citation graphs, and bibliographic metadata at inference time. This makes `related_papers` inherently difficult because the target itself is graph-based. Paper-level text or citation context could improve future systems, but would constitute a richer inference setting than the content-only formulation studied here.

8 Conclusion

We developed a modality- and relation-aware AstroCLIMB system that distinguishes exact figure–caption correspondence from paper-level provenance. The dedicated Qwen3-VL IC gate produced the largest early improvement; astronomy-specific CC embeddings subsequently improved OOF performance, and a learned II relatedness specialist contributed a smaller test-set gain. The final V8 Kaggle macro-F1 is 0.6096, compared with 0.3641 initially. Across branches, `related_papers` remains difficult because citation linkage is not equivalent to visible semantic similarity. The results favor specialized, content-based components, while the pair-level validation protocol and score-selection procedures limit generalization claims.

Limitations

Pair-level cross-validation does not guarantee disjoint training and validation sets at the image, caption, or source-paper level. We did not quantify such overlap; therefore, the local scores do not establish generalization to previously unseen papers. The citation-link relation may not be recoverable from isolated figure or caption content. Selection among multiple OOF CC variants and tuning the II blend on pooled OOF predictions may overstate prospective gains. The IC hybrid table uses cross-fitted Stage-2 predictions, but its residual blend weight was selected on OOF results including the 400 diagnostic examples, so it is not fully independent of model selection. Repeated Kaggle submissions provided performance feedback and the final score should be interpreted within that shared-task setting; the small V7–V8 gain is not a significance test. The retained gate-summary counts do not support a paired confidence interval for 4B versus 8B+4B.

Acknowledgments

During the preparation of this manuscript, the authors used ChatGPT to assist with language editing, including grammar correction, spelling checks, restructuring of sentences, and rephrasing to improve clarity, flow, and readability. The authors reviewed and edited all text after using the tool and take full responsibility for the content of the publication.

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. *arXiv preprint arXiv:2511.21631*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

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.

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 and astrophysics. *arXiv 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.

NeuML. 2026. AstroBERT small embeddings. Hugging Face model card. Model identifier: NeuML/astrobert-small-embeddings.

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, Singapore. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2502.14786*.

A Additional Ablations and Diagnostics

CC variant	$\mu \pm \sigma$	OOF
NeuML replace	.6327 ± .0094	.6328
NeuML added (V7)	.6386 ± .0079	.6388
AstroBERT replace	.5974 ± .0188	.5977
AstroBERT added	.6313 ± .0115	.6317

Table 7: Astronomy-specific CC representation ablations on identical stratified folds.

True class	Same	Rel.	Unrel.
Same paper	550	234	216
Related papers	275	310	415
Unrelated papers	87	251	662

Table 8: OOF confusion matrix for the frozen So400m II XGBoost model.

IC Stage 2 variant	Same	Rel.	Unrel.	Macro
So400m + SPECTER2	.4540	.3425	.4833	.4266
NeuML replaces SPECTER2	.4533	.3265	.4881	.4226
NeuML added	.4461	.3340	.4592	.4131

Table 9: Per-class OOF F1 of IC residual three-class models. These results exclude the `same_figure` gate.

True IC Stage 2	Same	Rel.	Unrel.
Same paper	451	296	253
Related papers	317	323	360
Unrelated papers	219	267	514

Table 10: Residual IC Stage-2 XGBoost OOF confusion matrix, before the Qwen exact-match gate.