Modality-Conditioned Ensembles and Zero-Shot Cross-Encoder Reranking for Relation Classification Between Astronomy-Paper Objects

M. Tahsan Zahid and Most. Atkia Farzana and Tonmoy Sarker Sourav
Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh

Abstract

The WASP 2026 (AstroCLIMB) shared task requires classifying the relation between a pair of objects drawn from astronomy papers — *same_figure*, *same_paper*, *related_papers*, or *unrelated_papers* — where each pair combines text (captions) and/or images (figures) in one of three disjoint modality configurations: image+text, text+text, and image+image. Since *same_figure* cannot occur outside the image+text subset, a single unified classifier is structurally sub-optimal; we instead adopt three independent per-modality classifiers whose predictions are pooled for the competition metric (pooled macro-F1). Starting from a hand-engineered baseline (CLIP, SciNCL and DINOv2 embeddings with lexical/OCR features and a LightGBM ensemble), we find that replacing the encoder with a joint vision-language embedding model (Qwen3-VL-Embedding-2B) produces the largest single improvement among the experiments we tested (public leaderboard 0.4808 → 0.62319), exceeding the combined effect of six feature-engineering ideas that were implemented and measured but produced less than 0.01 macro-F1 each. We further show that *zero-shot* cross-encoder rerankers (Qwen3-VL-Reranker-2B, Qwen3-Reranker-4B), applied with no task-specific fine-tuning, add a second large improvement on top of the engineered ensemble (leaderboard 0.64210 → 0.66299; a later configuration with high-resolution inputs and an out-of-fold-trained VL cross-encoder reaches 0.68759), and that an untrained text+text reranker alone already outperforms the fully engineered pipeline on that subset (macro-F1 0.7111 vs. 0.6704). We report a parallel distant-supervision branch that trains on 120,000 externally-labeled pairs mined from the competition's source corpus and reaches a comparable but lower leaderboard score (0.57713) through a different mechanism; under our tested data and modeling regimes, a better encoder helped more than either additional hand-engineered features or our particular distant-supervision recipe (a controlled data-scaling study is left to future work). Reported numbers are traced to raw out-of-fold prediction arrays or literal notebook output under a shared row-level `StratifiedKFold(5)` protocol (blend weights, class multipliers and the cascade are tuned on the same OOF predictions), and we document two negative results (a broken nested-honest validation split, and several feature-engineering ideas that measurably hurt) as concrete engineering findings.

1 Introduction

Astronomy papers make heavy use of figures — plots, spectra, images — that are frequently discussed, reused, and cross-referenced across papers. The WASP 2026 (AstroCLIMB) shared task (Grezes et al., 2026) poses this as a four-way relation classification problem over pairs of objects (a caption, a figure, or both): *same_figure* (the pair refers to the same physical figure), *same_paper* (different figures/captions from the same paper), *related_papers* (the source papers cite one another), and *unrelated_papers*. Object pairs are drawn from three disjoint modality configurations — image+text, text+text, image+image — and the competition metric pools predictions across all three subsets into a single macro-F1 score.

This paper documents an end-to-end study of that task: a staged, gate-driven data pipeline (§2), a progression of modeling choices from a hand-engineered baseline to a joint vision-language embedding model with zero-shot cross-encoder reranking (§3, §4), a parallel distant-supervision branch pursued independently on the same infrastructure (§5), and an error analysis that examines why, in our tested configurations, representation quality appeared to matter more than both feature engineering and additional supervision for this task (§6, §7). We deliberately report several negative results (§7.2) and one validation bug we discovered

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 112–121
November 10, 2026 ©2026 Association for Computational Linguistics

and fixed (§9.1), since we believe the negative evidence is as informative as the positive result for anyone attempting a similar task.

Our main claims are:

1. Swapping the visual/text encoder for a joint vision-language embedding model (Qwen3-VL-Embedding-2B) produced a larger single improvement (leaderboard $+0.1424$) than the sum of every hand-engineered feature we tried (six ideas, each < 0.01 macro-F1, one of them negative).

2. Zero-shot cross-encoder reranking, applied with *no* task-specific training, is a second independent lever of comparable importance: it moves the leaderboard score from 0.64210 to 0.66299 and, on the text+text subset alone, an untrained reranker outperforms the entire engineered pipeline (0.7111 vs. 0.6704 macro-F1).

3. A parallel branch that traded a better encoder for $12\times$ more (distantly supervised) training data reached a lower score (0.57713) through a different mechanism. Because that branch changes several factors at once (label source, sampling, pseudo-OCR augmentation, data size), we treat it as suggestive comparative evidence only, not an isolated test of supervision quantity versus representation quality.

4. Two of our reports are negative/engineering findings that we keep because they generalize: a nested-honest validation split silently broken by an unshuffled, label-sorted source CSV, and several feature-engineering ideas that measurably hurt.

All accuracy figures in this paper are traced to one of four sources: a number typed directly from a Kaggle leaderboard screen, a number independently recomputed from a raw out-of-fold prediction array shipped with a result archive, a number printed by an executed notebook cell, or a number recorded in a shared results log. Where a number could not be independently re-derived from a raw array in this pass, we mark it explicitly in the corresponding table caption.

2 Task and Data

2.1 Task definition

The task is 4-class relation classification between pairs of objects drawn from astronomy papers.

Subset	Train	Test	Classes
image+text	4000	4000	all 4
text+text	3000	3000	3
image+image	3000	3000	3

Table 1: Task subsets. The competition metric is *pooled* macro-F1 over all 10,000 test rows, *not* the mean of the three subset scores (§7).

Each pair falls into exactly one of three disjoint modality subsets, summarized in Table 1. *same_figure* is only possible in the image+text subset, since it requires exactly one image and one caption from the same physical figure; the other two subsets are 3-way.

2.2 A gate-driven, not architecture-first, pipeline

Rather than committing to one architecture up front, the preprocessing pipeline was organized into sequential stages (load $\rightarrow$ modality audit $\rightarrow$ imbalance profiling $\rightarrow$ decode/clean $\rightarrow$ encode $\rightarrow$ pair-features $\rightarrow$ split), each ending in a decision gate whose answer determined the next stage's design. Several architecture-defining facts were not knowable until the data was actually inspected; exploratory data analysis (EDA) overturned four a-priori plans: **(1)** SMOTE (Chawla et al., 2002) oversampling for rare classes was dropped, since the imbalance ratio is only $\sim$3 and class weighting was used instead (we report no SMOTE ablation); **(2)** object-grouped CV was not used, since train/test object overlap is only 1 object out of $\sim$29,000 (paper-level dependence between folds is not excluded, §8); **(3)** harvesting rare-class pairs from the HF source corpus was repurposed as an *external validation* set for generalization claims instead of training data (§5 uses it differently, for training); and **(4)** a single unified 4-way classifier was replaced by three independent per-modality classifiers, since *same_figure* is structurally impossible outside image+text.

2.3 Standing constraints

Three constraints, established during the modality/structural audit stage, hold for every experiment reported in this paper:

- `id` is never used as a feature: both `train.csv` and `test.csv` are sorted by label block, so `id` alone leaks the label.

- `solution.csv` (full test-set ground truth)

is never used for training, tuning, or thresholding — only for the offline analysis reported in this paper, and disclosed as such throughout.

- The floor baseline is **0.2896** macro-F1, the score of a modality-constraint-aware random predictor (one that respects *same_figure* being impossible outside image+text) — not 0.25, which ignores that constraint. Every model below is compared against 0.2896.

2.4 Caching architecture

Kaggle's GPU quota and session-crash constraints motivated a two-stage cache: a CPU stage decodes and resizes all images once to a shared 384px JPEG cache plus parquet tables of hashes, text, and metadata, persisted as a Kaggle Dataset; each subsequent embedding experiment reads from this cache rather than re-decoding the original base64-encoded CSVs (>10 GB), so swapping encoders costs only GPU embedding time. All embeddings are deduplicated by content hash (unique objects, not unique rows, since a caption or figure can recur across pairs) and L2-normalized on save so downstream cosine similarity reduces to a dot product.

3 Method

3.1 Modeling architecture

For each of the three modality subsets we train an independent ensemble of two LightGBM (Ke et al., 2017) models (one on scalar pairwise features, one on the full feature block) and a logistic-regression model, blended with tuned weights, followed by a per-class threshold correction (coordinate ascent) and, where beneficial, a binary related-vs-unrelated cascade model blended into the multiclass output. All three subsets share an identical `StratifiedKFold(5, shuffle=True, random_state=0)` protocol (Pedregosa et al., 2011), which keeps every stage's out-of-fold (OOF) predictions blendable (row-level folds; see §8 for the tuning caveats).

Implementation details. Per subset we build 45/31/42 scalar pair features (image+text/text+text/image+image) and, separately, embedding-pair blocks ($|e_1 - e_2|$ and $e_1 \odot e_2$ per embedding space; 5,632/7,168/7,168 dimensions), for 5,677/7,199/7,210 inputs in the full block. Scalar features comprise cosine and CSLS similarity ($2\cos - r_a - r_b$, where r is the mean top-10 cross-pool cosine), log-rank of the partner within the pool (hubness), word- and character-n-gram (3–5) TF-IDF cosine, rare-token (document frequency ≤ 10), entity, number and IDF-weighted overlap, 2-hop proxy similarities through each object's nearest cross-modal neighbors, text/OCR length, and reranker features (raw score, percentile rank, deviation from the object's mean score, and object degree). LightGBM on scalar features: 600 trees, learning rate 0.03, 15 leaves, column subsample 0.7, 3 bagging seeds; on the full block: 700 trees, learning rate 0.05, 31 leaves, column subsample 0.3, 2 bagging seeds; class-balanced weights. Logistic regression uses standardized inputs with $C = 0.05$. Blend weights are searched on a 0.1 grid, class multipliers by coordinate ascent, and the cascade (a binary LightGBM on related/unrelated rows) by a grid over blend weight $\times$ threshold.

3.2 Encoder progression

The pipeline's encoder stack evolved through the following stages, each evaluated with the same downstream LightGBM ensemble so that the encoder change is the only varying factor:

1. **Baseline**: CLIP ViT-L/14 (Radford et al., 2021) for images/cross-modal similarity, SciNCL (Ostendorff et al., 2022) for scientific-text embeddings, and DINOv2 (Oquab et al., 2023) for image-only similarity.

2. **+ lexical/OCR features**: TF-IDF, hubness-corrected similarity, and 2-hop graph features, plus OCR text extracted from figures.

3. **Encoder swap**: CLIP/SciNCL/DINOv2 replaced by a single joint image+text embedding model, Qwen3-VL-Embedding-2B (Qwen Team, 2026), used for all three modality subsets.

4. **+ zero-shot cross-encoder reranking**: scalar relevance scores from Qwen3-VL-Reranker-2B (image+text) and, in a later run, Qwen3-Reranker-4B (text+text) and Qwen3-VL-Reranker-2B (image+image) (Qwen Team, 2026; Zhang et al., 2025), none of which are fine-tuned on this task's training data (each pair is scored in both orders and averaged under a task instruction prompt; the image+image reranker receives each figure together with its OCR text), added as additional input features to the same LightGBM ensemble.

We additionally tested SigLIP-so400m (Zhai et al., 2023) as an added image/text encoder at several points in this progression; its effect was subset-dependent and is reported as a controlled ablation in §A, not as part of the main progression, since it was ultimately not carried into the best-performing pipeline.

4 Experimental Results

4.1 Headline progression

Table 2 traces the pooled macro-F1 and public leaderboard score across the encoder and reranking progression described above. Leaderboard (LB) scores are typed from the Kaggle screen. Pooled values marked [b]/[c] are the pipeline's *self-reported* scores (including the cascade re-decision); the footnotes give the separately *recomputed* value from the raw OOF arrays (`sklearn.metrics`, without the cascade re-decision). Other values come from notebook output or our results log without an independent re-check.

The largest jump among the stages we tested is the encoder swap (CLIP → Qwen3-VL-Embedding-2B): leaderboard 0.4808 → 0.62319. This exceeds the combined effect of every feature-engineering idea attempted before the swap (§7.2), suggesting that representation, not hand-built features, was the stronger bottleneck at that stage. The second largest jump is zero-shot cross-encoder reranking (§7), added with no task-specific training on top of the fully engineered pipeline.

Final submission configuration. Our best leaderboard score (0.68759) comes from a later run of the same ensemble with three additions: high-resolution (1280px) OCR text and a Qwen3-VL embedding of the OCR text, a second Qwen3-VL-Embedding image embedding computed from high-resolution images, and class-probability features from a Qwen3-VL cross-encoder trained with 2-fold out-of-fold predictions (in addition to the zero-shot rerankers). The run prints per-subset tuned macro-F1 of 0.6407/0.7310/0.5847 (image+text/text+text/image+image) and pooled OOF 0.6861, consistent with the leaderboard score; boundary AUCs (sp/rel, rel/unrel) are 0.749/0.783, 0.825/0.911 and 0.735/0.752. Its half-split check (class multipliers re-tuned on one half) gives 0.625/0.720/0.559. These numbers are taken from the run's printed output, not re-derived from raw arrays, and the individual contribution of each addition is not ablated.

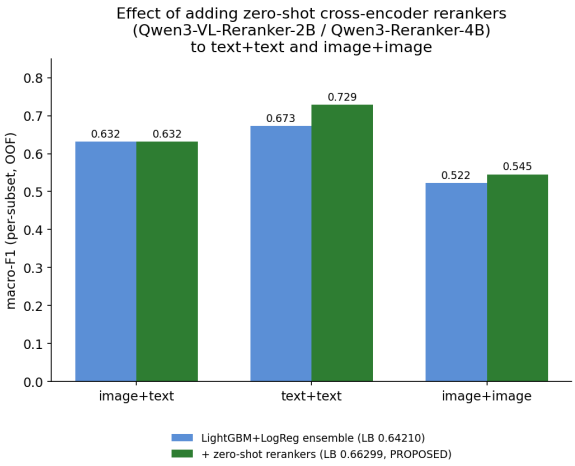


Figure 1: Per-subset macro-F1, with vs. without zero-shot cross-encoder reranking.

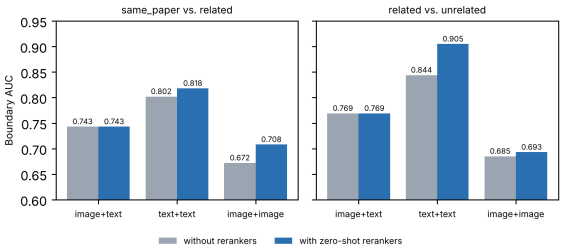


Figure 2: Boundary AUC per subset, with vs. without zero-shot cross-encoder reranking (values in Table 4).

4.2 Verified model trace

We trace exactly three result archives back to the notebooks that produced them (Table 3). The pooled and per-class numbers in Table 2 (footnotes b, c) and Table 5 are recomputed from the raw arrays inside those archives; all other numbers come from notebook output or our results log (§8).

Boundary AUC. Table 4 reports the two decision-boundary AUCs (*same_paper* vs. *related*, *related* vs. *unrelated*) from the proposed model's own report file, with and without the zero-shot reranker features.

Per-class results. Table 5 gives pooled per-class precision/recall/F1 across all 10,000 test rows for the proposed model, recomputed directly from `F_oof_{combo}.npy` arrays with `sklearn.metrics.classification_report`; Table 6 gives the per-class scores of the final configuration and Figure 3 the confusion matrix of the proposed model. Figure 1 shows per-subset F1 with and without reranking.

Stage	img+txt	txt+txt	img+img	Pooled	LB
Constraint-aware random floor	–	–	–	0.2896	–
Baseline: CLIP+SciNCL+DINOv2+LightGBM	0.3964	0.5529	0.4548	0.4806	0.4808
+ OCR/TF-IDF/hubness features	0.4031	0.6196	0.4705	–	–
+ per-class threshold	0.4186	0.6327	0.4792	0.5010[a]	–
Alt. branch: distant supervision (§5)	–	–	–	–	0.57713
Qwen3-VL-Embedding-2B + cascade	0.6031	0.6736	0.5242	0.6006[b]	0.62319
+ engineered feats. + reranker (img+txt)	0.6317	0.6728	0.5223	0.6511[c]	0.64210
+ zero-shot rerankers (txt+txt, img+img) — **proposed**	0.6317	**0.7286**	**0.5455**	0.6707[c]	**0.66299**
+ hi-res OCR/embeddings + OOF-trained VL cross-encoder	0.6407	0.7310	0.5847	0.6861[d]	**0.68759**

Table 2: Headline progression. "–" marks a value that does not apply (the random floor and the distant-supervision branch have no per-subset breakdown in our records) or was not recorded (the OCR/TF-IDF/hubness stage was not separately pooled or submitted to the leaderboard). [a]Weighted-average artifact from an early bookkeeping bug, kept for the historical record, not the pooled metric used elsewhere. [b]Self-reported pooled macro-F1; recomputed from raw OOF arrays (threshold applied, no cascade re-decision): 0.5965. [c]Self-reported pooled macro-F1; recomputed from raw OOF arrays: 0.6496 and 0.6683 respectively (see §4.2); these are ∼0.001–0.002 lower than the pipeline's self-reported number because the self-reported number additionally applies a cascade re-decision step whose own OOF was not saved to disk and so cannot be replayed here — the two numbers agreeing this closely is itself a sanity check. [d]Pooled OOF printed by the run (not re-derived from raw arrays).

Model	Result archive	What changed
Qwen3-VL-Embedding-2B + cascade	`report_A.json` (LB 0.62319)	Joint image/text embedding space, replacing CLIP+SciNCL+DINOv2; LightGBM×2 + logistic-regression blend; per-class threshold; related-vs-unrelated cascade.
+ engineered features + image+text reranker	`report_final.json` (run 2; LB 0.64210)	Adds the full engineered-feature block (CSLS-style local scaling, hubness, lexical/2-hop) on top of the above embeddings, plus the Qwen3-VL-Reranker-2B image+text score.
+ zero-shot rerankers on all subsets (**proposed**)	`report_final.json` (run 3; LB 0.66299)	Identical to the row above, plus zero-shot reranker scores added for text+text (Qwen3-Reranker-4B) and image+image (Qwen3-VL-Reranker-2B).

Table 3: Verified model trace. All three runs share the identical `StratifiedKFold(5, shuffle=True, random_state=0)` protocol per combination, so OOF arrays are directly comparable and blendable.

Subset	sp vs. rel	rel vs. unrel
image+text	0.743 → 0.743[*]	0.769 → 0.769
text+text	0.802 → **0.818**	0.844 → **0.905**
image+image	0.672 → **0.708**	0.685 → **0.693**

Table 4: Boundary AUC, without → with zero-shot reranker features. [*]image+text is unchanged since the zero-shot reranker addition in this stage targets text+text and image+image only.

Class	P	R	F1
same_figure	0.894	0.808	0.849
same_paper	0.678	0.597	0.635
related_papers	0.501	0.485	0.493
unrelated_papers	0.642	0.760	0.696
pooled macro-F1			**0.6683**

Table 5: Pooled per-class precision/recall/F1 computed from the proposed model's saved out-of-fold predictions (Qwen3-VL-Embedding-2B ensemble + zero-shot rerankers). The corresponding submission scored 0.66299 on the public leaderboard; these class-level values are OOF, not leaderboard test-set, metrics.

No result archive was generated for the Qwen3-VL-Embedding-2B-only stage that contains a raw OOF array (only summary JSON), so no confusion matrix or per-class chart is produced for that intermediate stage; Table 2 reports its summary numbers verbatim.

5 Alternative Branch: Distant Supervision from an External Corpus

In parallel with the encoder-progression branch above, a second branch pursued a different lever entirely: *more labeled supervision*, on the same feature-engineering base (CLIP/SciNCL/DINOv2 + OCR + lexical/2-hop features), rather than a better encoder. We report it as a separate branch rather than folding it into Table 2 because it does not build on the Qwen3-VL embeddings and reaches a comparable-but-lower leaderboard score through

Class	P	R	F1
same_figure	0.863	0.859	0.861
same_paper	0.708	0.597	0.648
related_papers	0.508	0.540	0.524
unrelated_papers	0.681	0.746	0.712
pooled macro-F1			**0.686**

Table 6: Pooled OOF per-class precision/recall/F1 of the 0.68759 configuration, as printed by the run (class multipliers applied, no cascade re-decision; not re-derived from raw arrays).

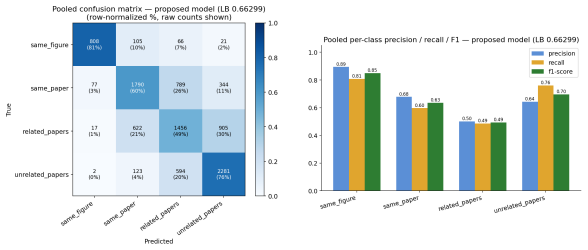


Figure 3: Pooled 4×4 confusion matrix (left) and pooled per-class precision/recall/F1 (right), both computed from out-of-fold predictions for the proposed model (Qwen3-VL-Embedding-2B ensemble + zero-shot rerankers; public leaderboard score 0.66299).

entirely different means — the comparison itself is informative.

Core idea. The competition's source corpus, `adsabs/AstroCLIMB` on HuggingFace, allows the same four relation labels to be reconstructed from metadata: *same_paper* from shared DOI, *related_papers* from an existing citation edge, and *unrelated_papers* from the absence of one. As downloaded and filtered, the corpus contains 94,059 figure rows over 9,998 papers with 26,995 citation edges; two DOIs touching a test caption were dropped before sampling.

Leakage audit. An audit printed by the run itself found 81.8% of Kaggle training captions present in the HF corpus, and 0.0% of Kaggle test captions present in it — the latter, safety-relevant figure is confirmed independently by the explicit DOI-drop step that runs regardless of the audit result.

Training data. 40,000 pairs were sampled per class (*same_paper* / *related_papers* / *unrelated_papers*), 120,000 pairs total ($12\times$ the size of the Kaggle training set). 16,000 of the *unrelated_papers* pairs are title-vocabulary hard negatives; 59,911 pairs (roughly half) have at least one side replaced with a pseudo-OCR noisy view, to better match the OCR-derived text used on the image side of the Kaggle task.

Metric	Value
Held-out macro-F1 (on HF pairs)	0.5929
Zero-shot transfer to Kaggle text+text	0.6242
Zero-shot transfer to Kaggle image+image	0.4158

Table 7: HF-corpus-trained model, transfer to the Kaggle task.

The 0.21 macro-F1 gap between clean-caption transfer (text+text) and OCR-represented transfer (image+image, Table 7) is exactly the motivation for the pseudo-OCR augmentation described above: a model trained only on clean captions transfers markedly worse to noisy OCR text.

Final result. The full per-subset blend (LightGBM on scalar and full features, logistic regression, and the HF-corpus-trained model's own prediction as a fourth blend arm for text+text/image+image) reaches an overall OOF macro-F1 of 0.5692 and a leaderboard score of **0.57713** — in close agreement (within 0.008), the same kind of local/leaderboard consistency seen in the main branch (§4.2).

Transductive post-processing (not scored). A test-time paper-clustering step (union-find over confident *same_paper*/ *same_figure* edges, pooling probabilities across test pairs connecting the same two clusters) found 7,922 paper clusters in the test set, of which 131 cross-cluster groups had more than one test pair between them (285 pairs total); 2,679 test pairs fell inside a single cluster. This was shipped as separate A/B submission variants rather than trusted into the base submission, and is not part of the reported 0.57713 score.

Comparative interpretation. $12\times$ more (distantly supervised) training data and several additional signal sources did not reach the level of a single better encoder: the encoder swap alone (§4.1) reached leaderboard 0.62319 using only the original 10,000 Kaggle pairs and one embedding model. This is consistent with our central claim (§7) that representation quality was the stronger bottleneck in our setting, but it is not a controlled comparison: the branch varies label source, sampling, augmentation and data size simultaneously, and its distant labels are noisy (a missing citation edge is treated as "unrelated", which can be wrong). We did not estimate that label noise; a learning-curve study at fixed representation (e.g., 10K/30K/60K/120K pairs) would be needed to isolate the effect of su-

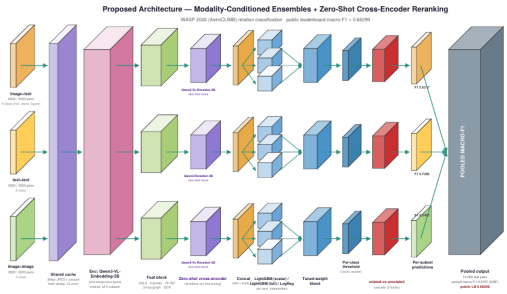


Figure 4: Proposed architecture: per-modality Qwen3-VL-Embedding-2B encoding, engineered features, zero-shot cross-encoder reranking, and a LightGBM/logistic-regression ensemble with per-class threshold and cascade post-processing, pooled into the leaderboard metric (LB 0.66299).

Boundary	**20%**	**40%**	**60%**
same_paper ↔ related	0.654	0.679	0.704
related ↔ unrelated	0.657	0.684	0.710

Table 8: Oracle analysis: pooled macro-F1 after fixing the given fraction of errors at one boundary.

sp/rel AUC	**rel/unrel AUC**	**Pooled F1**
0.730 (baseline)	0.766 (baseline)	0.6293
0.810	0.816	0.6531
0.869	0.862	0.6891
0.916	0.907	0.7369

Table 9: AUC-to-pooled-F1 mapping used to set feature/model priorities.

pervision quantity.

6 Error Analysis

At the Qwen3-VL-embedding stage (pooled OOF 0.6293), 78% of the 4041 total OOF errors fall in exactly two class boundaries: *same_paper*↔*related_papers* (1471 errors) and *related_papers*↔*unrelated_papers* (1690 errors); *same_figure* is substantially easier than the paper-level relation classes at this stage (F1 0.789), though not perfectly solved. Table 8 shows an oracle analysis: fixing a random fraction of errors at one boundary and re-measuring pooled macro-F1. These are diagnostic upper-bound sensitivity estimates, not scores attainable by any model we built.

Mapping this onto a target AUC (Table 9) turned the oracle analysis into an actionable target: any candidate feature or model was judged by whether it moved these two boundary AUCs, independent of how principled it looked on paper.

Per-subset AUC breakdown shows image+image as the consistently weakest link: text+text reaches 0.797/0.846 (sp/rel, rel/unrel), image+text 0.722/0.768, and image+image only 0.672/0.683 — the same subset that zero-shot reranking (§7) improves the least in absolute AUC terms (Table 4), suggesting that a stronger image-only encoder, rather than reranking, is the more promising future lever for that subset specifically.

7 Discussion

7.1 Untrained cross-encoders rival a hand-engineered pipeline

Three independent measurements motivated a fine-tuning phase for cross-encoder rerankers; apart from the out-of-fold-trained VL cross-encoder in the final configuration (§4.1, not independently verified), it was not completed to a verified result (§8): on image+text (*same_paper* vs. *related*), a single zero-shot reranker scalar reaches AUC 0.709, close to the full ∼5.7K-input engineered pipeline's 0.722, and combining both reaches 0.741; on text+text, Qwen3-Reranker-4B alone, with zero training, reaches macro-F1 0.7111, *exceeding* the full engineered pipeline's 0.6704; on image+image, the reranker alone (F1 0.509) roughly ties the pipeline (0.520). If an untrained cross-attention model matches or beats a heavily engineered bi-encoder-plus-tree-ensemble pipeline, a plausible reading is that much of the remaining signal is representational (requiring token-level cross-attention) rather than feature-space — this is our central interpretive hypothesis, not a proven conclusion, and the reason the encoder swap and zero-shot reranking additions were prioritized over further feature engineering.

Table 10 gives a partial ablation: the zero-shot reranker score alone (AUC on all training pairs; no fitting involved), the engineered embedding pipeline without rerankers (OOF AUC), and the

Subset	Boundary	Rerank.	Pipeline	Both
txt+txt	sp/rel	0.824	0.802	0.818
txt+txt	rel/unrel	0.892	0.844	0.905
img+img	sp/rel	0.710	0.672	0.708
img+img	rel/unrel	0.653	0.685	0.693

Table 10: Boundary AUC: zero-shot reranker alone (Qwen3-Reranker-4B for text+text; Qwen3-VL-Reranker-2B on figure+OCR for image+image), engineered pipeline without rerankers, and the final ensemble. Reranker AUCs are over all 3,000 training pairs per subset; pipeline and ensemble AUCs are OOF.

combined ensemble. The reranker alone beats the engineered pipeline on three of four text+text and image+image boundaries, and the ensemble improves on both single sources only at related-vs.-unrelated; at same_paper-vs.-related the reranker alone is as good as or slightly better than the ensemble (0.824 vs. 0.818 text+text; 0.710 vs. 0.708 image+image). This suggests the reranker largely supplants rather than complements the similarity features at that boundary. An embedding-only arm and a full factorial are not available.

7.2 Rejected feature-engineering ideas

Every idea in Table 11 was implemented, measured on real OOF predictions, and reverted; we report them as direct evidence against "just add more features" as a strategy for this task.

7.3 Model-side (structural) findings

Two structural changes, orthogonal to both the encoder and the feature set, measurably helped: a per-class multiplier/threshold via coordinate ascent ($+0.011$ honest, weighted; plain argmax is not macro-F1-optimal), and a binary related-vs-unrelated cascade blended into the multiclass output ($+0.006$ to $+0.011$ nested-honest) — the only successful structural change we found, and notably the simpler two-knob cascade beat a five-knob one-vs-rest variant, consistent with the overfitting pattern in Table 11's last free-parameter row.

8 Limitations

Validation and tuning. Blend weights, class multipliers and cascade parameters were tuned on the same row-level OOF predictions that are then scored, with one seed (`random_state=0`) and no confidence intervals, so the pooled OOF figures (0.6496/0.6683) are mildly optimistic. A half-split check that re-tunes only the class multipliers gives 0.626/0.663/0.509 (image+text/text+text/image+image) against 0.631/0.670/0.520 tuned, i.e., 0.004–0.011 lower (on a different half-sample); the final cascade OOF was not saved. Folds are not paper-grouped, so paper-level dependence between folds is not excluded. Our team made 17 leaderboard submissions across all branches, so leaderboard feedback may have influenced stage selection, and differences of $\sim$0.01 or less cannot be distinguished from noise.

Ablations. We lack a full factorial comparison (embedding only, reranker only, features only, and their combinations); Table 10 gives a partial one. The 0.68759 configuration (high-resolution OCR/embeddings and an out-of-fold-trained VL cross-encoder) is not ablated, and the cross-encoder's training is documented only in our released notebooks.

Verification and scope. Only the OOF-recomputed numbers marked in Table 2 are reproducible from saved arrays; the error analysis, rejected ideas and distant-supervision ablations come from logs or notebook output. Two fine-tuning notebooks (DeBERTa-v3-large; one early attempt) were not verified to completion and carry no claim. Distant-supervision label noise was not estimated. Feature definitions are in Section 3, and code and notebooks are in our public repository. All models are open-weight and run locally; no external LLM API was used.

Planned experiments. Our results would be strengthened by (i) a fully nested evaluation of all tuned parameters, (ii) paper- or citation-component-grouped cross-validation, (iii) the full reranker ablation, (iv) multiple seeds with confidence intervals, and (v) a controlled learning curve (e.g., 10K/30K/60K/120K pairs) at fixed representation for the distant-supervision branch. We did not run these before the camera-ready deadline.

9 Reproducibility and Data Statement

The following rules were applied to every result in this paper: `solution.csv` was never used for training, tuning, or threshold selection; `id` was never used as a feature (both CSVs are label-sorted, so `id` alone would leak the label); an identical `StratifiedKFold(5, shuffle=True, random_state=0)` (or its matching 2-fold variant) was applied across every

Idea	Measured effect	Why it failed
IPF/Sinkhorn class-count matching	-0.0087 pooled	Macro-F1's optimum is not class-balanced.
Graph/label-transitivity propagation	Would have hurt	Zero closed triangles in train graph; label correlation significant ($\chi^2 = 59.5$, $p < 10^{-4}$) but *anti-transitive* ($P(\textit{same_paper} \mid \textit{neighbor same_paper}) = 0.72$).
OCR text as a *same_paper* signal	AUC 0.51–0.53 (chance)	Mean OCR length only 54 chars.
SigLIP added on CLIP+SciNCL, text+text	$0.5529 \rightarrow 0.5420$ (-0.0109)	64-token limit truncates captions.
MLP blend instead of LightGBM	Always below LightGBM	Tree inductive bias fits this feature space.
One-vs-rest (5 knobs) vs. cascade (2 knobs)	$+0.0025$ vs. cascade's $+0.0063$ (honest)	More free parameters overfit the OOF used for tuning.
Zero-shot Qwen3-Reranker-4B on proxy captions (nearest retrieved caption) for image+image	AUC 0.575/0.701 (sp/rel, rel/unrel) vs. pipeline 0.672/0.685	Retrieved captions are noisy stand-ins; helps rel/unrel only and was not used in the proposed model.
Train/test object-overlap exploitation	1 shared object of $\sim$29,000	No leakage path available.

Table 11: Feature-engineering ideas that were implemented, measured, and reverted.

notebook contributing an OOF feature, so that OOF predictions can be blended on a shared protocol (the tuned scores are not unbiased estimates, §8); and, where nested-honest tuning was implemented and retained, threshold/blend weights were tuned on one random half and measured on the other. One exception: the cascade re-decision step's own OOF was not saved to disk, so its exact honest score cannot be replayed (Table 2).

9.1 Engineering findings

Two findings generalize beyond this competition. First, naively calling a vision-language model's full generation head (`Qwen3VLForConditionalGeneration.forward()`) for hidden states materializes a logits tensor sized for a 151,936-token vocabulary at every sequence position (212 MB for a 731-token sequence here) that is never used — the cause of every GPU OOM we observed; the fix is to call the inner transformer directly and read `last_hidden_state`. Second: an early nested-honest split naively dividing a CSV into "first half"/"second half" produced implausible macro-F1 of 0.1087/0.1478/0.1205 (image+text/text+text/image+image) against "final" scores of 0.60/0.67/0.52 on the identical run — a $\sim 5\times$ collapse revealing the source CSV was label-sorted, making the halves near-disjoint by class. The fix was to permute indices with a seeded RNG before splitting — a concrete illustration of why a nested-honest check needs its own sanity check.

10 Related Work

Our encoder progression draws on CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2023) and SciNCL (Ostendorff et al., 2022), combined via gradient-boosted trees (Ke et al., 2017). Our strongest results build on the Qwen3 embedding and reranker models (Qwen Team, 2026; Zhang et al., 2025; Yang et al., 2025), used zero-shot; the distant-supervision branch uses a vision-language model (Bai et al., 2025) for an (ultimately disabled) captioning path and similarity local scaling (Conneau et al., 2018). Cross-validation uses scikit-learn (Pedregosa et al., 2011).

11 Conclusion

We presented a multi-branch effort on the WASP 2026 (AstroCLIMB) relation-classification task, tracing every reported number to notebook output, a raw OOF array, or our results log. Under our tested regimes, representation quality was the stronger bottleneck: a single encoder swap outperformed six feature-engineering ideas combined, and zero-shot cross-encoder reranking rivaled or beat the fully engineered pipeline on two of three subsets. Our proposed submission reaches a public leaderboard score of 0.66299 (0.68759 with the later high-resolution and fine-tuned cross-encoder configuration). Task-specific cross-encoder fine-tuning (§8) is the most promising next step.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dai, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL technical report. ArXiv preprint.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Cross-domain similarity local scaling for bilingual lexicon induction. In *International Conference on Learning Representations (ICLR)*.

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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qwen Team. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *Preprint*, arXiv:2601.04720.

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. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Subset	Space(s)	Macro-F1
image+text	clip	0.3964
image+text	sig	0.4155
image+text	**clip+sig**	**0.4351** (+0.0387)
text+text	clip+sci	**0.5529**
text+text	sig	0.4327
text+text	clip+sci+sig	0.5420 (−0.0109)
image+image	clip+dino	0.4548

Table 12: SigLIP ablation. SigLIP helps the cross-modal subset but hurts the text-only subset when stacked on an already-good text space, consistent with a 64-token truncation explanation (text+text captions are the longest of the three subsets).

An Yang et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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)*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A SigLIP: a subset-dependent additional encoder

We separately measured SigLIP-so400m (Zhai et al., 2023) as an additional encoder, holding the LightGBM pipeline fixed and swapping one encoder combination at a time (Table 12).