# Parameter-Efficient Adaptation of Qwen3-VL for Astronomy Citation Linking at AstroCLIMB 2026

**Syed Mohaiminul Hoque**
Center for Computational and Data Sciences Lab
Independent University, Bangladesh

## Abstract

We describe a submission to the AstroCLIMB shared task at WASP 2026. The task requires classifying a pair of astronomy figures and/or captions as the same figure, from the same paper, from citation-related papers, or from unrelated papers. Our system casts classification as constrained next-token prediction with Qwen3-VL and adapts the model using 4-bit quantization and low-rank adapters. On the 10,000-pair Kaggle test set, three Qwen3-VL-8B models with language-attention adapters obtain 0.73230–0.73307 macro-F1 individually. Averaging their four-class probabilities raises the score to 0.73604, and masking the impossible same-figure class for caption–caption and image–image pairs produces 0.73725. This paper documents empirical system engineering rather than proposing a new learning algorithm; broader visual adaptation, restricted-loss blending, and citation-focused auxiliary training do not improve the result, and most remaining errors involve the `related_papers` class.

## 1 Introduction

Scientific figures condense results that are difficult to recover with text-only indexing. The AstroCLIMB shared task asks systems to partially reconstruct an astronomy citation graph using only figures and captions (Grezes et al., 2026). Unlike conventional document classification, the input is a pair of heterogeneous objects: either object may be an image or a caption, and the target is a symmetric relation between their source papers.

Modern vision–language models (VLMs) offer a natural interface for such mixed inputs. However, full fine-tuning remains expensive, especially on the two NVIDIA T4 GPUs available in a standard Kaggle environment. We therefore formulate the task as instruction-following with four possible answer tokens and use QLoRA (Dettmers et al., 2023) to adapt Qwen3-VL (Bai et al., 2025a).

We develop a uniform multimodal pipeline with random object swapping; compare scale, loss normalization, adapter coverage, and seed variation; and analyze both a balanced validation split and a negative auxiliary-training result. Our three-seed probability ensemble with a modality constraint reaches 0.73725 test macro-F1, with `related_papers`, which is confused with both neighbouring classes, remaining the main difficulty. Code, notebooks, and artifacts are available at `https://github.com/SyedT1/Astroclimb-sharedtask`.

## 2 Related Work

Scientific-figure research commonly studies caption generation or cross-modal retrieval. SciCap introduced a large figure–caption corpus and captioning baselines (Hsu et al., 2021); SciMMIR benchmarks retrieval over 530K scientific image–text pairs (Wu et al., 2024). Chart retrieval also shows value in combining visual, textual, and derendered table evidence (Nowak et al., 2024). AstroCLIMB instead predicts provenance relations rather than retrieving a matching caption.

SPECTER2 learns citation-informed scientific-document representations (Cohan et al., 2020; Singh et al., 2023); SigLIP2 and DINOv3 provide cross-modal and self-supervised visual features (Tschannen et al., 2025; Siméoni et al., 2025). We compare their frozen features with our VLM approach. Casting labels as answer tokens is related to pattern-based classification (Schick and Schütze, 2021); our contribution is a carefully evaluated shared-task system combining this interface with QLoRA, ensembling, and a task-valid constraint.



*Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, pages 104–111
November 10, 2026 ©2026 Association for Computational Linguistics

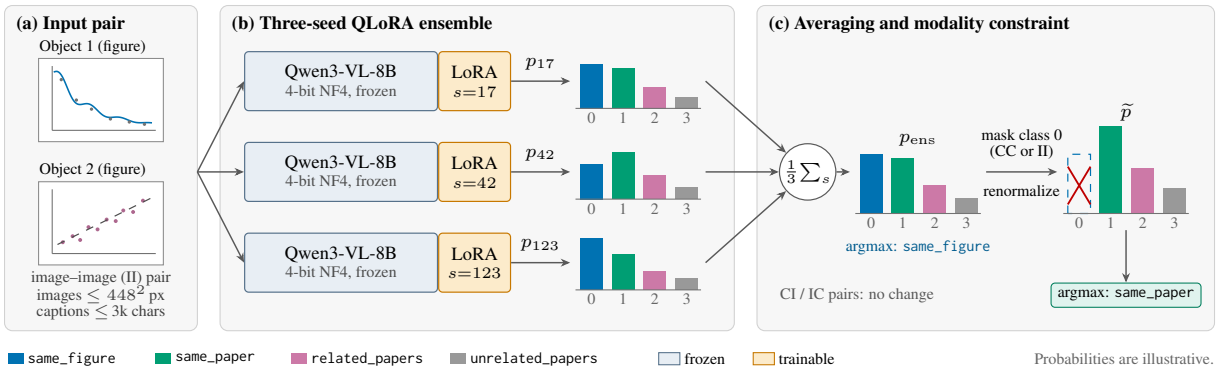



Figure 1: Inference pipeline of the best system, traced on an illustrative image–image pair. (b) Three full-vocabulary Qwen3-VL-8B models, each a frozen 4-bit backbone with a seed-specific language-attention LoRA adapter, produce four-class probabilities. (c) Equal averaging reduces seed variance; the deterministic modality mask then removes the impossible `same_figure` option from caption–caption and image–image pairs and renormalizes, here changing the prediction to `same_paper`.

## 3 Task and Data

Each example is a pair $x_i = (o_{i1}, o_{i2})$. An object is either an English figure caption or a base64-encoded scientific figure. The four mutually exclusive classes are: (0) `same_figure`, when the objects are the image and caption of one figure; (1) `same_paper`, when they are different objects from one paper; (2) `related_papers`, when one source paper cites the other; and (3) `unrelated_papers`. The relation is symmetric under exchanging the two objects.

The Kaggle release contains 10,000 labeled training pairs and 10,000 test pairs. Training labels are distributed $(1000, 3000, 3000, 3000)$ in class order. The official metric is macro-averaged F1,

$$F_{1,\text{macro}} = \frac{1}{4}\sum_{c=0}^{3} F_{1,c}, \tag{1}$$

so performance on the less frequent `same_figure` class contributes equally. The broader collection contains 94,233 figure–caption records with citation metadata (NASA Astrophysics Data System, 2026); we use it only for the auxiliary experiment in Section 5.4.

For controlled experiments, we reserve 200 examples per class using seed 42, producing a balanced 800-pair validation set and 9,200 training pairs. Full-data systems use all 10,000 labeled pairs. We do not use test metadata, external DOI lookup, or hidden labels.

## 4 System Description

### 4.1 Input representation

We use instruction-tuned Qwen3-VL-4B and Qwen3-VL-8B checkpoints. A system message defines the four relations and assigns them digits 0–3. The user message then presents Object 1 and Object 2 in their native modalities and asks for one digit. Captions are limited to 3,000 characters by retaining both their beginning and end. Images are decoded to RGB and processed within an area range of $256^2$ to $448^2$ pixels. This cap is important for fitting the 8B model on T4 hardware.

Because the relation is symmetric, each training pair is randomly swapped with probability 0.5 without changing its label. We do not use swap test-time augmentation in the submitted systems.

### 4.2 QLoRA adaptation

We load the backbone in 4-bit NormalFloat (NF4) format with double quantization and FP16 computation. Low-rank adaptation (Hu et al., 2022) modifies a frozen weight $W_0$ as

$$W_{\text{eff}} = Q_{\text{NF4}}(W_0) + \frac{\alpha}{r} BA, \tag{2}$$

where $r = 16$, $\alpha = 32$, and LoRA dropout is 0.05. Our main systems adapt language-attention q/k/v/o projections; we also test adding visual-merger projectors or adapting all eligible linear layers.

Training uses two-process distributed data parallelism, one process per T4. The per-device batch size is one with eight gradient-accumulation steps,

| Backbone | Training setup | Epochs | Adapter coverage | Test macro-F1 |
|---|---|---|---|---|
| Frozen encoders | SPECTER2/SigLIP2/DINO + features | — | modality CatBoost | 0.48279 |
| Qwen3-VL-4B | balanced 5K, $\mathcal{L}_{\text{vocab}}$ | 1.0 | language attention | 0.67856 |
| Qwen3-VL-4B | full 10K, $\mathcal{L}_{\text{vocab}}$ | 1.0 | language attention | 0.70751 |
| Qwen3-VL-4B | 9.2K, $\mathcal{L}_4$ | 2.0 | language attention | 0.71453 |
| Qwen3-VL-4B | full 10K, $\mathcal{L}_{\text{vocab}}$ | 1.0 | language + visual projectors | 0.71016 |
| Qwen3-VL-4B | full 10K, $\mathcal{L}_{\text{vocab}}$ | 1.0 | all linear layers | 0.71202 |
| Qwen3-VL-8B | full 10K, $\mathcal{L}_{\text{vocab}}$, seed 42 | 1.0 | language attention | 0.73230 |
| Qwen3-VL-8B | full 10K, $\mathcal{L}_{\text{vocab}}$, seed 123 | 1.0 | language attention | 0.73250 |
| Qwen3-VL-8B | full 10K, $\mathcal{L}_{\text{vocab}}$, seed 17 | 1.0 | language attention | 0.73307 |
| Qwen3-VL-8B | full 10K, $\mathcal{L}_4$ | 1.5 | language attention | 0.73032 |
| Qwen3-VL-8B | three-seed probability ensemble | 1.0 | language attention | 0.73604 |
| Qwen3-VL-8B | 75% ensemble + 25% restricted | — | probability blend | 0.73590 |
| Qwen3-VL-8B | ensemble + modality mask | — | deterministic constraint | **0.73725** |

Table 1: Final Kaggle leaderboard results. "Full 10K" denotes all labeled pairs; the 9.2K run retains 800 validation pairs. The final three rows are derived systems.

giving a global batch size of 16. The primary full-data systems train for one epoch with learning rate $10^{-4}$. The 8B restricted-loss refit uses $5\times10^{-5}$ for 1.5 epochs, selected on the fixed validation split.

### 4.3 Classification objectives

Let $t_c$ be the tokenizer ID for digit $c$, and let $z_{i,v}$ be the logit immediately before the answer. The full-vocabulary objective supervises only that position:

$$\mathcal{L}_{\text{vocab}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp z_{i,t_{y_i}}}{\sum_{v\in\mathcal{V}}\exp z_{i,v}}. \quad (3)$$

We also test $\mathcal{L}_4$, replacing the denominator by the four valid digit tokens. Inference always selects the largest of those four logits, avoiding unconstrained generation. Appendix A gives the full restricted-loss, inference, and swap definitions.

### 4.4 Probability ensemble and modality constraint

Figure 1 summarizes the best system. We independently train the full-vocabulary 8B configuration with seeds 17, 42, and 123. If $p_s(c \mid x)$ is the normalized four-class distribution from seed $s$, their equal-weight ensemble is

$$p_{\text{ens}}(c \mid x) = \frac{1}{3}\sum_{s\in\{17,42,123\}} p_s(c \mid x). \quad (4)$$

This is probability averaging rather than majority voting over hard predictions.

The `same_figure` label is valid only for a cross-modal figure–caption pair. For caption–caption (CC) and image–image (II) pairs, we impose the task constraint after ensembling:

$$\widetilde{p}_0 = 0, \qquad \widetilde{p}_c = \frac{p_{\text{ens},c}}{\sum_{j=1}^{3} p_{\text{ens},j}} \quad (c = 1, 2, 3). \quad (5)$$

Cross-modal probabilities remain unchanged. The mask modifies vectors for 6,000 of 10,000 test rows but changes only seven hard predictions. We do not use swap test-time augmentation, class-bias calibration, or the restricted-loss model in the final system.

## 5 Experiments

### 5.1 Zero-shot baselines

With the same prompt and 4-bit loading, Qwen3-VL-4B is the strongest zero-shot model at 0.47956 macro-F1; Qwen2.5-VL (Bai et al., 2025b) scores 0.46745. The frozen-encoder baseline in Table 1 reaches 0.48279; Appendix B details both baselines.

### 5.2 Fine-tuned systems

Table 1 summarizes the final-leaderboard results. Full-data 4B training raises macro-F1 from 0.47956 to 0.70751; the analogous 8B seed-42 model reaches 0.73230. Replications score 0.73250 and 0.73307, motivating probability averaging.

The best 8B model exceeds the comparable 4B model by 0.02556. Across three seeds, the mean is $0.73262 \pm 0.00040$ sample standard deviation (range 0.00077). Broader 4B adapter coverage adds only 0.00265–0.00451; a visual-projector-only run scores 0.48182 because caption–caption pairs bypass its trainable path.

The restricted-loss comparison also changes the split and epochs. Its selected 8B checkpoint ob-

tains 0.729164 validation macro-F1; a full-data refit scores 0.73032. Blending 25% of its probabilities into the ensemble scores 0.73590, 0.00014 below the raw ensemble.

Probability averaging improves the best seed by 0.00297; the mask adds 0.00121, reaching 0.73725. Seven predictions move from the impossible `same_figure` label to `same_paper`, with no extra inference.

Test labels are unavailable, so these scores carry no exact confidence intervals. Rescaling the validation bootstrap of Appendix B to 10,000 rows suggests a 95% half-width of roughly 0.008. The 4B-to-8B gain is therefore resolvable, whereas seed differences and the ensemble gain should be read as indicative. The mask gain differs in kind: all seven changed predictions were necessarily wrong before masking.

### 5.3 Controlled validation analysis

On the balanced 800-pair split, the selected restricted-loss checkpoint obtains 0.72916 macro-F1. `same_figure` is easiest and `related_papers` weakest (Appendix B). The split is small: a stratified bootstrap over the recorded validation confusion matrix gives a 95% interval half-width of 0.029. Adding language MLP targets reaches 0.729019, a difference of 0.00015 that this split cannot resolve; we keep attention-only adaptation because it is cheaper, not because it is more accurate. The split is also pair-level, so it may overestimate performance on unseen source papers.

### 5.4 Citation-focused auxiliary training

We construct 20,000 auxiliary examples from the public AstroCLIMB collection (NASA Astrophysics Data System, 2026), distributed as $(2000, 6000, 8000, 4000)$ in class order. Class 0 replays training examples with replacement; classes 1–3 pair captions from one DOI, directly citation-linked DOIs, or non-adjacent papers sharing a meaningful title token, with metadata used only for selection. UUID-pair keys enforce pair-level uniqueness for classes 1–3, and exact normalized-caption matches exclude recoverable validation and test DOIs, although image-only papers cannot be mapped. This auxiliary stage lowers macro-F1 from 0.72916 to 0.71405, losing 0.035 and 0.042 F1 on `same_figure` and `same_paper` while gaining at most 0.013 on the different-paper classes.

Four properties of the auxiliary data can explain this pattern. First, *modality shift*: all 18,000 auxiliary pairs in classes 1–3 are caption–caption, whereas competition pairs mix captions and images. Second, *prior shift*: the auxiliary labels follow a 10/30/40/20% split rather than 10/30/30/30%, over-weighting `related_papers`. Third, *invisible labels*: a citation edge does not imply that two randomly drawn captions from the linked papers share visible evidence, and title-token negatives are topically similar by construction, so classes 2 and 3 are often separable only through metadata. Fourth, *replay*: class 0 reuses each of the 800 training `same_figure` rows about 2.5 times. We did not ablate these factors individually. A better sampler would retain only pairs with shared visible evidence, such as object names or numerical values; include image objects; follow the competition label prior; and interleave auxiliary and competition batches rather than training them in sequence. Appendix B gives detailed results.

## 6 Error Analysis

**Where errors concentrate.** Of the 234 validation errors in the recorded confusion matrix, 113 (48%) confuse `same_paper` with `related_papers` and 74 (32%) confuse `related_papers` with `unrelated_papers`; 34 involve `same_figure` and `same_paper`, and only 13 involve other pairs. `related_papers` is thus confused with both neighbouring classes. Both confusions are nearly symmetric (62 vs. 51 and 36 vs. 38), which points to genuine ambiguity rather than a class-prior bias that recalibration could remove. On the unlabeled test set, the three seeds of the final system give a label-free view (Table 6). They disagree on 13.3% of rows; 47.8% of these disagreements are between `same_paper` and `related_papers`, and 42.1% are between `related_papers` and `unrelated_papers`. `related_papers` predictions are also the least confident, with a median maximum probability of 0.566 against 0.739–0.956 for the other classes.

**Why `same_paper` and `related_papers` collide.** Both relations are defined by metadata the model never sees: a shared DOI or a citation edge. Two figures from one paper can cover different sub-analyses, while figures from citing papers often share instruments, targets, and notation. Topical similarity therefore indicates that a pair is related, but says little about which relation holds. The decisive cues are sparse: an identical dataset, a

repeated object identifier, or a matching plotting style. Example-level validation predictions were not retained, so we cannot show annotated examples.

**Use of visual evidence.** Exact figure–caption matching remains tractable because shared entities, values, and panel labels align directly. The raw ensemble nevertheless predicts `same_figure` for seven same-modality pairs, which the mask corrects. Whether the model reads content inside figures is less clear. Validation macro-F1 falls from 0.728 on caption–caption to 0.671 on caption–image and 0.651 on image–image pairs. On the test set, seeds agree less often on cross-modal pairs (85.6%) than on same-modality pairs (87.5%). Our frozen-feature baseline already uses OCR-token overlap and perceptual hashing, yet reaches only 0.48279, so shallow figure-text matching is far from sufficient. That language-side adaptation beats broader visual adaptation does not show that figures are unimportant: 10,000 labels may be too few to tune many visual parameters, while the pretrained encoder already supplies useful features. Image-blanking and OCR-substitution ablations are needed to measure how much predictions depend on figure content.

## 7 Efficiency and Reproducibility

The restricted 8B run trains 15,335,424 parameters for 938 steps in 320.45 minutes, peaking at 11.53 GiB on rank 0. Two-T4 inference over 10,000 pairs takes 78.15 minutes. The final system requires three independently trained and inferred adapters, but ensembling and masking are CPU-only. We use Transformers 4.57.1, PEFT 0.17.1, Accelerate 1.10.1, and bitsandbytes 0.47.0. The split seed is 42; full-data seeds are 17, 42, and 123. Appendix C reports the optimizer, scheduler, checkpoint policy, and output validation.

## 8 Conclusion

Our language-attention QLoRA system constrains Qwen3-VL prediction to four tokens. One 8B model reaches 0.73307 test macro-F1, three-seed averaging reaches 0.73604, and the modality mask yields 0.73725. Remaining errors centre on `related_papers`, which is confused with both `same_paper` and `unrelated_papers`. Broader visual adaptation, restricted-loss blending, and citation-graph supervision do not improve the result. Future work should evaluate on source-paper-grouped splits, ablate figure content directly, and use evidence-visible, modality-matched auxiliary examples.

## Limitations

Most full-data configurations lack local validation, and differences are measured by single final-leaderboard scores. Test labels and example-level validation predictions are unavailable, so paired significance tests cannot be computed. Our confidence intervals come from the single recorded validation confusion matrix, and test intervals are approximate; the seed standard deviation measures run variability, not a confidence interval. The controlled split is pair-level and uses one seed. Kaggle pairs carry no paper identifiers, and DOIs can be recovered only for caption objects by exact caption matching, so we could not build a source-paper-grouped split. We did not run image-blanking or OCR-substitution ablations, so reliance on figure content is not quantified, and the proposed causes of the auxiliary-training drop are not individually ablated. Some comparisons vary multiple factors, including data size and epoch count. Caption truncation and image resizing can remove evidence. Finally, metadata-defined citation relations may be invisible in the presented objects, imposing a task-level ceiling that we do not estimate.

## Ethics Statement

The system processes public scientific figures and captions supplied for the shared task. It is a benchmark model, not an authoritative bibliographic database: predicted links can be wrong and should not assign credit, evaluate researchers, or alter citation records without verification. We use only released data and public task metadata and infer no sensitive personal attributes. GPU computation is the main environmental cost; 4-bit quantization, LoRA, and constrained one-step inference reduce it relative to full-model training and free-form generation.

## Acknowledgments

The author thanks the WASP 2026 organizers, NASA ADS, and AstroExplorer for the ASTRO-CLIMB benchmark.

## References


Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

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

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

NASA Astrophysics Data System. 2026. AstroCLIMB dataset. Hugging Face Datasets. Accessed 2026-09-18.

Averi Nowak, Francesco Piccinno, and Yasemin Altun. 2024. Multimodal chart retrieval: A comparison of text, table and image based approaches. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5488–5505.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–269.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. 2025. DINOv3. *arXiv preprint arXiv:2508.10104*.

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.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

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

## A Objective and Prompt Details

Because the ground-truth relation is symmetric, each training pair is swapped with probability 0.5:

$$(\tilde{o}_{i1}, \tilde{o}_{i2}) = \begin{cases} (o_{i2}, o_{i1}) & b_i = 1, \\ (o_{i1}, o_{i2}) & b_i = 0, \end{cases} \qquad (6)$$
$$b_i \sim \text{Bernoulli}(0.5).$$

The label is unchanged. For the restricted objective, the denominator contains only the four valid digit tokens:

$$\mathcal{L}_4 = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp z_{i,t_{y_i}}}{\sum_{c=0}^{3} \exp z_{i,t_c}}. \qquad (7)$$

Regardless of the training objective, inference reads only those digit logits:

$$\hat{y}_i = \arg \max_{c \in \{0,1,2,3\}} z_{i,t_c}. \qquad (8)$$

The system instruction used in the main runs is:

> You classify the relationship between two objects from astronomy papers.
> 0: The objects are the figure and caption of the same scientific figure.
> 1: The objects are from different figures in the same paper.
> 2: The objects are from different papers and one paper cites the other.
> 3: The objects are from unrelated papers.
> The relationship is symmetric. Output only one digit: 0, 1, 2, or 3.

Each object is introduced as either a "scientific figure" or a "figure caption." The final user instruction is: "Classify their relationship. Reply with one digit only."

## B Additional Experimental Results

Our non-generative baseline combines frozen SPECTER2 caption embeddings (Singh et al., 2023), SigLIP2 image–text embeddings (Tschannen et al., 2025), and DINO visual embeddings (Siméoni et al., 2025) with word/character TF–IDF, OCR overlap, and perceptual-hash similarities. Three modality-specific CatBoost classifiers handle caption–caption, mixed, and image–image pairs. It is therefore a frozen-feature classifier pipeline, not a SPECTER-only model. Its submission scores 0.48279 final-leaderboard macro-F1.

The auxiliary stage contains 20,000 pairs distributed as $(2000, 6000, 8000, 4000)$, followed by restricted-loss tuning on 9,200 competition pairs. Its confusion matrix (gold rows, predicted columns) is

| Model | Test macro-F1 |
|---|---|
| Qwen3-VL-4B-Instruct | **0.47956** |
| Qwen2.5-VL-7B-Instruct | 0.46745 |
| Qwen3-VL-8B-Instruct | 0.40112 |

Table 2: Zero-shot results on the 10,000-pair test set.

| Metric | F1 |
|---|---|
| `same_figure` | 0.92913 |
| `same_paper` | 0.66988 |
| `related_papers` | 0.53790 |
| `unrelated_papers` | 0.77975 |
| Macro average | **0.72916** |

Table 3: Fixed-split results for restricted-loss 8B QLoRA.

$$\begin{bmatrix} 165 & 30 & 5 & 0 \\ 4 & 131 & 62 & 3 \\ 0 & 51 & 113 & 36 \\ 0 & 5 & 38 & 157 \end{bmatrix}. \qquad (9)$$

The largest confusion is bidirectional between `same_paper` and `related_papers` (62 and 51 cases), followed by `related_papers` and `unrelated_papers` (36 and 38); together these account for 187 of the 234 errors. Auxiliary training improves the two different-paper classes but degrades `same_figure` and `same_paper`. Modality-level macro-F1 is 0.72845 on caption–caption, 0.67108 on caption–image, and 0.65102 on image–image pairs. These values diagnose the auxiliary model, not the full-data submission.

**Validation uncertainty.** Macro-F1 depends only on the confusion matrix. Resampling validation examples within each gold class (200 per class) is therefore equivalent to resampling each row of Eq. (9) multinomially. Over 20,000 replicates, the auxiliary model's macro-F1 has a 95% percentile interval of $[0.6845, 0.7431]$ (standard deviation 0.0149). The per-class intervals are $[0.860, 0.926]$, $[0.578, 0.678]$, $[0.487, 0.593]$, and $[0.751, 0.833]$. Confusion matrices of the other validation runs were not recorded, but models of similar accuracy on this split should have intervals of similar width. Because the interval is unpaired, it is conservative for comparisons between two models evaluated on the same examples. Scaling the standard deviation by $\sqrt{800/10{,}000}$ gives the approximate test half-width of 0.008 used in

| Metric | Baseline | +Auxiliary |
|---|---|---|
| Macro-F1 | **0.72916** | 0.71405 |
| `same_figure` | **0.92913** | 0.89431 |
| `same_paper` | **0.66988** | 0.62830 |
| `related_papers` | 0.53790 | **0.54067** |
| `unrelated_papers` | 0.77975 | **0.79293** |

Table 4: Per-class F1 before and after auxiliary training.

| Setting | Value |
|---|---|
| Backbone | Qwen3-VL-8B-Instruct |
| Component seeds | 17, 42, 123 |
| Training pairs | 10,000 per seed |
| Epochs | 1 per seed |
| Learning rate | $10^{-4}$ |
| Optimizer | paged AdamW, 8-bit |
| Scheduler | cosine; 5% warm-up |
| Weight decay | 0.01 |
| Gradient clipping | maximum norm 1.0 |
| Quantization | 4-bit NF4, double quantization |
| Compute dtype | FP16 |
| LoRA $(r, \alpha, p)$ | $(16, 32, 0.05)$ |
| LoRA targets | language $q/k/v/o$ projections |
| Global batch size | 16 |
| Maximum image area | $448 \times 448$ |
| Maximum caption length | 3,000 characters |
| Training augmentation | object swap, $p = 0.5$ |
| Ensemble | equal mean of three probability vectors |
| Postprocessing | mask class 0 for CC and II pairs |
| Inference | four digit logits, ensemble, mask, argmax |
| Hardware | 2× NVIDIA T4; CPU postprocessing |
| Checkpoint policy | final epoch; no validation or early stopping |

Table 5: Configuration of the best leaderboard system.

Section 5.2. This ignores the different test class prior and is only a rough guide.

**Auxiliary data properties.** Classes 1–3 of the auxiliary set contain only caption–caption pairs, and the auxiliary epoch precedes, rather than mixes with, competition tuning. Class 0 draws 2,000 rows with replacement from the 800 `same_figure` rows of the 9,200-pair training split. Relative to the baseline, the auxiliary model gains 0.003 and 0.013 F1 on `related_papers` and `unrelated_papers`, but loses 0.035 and 0.042 on `same_figure` and `same_paper`. This is consistent with a shift toward caption-level topical cues.

| Prediction | Share | Agree | Med. $p_{\max}$ | $<0.6$ |
|---|---|---|---|---|
| `same_figure` | 9.9% | 92.7% | 0.956 | 6.8% |
| `same_paper` | 24.9% | 85.4% | 0.739 | 26.4% |
| `related_papers` | 35.0% | 83.4% | 0.566 | 65.5% |
| `unrelated_papers` | 30.3% | 89.7% | 0.854 | 17.8% |
| All rows | 100% | 86.7% | | |

| Classes involved | Seed disagreements | Low-margin rows |
|---|---|---|
| 1 and 2 | 634 (47.8%) | 1,335 (58.2%) |
| 2 and 3 | 558 (42.1%) | 874 (38.1%) |
| 0 and 1 | 133 (10.0%) | 86 (3.7%) |
| 0, 1, and 2 | 1 (0.1%) | — |
| Total | 1,326 | 2,295 |

Table 6: Label-free diagnostics on the 10,000-pair test set. Top: final-system predictions by class, with the share of rows where all three seeds agree, median maximum probability, and the share below 0.6. Bottom: classes among the three seeds' hard predictions when they disagree, and the top-two classes of rows whose final top-two margin is below 0.2.

## C Primary Configuration and Reproducibility

The full-data seed models use the final one-epoch adapter without validation-based selection; intermediate recovery checkpoints are saved every 100 steps. The restricted validation experiment evaluates at quarter-run milestones and selects checkpoint 863 (1.5 epochs) by maximum validation macro-F1 before the locked full-data refit. The restricted 8B run trains 15,335,424 parameters for 938 optimizer steps in 320.45 minutes, with 11.53 GiB peak allocated memory on rank 0. Two-GPU inference over 10,000 pairs takes 78.15 minutes. Combining stored probabilities and applying the modality mask require no model inference. The implementation uses Transformers 4.57.1, PEFT 0.17.1, Accelerate 1.10.1, and bitsandbytes 0.47.0, and validates unique IDs, exact column order, finite normalized probabilities, binary targets, one prediction per row, and 10,000 output rows.