Public Scientific-Figure Adaptation for AstroCLIMB with a 4B Vision–Language Model

Irving Ernesto Quezada Ramírez
Independent Researcher
research@irvingernesto.com

Abstract

We adapt Qwen3.5-4B to the AstroCLIMB task, which asks whether two scientific figures or captions come from the same figure, the same paper, citation-linked papers, or unrelated papers. For caption–image pairs, we first adapt the model on 80,000 public pairs labeled automatically from figure, paper, and citation structure in the public corpus, then train on gold data. Across three gold-training seeds that share one public checkpoint, this initialization improves caption–image development macro-F1 by 0.067 on average over gold-only training under matched gold schedules, although with more total compute. Answer generation and the reinforcement-learning variants we tested do not beat the classifier. The combined system of three modality-specific classifiers reaches a pooled development macro-F1 of 0.756 and a Kaggle leaderboard score of 0.751. Repeated development selection and earlier project exposure to competition materials confine these findings to the evaluated setting, and we document both.

1 Introduction

Much of the evidence in an astronomy paper lives in its figures, yet figure content is hard to search: a light curve or a density cross-section is indexed, if at all, only through its caption. The AstroCLIMB shared task (Grezes et al., 2026; WASP Organizers, 2026) benchmarks whether multimodal models can recover relations between papers from figures and captions alone. Given two objects, a system must decide whether they come from the same figure, the same paper, two papers connected by a citation in either direction, or unrelated papers. Inputs come in three forms: caption–image (CXI), image–image (IXI), and caption–caption (CXC). The organizers frame this as a multimodal benchmark rather than a deployed citation-reconstruction tool, and we make no deployment claim.

The task is harder than its four labels suggest. A same-paper pair and a cited-paper pair can share a topic, the same astronomical objects, and the same plotting conventions, so surface similarity separates neither from the other. At the same time, the task offers an unusual resource. The public AstroCLIMB corpus records figure, paper, and citation structure for tens of thousands of figures, so relation labels can be derived from the graph without manual annotation. Our question is whether that free supervision helps a compact vision–language model before it sees the gold training data.

We answer this with a system of three modality-specific classifiers built on Qwen3.5-4B (Qwen Team, 2026) with low-rank adaptation (LoRA; Hu et al., 2021). Our contributions are:

- A public adaptation stage for the caption–image condition whose labels come from the public corpus graph, with content-matched exclusion of gold and test papers, followed by gold training. In a three-seed matched comparison it improves development macro-F1 by 0.067 on average over gold-only training (Section 4).
- Per-class development results for every input type, showing that related-paper pairs are the weakest class throughout and that same-paper versus related-paper confusion accounts for half of all errors.
- Comparisons that bound the recipe: longer public adaptation trades accuracy between classes, and answer generation with supervised or reinforcement-learning objectives does not surpass the classifier.
- An explicit account of prior project exposure to competition materials and of the dataset's label noise (Section 5), together with public source, prompts, and adapter weights.

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 154–165
November 10, 2026 ©2026 Association for Computational Linguistics

Input type	Training	Development
Caption–image (CXI)	2,230	794
Image–image (IXI)	1,670	590
Caption–caption (CXC)	1,649	602
Total	5,549	1,986

Table 1: Gold partitions after leakage filtering.

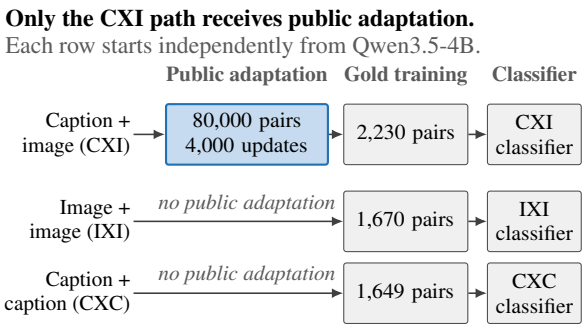


Figure 1: Training overview. One classifier per input type, each trained independently from the same frozen Qwen3.5-4B base (rank-16 LoRA adapters, four-class head). Only the caption–image (CXI) classifier receives public adaptation (80,000 public pairs; gold and test papers excluded) before gold training; IXI and CXC train on gold data alone.

Related work. Prior work has learned from scientific figures and their captions for correspondence (Gómez-Pérez and Ortega, 2019), aligned medical images with text (Subramanian et al., 2020), generated captions for scientific figures (Hsu et al., 2021), and pretrained multimodal scientific models at scale (Tarsi et al., 2024). SPECTER learns document representations from citation signals (Cohan et al., 2020); we instead derive figure-level pair labels from citation structure and use them to adapt a generative vision–language model.

2 Task and Data

Gold partitions. We inherit a gold split made earlier in the project and do not rebuild it. The split groups pairs by paper component, so pairs connected through shared papers form one gold group, and each group lies entirely in one of five folds. We train on folds 2–4 and select models on fold 1, the development set. Fold 0, with 2,001 pairs, was exposed during earlier project work; we exclude it from all fitting, selection, and evaluation, but it is not an untouched test set. After the duplicate and leakage checks in Appendix A, 5,549 training and 1,986 development pairs remain (Table 1).

Public supervision. We index all 114 shards of the public `adsabs/AstroCLIMB` release (94,233 figures) and label caption–image pairs from its recorded figure identity, paper identity, and citation edges. Because a paper can appear more than once in the release, we connect public papers with matching figure content or identifiers into groups, which we use together for exclusion and for the public train–validation split; pair labels use the original paper identities (Appendix A). Before drawing any pair we match public papers, by normalized caption and image hash, against every gold split and against the test figures and captions, and we drop every public component that matches. From the 21,258 figures eligible under the conditions in Appendix A, we draw 80,000 class-balanced training pairs and 2,000 validation pairs from disjoint paper components. Unrelated pairs mix random-paper pairs, same-journal pairs, and two-hop pairs: papers joined through a common third paper in the citation graph but with no citation edge between them. The same matches also prune gold groups that public papers connect to reserved gold, meaning gold held out of both training and development, including fold 0, or to test content. Test content serves only for exclusion and never carries labels. These labels are not manually verified (Section 5). Appendix A gives the procedure and counts.

Selection and reporting. The task metric is four-class macro-F1 computed from the pooled confusion matrix over all development pairs. Because IXI and CXC contain no same-figure pairs, we also report an active-class macro over the three classes present; the four-class macro scores the absent class as zero. Exploratory feasibility trials (Appendix G) were screened on fixed subsets of the gold training set declared before each trial, which we call TRAIN monitors. Public validation pairs never decided when adaptation stopped. All development numbers here are selection results rather than independent tests (Section 5).

3 System

Figure 1 shows the pipeline; the submission concatenates the three classifiers' predictions.

Classifiers. Each classifier adds rank-16 LoRA adapters and a four-class head to a frozen BF16 Qwen3.5-4B base, implemented with Transformers (Wolf et al., 2020) and PEFT (Mangrulkar et al., 2022). The adapters cover language-model MLP

and attention projections, about 21.2M trainable parameters; the vision tower stays frozen. The head reads the final non-padding representation and is trained with cross-entropy. No prompt contains paper identifiers or retrieved labels, and for IXI and CXC the same-figure logit is masked in training and prediction. Appendix B lists adapter placement, prompt handling, and hyperparameters.

Public adaptation for CXI. We adapt only the CXI classifier on public data because CXI is the cross-modal condition and the only one containing all four labels, so it permits one controlled comparison of public initialization against gold-only training without a masked class. The classifier is first trained on the 80,000 public pairs with a 10,000-update warmup–cosine schedule, and the checkpoint saved at update 4,000 is the parent from which gold training starts. Public adaptation for IXI and CXC was not evaluated in this compute-limited study; its absence reflects scope, not a demonstrated optimum.

Gold training and prediction. CXI gold training starts from the parent's adapter and head with a fresh optimizer; IXI and CXC start from the pristine base. All three use the same three-epoch gold schedule at seed 7, and we keep whichever end-of-epoch checkpoint has the highest development macro-F1. Each classifier predicts by argmax without imposing test class counts. For IXI we average class probabilities over both image orders, a choice fixed before evaluation; caption-order averaging did not help CXC, which uses the original order. The submission uses the development-selected models directly, without refitting on all labels. Inference runs at batch size one, because a larger batch changed a few predictions in an earlier model (Appendix C).

4 Results

Submitted system. Table 2 reports per-class development F1 for the submitted classifiers; the pooled four-class macro-F1 is 0.756. Same-figure pairs, which occur only in CXI, are nearly solved. Related-paper pairs are the weakest class in every input type, and the pooled confusion matrix (Appendix D) shows that same-paper versus related-paper confusions make up 281 of the 549 errors, or 51%. Frozen-feature baselines built before training the IXI and CXC classifiers score lower: a blend of a frozen SigLIP 2 (Tschannen et al., 2025) classifier and an OCR-topic model reaches an active-class macro of 0.541 on IXI, against 0.678 for the classifier, and logistic regression on frozen caption embeddings (Zhang et al., 2025; Beijing Academy of Artificial Intelligence, 2023; Xiao et al., 2023) reaches 0.673 on CXC, against 0.728 (Appendix F).

On the Kaggle leaderboard the submission scores 0.751, rank 2 among 17 ranked teams as of 12 October 2026. The leaderboard page states that it is calculated with all of the test data (Kaggle, 2026), and no separate private score is displayed, so no withheld private Kaggle subset awaits evaluation. Because entries were still being accepted, the rank is a dated snapshot rather than an organizer-confirmed final placement.

Matched gold-only controls. The central question is whether public adaptation helps. Under one CXI gold schedule, on one RTX 3090 runtime, with gold seeds 7, 19, and 37, the public arm starts gold training from the parent (the public checkpoint at update 4,000) and the gold-only arm from the pristine base. A fresh export of each run reproduces its selected and final predictions exactly. Public initialization improves every seed pair (Table 3; Figure 2b), by 0.067 on average at the development-selected checkpoint and 0.068 at the final epoch, and within each seed both arms select the same gold update. Several boundaries apply. The three public-arm runs share one parent, so the seeds vary only the gold stage. Total compute is not matched, since the public arm adds its adaptation to the same gold schedule. That parent was itself chosen on this development set among three durations, a choice the gold-only arm lacked, and the shared gold schedule was inherited from the public-initialized recipe rather than retuned for gold-only training, so the gain holds under this fixed schedule, not over an optimized gold-only baseline.

Historical versus fresh runs. The submitted CXI checkpoint, trained earlier on an L40S, scores 0.748, whereas the fresh seed 7 public arm on the RTX 3090 scores 0.756 from identical parent files, data hashes, and hyperparameters. Each run reproduces its own predictions exactly, so the gap lies between separate optimizer runs. Without a controlled cross-hardware test we cannot attribute it to hardware rather than to software or runtime (Appendix C), so we compare fresh arms only within one runtime, and the 0.008 gap shows that this single-seed setup cannot establish differences of a few thousandths as reliable improvements.

Input	SF	SP	REL	UNR	Macro (active)	Macro (4)	N
Caption–image	0.927	0.654	0.610	0.803	0.748	0.748	794
Image–image	—	0.712	0.554	0.768	0.678	0.508	590
Caption–caption	—	0.751	0.593	0.840	0.728	0.546	602
Pooled	0.927	0.708	0.586	0.804	0.756	0.756	1,986

Table 2: Per-class F1 of the submitted system on the development set (not training data). SF, SP, REL, and UNR denote same figure, same paper, related papers, and unrelated papers. A dash marks a class absent from that input type. Macro (active) averages over the classes present; Macro (4) averages over all four classes and scores an absent class as zero. The pooled row is computed from the pooled confusion matrix over all 1,986 pairs and is not the mean of the per-type macros.

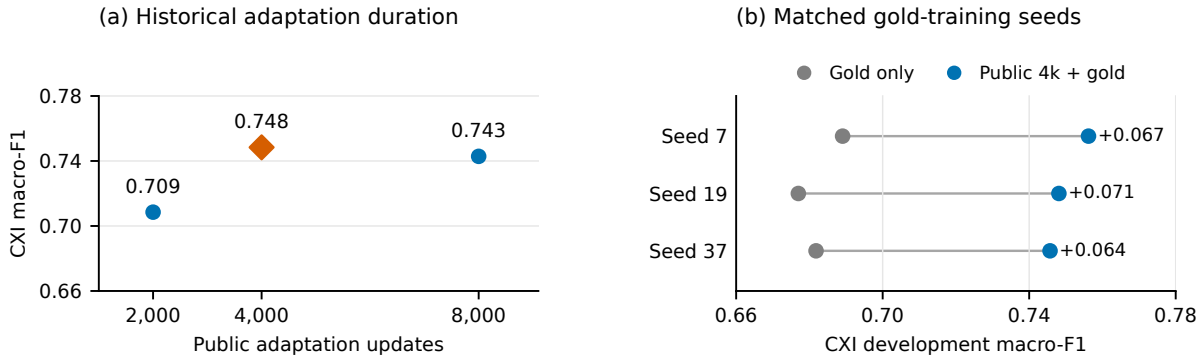


Figure 2: CXI development macro-F1 (794 pairs; zoomed axes). (a) Historical comparison of public adaptation duration with the gold schedule fixed; the diamond marks the submitted model. (b) Fresh runs paired by gold seed; public-initialized runs share one parent. Gold schedules match, total compute does not. Both panels use development-selected checkpoints.

Gold seed	Gold only	Public 4k	Gain
7	0.689	0.756	+0.067
19	0.677	0.748	+0.071
37	0.682	0.746	+0.064
Mean (selected)			+0.067
Mean (final epoch)			+0.068

Table 3: Matched CXI controls on 794 development pairs with epoch-boundary selection and one fixed public parent. These are fresh runs on one runtime, distinct from the submitted model.

Public adaptation duration. A historical single-seed comparison fixes the gold schedule and varies the number of public updates (Figure 2a). Moving from 2,000 to 4,000 updates raises CXI development macro-F1 from 0.709 to 0.748, mainly through the paper-relation classes, with 81 corrections against 49 regressions on the same 794 pairs. Doubling to 8,000 updates gives 0.743: same-paper F1 falls by 0.049 while related and unrelated F1 rise, with 45 corrections and 47 regressions. This is a redistribution across classes, not evidence of a capacity ceiling. At the pooled level the 8,000-update branch is less than 0.001 below the selected system, far less than the cross-runtime gap above, so we retained the 4,000-update branch on its CXI score alone. Appendix E gives the per-class changes and group-resampling intervals for these fixed historical predictions.

Generative and reinforcement-learning alternatives. We also asked whether answer generation could beat the classification head. From the classifier-adapted model, we taught canonical answer tags in a short supervised stage and then applied Group Sequence Policy Optimization (GSPO; Zheng et al., 2025a) with verifiable rewards. Neither the 32- nor the 512-update GSPO model beats the classifier on the pooled development set: GSPO at 512 updates scores 0.751, above answer-supervised fine-tuning at 0.746 and below the classifier at 0.756. Its corrections concentrate on related-paper cases, but without a supervised control matched at 512 updates the gain over answer-supervised training cannot be credited to reinforcement learning. In our group-relative reward setup, a prompt contributes a learning signal

only when its sampled answers disagree in reward, so we also tried selecting such mixed-outcome prompts beforehand, as in dynamic sampling (Yu et al., 2025); this raised the share of informative rollout groups but left a fixed TRAIN monitor essentially unchanged. OCR features added little. Appendices F and G give these experiments in full.

5 Dataset and Exposure Limitations

What the development numbers mean. Every development score here comes from a set consulted repeatedly for selection across the project, and the TRAIN monitors used for screening are training pairs already used to fit the classifier from which the screened models descend. Neither is an independent test. The Kaggle score is computed on all of the test data, but the earlier exposure described below limits how independent even that score is. All comparisons other than the three-seed matched controls rest on a single seed.

Prior project exposure. Earlier work on this project, predating the reported system, was exposed to competition materials in two ways that cannot be undone. First, a scoring-label file named `solution.csv`, distributed in the competition bundle, was inspected during earlier project work, and an identifier-only exploration inferred labels from the class-block ordering of pair identifiers. Second, an earlier synthetic caption corpus of 966,646 pairs overlapped 968 gold caption pairs and used journal-sampling proportions estimated from all gold unrelated examples, including held-out gold. The reported recipe starts from a pristine pretrained base, a new public corpus with gold and test content excluded by component, and gold training data only; every earlier fine-tuned model, feature set, and prediction affected by that exposure is excluded. Pair identifiers serve only to join records, never for prediction, and no test-label supervision or forced test class counts enter the system. We therefore do not claim a clean, independent benchmark result.

Public label noise. The public labels derive from an observed citation graph without manual verification, so a missing edge can make a truly related pair look unrelated, and the duplicate-identity mapping used for exclusion is incomplete. Zero detected overlap after filtering is not proof of independence. All three matched public arms descend from one parent, since we ran no independent public-stage seeds, and exposure of the base model to the underlying astronomy papers during pretraining is unknown.

Reproducibility. We hash source, data, model, and prediction artifacts, and we accept an export only after a fresh inference run reproduces it exactly. Source, prompts, configurations, and adapter and head weights are public at `https://github.com/Sekinal/AstroCLIMB`, which links the model downloads. The packaged inference script verifies the released files against a hash manifest before running but has not been validated on a GPU, and full retraining from the release has not been independently verified.

6 Conclusion

Public initialization from graph-derived caption–image labels improves a 4B vision–language classifier over gold-only training across three gold-stage seeds, with one public parent and extra compute. In our single-seed sweep, extending public adaptation from 4,000 to 8,000 updates traded performance between classes; generative and reinforcement-learning alternatives do not surpass classification, and related-paper pairs remain the hardest class in every input type. Future work includes public-stage seed replication, a compute-matched gold-only arm, public adaptation for the image–image and caption–caption conditions, a long supervised control for the 512-update GSPO run, and independent evaluation and retraining from the release.

Acknowledgments

GLM 5.3 Flash, GLM 5.3, DeepSeek Flash V4, GPT 6 Astra, and Muse Spark 1.3 Contributor assisted with literature discovery, experiment planning, code generation and debugging, experiment orchestration, evidence review, and drafting and revising all sections of the manuscript. Claude Fable 5.1 assisted with manuscript review and prose revision. Responsibility for the submitted claims, citations, code, and interpretation of results rests solely with the author.

The results figure was produced with Matplotlib (Hunter, 2007).

I thank Lium for providing GPU credits used in this work.

References

Beijing Academy of Artificial Intelligence. 2023. bge-base-en-v1.5 Model Card. Official software or model resource; accessed 2026-09-14.

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.

José Manuel Gómez-Pérez and Raúl Ortega. 2019. Look, read and enrich - learning from scientific figures and their captions. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 101–108. ACM.

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

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

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

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Kaggle. 2026. AstroCLIMB leaderboard. Accessed 12 October 2026. The page states that the leaderboard is calculated with all of the test data.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. *Preprint*, arXiv:2309.06180.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-RFT: Visual Reinforcement Fine-Tuning. *Preprint*, arXiv:2503.01785.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Qwen Team. 2026. Qwen3.5-4B model card. Model revision 851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *Preprint*, arXiv:2504.07615.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. MedICaT: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120, Online. Association for Computational Linguistics.

Tim Tarsi, Heike Adel, Jan Hendrik Metzen, Dan Zhang, Matteo Finco, and Annemarie Friedrich. 2024. SciOL and MuLMS-Img: Introducing a large-scale multimodal scientific dataset and models for image-text tasks in the scientific domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4548–4559.

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. *Preprint*, arXiv:2502.14786.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

WASP Organizers. 2026. AstroCLIMB: Shared task at WASP 2026.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2025. Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs. *Preprint*, arXiv:2506.14245.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-Pack: Packed Resources For General Chinese Embeddings. *Preprint*, arXiv:2309.07597.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *Preprint*, arXiv:2504.13837.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *Preprint*, arXiv:2506.05176.

Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng, Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou, Hongen Liu, Suyin Liang, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. 2026. PP-OCRv6: From 1.5M to 34.5M parameters, surpassing billion-scale VLMs on OCR tasks. *Preprint*, arXiv:2606.13108.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025a. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025b. DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning. *Preprint*, arXiv:2505.14362.

A Public Supervision Construction

Matching and exclusion. The goal of matching is to find every public figure that is also a gold or test figure, so that public supervision never reuses that content. We compare every public paper against every gold split and against the test figures and captions, using their content but not their labels. Two figures match if their normalized captions are equal, if their image bytes or decoded pixels hash to the same value, or if their 256-bit perceptual difference hashes differ in at most eight bits. Because one paper can appear more than once in the public release, we first connect public papers into groups, which we call components, whenever their figures share an exact normalized caption, image hash, or figure identifier, or have nearly matching perceptual hashes. These components serve only for exclusion and for splitting public papers between training and validation; pair labels use the original paper identities. We then forbid every component that contains a match. This forbids 7,500 public papers. A figure is eligible for pair sampling only if it meets four conditions: it lies outside the forbidden components, its paper has a nonempty DOI, its image-byte and difference hashes were computed successfully, and its normalized caption has at least 40 characters. Exclusion and these eligibility conditions together leave 21,258 eligible figures out of the 94,233 indexed in the 114 shards of `adsabs/AstroCLIMB`. Public training and validation pairs are drawn from disjoint paper components.

Removing overlap between splits. The same matches can link gold groups to content they must not share. A gold object that matches a public paper joins that paper's component, so gold groups can be linked indirectly through chains of public papers and other groups. We therefore take connected components of the combined graph of gold groups, gold objects, and public components, and remove whole gold groups by split (Figure 3):

- a training group is removed if its component contains any nontraining gold object (development, fold 0, or other reserved gold) or any test object;
- a development group is removed if its component contains any reserved gold object (gold outside both training and development, including the exposed fold 0) or any test object;
- if a training and a development group share a component with no other forbidden link, only the training group is removed.

This removes 450 training and seven development pairs, leaving 5,549 and 1,986, with no remaining intersections under these checks. The identity map behind the matches is incomplete, so zero detected overlap is not proof of independence.

Pair sampling. Same-figure, same-paper, and related pairs are labeled directly from figure identity, paper identity, and observed citation edges. Unrelated pairs are drawn from three sources: random pairs of papers, pairs of papers from the same journal, and two-hop pairs, meaning papers joined through a common third paper in the citation graph but with no citation edge between them. The same-journal and two-hop sources are intended to supply negatives that are closer in subject than random pairs; we did not test whether they are harder or whether they share a topic. The 80,000 training pairs are class-balanced. The 2,000 validation pairs only monitored training and never decided when adaptation stopped.

B Training Details

Adapters. Each classifier adds rank-16 LoRA adapters (alpha 32, effective scaling 2) and a four-class head on the final non-padding representation of a frozen BF16 Qwen3.5-4B base. The adapters cover the language-model MLP projections in all 32 layers and the query, key, value, and output projections in the eight full-attention layers, about 21.2M parameters; the vision tower and the linear-attention projections stay frozen. Training minimizes cross-entropy without adapter dropout.

Prompts. CXI prompts hold one image and one caption, IXI prompts hold Figure A and Figure B, and CXC prompts hold two separated captions. Public captions are lowercased and whitespace-normalized, gold captions retain their original case, and both are truncated at 2,400 characters. Images are decoded to RGB from verified original bytes and capped at 262,144 pixels. No prompt contains paper identifiers or retrieved labels. For IXI and CXC the same-figure logit is masked in training and prediction.

Public adaptation. The CXI public stage uses batch size four with two gradient-accumulation steps, so each update sees eight pairs. The adapter

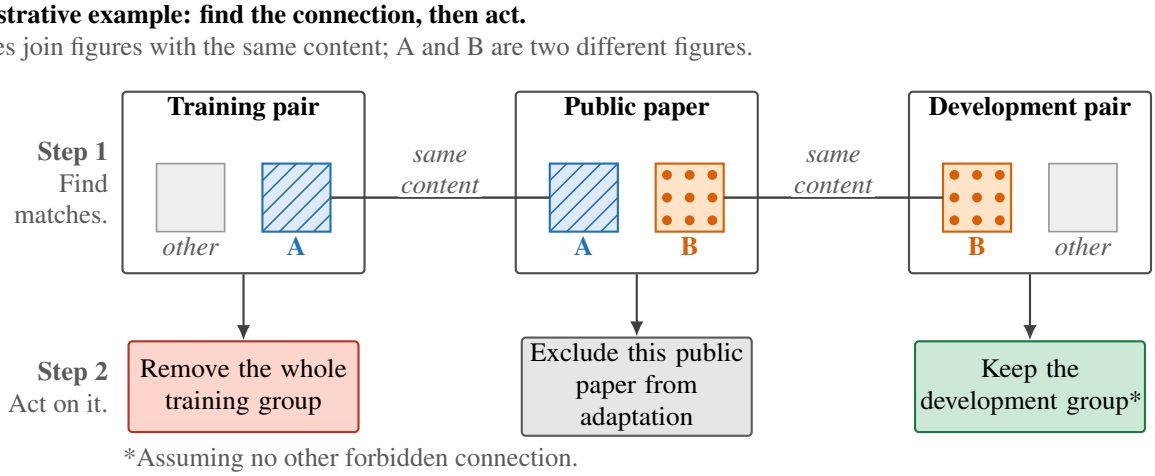


Figure 3: Illustrative example. A training pair and a development pair can overlap indirectly through different figures of one public paper: schematic tile A matches training to the paper, tile B the paper to development. The whole training group is removed, the public paper is excluded from adaptation, and the development group is kept only if no other forbidden link exists. Transitive checks and other removal rules are in the text.

and head learning rates are 10^{-4} and 3×10^{-4}, following a warmup–cosine schedule laid out over 10,000 updates. The submitted parent is the checkpoint saved at update 4,000 of that schedule, after 32,000 pair presentations. We did not rerun a schedule shortened to 4,000 updates, so the parent was taken partway through the cosine decay.

Gold training. CXI gold training starts from the parent's adapter and head with a fresh optimizer; IXI and CXC start from the pristine base. All three use batch size one with eight accumulation steps, three epochs, seed 7, and adapter and head learning rates of 3×10^{-5} and 10^{-4}. With these settings, three epochs amount to 837 updates for CXI, 627 for IXI, and 621 for CXC. We evaluate on the development set at the end of each epoch and keep the checkpoint with the highest development macro-F1, choosing the earlier checkpoint on ties. The submitted checkpoints are CXI at update 837, IXI at update 627, and CXC at update 414. The matched controls in Section 4 reuse this CXI schedule with gold seeds 7, 19, and 37 (Figure 4). All six runs complete 837 updates and are evaluated at the epoch boundaries, updates 279, 558, and 837. For seed 19 both arms select update 558; for seeds 7 and 37 both select update 837.

C Prediction and Reproducibility

Batch size. We run inference at batch size one, the same batch size used in training, because batching changed predictions in an earlier model. That model, trained from the 2,000-update public parent to gold update 837, changed 11 of its 794 CXI development predictions when scored at batch size two, with a class probability moving by as much as 0.380 for one pair. We did not isolate the numerical cause. Exporting at batch size one does not explain the effect; it removes it from our pipeline so that our exports replay exactly, which is not a general guarantee of reproducibility.

Cross-runtime difference. Two runs with the same recipe on different machines gave different scores. The submitted CXI checkpoint was trained on an NVIDIA L40S and scores 0.748 on the 794 development pairs. The fresh seed 7 public arm of the matched controls was trained on an RTX 3090 from the same public parent files, the same data hashes, and the same gold hyperparameters, and scores 0.756; the two configurations differ only in file paths and a wall-clock safety deadline. Both runs complete 837 updates. For each run, a fresh inference pass reproduces that run's own predictions exactly, which shows that each run is internally consistent but does not show that training is reproducible bit for bit across hardware. Different floating-point kernels and reduction orders in BF16, together with nondeterministic GPU execution, are plausible mechanisms. We ran no controlled cross-hardware test, so we cannot attribute the 0.008 gap to hardware rather than to software or runtime, and we do not claim that GPU type alone causes it.

Verification and release. We hash source, data, model, and prediction artifacts, meaning the code, inputs, weights, and prediction files on which an

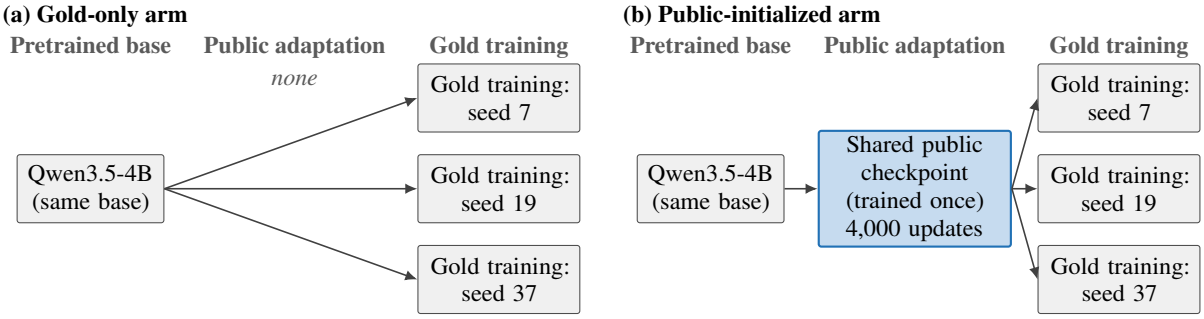


Figure 4: Design of the three-seed matched control (Section 4, Table 3). Both arms start from the same pretrained base and run the same CXI gold schedule on the same runtime with gold seeds 7, 19, and 37. (a) The gold-only arm trains each seed directly from the base. (b) The public-initialized arm first trains one public adaptation checkpoint, saved at update 4,000, and all three gold runs branch from that single checkpoint, so the seeds replicate only the gold stage, not public adaptation. Gold schedule and runtime match across paired seeds; the public arm uses extra adaptation compute. Arrows denote training progression.

export depends, and we accept an export of predictions only after a fresh inference run reproduces it exactly. Source, prompts, configurations, and adapter and head weights are released. The packaged inference script checks every released weight and configuration file against a hash manifest before it runs, but we have not validated it on a GPU from a clean install, and full retraining from the release has not been independently verified. Kaggle displays the leaderboard score as 0.75149.

D Confusion Matrices and Examples

Figure 5 gives the development confusion matrices of the submitted classifiers, with true classes in rows and predicted classes in columns. These describe repeatedly selected development results and are not an independent evaluation.

Two caption–caption errors. To illustrate the dominant confusion we selected, deterministically, the development pair with the smallest identifier in each direction of the same-paper versus related-paper confusion for the caption–caption classifier. We paraphrase the captions rather than quote them. In the first pair, the gold label is same paper and the classifier predicted related papers. One caption describes a horizontal cross-section of the initial density in a simulation; the other describes the time-integrated Poynting flux through a boundary in tuned and detuned simulations. The two captions concern different physical quantities from the same study, which illustrates how judging a same-paper relationship can be difficult when the captions describe different measurements. In the second pair, the gold label is related papers and the classifier predicted same paper. One caption discusses the radial structure, pressures, and luminosity of a simulated accretion disk; the other plots black-hole spin-up against spin and accretion rate from general-relativistic magnetohydrodynamic simulations. The captions share an accretion theme and a simulation vocabulary, but a shared theme does not establish a shared source paper. These are descriptive selected errors; they do not explain the model's internal reasoning or confirm that the gold labels are correct, and we draw no claim about the underlying papers from them.

E Public Adaptation Duration

The duration comparison in Section 4 is a historical single-seed sweep on an L40S with the gold schedule fixed. Development selection chose gold update 837 for the 2,000- and 4,000-update parents and update 279 for the 8,000-update parent. Per-class CXI F1 changes from 2,000 to 4,000 public updates are -0.005 for same figure, $+0.052$ for same paper, $+0.072$ for related papers, and $+0.041$ for unrelated papers, with 81 corrections against 49 regressions on the same 794 pairs. From 4,000 to 8,000, same-paper F1 falls by 0.049 while related and unrelated F1 rise by 0.018 and 0.011, with 45 corrections and 47 regressions. Both pooled scores round to 0.756; the 8,000-update branch is lower by less than 0.001.

To gauge how much these differences depend on which development pairs happen to be present, we resample the development set at the level of its 278 inherited gold groups. We draw the groups with re-

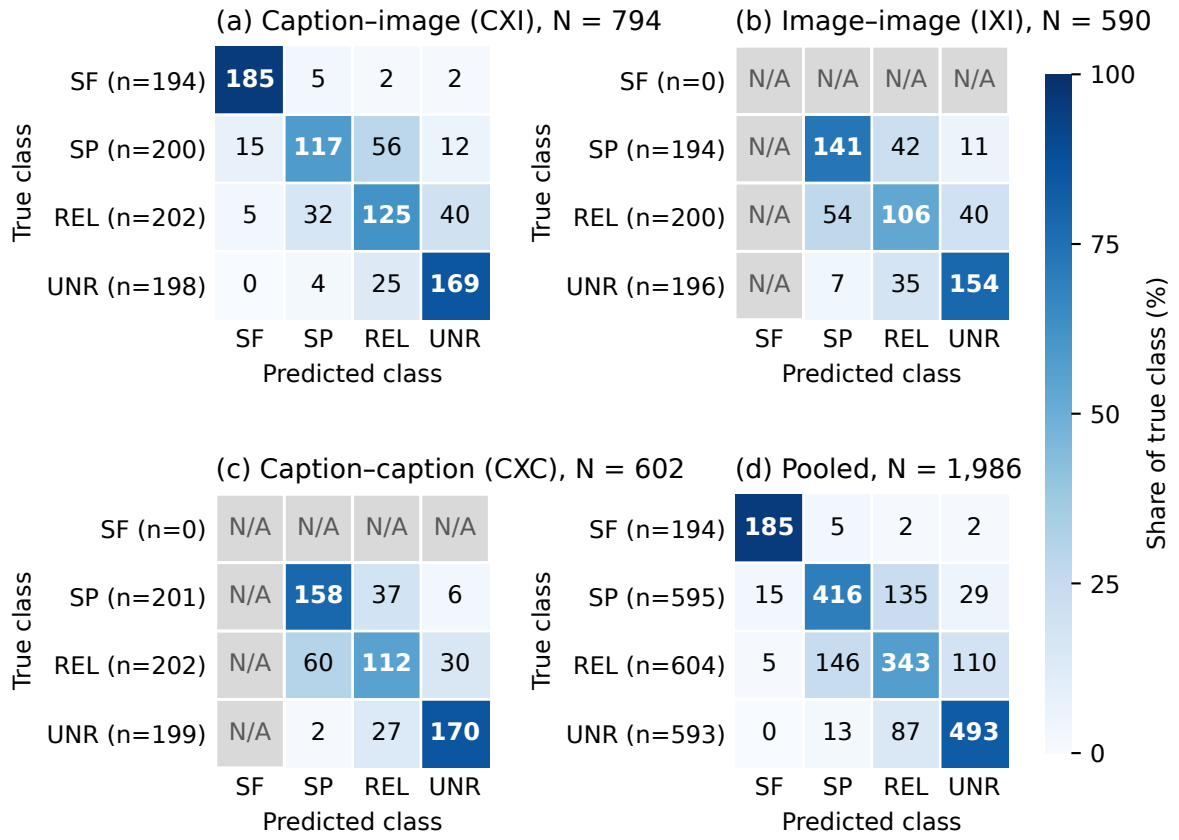


Figure 5: Development confusion matrices for (a) caption–image, (b) image–image, (c) caption–caption, and (d) all pairs pooled. Rows are true classes with support n; columns are predicted classes. Color is the row-normalized share of each true class; numbers are raw pair counts. Gray N/A marks the same-figure class, absent from IXI and CXC and masked at prediction. SF, SP, REL, and UNR denote same figure, same paper, related papers, and unrelated papers. The pooled panel sums counts rather than averaging them; the task metric is its four-class macro-F1.

placement 10,000 times, keep every model's predictions fixed, recompute CXI macro-F1 from each resampled confusion matrix, and report the 2.5th and 97.5th percentiles of the difference between models. The resulting 95% intervals are $[0.007, 0.074]$ for the 2,000-to-4,000 change, $[-0.036, 0.023]$ for the 4,000-to-8,000 change, and $[-0.029, 0.002]$ for GSPO-512 minus the classifier. These intervals capture only the variation from resampling fixed historical predictions. They do not account for model selection, for training variation across runs, for the pooled scores, or for the three-seed gain. The inherited groups may miss dependencies between pairs, and an interval that spans zero does not establish that two models are equivalent.

F Frozen-Feature Baselines, OCR, and Auxiliary Transfer

Frozen-feature baselines. Before training the IXI and CXC classifiers we built frozen-feature baselines. For IXI, a 75/25 blend of a multilayer perceptron on frozen SigLIP 2 (Tschannen et al., 2025) features and a logistic model on OCR-topic features reaches an active-class macro-F1 of 0.541. The IXI classifier reaches 0.655 with a single image order and 0.678 with image-order averaging. For CXC, logistic regression on frozen Qwen3-Embedding-8B (Zhang et al., 2025) and BGE-base-en-v1.5 (Beijing Academy of Artificial Intelligence, 2023; Xiao et al., 2023) caption features with TF-IDF and lexical statistics reaches 0.673; the CXC classifier reaches 0.728 in the original caption order and 0.727 with caption-order averaging, so we keep the original order.

OCR. OCR adds little. PP-OCRv6 (Zhang et al., 2026) features alone score 0.441 on CXI, and OCR-conditioned adaptation of the vision–language model falls slightly below the pristine gold-only model, although a fixed 50/50 combination lifts a superseded whole-system baseline from 0.690 to

0.695.

Frozen-feature IXI transfer. In a separate experiment with a different architecture and training procedure, public-to-gold transfer did not help a frozen-feature IXI classifier, even in a matched-duration follow-up. This negative result concerns frozen features, not the adapted vision–language model, and it did not determine the CXI-only scope of public adaptation in Section 4; it says nothing about whether adapting the IXI vision–language classifier on public data would help.

G Generative Training and Reinforcement Learning

Motivation. Visual-RFT and VLM-R1 (Liu et al., 2025; Shen et al., 2025) motivated trying verifiable visual rewards, studies of coverage and reasoning under verifiable rewards (Yue et al., 2025; Wen et al., 2025) motivated our sampled versus single-answer evaluation, and DeepEyes (Zheng et al., 2025b) motivated fixed-view inspection trials, which were negative. None of these pipelines was reproduced in full.

Answer supervision. The generative variants predict by writing a short answer tag instead of reading the classifier head. The classifier-adapted model initially produced only 13 strictly valid answer tags in 64 completions, so we first taught it the canonical tags with supervised fine-tuning on 1,024 gold training pairs for 128 updates; we call this stage answer-SFT. All generative models reported below predict from the generated answer, not from the classifier head.

GSPO. Group Sequence Policy Optimization (Zheng et al., 2025a) is a reinforcement-learning method that samples several answers for each prompt, which we call a context, scores each answer with a reward, and updates the model toward answers that score better than the others sampled for the same context. We apply it with TRL (von Werra et al., 2020) using batch size 16, four sampled completions per context, two uses of each rollout, learning rate 10^{-6}, rewards normalized within each context's group of samples, no penalty for drifting from a reference model, and asymmetric clipping bounds of 0.0003 and 0.0004 on the sequence-level importance ratio. The reward is one only for a correct canonical answer terminated by the end-of-sequence token, and zero otherwise. On the pooled development set with IXI and CXC held fixed, the submitted classifier scores 0.756, answer-SFT 0.746, GSPO after 32 updates 0.745, and GSPO after 512 updates 0.751. Of the 512-update model's 26 corrections relative to answer-SFT, 25 are related-paper cases, and all 19 regressions lie elsewhere. A short supervised control matches the 32-update GSPO run in parent, 64 TRAIN contexts, 32 updates, batch size, learning rate, and seed 7, but not in tokens, FLOPs, or wall time, and scores 0.747. Without a supervised control matched at 512 updates, the gain of the 512-update model over answer-SFT cannot be credited to reinforcement learning. The 32- and 512-update runs also differ in context pool, not only in update count.

Mixed-outcome selection. In our setup, when every sampled answer for a context receives the same reward, the group-relative advantage is zero and that context contributes no learning signal. Dynamic sampling (Yu et al., 2025) exploits this by discarding such contexts during training. We tested an offline version of the idea on training data only. We sampled eight answers per candidate context and selected 16 contexts whose answers had mixed outcomes, some correct and some not. As a control we selected 16 random contexts with the same class distribution. Each set received 32 GSPO updates with vLLM-assisted rollouts. Mixed selection raised the number of informative groups, those with nonzero advantage, from 14 to 24 of 32. The fixed TRAIN monitor of 256 contexts barely moved: 1,874 correct answers of 2,048, against 1,875 for the control. Neither trial met the prespecified screen of a two-percentage-point improvement. This was an offline selection motivated by dynamic sampling, not a full online implementation of that method.

Backends. vLLM (Kwon et al., 2023) generates the training rollouts, but all reported evaluation uses the native Transformers inference path, because the two backends do not produce identical predictions. For scoring and for checks that a reloaded checkpoint reproduces its predictions, we match the batch shape and the BF16 autocast setting.