GRAIL: Compact Multimodal Similarities for Astronomy Citation Linking

Egor Fedosov
Ural Federal University

Abstract

We present GRAIL, a system for classifying relations between astronomy figures and captions in the AstroCLIMB shared task. GRAIL combines frozen CLIP and SciNCL representations with optical character recognition (OCR), lexical overlap, and figure layout. A factorized gradient-boosted classifier predicts four relation classes from 190 symmetric pair features. The fixed system achieves 0.48 macro-F1 under cross-validation that separates exact objects across folds and 0.48 on Kaggle. With the same boosted classifier, the full features improve local macro-F1 by 0.068 over raw CLIP/SciNCL cosines. Nested classifier selection gives no observed aggregate gain. Removing OCR or corpus-weighted lexical features lowers local macro-F1 by 0.0081 and 0.013. Direct citation relations remain the hardest class.

1 Introduction

Scientific figures and captions contain evidence about what a paper studies: object names, measured quantities, instruments, and methods. Recovering relations between papers from this evidence requires more than recognizing a shared topic. Two papers may describe similar observations without citing one another. AstroCLIMB (Grezes et al., 2026) tests this distinction by asking systems to classify pairs of figures and captions according to their source relationship.

We ask how far a compact classifier can go when it combines several views of the same pair. CLIP (Radford et al., 2021) provides a shared space for images and text, while SciNCL (Ostendorff et al., 2022) provides representations learned from scientific documents and citation neighborhoods. We supplement these representations with explicit agreement between words, numbers, and identifiers. OCR makes some of this textual evidence available inside figures. The resulting system, GRAIL, learns from 190 pair features without fine-tuning either encoder.

Our experiments test whether explicit pair features add to raw cosine similarities, whether both CLIP backbones contribute, and whether additional model selection helps. Controlled feature ablations support OCR, corpus-weighted lexical evidence, and both CLIP blocks in the fixed classifier. Nested cross-validation evaluates selecting a classifier per modality and adjusting its class decisions. The submitted variants receive similar Kaggle scores, with the fixed model slightly ahead, so the added selection procedure gives no observed aggregate gain.

2 Task and Data

AstroCLIMB provides 10,000 training pairs and 10,000 scoring pairs. Each object is a figure or a caption. The four mutually exclusive labels identify a matching figure and caption (`same_figure`), different objects from the same paper (`same_paper`), objects from papers linked by a citation in either direction (`related_papers`), or unrelated papers (`unrelated_papers`). The evaluation metric is macro-F1, the unweighted mean of the four class F1 scores.

The training set contains 4,000 figure-caption, 3,000 figure-figure, and 3,000 caption-caption pairs. Each feasible class contributes 1,000 examples within its modality. The same-figure label is possible only for figure-caption pairs. Relations are symmetric, so predictions should not depend on object order.

We process the CSV files incrementally and identify objects by hashes of their original serialized strings. Across both splits, the cache contains 14,598 images and 14,581 captions. Images are resized to fit within 384 by 384 pixels without changing their aspect ratio and saved as JPEG at quality 88; transparent regions are composited

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 99–103
November 10, 2026 ©2026 Association for Computational Linguistics

onto white. Pair IDs and row order are excluded from the predictors. We use no scoring labels or external lookups of paper titles, authors, DOIs, or citation links. Frozen encoders process objects independently. Learned corpus statistics use only objects in each fitting partition; the scoring split is never used to fit them.

3 GRAIL

GRAIL stands for Group-safe Relational Alignment for Illustrations and Language. Here, grouping refers to exact objects: it does not guarantee separation by source paper. The pipeline extracts object representations, computes symmetric pair features, and classifies each pair with a model for its modality.

3.1 Object representations

We use frozen CLIP ViT-B/16 and ViT-B/32 encoders. Each produces normalized 512-dimensional image and caption vectors. Their different patch sizes provide two visual representations; controlled ablations quantify each block's contribution in the final classifier (Section 5). A white square canvas preserves the full figure before CLIP preprocessing. CLIP captions use at most 77 tokens. SciNCL supplies 768-dimensional caption vectors from sequences of up to 512 tokens.

We also compute 33 image descriptors that summarize dimensions, aspect ratio, color and grayscale statistics, histograms, pixel fractions, and edges. Apple's Vision recognizer extracts English text in fast mode from the cached JPEGs, whose longest side is at most 384 pixels. Reusing this compact cache limited local storage and preprocessing cost; it was an engineering choice, not an OCR accuracy optimum. The recognizer returns nonempty text for 8,859 of the 14,598 images. This is coverage, not a transcription-accuracy measure. Full-resolution or tiled OCR was not evaluated. For lexical comparisons, a figure is represented by its OCR text and a caption by its text after HTML unescaping, Unicode NFKC normalization, and whitespace normalization. OCR uses no task labels.

3.2 Symmetric pair features

The 190-dimensional representation consists of three modality indicators, 34 lexical features, 14 TF-IDF features, five serialized-object length features, 68 representation statistics, and 66 layout coordinates. Every feature is invariant to swapping the two objects. The five length features use original CSV strings, including base64 strings for images.

Lexical features measure exact or normalized matches, substring containment, and overlap between words, informative terms, numbers, acronyms, and alphanumeric identifiers. Jaccard similarity measures agreement across both texts; containment measures how much of the shorter text appears in the longer one. This distinction is useful when a short OCR string matches part of a caption. Inverse document frequency (IDF) gives more weight to uncommon shared terms. We retain at most 4,000 characters from each textual view.

Two TF-IDF spaces capture word unigrams and bigrams, and character trigrams and four-grams within word boundaries. Each vocabulary contains at most 40,000 entries; term frequencies are scaled sublinearly. Seven descriptors per space summarize cosine similarity, weighted Jaccard similarity, containment, sparsity, and text availability. Vocabularies and IDF weights are fitted on unique objects from the fitting partition.

Each of the two CLIP spaces, the SciNCL space, and the layout descriptor space contributes 17 summary statistics. These include cosine similarity, distances, quantiles of absolute differences, product statistics, sign agreement, and centered or variance-scaled cosine similarity. Means and standard deviations are estimated from fitting objects only, with standard deviations floored at 0.01. SciNCL comparisons apply to caption pairs. Missing representations are flagged explicitly. The layout block also contains the coordinate-wise absolute difference and mean of the two 33-dimensional descriptors.

3.3 Relation classifiers

We train a separate classifier for each pair modality. The primary system uses a factorized classifier with three binary decisions: whether the objects are the same figure, whether they come from the same paper otherwise, and whether different papers are citation-related. Let the corresponding conditional probabilities be q_f, q_p, and q_r. The

four class probabilities are

$$p_f = q_f, \qquad p_p = (1 - q_f)q_p, \tag{1}$$

$$p_r = (1 - q_f)(1 - q_p)q_r, \tag{2}$$

$$p_u = (1 - q_f)(1 - q_p)(1 - q_r). \tag{3}$$

For figure-figure and caption-caption pairs, we set $q_f = 0$. Each binary model uses histogram gradient boosting with at most seven leaves per tree.

We compare this factorization with flat multiclass models using seven or fifteen leaves. All models use 180 boosting iterations, a learning rate of 0.055, and a minimum of 35 examples per leaf. L2 regularization is 10 for the seven-leaf models and 15 for the fifteen-leaf model. We use neither early stopping nor class reweighting. The fixed system chooses the class with the largest predicted probability and applies no learned decision offsets.

4 Experimental Protocol

4.1 Grouping and model selection

Repeated objects can connect several pairs, so random row splits could expose a validation object during training. We instead construct a graph whose vertices are object hashes and whose edges are training pairs. Stratified group splitting assigns whole connected components to five outer folds. Each validation fold contains 2,000 pairs. Automated checks confirm zero exact-object overlap with the corresponding fitting partition.

All fixed classifiers use these same folds. To evaluate the additional selection procedure, we divide each outer training partition into three inner folds. Within every inner fit, we refit the text vocabularies, IDF weights, and representation statistics before training the candidate classifiers. Inner out-of-fold (OOF) macro-F1 over feasible classes selects one classifier per modality. We then fit four additive log-probability offsets to maximize overall macro-F1 on these inner predictions. Coordinate search considers offsets from -0.8 to 0.8 in steps of 0.05, for at most four rounds.

The chosen models and feature transformations are refitted on the outer training partition. Only then do they predict the held-out outer fold. Thus, outer labels do not determine either model selection or offsets. Pooled outer predictions give the nested development score. A 95% interval is estimated from 1,000 bootstrap resamples of exact-object components, keeping the predictions fixed. The outer-fold and estimator seed is 20260919; the inner-fold seed is 20260920.

4.2 Submissions and ablations

Both Kaggle submissions average class probabilities from five outer models. The fixed submission uses the factorized seven-leaf model throughout, without selection or offsets. The nested submission uses the models and offsets chosen within each outer training partition. It also averages probabilities across exact duplicate unordered scoring pairs before taking the highest-probability class; the fixed submission does not apply this step.

Feature ablations retain the fixed factorized model and the same outer folds. We remove either OCR-derived text or the TF-IDF and IDF-weighted lexical features, then refit the remaining pipeline. These comparisons assess feature sets without model selection or decision offsets. Removing OCR also changes the corpus used to estimate IDF, so this ablation measures the effect of the OCR pipeline as a whole.

Camera-ready comparisons use raw cosine similarities from L2-normalized CLIP and SciNCL vectors. SciNCL cosine applies only to caption-caption pairs and is zero otherwise; classifiers remain separate by modality. Logistic references use the same factorized decisions, fitting a standard scaler and L2-regularized logistic regression ($C = 1$, `liblinear`, at most 2,000 iterations) within each training fold. A matched-head reference replaces all 190 features with at most three cosines while retaining the seven-leaf boosted classifier and its settings.

These comparisons regenerate frozen representations for the 10,000 validated training pairs and reuse the saved outer folds and label-free OCR. CLIP ablations remove only the corresponding 17 summary features, keeping all other blocks and fitting procedures unchanged. The replayed full model is reported alongside these comparisons; its OOF score differs from the saved fixed run by 0.00022. Table 1 retains the original submission results. No comparison uses scoring data, model selection, or decision offsets.

5 Results and Discussion

We report scores and absolute differences to two significant digits. Differences are computed before rounding, so displayed scores may coincide despite a nonzero difference.

Additional selection gives no observed gain. The fixed model has a small numerical advantage: 0.0033 macro-F1 locally and 0.00071 on Kaggle,

System	Local OOF	Kaggle
Fixed factorized	0.48	0.48
Nested selection	0.48	0.48
Fixed minus nested	0.0033	0.00071

Table 1: Macro-F1 and absolute differences for the submitted systems. Kaggle values are the scores displayed on September 19, 2026, not final private-leaderboard results.

Classifier / features	Macro-F1	Drop
Flat, 7 leaves	0.47	0.013
Flat, 15 leaves	0.47	0.013
Factorized, 7 leaves	0.48	0
without OCR	0.47	0.0081
without TF-IDF / rarity	0.47	0.013

Table 2: Local macro-F1 on the same outer folds, without routing or decision offsets. Drop is the absolute decrease from the full factorized model, computed before rounding.

although the rounded scores in Table 1 coincide. These differences provide no evidence that the additional selection procedure improves aggregate performance. We do not claim a statistically established difference between the Kaggle scores. The nested local estimate has a conditional bootstrap interval of [0.47, 0.49]; that interval applies only to the nested predictions and must not be transferred to the fixed model.

Explicit textual evidence helps the fixed model. The factorized model outperforms both flat alternatives in local validation (Table 2). Removing OCR lowers its macro-F1 by 0.0081. Removing TF-IDF and rarity features lowers it by 0.013. These controlled comparisons support retaining both feature groups in this system. They do not establish that every individual feature is useful or that the same gains would hold for other classifiers. The OCR ablation compares cached low-resolution OCR with no OCR; it does not compare image resolutions or show that downsampling preserves all useful text.

Citation relations remain difficult. Related papers have the lowest class F1 in both systems (Table 3). The nested variant improves this class and same-figure detection, but lowers F1 for the other two classes. For the fixed model, feasible-class

Relation	Fixed	Nested
Same figure	0.51	0.52
Same paper	0.51	0.48
Related papers	0.37	0.39
Unrelated papers	0.53	0.52

Table 3: Per-class F1 from local OOF predictions. The small aggregate difference hides changes in which relations each system predicts well.

CLIP blocks	Features	F1	Drop
B/16 + B/32	190	0.48	0
B/32 only	173	0.47	0.0096
B/16 only	173	0.47	0.0076

Table 4: Controlled CLIP ablations of the replayed full model. The classifier, outer folds, and non-CLIP inputs are fixed. Drop is computed before rounding.

macro-F1 is 0.63 for caption-caption pairs, 0.47 for figure-figure pairs, and 0.35 for figure-caption pairs. These modality scores average over different feasible label sets, so they are neither directly comparable nor components of the overall four-class metric.

Both CLIP blocks contribute in the final classifier. Table 4 removes each CLIP block from the final boosted model. Excluding B/16 lowers macro-F1 by 0.0096; excluding B/32 lowers it by 0.0076. Each ablation retains 173 features, including SciNCL, OCR, lexical comparisons, and layout. The observed decreases support complementary predictive information from the two backbones in this development comparison, rather than complete redundancy. The gains are modest and are not a measured runtime or energy advantage.

Explicit pair features improve over raw cosine similarity. Table 5 reports simple cosine-only references. Combining B/16 and SciNCL with logistic heads gives 0.42 overall macro-F1; adding B/32 gives the same rounded score. SciNCL alone reaches 0.50 on caption-caption pairs, compared with 0.63 for the full features on that same three-class scope. Holding the boosted classifier fixed, replacing at most three cosines with the 190-feature representation improves overall macro-F1 from 0.41 to 0.48, an absolute gain of 0.068. This matched-head comparison isolates the combined feature design from a change of classifier within the stated development protocol.

Inputs / head	All	CC
B/16 cosine / LR	0.40	0.44
B/32 cosine / LR	0.39	0.43
SciNCL cosine / LR	-	0.50
B/16 + SciNCL / LR	0.42	0.52
Dual CLIP + SciNCL / LR	0.42	0.52
Dual CLIP + SciNCL / GB	0.41	0.50
190 features / GB	0.48	0.63

Table 5: Cosine-only references and the replayed full model. LR: logistic heads; GB: fixed seven-leaf boosted heads. All averages four labels over 10,000 pairs; CC averages three feasible labels over 3,000 caption-caption pairs. All inputs except the last row are raw cosines.

6 Conclusion

GRAIL combines frozen semantic representations with explicit textual and visual comparisons in a compact classifier for AstroCLIMB. The fixed factorized system reaches 0.48 macro-F1 on Kaggle. A matched-head cosine reference and controlled ablations support the richer pair features, both CLIP blocks, OCR, and corpus-weighted lexical evidence. Additional model selection yields no observed aggregate gain. The remaining difficulty is linking content similarity to a direct citation relation. Future evaluation should test whether these findings persist when source papers, rather than exact objects, are separated across folds.

References

Egor Fedosov. 2026. GRAIL for AstroCLIMB: Source code and experiment reports. GitHub repository. https://github.com/denzasikora-lab/astroclimb-grail.

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.

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*, page 11670-11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*, volume 139 of *Proceedings of Machine Learning Research*, page 8748-8763. PMLR.