

DiagramBank: A Quality-Audited Dataset of Scientific Schematic Diagrams with Multi-Level Document Context

Ling Yue¹ Tingwen Zhang¹ Jiaying Wang¹
Zhen Xu² Shaowu Pan^{1,*}

¹Rensselaer Polytechnic Institute

²University of Chicago

*Corresponding author: pans2@rpi.edu

Abstract

Scientific papers use schematic diagrams to communicate methods, workflows, and system structure, yet existing scientific-figure corpora often mix them with plots, screenshots, and photographs and rarely preserve document context. We introduce DiagramBank, a quality-audited dataset of 57,100 schematic diagrams curated from OpenReview-hosted AI/ML venues. Each record links a diagram image to its paper title, abstract, figure caption, in-text figure-reference spans, venue/year metadata, provenance fields, and filtering labels. DiagramBank is a reusable resource for scientific-document understanding, diagram retrieval, corpus analysis, and future benchmark construction. We describe its extraction and cascade-filtering pipeline, release schema, confidence-controlled views, dataset card, and indexing utilities. A manual blind audit of the released cascade-filtered records estimates 93.67% precision, and a separate CLIP threshold analysis characterizes the precision-coverage trade-off for simpler filtering views. We further provide lightweight metadata-indexing and authoring examples to illustrate downstream protocols without treating these utilities as standalone methods. The code is public at: <https://github.com/csml-rpi/DiagramBank>.

1 Introduction

Scientific papers often rely on schematic diagrams to convey objects, modules, and information flow before readers can understand the surrounding technical details. Such figures act as graphical abstracts in practice (Yang et al., 2019), but they are not ordinary plots or natural images: their meaning depends on compositional layout, arrows, labels, grouping, and community-specific notation. This structure-rich form motivates diagram understanding as a distinct problem (Kembhavi et al., 2016), and it is increasingly relevant as scientific-authoring systems begin to automate manuscripts

and figures (Lu et al., 2024; Yue et al., 2026; Zhu et al., 2026a,b).

For studying and building with such diagrams at scale, data is a central bottleneck. Existing scientific-figure resources provide valuable supervision for captioning, figure classification, and summarization, but they usually mix schematic diagrams with plots, photos, tables, or screenshots, and they often omit the body-text passages that explain how a figure is used in the paper. For resource construction, this distinction matters: a dataset for scientific diagrams should identify schematic figures, preserve paper-level and figure-level context, expose filtering uncertainty, and keep provenance available for audit and attribution.

We therefore frame DiagramBank as a dataset contribution, not as a new RAG or diagram-generation method. DiagramBank contains 57,100 cascade-filtered scientific schematic diagrams from OpenReview-hosted AI/ML venues, with each record coupling an image to title, abstract, caption, in-text figure-reference spans, venue/year metadata, quality-control labels, and attribution fields. The retrieval and authoring examples in Section 5 show how the released records can be used; they are deliberately framed as reference utilities rather than primary methodological claims. Our contributions are:

- **Diagram-specific corpus.** We release 57,100 scientific schematics with paper/figure context, provenance, venue/year metadata, and confidence views.
- **Measured quality.** We audit both filtering stages: 94.4% precision at $\tau = 0.85$ for the high-confidence CLIP view and 93.67% precision for the final release.
- **Reusable artifacts.** We provide records, labels, confidence views, a dataset card, indexing code, utility-check files, and reporting guidance for explicit subset and filter choices.

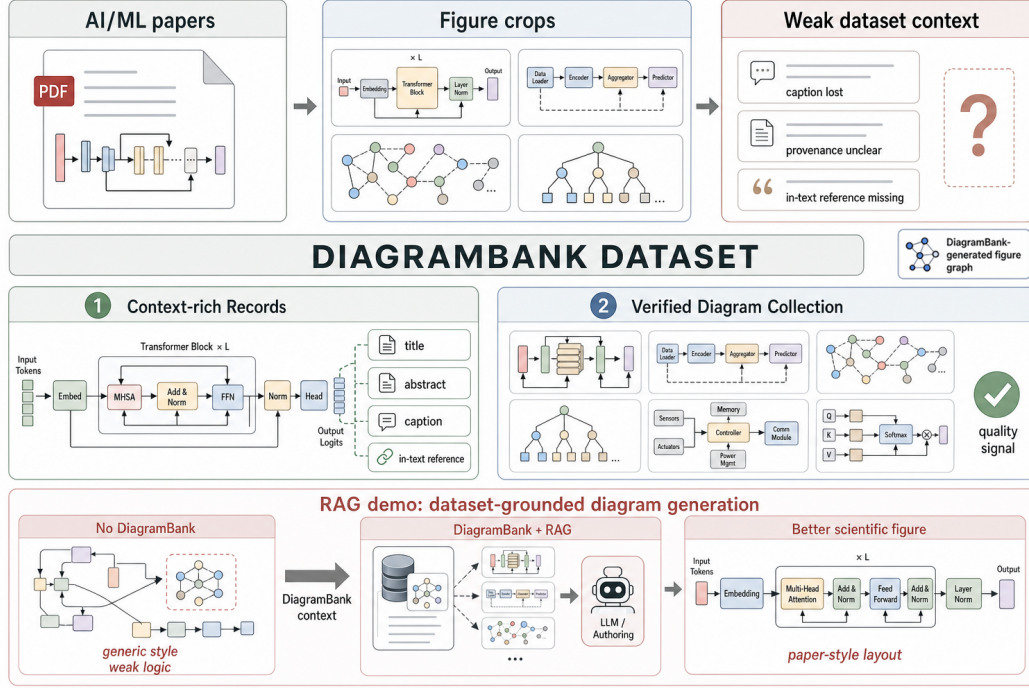


Figure 1: DiagramBank dataset and resource overview. This illustrative schematic contrasts weak figure-only extraction with context-rich records, a verified diagram collection, and dataset-grounded diagram-generation use.

2 Related Work and Resource Positioning

Scientific-figure datasets provide important starting points, but they are not organized as diagram-specific resources with multi-level document context. SciCap (Hsu et al., 2021) supports caption generation over heterogeneous scientific figures; DocFigure (Jobin et al., 2019) focuses on figure-type classification; ACL-Fig (Karishma et al., 2023) covers ACL Anthology figures and tables; and visual-summary work ranks representative figures from paper content (Yamamoto et al., 2021; Zhong et al., 2025). These resources show the value of figure-level supervision, but they either mix plots, tables, photos, and diagrams, omit the OpenReview-centered AI/ML venues studied here, or do not preserve the body-text context that explains how a schematic is used in the paper. Diagram-centric AI2D (Kembhavi et al., 2016) provides dense annotations, but its grade-school science domain differs from the conventions of modern AI/ML diagrams.

Recent automated authoring and diagram-generation systems, including AI Scientist (Lu et al., 2024), autonomous paper-generation work (Yue et al., 2026), Paper-Banana (Zhu et al., 2026a), AutoFigure (Zhu et al., 2026b), and diagram planners such as DiagrammerGPT and Sci-Doc2Diagrammer (Zala et al., 2023; Mondal et al., 2024), increase the need for auditable scientific-

figure resources. Grounded authoring and retrieval-augmented systems also depend on external evidence (Lewis et al., 2020), but applying retrieval to diagrams requires a corpus where the retrieval unit is a diagram, the textual evidence is preserved at multiple granularities, and filtering quality is measured rather than assumed. DiagramBank is positioned in this gap: it is neither a benchmark leaderboard nor a new retrieval algorithm, but it supplies the auditable records needed to build such evaluations. To our knowledge, no public, large-scale dataset simultaneously offers quality-audited scientific schematic diagrams, paper metadata for domain-aware retrieval, and figure-local context spans that connect a diagram to the surrounding scientific narrative.

3 Dataset Construction and Quality Assurance

The design of DiagramBank follows a four-stage pipeline: metadata acquisition, content extraction, schematic classification, and relational aggregation. The non-trivial parts of the pipeline are cross-venue schema normalization, noisy figure/caption extraction, figure-reference alignment in body text, and high-precision separation of schematic diagrams from plots, screenshots, photos, and other visual material.

Resource	Domain and scale	Schematic focus	Document context	Quality / retrieval support
SciCap	Scientific figures; 416,000 figures	Mixed scientific figures for caption generation	Partial paper metadata and captions; no body-reference context	Captioning supervision; not diagram-filtered
DocFigure	Scientific documents; 33,000 figures	Figure-type classification over document figures	Caption information only; no paper-level context	Type labels; not retrieval-ready
ACL-Fig	Computational-linguistics papers; 112,000 figures from 56,000 papers	Figures and tables from ACL papers	Paper metadata, captions, and limited body context	Classification-oriented; limited diagram-specific QA
AI2D	School science; 5,000 diagrams	Dense labels for educational diagrams	No scientific-paper metadata or caption context	Diagram annotations outside scholarly-paper domain
SMSMO	Scientific summaries; 24,000 papers	Representative figure selection for summaries	Paper metadata and captions; no extracted figure-reference spans	Summary evaluation rather than diagram retrieval
Generation systems	General or scientific diagrams; system dependent	Produced or planned diagrams	System-dependent context	Evaluate generation behavior, but do not release a diagram corpus
DiagramBank (ours)	AI/ML papers; 57,100 schematics	Scientific schematic diagrams filtered from mixed figures	Title, abstract, caption, in-text figure context, provenance, and quality labels	Manual blind audit, confidence views, cascade paths, dataset card, and indexing utilities

Table 1: Comparison with related figure and diagram resources. DiagramBank is differentiated by diagram-specific filtering, multi-level document context, measured release precision, and retrieval-ready metadata rather than by a new downstream method.

Data acquisition and schema normalization.

We source PDFs and metadata from OpenReview for ICLR 2017–2026, NeurIPS 2021–2022 and 2024–2025, ICML 2023–2025, and TMLR 2022–2026 using the OpenReview Python API (OpenReview Team, 2025). Because decisions, scores, subject areas, and TL;DR fields differ across venue-year schemas, we normalize paper-level attributes into a *Papers* relation containing bibliographic details, semantic metadata, platform IDs/URLs, decision metadata when available, and BibTeX entries for attribution.

Figure and context extraction. We use PDFFigures 2.0 (Clark and Divvala, 2016) to extract figures and captions while discarding table artifacts. To capture how authors explain each figure, we parse the PDF text layer with PyMuPDF (Artifex Software, Inc., 2024) and extract paragraphs that explicitly cite the figure number. The resulting *figure_context* field links each image to the paper’s argumentative flow and supplies a retrieval signal that caption-only corpora lack. Because PDF text order and figure references are noisy, the release keeps extracted context as evidence rather than treating it as a gold annotation; downstream users can filter records with missing or short context spans when a task requires complete local grounding.

Schematic classification. A core challenge is distinguishing explanatory diagrams from plots,

screenshots, photos, and miscellaneous visual material. We first run an ensemble-prompt CLIP pre-filter using OpenCLIP ViT-B/32 with DataComp-XL weights (Ilharco et al., 2021; Radford et al., 2021), assigning each extracted figure to *diagram*, *plot*, *photo*, or *other*. CLIP confidence is the softmax probability derived from image–prompt cosine similarity. The primary release then applies a vision-language-model (VLM) cascade (Table 2): CLIP-labeled diagrams enter a majority-vote path, low-confidence non-diagram candidates enter a rescue path, and a larger VLM resolves specified disagreements. We therefore use *cascade-filtered* to denote model-path acceptance and reserve precision claims for the held-out audit.

Data aggregation and filtering. We manage relational data with DuckDB (Raasveldt and Mühleisen, 2019) and join the *Papers* and *Figures* relations on *platform_id*. The final schema is deliberately denormalized so every retrieved exemplar carries its own title, abstract, caption, context spans, quality labels, and attribution fields without requiring runtime joins.

Following aggregation, we expose three views: all CLIP-labeled diagrams, high-confidence CLIP diagrams with `clip_confidence` ≥ 0.85 , and the primary cascade-filtered accepted-paper release. Rejected-paper records remain in the all-paper CLIP views so users can study venue distributions and construct their own filters.

Cascade stage	Candidate figures	Kept diagrams	Decision rule
Tier 1: CLIP predicts diagram	60,717	52,034	Two independent VLM judges classify each candidate; 46,524 unanimous and 3,645 majority cases are kept, and a larger VLM confirms 1,865 of 10,548 minority cases.
Tier 2: non-diagram with CLIP confidence < 0.70	33,459	5,066	Both smaller judges must independently relabel the figure as diagram; the larger VLM then confirms 5,066 of 5,172 consensus candidates.
Final cascade release	94,176	57,100	Accepted-paper candidate pool; final venue counts are ICLR 20,516, ICML 11,267, NeurIPS 19,655, and TMLR 5,662.

Table 2: Cascade funnel for the primary DiagramBank release. CLIP is the low-cost routing stage; the kept records are classified by independent VLM judgments and, where needed, confirmation by a larger VLM. Counts summarize the final cascade logs; exact model identifiers are reported in the release documentation.

Filter quality audits. We perform two complementary manual audits and use them for different claims. In both cases, annotators apply a binary schematic-versus-non-schematic rubric without seeing the model path or confidence score; the release documentation includes the rubric, sampling strata, and audited labels needed to reproduce the estimates. First, a 500-figure stratified sample validates the CLIP pre-filter used to expose simple confidence-controlled views. At the $\tau = 0.85$ operating point, CLIP reaches 94.4% precision with a 0.5% false-positive rate in the sampled non-diagram stratum; a threshold sweep shows that 0.85 gives a large precision gain over 0.80 with less retention loss than 0.90. In that sweep, precision increases from 64.0% at $\tau = 0.50$ to 77.3% at 0.70, 86.0% at 0.80, 94.4% at 0.85, and 96.0% at 0.90, while retention decreases from 100% to 55.7%.

Second, because the primary release is not identical to the high-confidence CLIP view, we separately audit the final cascade-filtered artifact. We draw 500 kept records stratified by cascade path and venue and estimate release precision with stratum weights rather than a simple unweighted average. This audit estimates 93.67% precision (95% CI 90.11–97.22) for the 57,100-record release. The dominant error mode is inclusion of screenshots or photo-like figures, which motivates exposing cascade paths and filtering labels so users can tighten subsets for precision-sensitive tasks.

4 Dataset Statistics and Analysis

Appendix Figure 4 summarizes macro-level characteristics across the 2017–2025 statistical analysis slice and venues. In this section, we provide quantitative breakdowns of (i) the cascade-filtered release and confidence-controlled intermediate subsets, (ii) the distribution of figure types over the

full extracted corpus, and (iii) venue-level visual density statistics that help contextualize retrieval and downstream diagram authoring.

4.1 Scale and Subset Sizes

DiagramBank is derived from a large pool of figures extracted from OpenReview PDFs. The initial classified pool contains 475,616 figures across the release coverage; the VLM cascade operates on the 94,176 accepted-paper candidates routed by the CLIP stage (Table 2). For the 2017–2025 non-table analysis slice summarized in the compact corpus-statistics graphic, we extract 452,339 non-table figures in total (plots, diagrams, photos, and others), of which 89,422 (19.8%) are classified as diagrams by the first-stage CLIP type classifier. Because automatic filtering inevitably introduces a precision–recall trade-off, we expose these intermediate CLIP scores while making the 57,100 cascade-filtered diagrams the primary release.

To support common usage patterns, we report two CLIP-based views: (1) an all-confidence view that includes all figures labeled as diagram by CLIP (no confidence threshold), and (2) a high-confidence view obtained by thresholding `clip_confidence` (here shown with ≥ 0.85), which improves precision for users who prefer a simple, cascade-free threshold. In addition, when paper decision metadata is available, we report accepted-paper counts (e.g., Accept/Poster/Spotlight/Oral). These counts are filters over the relevant view, not separate collection pipelines.

The primary cascade-filtered release contains 57,100 diagrams: 20,516 from ICLR, 11,267 from ICML, 19,655 from NeurIPS, and 5,662 from TMLR. This release is slightly smaller than the simple high-confidence CLIP view but is directly validated on its final kept paths and includes re-

Venue	All CLIP diagrams		High-confidence CLIP diagrams	
	All papers	Accepted papers	All papers	Accepted papers
ICLR	46,460	19,261	31,325	12,550
ICML	12,267	12,204	8,045	8,005
NeurIPS	21,569	20,582	14,226	13,533
TMLR	9,126	5,761	6,169	3,849
Total	89,422	57,808	59,765	37,937

Table 3: Intermediate CLIP-based subset sizes for the 2017–2025 analysis slice. Accepted-paper columns are filters over each CLIP view; Table 2 separately reports the vision-language-model cascade pool.

covered diagrams that CLIP initially labeled as non-diagrams. Figure 4(b) should be read as both a volume curve and a venue-coverage curve: early years are dominated by ICLR, with later increases partly reflecting broader OpenReview adoption and venue availability rather than publication growth alone. These views are related but distinct cuts: the CLIP analysis slice reports first-stage labels, while the primary release reports accepted-paper cascade outcomes from the documented routing pool.

4.2 Figure Type Distribution

Figure 2 reports the aggregate composition of the full extracted non-table figure corpus. Across all venues, plots dominate the visual content (65.2%), while diagrams account for 19.8% and photos for 11.5%. The aggregate diagram share indicates that schematic figures are a substantial and recurring component of modern AI/ML papers.

We also report the mean confidence score per predicted type. Plots tend to receive the highest confidence on average, while diagrams have lower (but still strong) mean confidence, reflecting the broader stylistic diversity of schematic figures (e.g., flowcharts, architectures, conceptual maps) compared to canonical chart types. This observation motivates exposing `clip_confidence` in the release so that users can tighten or relax filtering depending on their tolerance for noise.

4.3 Visual Density and Caption Characteristics

Figure 2(b) characterizes venue-level visual density using three indicators: figures per paper, visual breadth, and average caption length. Two patterns are particularly relevant for retrieval and diagram authoring.

First, visual breadth is consistently high (79.8–97.6%), meaning that most papers contain at least one extractable figure. This is important for re-

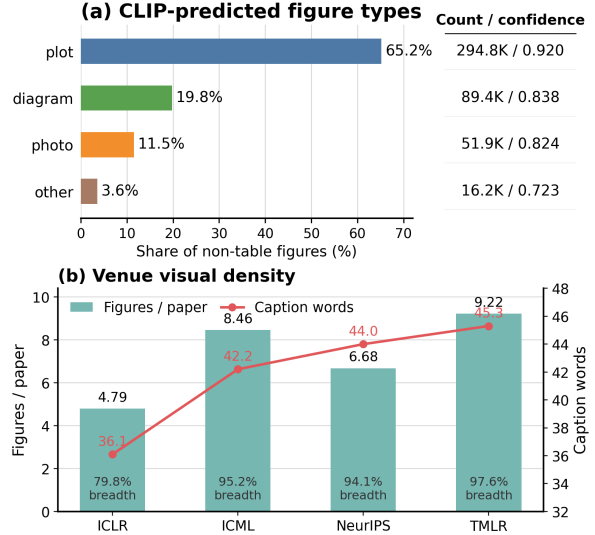


Figure 2: Compact corpus statistics. (a) CLIP-predicted figure-type distribution over 452,339 extracted non-table figures, annotated with counts and mean confidence scores. (b) Venue-level visual density: bars show figures per paper, the line shows average caption length, and bar labels show breadth, defined as the percentage of papers with at least one extracted non-table figure.

trieval because it ensures dense coverage of the literature and reduces the number of “empty” papers during indexing.

Second, there are clear venue-level differences in figure density and caption verbosity. ICLR exhibits the lowest figure density (4.79 figures/paper) and shortest captions on average, while TMLR has the highest figure density (9.22 figures/paper) and the longest captions in this slice. These differences imply that retrieval and analysis protocols must handle both sparse and visually dense papers; in particular, high figure density increases the number of within-paper candidates, motivating the decision to preserve separate paper-level and figure-level context fields.

Finally, Appendix Figure 4(a) complements these venue-level statistics by visually suggesting a mild trend toward shorter captions over time, while Figure 4(c) highlights that keyword/subject-area distributions vary across venues (e.g., venue-specific taxonomies), motivating normalization and multi-stage filtering for domain-aligned retrieval.

4.4 Release and Documentation

Release structure and schema. DiagramBank is packaged as denormalized metadata records, cascade-filtered release labels, optional confidence-controlled CLIP views, extracted diagram assets where redistribution is permitted, source-document

View / denominator	Scope	Count	Use in paper
Classified figure pool	Release coverage	475,616	Initial figure pool used for CLIP routing and cascade accounting across venue-specific OpenReview coverage.
Non-table analysis slice	2017–2025	452,339	Denominator for aggregate figure-type statistics shown in the compact corpus-statistics graphic.
All-CLIP diagram view	2017–2025	89,422	Intermediate view containing every figure whose first-stage CLIP label is diagram.
High-confidence CLIP view	2017–2025	59,765	Intermediate view applying <code>clip_confidence</code> ≥ 0.85 .
Accepted-paper cascade candidates	Release coverage	94,176	Candidate pool routed through the VLM cascade (60,717 Tier 1 + 33,459 Tier 2).
Primary DiagramBank release	Release coverage	57,100	Cascade-filtered schematic diagrams distributed as the main resource.

Table 4: Denominators and release views. The release coverage is venue-specific (ICLR 2017–2026, NeurIPS 2021–2022 and 2024–2025, ICML 2023–2025, and TMLR 2022–2026), while several descriptive statistics use a 2017–2025 non-table analysis slice.

links/BibTeX for attribution, example indexing scripts, and a dataset card. During review, access is provided through an anonymized supplementary archive or anonymized repository rather than through author-identifying project pages. The downstream fields fall into three groups: **paper context** (`platform_id`, venue/year, title, abstract, authors, keywords, subject areas, TL;DR, decision metadata, URL, and BibTeX), **figure context** (`figure_id`, asset path, caption, and body paragraphs that explicitly cite the figure), and **quality fields** (CLIP label/confidence, cascade decision/path, accepted-paper filter, and redistribution/attribution metadata). Keeping these groups in each denormalized record is intentional: a retrieved diagram should carry enough surrounding evidence to be inspected, cited, filtered, or excluded without requiring a user to reconstruct joins across paper tables, figure tables, and audit logs. The full field list appears in Appendix Tables 6 and 7.

Documentation and usability. The review package includes a dataset card/datasheet, release manifest, schema description, loading and indexing scripts, confidence-controlled metadata tables, and either extracted diagram assets or source links according to redistribution terms. The dataset card complements the construction and schema description with intended-use, split/evaluation, access, licensing, and attribution guidance summarized in Appendix D. We include these artifacts to make the resource usable for retrieval studies, scientific-figure analysis, and exemplar-conditioned authoring examples.

4.5 Resource Uses and Suggested Protocols

DiagramBank is not itself a fixed benchmark: it does not prescribe a single train/test split, metric, or leaderboard. Instead, its record design supports future benchmark and analysis protocols that require both visual and document-level evidence. Because each instance connects a diagram image to paper-level intent, figure-local semantics, body-reference context, venue metadata, quality labels, and source attribution, users can define stricter or broader evaluation subsets without changing corpus identity. The main use patterns are compact enough to describe inline:

- **Tasks:** images, captions, and figure-reference spans support tasks that ask whether a model can recover modules, relations, and the role of a schematic in the paper narrative. The body-reference spans are especially useful for tasks that go beyond object naming, because they expose how authors explain the figure’s contribution in prose.
- **Retrieval and filtering:** title, abstract, caption, context fields, first-stage labels, confidence scores, cascade paths, and audit estimates support paper-level, figure-level, coarse-to-fine, precision-oriented, and coverage-oriented protocols. A study can therefore compare broad candidate discovery against stricter high-confidence subsets without changing the identity of the underlying corpus.
- **Discovery and authoring:** venue/year metadata, subject areas, source links, citation entries, and retrieved exemplars provide provenance-aware comparable papers plus concrete layout, notation, grouping, and terminology references for human figure revision. This is useful both for

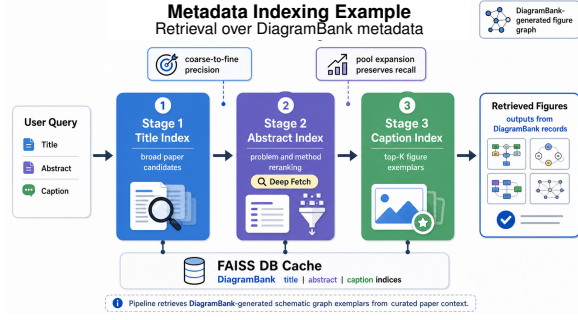


Figure 3: Example metadata-indexing use. DiagramBank metadata supports title retrieval, abstract reranking, and caption-level exemplar retrieval.

literature discovery and for authoring workflows where the desired output is a grounded set of neighboring examples rather than an automatically generated final figure.

The record structure turns these uses into auditable protocols: a diagram-understanding benchmark can hold out venues or years, a retrieval study can compare title-only, caption-only, and hierarchical indexing, and a literature-discovery workflow can require retrieved diagrams to carry BibTeX/source provenance. This benchmark-facing organization is why DiagramBank preserves intermediate confidence views and cascade paths alongside the primary release.

5 Example Metadata Uses

The following examples show what the released metadata makes possible; they are not presented as a new retrieval or RAG method. Given a user’s paper title/abstract and an intended figure caption (or detailed description), a reference pipeline can retrieve schematic exemplars that are inspected for layout, grouping, iconography, and terminology conventions. DiagramBank exposes enough metadata to support multiple retrieval strategies; hierarchical retrieval is one reproducible instantiation. Caption-only retrieval remains useful for some queries, while paper-level filtering helps when generic figure captions cause domain drift. The reported numbers below are a metadata-use sanity check, not a benchmark result: they verify that the released fields can support reproducible retrieval protocols and that paper-level and caption-level signals behave differently.

Reference index construction. Each diagram is paired with multi-level text: paper title, paper abstract, and figure caption. The reference implementation keeps three separate indices rather than mixing all text into one embedding: a title index

Table 5: Automatic retrieval sanity check. DiagramBank’s multi-level metadata supports complementary paper-level and caption-level search strategies.

Evaluation item	Value
Queries	20 scientific-diagram queries
Pool	Top-50 from both retrieval strategies
Judged pairs	1,862 unique query–diagram pairs
Relevance judges	Three VLM judges; exact model identifiers are in the release documentation
Rubric	0 = not relevant; 1 = marginal; 2 = high
Aggregation	Majority vote; Fleiss’ $\kappa = 0.425$; 52% unanimous
Precision@10	Model-judged: hierarchical 78.0%; caption-only 81.0%
Per-query wins	Hierarchical 7, caption-only 9, ties 4

for coarse topical/domain filtering, an abstract index for problem-setting and methodology matching, and a caption index for figure-level semantic matching over modules, components, and relations. We compute OpenAI text embeddings (OpenAI, 2024) for each title, abstract, and caption, store the vectors in FAISS for fast ANN search, and load the three indices once for reuse across all retrieval stages (Figure 3).

Example hierarchical retrieval. Let the query be $q = (q_{\text{title}}, q_{\text{abs}}, q_{\text{cap}})$. Retrieval proceeds from broad to specific. Stage 1 queries the Title Index and keeps $S_1 = \text{TopK}_{\text{title}}(q_{\text{title}}, k = t_1)$; in our released configuration, $t_1 = 100$. Stage 2 queries the Abstract Index within S_1 and keeps $S_2 = \text{TopK}_{\text{abs}}(q_{\text{abs}} \mid \text{paper} \in S_1, k = t_2)$ with $t_2 = 30$. Because ANN systems may lose recall under post-filtering, we use Deep Fetch: retrieve a larger pool from the full index, then filter by S_1 and keep the best t_2 ; our default multiplier is $\alpha = 100$. Stage 3 queries the Caption Index within S_2 and returns $\text{TopK}_{\text{cap}}(q_{\text{cap}} \mid \text{paper} \in S_2, k = K)$ with $K = 3$, using $\beta = 1000$ because each paper may contain multiple figures.

Table 5 summarizes the available sanity check. The key finding is complementarity rather than dominance: hierarchical retrieval helps when paper-level context disambiguates generic captions, while caption-only retrieval can find relevant diagrams from topically different papers.

Additional qualitative checks.

- **Code2MCP:** adjacent code-agent papers and source-level provenance for human revision, including neighboring work to inspect before choosing components and labels.
- **EagleFT:** speculative-decoding conventions such as frozen/trainable iconography and forecast-token notation.

- **LatentDreamer:** Dreamer-family terminology including recurrent state-space models plus self-distillation and masked-autoencoder names.
- **ProtBERT-3D:** protein-language-model conventions with Enzyme Commission labels plus query, key, and value cross-attention notation.

Release protocols and reporting guidance. The release is organized so reviewers and downstream users can inspect the resource without relying on author-side services. During review, the anonymous package contains metadata records, confidence-controlled views, cascade-path labels, retrieval/indexing code, a release manifest, a dataset card, and either redistributable diagram assets or source links with attribution metadata. The utility-check files include query prompts, pooled candidates, per-judge rubric labels, majority-vote labels, and scripts for recomputing Precision@10 and per-query wins. This structure separates three concerns that are often conflated in figure resources: what was extracted, why it was accepted into a particular view, and how a user should cite or filter it.

Reviewer-facing artifacts.

- **Records:** denormalized paper, figure, and context fields.
- **Views and labels:** all-confidence and high-confidence CLIP subsets plus kept/rejected cascade paths and decision metadata.
- **Audit materials:** sampling strata, rubric labels, held-out precision estimates, and error modes.
- **Indexing and release files:** title, abstract, and caption retrieval indices, together with the manifest and dataset card for scope, access, licensing, attribution, and reporting.

Each downstream evaluation should name the subset, confidence threshold, and cascade paths used in the experiment. Temporal splits can test robustness to changing visual conventions, venue holdouts can test domain transfer across publication cultures, and confidence-controlled evaluations can separate high-precision studies from broad-coverage analyses. These protocols are intentionally lightweight: DiagramBank is a reusable resource, so downstream tasks should be able to choose stricter or broader views without changing the released record format.

Audit interpretation. We use the term *cascade-filtered* for automatic model decisions and use *audit* only for held-out manual measurement on sampled

records. This distinction is important for interpreting the reported precision numbers: the cascade decides which records enter the primary release, while the audit estimates how often those kept records are actually schematic diagrams under the annotation rubric. The CLIP threshold $\tau = 0.85$ is a precision-oriented operating point for the high-confidence CLIP view: in the stratified audit sample it yields 94.4% precision (34/36 high-confidence diagram predictions) and a 0.49% false-positive rate (2/410 sampled negatives), while increasing the threshold to 0.90 gives only a small precision gain at a larger retention cost.

The final 57,100-record release combines CLIP-routed candidates, VLM majority decisions, a low-confidence rescue path, and larger-VLM confirmation for specified disagreements, so it is evaluated separately from the high-confidence CLIP view. The held-out cascade audit estimates 93.67% precision with a 95% confidence interval of 90.11–97.22. We also expose the weakest kept path and observed error modes so users can exclude higher-risk paths or inspect screenshot/photo-like records before building task-specific benchmarks. The dataset card further recommends task-specific checks for missing captions, very short captions, incomplete context spans, and extremely small images when downstream evaluations depend on fine visual details. Downstream reports should therefore name the release view, cascade paths, and any screenshot/photo filters, making quality control an explicit experimental variable alongside model and split settings.

6 Conclusion

This paper introduced DIAGRAMBANK, a quality-audited dataset of 57,100 scientific schematic diagrams with multi-level document context. Each record connects a diagram image to title, abstract, caption, in-text figure-reference spans, venue/year metadata, provenance, confidence scores, and cascade labels. The CLIP and cascade audits quantify the precision trade-offs of automatic diagram collection, while the release exposes confidence-controlled views so users can choose stricter or broader subsets. Together, these artifacts support scientific-document understanding, diagram retrieval, and controlled resource analysis.

Limitations

DIAGRAMBANK relies on automated PDF parsing, figure classification, and context extraction, so some records may contain imperfect captions, incomplete figure-reference spans, or diagram-plot classification errors. The primary release is precision-oriented and may exclude useful lower-confidence diagrams; users who need higher coverage should inspect the intermediate CLIP views and apply task-specific filtering. Coverage is limited to accessible OpenReview venues and years, which may bias the resource toward recent AI/ML publication norms. The retrieval utilities are compact metadata-use sanity checks, and the authoring example remains qualitative; downstream claims should be evaluated with task-specific human studies or benchmark protocols. Human verification and post-editing remain necessary before using generated or exemplar-conditioned figures in real publications.

Ethics Statement

DiagramBank is derived from public scholarly documents and can be used in authoring pipelines, so provenance, redistribution, privacy, bias, misuse, and AI-use disclosure matter. Figures originate from third-party papers and may be subject to copyright or venue-specific license terms; the release preserves source identifiers and BibTeX entries, distributes extracted images only where permitted, and otherwise provides metadata/source links or license-filtered subsets. The source-document collection does not intentionally target non-public personal data, but publication metadata and figure contents may still contain identifying information from source papers. The corpus reflects OpenReview-era AI/ML publication practices and should not be treated as demographically or scientifically representative. Generated figures using retrieved exemplars can be misleading if used without disclosure or checking; users should track provenance, cite both DiagramBank and the original source papers when specific diagrams inform downstream work, disclose AI assistance where appropriate, manually verify factual and visual details, and honor reasonable rights-holder takedown requests (Klug and Pietsch, 2024). We used AI-based writing assistance for grammar and clarity. Figures 1, 3, and 5 contain AI-generated illustrative or example-use graphics; exact model names, prompts, and post-editing steps are documented in the anonymous

supplement. Generative AI did not create source records; learned models are used only in the documented classification and example-use pipeline.

References

- Artifex Software, Inc. 2024. [Pymupdf: Python bindings for the mupdf library](#).
- Christopher Clark and Santosh Divvala. 2016. [Pdffigures 2.0: Mining figures from research papers](#). In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 143–152.
- 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.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. [Openclip](#). If you use this software, please cite it as below.
- KV Jobin, Ajoy Mondal, and CV Jawahar. 2019. [Docfigure: A dataset for scientific document figure classification](#). In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE.
- Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C Lee Giles. 2023. [Aclfig: A dataset for scientific figure classification](#). *arXiv preprint arXiv:2301.12293*.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. [A diagram is worth a dozen images](#). In *Euro-pean conference on computer vision*, pages 235–251. Springer.
- Julian Klug and Urs Pietsch. 2024. [Can artificial intelligence help for scientific illustration? details matter](#). *Critical Care*, 28(1):196.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Advances in neural information processing systems*, 33:9459–9474.
- Honglin Lin, Qizhi Pei, Xin Gao, Zhuoshi Pan, Yu Li, Juntao Li, Conghui He, and Lijun Wu. 2025. [Scaling code-assisted chain-of-thoughts and instructions for model reasoning](#). *Preprint*, arXiv:2510.04081.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. [The ai scientist: Towards fully automated open-ended scientific discovery](#). *arXiv preprint arXiv:2408.06292*.

- Dung Nguyen Manh, Thang Phan Chau, Nam Le Hai, Thong T. Doan, Nam V. Nguyen, Quang Pham, and Nghi D. Q. Bui. 2025. [CodeMMLU: A multi-task benchmark for assessing code understanding & reasoning capabilities of CodeLLMs](#). *Preprint*, arXiv:2410.01999.
- Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Lee Boyd-Graber. 2024. [Scidoc2diagrammer-maf: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13342–13375.
- OpenAI. 2024. [Openai text embeddings api](#). Model: text-embedding-3-large.
- OpenReview Team. 2025. [Openreview python client](#). Python library for the OpenReview API.
- Chaoqian Ouyang, Ling Yue, Shimin Di, Libin Zheng, Linan Yue, Shaowu Pan, Jian Yin, and Min-Ling Zhang. 2025. [Code2mcp: Transforming code repositories into mcp services](#). *arXiv preprint arXiv:2509.05941*.
- Mark Raasveldt and Hannes Mühleisen. 2019. [Duckdb: an embeddable analytical database](#). In *Proceedings of the 2019 international conference on management of data*, pages 1981–1984.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. [Learning transferable visual models from natural language supervision](#). In *International conference on machine learning*, pages 8748–8763. PmLR.
- Hongyuan Tao, Ying Zhang, Zhenhao Tang, Hongen Peng, Xukun Zhu, Bingchang Liu, Yingguang Yang, Ziyin Zhang, Zhaogui Xu, Haipeng Zhang, Linchao Zhu, Rui Wang, Hang Yu, Jianguo Li, and Peng Di. 2025. [Code graph model \(CGM\): A graph-integrated large language model for repository-level software engineering tasks](#). *Preprint*, arXiv:2505.16901.
- Shintaro Yamamoto, Anne Lauscher, Simone Paolo Ponzetto, Goran Glavaš, and Shigeo Morishima. 2021. [Visual summary identification from scientific publications via self-supervised learning](#). *Frontiers in Research Metrics and Analytics*, 6:719004.
- Sean T Yang, Po-Shen Lee, Lia Kazakova, Abhishek Joshi, Bum Mook Oh, Jevin D West, and Bill Howe. 2019. [Identifying the central figure of a scientific paper](#). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1063–1070. IEEE.
- Ling Yue, Ching-Yun Ko, Pin-Yu Chen, Shimin Di, and Shaowu Pan. 2026. [Building MCP-native hierarchical AI scientist ecosystems: A perspective on scaling multi-agent scientific discovery](#). *Frontiers in Artificial Intelligence*, 9:1820375.
- Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. 2023. [Diagrammergpt: Generating open-domain, open-platform diagrams via llm planning](#). *arXiv preprint arXiv:2310.12128*.
- Xinyi Zhong, Zusheng Tan, Shen Gao, Jing Li, Jiaxing Shen, Jingyu Ji, Jeff Tang, and Billy Chiu. 2025. [Smsmo: Learning to generate multimodal summary for scientific papers](#). *Knowledge-Based Systems*, 310:112908.
- Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, and Jinsung Yoon. 2026a. [Paper-Banana: Automating academic illustration for AI scientists](#). *Preprint*, arXiv:2601.23265.
- Minjun Zhu, Zhen Lin, Yixuan Weng, Panzhong Lu, Qiujie Xie, Yifan Wei, Sifan Liu, QiYao Sun, and Yue Zhang. 2026b. [AutoFigure: Generating and refining publication-ready scientific illustrations](#). In *The Fourteenth International Conference on Learning Representations*. ICLR 2026 poster.

A Supplementary Statistics

Figure 4 reports descriptive diagnostics for the released corpus rather than task metrics. Panel (a) contextualizes caption-text retrieval, panel (b) explains why year-based analyses should account for venue coverage changes, and panel (c) shows why subject metadata is useful for domain-aware filtering.

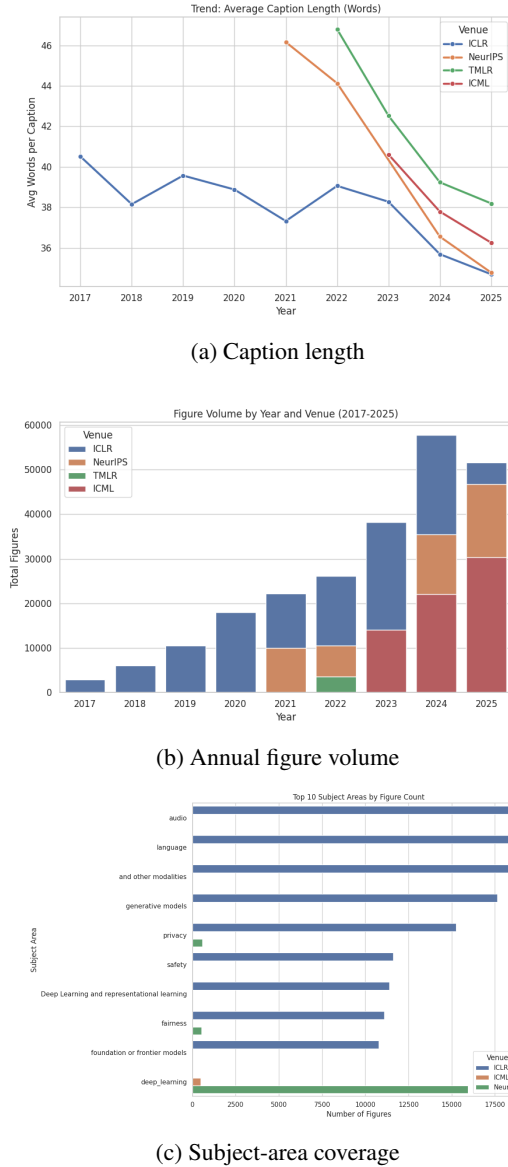


Figure 4: DiagramBank statistics. (a) Mean caption length by year and venue, showing venue-specific verbosity and a recent shift toward shorter captions in several venues. (b) Annual extracted-figure volume, where growth reflects both publication volume and the gradual expansion of OpenReview venue coverage. (c) Top subject areas by figure count, highlighting the venue-specific taxonomy differences that motivate preserving normalized metadata.

B Full Release Schema

Tables 6 and 7 list the denormalized fields exposed to downstream users. Paper-level fields make each record self-contained for retrieval and attribution, while figure-level and quality fields expose the extraction, filtering, and user-side subset controls.

Field	Type	Description
platform_id	string	Unique OpenReview identifier used to join paper- and figure-level information.
venue	string	One of {ICLR, ICML, NeurIPS, TMLR}.
year	integer	Publication year; release coverage is venue-specific from 2017–2026.
paper_title	string	Paper title.
paper_abstract	string	Paper abstract.
paper_authors	string or list	Source-paper author string or list (as available from OpenReview metadata), used for attribution.
paper_keywords	string or list	Keywords (when available).
paper_subject_areas	string or list	Primary subject areas or taxonomy tags (when available).
paper_tldr	string	TL;DR summary (when available).
paper_decision	string	Decision status (when available).
paper_scores	number or list	Aggregated reviewer scores (when available).
paper_url	string	Web address for the OpenReview paper and, when available, the paper file.
paper_bibtex	string	BibTeX entry for attribution to the original work.

Table 6: Paper-level metadata fields in each denormalized DiagramBank record.

Field	Type	Description
figure_id	string or integer	Figure identifier within the source paper file, such as the figure number.
figure_path	string	Relative path to the extracted figure image in the release.
caption	string	Figure caption extracted by PDFFigures 2.0.
figure_context	string	Paragraphs that cite the figure number, extracted from the paper text.
clip_label	string	First-stage CLIP type label, such as diagram, plot, photo, or other.
clip_confidence	float	Confidence score for the CLIP-based classification.
label_cascade	string or Boolean	Final cascade decision used for the primary release.
cascade_path	string	Kept/rejected path, enabling stricter user-side filtering.

Table 7: Figure-level, classification, and quality-control fields. Exact field names are documented in the dataset card and anonymous supplementary package.

C Qualitative Authoring Use Case

We illustrate one DiagramBank use with a case study on visualizing the “Code2MCP” framework (Ouyang et al., 2025). Figure 5 compares a text-only output with a DiagramBank-conditioned output, while Figure 6 shows the retrieved references. Together, they illustrate how exemplars can supply adjacent papers, visual terminology, grouping choices, and source-level provenance for human revision. The same released metadata also supports other query families: EagleFT retrieves speculative-decoding conventions such as

The diagram illustrates the Code2MCP Process flow, which is divided into three main stages: Supply Problem, Code2MCP Process, and Consumption/Usage.

- Supply Problem:** This stage represents the input data, categorized as GitHub Repos. It includes four specific examples: Bio Code (represented by a DNA helix icon), Math Lib (represented by a calculator icon), and Fluid Sim (represented by a wavy line icon).
- Code2MCP Process:** This central stage is represented by an orange box with a gear icon. It contains a circular workflow with four steps: Analyze (magnifying glass icon), Setup (wrench icon), Design (circuit board icon), and Generate (document with arrow icon). Arrows indicate a clockwise cycle between these steps.
- Consumption/Usage:** This final stage is represented by a green box with a plug icon. It shows the output being used by MCP Tools, specifically an LLM (Large Language Model, represented by a brain icon) and Tool Use (represented by a hand icon with a wrench).

The flow is indicated by arrows: a "Transform" arrow connects the Supply Problem to the Code2MCP Process, and an "Enable" arrow connects the Code2MCP Process to the Consumption/Usage stage.

Code Repository Supply Problem

Bioinformatics Repos

Mathematics Repos

Fluid Dynamics Repos

1

GitHub Repository

Code2MCP Framework

Multi-agent Workflow

Code Analysis

Environment Setup

Service Generation

Tool Function Design

2

MCP Service Output

MCP-Compatible Services

Standardized Service 1

Standardized Service 2

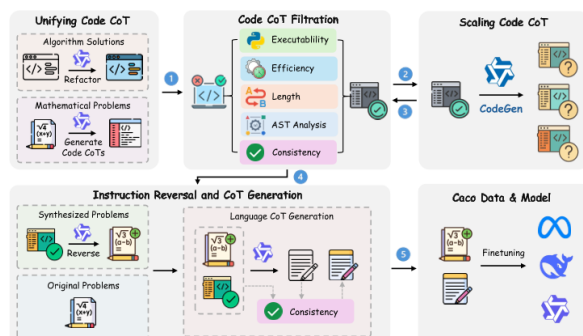
Tool Pool

LLM Integration

3

Automated Transformation Pipeline

Figure 5: Qualitative authoring use case. A text-only illustrative output (a) is compared with a DiagramBank-conditioned illustrative output (b). The comparison is not a visual-quality benchmark; it shows how retrieved exemplars can expose reusable diagram conventions for human inspection and revision.

[illegible][illegible]

(c) Code-graph model

D Dataset Card / Datasheet and Release Protocols

Intended and non-intended use. DiagramBank is intended to enable retrieval of semantically relevant schematic-diagram exemplars at paper, caption, and context granularity, support construction of future multimodal retrieval evaluations on scientific schematic figures, and serve as reference material for exemplar-driven figure authoring systems. It is not intended for identity inference, author profiling, or producing misleading scientific artifacts.

User-side filtering. Users should report which subset, confidence threshold, and cascade paths they use. Because extraction is automated, downstream studies should apply task-specific checks when needed, such as filtering records with missing captions, very short captions, incomplete context spans, or extremely small images.

Distribution, licensing, and attribution. During review, DiagramBank resources are provided through an anonymized supplementary archive or repository that does not track reviewer identities; the public hosting location can be de-anonymized after review. The release distinguishes metadata/code from extracted figure images: metadata and code carry explicit research licenses, while images are redistributed only when permitted and otherwise represented by source links and attribution metadata. Users are asked to cite both DiagramBank and the original source papers when using specific diagrams, and the maintainers will honor reasonable rights-holder takedown requests.