

DIAGRAMS: A Review Framework for Reasoning-Level Attribution in Diagram QA

Anirudh Iyengar Kaniyar Narayana Iyengar^{*,*}, Tampu Ravi Kumar^{*,*},
Manan Suri^{*,*}, Raviteja Bommireddy^{*,*}, Dinesh Manocha^{*,*},
Puneet Mathur^{*,*}, Vivek Gupta^{*,*}

^{*} Arizona State University [†] Adobe Research [‡] IIITDM [§] University of Maryland

[▶ Project Page](#) [■ Video](#) [🔗 Code](#)

akaniyar@asu.edu, traviku2@asu.edu, manans@umd.edu, cs23b2011@iiitdm.ac.in
dmanocha@umd.edu, puneetm@adobe.com, vgupt140@asu.edu

Abstract

Diagram question answering (Diagram QA) requires *reasoning-level attribution* that links each question-answer pair to all visual regions needed to derive the answer, rather than only the region containing the final response. Creating such structured evidence across diagrams, charts, maps, circuits, and infographics is time-consuming, and existing annotation tools tightly couple their interfaces to dataset-specific formats. We present **DIAGRAMS**, a lightweight, schema-driven review framework that decouples interface logic from dataset-specific JSON structures through an internal meta-schema and dataset adapters. Given an image and QA pair with optional candidate regions, the system performs QA-conditioned evidence selection and proposes the regions required for reasoning. When QA pairs or candidate regions are missing, it generates them and supports human verification and refinement. Across six Diagram QA datasets, model-suggested evidence achieves 85.39% precision and 75.30% recall against reviewer-final selections (micro-averaged). These results indicate that the review-first framework reduces the number of regions that annotators must create from scratch. Human reviewers accept, edit, or reject each proposed region before export, which structurally limits over-reliance on AI proposals. We release a public demo and installable package to support dataset auditing, grounded supervision creation, and grounded evaluation.

1 Introduction

Diagram question answering (Diagram QA) spans diverse visual domains, including scientific diagrams (Kembhavi et al., 2016), charts (Masry et al., 2022), circuit schematics (Mehta et al., 2024), infographics (Mathew et al., 2022), and map-based

benchmarks (Srivastava et al., 2025; Mukhopadhyay et al., 2025). Unlike pure text-based reasoning, Diagram QA requires models to interpret structured visual elements and relate them to linguistic queries. In many cases, answering a question does not depend on a single visual region, but on a chain of interconnected elements. For example, when a question asks about the neighbors of a highlighted state on a map, the reasoning process requires identifying the target state and examining all adjacent regions. The final answer emerges from multiple visual dependencies rather than from a single answer-containing box.

Despite this, most existing annotation efforts focus on answer-level grounding, where annotators mark only the region that directly contains the final response. Such annotations fail to capture the full reasoning path that supports the answer. Reasoning-level attribution links each question-answer pair to all visual regions required to derive the answer. This richer supervision enables grounded training, faithfulness evaluation, and systematic error analysis of vision-language models (VLMs). Without structured attribution, models may produce correct answers while relying on spurious correlations or hallucinated associations. Recent analyses of chart reasoning benchmarks demonstrate that models often struggle with distributed and compositional visual evidence, reinforcing the need for structured, region-level attribution (Iyengar et al., 2025). Annotators find reasoning-level attribution expensive. Annotators must analyze the diagram, identify all relevant regions, and create or refine precise bounding boxes. The burden increases further because Diagram QA datasets vary widely in JSON structure, coordinate conventions, region availability, and QA formatting. Many existing annotation interfaces tightly couple their design to a specific dataset format, making it difficult to reuse tools across heterogeneous resources. As a result, researchers often rebuild front-end logic for each

*These authors contributed equally to this work

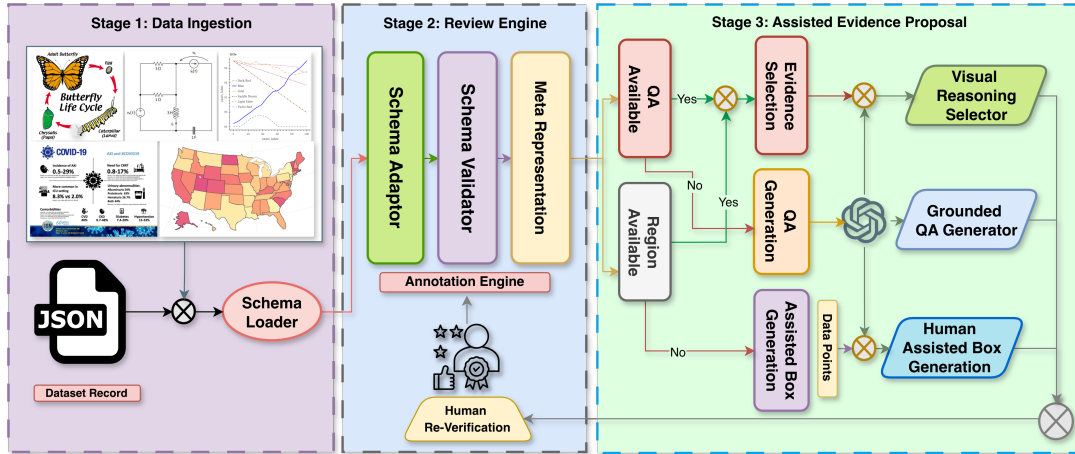


Figure 1: Overview of the **DIAGRAMS** architecture. The framework operates in three stages: (1) Data Ingestion, which converts dataset-specific records into a unified meta-schema; (2) Review Engine, which renders normalized QA items and candidate regions for human verification; and (3) Assisted Evidence Proposal, where a multimodal backend performs QA-conditioned evidence selection and optional QA or region generation.

new dataset, slowing progress and limiting reproducibility.

We introduce **DIAGRAMS**, a review-first framework for reasoning-level attribution in Diagram QA. Rather than centering the interface on dataset-specific schemas, **DIAGRAMS** converts each input record into an internal meta-schema representation that decouples annotation logic from dataset structure. A multimodal backend proposes QA-conditioned evidence regions, and the human reviewer verifies, edits, removes, or adds regions as needed. When candidate boxes or QA pairs are missing, the system proposes them and supports interactive refinement. **DIAGRAMS** targets researchers developing Diagram QA benchmarks, practitioners building grounded supervision datasets, and vision-language researchers studying reasoning-level attribution.

We evaluate **DIAGRAMS** across six Diagram QA datasets spanning charts, maps, diagrams, circuits, and infographics. Model-suggested evidence achieves high agreement with reviewer-final selections, demonstrating that a review-first pipeline shifts annotation effort from manual region creation to guided verification while preserving reasoning-level fidelity. We summarize our contributions as follows:

1. We present **DIAGRAMS**, a review framework that formalizes *reasoning-level* attribution for Diagram QA, explicitly distinguishing it from answer-only grounding.
2. We design a multimodal proposal pipeline that

performs QA-conditioned evidence selection, missing-region generation, and optional QA generation, enabling structured human-in-the-loop verification.

3. We provide empirical evaluation across six visual QA datasets, demonstrating strong proposal utility under a review-first workflow.

2 Related Work

Existing image annotation platforms provide extensive support for drawing and editing visual regions such as bounding boxes, polygons, and masks. Systems including LabelMe (Russell et al., 2008), VIA (Dutta and Zisserman, 2019), CVAT (Sekachev et al., 2019; CVAT.ai Corporation, 2023), and Label Studio (Tkachenko et al., 2020-2025) also incorporate model-assisted workflows to accelerate labeling. These platforms standardize region creation and editing but operate primarily at the object or instance level rather than supporting *QA-conditioned* evidence attribution, and they do not address the structural heterogeneity inherent in Diagram QA datasets. Prior work has further reduced bounding-box effort through interaction techniques such as boundary snapping (Snapper) (Williams et al., 2024) and iterative human-in-the-loop box refinement (Adhikari and Huttunen, 2021). In parallel, VQA research has explored answer grounding, attention supervision, and explanation alignment (Das et al., 2017; Wang et al., 2020; Chen et al., 2022; Pantazopoulos and Özyiğit, 2025; Selvaraju et al., 2017, 2019), often using region-level or attention-based signals to im-

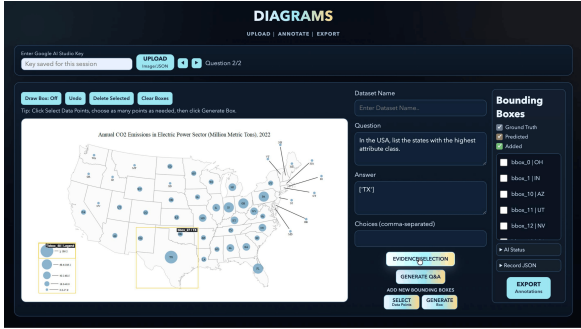


Figure 2: Selection-based attribution. The reviewer verifies QA-conditioned region proposals and inspects ground-truth, predicted, and added boxes.

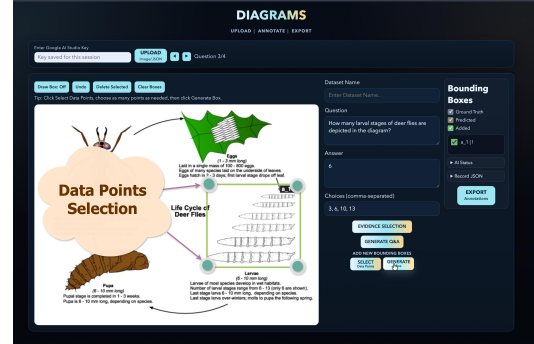


Figure 3: Proposal-based attribution with region generation. The reviewer draws or refines evidence boxes when candidate regions are unavailable.

prove model interpretability. However, these approaches typically annotate answer-containing regions or saliency maps rather than full reasoning chains. DIAGRAMS extends this line of work by targeting *reasoning-level* attribution and by providing a schema-driven review workflow that remains reusable across heterogeneous dataset formats.

3 DIAGRAMS: System Architecture and Workflow

DIAGRAMS operates in three coordinated stages: data ingestion and normalization, QA-conditioned evidence proposal, and structured human verification. The system first converts heterogeneous dataset records into a unified internal meta-schema representation. It then performs QA-conditioned evidence selection or generation using a multimodal backend. Finally, reviewers verify, refine, and export reasoning-level attributions through an interactive interface. Figure 1 illustrates this end-to-end workflow.

3.1 Architecture

The framework consists of the following components:

1. **Schema Loader:** The system ingests a JSON file and retrieves the corresponding images, questions/annotations.
2. **Schema Adapter:** The adapter maps dataset-specific fields into an internal meta-schema representation.
3. **Schema Normalizer:** The normalization layer standardizes heterogeneous bounding box formats (e.g., $[x, y, w, h]$, left/top/width/height).
4. **Internal Meta-Schema Representation:** The system stores images, QA pairs, candidate re-

gions, and attribution mappings in a unified internal structure.

5. **Annotation Engine:** The interface renders images, QA items, and regions, and tracks edits through standard region operations.
6. **Evidence Proposal Module:** The multimodal backend performs QA-conditioned evidence selection or generation.
7. **Human Re-Verification:** The reviewer verifies, edits, removes, or adds regions and QA items.
8. **Exporter:** The system saves finalized annotations in structured JSON format and optionally exports annotated images.

This modular design separates dataset-specific parsing from attribution logic and human interaction. By decoupling dataset-specific schema handling from attribution logic, the system remains reusable across heterogeneous Diagram QA datasets without requiring interface redesign.

3.2 Review-First Workflow

DIAGRAMS implements a review-first workflow that prioritizes structured verification over manual region creation. After loading and normalizing a record, the system invokes the evidence proposal module to generate QA-conditioned region suggestions. The reviewer inspects these proposals directly in the interface and verifies, edits, or removes them instead of drawing all regions from scratch. This interaction model shifts the annotation process from box creation to guided verification. We illustrate both selection-based and proposal-based workflows in Figures 2-4.

Depending on data availability, the system operates in two settings:

Selection-based attribution: When the dataset provides QA pairs and candidate regions, the sys-

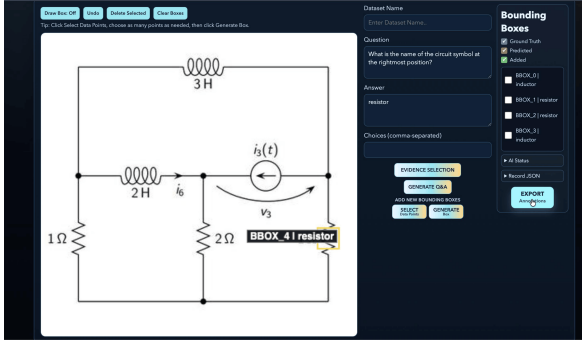


Figure 4: Human verification and export. The reviewer confirms, edits, or removes proposed regions before finalizing the annotation.

tem performs QA-conditioned evidence selection over the candidate region pool and initializes the selected subset in the interface.

Proposal-based attribution: When QA pairs and/or candidate regions are missing, the system proposes them automatically and supports interactive human refinement before finalization.

4 Evidence Proposal

We implement the evidence proposal module using two backend components. For QA-conditioned evidence selection, we use Gemini 2.5 Flash (Comanici et al., 2025) via the Gemini API, prompting the model with the diagram image, question, answer, and candidate bounding boxes to return the region identifiers that support the reasoning chain. For assisted region generation when candidate regions are absent, we deploy the Segment Anything Model (SAM) (Kirillov et al., 2023) locally, generating region proposals from reviewer-provided points on the image.

4.1 Evidence Selection from Candidate Regions

When both QA pairs and candidate regions are available, the backend performs QA-conditioned evidence selection over the candidate region pool. Given the question, answer, and candidate bounding-boxes, the multimodal model proposes the subset of regions required to support the reasoning chain as shown in Figure 2. The selected region identifiers are initialized in the interface. The reviewer verifies, edits, or augments the proposed evidence before export.

Dataset	Precision	Recall	F1
ChartQA	54.79	42.22	47.69
MapWise	94.49	64.10	76.38
AI2D	99.64	91.13	95.19
Circuit-VQA	66.97	49.50	56.92
InfographicsVQA	96.49	98.97	97.72
MapIQ	91.13	99.74	95.24
Overall	85.39	75.30	80.03

Table 1: Proposal utility after human verification (micro-averaged). Precision = $TP/(TP+FP)$, Recall = $TP/(TP+FN)$, F1 = harmonic mean.

4.2 Region Generation When Candidates Are Absent

When candidate regions are unavailable but QA pairs exist, the backend proposes new evidence regions directly on the image. The reviewer may accept, edit, or delete these proposals and may draw additional regions if required. Figure 3 illustrates manual refinement when candidate regions are absent or incomplete.

4.3 QA and Region Generation

When QA pairs and candidate regions are missing, the system first proposes QA items and then generates corresponding evidence regions. The reviewer verifies and refines both components before finalizing attribution, as shown in Figure 4.

5 Annotation Engine

The annotation engine renders images, QA items, and regions in an interactive interface. Reviewers can select, resize, delete, or draw regions, while the system tracks edits and maintains links between QA items and evidence regions. The interface exposes normalized metadata to preserve traceability between the internal meta-schema and the original dataset record. Figure 2 shows selection-based attribution, where candidate regions exist and the reviewer verifies proposed evidence.

6 Human Re-Verification and Export

During re-verification, reviewers confirm or modify proposed QA items and evidence regions. The system records whether each region was proposed, edited, removed, or newly created. Appendix A details the annotation guidelines and arbitration procedures. After verification, the exporter saves the finalized attribution in structured JSON format and optionally produces an annotated image. These outputs support downstream training, auditing, and

Dataset	Added GT	New Drawn
ChartQA	8308	215
MapWise	2700	1162
AI2D	1173	23
Circuit-VQA	1405	1846
InfographicsVQA	28	68
MapIQ	36	0
Overall	13650	3314

Table 2: Breakdown of reviewer-added evidence regions. “Added GT” denotes missed ground-truth boxes; “New Drawn” denotes boxes created during review.

evaluation workflows. Figure 4 shows reviewers finalizing attribution before export.

7 Results and Analysis

We organize our evaluation around four research questions. (RQ1) Does QA-conditioned evidence proposal reduce manual annotation effort while maintaining correctness? (RQ2) Does proposal utility generalize across heterogeneous Diagram QA datasets? (RQ3) What types of errors drive reviewer intervention? (RQ4) Is proposal verification reproducible across annotators? (RQ5) Does the AI-assisted workflow save per-instance annotation time relative to fully manual annotation? We answer these questions through large-scale proposal utility analysis across six datasets. Specifically, we use AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), Circuit-VQA (Mehta et al., 2024), InfographicVQA (Mathew et al., 2022), MapIQ (Srivastava et al., 2025), and MapWise (Mukhopadhyay et al., 2025) and a focused reliability study on ChartQA and Circuit-VQA.

7.1 Proposal Utility Across Six Datasets

We evaluate proposal utility by comparing model-suggested regions to reviewer-approved evidence using micro-averaged precision, recall, and F_1 . Appendix B provides formal metric definitions and implementation details.

Do model proposals reduce manual region creation? As shown in Table 1, the backend proposed 60,553 boxes, of which reviewers retained 51,709 (85.39% precision). Reviewers finalized 68,673 evidence boxes in total, and proposals achieved 75.30% recall and 80.03 F_1 (micro-averaged). These results show that the system pre-selects a large fraction of final evidence regions, converting large-scale manual box creation into

Dataset	New Drawn (%)
ChartQA	2.52
MapWise	30.10
AI2D	1.92
Circuit-VQA	56.79
InfographicsVQA	70.83
MapIQ	0.00
Overall	19.53

Table 3: Percentage of false negatives requiring newly drawn boxes. Higher values indicate greater need for novel region creation rather than missed ground-truth selection.

guided verification.

Does the system prioritize correctness over over-generation? High precision (85.39%) indicates that reviewers removed relatively few suggested boxes. The backend avoids aggressive over-proposal and instead surfaces regions that reviewers largely accept, demonstrating stable QA-conditioned selection.

Where does the system still miss evidence? Reviewers added 16,964 missing boxes (Table 2), indicating that recall remains the primary area for improvement. Among these, 13,650 correspond to missed ground-truth regions and 3,314 to newly drawn boxes, suggesting that missed candidate coverage drives most recall gaps.

7.2 Cross-Dataset Variability

Which datasets show strongest alignment? AI2D, InfographicsVQA, and MapIQ exhibit the strongest proposal alignment. AI2D achieves 99.64% precision and 91.13% recall (95.19 F_1), while InfographicsVQA reaches 97.72 F_1 . These results suggest that structured diagram layouts and clearer region boundaries support highly reliable proposal behavior.

Why is ChartQA challenging? ChartQA achieves 54.79% precision and 42.22% recall. Notably, 8,308 of 8,523 false negatives correspond to missed ground-truth boxes rather than newly drawn regions. This pattern suggests that evidence regions exist but are not selected by the model. Improving candidate selection strategy could significantly raise recall in chart-heavy domains.

Why does Circuit-VQA require more intervention? Circuit-VQA shows moderate precision (66.97%) and recall (49.50%). Over half of its

Dataset	Criterion	Agreement (%)	κ
ChartQA	CVR	87.0	0.658
ChartQA	CEA	83.0	0.589
Circuit-VQA	CVR	73.0	0.464
Circuit-VQA	CEA	85.0	0.540

Table 4: Inter-annotator agreement on 100 instances per dataset under two verification criteria.

false negatives (56.79%) required newly drawn regions, suggesting incomplete candidate coverage or fine-grained component requirements in circuit diagrams.

7.3 Human-AI Verification

Does the system primarily fail due to selection or region absence? The breakdown in Table 2 shows that most false negatives (13,650/16,964) correspond to missed ground-truth boxes rather than newly drawn regions. Therefore, incomplete QA-conditioned evidence selection, rather than region absence, explains most recall gaps. Improving the evidence selection mechanism could further reduce reviewer workload.

How often must reviewers create entirely new boxes? Across all datasets, only 19.53% of false negatives required newly drawn boxes (Table 3). In most cases, evidence regions already exist but were not proposed. This finding confirms that the review-first workflow primarily involves selecting from available regions rather than drawing from scratch.

Does proposal utility remain stable across heterogeneous visual domains? Proposal utility remains strong across heterogeneous visual domains, with performance varying by dataset. AI2D and MapIQ show high alignment, while ChartQA and Circuit-VQA exhibit lower recall. Despite these differences, the overall 80.03 F_1 indicates that QA-conditioned evidence selection generalizes across charts, maps, circuits, and infographics under heterogeneous dataset formats.

7.4 Inter-Annotator Reliability (Selection-Based Proposals)

We assess annotation reliability for selection-based evidence proposals on 100 instances per dataset. We evaluate agreement under two complementary criteria: Complete Visual Reasoning (CVR), which requires full reasoning-chain coverage, and Core Evidence Alignment (CEA), which requires correct

Dataset	AI _{ES}	SAM _{RP}	Manual _{ES} +SAM _{RP}	Δ Time (saved)
AI2D	~13s	~51s	~91s	~27s
ChartQA	~10s	~28s	~90s	~52s
Circuit-VQA	~12s	~9s	~40s	~19s
MapWise	~14s	~20s	~60s	~26s
MapIQ	~20s	~23s	~60s	~17s
InfographicsVQA	~40s	~33s	~80s	~7s

Table 5: Average per-instance annotation time (seconds) on 20 instances per dataset across three annotation conditions. Δ Time denotes time saved per instance by the AI-assisted workflow over fully manual annotation.

grounding of the primary evidence elements.

Do reviewers agree on proposal verification?

Yes. Under CVR, ChartQA achieves 87.0% agreement ($\kappa = 0.658$) and Circuit-VQA achieves 73.0% agreement ($\kappa = 0.464$) (Table 4). Under CEA, agreement remains high across both datasets (83.0% and 85.0%, respectively). These results indicate reproducible selection-level verification across structured charts and spatially complex circuit diagrams. We report full criterion definitions and extended discussion in Appendix C.

7.5 Annotation Efficiency

We conduct a timing study across 120 instances, sampling 20 instances per dataset, under three conditions: (1) AI_{ES}, where we measure the average time Gemini 2.5 Flash takes to generate evidence region proposals per image and QA pair; (2) SAM_{RP}, where the reviewer verifies AI-proposed evidence and uses SAM to draw bounding boxes for any missing regions; and (3) Manual_{ES}+SAM_{RP}, where the reviewer selects evidence without AI proposals and uses SAM only for bounding box drawing. We compute Δ Time as the per-instance time saved by the AI-assisted workflow over fully manual annotation, defined as Manual_{ES}+SAM_{RP} minus (AI_{ES}+SAM_{RP}). Table 5 reports average per-instance times across all three conditions. We discuss the limitations of this study and the risk of AI-induced annotation bias in Appendix E.

Does the review-first workflow reduce per-instance annotation time?

Yes. The AI-assisted workflow saves 7 to 52 seconds per instance across all six datasets, with ChartQA showing the largest gain (~52s) and InfographicsVQA the smallest (~7s). Gemini 2.5 Flash generates evidence proposals in 10 to 40 seconds per instance, while

Manual_{ES}+SAM_{RP} requires 40 to 91 seconds, confirming that the review-first workflow consistently reduces per-instance annotation time across heterogeneous diagram domains.

Where does reviewer effort concentrate under AI assistance? Datasets with sparser candidate coverage, such as Circuit-VQA and MapWise, require reviewers to draw more new boxes via SAM after AI evidence selection, which narrows the time saved relative to fully manual annotation (~ 19 s and ~ 26 s respectively). InfographicsVQA shows the smallest Δ Time at ~ 7 s, where denser visual layouts increase both proposal generation time and SAM-assisted drawing effort. This pattern aligns with the false negative breakdown in Table 2, where Circuit-VQA records the highest rate of newly drawn regions at 56.79%.

8 Demo and Reproducibility

We release DIAGRAMS as an open-source MIT-licensed repository and provide the complete review interface, including dataset loading, meta-schema normalization, QA-conditioned proposal visualization, interactive region editing, and export of reasoning-level annotations. The repository contains the unified meta-schema, dataset adapters for all evaluated datasets, and example input–output records. Researchers can load supported Diagram QA datasets, conduct human-in-the-loop verification, and export standardized attributions. Appendix B details the proposal utility metrics and evaluation procedure.

9 Conclusion

We presented DIAGRAMS, a review-first framework for reasoning-level attribution in Diagram QA. The system links each question–answer pair to the complete set of visual regions required for reasoning and separates dataset-specific parsing from attribution logic through a unified meta-schema representation. By combining QA-conditioned proposal models with structured human verification, DIAGRAMS shifts the workflow from manual region creation to guided review. Across six heterogeneous Diagram QA datasets, model-suggested evidence achieved 85.39% precision and 75.30% recall (micro-averaged) against reviewer-final selections, demonstrating strong alignment under a review-first workflow. A timing study across 120 instances further shows that the AI-assisted workflow saves 7 to 52 seconds per instance relative

to fully manual annotation across all six datasets. Our analysis shows that most recall gaps arise from missed candidate selection rather than absence of regions, indicating that improved QA-conditioned selection could further reduce reviewer effort. The released framework, schema, and dataset adapters support structured reasoning-level attribution across diverse Diagram QA resources and enable grounded supervision and dataset auditing for diagram understanding.

Limitations

DIAGRAMS improves reasoning-level attribution efficiency but has several limitations. First, evidence quality depends on the underlying multimodal models, and performance may degrade for densely packed or visually cluttered diagrams that require fine-grained localization. Second, the current system represents evidence primarily as bounding boxes, which may not accurately capture thin lines, irregular shapes, or non-rectangular regions common in circuits and infographics. Third, our evaluation measures proposal utility under a review-first workflow and includes a preliminary timing study across 120 instances, but does not quantify annotation time reduction through a controlled user study with statistical rigor. Finally, the framework assumes structured QA inputs and does not support chain-of-thought supervision. Future work may incorporate polygon-level annotation, latency analysis, broader human studies, and structured reasoning trace integration.

Ethics Statement

We designed DIAGRAMS to support transparent and responsible human-in-the-loop annotation. All datasets used in this work originate from publicly available research benchmarks under their respective licenses. The framework does not introduce new personal or sensitive data. During annotation, we provided clear written guidelines and arbitration procedures to ensure consistency and fairness. Annotators participated voluntarily and followed standardized review protocols. Because DIAGRAMS integrates multimodal models for proposal generation, users must carefully verify all model-suggested regions and QA items before finalizing annotations. The system does not assume model correctness and places decision authority entirely with the human reviewer. We release the schema specification, adapters, and evaluation scripts to promote reproducibility and transparency.

We encourage users to respect original dataset licenses when redistributing derived annotations.

Acknowledgements

We conducted this research as a collaborative effort at Arizona State University. We thank CoRAL ASU within the School of Augmented Intelligence at Arizona State University for providing computational resources and institutional support. We also thank Neha Valeti and Gaurav Najpande for their assistance during the annotation phase of this project. Finally, we are grateful to the anonymous reviewers for their thoughtful feedback and constructive suggestions, which helped improve the clarity and quality of this work.

References

- Bishwo Adhikari and Heikki Huttunen. 2021. Iterative bounding box annotation for object detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4040–4046. IEEE.
- Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- CVAT.ai Corporation. 2023. [Computer Vision Annotation Tool \(CVAT\)](#).
- Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.
- Abhishek Dutta and Andrew Zisserman. 2019. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2276–2279.
- Anirudh Iyengar Kaniyar Narayana Iyengar, Srija Mukhopadhyay, Adnan Qidwai, Shubhankar Singh, Dan Roth, and Vivek Gupta. 2025. [Interchart: Benchmarking visual reasoning across decomposed and distributed chart information](#). In *IJCNLP-AACL*.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279. Association for Computational Linguistics.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Rahul Mehta, Bhavyajeet Singh, Vasudeva Varma, and Manish Gupta. 2024. Circuitvqa: A visual question answering dataset for electrical circuit images. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 440–460. Springer.
- Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. 2025. Mapwise: Evaluating vision-language models for advanced map queries. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9348–9378.
- Georgios Pantazopoulos and Eda B Özyiğit. 2025. Towards understanding visual grounding in visual language models. *CoRR*.
- Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173.
- Boris Sekachev, M Nikita, and Z Andrey. 2019. Computer vision annotation tool: A universal approach to data annotation. 2019. URL <https://github.com/openvinotoolkit/cvat>.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.
- Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600. IEEE.

- Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. 2025. Mapiq: Evaluating multimodal large language models for map question answering. In *Second Conference on Language Modeling*.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. [Label Studio: Data labeling software](https://github.com/HumanSignal/label-studio). Open source software available from <https://github.com/HumanSignal/label-studio>.
- Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. 2020. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10123–10132.
- Alex C Williams, Min Bai, Jonathan Buck, Tristan J McKinney, Amy Rechkemmer, Koushik Kalyanaraman, Matthew Lease, Patrick Haffner, Xiong Zhou, and Erran Li. 2024. Snapper: Accelerating bounding box annotation in object detection tasks with find-and-snap tooling. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 471–488. ACM.

A Annotation Instructions

To ensure consistency and reliability, we provided annotators with standardized written guidelines defining task objectives, labeling rules, and arbitration procedures. Below we summarize the core instructions.

Objective Annotators verified the correctness and clarity of each question–answer (QA) pair and identified all visual regions required to derive the answer. For each QA item, annotators selected bounding boxes corresponding to the interconnected visual elements necessary for reasoning, rather than marking only the region containing the final answer.

Procedure

1. Read the question carefully and locate all referenced visual elements in the image.
2. Use the verified answer (and answer choices, if provided) to select the bounding boxes that support the reasoning process.
3. If a question is ambiguous, inconsistent, or unanswerable from the image, flag it for arbitration rather than guessing.

Annotation Policy

1. Two independent annotators labeled each item (six annotators participated overall).
2. A senior annotator reviewed disagreements following predefined arbitration guidelines.
3. We retained only consensus or majority-agreed annotations in the final dataset.

Examples We provided dataset-specific examples illustrating reasoning-level attribution across diagrams, charts, maps, circuits, and infographics. We will release the full instruction documents, templates, and arbitration notes as supplementary materials to support transparency and reproducibility.

B Evaluation

B.1 Proposal Utility Metrics

We evaluate the usefulness of model-suggested evidence regions in a review-first workflow by comparing the proposed box set P to the final reviewer-approved evidence set H .

We define:

- True positives: $TP = |P \cap H|$ (proposed boxes retained after review),
- False positives: $FP = |P \setminus H|$ (proposed boxes removed by the reviewer),
- False negatives: $FN = |H \setminus P|$ (evidence boxes added during review).

In our logs, TP corresponds to `retained_pred_count`, FP to `effective_removed_count`, and FN to the sum of `added_gt_count` and `new_drawn_count`. We report micro-averaged precision $TP/(TP + FP)$, recall $TP/(TP + FN)$, and F_1 . We do not define true negatives because there is no closed set of non-evidence regions.

B.2 Discussion

The review-first workflow generalizes across diagram domains, though datasets differ in whether errors arise from *distracting proposals* (higher FP) or *coverage gaps* (higher FN). ChartQA presents the greatest challenge due to dense chart elements and small text, while AI2D, InfographicsVQA, and MapIQ show strong alignment with few removals. These findings motivate future work on confidence-based filtering and improved QA-conditioned selection strategies.

C Inter-Annotator Agreement Details

C.1 Evaluation Scope

We evaluate reliability for the selection-based proposal workflow only, where the system selects evidence from pre-existing candidate regions. We exclude newly drawn regions from this analysis to isolate agreement on evidence selection quality without conflating variability introduced by bounding-box generation.

C.2 Evaluation Criteria

Complete Visual Reasoning (CVR) Reviewers mark an instance as correct only when the selected regions capture the entire visual reasoning chain required to derive the answer. If any required intermediate evidence element is missing, reviewers mark the instance as incorrect.

Core Evidence Alignment (CEA) Reviewers mark an instance as correct when the selected regions capture the primary visual evidence necessary to answer the question, even if minor intermediate reasoning components are omitted. This

criterion evaluates agreement on essential grounding rather than exhaustive coverage. This design isolates agreement on evidence selection quality without introducing variability from newly drawn regions.

D Input and Output Data Format

D.1 Input Record Structure

The tool accepts a JSON array containing one or more image records. Each record stores metadata, optional candidate regions, predicted regions, and one or more question–answer pairs. The system loads one image record at a time and saves annotations independently.

Example Input Format (Multi-Image JSON)

```
[
  {
    "image_uid": "img_001",
    "image_path": "path_to_image",
    "annotation_path": "path_to_json",
    "bbox": [...],
    "predicted_boxes": [],
    "questions": [
      {
        "question_text": "...",
        "answer_text": "...",
        "choices": []
      }
    ]
  },
  {
    "image_uid": "img_002",
    "image_path": "path_to_image",
    "annotation_path": "path_to_json",
    "bbox": [],
    "predicted_boxes": [],
    "questions": []
  }
]
```

The loader automatically detects dataset structure and converts each record into a unified internal meta-schema before rendering.

D.2 Unified Output Meta-Schema

After review, the system exports one JSON file per image using a standardized meta-schema. This representation ensures consistent reasoning-level attribution across heterogeneous Diagram QA datasets.

Unified Output Format (Per Image)

```
{
  "dataset_type": "",
  "image": "path_to_image",
  "qa": {
    "question": "...",
    "answer": "...",
    "choices": []
  },
  "annotations": [
    {
      "id": "a_1",
      "bbox": [x, y, w, h],
      "label": "",
      "meta": {
        "source": "added",
        "kind": "bbox"
      }
    }
  ],
  "metadata": {
    "annotation_path": "path_to_json",
    "ground_truth_path": "",
    "answers": {}
  }
}
```

This design enables multi-image ingestion while exporting standardized, per-image reasoning-level annotations.

E Annotation Bias Discussion

We do not conduct a controlled blind study isolating AI-induced annotator bias, and we acknowledge this as a limitation. A key risk in review-first annotation is that human reviewers may anchor on AI-proposed regions and accept partial reasoning chains without exhaustively searching the diagram for missing interconnected elements, potentially lowering the ceiling of annotation quality. However, the review-first design requires reviewers to actively accept, edit, or reject each proposed region rather than passively confirm AI output. Reviewers may also draw entirely new boxes when proposals are absent or incomplete, as evidenced by the 3,314 newly drawn regions across all datasets (Table 2). We treat the timing study in Section 7.5 as a preliminary efficiency signal and encourage future work to quantify bias through controlled annotation experiments comparing reviewer decisions with and without AI assistance on the same instances.