SciEGQA: A Dataset for Scientific Evidence-Grounded Question Answering and Reasoning

Wenhan Yu*
yuwenhan@buaa.edu.cn
Beihang University
Beijing, China

Zhaoxi Zhang*
zhaoxizhang25@stu.pku.edu.cn
Peking University
Beijing, China

Wang Chen*
wchen22@connect.hku.hk
The University of Hong Kong
Hong Kong, China

Guanqiang Qi
qiguanqiang@baidu.com
Baidu Inc.
Beijing, China

Weikang Li†
wavejkd@pku.edu.cn
Peking University
Beijing, China

Lei Sha†
shalei@buaa.edu.cn
Beihang University
Beijing, China

Deguo Xia
xiageguo@baidu.com
Baidu Inc.
Beijing, China

Jizhou Huang
huangjizhou@baidu.com
Baidu Inc.
Beijing, China

Abstract

Scientific documents contain complex multimodal structures, which makes evidence localization and scientific reasoning in Document Visual Question Answering particularly challenging. However, most existing benchmarks evaluate models only at the page level without explicitly annotating the evidence regions that support the answer, which limits both interpretability and the reliability of evaluation. To address this limitation, we introduce SciEGQA, a scientific document question answering and reasoning dataset with semantic evidence grounding, where supporting evidence is represented as semantically coherent document regions annotated with bounding boxes. SciEGQA consists of two components: a human-annotated fine-grained benchmark containing 1,623 high-quality question–answer pairs, and a large-scale automatically constructed training set with over 30K QA pairs generated through an automated data construction pipeline. Extensive experiments on a wide range of Vision-Language Models (VLMs) show that existing models still struggle with evidence localization and evidence-based question answering in scientific documents. Training on the proposed dataset significantly improves the scientific reasoning capabilities of VLMs. The project page is available at https://yuwenhan07.github.io/SciEGQA-project/.

CCS Concepts

• **Information systems** → **Multimedia databases**; • **Computing methodologies** → **Computer vision tasks**; *Visual content-based indexing and retrieval.*

Keywords

Scientific Document Question Answering and Reasoning, Benchmark and Dataset, Evidence Region Grounding, VLMs

1 Introduction

With the rapid development of Multimodal Large Language Models (MLLMs) [5, 14, 33, 34], Vision–Language Models (VLMs) have achieved remarkable progress in tasks such as image understanding, visual question answering, and multimodal reasoning [31]. However, compared with natural images, scientific documents present fundamentally different challenges. A page of a scientific paper often contains dense multimodal content, including textual paragraphs, tables, figures, and cross-page references [13, 32]. These characteristics make Document Visual Question Answering (DocVQA) [19] particularly challenging, as models must not only accurately localize the relevant evidence regions but also integrate information across modalities to perform reasoning. In this work, we study scientific documents directly as rendered page images rather than as OCR transcripts or parsed text. This image-based setting preserves both visual content and spatial layout [8, 12, 23, 30].

In recent years, a variety of DocVQA benchmarks have been proposed to evaluate document understanding capabilities [18, 19, 21, 27]. Nevertheless, existing scientific document QA datasets still exhibit several limitations in how supporting evidence is represented. Datasets that only provide **doc/page-level** evidence annotations are often too coarse-grained for scientific documents, where a single page may contain multiple dense technical elements such as paragraphs, figures, and tables. As a result, page-level supervision cannot precisely identify the specific regions that support an answer [3, 13, 16, 23]. Furthermore, some datasets construct question answering tasks around **specific structured components**, such as a single table or figure [17, 20, 32, 35]. While useful for studying individual modalities, these settings isolate individual visual elements from the broader document context and therefore cannot capture the complete semantic information of the document. In contrast, **token-level** annotations, although more fine-grained, often fragment semantically coherent units and lose important contextual information, making systematic reasoning difficult and reducing interpretability [9, 25, 28].

*This work was done during the internships of Wenhan Yu, Zhaoxi Zhang, and Wang Chen at Baidu Inc.
†Corresponding author

However, incorporating precise evidence region localization into the reasoning process has been shown to be effective in visual question answering [10, 36, 38], yet this capability is rarely systematically evaluated in existing scientific DocVQA benchmarks. Therefore, an effective evidence representation should support **semantic region-level grounding**, which preserves semantic completeness while enabling precise spatial localization of the supporting evidence.

To address these limitations, we introduce **SciEGQA**, a visual question answering dataset for scientific documents that incorporates *semantic evidence grounding*. In SciEGQA, the supporting evidence for each answer is represented as **semantically coherent document regions** annotated with bounding boxes. This design provides an intermediate granularity between page-level and token-level annotations, enabling precise spatial localization while preserving semantic completeness, which is more suitable for multimodal reasoning over scientific documents.

SciEGQA consists of two complementary components: a **human-annotated benchmark** constructed from 80 scientific papers across eight research domains with 1,623 high-quality QA pairs, and a **large-scale automatically constructed training set** generated from more than 3.6K papers containing over 30K QA pairs. The dataset supports three reasoning scenarios with increasing complexity: **Single-Page Single-Region (SPSR)**, **Single-Page Multi-Region (SPMR)**, and **Multi-Page Multi-Region (MPMR)**. We further demonstrate through supervised fine-tuning experiments that the automatically constructed training set provides effective supervision for improving scientific reasoning capability.

Based on SciEGQA, we conduct comprehensive evaluations on a wide range of state-of-the-art vision–language models. Experimental results reveal that even the most advanced models still struggle with evidence localization and evidence-based reasoning in scientific documents. In particular, most models achieve less than 40% IoU in evidence grounding tasks, highlighting the challenging nature of scientific document understanding.

The main contributions of this work are summarized as follows:

- We introduce **SciEGQA**, a new benchmark and training set for evidence-grounded question answering and reasoning on scientific documents.
- We propose **semantic region-level evidence grounding**, an intermediate annotation granularity that preserves semantic completeness while enabling precise spatial localization for reasoning.
- We propose a **Grounding–Crop–then–Answer** evaluation protocol that disentangles evidence localization from reasoning ability.

2 Related Work

Document Visual Question Answering (DocVQA) has attracted significant attention in recent years, leading to the development of various benchmark datasets for evaluating document understanding capabilities. While many datasets focus on general document types such as receipts, forms, and infographics, scientific documents introduce additional challenges due to their domain-specific characteristics and complex multimodal structures. As illustrated in Figure 1 and Table 1, existing Scientific DocVQA datasets can be broadly categorized into three types: doc/page-level datasets, structured component-level datasets, and token-level datasets.

Table 1: Comparison of Scientific DocVQA benchmarks across doc/page-level, component-level, and token-level datasets and our SciEGQA benchmark.

Dataset	#QA	Evidence	Multi-page	BBox	Evidence Granularity
DocVQA [19]	5.19k	✓	✗	✗	Page Level
MP-DocVQA [27]	5.02k	✓	✓	✗	Page Level
TAT-DQA [39]	1.66k	✓	✗	✗	Page Level
SciDQA [23]	2.94k	✗	✓	✗	Doc Level
PeerQA [3]	0.57k	✗	✓	✗	Doc Level
MMLongBench-Doc [16]	1.09k	✓	✓	✗	Page Level
Charxiv [32]	1.32k	✓	✗	✗	Component (Chart)
SPIQA [20]	0.66k	✓	✗	✗	Component (Table)
AnaFig [35]	10k	✗	✗	✗	Component (Figure)
VisualMRC [25]	6.71k	✗	✗	✓	Token Level
DUDE [28]	41k	✗	✓	✗	Token Level
BoundingDocs [9]	4.83k	✗	✗	✓	Token Level
SciEGQA (Ours)	**1.62k**	✓	✓	✓	**Semantic Region Level**

Doc/Page-level datasets. Early datasets such as DocVQA [19] primarily focus on question answering over single-page documents. MP-DocVQA [27] extends this setting to multi-page scenarios, introducing more complex challenges such as cross-page reasoning and long-context understanding. In addition, datasets including SciDQA [23], PeerQA [3] and TAT-DQA [39] further explore question answering over scientific paper-style documents. More recently, benchmarks such as MMLongBench-Doc [16] have been proposed to systematically evaluate long-document understanding capabilities of multimodal models. However, these datasets typically provide only page-level evidence annotations, lacking precise localization of fine-grained answer-supporting regions.

Structured component-level datasets. Another line of work further focuses on structured visual components with rich semantic information in documents, such as charts, tables, and figures. Representative benchmark datasets include SPIQA [20], Charxiv [32], and AnaFig [35]. These datasets typically construct question answering tasks around specific document components, particularly figures or charts extracted from scientific papers or technical reports. While this setting facilitates fine-grained reasoning over individual components, it often isolates these elements from the broader document context, making it difficult to evaluate comprehensive document understanding and cross-component information integration.

Token-level datasets. Another line of work adopts token-level supervision by aligning answers with OCR-extracted text tokens. Datasets such as VisualMRC [25], DUDE [28], and BoundingDocs [9] support fine-grained evidence localization at the textual token level. Although this approach provides precise localization signals, the annotations are typically limited to individual tokens or short phrases, lacking complete semantic context and making it difficult to capture hierarchical structures and compositional semantics within documents.

Overall, existing datasets still exhibit notable limitations in fine-grained document modeling. Doc/Page-level methods lack precise localization, token-level methods lack semantic completeness, and component-level methods often focus on isolated visual elements rather than holistic document reasoning. These limitations are

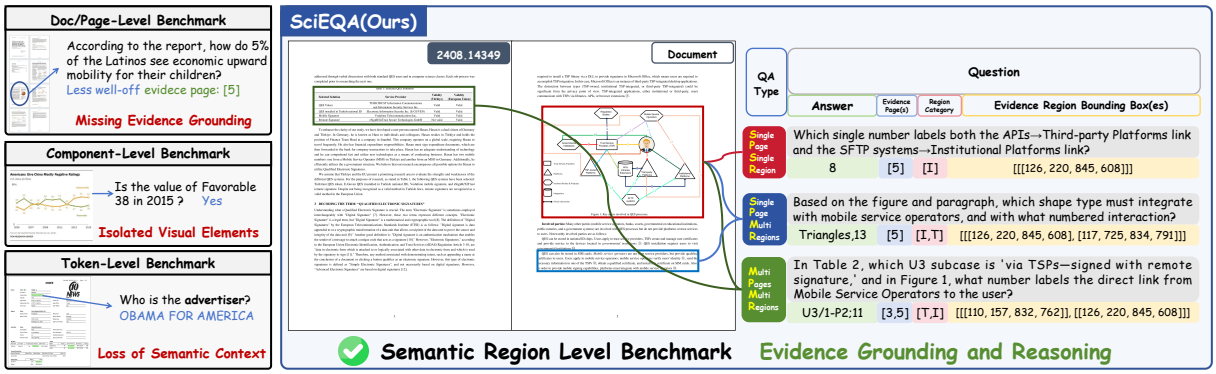


Figure 1: Comparison of evidence annotation granularity in document QA datasets. Existing benchmarks provide page-level QA, token-level grounding, or component-level reasoning. In contrast, SciEGQA supports page-level question answering with semantic region grounding via bounding-box evidence.

particularly evident in scientific documents, where answering questions often requires integrating information across text, figures, tables, and multiple document regions. To address these challenges, we introduce the **SciEGQA** dataset, which models supporting evidence as **semantic regions** annotated with bounding boxes, enabling both precise evidence localization and multimodal reasoning across document regions.

3 SciEGQA Dataset

3.1 Semantic Regions and Evidence Units

We define a **Semantic Region (SR)** as the smallest spatially contiguous document region that preserves a complete semantic unit. An SR is a reusable document-level grounding unit rather than a geometry-driven layout block or a question-specific crop.

As shown in Figure 2, OCR and PDF-parsing outputs do not reliably form complete semantic units [4, 24]. OCR omits non-textual cues, while object-level parsing can lose relationships encoded by page layout [7, 11, 15]. SR boundaries are instead determined by semantic completeness.

SR determination follows two principles: *semantic completeness* and *minimality*. A region must contain all content needed to interpret the unit while excluding adjacent content that is not required.

Text SRs are paragraphs or lists. Figure SRs include necessary captions; table SRs include the title, body, and notes; and equations include the surrounding text needed for interpretation.

Unrelated columns are not merged. Cross-page evidence is represented by multiple page-local SRs, each preserving a complete semantic unit. These rules preserve necessary context while maintaining precise localization.

For a question–answer pair, an **Evidence Unit (EU)** is an SR selected because it provides information needed to derive the answer. Thus, SR denotes a reusable document unit, whereas EU denotes its question-specific evidential role.

A question may require one or multiple EUs. Its supporting evidence is represented as a set of EUs, including multiple boxes when reasoning spans regions or pages.

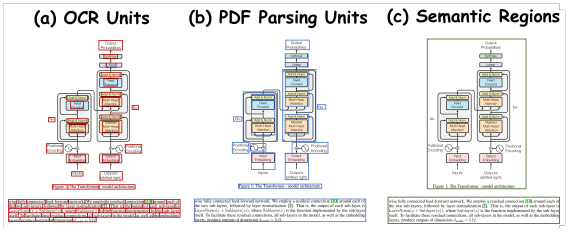


Figure 2: Comparison of OCR, PDF parsing, and semantic-region representations. OCR produces fine-grained text boxes, PDF parsing organizes content according to file objects, while semantic regions preserve coherent evidence units.

3.2 Overview of the SciEGQA Dataset

SciEGQA is designed to evaluate models on the task of **scientific document understanding with evidence grounding**. Unlike conventional DocVQA datasets that focus only on predicting the correct answer [16, 17, 19], SciEGQA explicitly models the supporting evidence required for answering each question.

Formally, each sample S in SciEGQA can be represented as

$$S = (D, Q, A, T, E),$$

where D is the source document, Q is the question, A is the answer, and T is the reasoning type. Let $R(D)$ denote the collection of annotated SRs in D.

The evidence set satisfies $E = \{e_i\}_{i=1}^{n} \subseteq R(D)$, where each EU $e_i = (p_i, b_i, c_i)$ records its page p_i, bounding box $b_i = (x_1, y_1, x_2, y_2)$, and semantic category c_i (e.g., text, figure, or table).

SciEGQA is designed with three key principles: **Semantic completeness.** Each EU inherits the complete and minimal boundary of its underlying SR rather than using fragmented tokens or a question-specific crop. **Multimodal coverage.** Scientific documents contain diverse modalities including text, tables, and figures. SciEGQA explicitly models evidence across these modalities to support realistic multimodal reasoning scenarios. **Reasoning diversity.** To reflect the complexity of scientific document understanding, SciEGQA includes three reasoning scenarios with increasing difficulty:

Table 2: Dataset scale and reasoning-type distributions of the SciEGQA benchmark and training set.

Dataset	#Paper	#Page	#QA	SPSR	SPMR	MPMR
Benchmark	80	1,941	1,623	46.15%	34.26%	19.59%
Training Set	3,671	42,380	30,780	37.91%	24.41%	37.69%

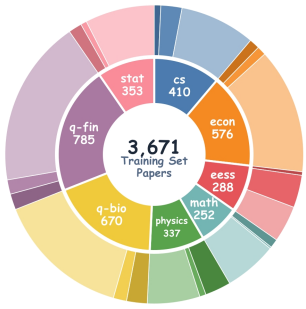
(a) SciEGQA Benchmark: Subject and Publication Type Distribution



Breakdown (Papers)

Subject	Journal	Conference	Preprint	Total
cs	3	2	5	10
econ	0	1	9	10
eess	0	6	4	10
math	3	0	7	10
physics	5	0	5	10
q-bio	2	0	8	10
q-fin	0	3	7	10
stat	4	0	6	10
Total	17	12	51	80

(b) SciEGQA Training Set: Subject and Publication Type Distribution



Breakdown (Papers)

Subject	Journal	Conference	Preprint	Total
cs	26	84	300	410
econ	41	36	499	576
eess	14	129	145	288
math	35	7	210	252
physics	103	25	209	337
q-bio	82	53	535	670
q-fin	63	57	665	785
stat	53	24	276	353
Total	417	415	2,839	3,671

Figure 3: Subject and publication-type distributions of the source papers. (a) The human-annotated benchmark contains 80 papers, balanced with 10 papers per domain. (b) The automatically constructed training set contains 3,671 papers with broader domain and publication-type coverage.

Single-Page Single-Region (SPSR), **Single-Page Multi-Region (SPMR)**, and **Multi-Page Multi-Region (MPMR)**. Each question uses one or multiple EUs, depending on whether its answer requires a single SR, multiple SRs on one page, or SRs across pages.

The dataset consists of two complementary components: a **human-annotated benchmark** designed for rigorous evaluation and a **large-scale automatically constructed training set** that supports scalable model training.

3.3 Data Collection

SciEGQA collects scientific papers from three publication types: **journals**, **conference proceedings**, and **preprint repositories**. This multi-source composition extends the dataset beyond a single publication venue, ensuring broad coverage across different publication stages and document types.

As illustrated in Figure 3 and Table 2, the dataset spans eight research domains: computer science, economics, electrical engineering and systems science, mathematics, physics, quantitative biology, quantitative finance, and statistics.

For the benchmark, domain experts manually selected **10 representative papers from each domain**. The selection covers diverse document layouts and publication types, including single-column and double-column formats, as well as journal articles, conference papers, and preprints, to comprehensively represent the common organizational structures and formatting styles of scientific literature. This process resulted in a balanced benchmark consisting of **80 papers**, with equal representation across all eight domains.

For the automatic construction of the training set, we collected **3,671 papers** from the same research domains and publication types. Unlike the benchmark, the training corpus preserves the natural distribution of the source data to further improve its scale, diversity, and representativeness.

Figure 3 summarizes both the domain distribution and publication-type distribution of the dataset. Specifically, the benchmark contains **17 journal articles, 12 conference papers, and 51 preprints**, while the training corpus consists of **417 journal articles, 415 conference papers, and 2,839 preprints**.

During preprocessing, all PDF pages are rendered as PNG images at **300 dpi**. This resolution preserves fine-grained visual details, including small text, mathematical symbols, and thin table or figure structures, while keeping the computational and storage costs of region segmentation manageable. We apply the same resolution to all documents to ensure consistent rasterization throughout the annotation and construction pipelines.

3.4 Fine-grained Benchmark Construction

As illustrated in the upper part of Figure 4, we construct a fine-grained benchmark by manually annotating the **80 sampled scientific papers** described in Section 3.3.

During the annotation process, **12 annotators** with backgrounds in document understanding or related fields participate in the data construction. In addition, **three domain experts for each research field** are involved in question design and annotation verification, resulting in a total of **24 domain experts**.

Question-independent SR annotation. Before any questions are written, annotators inspect each page and draw bounding boxes around candidate SRs. This order prevents region boundaries from being tailored to a particular QA instance.

Annotators apply the completeness and minimality criteria defined above. A box is expanded when a caption, title, note, or explanatory text is required, and is tightened or split when it includes unrelated content.

Cross-page content is never enclosed by one box. Each complete unit on a page is annotated as a separate SR, allowing later questions to combine multiple SRs without sacrificing localization precision.

Furthermore, to preserve semantic completeness, we intentionally exclude fragments that cannot independently convey complete semantics, such as cross-page text fragments and table content whose interpretation depends on missing context. While this design inevitably discards a small amount of cross-page information, it substantially reduces evidence ambiguity, improves annotation

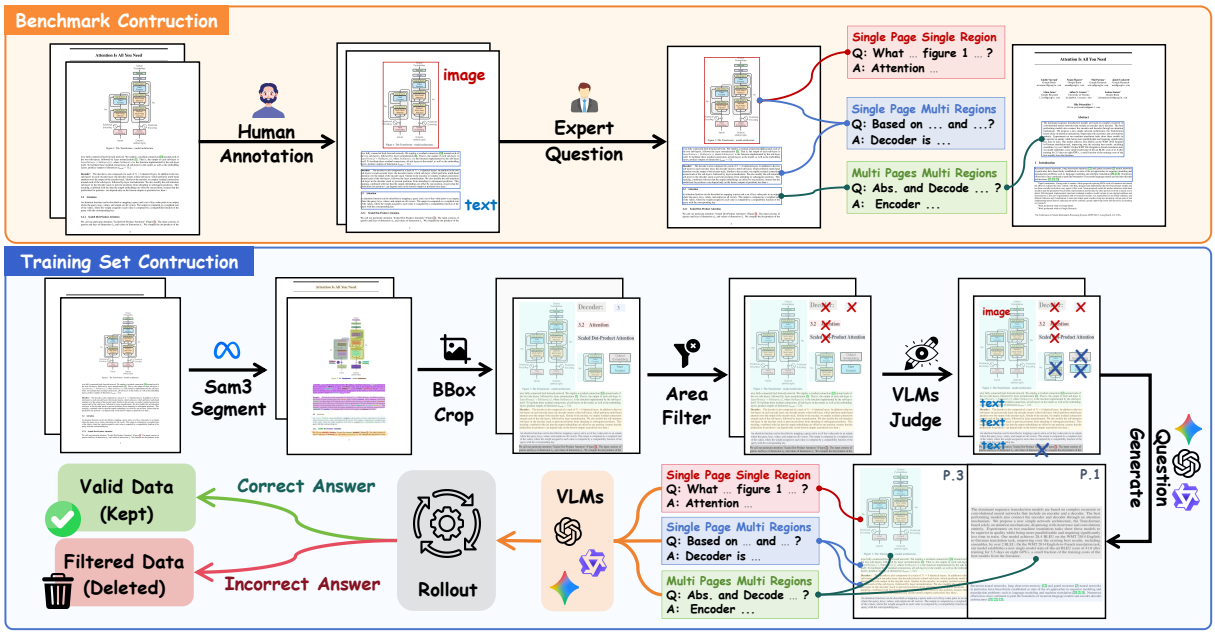


Figure 4: Overview of SciEGQA dataset construction. The dataset is built through two complementary processes: (top) benchmark construction with human annotation and (bottom) large-scale training set construction via an automated pipeline.

consistency, and leads to a more reliable benchmark for evidence grounding evaluation.

Boundary agreement and adjudication. Each paper is independently annotated by two annotators. Matched SR boxes are compared using Intersection over Union (IoU) to measure boundary agreement.

If the IoU is below **0.9**, a third annotator reviews both the boundary and semantic completeness of the region. The adjudicated boxes form the finalized SR pool for the document.

Evidence-unit selection and QA authoring. After the SR pool is finalized, domain experts design questions and select the SRs needed to answer each one. The selected SRs become the EUs for that QA instance.

The questions cover three reasoning scenarios: **SPSR**, **SPMR**, and **MPMR**, with an approximate ratio of **5:3:2**. Each question must be answerable *solely and completely* from its EUs, ensuring a clear correspondence between the answer and its supporting SRs.

This process separates reusable SR annotation from question-specific EU selection. The resulting benchmark supports evaluation of both *document question answering* and *evidence region localization*.

3.5 Automatic Training Set Construction

To construct a large-scale training set, we design an automated data generation pipeline, as illustrated in the lower part of Figure 4. Unlike the benchmark dataset, which is entirely constructed through human annotation, the training set is generated automatically to substantially increase the dataset scale while maintaining reasonable data quality.

Region Segmentation. For each document page, we use **SAM3** [6] with its default segmentation settings to generate candidate visual segments directly from the rendered page image. SAM3 serves as a category-agnostic proposal generator rather than a document parser.

Because it operates on visual content, SAM3 preserves both non-textual information and layout cues without imposing predefined document categories. These cues provide a stronger basis for assessing whether candidate segments form complete semantic units.

The resulting segments are not treated as SRs directly. They are proposals to be evaluated and, when necessary, combined in the next stage under the same completeness and minimality criteria used for human annotation.

Semantic Region Formation. We first filter the segmentation proposals according to their relative page area, removing candidates that occupy less than **5%** or more than **70%** of the page. This step eliminates degenerate regions that are excessively small, as well as overly large segments that are unlikely to represent valid semantic units. These thresholds are selected through small-scale manual validation to achieve a reasonable balance between retaining meaningful regions and removing anomalous proposals. We then apply semantic filtering with random strong vision–language models and retain proposals that already constitute complete and minimal semantic units.

Spatially adjacent proposals are combined when they are semantically dependent, such as a figure and its caption or a table title and body. Proposals that remain incomplete or noisy are discarded, and unrelated columns are never grouped.

The output of this stage is an automatically constructed SR pool. Thus, SAM3 supplies visual boundaries, while semantic filtering and grouping determine which boundaries satisfy the SR definition.

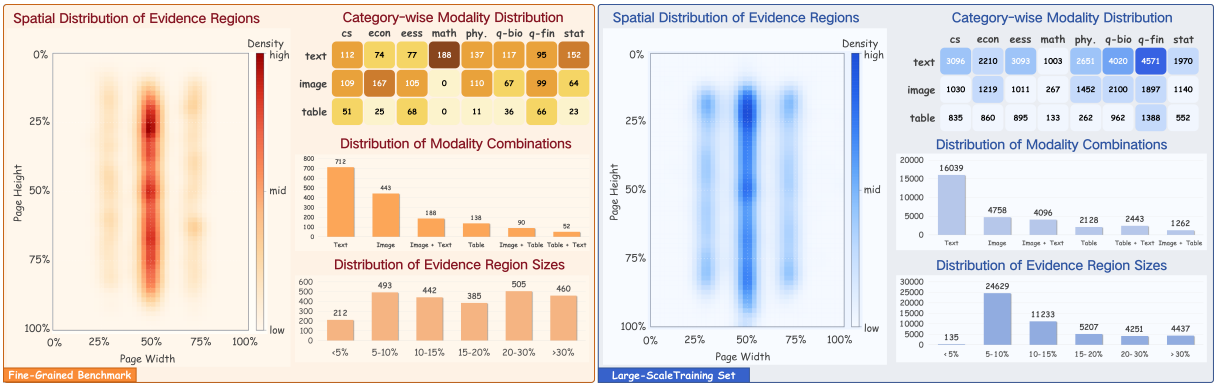


Figure 5: Dataset statistics. Comparison of the fine-grained benchmark (left) and the automatically constructed training set (right). The figure shows distributions across category–modality combinations, question and answer lengths, spatial locations of evidence bounding boxes, modality combinations, and region sizes, illustrating the diversity of document types, evidence modalities, and spatial layouts in the dataset.

Evidence-set Construction. After obtaining the SR pool, the automated pipeline samples one or more semantic regions according to their spatial distribution and semantic relationships to construct candidate evidence sets, covering three reasoning scenarios: single-region, single-page multi-region, and multi-page multi-region reasoning. Once assigned to a QA instance, the selected SRs become the corresponding Evidence Units (EUs), forming **SPSR**, **SPMR**, or **MPMR** evidence sets.

Question–Answer Generation. Given the selected EUs, we employ both open-source and proprietary vision–language models with strong visual reasoning capabilities, including **Qwen3-VL**, **GPT-5**, and **Gemini 3**, to generate question–answer pairs. During generation, the models are explicitly instructed to construct questions and answers solely based on the provided EUs, ensuring that each answer is grounded in the corresponding evidence. Depending on the composition of the evidence set, the generated questions are categorized into the three reasoning types: **SPSR**, **SPMR**, and **MPMR**, covering single-region, single-page multi-region, and multi-page multi-region reasoning.

Automatic Answer Verification. To reduce noise introduced during automatic data generation, we perform an additional answer verification stage. Given the selected EUs and the generated question, two other vision–language models independently generate answers without access to the original answer. A QA instance is retained only if the verified answers are consistent with the original answer and the question can be answered solely from the selected EUs.

Human Quality Audit. To assess the quality of the automatically constructed training set, we randomly sample 500 QA instances together with their associated evidence regions for manual review. Each instance is evaluated on four criteria: evidence-region validity, question clarity, answer correctness, and evidence sufficiency. Two annotators independently review each instance, with disagreements resolved by a third annotator. Inter-annotator agreement is measured using Cohen's κ.

The automatically generated samples achieve pass rates of 96.8%, 97.6%, 96.2%, and 95.4% on the four criteria, respectively, with an overall acceptance rate of 91.8%. Compared with the manually annotated benchmark (93.4%), the difference is only 1.6 percentage points (95% CI: $[-4.8, 1.6]$), and the inter-annotator agreement reaches $\kappa = 0.86$.

These results indicate that the automatically generated training data approaches the quality of manual annotations and provides reliable supervision for large-scale model training.

3.6 SciEGQA Statistics

The main statistics of the SciEGQA dataset are summarized in Table 2. SciEGQA consists of a manually annotated *fine-grained benchmark* and a large-scale automatically constructed *training set*. The benchmark contains **1,623 question–answer pairs** with precise evidence region annotations, while the training set is generated through an automated data construction pipeline and includes approximately **30K question–answer pairs** to support large-scale model training.

Figure 5 presents the statistical distributions of the dataset across multiple dimensions. The results show that the evidence annotations in SciEGQA cover diverse visual elements, including text blocks, figures, and tables, and exhibit a diverse spatial distribution across the page layout.

Overall, these statistics indicate that solving the SciEGQA task requires models not only to perform document-level question answering, but also to identify semantically relevant evidence regions within complex scientific documents and conduct reasoning across multiple regions or even multiple pages. This makes SciEGQA the first scientific document QA dataset that simultaneously supports *semantic region grounding* and *multi-page reasoning*.

4 Experiment Setup

4.1 Evaluation Metrics

We employ **Answer Accuracy** as the primary evaluation metric. Additionally, **Intersection over Union (IoU)** is utilized to assess the precision of predicted bounding boxes.

Table 3: Main results on the SciEGQA benchmark. Task 1 evaluates visual grounding via the Grounding–Crop–then–Answer pipeline, while Task 2 evaluates QA accuracy under different evidence granularities. Bold and underline denote the best and second-best results; ↑ indicates accuracy gains over the previous granularity. More details are available on our project webpage.

	Task 1: Grounding-Crop-then-Answer						Task 2: Evidence-Granularity-QA		
Model	Mean IoU	IoU@0.3	IoU@0.5	IoU@0.7	Valid Ratio	Acc.	Document	Page	Region
Qwen3-VL-8B	21.43%	24.77%	12.14%	7.27%	87.06%	32.04%	9.30%	60.51%$^{\uparrow 51.20}$	65.87%$^{\uparrow 5.36}$
Qwen3-VL-32B	31.53%	44.55%	18.55%	11.95%	98.15%	46.46%	16.14%	73.14%$^{\uparrow 56.99}$	75.54%$^{\uparrow 2.40}$
Qwen3-VL-235B-A22B	25.22%	33.52%	10.72%	4.87%	88.66%	38.63%	40.30%	70.49%$^{\uparrow 30.19}$	75.42%$^{\uparrow 4.93}$
Qwen3.5-27B	36.04%	47.50%	<u>29.39</u>%	19.90%	99.82%	<u>51.26</u>%	42.95%	69.69%$^{\uparrow 26.74}$	77.02%$^{\uparrow 7.33}$
Qwen3.5-122B-A10B	30.76%	38.08%	21.87%	14.11%	92.48%	46.95%	37.58%	68.95%$^{\uparrow 31.36}$	73.01%$^{\uparrow 4.07}$
Qwen3.5-397B-A17B	**38.89**%	52.19%	**32.96**%	**20.64**%	99.51%	50.15%	36.23%	68.08%$^{\uparrow 31.85}$	73.75%$^{\uparrow 5.67}$
InternVL3-38B	5.23%	3.39%	0.62%	0.12%	89.03%	5.05%	10.23%	48.18%$^{\uparrow 37.95}$	64.14%$^{\uparrow 15.96}$
GLM-4.6V-Flash	19.90%	21.26%	9.67%	5.85%	77.94%	35.92%	50.16%	72.03%$^{\uparrow 21.87}$	77.94%$^{\uparrow 5.91}$
Ernie-5	14.57%	19.10%	9.55%	4.00%	84.90%	22.12%	38.51%	77.39%$^{\uparrow 38.88}$	81.89%$^{\uparrow 4.50}$
Kimi-K2.5	7.17%	7.16%	5.49%	2.23%	96.30%	7.46%	**68.02**%	<u>79.98</u>%$^{\uparrow 11.95}$	81.21%$^{\uparrow 1.23}$
Claude–4.6–Sonnet	15.93%	17.25%	4.81%	0.68%	93.84%	19.29%	43.99%	68.45%$^{\uparrow 24.46}$	74.31%$^{\uparrow 5.85}$
GPT-5.2	31.63%	49.11%	18.55%	4.81%	<u>99.88</u>%	41.22%	<u>65.43</u>%	**81.39**%$^{\uparrow 15.96}$	**85.15**%$^{\uparrow 3.76}$
Qwen3-VL-8B-SFT	<u>38.66</u>%	**57.92**%	28.77%	17.25%	**99.95**%	49.97%	32.87%	69.43%$^{\uparrow 36.56}$	76.89%$^{\uparrow 7.46}$
GLM-4.6V-Flash-SFT	37.23%	<u>54.16</u>%	27.05%	14.23%	99.81%	**52.80**%	63.09%	78.37%$^{\uparrow 15.28}$	<u>85.10</u>%$^{\uparrow 6.73}$

Evidence grounding. Valid Ratio is the proportion of predicted page–box pairs that can be parsed with a valid page index and coordinates satisfying $0 \le x_1 < x_2 \le 1000$ and $0 \le y_1 < y_2 \le 1000$. For each sample, predicted and ground-truth boxes are grouped by page. On page p, we compute their pairwise IoUs and use dynamic programming to find the one-to-one maximum-IoU matching

$$M_p^* = \arg\max_{M_p} \sum_{(i,j)\in M_p} \mathrm{IoU}(P_{p,i}, G_{p,j}). \tag{1}$$

Each box can be matched at most once. We normalize by

$$d_x = \max\left(\sum_p |P_p|, \sum_p |G_p|, 1\right).$$

The sample score is

$$\mathrm{MeanIoU}(x) = \frac{1}{d_x} \sum_p \sum_{(i,j)\in M_p^*} \mathrm{IoU}(P_{p,i}, G_{p,j}), \tag{2}$$

$$\mathrm{IoU@}\,\tau(x) = \mathbb{I}[\mathrm{MeanIoU}(x) \ge \tau], \tag{3}$$

where $\tau \in \{0.3, 0.5, 0.7\}$. Unmatched predictions and references contribute zero through d_x, penalizing both redundant and missing boxes. Mean IoU is averaged over samples with successfully parsed predictions; IoU@τ is averaged over all samples, with parse failures counted as zero.

Answer accuracy. Exact string matching is inadequate because many questions require interpretation. Five distinct LLM judges independently compare the question, reference answer, and model answer, each casting a binary vote $v_k \in \{0, 1\}$.

An answer is correct when at least three judges vote positively:

$$\mathrm{Acc}(x) = \mathbb{I}\left[\sum_{k=1}^{5} v_k \ge 3\right]. \tag{4}$$

The judges accept semantically equivalent paraphrases and numerical forms, but reject contradictory or incomplete answers. Answer Accuracy is the mean of $\mathrm{Acc}(x)$ over all samples.

4.2 Task Definitions

We design **two evaluation tasks** on the SciEGQA benchmark to systematically evaluate a diverse set of Vision-Language Models (VLMs), using **zero-shot inference** for base models and separately evaluating task-specific SFT checkpoints.

Task 1: Grounding–Crop–then–Answer. Given the ground-truth evidence page(s) and the corresponding question, the model is required to predict a set of **bounding boxes** that localize the regions relevant to answering the question. This oracle-page setting isolates region localization from page retrieval. Each bounding box is represented as (x_1, y_1, x_2, y_2), with coordinates normalized to the page coordinate space of $[0, 1000]$.

Based on the predicted bounding boxes, we crop the corresponding regions from the page images and feed the cropped regions into a fixed question answering model, **Qwen3.5-27B**, to generate the final answer. Since all methods share the same answering model, this pipeline controls for differences in reasoning capability across models. We evaluate the grounding quality of the predicted regions using both **IoU** and the resulting **answer accuracy**.

Task 2: Evidence-Granularity-QA. We first evaluate model performance under different levels of input granularity. Specifically, we provide the model with three types of inputs: (1) the **entire document**, (2) the **evidence page(s)** containing the answer, and (3) the **cropped evidence region(s)**. By comparing answer accuracy across these different input settings, we analyze how different evidence granularities influence the question answering capability of VLMs.

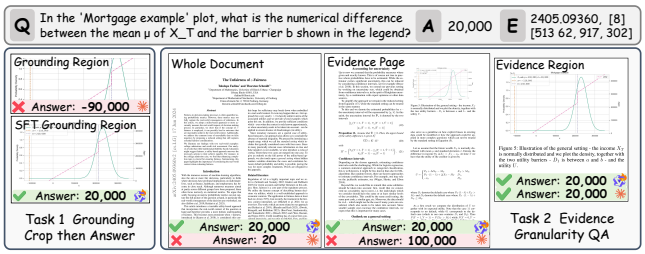


Figure 6: An SPSR case study of Claude–4.6–Sonnet and Qwen3-VL-8B-SFT on Task 1 Grounding–Crop–then–Answer and Task 2 Evidence–Granularity–QA.

4.3 Evaluation Suite

We evaluate a comprehensive range of state-of-the-art VLMs, including both open-source and closed-source models. This includes the **Qwen3-VL** [2] series (8B, 32B, and the 235B–A22B variant) and the **Qwen3.5** [2] series (27B, 122B–A10B, and 397B–A17B). Furthermore, we evaluate **InternVL3-38B** [40], **Ernie-5** [29], **Kimi-K2.5** [26], **Claude-4.6-Sonnet** [1] and **GPT-5.2** [22].

To evaluate the effectiveness of the training dataset, we perform supervised fine-tuning on **Qwen3-VL-8B** for two tasks separately using the **LLaMA-Factory** framework [37], and additionally fine-tune **GLM-4.6V-Flash** on the same SciEGQA training set. Their task-specific checkpoints are denoted as **Qwen3-VL-8B-SFT** and **GLM-4.6V-Flash-SFT**, respectively. Task 1 retains the fixed Qwen3.5-27B answerer; Task 2 evaluates each model's own QA performance.

Models with fewer than 38B parameters run locally, while larger models are accessed via official APIs. For fair comparison, all models are evaluated using greedy decoding without sampling, and each question is answered with a single inference pass.

5 Results and Analysis

Table 3 presents the performance of various VLMs on the SciEGQA benchmark. Overall, current state-of-the-art models still struggle to simultaneously achieve reliable evidence localization and accurate question answering in scientific documents, indicating that SciEGQA remains a challenging benchmark.

Grounding Capability. The left part of Table 3 evaluates visual grounding and reasoning using the **grounding-crop-then-answer** pipeline. Most models achieve relatively low IoU scores, with the best model reaching only **38.89%** Mean IoU, indicating that precise evidence localization in scientific documents remains challenging. Since answering scientific questions often requires reasoning over specific document regions, inaccurate grounding can easily disrupt the reasoning process. Moreover, QA accuracy generally improves with better IoU, suggesting that inaccurate evidence localization is a key factor limiting model performance.

Error Decomposition. Our controlled evaluation separates errors caused by evidence localization from those caused by downstream reasoning. When ground-truth evidence regions are provided, the fixed answering model Qwen3.5-27B achieves 77.02% accuracy. In contrast, the best Grounding–Crop–then–Answer result reaches only 52.80%, leaving a 24.22-point gap associated with imperfect localization and cropping. Meanwhile, the remaining 22.98% error under ground-truth regions indicates that precise evidence alone does not eliminate failures in scientific interpretation and reasoning. Thus, current models are limited by both evidence localization and evidence-conditioned reasoning, rather than by either component alone.

Effect of Evidence Granularity. The right part of Table 3 reports QA accuracy under three evidence granularities. Across these three input settings, **more focused evidence leads to higher QA accuracy**. When the entire document is provided as input, all models achieve relatively low accuracy, indicating that locating relevant evidence within large-scale context remains challenging for current models. These results highlight the importance of precise evidence localization in scientific document QA, as it enables models to focus on relevant context and perform more reliable reasoning. Further analysis shows that multi-page, multi-region questions generally remain more challenging. Appendix D.4 additionally compares SR inputs with OCR text and layout-block images.

Effectiveness of the Training Data. After fine-tuning on the SciEGQA Training Set, Qwen3-VL-8B-SFT achieves **38.66%** Mean IoU and **49.97%** QA accuracy under the grounding–crop–then–answer evaluation. Meanwhile, under **Evidence–Granularity-QA**, the model also shows consistent improvements across all evidence granularities compared with the base model. GLM-4.6V-Flash also improves after SFT: Mean IoU rises from 19.90% to 37.23%, and Task 1 QA accuracy from 35.92% to 52.80%, with gains under all three Task 2 input granularities. These results support the effectiveness of SciEGQA supervision beyond the Qwen family.

Case Study. Figure 6 illustrates how evidence granularity and grounding accuracy affect QA performance. In **Task 1**, inaccurate grounding leads to incorrect cropped regions and consequently wrong answers. After fine-tuning on SciEGQA, Qwen3-VL-8B-SFT correctly localizes the evidence region and produces the correct answer, while Claude–4.6–Sonnet fails. In **Task 2**, Claude–4.6–Sonnet fails under document- and page-level inputs due to distracting context, but succeeds when the precise evidence region is provided. This example highlights both the challenging nature of SciEGQA and the effectiveness of the training data.

6 Conclusion

In this paper, we introduce **SciEGQA**, a dataset for visual evidence-grounded question answering on scientific documents, which provides **semantic region-level evidence annotations** in the form of bounding boxes. The dataset consists of a human-annotated benchmark and a large-scale automatically constructed training set, and supports multiple reasoning scenarios in scientific documents. Experimental results show that current Vision-Language Models still struggle with accurate evidence localization and evidence-grounded reasoning in scientific documents. We hope SciEGQA can serve as a challenging benchmark to promote future research on evidence-grounded multimodal reasoning and scientific understanding in complex documents. Meanwhile, the SciEGQA training set can serve as a valuable resource for improving the reasoning capabilities of Vision-Language Models, including future reinforcement learning–based optimization.

7 Ethical Considerations

SciEGQA is constructed from publicly accessible scientific papers and does not intentionally collect private or sensitive data. To respect copyright, source documents are redistributed only when permitted by their licenses; otherwise, we release derived annotations and source identifiers. SciEGQA is intended for research on scientific document understanding and evidence grounding. Its outputs should not be treated as verified scientific claims or used for high-stakes decisions without expert review.

References

[1] Anthropic. 2026. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-03-21.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] https://arxiv.org/abs/2511.21631

[3] Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. 2025. Peerqa: A scientific question answering dataset from peer reviews. 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)*. 508–544.

[4] Lukas Blecher, Guillem Cucurull, Thomas Scialom, Robert Stojnic, and M Ai. [n. d.]. Nougat: Neural optical understanding for academic documents, 2023. *arXiv preprint arXiv:2308.13418* ([n. d.]).

[5] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024* (2024), 13590–13618.

[6] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025).

[7] 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*. 143–152.

[8] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4599–4610.

[9] Simone Giovannini, Fabio Coppini, Andrea Gemelli, and Simone Marinai. 2025. Boundingdocs: a unified dataset for document question answering with spatial annotations. *arXiv preprint arXiv:2501.03403* (2025).

[10] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. 2025. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271* (2025).

[11] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*. 4083–4091.

[12] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*. PMLR, 19036–19052.

[13] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231* (2024).

[14] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan S Kankanhalli. 2024. Improving Context Understanding in Multimodal Large Language Models via Multimodal Composition Learning. In *ICML*, Vol. 3. 7.

[15] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15630–15640.

[16] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems* 37 (2024), 95963–96010.

[17] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244* (2022).

[18] 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*. 1697–1706.

[19] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2200–2209.

[20] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems* 37 (2024), 118807–118833.

[21] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries* 23, 3 (2022), 289–301.

[22] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267* (2025).

[23] Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 20908–20923.

[24] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633.

[25] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13878–13888.

[26] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. 2026. Kimi K2.5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276* (2026).

[27] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition* 144 (2023), 109834.

[28] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19528–19540.

[29] Haifeng Wang, Hua Wu, Tian Wu, Yu Sun, Jing Liu, Dianhai Yu, Yanjun Ma, Jingzhou He, Zhongjun He, Dou Hong, et al. 2026. ERNIE 5.0 Technical Report. *arXiv preprint arXiv:2602.04705* (2026).

[30] Jui-I Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2025. Mesaqa: A dataset for multi-span contextual and evidence-grounded question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*. 10891–10901.

[31] Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. 2024. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 8296–8305.

[32] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems* 37 (2024), 113569–113697.

[33] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

[34] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (2024), nwae403.

[35] Tan Yue, Xuzhao Shi, Rui Mao, Zilong Song, Zonghai Hu, and Dongyan Zhao. 2025. AnaFig: A Human-Aligned Dataset for Scientific Figure Analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 12837–12843.

[36] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. 2025. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630* (2025).

[37] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*. 400–410.

[38] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362* (2025).

[39] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4857–4866.

[40] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring

advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).

A Benchmark Construction Details

A.1 Document Selection and Split Integrity

SciEGQA contains 80 scientific papers and 1,941 rendered pages. We sample 10 papers from each of eight domains: computer science, economics, electrical engineering and systems science, mathematics, physics, quantitative biology, quantitative finance, and statistics.

The collection comprises 17 journal articles, 12 conference papers, and 51 preprints, with both single- and double-column layouts. Before annotation, each paper is checked for stable page rendering and sufficient scientific content for evidence-grounded QA.

Each released record retains the document identifier, domain, question, reference answer, evidence-page indices, absolute and normalized bounding boxes, and evidence-modality labels. Page images are organized by domain and document identifier to preserve provenance.

We include publicly accessible papers from the target domains that contain verifiable scientific claims, experiments, derivations, quantitative analyses, or structured figures and tables. Papers are excluded if rendering fails, key content is missing or illegible, or the material is non-research or largely duplicates another candidate.

We also perform version-level deduplication. Potential arXiv, conference, and journal versions are matched using titles, authors, abstracts, and DOIs, followed by manual comparison of the main text and section structure. For the same study, we retain only the public version with the most complete results and stable layout.

All splits are isolated at the document level rather than the question, page, or evidence-region level. Test papers do not overlap with documents used for training, prompt construction, or development, and different versions of the same work are never assigned to different splits.

Finally, we check for contamination using title–author matching and text-similarity retrieval. Test QA pairs are manually compared with public QA datasets. Evidence-page and region identifiers are also checked for uniqueness and traceability.

A.2 Semantic Region Annotation Guidelines

The benchmark is constructed without automatically generated questions or answers. Annotators first identify Semantic Regions (SRs) without reference to any question. Domain experts then write questions and associate the required SRs with each answer. This ordering prevents region boundaries from being tailored to a particular question.

For each page, annotators identify a semantic anchor, such as a paragraph, list, figure, table, or equation, and apply three operational checks. The *completeness check* expands a candidate when a caption, title, note, legend, or explanatory sentence is necessary for interpretation. The *minimality check* removes adjacent material that can be excluded without changing the meaning of the unit. The *separation check* prevents unrelated columns or neighboring components from being merged.

Text SRs generally correspond to complete paragraphs or lists. Figure SRs include the figure and the caption or legend required to interpret it. Table SRs include the title, body, and necessary notes, while equation SRs include the surrounding text needed to identify the variables and the role of the equation. Cross-page evidence is represented by multiple page-local SRs, each preserving a complete semantic unit; one bounding box is never allowed to cross pages.

Table 4: Example of semantic-region boundary decisions.

Candidate	Action	Reason
Plot only	Expand	The caption is required to identify the setting.
Plot and caption	Retain	The unit is complete and minimal.
Plot, caption, and neighboring paragraph	Shrink	The paragraph discusses an unrelated experiment.

Table 4 gives a representative boundary decision. It illustrates that geometric tightness alone is insufficient: a plot-only box may be visually compact but semantically incomplete when its caption defines the experimental setting.

A.3 QA Authoring and Evidence Linking

After the SR pool is finalized, a domain expert selects a scientific fact or relationship to query, writes a concise reference answer, and links every SR required to derive that answer. An SR selected for a particular question becomes an Evidence Unit (EU). The expert verifies that the question has one unambiguous answer, every selected EU is necessary, no required evidence is omitted, and the answer does not depend on unselected content or external knowledge.

Questions supported by one EU are labeled Single-Page Single-Region (SPSR). Questions requiring multiple EUs on the same page are labeled Single-Page Multi-Region (SPMR), while questions requiring EUs from different pages are labeled Multi-Page Multi-Region (MPMR). Thus, the reasoning type is determined by the finalized evidence set rather than by surface cues in the question. The benchmark contains 46.15% SPSR, 34.26% SPMR, and 19.59% MPMR instances.

The released annotation uses the following core fields: `query` and `answer` store the QA pair; `doc_name` and `category` identify the source paper; `evidence_page` records the relevant pages; `rel_bbox` stores normalized coordinates; and `subimg_type` records whether an EU is text, a table, or an image. Bounding boxes are grouped first by evidence page and then by region, preserving the structure required for SPMR and MPMR instances.

A.4 Annotation Agreement and Adjudication

Each benchmark paper is independently processed by two annotators. Corresponding SR proposals are matched and their Intersection over Union (IoU) is computed. Cases with IoU below 0.9 are sent to a third annotator for adjudication. The adjudicator determines whether the disagreement is caused by missing context, excessive surrounding content, or an incorrect merge or split, and the adjudicated boundary replaces the two initial proposals in the final SR pool.

Question–answer instances are also checked for answer correctness and evidence sufficiency. In particular, the reviewer confirms

that the selected EUs are jointly sufficient and that the reference answer can be derived without consulting any unselected region.

B Automatic Training-Set Construction Details

Implementation details of the automated construction pipeline are available in our GitHub repository; here, we clarify several key aspects of training-set construction.

B.1 Pipeline and Model Configuration

The automatic construction pipeline processes each paper in five stages: (1) PDF rasterization, (2) candidate-region segmentation, (3) semantic-region screening and deduplication, (4) question–answer generation and independent verification, and (5) conversion to the released JSONL format. All PDF pages are rendered as PNG images at 300 dpi before visual processing.

Following the procedure described in the main text, we use SAM3 to generate candidate masks with the default checkpoint configuration and inference parameters provided by its official implementation, and subsequently convert these masks into rectangular proposals. Before semantic screening, proposals whose bounding-box area covers less than 5% or more than 70% of the page are removed.

For each sample, the VLM responsible for semantic-region screening and question–answer generation is randomly selected from a predefined pool consisting of `Qwen3-VL-235B-A22B`, `GPT-5`, and `Gemini-3`. The selected model checks whether each crop forms a complete and minimal semantic unit and assigns it to one of three modality categories: `text`, `table`, or `image`. During independent verification, the other two VLMs independently answer the generated question using only the selected EUs, without access to the generated answer. We retain a sample only when both verification answers are consistent with the generated answer and the evidence is sufficient, as described in the main text.

B.2 Prompt Templates

These templates cover region screening and region-aware QA generation. Bracketed expressions specify the evidence scope; image tokens and serialization details are omitted.

Crop-screening template

Determine whether the supplied crop is a complete and sufficiently fine-grained semantic region. Set `keep=true` only if it contains one complete logical unit, includes the context required for interpretation, and excludes unrelated content. Reject blank, illegible, truncated, or mixed-content crops. Classify the primary modality as `text`, `table`, or `image`. Return only `{"keep": true/false, "type": "..."}`; if uncertain, set `keep=false`.

Region-aware QA-generation template

You are given one or more evidence regions extracted from a scientific document. Region boundaries are represented by one of the following scopes:
`[SPSR: page` p`; regions` r_1`]`
`[SPMR: page` p`; regions` $r_1, \ldots, r_n$`]`
`[MPMR: page` p_1`; regions` $r_1, \ldots, r_i$`;`
`...; page` p_m`; regions` $r_j, \ldots, r_n$`]`
Use only the content inside the specified scope. Generate three diverse question–answer pairs that match its structure. For SPSR, one region must be sufficient. For SPMR, the answer must require multiple regions on the same page. For MPMR, it must require evidence from multiple regions across multiple pages. Do not generate a multi-region question if any single region is sufficient.
Prefer questions involving comparisons, numerical trends, relationships, hierarchy, or explicit scientific statements. For each pair, return `question`, a concise `answer`, `region_scope`, an answer type from `[span, number, date, yesno, categorical]`, and a difficulty label from `[easy, medium, hard]`. If the evidence does not support a valid pair, return `NO_VALID_QA`.

B.3 Filtering and Failure Handling

The construction code applies conservative filtering at each stage. Candidate crops that fail the area constraint are removed before any VLM call. Crop-screening output is parsed as strict JSON; malformed or unrecognized outputs default to `keep=false` and are assigned the `unknown` modality rather than being retained.

After semantic screening, highly overlapping crops of the same known modality are deduplicated. For two regions whose intersection covers at least 90% of the smaller region, the smaller crop is retained for text, reflecting minimality, while the larger crop is retained for tables and figures, reflecting completeness. Overlapping regions with different or unknown modality labels are preserved to avoid deleting potentially valid evidence.

If a batch fails during screening or generation, the pipeline processes its images individually where possible. Unreadable images, missing crop files, invalid boxes, and unusable evidence crops are skipped and recorded. The parser accepts tagged, JSON-like, and line-based QA variants, removes duplicate questions, and discards entries without both a question and an answer. Raw generation outputs are retained alongside parsed records for later inspection.

C Evaluation and Implementation Details

C.1 Input Formatting and Inference Prompts

The two main benchmark tasks operate on page images, not OCR transcripts or parsed text. Each instance uses a ShareGPT-style multimodal format containing image tokens, an instruction, and a question. The code builds four views from each annotation: `document` supplies every page in document order; `page` supplies the annotated evidence pages; `crop` supplies the ground-truth evidence regions; and `bbox` supplies the evidence pages and asks the model to predict normalized region coordinates.

QA prompt

Answer the question using only the document image(s). Return only the final answer with no explanation.

Grounding prompt

Locate the region or regions needed to answer the question in each document image. Return only JSON grouped by page in the format

$[[[x_1, y_1, x_2, y_2], \ldots], \ldots]$. Each outer item corresponds to one input image or page in order, and every coordinate is normalized to $[0, 1000]$.

For Task 1, predicted regions are cropped from the supplied evidence pages and passed, together with the original question, to the fixed Qwen3.5-27B answering model. For Task 2, the same QA instruction is used for document-, page-, and region-level inputs so that the primary controlled variable is evidence granularity.

C.2 Grounding Output Parsing and Region Cropping

The evaluator first attempts to parse grounding output as JSON and also supports JSON enclosed in code fences, Python-style list literals, coordinate strings, and dictionaries with common keys such as `bbox` or `coordinates`. Page-grouped predictions are matched only against references on the corresponding page. When a flat prediction contains exactly one box for each page of a multi-page single-box example, it is promoted to page structure in input order; otherwise, the evaluator applies global flat matching and records this fallback mode. An output that cannot be parsed is counted as a parse failure and contributes zero to IoU-threshold rates; Mean IoU is averaged over successfully parsed samples.

Within each page, predicted and reference boxes are matched one-to-one to maximize the sum of IoUs. The sample denominator is the larger of the number of predicted and reference boxes, so both missing and redundant regions reduce the score. IoU@0.3, IoU@0.5, and IoU@0.7 report the fractions of all benchmark samples whose normalized sample-level scores reach the corresponding thresholds.

For the crop–then–answer stage, normalized coordinates are mapped back to source-image dimensions and clamped to valid boundaries. The released reconstruction utility rejects crops whose width or height is smaller than 28 pixels and does not generate answers for samples with no usable crop. Such samples remain in the Task 1 QA denominator of 1,623 and are counted as incorrect. Multiple valid crops are retained as separate images and are presented in page order and then region order.

C.3 LLM-as-a-Judge Protocol

Strict string matching cannot reliably evaluate answers that allow semantically equivalent expressions, alternative numerical forms, or different wording. We therefore adopt an LLM-as-a-Judge protocol for automatic answer evaluation.

Each judge receives the complete question, reference answer, and candidate response, and independently returns a binary decision. The instruction used by the released evaluation utility is shown below.

LLM-as-a-Judge prompt

Evaluate whether the model response correctly answers the question based on the reference answer. The output must be strictly either `correct` or `incorrect`. Accept semantically equivalent answers and reject responses that contradict the reference, omit required information, introduce a substantive error, or fail to answer the question. Return only one word.

For the final evaluation, we use five distinct LLMs as independent judges to reduce bias from any single evaluator: `DeepSeek-V4`, `Gemini-3`, `GPT-5`, `Claude-4.6`, `Qwen3-235B-A22B`.

All judges receive identical inputs and use the same evaluation instruction and generation configuration, ensuring comparability across evaluators.

Let $v_k(x) \in \{0, 1\}$ denote the decision of judge k for sample x, where 1 indicates correct and 0 indicates incorrect. Sample-level correctness is determined by majority vote:

$$\mathrm{Acc}(x) = \mathbb{I}\left[\sum_{k=1}^{5} v_k(x) \geq 3\right]. \tag{5}$$

Because five judges are used, ties cannot occur. Dataset-level Answer Accuracy is the mean of these binary sample-level labels.

To validate the protocol, we randomly sample 200 test instances for human review. Human annotators independently assess each candidate response, and we compare their decisions with the LLM-judge majority vote.

The two evaluations agree on 184 of 200 instances, yielding an Agreement of **92.0%** and Cohen's κ of **0.802**. This strong agreement indicates that the automatic protocol reliably approximates human evaluation.

C.4 Supervised Fine-Tuning Details

We use LLaMA-Factory to fine-tune Qwen3-VL-8B and GLM-4.6V-Flash on the same SciEGQA training set. For each model, we train two separate checkpoints: one for grounding and one for QA. These task-specific checkpoints are denoted as Qwen3-VL-8B-SFT and GLM-4.6V-Flash-SFT, respectively, in the main paper. Both model families use LoRA-based parameter-efficient fine-tuning with the same training settings. Training is performed with the AdamW optimizer and a cosine learning-rate scheduler, using an initial learning rate of 1×10^{-3}. Each checkpoint is trained for 3 epochs with an effective batch size of 64 on 8 NVIDIA A100 GPUs with 80GB memory each.

Training examples follow the same multimodal message format and task instructions used during evaluation. Optimization settings and data formatting are kept consistent across both model families and the grounding and QA tasks. For Task 1, each grounding checkpoint predicts regions that are passed to the fixed Qwen3.5-27B answerer; Task 2 evaluates the corresponding QA checkpoint under document-, page-, and region-level inputs.

D Additional Results and Analysis

This section provides complementary analyses of the two benchmark tasks. We examine how localization quality propagates to answer accuracy, how the amount of visual context affects QA, and how these effects vary across the three reasoning structures in SciEGQA.

D.1 Grounding Quality and Answer Accuracy

For the models shown in Figure 7, grounding quality and final answer accuracy are tightly coupled, with a Pearson correlation of $r = 0.98$. The predicted boxes directly determine the visual evidence passed to the fixed QA model.

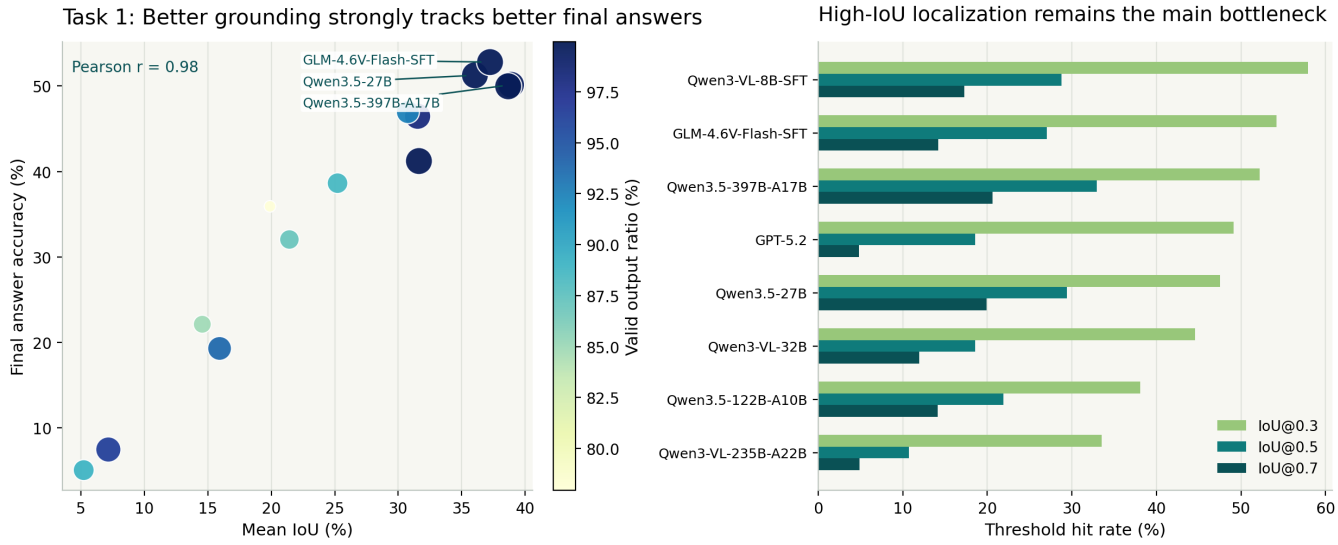


Figure 7: Task 1 grounding analysis. Left: mean IoU is strongly correlated with downstream answer accuracy across evaluated models; marker color indicates the valid-output ratio. Right: localization hit rates at IoU thresholds of 0.3, 0.5, and 0.7.

Because the downstream answering model and its prompt are fixed for all grounding systems, this comparison isolates the practical effect of the predicted crops. Improvements in end-to-end accuracy therefore depend not only on whether a model emits valid coordinates, but also on whether those coordinates preserve the evidence needed to answer the question.

When a predicted region deviates from the reference evidence, its crop may omit necessary content or introduce irrelevant information. The resulting evidence corruption can therefore turn an otherwise answerable question into an incorrect prediction.

The sharp decline from IoU@0.3 to IoU@0.7 further shows that precise localization remains substantially harder than coarse localization. Current models often identify the general area but fail to recover a semantically complete region boundary.

The valid-output ratio and IoU capture two distinct failure modes. A syntactically valid box may still exclude a caption, legend, table note, or explanatory sentence, while a malformed output prevents a usable crop from being constructed at all. The threshold results consequently distinguish broad evidence discovery from boundary-level semantic completeness.

D.2 Effect of Evidence Granularity

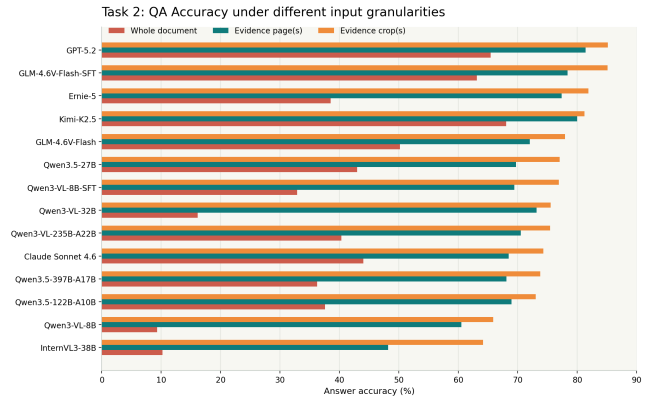


Figure 8: Task 2 answer accuracy under whole-document, evidence-page, and evidence-crop inputs. More focused evidence consistently improves QA accuracy across the evaluated models.

As shown in Figure 8, QA accuracy increases as the input narrows from the whole document to the evidence page or pages and finally to the evidence crop or crops. This pattern is consistent across the evaluated models.

The three settings form a controlled continuum of evidence access. Whole-document input requires the model to search the complete paper before reasoning; evidence-page input removes cross-page retrieval but retains unrelated material on the selected pages; evidence-crop input provides the annotated semantic regions directly.

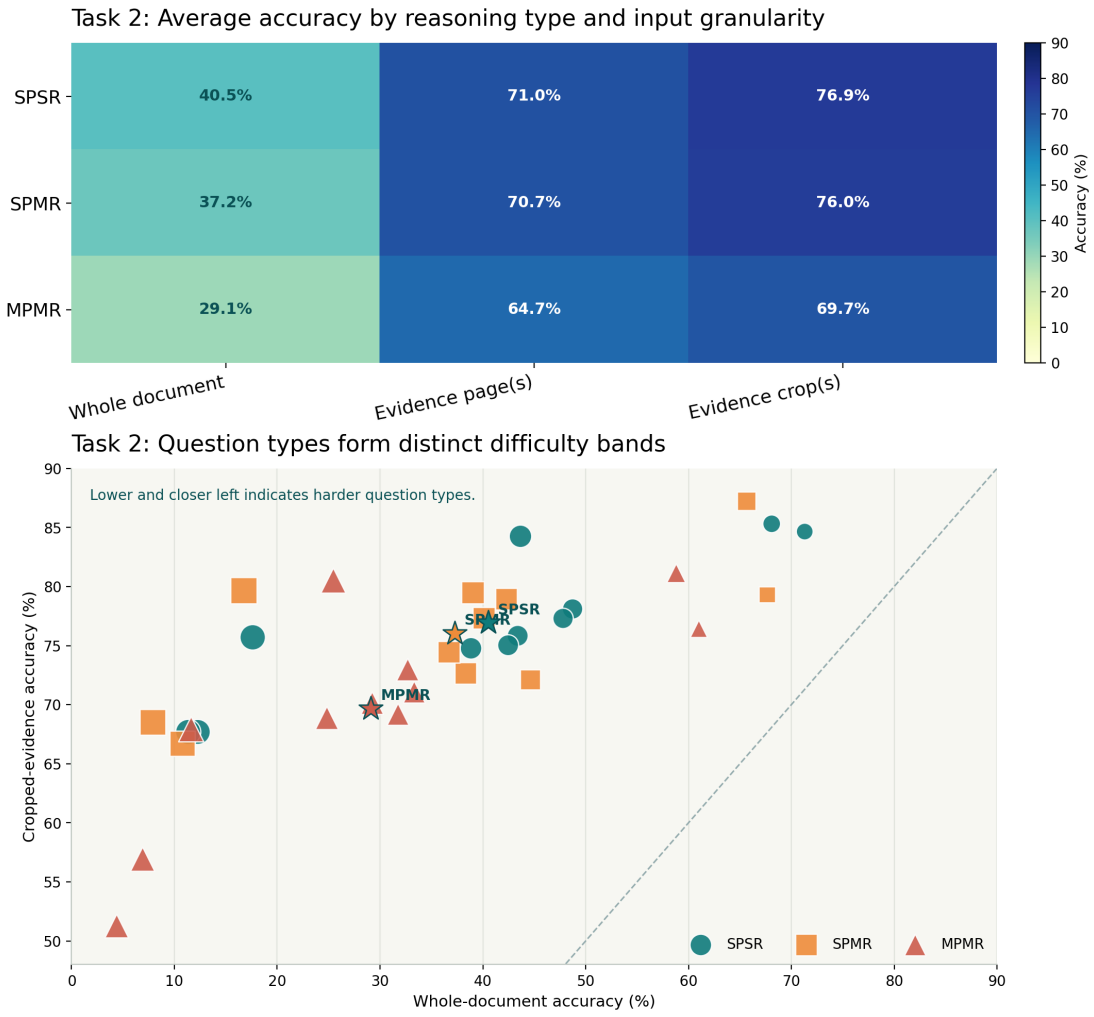


Figure 9: Task 2 results by reasoning type for the 11 models with available per-type scores; GLM and SFT models are excluded. Top: average accuracy for SPSR, SPMR, and MPMR under three input granularities. Bottom: per-model whole-document and cropped-evidence accuracy, with stars marking the average of each reasoning type.

Reducing irrelevant context allows a model to attend more directly to the information needed for reasoning. The result also explains why grounding quality is consequential: accurate crops provide the strongest QA input, whereas inaccurate crops can remove that advantage.

The monotonic trend across different model families indicates that sensitivity to irrelevant visual context is not confined to a particular architecture or model scale. At the same time, accuracy remains below perfect even with ground-truth crops, showing that precise evidence access does not eliminate errors in scientific interpretation, comparison, and numerical reasoning.

D.3 Results by Reasoning Type

Figure 9 shows distinct difficulty bands. Mean accuracy with evidence crops decreases from 76.9% for SPSR to 76.0% for SPMR and 69.7% for MPMR as evidence aggregation becomes more complex.

The same ordering holds for whole-document input (40.5%, 37.2%, and 29.1%) and evidence-page input (71.0%, 70.7%, and 64.7%), showing that MPMR is especially sensitive to irrelevant context.

SPSR and SPMR become close with oracle pages or crops, suggesting that within-page integration is manageable after localization. MPMR remains lower because evidence must be preserved and combined across page boundaries.

Thus, the three categories separately test localized, within-page, and cross-page evidence integration and should not be treated as a homogeneous QA set.

Together, these analyses identify three bottlenecks: valid evidence localization, distraction control, and cross-region integration. SciEGQA therefore requires joint progress in grounding and reasoning.

Table 5: QA accuracy (%) for alternative evidence representations on the balanced 300-question subset. Each reasoning type contains 100 questions.

Input condition	SPSR	SPMR	MPMR	Overall
Evidence-page images	86	83	85	84.67
Gold SR images	**92**	**89**	**90**	**90.33**
Page OCR text	66	55	42	54.33
Region OCR text	71	60	44	58.33
All layout-block images	79	78	57	71.33
Gold-overlap layout blocks	89	84	85	86.00

D.4 Comparison with Other Fine-Grained Evidence Representations

We compare six evidence representations on the same 300-question diagnostic subset, with 100 questions each from SPSR, SPMR, and MPMR. Sampling is stratified by modality within each reasoning type and spreads questions across source papers. This balanced subset differs from the full benchmark distribution. All conditions use ground-truth evidence pages as their source and Qwen3.5-27B as the fixed answerer, with a common QA instruction and generation settings. PaddleOCR supplies OCR text, while PP-StructureV3 supplies layout-block images. The gold-overlap condition retains blocks satisfying $|B \cap G|/\min(|B|, |G|) \geq 0.5$ for a ground-truth evidence box G and parsed block B; crop boundaries remain those of the parser. This supplementary comparison uses a single DeepSeek-V4-Flash judge, with all 300 questions retained in the accuracy denominator and failed predictions counted as incorrect.

Table 5 shows that SR inputs achieve the highest accuracy in all three reasoning types, exceeding full evidence pages by 5.66 percentage points and gold-overlap layout blocks by 4.33 points overall. Selecting relevant layout blocks improves accuracy over providing all blocks, indicating that finer decomposition alone does not ensure better QA. These results support SRs as an effective evidence granularity under this protocol. The differences also reflect OCR information loss, parsing quality, and input-budget failures, rather than isolating semantic completeness alone.