BBox-DocVQA: A Large-Scale Bounding-Box–Grounded Dataset for Enhancing Reasoning in Document Visual Question Answer

Wenhan Yu *
Baidu Inc.
Beihang University
yuwenhan@buaa.edu.cn

Wang Chen*
Baidu Inc.
The University of Hong Kong
wchen22@connect.hku.hk

Guanqiang Qi
Baidu Inc.
qiguanqiang@baidu.com

Weikang Li
Peking University
wavejkd@pku.edu.cn

Yang Li †
Baidu Inc.
liyang164@baidu.com

Lei Sha
Beihang University
shalei@buaa.edu.cn

Deguo Xia
Baidu Inc.
xiadeguo@baidu.com

Jizhou Huang
Baidu Inc.
huangjizhou01@baidu.com

Abstract

Document Visual Question Answering (DocVQA) is a fundamental task for multimodal document understanding and a key testbed for vision–language reasoning. However, most existing DocVQA datasets are limited to the page level and lack fine-grained spatial grounding, constraining the interpretability and reasoning capability of Vision–Language Models (VLMs). To address this gap, we introduce **BBox-DocVQA**—a large-scale, bounding-box–grounded dataset designed to enhance spatial reasoning and evidence localization in visual documents. We further present an automated construction pipeline, **Segment–Judge–and–Generate**, which integrates a segment model for region segmentation, a VLM for semantic judgment, and another advanced VLM for question–answer generation, followed by human verification for quality assurance. The resulting dataset contains **3.6 K** diverse documents and **32 K** QA pairs, encompassing single- and multi-region as well as single- and multi-page scenarios. Each QA instance is grounded on explicit bounding boxes, enabling fine-grained evaluation of spatial–semantic alignment. Benchmarking multiple state-of-the-art VLMs (e.g., GPT-5, Qwen2.5-VL, and InternVL) on BBox-DocVQA reveals persistent challenges in spatial grounding and reasoning accuracy. Hence, our dataset can contribute to advancing research on interpretable and spatially grounded vision–language reasoning.

1 Introduction

Document Visual Question Answering (DocVQA) has emerged as a crucial benchmark for advancing multimodal document understanding, requiring models to reason over textual, structural, and

*Co-First Authors
†Corresponding author

Dataset / benchmark	#Queries	#Docs	#Images	#I/Q	Avg. Q Len	Avg. Ans Len	BBox	MP
DUDE Van Landeghem et al. [2023]	41 K	5 K	28.7 K	1.00	8.65	3.35	Yes	No
BoundingDocs Giovannini et al. [2025]	249 K	48 K	237 K	1.00	–	–	Yes	No
SlideVQA Tanaka et al. [2023]	14.5 K	–	52 K	1.26	–	–	Yes	Yes
SPIQA Pramanick et al. [2024]	270 K	25.9 K	270 K	1.00	12.98	14.56	–	No
PlotVQA Methani et al. [2020]	29 M	–	224 K	1.00	43.54	–	–	No
ArXivQA Li et al. [2024]	100 K	16.6 K	32 K	1.00	16.98	7.59	–	No
ChartQA Masry et al. [2022]	33 K	–	22 K	1.00	13.18	1.08	–	No
TAT-DQA Zhu et al. [2022]	16 K	2.7 K	3 K	1.07	12.54	3.44	No	No
VisualMRC Tanaka et al. [2021]	30 K	10 K	10 K	1.00	10.55	9.55	No	No
InfographicsVQA Mathew et al. [2022]	30 K	5.4 K	5.4 K	1.00	11.54	1.60	No	No
DocVQA Mathew et al. [2021]	50 K	6 K	12 K	1.00	9.49	2.43	No	No
MP-DocVQA Tito et al. [2023]	46 K	6 K	48 K	8.27	9.90	2.20	No	Yes
BBox-DocVQA (ours)	32 K	3.6 K	44K	1.38	10.80	2.78	Yes	Yes

Table 1: Comparison between the BBox-DocVQA dataset and the existing DocVQA datasets. Docs, I, Q, Ans, BBox, and MP denote documents, images, queries, answers, bounding boxes, and multiple pages, respectively.

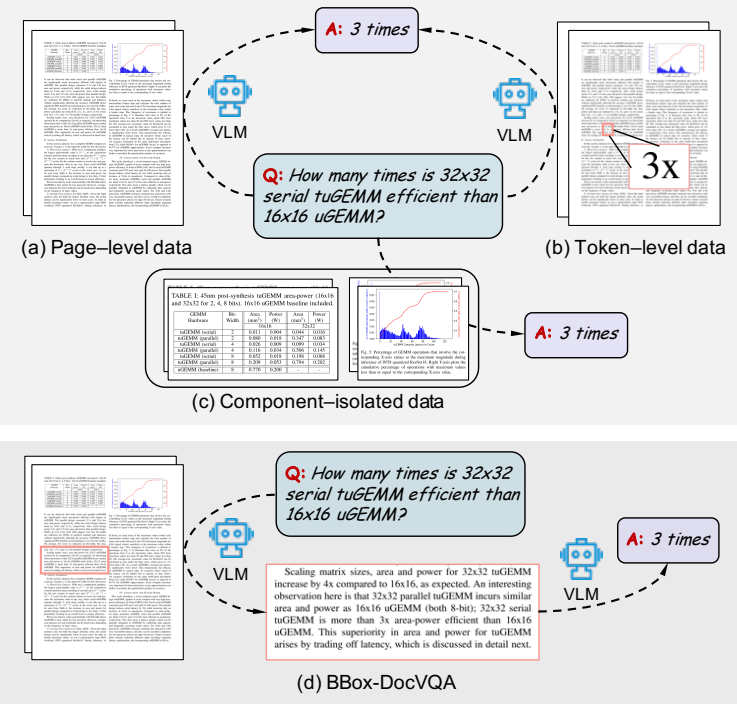

Figure 1: Comparison among (a) page–level data, (b) token–level data, (c) component–isolated data, and (d) our proposed BBox-DocVQA.

visual information jointly. Recent progress in vision–language models (VLMs) [Alayrac et al., 2022, Radford et al., 2021, Bai et al., 2025, Yang et al., 2025, Achiam et al., 2023] has greatly improved their ability to process and comprehend complex document layouts. However, the majority of existing DocVQA datasets, such as DocVQA [Mathew et al., 2021], MP-DocVQA [Tito et al., 2023], and VisualMRC [Tanaka et al., 2021], are constructed at the *page level*. Correspondingly, most existing methods Chen et al. [2023], Ma et al. [2024a], Hu et al. [2025], Chen et al. [2022], Lin et al. [2023], Zhang et al. [2024], Qi et al. [2024a], Tian et al. [2025], Qi et al. [2024b], Cho et al. [2024], Wang et al. [2025a], Li et al. [2025], Wasserman et al. [2025], Faysse et al. [2025], Chen et al. [2025b], Yu et al. [2025], Wang et al. [2025b], Chen et al. [2025c] focus on generating the final answer based on one or multiple document pages, without explicit modeling of where the evidence originates. This absence of spatial grounding leads to reduced interpretability and may degrade the

generation accuracy of VLMs, as models can yield seemingly correct answers while attending to irrelevant or incomplete regions.

Human document comprehension typically involves two steps: first, identifying the relevant pages, and second, locating specific *evidence components*—such as text paragraphs, tables, or figures—that support the final answer. Current page-level DocVQA datasets (as shown in Figure 1(a)), however, only support the first step, offering page-level evidence without bounding-box annotations. On the other hand, a few token-level datasets (as shown in Figure 1(b)), such as DUDE [Van Landeghem et al., 2023] and BoundingDocs [Giovannini et al., 2025], provide OCR–based token-level bounding boxes to localize answers. While these datasets allow precise token grounding, they lack contextual semantic integrity, as each annotation typically corresponds to an isolated word or phrase rather than a complete semantic unit. Also, as shown in Figure 1(c), a few component–isolated datasets have been proposed to enhance the visual understanding ability of VLMs, but still lack contextual semantic information. As a result, existing DocVQA resources cannot fully capture the fine-grained reasoning process required for spatially interpretable document understanding.

To bridge these gaps, as shown in Figure 1(d), we introduce **BBox-DocVQA** (Bounding-Box–grounded Document Visual Question Answering), the first large-scale DocVQA dataset that explicitly annotates evidence bounding boxes across both single- and multi-page settings. Each question–answer (QA) pair is linked to one or multiple bounding boxes representing coherent semantic regions (paragraphs, tables, or figures) that provide the necessary evidence for answering the query. In addition, we develop an automated dataset construction pipeline, named **Segment–Judge–and–Generate**, which combines segmentation, visual–semantic evaluation, and automatic QA generation. Specifically, our pipeline first employs a segmentation model (SAMKirillov et al. [2023]) to detect visual and textual regions, then uses a VLM to judge the effectiveness of the segments, and finally prompts an advanced VLM to generate the corresponding QA pairs. This pipeline enables scalable and consistent dataset generation while maintaining semantic diversity and spatial precision.

We conduct extensive experiments on the proposed BBox-DocVQA dataset using a range of state-of-the-art VLMs, including GPT-5 Achiam et al. [2023], Qwen-VL Bai et al. [2025], Yang et al. [2025], and the InternVL Chen et al. [2024] series. These models are evaluated in terms of both spatial grounding accuracy (bounding-box localization) and answer correctness. Experimental results reveal that even advanced VLMs often struggle to accurately identify the true evidence regions, underscoring a persistent weakness in spatial grounding and interpretability. The main contributions of this paper are summarized as follows:

1. We introduce **BBox-DocVQA**, the first large-scale bounding-box–grounded DocVQA dataset, which covers 3.6 K diverse visual documents and 32 K QA pairs, encompassing both single- and multi-page, as well as single- and multi-region scenarios.
2. We propose an automated dataset construction pipeline, **Segment–Judge–and–Generate**, integrating segmentation, visual–semantic judgment, and large VLM-based QA generation, ensuring scalability and annotation consistency.
3. We conduct extensive benchmarking experiments with multiple state-of-the-art VLMs, demonstrating that the existing VLMs cannot accurately predict spatial grounding regions.
4. The BBox-DocVQA dataset provides a powerful supervision signal for strengthening evidence localization and spatial reasoning in VLMs, enabling both improved answer accuracy and more interpretable decision pathways.

2 Related Work

Visual document understanding has been widely explored through diverse dataset paradigms, which differ in annotation granularity, component coverage, and interpretability. In this section, we review three major types of document visual question–answering datasets—*token-level*, *component-*

isolated, and *page-level*—and clarify how our proposed **BBox-DocVQA** bridges their existing limitations. A detailed statistical comparison is provided in Table 1.

Token-level Datasets. A branch of VQA datasets, such as DUDE Van Landeghem et al. [2023] and BoundingDocs Giovannini et al. [2025], and SlideVQA Tanaka et al. [2023], provides token-level annotation, where bounding boxes are aligned with OCR tokens and serve as ground-truth evidence for question answering. These datasets directly associate answers with OCR-extracted tokens, thereby facilitating fine-grained text localization. However, they often lack contextual semantic information, since each answer is limited to a single token or phrase rather than a semantically complete region. Consequently, token-level datasets may fail to capture the hierarchical and compositional nature of document understanding.

Component–isolated Datasets. Another line of research focuses on component–isolated visual question answering, targeting specific structured elements such as tables, charts, and figures. Representative examples include SPIQA Pramanick et al. [2024], PlotQA Methani et al. [2020], ChartQA Masry et al. [2022], and ArXivQA Li et al. [2024]. These datasets generate massive numbers of QA pairs by isolating document components and synthesizing questions from them. Such datasets can enhance the reasoning ability of vision–language models (VLMs) over structured visual content. Nonetheless, their task formulation deviates from DocVQA: they do not capture inter-component or cross-page reasoning, and the generated QA pairs typically remain confined to single visual objects rather than full document contexts.

Page-level Datasets. To address holistic understanding, page-level DocVQA datasets annotate question–answer pairs at the document or page level, requiring models to comprehend both text layout and visual structure. For example, the DocVQA Mathew et al. [2021] dataset and its multi-page extension MP-DocVQA Tito et al. [2023] comprise 50 K questions over 12 K single-page documents and 46 K questions spanning 6 K multi-page documents, respectively. Similarly, VisualMRC Tanaka et al. [2021], InfographicsVQA Mathew et al. [2022], and TAT-DQA Zhu et al. [2022] extend document-level comprehension to specific domains. Also, a few benchmarks, including OK-VQA Marino et al. [2019], A-OKVQA Schwenk et al. [2022], WebQA Chang et al. [2022], UDA Hui et al. [2024], Dyn-VQA Li et al. [2025], MMLongBench Ma et al. [2024b], REAL-MM-RAG Wasserman et al. [2025], ViDoRe Faysse et al. [2025], ViDoSeek Wang et al. [2025a], M3DoCVQA Cho et al. [2024], OpenDocVQA Tanaka et al. [2025], VisR-Bench Chen et al. [2025a], and VisDoMBench Suri et al. [2025], have been proposed to evaluate the VLMs' understanding and reasoning capabilities. These datasets or benchmarks advance layout-aware VQA; however, they do not provide explicit bounding-box annotations for evidence localization, limiting explainability and making it difficult to evaluate spatial grounding.

To fill this gap, we introduce **BBox-DocVQA**, a large-scale, bounding-box–grounded dataset for visual document question answering. Unlike prior datasets, BBox-DocVQA explicitly annotates evidence bounding boxes, enabling fine-grained evaluation of both *spatial grounding* and *semantic correctness*. It further extends to multi-page scenarios and includes questions whose answers depend on multiple bounding boxes across different pages. As shown in Table 1, BBox-DocVQA combines the interpretability of token-level datasets with the holistic reasoning of page-level benchmarks, achieving a balance between accuracy, scalability, and explainability. This design not only establishes a new benchmark for VLM spatial reasoning but also enables model evaluation on interpretable, evidence-based document-understanding tasks.

3 BBox-DocVQA Dataset

In this section, we introduce the collection and construction details of BBox-DocVQA, along with its statistics.

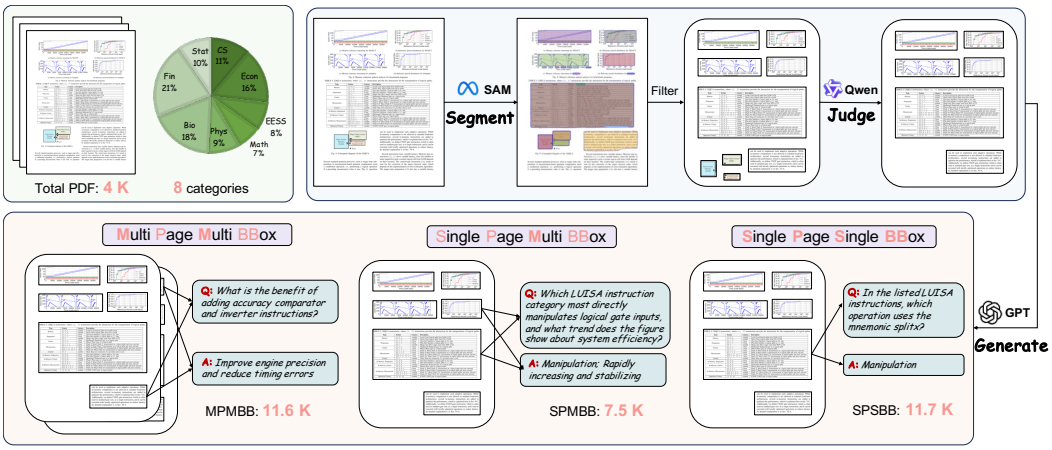

Figure 2: An overview of the collected dataset and the proposed pipeline. We collect 4K PDFs across eight academic domains, segment page regions using SAM, filter and validate them with Qwen, and generate grounded QA pairs with GPT. We consider three annotation granularities—MPMBB, SPMBB, and SPSBB—covering multi-page, multi-region, and single-region reasoning.

3.1 Segment-Judge-and-Generate Framework

Existing approaches based on either standalone Vision-Language Models (VLMs) or conventional Computer Vision (CV) methods struggle to accurately perform fine-grained structure extraction from documents. These methods often fail to produce bounding boxes (bboxes) that are both precise and semantically complete. To address this issue, we propose an automated *Segment-Judge-and-Generate* framework to efficiently construct a large-scale, bounding-box-grounded document question answering (DocQA) dataset. The core idea is to first detect fine-grained semantic components (*segments*) within document pages, and then generate reasoning-oriented question–answer (QA) pairs conditioned on these localized regions. This enables an explicit alignment between visual content, spatial layout, and semantic reasoning. The overall pipeline and case study are illustrated in Figure 2.

Data Collection. We collected approximately **4,000** PDF documents from the *arXiv* platform, spanning the years 2023 to 2025, with a total of about **137,000** pages. The corpus covers eight major academic domains, including Computer Science, Mathematics, Physics, Biology and other disciplines, representing all top-level categories in arXiv. To facilitate subsequent visual analysis and multimodal reasoning, each page was converted into a high-resolution image to preserve layout structure, textual organization, and figure/table details.

Segment Components. We adopt the Segment Anything Model Kirillov et al. [2023] (SAM, ViT-H variant) to automatically detect layout components on each page image. The model first produces multiple semantic masks, which are then converted into compact bounding boxes. To remove trivial or redundant regions, we filter the bounding boxes based on their area ratio relative to the page size, retaining only those with coverage between 5% and 70%. This effectively discards tiny noisy fragments and full-page detections.

Judge and Classify Components. The detected regions typically correspond to independent semantic units such as paragraphs, tables, or figures, but overlapping or ambiguous boundaries may still occur. To ensure higher quality and semantic consistency, we employ **Qwen2.5-VL-72B** to evaluate each cropped region in terms of completeness, information density, and visual cleanliness, and to predict its primary content type (`text`, `table`, or `image`). Furthermore, to eliminate redundant detections, we design an overlap-based deduplication strategy: if the overlapping area between two regions exceeds 90% of the smaller one, we retain different regions depending on their type —

for `text`, we preserve the smaller region to capture fine-grained semantics; for `table` and `image`, we keep the larger region to ensure contextual and structural integrity.

Question–Answer Generation. After obtaining high-quality fine-grained regions, we employ **GPT-5** to generate diverse and reasoning-oriented QA pairs. We first perform page-level summarization to provide contextual background, ensuring semantic consistency across regions. The generated QA samples are categorized into three types: *(1) Single-Page Single-BBox (**SPSBB**)*, *(2) Single-Page Multi-BBox (**SPMBB**)*, and *(3) Multi-page Multi-BBox (**MPMBB**)*, covering a wide range of document-level reasoning scenarios. For single-region cases, the model receives the page summary and the cropped sub-image, and generates QA pairs based strictly on visible content. For multi-region cases, the model first evaluates the semantic relevance among sub-images and only produces questions when meaningful relationships are detected, thereby avoiding incoherent or trivial QAs. Each valid region is instructed to yield one QA pair through multi-step reasoning and factual grounding. Every QA sample contains a concise question and an accurate answer.

In summary, the **Segment-Judge-and-Generate** framework provides an efficient and scalable automated solution for constructing fine-grained, spatially grounded, and semantically rich document QA datasets, laying a solid foundation for region-level DocQA model training and evaluation.

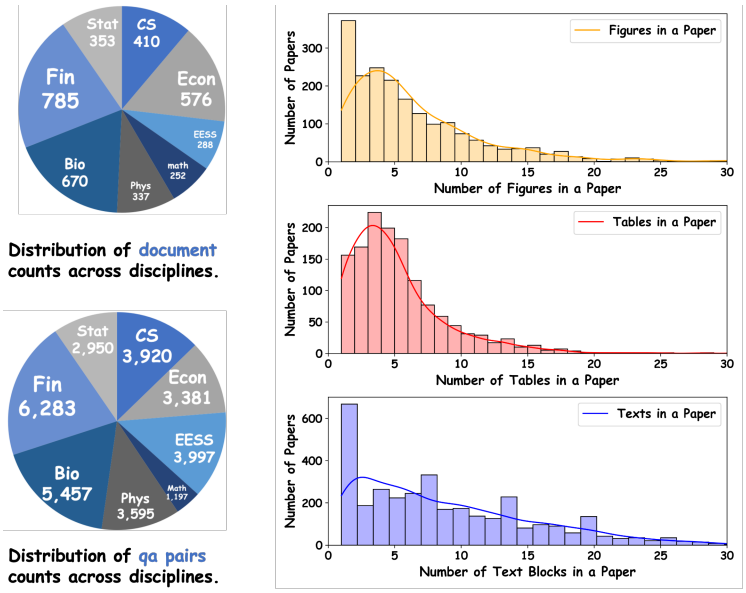

Figure 3: The left panels show the proportions of papers across research domains for (top) the document dataset and (bottom) the generated QA pairs. The right panels display the distributions of the number of figures, tables, and text blocks per paper.

3.2 Fine-grained Benchmark

To further evaluate the effectiveness of our proposed **Segment-Judge-and-Generate** framework, we construct a high-quality manually annotated fine-grained benchmark, referred to as the **Fine-grained Benchmark**.

Data Sources and Selection. We randomly selected **10 PDF documents** from each of the eight major academic domains on the *arXiv* platform (Computer Science, Mathematics, Physics, Quantitative Biology, Quantitative Finance, Electrical Engineering and Systems Science , Economics, and Statistics), resulting in a total of **80** research papers as the benchmark corpus. During the initial filtering phase, we conducted manual inspection to ensure that the sampled documents exhibit sufficient diversity and representativeness in layout structure, page count, content type (text, table, and image), and subject area.

Manual Annotation and Verification. Domain experts were invited to participate in the annotation process. Each page was manually cropped to precisely delineate every semantic unit with bounding boxes (bboxes). To guarantee the reliability and consistency of the annotations, we adopted a multi-stage verification protocol: each sample was independently annotated by at least two experts and subsequently reviewed by a third annotator to resolve inconsistencies, yielding a set of high-confidence bounding boxes.

Question Generation and Quality Control. After obtaining the verified bbox regions, we employed the state-of-the-art multimodal large model **GPT-5** to generate question–answer (QA) pairs for each region. The generation process strictly follows the same protocol as in the training data, covering three QA formats: *Single-Page Single-BBox (SPSBB)*, *Single-Page Multi-BBox (SPMBB)*, and *Multi-Page Multi-BBox (MPMBB)*. Human experts further conducted multiple rounds of validation to assess the logical soundness of questions, the factual correctness of answers, and the overall reasoning quality, ensuring that each QA pair is interpretable, accurate, and challenging.

This benchmark covers a wide variety of document structures and semantic types, providing a comprehensive evaluation resource for testing the performance and generalization ability of Vision–Language Models (VLMs) on fine-grained document understanding tasks.

3.3 Data Statistics and Analysis

This section summarizes the key statistics of the BBox-DocVQA dataset, including the automatically constructed training set and the manually curated fine-grained benchmark. By presenting the data distribution, region composition, and domain coverage in a concise tabular form, we highlight the structural alignment between the two subsets and their respective roles in model training and evaluation.

3.3.1 Training Set Statistics

The training set contains 30,780 automatically generated QA pairs derived from 42,380 pages of 3,671 arXiv papers. As summarized in Table 3.3.2 and visualized in Figure 3, the dataset features a balanced distribution across the three task types (SPSBB, SPMBB, MPMBB), along with a rich mixture of text, image, and table regions dominated by textual content. The scientific domain coverage naturally follows arXiv's real-world submission distribution, while the long-tailed variation in the number of figures, tables, and text blocks per paper reflects authentic heterogeneity in scientific document layouts. These characteristics collectively provide a large-scale and structurally diverse corpus suitable for training models on multi-region and multi-modal document understanding.

3.3.2 Benchmark Statistics

The Fine-grained Benchmark consists of 1,623 high-quality manually annotated QA samples spanning 1,941 pages from 80 papers. Table 3.3.2 provides a detailed summary of its composition. Similar to the training set, the benchmark includes all three task types and a multimodal mixture of region categories, but differs in its uniformly sampled domain distribution and precisely curated bounding boxes and QA annotations. The benchmark therefore serves as a reliable evaluation suite with balanced domain coverage, stringent annotation consistency, and a suitable level of task difficulty for assessing fine-grained document understanding.

3.3.3 Comparison and Summary

The training set is large and diverse, making it well suited for learning multimodal document structures and multi-region reasoning patterns. The benchmark, while smaller, offers highly reliable manual annotations and balanced domain sampling, making it ideal for robust and fair performance evaluation. Their aligned distributions in task types and region types ensure consistent training–evaluation conditions and meaningful performance interpretation.

Item	Training Set	Benchmark
Total QA samples	30,780	1,623
Total pages	42,380	1,941
Total papers	3,671	80
Task types		
SPSBB	11,668 (37.91%)	749 (46.15%)
SPMBB	7,512 (24.41%)	556 (34.26%)
MPMBB	11,600 (37.69%)	318 (19.59%)
Region types		
Text	30,424 (60.98%)	1,247 (49.94%)
Image	12,542 (25.14%)	916 (36.68%)
Table	6,926 (13.88%)	334 (13.38%)
Avg. bbox area ratio	14.26%	19.08%

Table 2: Comparison of Training Set and Fine-grained Benchmark Statistics

Models	gt_subpage_answer	gt_page_answer	pages_bbox_answer	Delta 1	Delta 2
Qwen2.5VL-3B	56.38%	45.72%	30.87%	10.66%	**25.51**%
Qwen2.5VL-7B	68.58%	57.18%	51.57%	11.40%	17.01%
Qwen2.5VL-32B	82.01%	73.01%	63.46%	9.00%	18.55%
Qwen2.5VL-72B	79.42%	69.32%	68.64%	10.10%	10.78%
Qwen3VL-4B	72.34%	67.22%	68.95%	5.12%	3.39%
Qwen3VL-8B	74.74%	72.27%	67.59%	2.47%	7.15%
Qwen3VL-32B	**82.01**%	**80.28**%	**77.10**%	1.73%	4.87%
InternVL3-2B	41.59%	32.04%	34.20%	9.55%	7.39%
InternVL3-8B	65.80%	53.42%	50.46%	**12.38**%	15.34%

Table 3: Performance comparison among three task strategies across different models. **Delta 1** and **Delta 2** denote the accuracy differences between **gt_subpage_answer** and **gt_page_answer** and **gt_subpage_answer** and **pages_bbox_answer**, respectively.

4 Experiment

4.1 Experiment setup

Bounding-box-grounded Doc-VQA task. Given a document D and a query q, a retriever $\mathcal{R}$ first fetches relevant pages $P_q = \mathcal{R}(q, D)$. A generator $\mathcal{G}$ then produces the final answer $\hat{a}$ based on a prompt $\mathcal{P}$. We explore three strategies that integrate bounding-box ($\hat{b}$) prediction: (1) **Direct Answer Generation:** The answer is generated directly without explicit localization, (2) **Simultaneous Generation:** The model generates the answer and its bounding box concurrently, and (3) **Sequential Generation:** A two-stage, coarse-to-fine approach involving localization, cropping, and then answer generation. The following equations formally represent these strategies:

$$\hat{a} = \mathcal{G}(\mathcal{P}(q, P_q)), \tag{1}$$

$$\hat{a}, \hat{b} = \mathcal{G}(\mathcal{P}(q, P_q)), \tag{2}$$

$$\begin{aligned} &\hat{b} = \mathcal{G}(\mathcal{P}(q, P_q)), \qquad P_q^{\text{crop}} = \text{CROP}(P_q, \hat{b}), \\ &\hat{a} = \mathcal{G}(\mathcal{P}(q, P_q^{\text{crop}})) \end{aligned} \tag{3}$$

In this study, sequential generation is achieved by deriving the final answer from ground-truth subimages, with bounding boxes $\hat{b}$ annotated in advance.

Evaluation metrics. We evaluate the model performance on both visual grounding and answer generation. For bounding-box prediction, we use the **Intersection over Union (IoU)** as the metric.

Model	Good Ratio	Mean IoU	SPSBB (749)	SPMBB (556)	MPMBB (318)
Qwen2.5VL-3B	51.4%	4.70%	3.80%	5.60%	4.90%
Qwen2.5VL-7B	73.0%	11.30%	14.60%	8.80%	8.00%
Qwen2.5VL-32B	94.0%	20.00%	22.20%	20.40%	14.30%
Qwen2.5VL-72B	**99.0**%	**35.20**%	**40.10**%	**33.20**%	**27.20**%
Qwen3VL-4B	95.87%	18.70%	18.90%	17.00%	21.30%
Qwen3VL-8B	92.7%	14.40%	17.60%	9.20%	15.90%
Qwen3VL-32B	94.5%	20.40%	22.60%	17.30%	21.00%
InternVL3-2B	80.90%	0.10%	0.00%	0.20%	0.10%
InternVL3-8B	97.97%	0.30%	0.10%	0.50%	0.50%
GPT-5	99.9%	0.90%	0.10%	1.60%	1.20%

Table 4: Mean IoU under the simultaneous perception–and–answering setting, reported across SPSBB, SPMBB, and MPMBB task types. **Good Ratio** denotes the ratio of the number of accurate instruction-following samples to the number of all samples.

Model	Mean Acc.	SPSBB (749)	SPMBB (556)	MPMBB (318)
Qwen2.5VL-3B	30.87%	31.38%	33.27%	25.47%
Qwen2.5VL-7B	51.57%	57.01%	53.42%	35.53%
Qwen2.5VL-32B	63.46%	67.69%	66.55%	48.11%
Qwen2.5VL-72B	68.64%	71.03%	71.58%	57.86%
Qwen3VL-4B	68.95%	70.49%	72.12%	59.75%
Qwen3VL-8B	67.59%	70.23%	73.20%	51.57%
Qwen3VL-32B	77.14%	81.04%	**84.35**%	55.35%
InternVL3-2B	34.20%	39.12%	33.27%	24.21%
InternVL3-8B	50.46%	53.14%	53.06%	39.62%
GPT-5	**81.45**%	**82.64**%	83.63%	**74.84**%

Table 5: Answer accuracy under the simultaneous perception–and–answering setting, where the model must localize evidence regions and generate answers in a single unified step. Results are reported for SPSBB, SPMBB, and MPMBB tasks.

Specifically, for **Single-Page Single-BBox (SPSBB)** cases, IoU is computed directly between the predicted and ground-truth boxes; for **Single-Page Multiple-BBox (SPMBB)** cases, we calculate the maximum IoU for each ground-truth box and then average across all boxes; for **Multi-Page Multiple-BBox (MPMBB)** cases, IoU is first computed for each page (following the SPSBB or SPMBB rule) and then averaged across pages. For answer evaluation, we employ a **large language model (LLM) as a semantic judge** to assess whether the generated response is consistent with the reference answer in meaning and context. This LLM-based evaluation captures semantic correctness beyond surface-level text similarity, offering a more reliable measure for open-ended question answering.

Implementation details. All experiments are run on NVIDIA A800 GPUs with a fixed decoding temperature of 0 to ensure deterministic outputs. Answer correctness is evaluated using **DeepSeek-v3.1** as an automatic judge, which compares the predicted answer with the ground truth and decides whether they are semantically consistent.

4.2 Main Results

Table 3 reports answer accuracy under the three task strategies. Across all models, **gt_subpage_answer** yields the highest accuracy, demonstrating that providing the exact evidence region substantially strengthens reasoning. The accuracy gap between **gt_subpage_answer** and **gt_page_answer** (Delta 1) ranges from 1.7 pp to 12.4 pp, confirming that full-page inputs introduce visual noise that weakens answer reliability. Performance drops even further when models must both localize and answer (**pages_bbox_answer**): Delta 2 reaches as high as **25.51 pp** (Qwen2.5VL-3B), indicating that weak grounding severely impairs reasoning. Larger and newer models narrow this gap—e.g., Qwen3VL-32B shows only 4.87 pp difference—but none eliminate it completely. This consistent pattern highlights a key conclusion: accurate evidence localization is a prerequisite for stable and reliable document reasoning, and current VLMs still lack robust spatial–semantic alignment.

Table 4 presents the visual grounding results under the simultaneous perception–and–answering setting. Qwen2.5VL-72B achieves the highest localization performance with a **35.20%** mean IoU across all task types, while Qwen3VL models obtain moderate IoU scores (14–20%). On the contrary, InternVL and GPT-5 exhibit very low IoU despite producing reasonable answers in some cases, likely because their internal image preprocessing (e.g., resizing) disrupts precise coordinate prediction (see Appendix for analysis). When broken down by task type, all models follow a similar trend: grounding is easiest in the **SPSBB** setting, becomes less reliable in the **SPMBB** setting where multiple regions must be matched, and degrades further in the **MPMBB** setting where reasoning spans multiple pages. This progressive decline highlights the increasing difficulty of locating coherent semantic units as the spatial search space grows. Overall, the discrepancy between answer correctness and localization accuracy which reveal an important insight: **correct answers do not guarantee correct reasoning paths**. Many models produce plausible answers while attending to irrelevant regions, suggesting reliance on global textual cues rather than truly grounded evidence. Although high Good Ratios indicate stable instruction following, they do not translate into accurate spatial grounding. These findings reinforce that **spatial grounding remains a critical bottleneck** for modern VLMs, underscoring the need for improved evidence localization to support trustworthy multimodal reasoning.

Overall, the results from both tables show a clear trend: models perform best when evidence is provided, degrade moderately with full-page inputs, and struggle most when required to localize and reason jointly. This progression reinforces the motivation behind BBox-DocVQA: to isolate, evaluate, and ultimately improve the grounding component that underpins faithful reasoning in document QA systems.

4.3 Analysis

Accuracy Analysis Across SPSBB, SPMBB, and MPMBB. Table 5 reports answer accuracy under the simultaneous perception–and–answering setting for the three types of QA. Overall, larger and more recent models achieve substantially higher accuracy, with GPT-5 and Qwen3VL-32B leading across all categories. A consistent pattern emerges across model families: performance is highest on **SPMBB** questions, slightly lower on **SPSBB**, and lowest on **MPMBB**. The strong performance on SPMBB is noteworthy; although these questions involve multiple bounding boxes, they force VLMs to focus on richer evidence within a single page, facilitating models to generate a stable answer. On the contrary, **MPMBB** questions require reasoning across multiple pages, expanding the search space and increasing the likelihood of grounding errors, which leads to a considerable accuracy drop (e.g., 84.35%→55.35% for Qwen3VL-32B). These results indicate that while newer VLMs are steadily improving in aggregating intra-page evidence, **multi-page reasoning remains challenging**, and precise cross-page evidence localization is still a major limitation in current models.

Case study. To further understand model behavior under different input settings, we perform representative case studies and visualize them in Figure 4. These examples reveal that, when processing full-page inputs, many errors arise from the model's inability to correctly align the question with

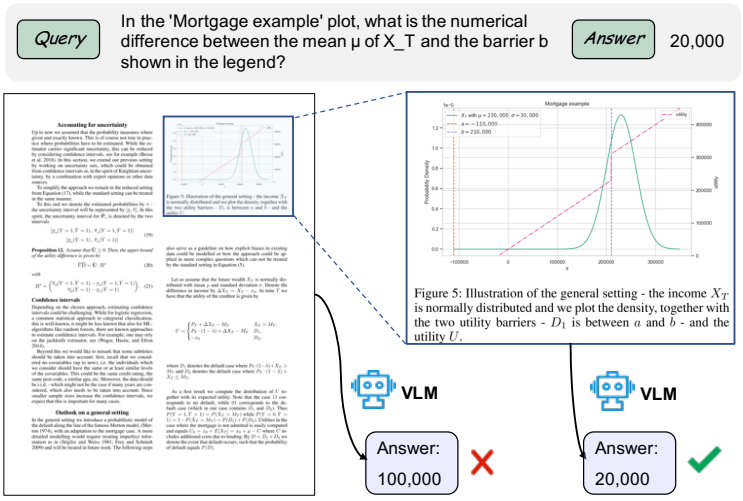

Figure 4: A case study demonstrating that the fine-grounded subimage can help the VLM to generate answers more accurately. The VLM is Qwen2.5VL-7B.

the intended target region, leading to confusion with adjacent text blocks, nearby table cells, or visually similar graphical elements. Such region-level misalignment becomes particularly severe in documents with complex structures. On the contrary, when the input is restricted to the cropped evidence region, the model is able to better capture local details—such as small numerical differences, label semantics, or localized diagram relationships—resulting in substantially improved accuracy. Notably, some questions remain challenging even with full-page information provided, reflecting inherent limitations of current VLMs in fine-grained visual understanding, structured document reasoning, and cross-modal alignment. Overall, these case studies demonstrate that *spatial grounding* not only improves answer quality but also provides greater interpretability and clearer error patterns, further highlighting the necessity of explicit region alignment in document understanding tasks.

4.4 Discussion

Our results underscore that reasoning in document understanding using a two-step process—a model first determines where the relevant evidence lies before it can accurately infer what the answer should be—can significantly increase the generation accuracy. Current VLMs, however, often collapse these two steps, attempting to reason directly from an entire page and consequently attending to visually dominant but semantically irrelevant content. This leads to unstable reasoning paths and incorrect answers, especially in documents containing dense layouts or multiple candidate regions. On the contrary, when the correct evidence box is provided, answer accuracy increases substantially, demonstrating that effective reasoning critically depends on precise spatial grounding. Yet, even state-of-the-art models struggle to identify coherent semantic units such as paragraphs, tables, or figures, revealing that their underlying reasoning mechanisms are insufficiently aligned with the structure of real documents. BBox-DocVQA fills this gap by offering bounding-box–grounded supervision that explicitly links each question to its reasoning substrate, enabling models to learn not only to produce an answer but also to locate and utilize the correct evidence. This establishes a more faithful approximation of human-like document reasoning and provides a diagnostic benchmark for evaluating whether VLMs truly reason or merely guess.

5 Conclusion

This work presents BBox-DocVQA, a large-scale dataset designed to advance reasoning-driven document understanding by grounding every QA pair in explicit evidence regions. Built through our automated Segment–Judge–and–Generate pipeline and verified through human inspection, the

dataset spans 3.6 K documents and 32 K QA pairs covering multi-page and multi-region reasoning scenarios. Our benchmarking results show that even the strongest VLMs exhibit significant failures in grounding—failures that directly impair their reasoning accuracy—demonstrating that current models are far from achieving robust spatial–semantic reasoning. By providing high-quality, semantically coherent bounding-box annotations, BBox-DocVQA offers the supervision necessary for training and evaluating models that can perform structured, evidence-based reasoning rather than relying on superficial correlations or global textual cues. We believe this dataset will catalyze future advances in grounding-aware model architectures, multi-step reasoning paradigms, and transparent document AI systems capable of explaining not only what they answer but why.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A Rossi, Changyou Chen, and Tong Sun. Sv-rag: Lora-contextualizing adaptation of mllms for long document understanding. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Wang Chen, Guanqiang Qi, Weikang Li, Yang Li, Deguo Xia, and Jizhou Huang. Pairs: Parametric-verified adaptive information retrieval and selection for efficient rag. *arXiv preprint arXiv:2508.04057*, 2025b.

Wang Chen, Wenhan Yu, Guanqiang Qi, Weikang Li, Yang Li, Lei Sha, Deguo Xia, and Jizhou Huang. Cmrag: Co-modality-based visual document retrieval and question answering. *arXiv preprint arXiv:2509.02123*, 2025c.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, 2022.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yulong Hui, Yao Lu, and Huanchen Zhang. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *Advances in Neural Information Processing Systems*, 37: 67200–67217, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

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

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, et al. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations*, 2025.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840, 2023.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, 2024a.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

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

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

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536, 2020.

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

Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. Rora-vlm: Robust retrieval-augmented vision language models. *arXiv preprint arXiv:2410.08876*, 2024a.

Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. Long2rag: Evaluating long-context & long-form retrieval-augmented generation with key point recall. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4852–4872, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. 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 6088–6109, 2025.

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

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645, 2023.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24827–24837, 2025.

Yang Tian, Fan Liu, Jingyuan Zhang, V. W., Yupeng Hu, and Liqiang Nie. CoRe-MMRAG: Cross-source knowledge reconciliation for multimodal RAG. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32967–32982, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1583. URL `https://aclanthology.org/2025.acl-long.1583/`.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

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

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*, 2025a.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*, 2025b.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, et al. mr ^{2} ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*, 2024.

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

1 Details of Dataset Construction

1.1 Data Collection: Large-Scale Scientific Document Crawling Based on arXiv

To construct a scientific document understanding dataset that spans multiple disciplines and diverse formats, we first performed automated paper crawling based on arXiv. The entire data collection process relies on the open-source tool `arxiv-dl`, which obtains paper metadata through the official arXiv API and supports automatic PDF downloading, failure retry mechanisms, and discipline-level sampling control. We covered eight major subject areas, including Computer Science, Mathematics, Physics, Economics, Statistics, Quantitative Biology, Quantitative Finance, and Electrical Engineering and Systems Science. Ultimately, we built a structured and scalable raw PDF corpus that serves as the foundational input for the subsequent automatic annotation pipeline.

1.2 PDF Conversion and Preprocessing

In the first stage of the pipeline, each PDF page is rendered into a 300 DPI PNG image. High-resolution rendering is essential because clear pixel-level details significantly enhance the performance of later modules such as segmentation, recognition, and classification. Fine-grained visual information—including table lines, mathematical symbols, and small-font text—is more accurately preserved at higher resolutions. This step provides a critical high-quality visual basis for the entire cropping and understanding workflow

1.3 Step 1: Extracting Candidate Cropping Regions via SAM Segmentation

After preprocessing, we first apply the Segment Anything Model (SAM) to perform fine-grained segmentation on each page image, enabling the automatic discovery of potential "logical content blocks" such as paragraphs, tables, figures, and charts. In our implementation, we adopt the official model checkpoint `sam_vit_h.pth`, which offers strong segmentation accuracy and generalization capabilities in complex document scenarios.

Given a page image, we perform automatic segmentation using SAM to obtain potential semantic regions. Specifically, we invoke SAM's automatic segmentation interface on each full-page image, which produces a set of high-quality segmentation masks covering the entire page. These masks typically align naturally with major structural components within the document, including paragraph text blocks, table regions, and various figure or chart elements. This provides the foundation for constructing candidate cropping regions.

After obtaining the segmentation masks, we convert each mask into a tight bounding box. For each mask, we compute the minimum and maximum foreground pixel coordinates along both the x and y axes, resulting in a bounding box defined as $[x_1, y_1, x_2, y_2]$ in the standard `bbox_xyxy` format. Bounding boxes extracted via minimal enclosing rectangles closely capture the true content regions, effectively reducing the inclusion of irrelevant background and improving region localization for subsequent filtering and question–answer pair generation.

To avoid extremely small noise regions (such as single characters or decorative icons) as well as excessively large regions that nearly cover an entire page, we further filter the bounding boxes based on their relative area ratio. Let the page height and width be H and W. For a bounding box $\text{bbox} = (x_1, y_1, x_2, y_2)$, its area is computed as

$$\text{area} = (x_2 - x_1) \times (y_2 - y_1).$$

We define the area ratio of the bounding box relative to the entire page as

$$r = \frac{\text{area}}{H \times W}.$$

To filter out invalid extremely small or nearly full-page regions, we empirically set lower and upper area-ratio thresholds:

$$r_{\min} = 0.05, \qquad r_{\max} = 0.70.$$

A bounding box is retained as a valid candidate region only when it satisfies

$$r_{\min} \leq r \leq r_{\max}.$$

Finally, to reduce the risk of cutting off text edges or table borders during cropping, we expand each bounding box by a uniform padding of 10 pixels on all sides (while ensuring that the padded coordinates remain within image boundaries). This padding preserves logical region completeness even under slight localization inaccuracies, which is particularly important for table grid lines, paragraph line endings, and other edge-sensitive content.

Through the steps above, we automatically extract a set of candidate cropping regions from each document page without relying on any handcrafted rules. These regions exhibit high content density and sufficient granularity, forming a strong foundation for the subsequent quality assessment and high-quality question–answer generation process.

1.4 Step 2: Judge and Classify

After obtaining the initial set of candidate cropping boxes, we employ the multimodal large language model Qwen2.5-VL-72B to automatically score and classify each cropped region. The objective of this step is to preserve fine-grained supervision signals while strictly filtering out incomplete, redundant, or noisy regions, ensuring that all crops entering the question–answer generation stage possess high annotation value.

Model Judgment Objectives and Output Format For each candidate cropping region, we provide Qwen2.5-VL-72B with a carefully designed instruction that enables the model to perform two tasks in a single inference pass.

First, the model must evaluate the quality of the crop by determining whether the region constitutes a "correct and sufficiently fine-grained" semantic block. It outputs a binary quality label `keep`: when `keep=true`, the crop should be preserved; when `keep=false`, the crop should be discarded.

Second, the model must classify the crop according to its dominant content type, assigning it to one of three categories:

- `text`: Regions primarily containing linear textual structures such as paragraphs or lists, without visible table grid lines.
- `table`: Regions exhibiting clear row–column structures, cell boundaries, or matrix-like layouts.
- `image`: Regions containing figures, illustrations, schematic diagrams, curves, bar charts, or other non-tabular visual elements.

Through this process, we achieve accurate semantic characterization and filtering of cropped regions while maintaining the efficiency of a fully automated pipeline. This provides precise type annotations for subsequent redundancy filtering and question–answer generation.

Positive and Negative Case Criteria During instruction design, we explicitly define detailed criteria for positive and negative examples to ensure that the model's judgments are interpretable and reproducible. As shown in Figure 5, we define explicit criteria for positive and negative cases. By clearly specifying these conditions within the prompt, Qwen Judge can systematically distinguish high-quality crops from invalid fragments, significantly reducing the burden of subsequent manual filtering.

Redundancy Removal via Overlap Ratio After the first round of quality judgment, we retain only those cropping regions for which the model outputs `keep=true`, forming the candidate set. To further reduce redundant crops and improve data usefulness, we compute pairwise overlap ratios for every pair of regions within the candidate set. Specifically, for any two bounding boxes, let I denote

Prompt for Qwen Judge

Please determine whether this image represents a correct and sufficiently fine-grained content crop.
Positive case (keep=true) must satisfy ALL of the following:
A. Contains only **one** complete logical block (a full paragraph, full table, or full chart) with no obvious truncation at the borders;
B. The meaningful content occupies a **significant portion** of the image (approximately $\geq 30\%$), not dominated by blank space;
C. The image is clean — no headers, footers, page numbers, watermarks, or decorative frames unrelated to the main content.
Negative case (keep=false) if ANY of the following is true:
– The image is mostly blank or has excessive whitespace;
– It only contains a header, footer, page number, year, or report title (e.g., "4 2006 Annual Report");
– It contains only a caption or title **without** the corresponding figure or table body;
– Multiple logical blocks appear simultaneously (e.g., double columns, cross-page layout, or figure + text mixed together);
– The content is truncated (e.g., only part of a figure, incomplete table borders, or a paragraph cut at the top/bottom);
– The cropped content is too small, blurry, or unreadable;
– Large margins, edges, or decorative borders occupy most of the frame.
In addition to `keep`, classify the primary content type of the crop as one of: `"text"`, `"table"`, or `"image"`.
– `text`: a body paragraph or list; predominantly textual without tabular gridlines.
– `table`: a grid-like table or matrix with visible rows/columns.
– `image`: a chart/figure/graphic/photo/diagram (including plots and bar/line charts).
Output a strict JSON object on a single line:
{`"keep"`: true—false, `"type"`: "text"—"table"—"image"}
If uncertain, set keep=false and choose the most plausible type.

Figure 5: Prompt used for automatic crop judging and classification.

the area of their intersection, and let $S_{\min}$ denote the area of the smaller box. The overlap ratio is then defined as

$$\text{overlap} = \frac{I}{S_{\min}}.$$

When overlap ≥ 0.90, we consider the two crops to be semantically near-duplicates and apply additional filtering rules. Based on the content type, we design differentiated retention strategies: for `text` regions, in cases of high overlap we preferentially retain the smaller bounding box in order to obtain more fine-grained textual supervision; for `table` and `image` regions, we instead retain the larger crop, ensuring that the full structural layout of the table or figure is preserved.

It is worth noting that in cases of high overlap between regions of *different* types (e.g., partial overlap between a text region and an image region), we do not enforce forced removal. This prevents the inadvertent deletion of potentially important content and maximizes the preservation of complementary supervision signals across different views of the page.

Through this overlap analysis and type-sensitive filtering mechanism, Step 2 ultimately yields a cleaner, more fine-grained, and semantically well-defined set of cropping regions, providing reliable inputs for generating high-quality single-region question–answer pairs.

1.5 Step 3: Question–Answer Generation

After completing quality filtering and type classification of the cropped regions, we further construct fine-grained question–answer supervision based on these candidate regions. In this stage, we first ap-

ply a type- and frequency-controlled sampling strategy to all retained crops, dividing them into three categories: SPSBB (Single-Page Single-Block Boxes), SPMBB (Single-Page Multi-Block Boxes), and MPMBB (Multi-Page Multi-Block Boxes). During sampling, we randomly select combinations from these categories for QA generation, while appropriately increasing the sampling probability of `image` and `table` crops to enhance supervision coverage over non-textual structures. The resulting sampled crop combinations may come from a single page or span multiple pages.

Based on the sampled crop sets, we instruct the multimodal model using a unified constraint-driven prompt to ensure that the generated QA pairs rely strictly on the fine-grained local details visible within the selected crops. The prompt is shown in Figure 6.

Prompt for Single-Crop QA Generation

You are an expert-level question-and-answer generator. Based strictly on the visible content of **ALL provided crop images** (they may come from **multiple pages**), generate exactly **one** challenging, grounded question with its answer.

Primary objective: The question must rely on fine-grained details visible *only* in the crops; a reader using the full pages without inspecting these precise crops should likely answer incorrectly.

Rules:

1. Use only the crops' visible content—no outside knowledge, hallucination, or assumptions. Do not infer text that is not clearly visible.
2. Prefer questions that integrate or compare details across multiple crops (e.g., numbers, labels, legends, figure indices, footnotes, coordinates). If only one crop is available, the question must hinge on a subtle or localized detail within that crop.
3. Avoid generic or high-level questions answerable from page-level context, titles, or common sense. The question should be easy *only* when the crops are carefully inspected.
4. Make the question as challenging and reasoning-intensive as possible while remaining fully grounded in the crops.
5. The answer must be concise (no more than 10 words), precise, and directly supported by the crops.
6. Output **valid JSON only** in the following exact format:
   {"q": "...","a": "..."}
7. If the crops lack sufficient information to form such a question, output exactly:
   {"q": "", "a": ""}

Figure 6: Prompt used for generating fine-grained single-crop QA pairs.

2 Prompts of Experiment

SYSTEM_PROMPT_IMAGE

You will be given:

1. One or more page images (each page as a separate input image).
2. A natural language question `q`.

Your task:

1. Analyze the visual content carefully (e.g., text, patterns, symbols, diagrams, or objects) and determine the correct answer.

Output requirements:

1. Include one concise `answer` field summarizing the final textual answer.

Return strictly in this JSON format:

```
{"answer": "<your concise answer here>"}
```

Do not include explanations, reasoning, or any other text outside the JSON.

Figure 7: Prompt used for answering questions from images.

SYSTEM_PROMPT_IMAGE_BBOX

You will be given:

1. One or more page images (each page as a separate input image).
2. A natural language question `q`.

Your task:

1. Locate all regions (bounding boxes) containing the information needed to answer the question.
2. Each bounding box must capture one complete logical block (full paragraph, full table, or full chart) without truncation.
3. Derive the concise textual answer using only the content inside those bounding boxes.

Bounding box format:

- Pixel coordinates relative to the top-left of the page.
- Format: `[left_x, top_y, right_x, bottom_y]`

Output requirements:

- Return one or more bounding boxes per page if needed.
- Pages with no relevant content must return an empty list `[]`.
- Include one concise `answer` field.

Return strictly in this JSON format:

```
{
  "bboxes": [[[x1, y1, x2, y2],[x1, y1, x2, y2]],[[x1, y1, x2, y2]]],
  "answer": "<your concise answer here>"
}
```

Do not include explanations, reasoning, or any other text outside the JSON.

Figure 8: Prompt used for answering questions from images with bounding boxes.

SYSTEM_PROMPT_GEN_BBOX

You will be given:

1. One or more page images (each page as a separate input image).
2. A natural language question `q`.

Your task:

1. Identify all regions (bounding boxes) on the pages that contain information relevant to the question.
2. Each bounding box must represent **one complete logical block** (e.g., paragraph, table, figure) with **no truncation**.

Bounding box format:

- Coordinates are absolute pixel values from the top-left corner.
- Format: `[left_x, top_y, right_x, bottom_y]`

Output requirements:

- Return one or more bounding boxes per page if applicable.
- Pages with no relevant content must return an empty list `[]`.

Return strictly in this JSON format:

```
{
  "bboxes": [[[x1, y1, x2, y2],[x1, y1, x2, y2]],[[x1, y1, x2, y2]]]
}
```

Do not include explanations, reasoning, or any other text outside the JSON.

Figure 9: Prompt used for generating bounding boxes relevant to the question.

3 Details of Experiment Results

Model	Mode	Mean Acc.	SPSBB(749)	SPMBB(556)	MPMBB(318)
Qwen2.5VL-3B	pages_bbox_answer	30.87%	31.38%	33.27%	25.47%
	gt_subpage_answer	56.38%	61.28%	57.01%	43.71%
	gt_page_answer	45.72%	50.20%	46.22%	34.28%
Qwen2.5VL-7B	pages_bbox_answer	51.57%	57.01%	53.42%	35.53%
	gt_subpage_answer	68.58%	73.03%	68.71%	57.86%
	gt_page_answer	57.18%	61.01%	58.63%	45.60%
Qwen2.5VL-32B	pages_bbox_answer	63.46%	67.69%	66.55%	48.11%
	gt_subpage_answer	**82.01**%	84.51%	83.27%	73.90%
	gt_page_answer	73.01%	76.37%	73.02%	65.09%
Qwen2.5VL-72B	pages_bbox_answer	68.64%	71.03%	71.58%	57.86%
	gt_subpage_answer	79.42%	83.04%	80.40%	69.18%
	gt_page_answer	69.32%	70.76%	73.20%	59.12%
Qwen3VL-4B	pages_bbox_answer	68.95%	70.49%	72.12%	59.75%
	gt_subpage_answer	72.34%	74.10%	74.28%	64.78%
	gt_page_answer	67.22%	66.89%	70.32%	62.58%
Qwen3VL-8B	pages_bbox_answer	67.59%	70.23%	73.20%	51.57%
	gt_subpage_answer	74.74%	75.97%	77.34%	67.30%
	gt_page_answer	72.27%	72.36%	75.54%	66.35%
Qwen3VL-32B	pages_bbox_answer	77.14%	81.04%	84.35%	55.35%
	gt_subpage_answer	**82.01**%	82.38%	84.71%	76.42%
	gt_page_answer	**80.28**%	79.44%	84.53%	74.84%
InternVL3-2B	pages_bbox_answer	34.20%	39.12%	33.27%	24.21%
	gt_subpage_answer	41.59%	37.25%	31.47%	20.75%
	gt_page_answer	32.04%	48.20%	38.85%	30.82%
InternVL3-8B	pages_bbox_answer	50.46%	53.14%	53.06%	39.62%
	gt_subpage_answer	65.80%	57.41%	53.24%	44.34%
	gt_page_answer	53.42%	67.56%	65.11%	62.89%
GPT-5	pages_bbox_answer	**81.45**%	82.64%	83.63%	74.84%

Table 6: Answer accuracy on the BBox DocVQA benchmark under different modes. We evaluate the simultaneous perception–and–answering setting, where the model must localize evidence regions and generate answers in a single unified step. Results are reported for SPSBB, SPMBB, and MPMBB tasks.