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

Anonymous CVPR submission

Paper ID 3031

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. Furthermore, fine-tuning on BBox-DocVQA substantially improves both bounding-box localization and answer generation, validating its effectiveness for enhancing the reasoning ability of VLMs. Our dataset and code will be publicly released to advance 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 visual information jointly. Recent progress in vision–language models (VLMs) [1–3, 28, 40] has greatly improved their ability to process and comprehend complex document layouts. However, the majority of existing DocVQA datasets, such as DocVQA [22], MP-DocVQA [35], and VisualMRC [31], are constructed at the *page level*. Correspondingly, most existing methods [6, 7, 9, 10, 12, 16–18, 26, 27, 34, 37–39, 41, 42] 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.

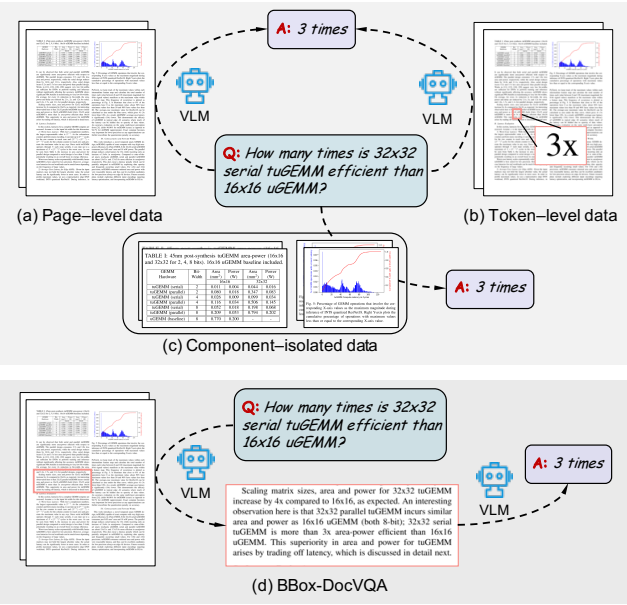

Figure 1. Comparison among (a) page–level data, (b) token–level data, (c) component–isolated data, and (d) our proposed BBox-DocVQA.

Human document comprehension typically involves two steps: first, identifying the relevant pages, and second, locating specific *evidence components*—such as text para-

Dataset / benchmark	#Queries	#Docs	#Images	#I/Q	Avg. Q Len	Avg. Ans Len	BBox	MP
DUDE [36]	41 K	5 K	28.7 K	1.00	8.65	3.35	Yes	No
BoundingDocs [11]	249 K	48 K	237 K	1.00	–	–	Yes	No
SlideVQA [32]	14.5 K	–	52 K	1.26	–	–	Yes	Yes
SPIQA [25]	270 K	25.9 K	270 K	1.00	12.98	14.56	–	No
PlotVQA [24]	29 M	–	224 K	1.00	43.54	–	–	No
ArXivQA [15]	100 K	16.6 K	32 K	1.00	16.98	7.59	–	No
ChartQA [21]	33 K	–	22 K	1.00	13.18	1.08	–	No
TAT-DQA [43]	16 K	2.7 K	3 K	1.07	12.54	3.44	No	No
VisualMRC [31]	30 K	10 K	10 K	1.00	10.55	9.55	No	No
InfographicsVQA [23]	30 K	5.4 K	5.4 K	1.00	11.54	1.60	No	No
DocVQA [22]	50 K	6 K	12 K	1.00	9.49	2.43	No	No
MP-DocVQA [35]	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.

graphs, tables, or figures—that support the final answer. Current page-level DocVQA datasets, however, only support the first step, offering page-level evidence without bounding-box annotations. On the other hand, a few token-level datasets, such as DUDE [36] and BoundingDocs [11], 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. As a result, existing DocVQA resources cannot fully capture the fine-grained reasoning process required for spatially interpretable document understanding.

To bridge these gaps, 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 (SAM[14]) 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 [1], Qwen2.5-VL [3], and the InternVL [8] 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. To further explore the utility of our dataset, we fine-tune Qwen2.5-VL on the BBox-DocVQA training set and observe substantial improvements in both bounding-box localization and answer generation. This demonstrates that bounding-box–grounded supervision can significantly enhance the reasoning transparency and accuracy of multimodal models.

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 and fine-tuning experiments with multiple state-of-the-art VLMs, demonstrating that BBox-DocVQA substantially enhances models' spatial grounding and answer accuracy, offering a new foundation for interpretable document understanding research.

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 [36] and BoundingDocs [11], and SlideVQA [32], 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 [25], PlotQA [24], ChartQA [21], and ArXivQA [15]. 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 [22] dataset and its multi-page extension MP-DocVQA [35] comprise 50 K questions over 12 K single-page documents and 46 K questions spanning 6 K multi-page documents, respectively. Similarly, VisualMRC [31], InfographicsVQA [23], and TAT-DQA [43] extend document-level comprehension to specific domains. Also, a few benchmarks, including OK-VQA [20], A-OKVQA [29], WebQA [4], UDA [13], Dyn-VQA [16], MMLongBench [19], REAL-MM-RAG [39], ViDoRe [10], ViDoSeek [37], M3DoCVQA [9], OpenDocVQA [33], VisR-Bench [5], and VisDoMBench [30], 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.

3.1. Segment-Judge-and-Generate

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 Fig. 2 and 3.

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 [14] (SAM, ViT-H variant) to automatically detect layout components on each page image. The model

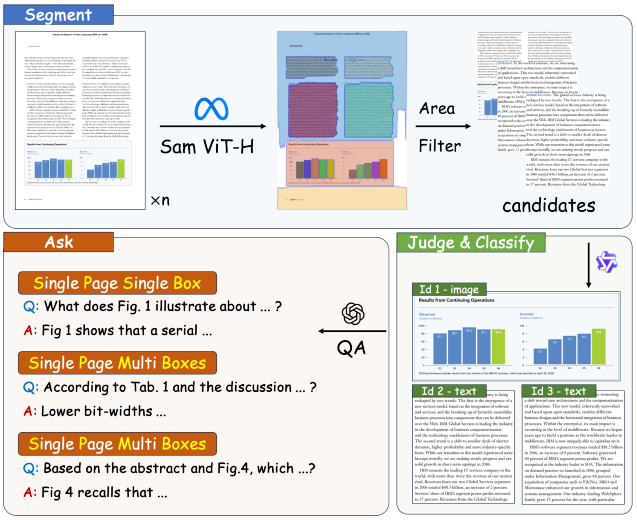

Figure 2. The proposed **Segment-Judge-and-Generate** pipeline for constructing a bounding-box-grounded DocQA dataset. It includes three stages: (1) Segment — detect fine-grained layout components via SAM (ViT-H); (2) Judge and Classify — evaluate and categorize regions with Qwen2.5-VL-72B; and (3) Generate — generate reasoning-oriented QA pairs with GPT-5.

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.

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,

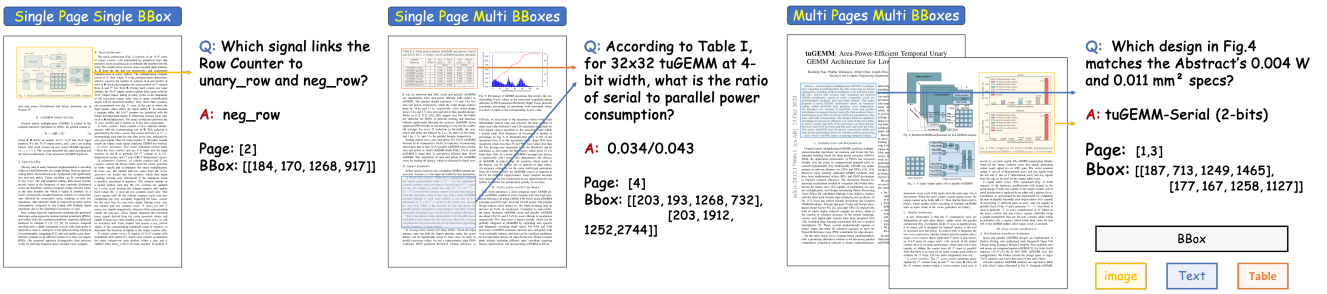

Figure 3. Case study illustrating different document understanding settings in the BBox-DocVQA dataset. Examples include (left) single-page single-bbox reasoning, (middle) single-page multi-bbox reasoning, and (right) multi-page multi-bbox reasoning, where questions and answers are grounded on one or more spatially localized regions (bboxes) across document pages.

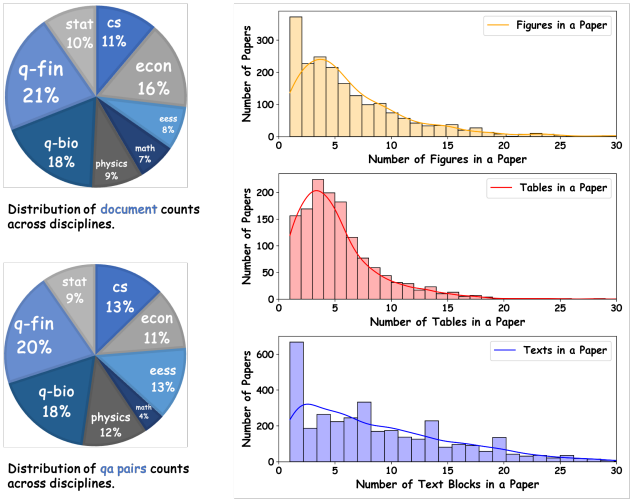

Figure 4. The left panels show the proportions of papers across research domains for (top) the document dataset and (bottom) the generated QA pairs. The right panels display the distributions of the number of figures, tables, and text blocks per paper.

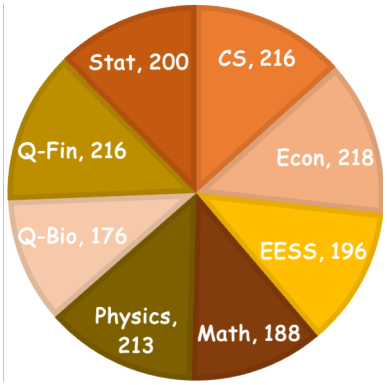

Figure 5. Domain-wise distribution of QA samples in the Fine-grained Benchmark.

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 2, 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 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 anno-

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

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%

tations. 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 Between Training Set and Benchmark

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.

4. Task and model training

In this section, we introduce the DocVQA task definition and the proposed two-stage training strategy using the BBox-DocVQA dataset.

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:

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

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

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

Two-stage training strategy. We adopt a two-stage training strategy to progressively align the model's visual grounding and reasoning capabilities. In the first stage, a LoRA fine-tuning is performed to teach the model to localize the evidence region (BBox) corresponding to each query. This step focuses purely on spatial grounding, allowing the model to build a stable correspondence between textual prompts and visual layouts before introducing semantic supervision. In the second stage, the model is further trained to jointly predict the bounding box and generate the textual answer. This joint optimization bridges grounding and generation, encouraging the model to reason based on localized visual evidence rather than relying on global context. All implementation details, including hyperparameters, optimization settings, and loss composition, are provided in the Appendix. Although the constructed dataset can also benefit the retriever by offering explicit evidence supervision, we leave retriever-level enhancement for future work and focus here on improving the generation stage.

5. Experiment

5.1. Experiment setup

Baselines.

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. 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** (i.e., DeepSeek-v3.1) 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.

5.2. Main results

5.3. Analysis and discussion

6. Conclusion

Model	Mean	SPSB (749)	SPMB (556)	MPMB (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 3. 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 SPSB, SPMB, and MPMB tasks.

Model	Mean IoU	SPSB (749)	SPMB (556)	MPMB (318)
Qwen2.5VL-3B	4.70%	3.80%	5.60%	4.90%
Qwen2.5VL-7B	11.30%	14.60%	8.80%	8.00%
Qwen2.5VL-32B	20.00%	22.20%	20.40%	14.30%
Qwen2.5VL-72B	**35.20%**	**40.10%**	**33.20%**	**27.20%**
Qwen3VL-4B	18.70%	18.90%	17.00%	21.30%
Qwen3VL-8B	14.40%	17.60%	9.20%	15.90%
Qwen3VL-32B	20.40%	22.60%	17.30%	21.00%
InternVL3-2B	0.10%	0.00%	0.20%	0.10%
InternVL3-8B	0.30%	0.10%	0.50%	0.50%
GPT-5	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.

Model	GT Subpage Ans.	GT Page Ans.
Qwen2.5VL-3B	56.38%	45.72%
Qwen2.5VL-7B	68.58%	57.18%
Qwen2.5VL-32B	82.01%	73.01%
Qwen2.5VL-72B	79.42%	69.32%
Qwen3VL-4B	72.34%	67.22%
Qwen3VL-8B	74.74%	72.27%
Qwen3VL-32B	**82.01%**	**80.28%**
InternVL3-2B	41.59%	32.04%
InternVL3-8B	65.80%	53.42%

Table 5. Answer accuracy when ground-truth evidence regions are provided (*GT Subpage*) compared to answering based only on the entire page image (*GT Page*). Across all models, supplying the precise evidence region consistently improves accuracy, demonstrating that **localization substantially enhances answer quality**.

References

[1] 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. 1, 2

[2] 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.

[3] 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. 1, 2

[4] 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. 3

[5] 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*, 2025. 3

[6] 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. 1

[7] 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. 1

[8] 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. 2

[9] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024. 1, 3

[10] 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. 1, 3

[11] 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. 2, 3

[12] 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. 1

[13] 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. 3

[14] 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. 2, 3

[15] 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. 2, 3

[16] 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. 1, 3

[17] 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.

[18] 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, 2024. 1

[19] 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, 2024. 3

[20] 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. 3

[21] 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. 2, 3

[22] 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. 1, 2, 3

[23] 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. 2, 3

[24] 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. 2, 3

[25] 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. 2, 3

[26] 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*, 2024. 1

[27] 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, 2024. 1

[28] 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. 1

[29] 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. 3

[30] 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. 3

[31] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 1, 2, 3

[32] 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*, pages 13636–13645, 2023. 2, 3

[33] 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. 3

[34] Yang Tian, Fan Liu, Jingyuan Zhang, V. W., Yupeng Hu, and Liqiang Nie. CoRe-MMRAG: Cross-source knowledge reconciliation for multimodal RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32967–32982, Vienna, Austria, 2025. Association for Computational Linguistics. 1

[35] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 1, 2, 3

[36] 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. 2, 3

[37] 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*, 2025. 1, 3

[38] 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*, 2025.

[39] 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. 1, 3

[40] 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. 1

[41] 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. 1

[42] 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. 1

[43] 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. 2, 3

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

Supplementary Material

1. Details of dataset construction