

Doc-CoB: Enhancing Document Understanding with Visual Chain-of-Boxes Reasoning

Ye Mo^{1*}, Kai Ye^{1*}, Xianwei Mao^{1*}, Zirui Shao¹, Gang Huang², Bo Zhang³,
Hangdi Xing¹, Kehan Chen², Huan Zhou², Zixu Yan², Jiajun Bu¹, Sheng Zhou^{1†}

¹ Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems,
Zhejiang University ²Alibaba Group, ³Shanghai AI Laboratory
{moye017, mercury0926, maoxianwei, zhousheng_zju}@zju.edu.cn

Abstract

Document understanding aims to perform question answering and information extraction over document images, where the visual content is highly information-dense and most queries rely on only a few relevant layout regions. However, existing methods either adopt a one-pass strategy that implicitly assumes all layouts are equally important, or focus excessively on small regions at the cost of losing critical layout information. To address these limitations, we introduce **Doc-CoB** (Chain-of-Boxes), a simple-yet-effective framework that integrates coarse-to-fine layout-aware visual reasoning into multimodal large language models. Instead of directly zooming into small regions, Doc-CoB progressively focuses on query-relevant layouts while preserving global document information. Specifically, it first selects key layout boxes and then focuses on them for further understanding with visual prompting. To support this paradigm, we introduce two reasoning tasks for *box recognition* and *box reasoning*, with an automatic pipeline that constructs 249k training samples with intermediate visual supervision. Experiments on seven benchmarks with four popular models show that Doc-CoB significantly improves performance, demonstrating its effectiveness and wide applicability. The code and the data are available at <https://github.com/Doc-CoB/Doc-CoB>.

1 Introduction

Document understanding (Cui et al., 2021; Xu et al., 2020) is a fundamental task that involves question answering and information extraction based on document images, such as forms, receipts, reports, and scanned pages. In recent years, multimodal large language models (MLLMs) (Bai et al., 2025a; Gemini Team, 2025; Llama Team, 2024; Cai et al., 2025) have demonstrated promising capabilities on document understanding (Hu et al.,

2024; Ding et al., 2026; Zhou et al., 2024; Luo et al., 2024; Mo et al., 2025), driving progress in both academic research and industrial applications.

Most existing MLLMs generate responses to queries by processing entire images in a uniform manner (Chen et al., 2025; Fu et al., 2025), implicitly assuming that all regions are equally important. However, this overlooks the information density (Cui et al., 2021) of document images, where only a few layout regions are task-relevant (Cao et al., 2023), while the remainder is redundant. As shown in Figure 1 (a), while not all regions are relevant to the query “Where does the applicant live?”, the model is misled by excessive redundant information and infers the answer from a wrong region.

Despite the emergence of coarse-to-fine strategies in general visual QA (Yu et al., 2024), many existing approaches have drifted toward the opposite extreme. Specifically, some works perform direct cropping or zoom-in on selected regions after a localization stage (Luan et al., 2024; Shao et al., 2024). While such designs reduce visual redundancy, they inevitably discard layout information that is unique and critical to document images, such as spatial relationships, relative positioning, and cross-region dependencies. As a result, these methods may struggle with layout-sensitive queries and complex document structures. This raises a research question: *How to follow the coarse-to-fine paradigm while preserving layout information for document understanding?*

To address the above challenge, we propose **Doc-CoB** (Chain-of-Boxes), a framework that enables MLLMs to perform document understanding through coarse-to-fine, layout-aware visual reasoning without modifying their architecture. We decompose the process into two stages: (S1) *Key Box Selection* and (S2) *Focused Answering*, as illustrated in Figure 1 (b). Both stages employ visual prompting (Yang et al., 2023b; Shtedritski et al., 2023) to explicitly guide the model’s attention to

* Equal contribution.

† Corresponding author.

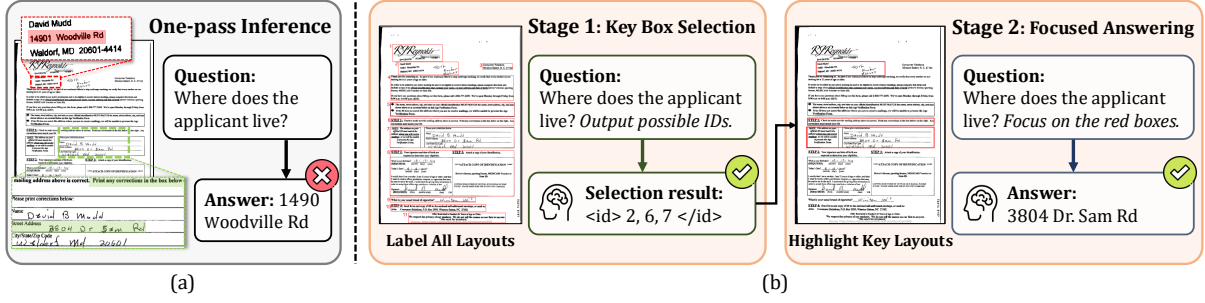


Figure 1: (a) Current MLLM attends to a wrong region and produces an incorrect answer, instead of obtaining the answer from the correct layout region below. (b) Overview of Doc-CoB. S1: Key Box Selection identifies boxes relevant to the query. S2: Focused Answering generates the final answer from the visually prompted image.

key layout (box) while suppressing visual redundancy. We then introduce two reasoning tasks: (T1) *Box Recognition* and (T2) *Box Reasoning*, which aim to improve the model’s understanding of document layout and its associated semantic information. Finally, to address the challenge of limited training data, we construct a large-scale dataset with intermediate visual supervision tailored to Doc-CoB. Built upon an automatic data generation pipeline, the dataset has 249K training samples collected from nine document datasets, where layout structure and semantic roles are jointly encoded.

Extensive experiments and ablation studies examine Doc-CoB from multiple complementary perspectives. All MLLMs achieve improved performance when integrated with Doc-CoB, with an 8B model even surpassing GPT-4o. Further in-depth analysis reveals that the explicit Doc-CoB reasoning paradigm and the proposed reasoning tasks are both critical. Beyond in-domain evaluation, Doc-CoB exhibits strong generalization, remaining effective under out-of-domain settings. Our contributions are as follows:

- We introduce Doc-CoB, a framework that enables MLLMs to perform document understanding through coarse-to-fine, layout-aware visual reasoning.
- We propose two reasoning tasks to further enhance the reasoning capabilities of MLLMs. In addition, we construct a dataset of 249k samples across nine document datasets, each annotated with intermediate visual supervision tailored for Doc-CoB.
- Doc-CoB improves performance on seven benchmarks, surpasses strong baselines, and remains compatible with various MLLM architectures and scales.

2 Related Work

Multimodal Reasoning. A critical step in advancing AI toward human-level intelligence is enabling the transition from basic perception to complex reasoning (Wang et al., 2025). Recent works (OpenAI, 2024b; DeepSeek-AI, 2025) have shown that incorporating reasoning can significantly enhance model performance. One widely used class of methods is Chain-of-thought (CoT) (Wei et al., 2022), which encourages large language models (LLMs) to clarify reasoning, specifically by adding thinking processes (Kang et al., 2025; Li et al., 2025; Miao et al., 2024; Mo et al., 2026).

Meanwhile, progress has been made in reasoning for MLLMs. MM-CoT (Zhang et al., 2024b) proposed a two-stage reasoning framework by first generating rationales and then generating answers. Subsequent works (He et al., 2024; Wang et al., 2024b) further improve the design on vision-language fusion mechanisms to automate the construction of CoT. In recent years, visual prompts in MLLMs have attracted significant attention (Xu et al., 2025; Yang et al., 2023a). Researchers (Luan et al., 2024) propose pioneering work to implement visual prompt for visual question answer task. However, the text form output of MLLMs significantly constrains accuracy of the prompt.

MLLMs for Document Understanding. Document understanding (Zhang et al., 2024a; Luo et al., 2025) is an increasingly prominent research area driven by the growing industrial demand for efficient information processing. Recent advancements (Chen et al., 2024b; Bai et al., 2025b) in general MLLMs have enhanced the encoding resolution of document images, markedly improving the efficacy of document understanding tasks. A number of MLLMs have been specifically developed to tackle challenges in document understanding. For

instance, mPLUG-DocOwl series (Ye et al., 2023; Hu et al., 2024) unify task processing across five types of document images, achieving notable performance in document comprehension. However, despite the significant progress made by existing MLLMs, some findings indicate that these models do not focus on relevant areas when responding to queries (Yu et al., 2024), especially problematic for complex documents.

Layout-aware Document Understanding. Recent studies incorporate document layout information from several complementary perspectives. GeoLayoutLM (Luo et al., 2023) models geometric relationships among layout entities. Fox (Liu et al., 2024) improves fine-grained regional perception by enabling MLLMs to operate on user-specified document regions through positional visual prompts. LayoutLLM (Luo et al., 2024) moves further toward explicit document reasoning by decomposing document understanding into multiple reasoning steps and representing the reasoning process with textual CoT. Unlike prior methods that focus on layout representation, regional perception, or textual reasoning, Doc-CoB explicitly models query-conditioned layout relevance, and reasons over the layouts that contribute to the downstream task.

3 Methodology

The Doc-CoB framework (Algorithm 1) is designed to address the challenge of balancing key layout regions and global contextual information in document understanding. Similar to human document reading behavior, document processing usually starts with segmenting the layout into “boxes”, and performing knowledge-guided saccadic scanning to select relevant boxes for focused understanding. Inspired by this reading behavior, we define Doc-CoB as a two-stage paradigm: *S1: Key Box Selection* and *S2: Focused Answering*.

3.1 S1: Key Box Selection

Given a document image x_I and a query Q , Doc-CoB uses a layout analyzer $\Lambda(\cdot)$ to generate a set of candidate layout boxes $B = \Lambda(x_I)$, where each box $b \in B$ is assigned a unique index id . Then we construct a visually prompted image X_I^{S1} by overlaying all candidate boxes B on x_I , and labeling each box with its index id . Then the MLLM $f(\cdot)$ selects the subset of boxes B^{key} that are most relevant to Q :

$$\mathcal{I}_{\text{key}} = f([X_I^{S1}; B], Q, P_{S1}), \quad (1)$$

Algorithm 1 Doc-CoB Algorithm

Input: Document image x_I and question Q

Output: Answer A

- 1: $B \leftarrow \Lambda(x_I)$ $\triangleright$ Layout analyzer extracts candidate boxes
 - 2: $X_I^{S1} \leftarrow \text{RENDERID}(x_I, B)$ $\triangleright$ Overlay each box with a unique ID
 - 3: $\mathcal{I}_{\text{key}} \leftarrow \text{SELECTIDS}(f([X_I^{S1}; B], Q, P_{S1}))$ $\triangleright$ Stage 1: select key box IDs
 - 4: $B_{\text{key}} \leftarrow \{b_i \mid i \in \mathcal{I}_{\text{key}}\}$ $\triangleright$ Retrieve key boxes by IDs
 - 5: $X_I^{S2} \leftarrow \text{DRAWBOX}(x_I, B_{\text{key}})$ $\triangleright$ Highlight selected boxes
 - 6: $A \leftarrow f([X_I^{S2}; B_{\text{key}}], Q, P_{S2})$ $\triangleright$ Stage 2: generate answer conditioned on key boxes
 - 7: **return** A
-

$$B^{\text{key}} = \{b \in B \mid id(b) \in \mathcal{I}_{\text{key}}\}. \quad (2)$$

where $\mathcal{I}_{\text{key}} \subseteq \{1, \dots, |B|\}$, and P_{S1} is the prompt used in the key box selection stage (see Appendix A). In practice, we use a layout analyzer and formulate this stage as a multiple-choice prediction problem over the candidate boxes.

3.2 S2: Focused Answering.

After obtaining B^{key} , we perform inference by directing the MLLM’s attention to the selected boxes. Specifically, we generate another visually prompted image X_I^{S2} by marking the selected boxes with red borders, without altering the remaining visual content. This design preserves the global document context while emphasizing relevant layout regions, which is crucial for layout-dependent queries (e.g., “What is the first bullet on the right side?”). The final answer A is produced by

$$A = f([X_I^{S2}; B^{\text{key}}], Q, P_{S2}), \quad (3)$$

where P_{S2} is a short instruction “Please pay more attention to the red boxes”.

3.3 Reasoning Tasks

Beyond proposing a coarse-to-fine framework tailored for document, we also need to enhance the model’s visual and semantic understanding of layouts. However, we observe that the intrinsic capabilities of existing MLLMs are still insufficient to fully support the Doc-CoB framework. First, the model may fail to align a box b with its id in the visually prompted image. Second, even when the box-id mapping is correct, the model frequently struggles to identify the semantic role of each layout box when answering the query.

To address these limitations, we introduce two reasoning tasks: *Box Recognition* and *Box Reasoning*. In the Box Recognition task, the MLLM is

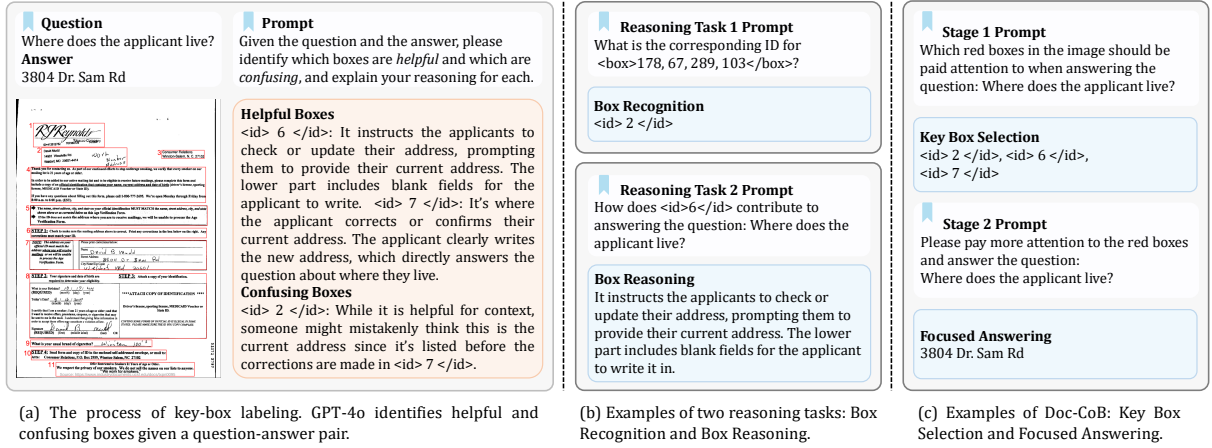


Figure 2: Training data generation pipeline and examples of two reasoning tasks and Doc-CoB training data.

given X^{S1} and a box b_i and is prompted, “What is the index of b_i ?”. It should return the corresponding id_i . In Box Reasoning task, the model is given X^{S1} , a question Q , and an index id_i , and is prompted, “What role does id_i play in answering Q ?”. It should describe the role of b_i with respect to Q . These two reasoning tasks inject layout awareness and query-conditioned reasoning into the model, thereby supporting Doc-CoB reasoning. We construct the training samples for these two reasoning tasks based on nine datasets (Sec. 3.4).

3.4 Training Dataset Generation

Existing document understanding corpora lack the annotations required by Doc-CoB. To address this, we construct an automatic data generation pipeline (Figure 2) that uses GPT-4o (OpenAI, 2024a) as the linguistic annotator and MinerU (Wang et al., 2024a), a layout analysis tool, as the visual annotator. Starting from widely used document datasets, we retain their images X_I , question Q and answer A . For each X_I , we apply MinerU to obtain a set of layout boxes $B = \{b_i\}_{i=1}^{|B|}$. These boxes serve as the visual primitives for subsequent reasoning.

The key annotation step is to identify the subset of key boxes $B^{\text{key}} \subseteq B$. Given (X_I^{S1}, B, Q, A) , GPT-4o returns two disjoint subsets: the *helpful box* B^H and the *confusing box* B^C , such that $B^{\text{key}} = B^H \cup B^C$. It also generates a natural language description $D(b)$ for each $b \in B^{\text{key}}$, explaining its function in answering Q . Boxes in B^H contain the useful evidence for A , while boxes in B^C resemble the query in wording, format, or position but do not contain the answer.

We design a prompt that specifies the task requirements, defines helpful and confusing boxes,

and outlines the output format to ensure consistent annotations. Data quality is maintained through automatic sanity checks, including verification that A is entailed in B_H , as well as targeted manual review. Details of the prompt and quality assurance process are provided in the Appendix B.

Specifically, we select nine widely used document datasets: DocVQA (Mathew et al., 2021), DUDE (Van Landeghem et al., 2023), DeepForm (Borchmann et al., 2021), FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), VRDU-Ad-Buy (Wang et al., 2023), VRDU-Registration-Form (Wang et al., 2023), FeTaQA (Nan et al., 2022), and PubLayNet (Zhong et al., 2019). We directly adopt the question-answer pairs provided by DocVQA and DUDE. We convert the annotations of FUNSD, SROIE, FeTaQA, and PubLayNet into question-answer pairs following Luo et al. (2024), and those of DeepForm, VRDU-Ad, and VRDU-RF following Hu et al. (2024). As a result, we construct the training datasets for Doc-CoB (249,601 samples), Box Recognition (126,970 samples), and Box Reasoning (269,108 samples). Details are in the Appendix C.

4 Experiments

In this section, we investigate four research questions to evaluate the effectiveness and scalability of Doc-CoB through experimentation:

Model performance. Can Doc-CoB improve document understanding performance? (Sec. 4.2)

Ablation analysis. Verify the contribution of each component. (Sec. 4.3, Sec. 4.4)

Generalization. Can Doc-CoB generalize to unseen domains and settings? (Sec. 4.5, Sec. 4.6)

Efficiency and practicality. What is the accu-

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
Qwen2.5-VL-3B*	38.20	94.99	75.94	63.88	68.30	82.43	90.63
Qwen2.5-VL-3B-R	42.11	95.52	77.50	62.73	69.99	82.79	90.79
Qwen2.5-VL-3B-CoB	79.20	96.07	80.55	64.36	88.66	90.83	91.11
DocOwl 1.5*	74.98	95.67	73.33	57.21	88.02	85.61	81.53
DocOwl 1.5-R	77.86	96.35	76.00	57.52	89.71	86.54	82.02
DocOwl 1.5-CoB	79.33	96.52	81.77	57.87	93.35	90.92	82.88
InternVL2-2B*	34.15	87.11	73.79	58.84	54.94	63.09	85.04
InternVL2-2B-R	48.49	88.02	72.33	58.31	70.94	73.94	85.27
InternVL2-2B-CoB	78.57	95.94	81.87	59.92	92.63	90.70	85.61
InternVL2-8B*	36.33	90.39	75.17	64.12	58.66	67.58	90.27
InternVL2-8B-R	48.49	94.31	73.91	63.42	77.77	80.99	91.00
InternVL2-8B-CoB	80.06	97.18	82.95	65.93	93.76	92.64	91.16
<i>Much Larger MLLMs</i>							
InternVL2-40B	43.71	92.03	75.87	65.80	64.81	69.17	93.86
GPT-4o	44.75	91.87	80.21	65.57	78.90	72.96	91.05
<i>Visual Prompting Methods</i>							
TextCoT	42.45	81.62	58.60	43.68	54.58	59.10	68.29
Visual CoT	44.76	88.78	74.90	53.41	69.14	77.98	81.16

Table 1: Model performance (%). **Bold** numbers denote the best within each group. Underline numbers indicate the best per column. * indicates SFT baseline; -R adds reasoning tasks, and -CoB denotes models with Doc-CoB.

#	R1	R2	CoB SFT	CoB Paradigm	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
1					36.33	90.39	75.17	64.12	58.66	67.58	90.27
2	✓	✓			48.49	94.31	73.91	63.42	77.77	80.99	91.00
3			✓	✓	64.17	96.37	79.21	64.32	83.39	84.44	91.09
4		✓	✓	✓	48.55	93.65	71.82	63.78	74.66	80.91	89.16
5	✓		✓	✓	45.82	91.21	72.51	63.34	72.97	77.29	90.79
6	✓	✓	✓		72.17	96.50	80.76	64.88	87.29	85.26	91.12
7	✓	✓	✓	✓	80.06	97.18	82.95	65.93	93.76	92.64	91.16

Table 2: Ablation study based on InternVL2-8B. R1, R2 and CoB SFT mean the models are supervised fine-tuned on these training datasets. CoB Paradigm means the models adopt the two-stage CoB paradigm during inference.

racy–efficiency trade-off when integrating Doc-CoB into existing MLLMs? (Sec. 4.7)

4.1 Experimental Setup

4.1.1 Datasets and Metrics

We evaluate Doc-CoB on seven widely used datasets, categorized into **Document QA** (DocVQA (Mathew et al., 2021), DUDE (Van Landeghem et al., 2023)) and **Document IE** (DeepForm (Borchmann et al., 2021), FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), VRDU-Ad-Buy, and VRDU-Registration-Form (Wang et al., 2023)). Detailed statistics of the test data are provided in the Appendix D. For evaluation metrics, we use Average Normalized

Levenshtein Similarity (ANLS) (Biten et al., 2019) to assess text similarity on DocVQA, DUDE, FUNSD, and SROIE, following Luo et al. (2024). For DeepForm, we report F1 scores following Hu et al. (2024). For VRDU, we adopt micro F1 with type-aware fuzzy matching, as proposed by Wang et al. (2023), to accommodate minor variations in answer formats.

4.1.2 Implementation Details.

We select four MLLMs with varying sizes and architectures: InternVL2-2B (Chen et al., 2024a), InternVL2-8B, Qwen2.5-VL-3B (Bai et al., 2025b), and DocOwl1.5-8B (Hu et al., 2024). The InternVL2 series and Qwen2.5-VL-3B are general-purpose models, and DocOwl1.5-8B is an MLLM

focusing on document understanding.

During training, we adopt a two-stage training strategy: first, we train the models on two reasoning tasks (Sec. 3.3, 396k samples), followed by training on Doc-CoB dataset (Sec. 3.4, 249k samples). Throughout training, the visual encoder of each model remains frozen, and only the language model is updated. InternVL2 series, Qwen2.5-VL, and DocOwl1.5 are trained with learning rates of $1e-6$, $1e-5$, and $2e-5$, respectively, for one epoch on 8 NVIDIA A100 GPUs. All randomness-inducing hyperparameters are disabled during inference to ensure consistent results. For a fair comparison, we fine-tune each model on the training set of the datasets, and report these results as SFT baselines.

4.2 Main Results

Table 1 presents the evaluation results. The results show that Doc-CoB consistently improves performance across all seven datasets and all four MLLMs. This demonstrates that Doc-CoB is a general-purpose solution applicable to a wide range of model sizes and architectures. Notable gains appear on DeepForm, SROIE, FUNSD, and VRDU. For instance, InternVL2-8B-CoB achieves 80.06% on DeepForm, exceeding the 36.33% SFT baseline by more than 43 percentage points. We also present results for variants trained only on the reasoning tasks and evaluated with one-pass inference (-R). These settings still yield clear gains, indicating that the two tasks inject layout awareness and semantic reasoning into the MLLM and thus contribute independently to document-understanding performance.

Comparison with Much Larger MLLMs. We further compare with larger MLLMs, namely the leading open-source model InternVL2-40B and the leading closed-source model GPT-4o (prompts are provided in the Appendix E). With the support of Doc-CoB, the smaller InternVL2-8B-CoB achieves competitive or superior performance: it surpasses InternVL2-40B on six datasets and outperforms GPT-4o on all seven benchmarks, underscoring the effectiveness of our method.

Comparison with Visual Prompt Methods. Several studies investigate visual prompting and demonstrate its effectiveness on natural scene vision tasks, such as Visual CoT (Shao et al., 2024) and TextCoT (Luan et al., 2024). Although effective in natural scenes, these methods have not been explored for document understanding. To address this gap, we implement Visual CoT and TextCoT

using InternVL2-8B. Training and inference procedures strictly follow Sec. 4.1.2. As shown in Table 1, these methods underperform on all benchmarks. We attribute this performance gap to inaccurate ROI grounding, which fails due to neglecting the critical role of document layout.

4.3 Ablation Study

Doc-CoB Design. Table 2 reports the ablation experiment results of the InternVL2-8B. We evaluate a one-pass variant of InternVL2-8B-CoB (#6), which is trained with the CoB dataset but answers directly without the two-stage paradigm. While (#6) surpasses all other baselines (#1~#5) across all benchmarks, it consistently underperforms the complete CoB (#7), indicating that the performance gain primarily stems from the change in reasoning paradigm rather than the training data alone.

Although the results in Table 1 demonstrate the benefits of adding two reasoning tasks, we further investigate a setting where the reasoning tasks are removed. The performance gap between (#3) and the full CoB (#7) suggests that the intrinsic reasoning capabilities of current MLLMs are still insufficient to fully exploit the potential of Doc-CoB. The results show that Doc-CoB’s full effectiveness arises from the combination of explicit reasoning paradigm and targeted supervision.

Impact of Global Context. To demonstrate the necessity of preserving global context in Stage 2, we further evaluate a cropping variant that replaces the whole-page input with the cropped area. As shown in Table 4, cropping consistently degrades InternVL2-8B-CoB across all seven datasets: the drop ranges from 5.42 to 15.26 points, and exceeds 8 points on six datasets.

This suggests that the selected area is insufficient when treated as an isolated snippet. Many document visual question answer tasks rely on global spatial context, such as references to “the middle table”, “the last image”, or “the third paragraph in the right column”. Cropping removes these layout cues and disrupts spatial relationships across regions. Table 4 also shows that cropping reduces visual tokens by 13.06%–44.27%, but still retains 55.73%–86.88% of the original tokens. Therefore, cropping provides only moderate computational savings relative to its substantial accuracy loss, making the visually prompted whole-image input a better trade-off for Stage 2.

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA	Avg.
Qwen2.5-VL-CoB	98.13	99.06	96.18	91.25	97.43	96.88	92.71	95.95
DocOwl1.5-CoB	98.28	99.11	98.43	91.36	97.74	96.67	93.13	96.39
InternVL2-2B-CoB	98.43	98.86	98.13	90.87	96.30	97.15	92.66	96.06
InternVL2-8B-CoB	99.01	99.57	99.23	93.42	98.08	97.79	95.82	97.56

Table 3: Performance on S1: Key Box Selection.

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
<i>Model Performance</i> $\uparrow$							
Cropped Image	66.06	91.76	70.29	57.89	80.26	77.38	82.44
Whole Image	80.06	97.18	82.95	65.93	93.76	92.64	91.16
<i>Visual Token Usage</i> $\downarrow$							
Cropped Image	1483.15 (60.09%)	1292.76 (86.88%)	1535.67 (70.96%)	918.93 (70.40%)	1391.12 (55.73%)	1592.89 (73.90%)	3448.78 (75.95%)
Whole Image	2468.80	1487.00	2163.34	1304.20	2496.32	2154.14	4540.47

Table 4: Comparison of model performance and visual token usage for full-image and cropped inputs in Stage 2.

4.4 Key Box Selection Performance

Given that Doc-CoB adopts a two-stage approach, we isolate Stage 1, Key Box Selection, to assess its impact on the overall performance. As defined in Sec. 3.1, a “helpful box” contains evidence useful for answering a question. Therefore, we report the model’s F1 score for selecting helpful boxes, denoted as $F1_H$. As shown in Table 3, on datasets where Doc-CoB achieves substantial performance improvements (such as DeepForm, SROIE and VRDU), the average $F1_H$ consistently exceeds 97%.

In contrast, on datasets where Doc-CoB has a less pronounced impact, such as DUDE, the $F1_H$ is relatively lower. This suggests that high-quality key-box selection tends to correlate with improved overall performance, although Table 1 also shows that Doc-CoB can still offer benefits even when $F1_H$ is relatively modest. This suggests that high-quality intermediate visual reasoning effectively supports document understanding.

4.5 Zero-shot Performance

To further validate the cross-domain scalability of our method, we conduct zero-shot evaluations of InternVL2-8B-CoB on InfographicVQA (Mathew et al., 2022) and MP-DocVQA (Tito et al., 2023). For MP-DocVQA, we numbered the boxes sequentially and fed all pages into the model for box selection. Then, we removed unselected pages, generated visual prompts based on the selected boxes, and asked the model to answer the question. Since

	InfographicVQA	MP-DocVQA
InternVL2-8B*	74.71	78.29
InternVL2-8B-R	75.74	78.82
InternVL2-8B-CoB	77.14	81.43

Table 5: Zero-shot performance of Doc-CoB.

multiple images were input simultaneously, we had to reduce the resolution hyperparameter to avoid OOM errors. In this setting, the advantage of CoB is that by emphasizing key areas, it can offset the model’s need for high effective resolution, thereby improving performance on multi-page data. The results demonstrate that our model can also be extended to infographics and multi-page documents.

4.6 Robustness to Layout Analyzers

According to Sec. 3.1, Doc-CoB reasoning uses a layout analysis tool to obtain all layout boxes in a given document image. To examine whether Doc-CoB depends on a specific analyzer, we repeat the experiments using two alternative tools. The first, we apply the K-means algorithm to spatially cluster OCR-extracted bounding boxes and is referred to as “OCR + K-means”. The second is Marker¹, an open-source layout analyzer similar to MinerU. All experiments use the InternVL2-8B-CoB checkpoint, and only the analyzer is changed without retraining the model. The results in Table 6 show that Doc-CoB consistently improves MLLM performance regardless of the analyzer used. This

¹<https://github.com/VikParuchuri/marker>

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
InternVL2-8B*	36.33	90.39	75.17	64.12	58.66	67.58	90.27
<i>InternVL2-8B-CoB (layout analyzer variations)</i>							
OCR+K-means	78.56	96.82	83.97	65.03	93.91	93.42	90.77
Marker	79.61	97.03	83.75	65.32	93.01	92.32	90.48
MinerU	80.06	97.18	82.95	65.93	93.76	92.64	91.16

Table 6: Performance of InternVL2-8B-CoB with different layout analyzers.

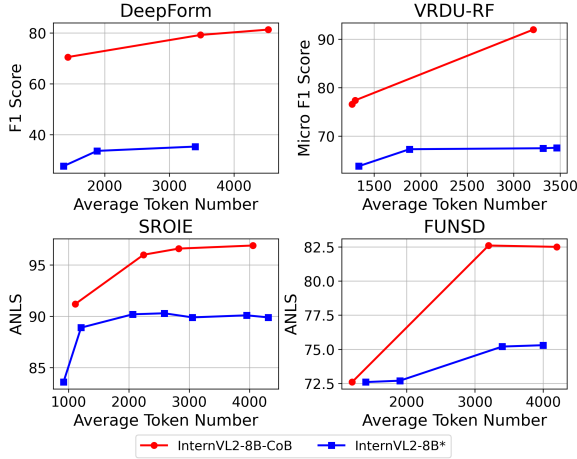


Figure 3: Performance of InternVL2-8B* (blue) and InternVL2-8B-CoB (red) on four datasets, plotted against the average token numbers.

suggests that Doc-CoB is practical in real-world applications, as any analyzer that provides coarse segmentation is sufficient.

4.7 Accuracy-Efficiency Trade-off

It is widely acknowledged that image tokens constitute the majority of input tokens in MLLMs designed for high-resolution images (Chen et al., 2024b; Hu et al., 2024). Doc-CoB reasoning appears to double the total token budget, as two images are processed. However, we hypothesize that Doc-CoB offsets this overhead by guiding the model’s attention to key regions, thereby reducing the effective resolution requirement.

To validate this, we conduct additional experiments using InternVL2-8B* and InternVL2-8B-CoB, adjusting the hyperparameters that control image resolution (i.e., image tokens). For InternVL2-8B*, the token count includes tokens used for vanilla inference, while for InternVL2-8B-CoB, it includes the total input tokens across both stages. As shown in Figure 3, Doc-CoB consistently outperforms vanilla inference while using an equal or even fewer tokens. Although two-stage inference introduces additional latency, accuracy is often

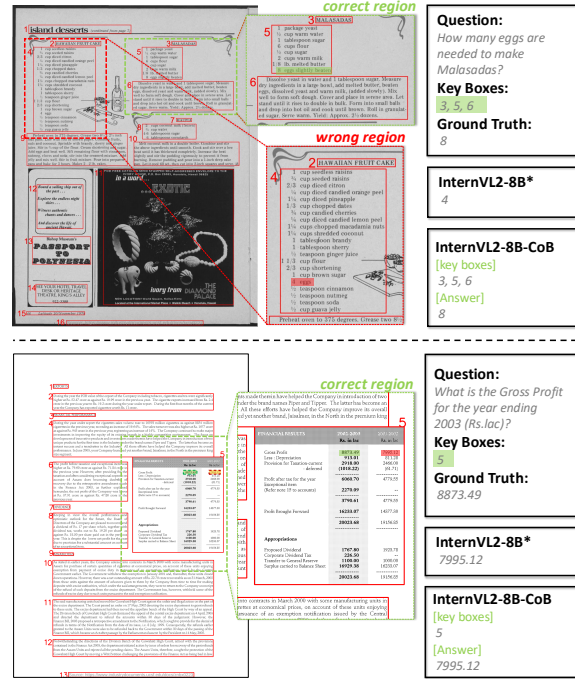


Figure 4: Top: Successful case. Bottom: Failure case.

more important in practical document understanding tasks such as compliance review. Therefore, Doc-CoB’s performance gains justify the moderate additional inference cost.

4.8 Case Study

Figure 4 presents two cases generated by InternVL2-8B* and InternVL2-8B-CoB (Doc-CoB). In the top example, the document image contains substantial irrelevant information. InternVL2-8B* outputs “4”. In contrast, InternVL2-8B-CoB suppresses the noise by selecting only the key boxes and outputs the correct answer “8”. The bottom example exposes a limitation of Doc-CoB, that currently provides no further granularity to resolve intra-box ambiguity. More cases and failure analysis are in the Appendix F.

5 Conclusion

In this paper, we propose Doc-CoB, a coarse-to-fine layout-aware document understanding frame-

work that leverages visual reasoning without modifying model architecture. By formulating cognitive processes into a two-stage MLLM reasoning paradigm, Doc-CoB significantly enhances the document understanding capabilities of MLLMs. We also create a tailored dataset with intermediate visual reasoning supervision and introduce two reasoning tasks to support our approach. Experimental results demonstrate that our method not only outperforms strong baselines but is also transferable across different MLLMs, exhibiting strong generalization capabilities in real-world application.

Limitations

While multiple strategies are adopted during data generation to suppress noise introduced by layout analysis, such noise cannot be fully eliminated at inference time and may accumulate across stages. In addition, the current pipeline performs only a single round of box selection; introducing a self-verification mechanism that iteratively refines the selected key boxes may further improve robustness and performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No: 62372408), Ningbo Natural Science Foundation (Grant No: 2025J001).

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Ali Furkan Biten, Ruben Tito, Andres Mafra, Lluís Gomez, Marçal Rusinol, Minesh Mathew, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. ICDAR 2019 competition on scene text visual question answering. In *ICDAR*.
- Łukasz Borchmann, Michał Pietruszka, Tomasz Stanisławek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Gralinski. 2021. DUE: End-to-end document understanding benchmark. In *NeurIPS Datasets and Benchmarks Track*.
- Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, and 10 others. 2025. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*.
- Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, and Xing Sun. 2023. Attention where it matters: Rethinking visual document understanding with selective region concentration. In *ICCV*.
- Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. Multimodal Language Models See Better When They Look Shallower. In *EMNLP*.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024a. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *SCIS*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*.
- Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Qiang Sun, Geoffrey Martin, Wei Liu, and Yifan Peng. 2026. A survey on MLLM-based visually rich document understanding: Methods, challenges, and emerging trends. In *Findings of ACL*.
- Pei Fu, Tongkun Guan, Zining Wang, Zhentao Guo, Chen Duan, Hao Sun, Boming Chen, Jiayao Ma, Qianyi Jiang, Kai Zhou, and Junfeng Luo. 2025. Multimodal Large Language Models for text-rich image understanding: A comprehensive review. *arXiv preprint arXiv:2502.16586*.
- Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

- Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *AAAI*.
- Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In *Findings of EMNLP*.
- Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *ICDAR*.
- Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDARW*.
- Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3oT: Generating shorter chain-of-thought without compromising effectiveness. In *AAAI*.
- Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025. Structured chain-of-thought prompting for code generation. *TOSEM*.
- Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Focus Anywhere for Fine-grained Multi-page Document Understanding. *arXiv preprint arXiv:2405.14295*.
- Llama Team. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. 2024. TextCoT: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*.
- Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. GeoLayoutLM: Geometric pre-training for visual information extraction. In *CVPR*.
- Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. In *CVPR*.
- Chuwei Luo, Guozhi Tang, Qi Zheng, Cong Yao, Lianwen Jin, Chenliang Li, Yang Xue, and Luo Si. 2025. Bi-VLDoc: Bidirectional vision-language modeling for visually-rich document understanding. *IJDAR*.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *WACV*.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *WACV*.
- Jing Miao, Charat Thongprayoon, Supawadee Supadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*.
- Ye Mo, Gang Huang, Liangcheng Li, Dazhen Deng, Zhi Yu, Yilun Xu, Kai Ye, Sheng Zhou, and Jiajun Bu. 2025. Tablenarrator: Making image tables accessible to blind and low vision people. In *CHI*.
- Ye Mo, Chuan Zhou, Fengxiang Cheng, Jialin Yu, Liangming Pan, Fenrong Liu, Sheng Zhou, Haoxuan Li, Zhouchen Lin, and Philip Torr. 2026. OpenWorld LLM logical reasoning. In *ICML*.
- Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *TACL*.
- OpenAI. 2024a. Gpt-4o system card.
- OpenAI. 2024b. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*.
- Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *ICCV*.
- Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multi-page DocVQA. *PR*.
- Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Joziak, Rafał Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanislawek. 2023. Document understanding dataset and evaluation (DUDE). In *ICCV*.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. MinerU: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.
- Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024b. T-SciQ: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *AAAI*.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, William Wang, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. VRDU: A benchmark for visually-rich document understanding. In *KDD*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Chen Xu, Yuhao Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. 2025. Progressive visual prompt learning with contrastive feature re-formation. *IJCV*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *KDD*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023a. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Lingfeng Yang, Yuezhe Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2023b. Fine-grained visual prompting. In *NeurIPS*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mPLUG-DocOwl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024. Attention prompting on image for large vision-language models. In *ECCV*.

Renshan Zhang, Yibo Lyu, Rui Shao, Gongwei Chen, Weili Guan, and Liqiang Nie. 2024a. Token-level correlation-guided compression for efficient multimodal document understanding. *arXiv preprint arXiv:2407.14439*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal Chain-of-Thought Reasoning in Language Models. *TMLR*.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In *ICDAR*.

Yinan Zhou, Yuxin Chen, Haokun Lin, Yichen Wu, Shuyu Yang, Zhongang Qi, Chen Ma, Li Zhu, and Ying Shan. 2024. DOGR: Towards versatile visual document grounding and referring. *arXiv preprint arXiv:2411.17125*.

A Doc-CoB Stage 1 Prompt

Which red boxes should be paid attention to when answering the following question: `{question}`? Output the box ID near the red box.

B Key Boxes Labeling Prompt and Quality Assurance

You are presented with an image containing multiple pre-labeled boxes, each identified by a unique number. You will receive a series of question-answer (QA) pairs. For each question, determine which labeled boxes in the image contain the information needed to arrive at the given answer, adhering to the following rules:

1. ****If the Number of Boxes in the Image Exceeds 3, Output at Least Three Boxes****
For each question, list all boxes that are genuinely helpful to answer the question. If the number of truly helpful boxes is less than three, please output several boxes that are most likely to cause confusion in answering the question to ensure that at least three boxes are output.

2. ****If the Number of Boxes in the Image is At Most 3, Output Only the Boxes Helpful for Answering the Question****
For each question, list all boxes that are genuinely helpful to answer the question. Do not output boxes that might cause confusion.

3. ****Output Reason and Content****
After listing the boxes, for each box, output the reasons (from semantic, layout, etc. perspectives) why it helps or doesn't help answer the question.

4. ****Other Details****
When referring to any box, use the notation `<box>num </box>`.

5. ****Output Format****
For each QA pair, output only the question ID (Q1, Q2, etc.). Output in this strict format:

Q1:

HELPFUL BOX: [<box>num</box>(s)]

CONFUSING BOX: [<box>num</box>(s)]

Reason for <box>num</box>: **30-50 word explanation**

Q2:

HELPFUL BOX: [<box>num</box>(s)]

CONFUSING BOX: [<box>num</box>(s)]

Reason for <box>num</box>: **30-50 word explanation**

Below is an example of the exact format expected:

Q1: HELPFUL BOX: [<box>16</box>]

CONFUSING BOX: [<box>15</box>, <box>19</box>]

Reason for <box>16</box>: **30-50 word explanation**

Reason for <box>15</box>: **30-50 word explanation**

Q2:

HELPFUL BOX: [<box>2</box>, <box>3</box>, <box>4</box>]

CONFUSING BOX: []

Reason for <box>2</box>: **30-50 word explanation**

Reason for <box>3</box>: **30-50 word explanation**

Reason for <box>4</box>: **30-50 word explanation**

Here are the QA pairs:

{QA_Pairs}

The prompt is utilized for data annotation using GPT-4o, where *QA_Pairs* represent all question-answer pairs corresponding to the given image. The quality assurance process consists of two sequential steps. First, a textual matching rule is applied to validate the format and confirm whether the ground truth appears within the target helpful bounding boxes. Second, if the annotation of a *QA_Pair* does not pass the rule-based filtering stage, we use Qwen2-VL-72B-instruct to classify the remaining data, filtering out samples with obvious reasoning or understanding errors. Samples deemed potentially correct are manually reviewed by four researchers for final inclusion.

C Training Dataset Details

DeepForm (Borchmann et al., 2021) is a document understanding dataset designed for structured information extraction from visually rich documents. It focuses on modeling the hierarchical and relational structure of document elements such as text blocks, tables, and key-value pairs. The dataset supports research on form parsing, layout-aware representation learning, and joint modeling of visual, textual, and structural information.

SROIE (Huang et al., 2019) targets key information extraction from scanned receipt images. It includes annotations for crucial semantic fields such as company name, date, address, and total amount. SROIE is widely used to evaluate end-to-end document understanding systems that combine OCR, layout analysis, and semantic entity recognition.

FUNSD (Jaume et al., 2019) is designed for semantic understanding of scanned forms. It provides word-level annotations with labels and explicit linking relationships between entities, enabling research on both entity classification and entity linking. FUNSD is commonly used to benchmark layout-aware language models for form understanding tasks.

DUDE (Van Landeghem et al., 2023) is a large-scale dataset aimed at extracting structured information from diverse document types. It emphasizes robustness to layout variation and visual noise, making it suitable for studying generalizable document information extraction methods. The dataset supports tasks such as key-value extraction and document-level semantic understanding.

VRDU (Wang et al., 2023) is a visually rich document understanding dataset that integrates textual content with layout and visual cues. It is designed to evaluate models on complex document layouts, including multi-column text and mixed content types. VRDU facilitates research on multimodal representation learning for documents by combining vision and language features.

DocVQA (Mathew et al., 2021) is a benchmark dataset for visual question answering on document images. It requires models to answer natural language questions based on both the textual content and visual layout of documents. DocVQA is widely used to assess document-level reasoning, cross-modal alignment, and the ability of models to locate and interpret relevant information within documents. Detailed statistics are provided in Table 7, Table 8, Table 9.

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
Images	813	626	146	5239	369	1451	10194
Questions	3407	2499	1676	10111	1972	3508	39463

Table 7: Statistics of the number of images and questions in the **original** training datasets across seven datasets. Each column represents a different dataset, and the rows provide counts of images and corresponding questions used in the training phase.

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA	FeTaQA	PubLayNet
Images	4220	3124	1822	15350	2341	4959	78626	39478	7597
Questions	14191	10534	5782	58525	8483	19540	61832	72438	27167

Table 8: Statistics of the number of images and questions in the training datasets used for **reasoning tasks** across nine datasets. Each column represents a different dataset, and the rows provide counts of images and corresponding questions used in the training phase. This includes both Box Recognition Enhancement and Box Reasoning Enhancement.

D Test Dataset Statistics

Details are provided in Table 10.

E GPT-4o and InternVL2-40B Evaluation Prompt

Below is the prompt for InternVL2-40B on **all** datasets.

{question} Answer the question using a single word or phrase.

Below is the prompt for GPT-4o on the **Document QA** datasets.

You are provided with an image of a document and a question related to it. Please carefully read the content of the document and answer the question based solely on the information in the image. The answers to questions are short text spans.

Question: {question}

Directly extract the answer of the question from the document with few words.

Answer:

Below is the prompt for GPT-4o on the **Document IE** datasets.

You are asked to answer questions asked on a document image. The answers to questions are short text spans taken verbatim from the document. This means that the answers comprise a set of contiguous text tokens present in the document.

Question: {question}

Directly extract the answer of the question from the document with few words.

Answer:

F Failure Analysis and Cases of Doc-CoB

Below is our failure analysis, which resulted in two types of errors. **Hallucination Errors (majority):** The MLLM misinterprets the question or struggles to understand the document content, resulting in unrelated outputs. **OCR Recognition Errors (minority):** Due to handwritten fonts, image blurriness, or other issues, the model’s perception fails, and it cannot correctly output the textual answer. Additionally, we observe that InternVL2-8B-CoB corrects 64% of the errors made by InternVL2-8B, and the performance gains of CoB primarily stem from correcting inter-box errors, rather than improving intra-box accuracy. These case statistics align with the main design goal of Doc-CoB. CoB is intended to enhance the MLLM’s coarse-to-fine reasoning ability at the layout level, rather than at the word-level OCR granularity.

We observed that CoB exhibits minimal per-

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
Images	1626	1250	292	10461	736	2880	20293
Questions	10221	7497	5028	30255	5906	10396	117619

Table 9: Statistics of the number of images and questions in the training datasets used for **Doc-CoB** across seven datasets. Each column represents a different dataset, and the rows provide counts of images and corresponding questions used in the training phase.

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
Images	344	347	47	1295	543	2181	1287
Questions	1423	1388	467	2551	3014	5318	5188

Table 10: Statistics of the number of images and questions in the test datasets across seven datasets. Each column represents a different dataset, and the rows provide counts of images and corresponding questions used in the testing phase.

formance improvement on DUDE and DocVQA, but achieves significant gains on DeepForm; therefore, we conducted a detailed analysis of the case and found that most of the errors on DUDE and DocVQA are not related to key boxes selection. In other words, the model correctly selected the box ID in the first stage, but due to the noise or complex content within the box, it caused the model to misinterpret the information, resulting in incorrect answers. Specifically, among all incorrect cases of DocVQA, only 14.98% involve wrong box selection, compared to 17.38% (DUDE), 21.74% (VRDU-RF), and 29.58% (DeepForm). This helps explain why the model shows only a small improvement over the baseline on DocVQA (+0.89), a larger improvement on DUDE (+1.81), and substantial improvements on the other five datasets (averaging +23.69).



Question:

What is the heading of the page?

Key Boxes:

1, 2

Ground Truth:

ITC's Brands: Creating Sustainable Livelihoods through a Robust Distribution Network

Doc-CoB:

[key boxes]

1, 2

[Answer]

ITC's Brands: Creating Sustainable Livelihoods through a Robust Distribution Network

Proposed Budget for One Year Period			
Personnel	Effort	Salary	fringe Benefits + Total
Dr. H. S. Bhatnagar, Principal Investigator	10	-	-
Yee Ping Huang, Res. Assoc.	20	7,710	8,200
Yamir C. Lee, Res. Tech.	20	4,200	4,400
S. C. Lee, Animal Tech.	20	3,450	3,650
Total Personnel:			\$ 20,250
Supplies			
Animal purchases	200		4,100
Maintenance	100		1,100
Special diets	100		1,100
Defecations	100		1,100
Miscellaneous laboratory supplies	1,000		1,100
Total Supplies:			8,500
Total Direct Costs:			28,750
Overhead 15%:			4,313
Total Amount Requested:			33,063
Fringe Benefits 20%			

Question:

Who is receiving the highest 'salary'?

Key Boxes:

3, 4, 5, 6

Ground Truth:

Yee Ping Huang

Doc-CoB:

[key boxes]

4, 5, 6

[Answer]

Yee Ping Huang

Figure 5: Cases showing intermediate and final outputs of Doc-CoB.

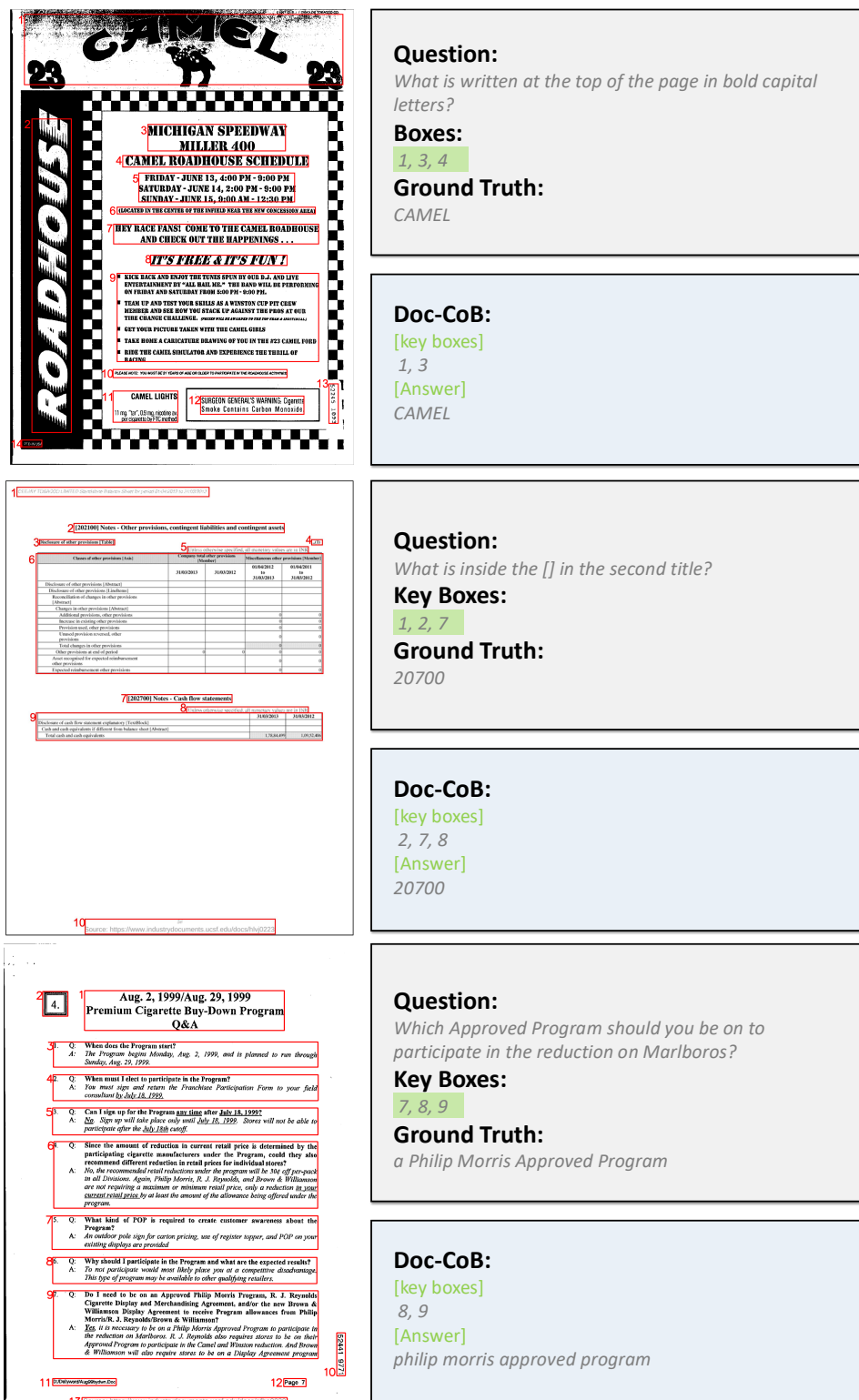


Figure 6: Cases showing intermediate and final outputs of Doc-CoB.

Figure 7: Cases showing intermediate and final outputs of Doc-CoB.

1. The recorder in the original design was replaced by a 0-100 S2P94 mass flow meter (Hassag 42008) to provide more precise flow readings. As in the original design, flow is set below each needle run by manual adjustment of a 3/8" needle valve (Whitney S540X7Y12-F14) and the flowmeter is bypassed during smoke measurements to minimize contamination. In the new design the mass flowmeter is bypassed during smoke runs by the parallel combination of a 3/8" needle valve (Whitney S540X7Y12-F14) and a length of aeromated 1/2" piping. The flowmeter is adjusted to give an identical pressure drop (as indicated by a 0-25 cm H₂O Dwyer Magnehelic gauge) to that of the mass flowmeter, this maintaining the flow when the meter is removed from the circuit.

2. The single run sensitivity potentiometer supplied on the Otek #0075 amplifier was replaced with a ten turn potentiometer to allow more precise settings of the instrumentation level.

3. To reduce noise and jitter in the computer data acquisition system, the output voltage divider internal to the Otek amplifier was altered to increase the instrument's output by a factor of 100. The sensitivities of the computer data system and the parallel only chart recorder were correspondingly modified.

4. Signal leads were re-routed and shielded to minimize ground loops and RF pickup. The exposed terminal board was eliminated.

5. A warning display of output voltage was added to the computer screen as a visual aid to the operator.

6. Changes to the computer program were made by Randy Grosse to eliminate minor operational problems. The changes did not affect the calculation of any reported parameters.

7. P. N. Grosse
A. C. Grosse
A. C. Lidy
D. D. McRae
J. E. Swales
J. E. Wickham
Control Files

12. 13. <https://www.industrydocuments.ucsf.edu/docs/qv0001>

Question:

How many points are there in modifications to readout instrumentation?

Boxes:

3, 4, 5, 6, 7, 8

Ground Truth:

5

Doc-CoB:

[key boxes]

3, 4, 5, 6, 7, 8

[Answer]

5

1. RESTAURANT
TOM ZONE'S ALUM TOWNAL CITY
2400 CAMPBELL ST. (CROSS-ST.)
LOT 1212, ALUM TOWNAL CITY
NO. 1, ALUM TOWNAL CITY
ALUM TOWNAL CITY

2. 1014 TABLE
TBL43 / 1

3. 1014 TABLE
TBL43 / 1

4. 1014 TABLE
TBL43 / 1

5. 1014 TABLE
TBL43 / 1

6. 1014 TABLE
TBL43 / 1

7. 1014 TABLE
TBL43 / 1

8. 1014 TABLE
TBL43 / 1

9. 1014 TABLE
TBL43 / 1

10. 1014 TABLE
TBL43 / 1

11. 1014 TABLE
TBL43 / 1

12. 1014 TABLE
TBL43 / 1

13. 1014 TABLE
TBL43 / 1

Question:

What is the "company" in the given document?

Key Boxes:

1, 5

Ground Truth:

GRAND COMPANIONS SDN BHD

Doc-CoB:

[key boxes]

1, 2

[Answer]

GRAND COMPANIONS SDN BHD

1. RIVERIDE REGION (Short Term Problem)

2. R.R. Performance

3. The Riverside region is a short term problem market for R.R. After accounting for industry volume change, Riverside's contribution to share was .019 below the expected level for the five month period ending April, 1984. All brands except COME contributed less than expected.

4. RIVERIDE REGION PERFORMANCE

5. 1984 EXPECTED VS. ACTUAL CONTRIBUTION TO SHARE (Including Industry Volume Effects)

6. 1984 COMPETITIVE PERFORMANCE

7. RIVERIDE REGION PERFORMANCE

8. Factors negatively affecting R.R.'s performance in Riverside during the period include:

9. (Strong performance by Marlboro and Playera)

10. (Solid performance by Newport and Marl)

11. 1984 COMPETITIVE PERFORMANCE

12. 1984 COMPETITIVE PERFORMANCE

13. 1984 COMPETITIVE PERFORMANCE

Question:

Which is the only RJR brand that contributed more than expected (CTS)?

Key Boxes:

2, 5, 6

Ground Truth:

Camel

Doc-CoB:

[key boxes]

4, 5, 6

[Answer]

Camel

Figure 8: Cases showing intermediate and final outputs of Doc-CoB.