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

Anonymous ACL submission

Abstract

Document understanding aims to perform question answering and information extraction over document images, where the visual content is highly information-dense and most task queries depend on only a few relevant regions. However, existing methods either adopt a one-pass strategy that implicitly assumes all regions are equally important, or focus excessively on small regions at the cost of losing critical layout information, leading to unfaithful responses. Following the human reading pattern, we introduce Doc-CoB (Chain-of-Box), a simple-yet-effective mechanism that integrates coarse-to-fine visual reasoning into MLLM without modifying its architecture. Our method allows the model to autonomously select the set of layouts most relevant to the query, and then focus on them for further understanding. To support this paradigm, we design two enabling tasks that improve box identification and box–query reasoning, facilitating layout-aware document understanding. We also design an automatic pipeline, integrating a commercial MLLM with a layout analyzer, to generate 249k training samples with intermediate visual reasoning supervision. Extensive experiments on seven benchmarks with four popular models show that Doc-CoB significantly improves performance, demonstrating its effectiveness and wide applicability. All code, data, and models will be released.

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., 2025; Comanici et al., 2025; Grattafiori et al., 2024; Cai et al., 2025) have demonstrated promising capabilities on document understanding (Hu et al., 2024; Ding et al., 2025; Zhou et al., 2024; Luo et al., 2024), 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 visual regions are equally important. However, this overlooks the information density (Cui et al., 2021) of document images, where only a subset of regions is task-relevant (Cao et al., 2023), while the remainder is largely redundant. Consequently, this uniform strategy is prone to producing unfaithful responses and hallucinations. As shown in Figure 1 (a), while not all regions are relevant to the query *"Where does the applicant live?"*, the model derived its answer from a wrong region.

Despite the emergence of coarse-to-fine strategies in general visual QA (Yu et al., 2025), many existing approaches have drifted toward the opposite extreme. Specifically, some works perform direct cropping or zoom-in on selected regions after a coarse localization stage (Luan et al., 2024; Shao et al., 2024). While such designs are effective in reducing 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 accurate and faithful document understanding?*

To address the above challenge, we first propose **Doc-CoB** (Chain-of-Boxes), a mechanism following human reading behavior that enables MLLMs to perform document understanding in a coarse-to-fine reasoning manner without modifying their architecture. We formulate cognitive processes into a two-stage MLLM reasoning paradigm consisting of *(S1) Key Box Selection* and *(S2) Focused An-*

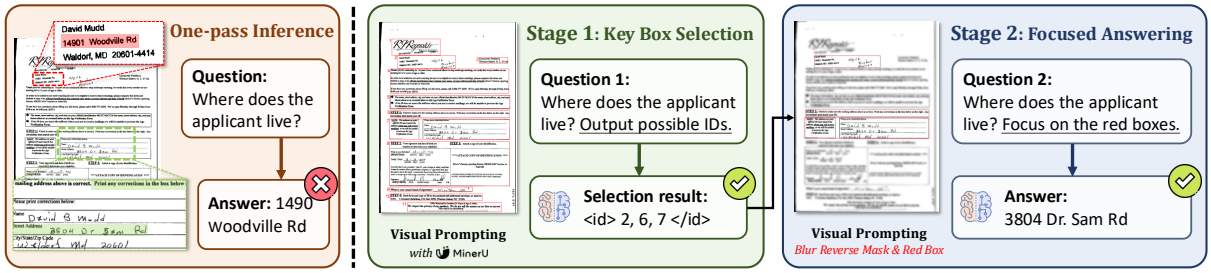


Figure 1: (a) Current MLLM attends to a wrong region and produces an unfaithful answer, instead of obtaining the answer from the correct region below. (b) Overview of Doc-CoB. The layout analyzer segments the document into boxes. (S1) Key box selection identifies boxes relevant to the query, followed by blur-reserve masking. (S2) Focused answering generates the final answer from the masked image.

swering as illustrated in Figure 1 (b). Both stages employ visual prompting (Yang et al., 2024b; Shtedritski et al., 2023) to explicitly guide the model's attention to key layout (box) while suppressing visual redundancy. We then introduce two enabling tasks: *(E1) Box-ID Recognition Enhancement* and *(E2) Box-Query Reasoning Enhancement*, which aim to improve the model's understanding of document layout and its associated semantic information. Finally, we develop an automatic data generation pipeline to address the challenge of limited training data, as existing document datasets lack layout-level annotations that jointly encode visual structure and semantic roles. This pipeline enables us to collect 249K training samples from nine document datasets with intermediate visual supervision tailored to Doc-CoB.

Extensive experiments examine Doc-CoB from multiple complementary perspectives. All MLLMs achieve improved performance when integrated with Doc-CoB, with a 8B model even surpassing GPT-4o. Further ablation analysis reveal that the explicit Doc-CoB reasoning paradigm and the proposed enabling tasks are both critical. Beyond in-domain evaluation, Doc-CoB exhibits strong generalization, remaining effective under zero-shot settings and robust to different layout analyzers. Our main contributions are as follows:

• We introduce Doc-CoB, a mechanism that enables MLLMs to perform document understanding in a human-like, coarse-to-fine reasoning manner.

• We construct a dataset of 249k samples across nine document datasets, each annotated with intermediate visual supervision tailored for Doc-CoB. In addition, we propose two enabling tasks to further enhance the reasoning capabilities of MLLMs.

• Doc-CoB improves performance on seven benchmarks, surpasses strong baselines, and remains compatible with various MLLM architectures and scales. All training data and models will be released publicly to support further research.

2 Related Work

Multimodal Reasoning. A critical step in advancing AI toward human-level intelligence is enabling the transition from basic perception to complex cognitive reasoning (Wang et al., 2025). Recent works (Jaech et al., 2024; Guo et al., 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 LLM to clarify reasoning, specifically by adding logical thinking processes (Kang et al., 2025; Li et al., 2025; Miao et al., 2024). Meanwhile, progress has been made in reasoning for MLLM. MM-CoT (Zhang et al., 2023) proposed a two-stage reasoning framework by firstly 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. However they are not specially designed for document characteristics.

Visual Prompting in MLLMs. In recent years, with the advancements in MLLMs, visual prompt techniques have attracted great attentions (Xu et al., 2025; Parmar et al., 2025). Visual prompt is a technique used in image-language tasks, where markers such as colorful boxes or circles are added onto an image to direct the attention of MLLMs toward the desired targets while preserving the global context. Nonetheless, compared to natural scene images, visual prompting in document image designs are rarely explored. Researchers (Luan et al., 2024) propose the pioneer work, TextCoT, to implement visual prompt for document understanding task. It

first requires the model to locate key area in the input document and zoom-in by cropping. However, the text form output of MLLMs significantly constrains accuracy of the prompt.

MLLMs for Document Understanding. Document understanding(Zhang et al., 2024; Luo et al., 2025, 2023, 2024) is an increasingly prominent research area driven by the growing industrial demand for efficient information processing. Recent advancements (Chen et al., 2024b; Yang et al., 2024a) 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 do when responding to queries (Yu et al., 2025), especially problematic for complex documents.

3 Methodology

3.1 Doc-CoB Reasoning Paradigm

Many studies in psychology and cognitive science (Treisman et al., 1964; Yli-Krekola et al., 2009), show that when the brain processes information, humans instinctively focus on certain stimuli while ignoring others. Specifically, in the context of document processing, we usually first segment the layout into semantic "boxes", then use saccades to scan the boxes based on prior knowledge and select boxes for focused attention. Following those manners, we define Doc-CoB as a two-stage paradigm: *(S1) Key Box Selection* and *(S2) Focused Answering*. Algorithm 1 provides the detail.

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:

$$B = \Lambda(x_I), \tag{1}$$

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:

$$\mathcal{I}_{\mathrm{key}} = f\big([X_I^{\mathrm{S1}}; B],\, Q,\, P_{\mathrm{S1}}\big), \tag{2}$$

Algorithm 1 Doc-CoB Reasoning Paradigm

Input: Document image x_I and question Q
Output: Answer A

$B \leftarrow \Lambda(x_I)$ ▷ *Layout analyzer extracts candidate boxes*
$X_I^{S1} \leftarrow \textsc{RenderID}(x_I, B)$ ▷ *Overlay each box with a unique ID*
$\mathcal{I}_{\mathrm{key}} \leftarrow \textsc{SelectIDs}\big(f([X_I^{S1}; B], Q, P_{S1})\big)$ ▷ *Stage 1: select key box IDs*
$B_{\mathrm{key}} \leftarrow \{b_i \mid i \in \mathcal{I}_{\mathrm{key}}\}$ ▷ *Retrieve key boxes by IDs*
$X_I^{S2} \leftarrow \textsc{BlurMask}(x_I, B_{\mathrm{key}})$ ▷ *Blur non-key areas and highlight selected boxes*
$A \leftarrow f([X_I^{S2}; B_{\mathrm{key}}], Q, P_{S2})$ ▷ *Stage 2: generate answer conditioned on key boxes*
return A

$$B^{\mathrm{key}} = \{\, b \in B \mid id(b) \in \mathcal{I}_{\mathrm{key}} \,\}. \tag{3}$$

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

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} using Blur Reverse Mask (Yang et al., 2024b), which blurs all areas outside B^{key} while highlighting the selected boxes with a red border. This helps the model focus on relevant layout while preserving global context, 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

$$A = f\big([X_I^{\mathrm{S2}}; B^{\mathrm{key}}],\, Q,\, P_{\mathrm{S2}}\big), \tag{4}$$

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

3.2 Enabling Tasks

Beyond proposing a coarse-to-fine paradigm 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 Doc-CoB reasoning. First, the MLLM 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 MLLM frequently struggles to identify the semantic role of each layout box when answering the query.

To address these limitations, we introduce two enabling tasks: *Box–ID Recognition Enhancement* and *Box–Query Reasoning Enhancement*. In the first task, the MLLM is given X^{S1} and a box b_i and

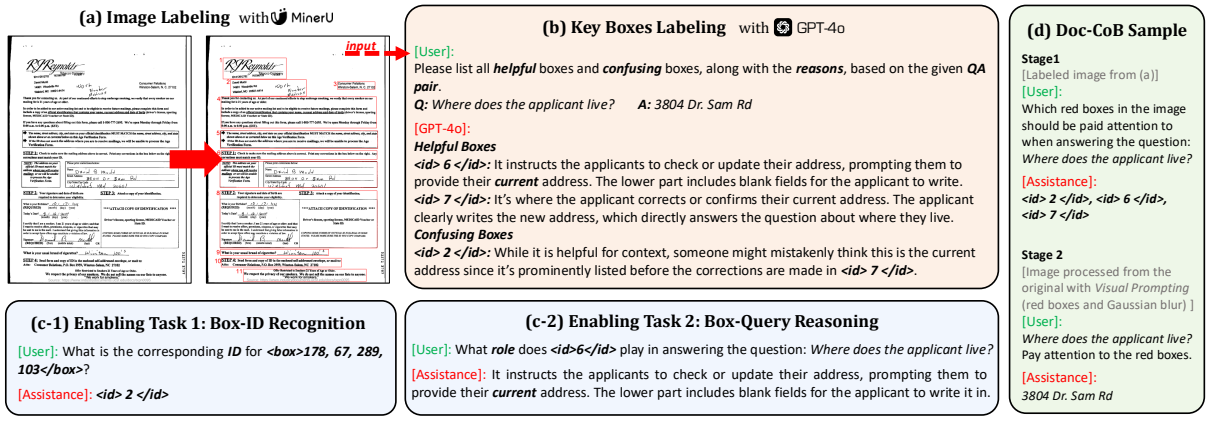


Figure 2: Automatic data generation pipeline. (a) Image labeling: Layout analyzer segments document into boxes. (b) Key-box labeling: GPT-4o identifies helpful and confusing boxes given a question-answer pair. (c-1) Task 1: Box-ID Recognition. (c-2) Task 2: Box-Query Reasoning. The model explains how the specified box contributes to answering the question. (d) Doc-CoB Training Sample.

is prompted, *"What is the index of b_i?"*. It should return the corresponding id_i. In the second 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 must describe the role of b_i with respect to Q. These two enabling tasks inject layout awareness and query-conditioned reasoning into the MLLM, thereby supporting Doc-CoB inference. We construct the training samples for these two enabling tasks based on nine datasets (Sec. 3.3), resulting in 126,970 samples for Box–ID Recognition Enhancement and 269,108 samples for Box–Query Reasoning Enhancement.

3.3 Data 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 a reliable commercial MLLM as the linguistic annotator and a robust 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 (Wang et al., 2024a), an high-precision layout analysis tool, to obtain a set of layout boxes $B = \{b_i\}_{i=1}^{|B|}$. These boxes serve as the visual primitives for subsequent reasoning.

As defined in Sec. 3.1, the key annotation step is to identify the subset of key boxes $B^{\mathrm{key}} \subseteq B$. This task requires complex multimodal semantic understanding, we use GPT-4o as the linguistic annotator. Given $(X_I^{\mathrm{S1}}, B, Q, A)$, where X_I^{S1} is defined in Sec. 3.1, GPT-4o returns two disjoint subsets: the *helpful box* B^H and the *confusion box* B^C, such that $B^{\mathrm{key}} = B^H \cup B^C$. It also generates a natural language description $D(b)$ for each $b \in B^{\mathrm{key}}$, explaining its function in answering Q. Boxes in B^H contain the exact 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 comprehensive 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 A.

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-adbuy (Wang et al., 2023a), VRDU-Registration-Form (Wang et al., 2023a), 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 Doc-CoB training dataset, which contains 249,601 samples. Each query is annotated with an average of three key boxes. Detailed dataset information are provided in the Appendix D.

4 Experiments

In this section, we investigate four key 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 efficiency–effectiveness 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., 2023b)). Detailed statistics of the test data are provided in the Appendix E. 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. (2023b), to accommodate minor variations in answer formats.

4.1.2 Implementation Details

We select four MLLMs with varying sizes and architectures: InternVL2-2B[1] (Chen et al., 2024a), InternVL2-8B[2], Qwen2.5-VL-3B[3] (Bai et al., 2023), and DocOwl1.5-8B[4] (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 the document understanding.

During training, we initialize all models with weights from HuggingFace. We adopt a two-stage training strategy: first, we train the models on the enabling task (Sec. 3.2, 126k samples), followed by training on Doc-CoB dataset (Sec. 3.3, 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 eight NVIDIA A100 GPUs. All randomness-inducing hyperparameters are disabled during inference to ensure consistent results. Additionally, MinerU is used as the layout analysis tool in Doc-CoB reasoning.

4.2 Main Results

Table 1 presents the evaluation results. For a fair comparison, we fine-tune each model on the training set of the datasets introduced in Sec. 4.1.1, and report these results as SFT baselines.

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 80.05% on DeepForm, exceeding the 36.33% SFT baseline by more than 43%. We also present results for variants trained only on the enabling tasks and evaluated with one-pass inference (-E). 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.

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 C). 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.

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 re-implement Visual CoT and TextCoT using InternVL2-8B, treating the helpful box (Section 3.3) as the ROI. Training and inference pro-

[1] https://huggingface.co/OpenGVLab/InternVL2-2B

[2] https://huggingface.co/OpenGVLab/InternVL2-8B

[3] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct

[4] https://huggingface.co/mPLUG/DocOwl1.5

	DeepForm	SROIE	FUNSD	DUDE	VRDU-Ad	VRDU-RF	DocVQA
Qwen2.5-VL*	38.20	94.99	75.94	63.88	68.30	82.43	90.63
Qwen2.5-VL-E	42.11	95.52	77.50	62.73	69.99	82.79	90.79
Qwen2.5-VL-CoB	**79.18**	**96.05**	**80.57**	**64.35**	**88.65**	**90.83**	**91.11**
DocOwl 1.5*	74.98	95.67	73.33	57.21	88.02	85.61	81.53
DocOwl 1.5-E	77.86	96.35	76.00	57.52	89.71	86.54	82.02
DocOwl 1.5-CoB	**79.33**2nd	**96.51**2nd	**81.78**2nd	**57.89**	**93.36**2nd	**90.91**2nd	**82.89**
InternVL2-2B*	34.15	87.11	73.79	58.84	54.94	63.09	85.04
InternVL2-2B-E	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-E	48.49	94.31	73.91	63.42	77.77	80.99	91.00
InternVL2-8B-CoB	**80.05**1st	**97.17**1st	**82.92**1st	**65.90**1st	**93.66**1st	**92.60**1st	**91.17**2nd
Much Larger MLLMs							
InternVL2-40B	43.71	92.03	75.87	65.80 2nd	64.81	69.17	93.86 1st
GPT-4o	44.75	91.87	80.21	65.57	78.90	72.96	91.05
Visual Prompting Methods							
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. 1st and 2nd indicate the best and second best per column. * indicates SFT baseline; -E adds enabling tasks, and -CoB denotes models with Doc-CoB.

#	E1	E2	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 2nd	96.50 2nd	80.76 2nd	64.88 2nd	87.29 2nd	85.26 2nd	91.12 2nd
7	✓	✓	✓	✓	80.05 1st	97.17 1st	82.92 1st	65.90 1st	93.66 1st	92.60 1st	91.17 1st

Table 2: Ablation study based on InternVL2-8B. E1, E2 and CoB SFT means the models are supervised fine-tuned on these training datasets. CoB Paradigm means the models adopt the two-stage CoB paradigm during inference.

cedures strictly follow Sec. 4.1.2. As shown in Table 1, these natural-scene 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

Table 2 reports the ablation experiment results of the InternVL2-8B. We first isolate the effect of the Doc-CoB reasoning paradigm itself. 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 baseline (#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 enabling tasks, we further investigate a setting where the enabling 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.

4.4 Key Box Selection Performance

Given that Doc-CoB adopts a two-stage approach, we isolate Stage 1, Key Box Selection, to assess

	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.

	InfographicVQA	MP-DocVQA
InternVL2-8B*	74.71	78.29
InternVL2-8B-E	75.74	78.82
InternVL2-8B-CoB	**77.13**	**81.45**

Table 4: Zero-shot performance.

its impact on the overall performance. As defined in Sec. 3.1, a "helpful box" contains the exact evidence required to answer 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 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.

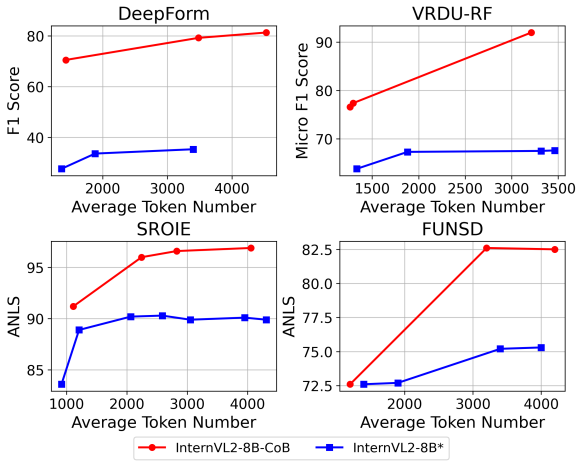


Figure 3: Performance of InternVL2-8B* (blue) and InternVL2-8B-CoB (red) on four datasets, plotted against the average token numbers.

4.6 Discussion of 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, following Syakur et al. (2018), applies the K-means algorithm to spatially cluster OCR-extracted bounding boxes and is referred to as "OCR + K-means". The second is Marker[5], 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 5 show that Doc-CoB consistently improves MLLM performance regardless of the analyzer used. These findings suggest that Doc-CoB is practical in real-world applications, as any mainstream analyzer that provides coarse segmentation is sufficient.

4.7 Efficiency-Effectiveness 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.,

[5] 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
InternVL2-8B-CoB (layout analyzer variations)							
OCR+K-means	78.57	96.80	83.96 1st	65.08	93.90 1st	93.40 1st	90.79 2nd
Marker	79.62 2nd	97.02 2nd	83.77 2nd	65.32 2nd	93.11	92.45	90.50
MinerU	80.05 1st	97.17 1st	82.92	65.90 1st	93.66 2nd	92.60 2nd	91.17 1st

Table 5: Performance of InternVL2-8B-CoB with different layout analyzer.

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 on DeepForm, VRDU-RF, SROIE, and FUNSD 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. These findings indicate that Doc-CoB not only enhances efficiency but also has substantial potential to improve effectiveness.

4.8 Case Study

To further investigate the effectiveness of Doc-CoB, 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. Faced with the question *"How many eggs are needed to make Malasadas?"* InternVL2-8B* applies one-pass attention, is distracted by visually salient but semantically irrelevant numerals that appear with the word *"eggs"*, and therefore outputs *"4"*. In contrast, InternVL2-8B-CoB, supported by Doc-CoB, suppresses this noise by selecting only the key boxes (3, 5, 6)—the three blocks headed *"MALASADAS"*. The model then focuses on these regions and returns the correct answer. This case confirms that Doc-CoB reliably filters noise and directs attention to key regions. The bottom example exposes a current limitation of Doc-CoB. Although InternVL2-8B-CoB correctly identifies the relevant financial table, it still produces an incorrect answer. The selected box aggregates several numerical entries,

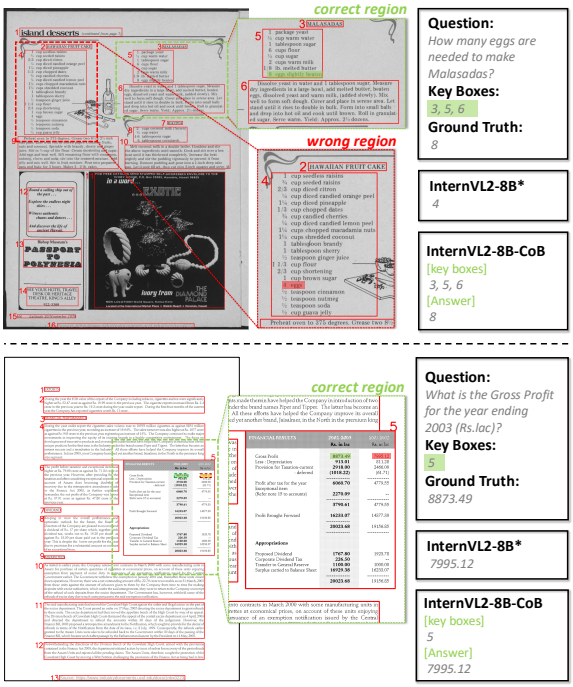


Figure 4: Top: Successful case. Bottom: Failure case.

and Doc-CoB currently provides no further granularity to resolve such intra-box ambiguity. More cases and failure analysis are presented in the Appendix F.

5 Conclusion

In this paper, we propose Doc-CoB, a human-style, coarse-to-fine document understanding mechanism that leverages visual reasoning manner without modifying model architecture. By formulate 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 enabling 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 layout analyzers.

Limitations

During inference, our method relies on an external layout analyzer for document layout detection. A promising direction for future work is to incorporate layout detection into the MLLM itself, enabling end-to-end layout-aware reasoning and reducing reliance on external tools. While multiple strategies are adopted during data generation to suppress noise introduced by layout analysis (Sec. 3.3), such noise cannot be fully eliminated at inference time and may accumulate across stages, making the performance of Doc-CoB partially dependent on the quality of the analyzer. 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.

References

Jinze Bai, Shuai Bai, and 1 others. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Ali Furkan Biten, Ruben Tito, and 1 others. 2019. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE.

Łukasz Borchmann, Michał Pietruszka, and 1 others. 2021. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19517–19527.

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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6677–6695, Suzhou, China. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, and 1 others. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Zhe Chen, Jiannan Wu, and 1 others. 2024b. 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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Lei Cui, Yiheng Xu, and 1 others. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Geoffrey Martin, and Yifan Peng. 2025. A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. *arXiv preprint arXiv:2507.09861*.

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. *Preprint*, arXiv:2502.16586.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Liqi He, Zuchao Li, and 1 others. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18180–18187.

Anwen Hu, Haiyang Xu, and 1 others. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Zheng Huang, Kai Chen, and 1 others. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Guillaume Jaume, Hazim Kemal Ekenel, and 1 others. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24312–24320.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23.

Bozhi Luan, Hao Feng, and 1 others. 2024. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*.

Chuwei Luo, Changxu Cheng, and 1 others. 2023. Geolayoutlm: Geometric pre-training for visual information extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7092–7101.

Chuwei Luo, Yufan Shen, and 1 others. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640.

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. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–12.

Minesh Mathew, Viraj Bagal, and 1 others. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and 1 others. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1):148.

Linyong Nan, Chiachun Hsieh, and 1 others. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Gaurav Parmar, Or Patashnik, Kuan-Chieh Wang, Daniil Ostashev, Srinivasa Narasimhan, Jun-Yan Zhu, Daniel Cohen-Or, and Kfir Aberman. 2025. Object-level visual prompts for compositional image generation. *arXiv preprint arXiv:2501.01424*.

Hao Shao, Shengju Qian, and 1 others. 2024. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*.

Aleksandar Shtedritski, Christian Rupprecht, and 1 others. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997.

Muhammad Ali Syakur, B Khusnul Khotimah, EMS Rochman, and Budi Dwi Satoto. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, volume 336, page 012017. IOP Publishing.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multi-page docvqa. *Preprint*, arXiv:2212.05935.

Anne M Treisman and 1 others. 1964. Selective attention in man. *British Medical Bulletin*.

Jordy Van Landeghem, Rubén Tito, and 1 others. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Bin Wang, Chao Xu, and 1 others. 2024a. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Lei Wang, Yi Hu, and 1 others. 2024b. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170.

Wenjin Wang, Yunhao Li, and 1 others. 2023a. Layout and task aware instruction prompt for zero-shot document image question answering. *arXiv preprint arXiv:2306.00526*.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, 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, and 1 others. 2023b. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23. ACM.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. 2025. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, 133(2):511–526.

Yiheng Xu, Minghao Li, and 1 others. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

An Yang, Baosong Yang, and 1 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Lingfeng Yang, Yueze Wang, and 1 others. 2024b. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36.

Jiabo Ye, Anwen Hu, and 1 others. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Antti Yli-Krekola, Jaakko Särelä, and Harri Valpola. 2009. Selective attention improves learning. In *Artificial Neural Networks–ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part II 19*, pages 285–294. Springer.

Runpeng Yu, Weihao Yu, and 1 others. 2025. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pages 251–268. Springer.

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

Zhuosheng Zhang, Aston Zhang, and 1 others. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Xu Zhong, Jianbin Tang, and 1 others. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

Yinan Zhou, Yuxin Chen, and 1 others. 2024. Doge: Towards versatile visual document grounding and referring. *arXiv preprint arXiv:2411.17125*.

A 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 Less Than 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:

• A textual matching rule is first applied to validate the format and confirm whether the ground truth appears within the target helpful bounding boxes.

• If the annotation of a *QA_Pair* does not pass the rule-based filtering stage, we use Qwen2-VL-72B-instruction 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.

B Doc-CoB Stage 1 Prompt

Which red box in the given image contains the answer to the following question: {*question*}? Use the box ID near the red box to answer the question.

C GPT-4o & 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:

D 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., 2023b) 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 6, Table 7, Table 8.

E Test Dataset Statistics

Details are provided in Table 9.

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 performance 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

	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 6: 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 7: Statistics of the number of images and questions in the training datasets used for **enabling 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–ID Recognition Enhancement and Box–Query Reasoning Enhancement.

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 a relatively small improvement over the baseline on DocVQA (+0.9), better on DUDE (+1.78), and significantly well on other five datasets (averaging +23.65).

	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

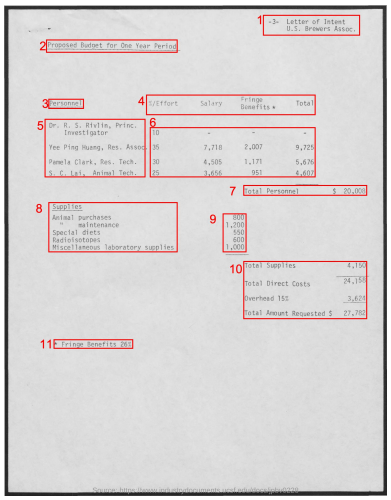
Table 8: Statistics of the number of images and questions in the training datasets used for **Doc-CoB** across seven datasets. Each column represents a different dataset, and the rows provide counts of images and corresponding questions used in the training phase.

	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 9: 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.

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

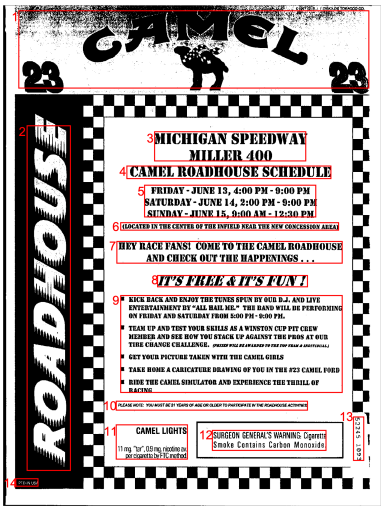
Doc-CoB:
[key boxes]
1, 2
[Answer]
ITC's Brands: Creating Sustainable Livelihoods through a Robust Distribution Network



Question:
Who is receiving the highest 'salary'?
Key Boxes:
3, 4, 5, 6
Ground Truth:
Yee Ping Huang

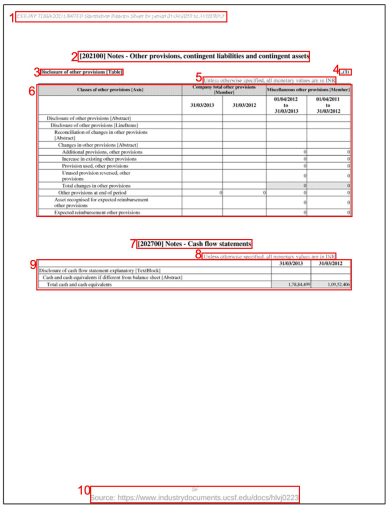
Doc-CoB:
[key boxes]
4, 5, 6
[Answer]
Yee Ping Huang

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



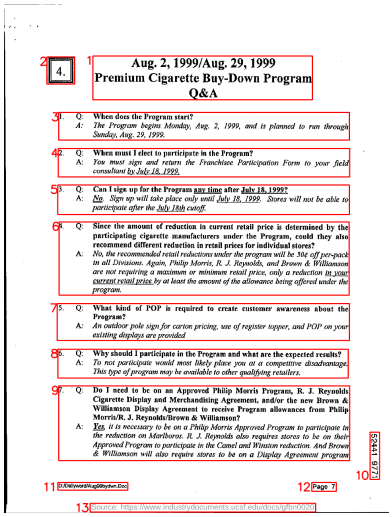
Question:
What is written at the top of the page in bold capital letters?
Boxes:
1, 3, 4
Ground Truth:
CAMEL

Doc-CoB:
[key boxes]
1, 3
[Answer]
CAMEL



Question:
What is inside the [] in the second title?
Key Boxes:
1, 2, 7
Ground Truth:
20700

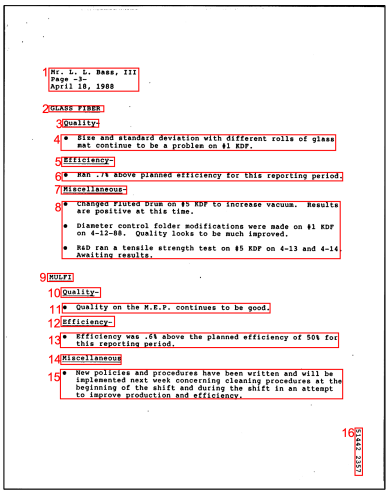
Doc-CoB:
[key boxes]
2, 7, 8
[Answer]
20700



Question:
Which Approved Program should you be on to participate in the reduction on Marlboros?
Key Boxes:
7, 8, 9
Ground Truth:
a Philip Morris Approved Program

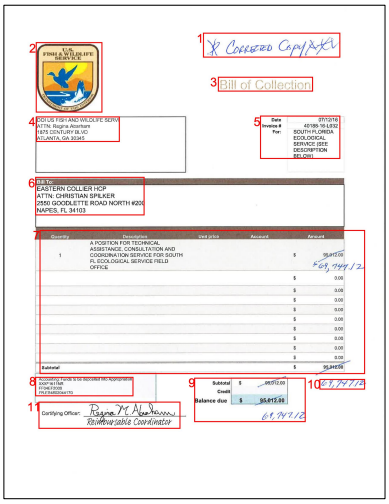
Doc-CoB:
[key boxes]
8, 9
[Answer]
philip morris approved program

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



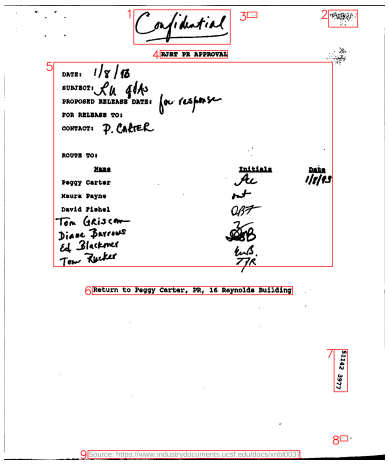
Question:
Which product, glass fiber of Mulfi is reported to have a continuing problem in 'Quality'?
Boxes:
2, 3, 4, 9, 11
Ground Truth:
Glass fiber

Doc-CoB:
[key boxes]
2, 4, 9, 11
[Answer]
glass fiber



Question:
For whom this invoice is made?
Key Boxes:
5, 6, 7
Ground Truth:
South Florida Ecological

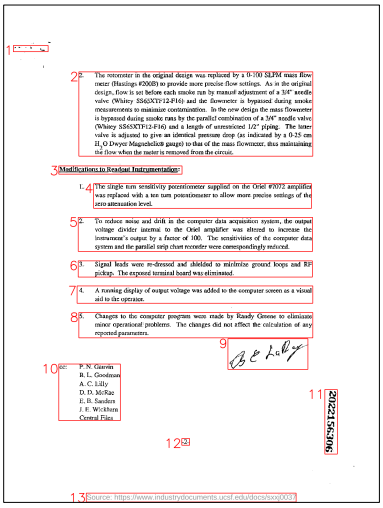
Doc-CoB:
[key boxes]
5, 6
[Answer]
South Florida Ecological



Question:
what is the contact person name mentioned in letter?
Key Boxes:
5, 6
Ground Truth:
p. carter

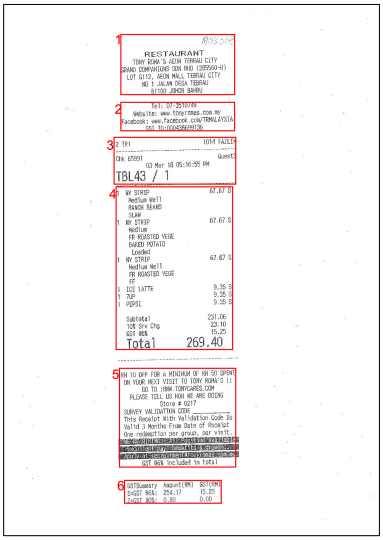
Doc-CoB:
[key boxes]
5, 6
[Answer]
p. carter

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



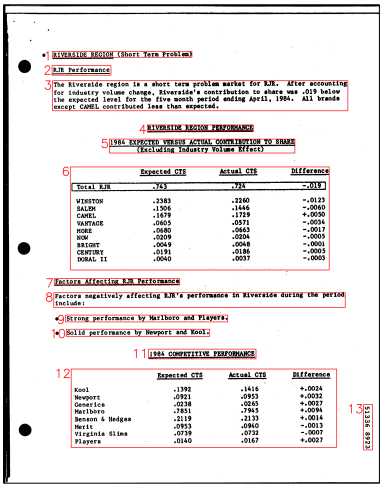
Question:
How many points are there in modifications to readout instrumentation?
Boxes:
3, 4, 5, 6, 7, 8
Ground Truth:
5

Doc-CoB:
[key boxes]
3, 4, 5, 6, 7, 8
[Answer]
5



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



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.