

DoCoG: Mask-based Multi-Type Grounded Chain-of-Thought for Document QA

Sai Madhusudan Gunda¹^{*}, Jyothi Swaroopa Jinka¹^{*}, Hrithik Sagar Rachakonda¹^{*}, Aryan Jain¹[†], Venkata Kesav Venna²[†], Anirudh Srinivasan²^{*}, and Ravi Kiran Sarvadevabhatla^{1,2}^{*}

¹ IIIT Hyderabad

{sai.gunda,jinka.swaroopa,hrithik.rachakonda,aryan.j}@research.iiit.ac.in
ravi.kiran@iiit.ac.in

² BharatGen

{venkat.kesav,anirudh.srinivasan}@bharatgen.com

Abstract. Existing methods for Visual Document Question Answering (DocVQA) often predict answers without explicit intermediate analysis or produce ungrounded intermediate steps. The few approaches which employ grounding produce rectangular bounding boxes which fail to accurately represent complex structures such as pie-chart wedges, flowchart nodes, and curved text. We present DoCoG, a mask-based multi-type grounded step-wise Chain-of-Thought architecture for Document QA. The model supports multi-type grounding, i.e., it localizes both textual content and fine-grained graphical elements such as legends, pie slices, and flowchart nodes. It generates step-wise intermediate outputs and final answers grounded through precise segmentation masks, enabling consistent evidence alignment. DoCoG integrates a Vision-Language Model with a promptable segmenter via a novel Grounding Interaction Module (GIM). During decoding, special [GND] tokens trigger mask prediction, linking each step and the final answer to supporting evidence. GIM refines [GND] tokens during decoding to improve spatial consistency and alignment. We introduce DoCoG-QA, a large-scale diverse dataset comprising 325K documents and 1.5M step-wise grounded QA pairs. We further release DoCoG-PQA with 20K preference pairs to improve alignment using Direct Preference Optimization (DPO), and DoCoG-Bench, a human-verified benchmark. Experiments show that DoCoG achieves state-of-the-art performance across multiple benchmarks, improving answer accuracy and grounding precision.

Keywords: Document Question Answering · Chain-of-Thought (CoT) · Grounding

^{*}Equal contribution

[†]Equal contribution

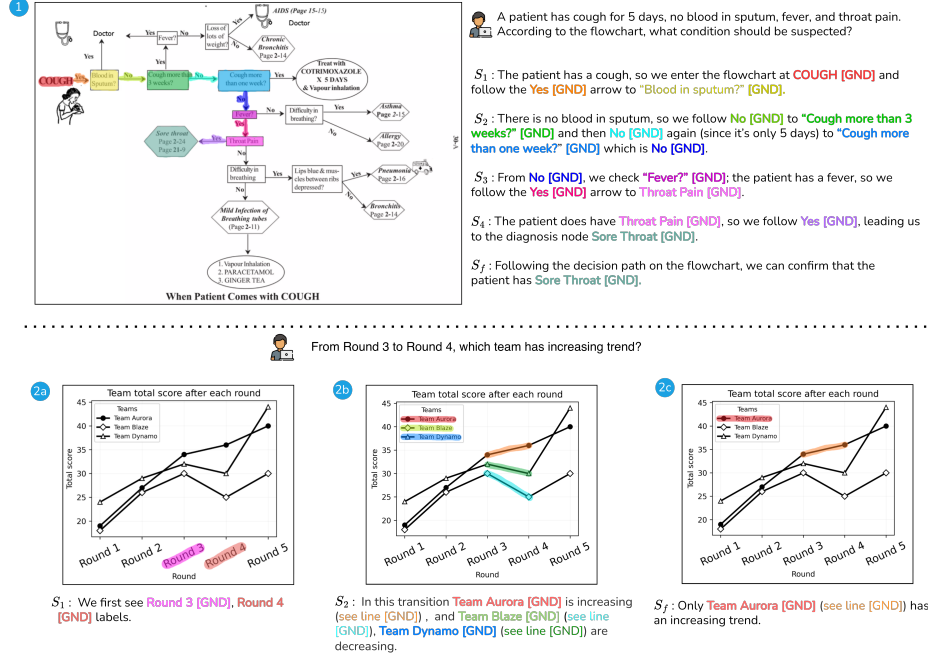


Fig. 1: Mask-based Multi-type grounded CoT. Each example shows a QA pair where the model interleaves [GND] tokens within the CoT (2a, 2b) and final answer (2c). Each [GND] predicts a segmentation mask that localizes the evidence used (1, 2). Evidence spans both text and graphical elements (e.g., flowchart connectors, nodes in 1 chart lines in 2).

1 Introduction

Visual Document Question Answering (DocVQA) requires models to answer queries over complex layouts containing both textual and graphical elements across heterogeneous document types such as reports, tables, charts, and plots. In this context, recent Vision-Language Models (VLMs) [8, 28, 44, 48] demonstrate strong answer generation and Chain-of-Thought (CoT) capabilities.

Grounding-based approaches for DocVQA remain comparatively few [6, 52, 64]. These methods either operate on a single document type [52] or rely on rectangular bounding boxes and often focus solely on grounding textual regions [6, 64]. However, box-based localization is coarse and inadequate for irregular structures such as curved text and graphical elements such as pie-chart wedges and line plots.

Many document questions require multi-step analysis, including performing arithmetic operations, validating claims and interpreting document content. Such analysis often depends on intermediate evidence extracted from specific regions in the document. While existing CoT systems enable step-wise analysis, the intermediate steps are typically not grounded [35] which leads to incorrect

answers. Overall, CoT with step-wise grounding for diverse document structures remains largely unexplored.

To address these limitations, we introduce *DoCoG*, an architecture that integrates a VLM and a promptable mask segmenter via a novel Grounding Interaction Module (GIM) (see Fig. 2). The model jointly generates step-wise CoT answers with segmentation masks. These masks precisely ground the evidence, moving beyond coarse box-based localization. Special [GND] tokens corresponding to each evidence region are interleaved within the intermediate steps and final answer. The GIM refines these [GND] token embeddings by enabling interaction among them. This allows the tokens to attend to each other and improve the consistency of these embeddings (Sec. 3.1).

Documents are inherently heterogeneous, containing textual spans, tables, charts, and diagrammatic components within a single layout (Fig. 2). Correspondingly, document QA questions may require grounding in purely textual evidence (e.g., extracting a field value), purely graphical evidence (e.g., identifying a chart component), or a combination of both (e.g., matching a legend entry to a graphical region) (Fig. 3 ②). Effective grounding must therefore support multiple evidence types. We refer to this as multi-type grounding, where evidence may originate from textual or graphical elements, or jointly from both. DoCoG supports grounding of *multi-type evidence* by predicting fine-grained segmentation masks across diverse document structures (Fig. 1).

Existing document grounding QA datasets [6, 60, 64] lack step-wise grounded CoT (see Tab. 1), are limited to specific domains, and rely on coarse rectangular bounding boxes. To support large-scale training of step-wise grounded CoT, we develop a data engine that generates pixel-level masks for grounding in intermediate steps and final answer. Using this engine, we introduce *DoCoG-QA*, a large-scale dataset comprising 325K documents and 1.5M step-wise grounded CoT QA pairs spanning diverse and comprehensive document domains with grounding over multi-type evidence (Sec. 4).

We curate *DoCoG-PQA*, a preference dataset that ranks candidate CoT outputs based on answer correctness and step-wise evidence grounding, enabling preference optimization beyond fixed-style supervision (Sec. 4.2). To the best of our knowledge, this is the first application of Direct Preference Optimization (DPO) in the Grounded Document VQA setting. We also curate *DoCoG-Bench*, a benchmark dataset with human-verified annotations to assess step-wise CoT quality, answer accuracy, and grounding precision (Sec. 4.3). We conduct extensive evaluations across multiple Document VQA benchmarks spanning diverse document domains (Sec. 5). DoCoG achieves significant improvement in answer accuracy and in grounding precision, establishing a new state-of-the-art in grounded document VQA (Sec. 6). More details at docog-eccv.github.io

2 Related Work

Grounding has become an important mechanism for improving answer transparency in multimodal models [3, 31, 33, 46, 55, 56]. These models demonstrate

Table 1: Comparison of DocVQA datasets across supported capabilities and document types (C: Charts, F: Flowcharts, T: Tables, D: Documents, G: Graphics, All: All type of documents, **txt**: textual grounding, **g**: graphical grounding, **t+g**: multi type grounding.)

Capabilities	ChartQA [20]	CharXiv [49]	ChartQA-Pro [21]	ChartBench [53]	ChartPoint [52]	FlowVQA [38]	AP2D [15]	MISS-QA [60]	TabFact [50]	WTQ [30]	MMTAB [62]	DocVQA [24]	DUDE [45]	VisualMRC [40]	MMDocBench [66]	Boundingdocs [6]	DOGR [64]	VisualCoT [35]	DoCoG-QA (Ours)
CoT	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
Grounding	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
Grounded CoT	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Mask	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Multi-type Grounding	✗	✗	✗	✗	t+g	✗	✗	✗	✗	✗	✗	✗	✗	txt	txt	txt	txt	txt	t+g
Document Type	C	C	C	C	C	F	F	F	T	T	T	D	D	D	D	D	All	All	All

strong grounding capabilities but are primarily designed for natural images and do not generalize grounding over diverse documents, where interpretation involves multi-type evidence regions.

Recent models such as Qwen3-VL [44], InternVL 3.5 [48] and DOGR [64] support grounded document QA. However, they often rely on coarse bounding-boxes to localize evidence. Although effective for simple layouts, bounding-box based grounding may not precisely capture fine-grained structures like curved text regions, irregular layouts and graphical structures such as pie chart wedges, table cells, flow diagram nodes etc. Moreover, these models primarily support grounding only over textual elements.

Existing CoT-based document QA methods [35] typically provide grounding only for the final answer, leaving intermediate steps unverified. This makes it difficult to assess whether a final answer is supported by an appropriate evidence region. While ChartPoint [52] attempts step-wise grounding, it is restricted to bounding-box supervision within the charts domain. Step-wise grounded CoT models across diverse document domains remain underexplored. Moreover, document grounding QA datasets [6, 64] typically rely on bounding-box supervision and are limited to textual regions. Publicly available CoT-based document QA datasets [35] emphasize answer generation with grounding only for the final answer (see Tab. 1). DLaVA [26] performs two-stage bounding-box grounding over OCR-cropped regions but grounds only the final answer. DocVXQA [39] learns a question-guided relevance mask via post-processing for answer localization. In contrast, DoCoG directly grounds both intermediate reasoning steps and the final answer with pixel-level masks on the original document, without OCR pre-processing or post-hoc localization. Publicly available CoT-based document QA datasets [35] emphasize answer generation with grounding only for the final answer (see Tab. 1).

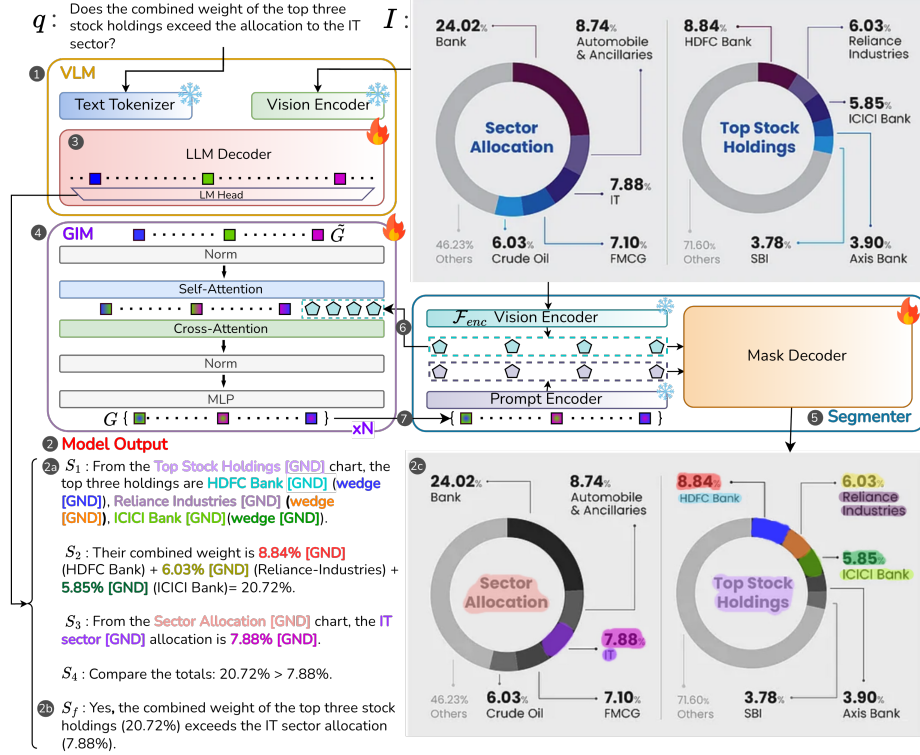


Fig. 2: DoCoG architecture: Given a document image I and question q , the VLM ① encodes visual features and, through its LM head, autoregressively generates step-wise outputs and the final answer interleaved with [GND] tokens ②. The hidden states of the last layer of the [GND] tokens (■ ■ ■) are passed to the Grounding Interaction Module (GIM) ④, which refines them using visual features (■ ■) ⑥ from segmenter ⑤ to produce grounding-aware prompt embeddings. These embeddings are then processed by the segmenter’s prompt encoder ⑦ and mask decoder of the segmentation module to generate evidence masks (2c), grounding each step in the document.

3 Method

Given a document image I and a question q , the goal is to generate step-wise CoT outputs along with the final answer, while grounding the supporting evidence with precise pixel-level masks (see Fig. 2 ②). DoCoG outputs an ordered sequence of steps $S = (S_1, S_2, \dots, S_n, S_f)$, where $S_1, \dots, S_n$ denote intermediate CoT steps ②a and S_f denotes the final answer ②b. [GND] tokens (a special token added to the vocabulary) are interleaved within intermediate steps, with each token predicting the corresponding evidence mask. Each mask corresponds to a multi-type evidence region (textual and graphical elements). Each step can contain multiple [GND] tokens. Across the steps, the model produces a total set

of K grounding masks $\hat{M} = \{\hat{M}_1, \hat{M}_2, \dots, \hat{M}_K\}$ 2c. If there are no supporting evidence regions, then $\hat{M} = \emptyset$ and $K = 0$.

3.1 DoCoG Architecture

DoCoG combines a vision-language backbone, a promptable segmenter with *Grounding Interaction Module* (GIM). The backbone takes (I, q) as input (see 1 in Fig. 2) and autoregressively generates the step sequence S , interleaving text tokens with [GND] tokens whenever grounding is required 2. Whenever the decoder emits a [GND] token, we extract its final-layer hidden state $\tilde{G}_k$ as its grounding representation. Let $\tilde{G} = \{\tilde{G}_k\}_{k=1}^K$ denote final-layer hidden states corresponding to all emitted [GND] tokens 3.

Since evidence regions across the generated steps are often not independent, multiple [GND] tokens may be related, complementary, or mutually constraining. To ensure that [GND] tokens remain mutually consistent and are aligned with the correct image regions, we refine the [GND] embeddings using the Grounding Interaction Module (GIM) 4. GIM is a transformer encoder that enables interaction between [GND] embeddings and visual features from the segmenter’s vision encoder. Specifically, GIM alternates self-attention among the [GND] embeddings and cross-attention with image features extracted by the segmenter encoder: $x = \mathcal{F}_{enc}(I)$ 6, producing refined grounding representations: $G = GIM(\tilde{G}, x)$ 7 which serve as input to segmenter’s prompt encoder. The segmenter 5 predicts masks $\hat{M}$ from image features (x) and refined embeddings (G).

3.2 DoCoG Training

We first train DoCoG with supervised fine-tuning to learn grounded step-wise CoT answers, followed by Direct Preference Optimization (DPO) [32] training to further improve performance.

Supervised Fine-Tuning (stage 1): We jointly learn step-wise text generation and pixel-level evidence grounding. The training objective is a weighted combination of language modeling and segmentation losses: $\mathcal{L}_{\text{sft}} = \lambda_{\text{lm}}\mathcal{L}_{\text{lm}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}$. Here $\mathcal{L}_{\text{lm}}$ denotes the cross-entropy loss over the generated output. The segmentation loss $\mathcal{L}_{\text{seg}}$ supervises each predicted evidence mask ($\hat{M}_k$) against its ground-truth mask (M_k) using a combination of Dice [25] and binary cross-entropy loss [34]: $\mathcal{L}_{\text{seg}} = \sum_{k=1}^K \left(\lambda_{\text{dice}}\ell_{\text{dice}}(\hat{M}_k, M_k) + \lambda_{\text{bce}}\ell_{\text{bce}}(\hat{M}_k, M_k) \right)$

Direct Preference Optimisation (stage 2): To improve robustness and CoT quality, we further train DoCoG using DPO [32] on a preference dataset constructed by ranking sampled grounded CoT outputs based on answer correctness and grounding quality [5] (more details in Sec. 4.2).

DPO optimizes preferences over generated CoT outputs ranked by a reward function. While the reward can reliably rank textual CoT outputs, defining preference pairs for pixel-level masks is ill-posed: multiple masks can validly explain the same evidence region, making pairwise preference ambiguous. Therefore, we

apply DPO only to the textual outputs by updating only VLM parameters, keeping the GIM and segmenter unchanged. However, updating the VLM shifts the hidden state space, causing the [GND] token representations to drift from those seen during SFT. Since these representations serve as the grounding representation to GIM and segmenter, this drift degrades mask quality. Therefore, we introduce *anchor loss* that constrains the [GND] token hidden states to remain close to those produced by the SFT model. We define anchor loss as:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{K} \sum_{k=1}^K \|G_k^\theta - G_k^{\text{ref}}\|_2^2 \quad (1)$$

where $G_k^{\text{ref}} \in G^{\text{ref}}$ and $G_k^\theta \in G^\theta$ denote the k -th GIM refined prompt embeddings from the frozen SFT model and the DPO optimised model, respectively. Together, we optimise it as:

$$\mathcal{L}_{\text{opt}} = \mathcal{L}_{\text{DPO}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}} \quad (2)$$

where $\mathcal{L}_{\text{DPO}}$ denotes traditional DPO [32] loss.

4 Dataset

Prior grounded DocVQA datasets are typically restricted to a single document type, focus on final answer prediction, and rely on coarse bounding boxes for evidence localization. They also lack support for grounding multiple evidence types (textual and graphical) within the same document and do not provide supervision for grounded step-wise CoT analysis (Tab. 1). Therefore, we introduce DoCoG-QA, a diverse large-scale dataset designed for mask based grounded document QA with step-wise CoT supervision. The dataset comprises 325K documents and 1.5M QA pairs, where each QA pair includes step-wise chain-of-thought analysis, the final answer, and grounding annotations linked to the supporting document elements.

The dataset covers a diverse set of document types, including charts, tables, flowcharts, reports, scientific documents, forms, webpages, infographics, and timelines. Specifically, the dataset contains 75K charts, 75K tables, and 75K flowcharts while the remaining 100K are diverse documents. This enables coverage of questions across diverse QA taxonomies, including extraction, comparison, aggregation, validation, interpretation, multi-step analysis (for examples, see Fig. 3).

Each document image is paired with structured metadata containing high-precision element-level bounding boxes, from which segmentation masks and associated attributes are derived. These annotations capture both textual elements (e.g., words, phrases, paragraphs) and graphical elements (e.g., chart components, flowchart nodes, table cells), enabling accurate grounding across diverse document structures. Each mask is assigned a unique identifier in the metadata. We build a data engine that takes the document image and metadata

as input and generates QA pairs with step-wise CoT while referencing mask identifiers as evidence. The identifiers are mapped back to the corresponding masks, ensuring precise mask-based grounding and preventing hallucinated coordinates.

4.1 Data Verification

We apply a multi-stage filtering pipeline to ensure structural consistency, answer correctness, and grounding reliability. First, we enforce structural constraints by requiring each QA pair to contain a valid final answer and at least one grounding reference; pairs that violate this condition are removed. Next, we verify answer correctness using `gpt-oss-120B` [17, 27], a verifier from a different model family than DoCoG’s Qwen3-VL backbone which evaluates whether the generated answer is supported by the document metadata and question;; QA pairs with scores below a threshold are discarded. We further assess the quality of step-wise CoT using ROSCOE [7], which evaluates step correctness, coherence, and consistency with the final answer, removing pairs below a threshold. Finally, we validate grounding correctness and mask quality. For textual elements, we apply OCR to the referenced mask region and compare the extracted text with the answer span. For graphical elements, grounding consistency is verified using graphical–text correspondence in the metadata. Mask quality for graphical elements is assessed using masks generated by a SOTA segmentation model [2] and manually verified over a stratified subset of 33,500 masks spanning diverse document domains and visual categories (e.g., pie wedges, flowchart nodes). The masks closely align with intended document elements with minimal spillover and are therefore used as grounding regions. This pipeline ensures that the final dataset contains structurally consistent, semantically correct, and reliably grounded QA pairs.

4.2 DPO Dataset and Reward Design

We build *DoCoG-PQA*, a 20K sampled DPO preference dataset, by sampling N grounded CoT outputs per (I, q) using the SFT reference model and ranking them with a reward that combines final-answer correctness, CoT quality, grounding quality, and format validity. *Format Reward* (FR) verifies structural validity by parsing (i) intermediate steps, (ii) the final answer, and (iii) evidence references produced via [GND] tokens; outputs where parsing fails receive FR=0, otherwise FR=1. *Final Answer Quality* (AQ) assigns a positive score when the predicted answer matches the ground-truth [29, 64] after text normalization and numeric tolerance. *CoT Quality* (CQ) measures reasoning quality and step-level faithfulness by first computing a normalized ROSCOE score [7]; if this score is below threshold τ_{ros} , we set CQ = 0. Otherwise, we measure alignment with ground-truth reasoning using maximum bipartite matching based on semantic similarity and numeric consistency [47], counting matched steps as true positives, unmatched predicted steps as false positives, and unmatched ground-truth steps as false negatives, from which an F1 score is computed as CQ. *Grounding Quality* (GQ) evaluates spatial and contextual grounding by matching each predicted

Table 2: DoCoG on chart understanding benchmarks. ChartQA [20], ChartQA-Pro [21], ChartBench [53], CharXiv [49], (all scores on official metrics) and the Charts split of DoCoG-Bench. Models include domain-specialized and representative general models. **Bold**: best among open-source models. Underline: best among commercial models. “-”: unsupported. “n/a”: indicates lack of results due to model unavailability.

Model	Params	ChartQA	ChartQA-Pro	ChartBench	CharXiv	DoCoG-Bench	
						AQ	GQ _{fa}
GPT-5 [28]	-	<u>89.0</u>	42.7	<u>69.4</u>	76.3	82.4	3.8
Gemini 3 pro [9]	-	88.5	<u>51.1</u>	61.2	<u>80.3</u>	<u>83.2</u>	<u>11.2</u>
TinyChart [57]	3B	78.9	11.7	22.5	12.2	54.8	-
Chart-R1 [4]	8B	91.0	44.0	34.1	46.2	74.2	-
ChartLlama [10]	13B	69.7	03.4	22.3	16.7	47.6	-
ChartGemma [22]	3B	80.2	09.8	38.5	16.9	59.3	-
ChartMoE [54]	8B	86.2	41.6	29.7	37.6	67.4	-
ChartPoint [52]	7B	87.7	n/a	67.9	n/a	n/a	n/a
InternVL 3.5 [48]	8B	86.7	34.7	51.4	68.8	73.4	4.1
Qwen3-VL [44]	8B	89.6	32.9	65.3	64.7	75.2	3.9
DoCoG(Ours)	9B	92.7	52.1	79.4	73.1	83.5	80.3

mask to the nearest ground-truth mask using IoU; a match is valid only if (i) IoU ≥ 0.5 , (ii) the element type matches (e.g., text span, table cell, chart wedge), and (iii) it appears in the correct step context. We compute two F1 variants: GQ_{all} over masks from all intermediate steps and the final answer, and GQ_{fa} using only masks required for the final answer, with GQ defined as their average. **Reward Function:** Samples with FR= 0 are rejected; for the remaining outputs we compute AQ, CQ, and GQ, combine them into a total reward, and construct DPO pairs by selecting the highest-scoring output as the *chosen* sample and the lowest-scoring output as the *rejected* sample.

4.3 DoCoG-Bench

Existing document QA benchmarks provide either bounding-box based grounding or step-wise CoT answers, but do not jointly evaluate grounded step-wise analysis with fine-grained segmentation masks. Therefore, we introduce DoCoG-Bench, a segmentation-based benchmark for evaluating step-wise CoT answers aligned with precise mask-level grounding. DoCoG-Bench consists of 3K document images and 10K QA pairs across multiple document types, including charts, tables, reports, webpages, infographics, flowcharts, and documents with curved or complex text layouts. We curate documents from multiple sources, including existing document QA test sets, to ensure structural diversity. The QA pairs are generated using our data engine, filtered through the data verification pipeline in Section 4.1, and human-verified through a two-stage process. Primary annotators trained on mask-annotation protocol, element-type definitions, and QA correctness criteria checked answer correctness, grounding relevance, and mask boundary accuracy. The benchmark includes complex multi-step questions across varied document structures.

Table 3: DoCoG on flowchart and diagram understanding benchmarks. FlowVQA [38], AI2D [15], MISS-QA [60], (all scores on official metrics) and the flowcharts split of DoCoG-Bench.

Model	Params	FlowVQA	AI2D	MISS-QA	DoCoG-Bench	
					AQ	GQ _{fa}
GPT-5 [28]	–	<u>91.9</u>	89.5	<u>71.7</u>	78.1	4.0
Gemini 3 pro [9]	–	88.1	<u>91.2</u>	67.3	<u>79.4</u>	<u>10.1</u>
CogAgent [11]	18B	38.8	58.2	56.7	53.8	–
Kimi VL [43]	16B	40.9	84.9	63.8	57.2	–
IXC 2.5 [58]	7B	43.3	82.3	40.6	59.4	–
InternVL 3.5 [48]	8B	78.4	84.0	48.5	67.2	1.5
Qwen3-VL [44]	8B	81.0	85.7	46.2	69.4	11.0
DoCoG (Ours)	9B	85.5	88.3	78.4	82.2	79.2

Table 4: DoCoG on table understanding benchmarks. TabFact [50], WTQ [30], TabMWP [18], (all scores on official metrics) and the Tables split of DoCoG-Bench.

Model	Params	TabFact	WTQ	TabMWP	DoCoG-Bench	
					AQ	GQ _{fa}
GPT-5 [28]	–	87.7	<u>78.8</u>	<u>93.7</u>	84.3	6.4
Gemini 3 pro [9]	–	<u>88.5</u>	76.2	91.5	<u>86.0</u>	<u>12.6</u>
Table-LLaVA [62]	7B	59.8	18.4	57.8	49.6	–
IXC 2.5 [58]	7B	85.2	53.6	26.6	69.8	–
DocOwl 2 [12]	8B	78.2	36.5	15.2	38.6	–
SynTab-LLaVA [63]	7B	70.8	39.6	88.3	62.4	–
TabPedia [59]	7B	28.8	20.4	12.3	33.7	–
InternVL 3.5 [48]	8B	78.1	55.6	55.7	78.0	3.2
Qwen3-VL [44]	8B	81.8	57.4	65.6	79.5	0.9
DoCoG(Ours)	9B	89.4	64.4	80.7	87.8	82.7

5 Experiments

Implementation Details. We instantiate DoCoG using Qwen3-VL-8B [44] as the VLM backbone, the GIM module a 3-layer transformer encoder and SAM [16] as the segmenter. During Supervised Fine-tuning (SFT), DoCoG is trained jointly end-to-end. Training uses AdamW with a learning rate of 10^{-5} , 3% linear warmup followed by cosine decay. We empirically set weights: during SFT, we set $\lambda_{lm} = 1.0$, $\lambda_{seg} = 2.0$, $\lambda_{dice} = 1.0$ and $\lambda_{bce} = 1.0$. These weights are kept fixed across all experiments. Next, we apply DPO while freezing segmenter and GIM. We set $\lambda_{anchor} = 1.0$. We empirically set a reduced learning rate of 10^{-6} during preference optimization. All experiments are trained on 32×H100 GPUs using DeepSpeed ZeRO-3 with an effective batch size of 256. On a single H100 GPU, DoCoG runs at 580 ms latency with 24.08 GB memory and 1.72 samples/s.

Training Data: DoCoG is first trained on DoCoG-QA with additional subsets of documents, chart, table, and flowchart datasets [20, 23, 24, 35, 45, 51, 64] and then on DoCoG-PQA. This combined training enables grounded step-wise CoT without compromising document QA ability. All baselines are fine-tuned on the same training protocol to maintain a consistent and a fair comparison.

Table 5: DoCoG on document understanding benchmarks. Results on DOGR [64], MMDocBench (MMDB) [66], VisualCoT-Bench (VCoT) [35], and BoundingDocs (BD) [6], and Documents split of DoCoG-Bench. BD and DOGR, grounding is measured by $F1_g$ (bbox IoU>0.5) and AQ by G_{eval} [17]. MMDB (official metrics). VCoT (official metric) includes general Document VQA [14, 23, 24, 36, 37, 45]. “n/a” indicates lack of results due to code-level issues in source repo [65]

Model	Params	DOGR		MMDB			VCoT	BD		DoCoG-Bench	
		$F1_g$	AQ	EM	$F1_{txt}$	IoU		$F1_g$	AQ	AQ	GQ_{fa}
GPT-5 [28]	–	3.6	87.0	74.3	76.6	9.3	89.7	5.4	89.3	86.7	5.0
Gemini 3 pro [9]	–	65.3	91.8	83.8	85.4	55.4	87.4	73.5	92.2	88.1	11.7
Kosmos 2.5 [19]	1.3B	0.0	36.1	12.6	13.8	0.0	61.2	0.70	33.7	34.6	0.6
DocOwl 2 [12]	9B	–	43.3	15.9	17.3	–	63.5	–	69.0	43.9	–
CogAgent [11]	18B	2.6	69.6	29.9	27.2	0.0	67.4	3.4	74.7	71.4	0.0
Kimi-VL [43]	16B	7.2	68.0	59.9	61.6	3.5	77.8	8.1	68.2	72.6	0.5
DOGR [64]	8B	66.4	78.3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
InternVL 3.5 [48]	8B	12.5	80.9	65.7	67.4	15.5	83.2	41.5	79.8	77.1	7.2
Qwen3-VL [44]	8B	27.6	82.6	69.8	71.3	28.7	86.7	44.5	80.3	78.6	6.3
DoCoG (Ours)	9B	75.4	86.2	73.1	76.3	70.5	88.9	85.4	89.3	89.7	85.4

Baselines: We compare DoCoG against a comprehensive suite of baselines spanning chart and table understanding, flowchart analysis, document VQA, and large-scale multimodal language models (MLLMs). For chart-centric reasoning, we compare against specialized models [4, 9, 10, 22, 28, 44, 48, 52, 54, 57]. In the documents domain, we compare against representative document VQA systems [9, 11, 12, 19, 28, 43, 44, 48, 64]. For table understanding, we benchmark against approaches [12, 58, 59, 62, 63] designed for tabular question answering. To ensure comparison with strong general-purpose models, we further include a diverse set of MLLMs from commercial models [1, 8, 9, 28], open source models [12, 19, 35, 42, 44, 48, 55] and finetuned models.

DoCoG-Bench Metrics: Evaluating step-wise grounded CoT answers requires assessing the final answer, the correctness and consistency of intermediate steps, and the faithfulness of grounded evidence. We therefore use the same components defined in Sec. 4.2 for evaluation, computed deterministically against ground-truth. Specifically, we report Final Answer Quality (AQ), CoT Quality (CQ), and Grounding Quality (GQ_{fa} , GQ_{all}). In addition to these deterministic measures, we include an LLM-based [61] score under the GEval framework [17] to assess the overall quality of the generated answer beyond exact matching.

Other benchmarks: We evaluate DoCoG across diverse document types, benchmarking against chart QA benchmarks [20, 21, 49, 53], document centric benchmarks [6, 35, 64, 66], flowchart benchmarks [15, 41, 60], table QA benchmarks [18, 30, 50]. All benchmark results are computed using their official metric implementations.

Mask Box Alignment: To ensure consistent evaluation across models with different grounding outputs, we standardize predictions to match the annotation format of each benchmark. When a model outputs bounding boxes but the ground-truth is provided as segmentation masks, we convert each predicted box into a binary mask and then compute mask-based IoU. Conversely, when a

Table 6: Unified evaluation across document types on DoCoG-Bench. AQ and GQ_{fa} reported per split (Charts, Flowcharts, Tables and Documents), The All block reports AQ, G_{eval}, CQ, GQ_{all}, and GQ_{fa}. Results for Commercial models , Open-source models , Fine-tuned models variants are shown. **Bold**: best among open-source and fine-tuned models. Underline: best among commercial models. “-”: model does not support this output type.

Model	Params	Charts		Flowcharts		Docs		Tables		All				
		AQ	GQ _{fa}	AQ	GQ _{fa}	AQ	GQ _{fa}	AQ	GQ _{fa}	AQ	G _{eval}	CQ	GQ _{all}	GQ _{fa}
GPT-5 [28]	–	82.4	3.8	78.1	4.0	86.7	5.0	84.3	6.4	82.9	<u>84.8</u>	<u>76.8</u>	2.4	4.8
Sonnet 4.5 [1]	–	79.1	5.7	75.3	<u>13.3</u>	83.8	6.8	81.6	2.6	79.9	80.5	71.3	3.3	7.1
Gemini 3 pro [9]	–	<u>83.2</u>	11.2	<u>79.4</u>	10.1	<u>88.1</u>	11.7	<u>86.0</u>	12.6	<u>84.2</u>	84.1	75.2	15.4	11.4
Gemini 3 flash [8]	–	79.6	<u>12.7</u>	75.8	12.5	84.5	<u>35.3</u>	82.4	<u>33.9</u>	80.6	82.8	73.4	<u>32.8</u>	<u>23.6</u>
Kosmos2.5VL [19]	1.3B	22.8	2.3	15.3	0.0	34.6	0.6	28.4	0.2	25.3	42.3	0.0	0.0	0.8
DocOwl2 [12]	9B	32.4	–	23.7	–	43.9	–	38.6	–	34.7	35.6	–	–	–
Gemma3 [42]	12B	57.1	1.8	49.3	1.0	66.4	3.5	61.8	2.9	58.7	62.4	28.4	0.0	2.3
InternVL 3.5 [48]	8B	73.4	4.1	67.2	1.5	77.1	7.2	78.0	3.2	73.9	78.2	33.2	2.9	4.0
Qwen3-VL [44]	8B	75.2	3.9	69.4	11.0	78.6	6.3	79.5	0.9	75.7	77.6	35.9	3.5	5.5
LISA++ [55]	7B	9.4	1.2	7.2	0.4	16.3	2.8	14.3	3.2	11.8	21.4	0.0	0.5	1.9
VisualCoT [35]	7B	52.3	5.5	44.6	3.9	59.7	3.6	55.4	1.0	53.0	52.1	51.2	2.4	3.5
InternVL 3.5 FT	8B	74.6	65.0	68.8	67.9	82.0	64.3	79.1	61.2	76.1	80.4	72.2	61.4	64.6
Qwen3-VL FT	8B	75.8	63.9	70.2	66.9	83.4	61.2	80.6	55.6	77.5	79.8	74.8	57.8	61.9
LISA++ FT	7B	28.4	21.3	26.8	19.6	41.2	22.4	38.4	21.9	33.7	39.8	17.8	19.3	21.3
DoCoG (Ours)	9B	83.5	80.3	82.2	79.2	89.7	85.4	87.8	82.7	85.8	85.2	82.0	78.2	81.9

model produces segmentation masks but the benchmark annotations are given as bounding boxes, we derive the minimal axis-aligned rectangle tightly enclosing the predicted foreground region and perform IoU evaluation in the box domain.

6 Results

Across all evaluated benchmarks, DoCoG consistently outperforms both commercial and open-source baselines. While commercial models such as GPT-5 and Gemini 3 Pro often achieve comparable answer accuracy (AQ), their grounding quality (GQ) remains substantially lower (Tab. 6). In contrast, DoCoG maintains strong performance on both AQ and GQ, indicating that explicit step-wise grounding improves spatial faithfulness and evidence alignment. We attribute this gap to the nature of DoCoG-Bench: many questions require grounding over non-textual evidence (graphic primitives, connectors, arrows, plot marks) and text cases (curved/rotated), where baselines fail.

On **chart** benchmarks (Tab. 2), DoCoG surpasses strong chart-specialized systems. Despite Chart-R1 being specifically designed for chart reasoning QA, our model outperforms it across all benchmarks (AQ: 74.2 → 83.5), highlighting the benefit of step-wise grounding for multi-type elements.

On **flowchart and diagram** reasoning benchmarks (Tab. 3), DoCoG consistently outperforms prior open-source models. It achieves the best results on FlowVQA and MISS-QA, and remains competitive on AI2D. On the flowchart split of DoCoG-Bench, DoCoG improves AQ(69.4 → 82.2) and GQ_{fa} (11.0 → 79.2) compared to Qwen3-VL. These large gains indicate that step-wise grounded

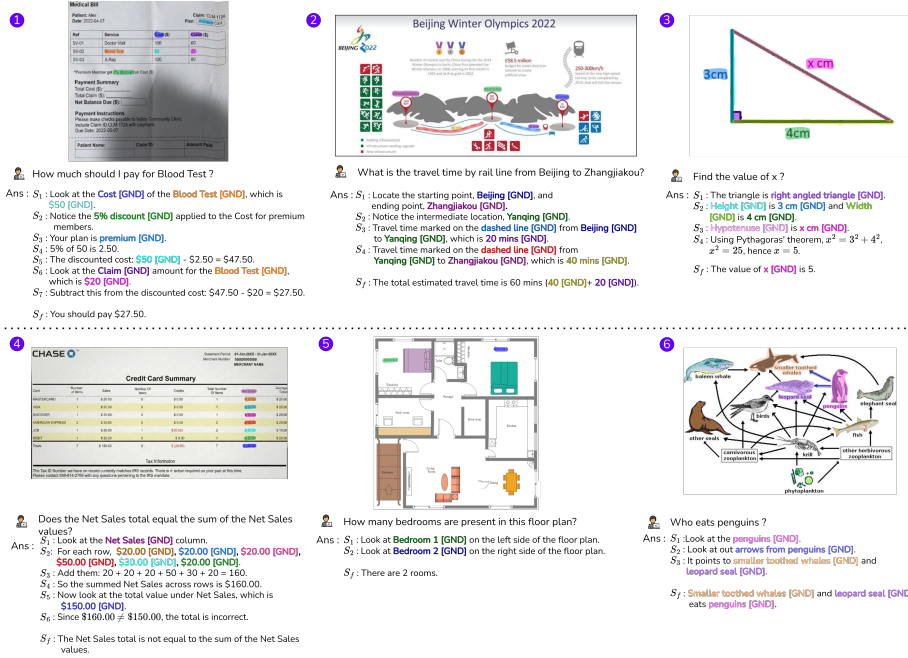


Fig. 3: Qualitative examples (Model outputs). Each example shows DoCoG’s step-wise CoT with mask-based, multi-type grounding over both textual and graphical elements. The outputs demonstrate how DoCoG answers complex questions while grounding evidence at intermediate steps and in the final answer.

reasoning enables the model to follow multi-hop structural dependencies while grounding intermediate steps to the correct diagram elements.

On **table** reasoning benchmarks (Tab. 4), DoCoG achieves the best overall performance across TabFact, WTQ, and consistently leads on DoCoG-Bench metrics. Notably, on TabMWP, a math-intensive dataset, DoCoG remains competitive. This suggests that step-wise grounded CoT transfers effectively to numerically intensive multi-step problems without task-specific specialization.

On **document** understanding benchmarks (Tab. 5), DoCoG achieves the best overall performance across DOGR, MMDocBench, VCoT, BoundingDocs, and the documents split of DoCoG-Bench compared to strong open-source VLM baselines such as InternVL 3.5 and Qwen3-VL. DoCoG significantly improves on GQ for both open-source and proprietary models. On standard DocVQA benchmarks [14, 23, 24, 45] evaluated through the Visual-CoT (VCoT) document split, DoCoG also maintains strong general document QA performance, showing that step-wise grounded CoT preserves general DocVQA capability while providing explicit evidence localization.

Finally, Tab. 6 reports unified DoCoG-Bench results across document types. On DoCoG-Bench, which measures AQ, CQ, and GQ, DoCoG achieves the strongest aggregated performance. Models lacking step-wise CoT or pixel-level

mask support (e.g., DOGR) have CQ and GQ_{all} undefined on DoCoG-Bench; we report only verifiable comparisons for such baselines. On DoCoG-Bench, which measures AQ, CQ, and GQ, DoCoG achieves the strongest aggregated performance. Zero-shot commercial and open-source models often obtain reasonable AQ but exhibit very low grounding scores, indicating that models can produce correct answers without correctly localizing supporting evidence. Fine-tuning InternVL 3.5 and Qwen3-VL on DoCoG-QA substantially improves their grounding across domains, validating the effectiveness of our dataset and training pipeline. DoCoG further improves over these fine-tuned variants, with its architectural contributions (GIM, mask supervision) across Charts, Flowcharts, Documents, and Tables, demonstrating consistent gains in answer quality, step consistency, and grounding fidelity.

Ablations: We ablate DoCoG along grounding, preference optimization, CoT, GIM design, and training choices. Results are in Tab. 7. *Progressive build-up:* Moving from no grounding to final-answer-only grounding already improves answer quality (DoCoG-Bench AQ: 78.1 $\rightarrow$ 81.5). Extending grounding to CoT steps brings further gains in both answer accuracy (DoCoG-Bench AQ: 81.5 $\rightarrow$ 84.1) and step consistency, confirming that grounding intermediate evidence improves the step-level consistency. Adding DPO yields the best overall performance across all metrics. *CoT build-up:* We additionally train a no-CoT variant that predicts only the final answer. Removing CoT consistently decreases performance across all metrics, showing the need for step-wise CoT reasoning and grounding accuracy. *GIM:* Removing GIM causes a large drop in grounding quality (GQ_{fa}: 81.9 $\rightarrow$ 69.6) with only minor impact on answer accuracy, showing that explicit refinement of [GND] representations is critical for stable evidence localisation. Disabling image cross-attention with [GND] tokens inside GIM also degrades grounding metrics (GQ_{all}: 78.2 $\rightarrow$ 73.5; GQ_{fa}: 81.9 $\rightarrow$ 77.0), indicating that grounding tokens must directly incorporate visual context, not only mutual token consistency, to remain spatially aligned. *Anchor loss:* Removing anchor loss causes a severe drop on grounding (IoU: 70.5 $\rightarrow$ 40.3; GQ_{fa}: 81.9 $\rightarrow$ 41.7), showing that DPO shifts hidden-state space of [GND] tokens in ways that break mask fidelity. *PEFT:* LoRA [13] underperforms across all metrics, particularly on grounding, as limited rank updates cannot update the model parameters required for [GND] embedding.

7 Conclusion

DoCoG is a novel mask-based multi-type grounded Chain-of-Thought approach for Visual Document Question Answering. By interleaving special [GND] tokens during decoding, DoCoG jointly produces step-wise reasoning and pixel-accurate evidence masks, enabling verifiable intermediate steps. Our Grounding Interaction Module (GIM) refines grounding-token representations via contextual interactions with visual features, improving grounding performance. Alongside the model, we release DoCoG-QA for large-scale training of step-wise grounded CoT, DoCoG-PQA for preference optimization, and DoCoG-Bench for unified evalu-

Table 7: Ablations. We compare across representative charts, documents, flowcharts, tables, and DoCoG-Bench (AQ, G_{eval} , CQ, GQ_{all} , GQ_{fa}). **GIM**: Grounding Interaction Module; **img attn**: image cross-attention with [GND] tokens in GIM.

Ablation setting	MMDB			CharXiv	FlowVQA	WTQ	DoCoG-Bench				
	EM	F1 _{txt}	IoU				AQ	G_{eval}	CQ	GQ_{all}	GQ_{fa}
PROGRESSIVE BUILD-UP											
No Grnding	66.2	69.4	—	66.5	68.7	58.6	78.1	77.5	74.6	—	—
+ Only Final Ans Grnding	69.2	72.5	66.9	69.4	71.7	61.2	81.5	80.9	77.9	—	77.8
+ Step and Final Ans Grnding	71.3	74.8	69.1	71.6	74.0	63.1	84.1	83.5	80.4	76.6	80.3
+ DPO (Ours)	73.1	76.3	70.5	73.1	85.5	64.4	85.8	85.2	82.0	78.2	81.9
CoT BUILD-UP											
w/o CoT	65.1	69.3	62.5	66.7	75.4	58.0	73.3	71.4	—	—	69.3
GIM ABLATIONS											
w/o GIM	68.4	71.7	59.9	68.7	71.0	60.5	80.7	80.1	77.1	66.5	69.6
GIM w/o img attn	70.6	74.0	66.3	70.9	73.2	62.5	83.2	82.6	79.5	73.5	77.0
LOSSES											
w/o $\mathcal{L}_{\text{anchor}}$ (Eq. 1)	65.8	68.6	40.3	64.2	67.9	56.6	78.1	75.3	74.6	39.7	41.7
PEFT											
LoRA	55.0	57.0	23.0	51.8	36.2	40.6	68.6	68.2	65.6	62.5	65.5
DoCoG (SFT + DPO)	73.1	76.3	70.5	73.1	85.5	64.4	85.8	85.2	82.0	78.2	81.9

ation of answer accuracy, step consistency, and grounding faithfulness. Extensive experiments demonstrate that DoCoG achieves state-of-the-art performance across charts, tables, flowcharts, and document-centric benchmarks, with particularly strong gains in grounding quality and step-level consistency.

Acknowledgements

Our sincere gratitude goes to Srihari Bandrupalli, Sai Sagar Jinka and late B.V. Khadiravana for their invaluable support. The project was supported by BharatGen.

References

1. Anthropic: Claude sonnet 4.5 system card. <https://www.anthropic.com/claude-sonnet-4-5-system-card> (Oct 2025), accessed 2026-02-24
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm’s referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
4. Chen, L., Zhao, X., Zeng, Z., Huang, J., Zhong, Y., Ma, L.: Chart-r1: Chain-of-thought supervision and reinforcement for advanced chart reasoner. CoRR **abs/2507.15509** (July 2025), <https://doi.org/10.48550/arXiv.2507.15509>
5. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)

6. Giovannini, S., Coppini, F., Gemelli, A., Marinai, S.: Boundingdocs: a unified dataset for document question answering with spatial annotations: S. giovannini et al. International Journal on Document Analysis and Recognition (IJDAR) pp. 1–16 (2025)
7. Golovneva, O., Chen, M.P., Poff, S., Corredor, M., Zettlemoyer, L., Fazel-Zarandi, M., Celikyilmaz, A.: ROSCOE: A suite of metrics for scoring step-by-step reasoning. In: The Eleventh International Conference on Learning Representations (2023), <https://openreview.net/forum?id=xYlJRpzZtsY>
8. Google: Gemini 3 flash. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-flash> (2025), accessed 2026-02-25
9. Google DeepMind: Gemini 3 pro. <https://deepmind.google/models/gemini/pro/> (2025), accessed 2026-02-25
10. Han, Y., Zhang, C., Chen, X., Yang, X., Wang, Z., Yu, G., Fu, B., Zhang, H.: Chartllama: A multimodal llm for chart understanding and generation. arXiv preprint arXiv:2311.16483 (2023)
11. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., et al.: Cogagent: A visual language model for gui agents. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14281–14290 (2024)
12. Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5817–5834 (2025)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
14. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520. IEEE (2019)
15. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
17. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: NLG evaluation using gpt-4 with better human alignment. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2511–2522. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.emnlp-main.153>, <https://aclanthology.org/2023.emnlp-main.153/>
18. Lu, P., Qiu, L., Chang, K.W., Wu, Y.N., Zhu, S.C., Rajpurohit, T., Clark, P., Kalyan, A.: Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In: International Conference on Learning Representations (ICLR) (2023)
19. Lv, T., Huang, Y., Chen, J., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., Dong, L., Luo, W., Wu, S., Wang, G., Zhang, C., Wei, F.: Kosmos-2.5: A multimodal literate model (September 2023), <https://www.microsoft.com/en-us/research/publication/kosmos-2-5-a-multimodal-literate-model/>

20. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
21. Masry, A., Islam, M.S., Ahmed, M., Bajaj, A., Kabir, F., Kartha, A., Laskar, M.T.R., Rahman, M., Rahman, S., Shahmohammadi, M., et al.: Chartqapro: A more diverse and challenging benchmark for chart question answering. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19123–19151 (2025)
22. Masry, A., Thakkar, M., Bajaj, A., Kartha, A., Hoque, E., Joty, S.: ChartGemma: Visual instruction-tuning for chart reasoning in the wild. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Darwish, K., Agarwal, A. (eds.) Proceedings of the 31st International Conference on Computational Linguistics: Industry Track. pp. 625–643. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), <https://aclanthology.org/2025.coling-industry.54/>
23. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
24. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
25. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
26. Mohammadshirazi, A., Neogi, P.P.G., Lim, S.N., Ramnath, R.: Dlava: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness. arXiv preprint arXiv:2412.00151 (2024)
27. OpenAI: gpt-oss-120b & gpt-oss-20b model card (2025), <https://arxiv.org/abs/2508.10925>
28. OpenAI: Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/> (2025), accessed: 2025-10-25
29. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). <https://doi.org/10.3115/1073083.1073135>, <https://aclanthology.org/P02-1040/>
30. Pasupat, P., Liang, P.: Compositional semantic parsing on semi-structured tables. In: Zong, C., Strube, M. (eds.) Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1470–1480. Association for Computational Linguistics, Beijing, China (Jul 2015). <https://doi.org/10.3115/v1/P15-1142>, <https://aclanthology.org/P15-1142/>
31. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
32. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)

33. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024)
34. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *nature* **323**(6088), 533–536 (1986)
35. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems* **37**, 8612–8642 (2024)
36. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European conference on computer vision. pp. 742–758. Springer (2020)
37. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
38. Singh, S., Chaurasia, P., Varun, Y., Pandya, P., Gupta, V., Gupta, V., Roth, D.: FlowVQA: Mapping multimodal logic in visual question answering with flowcharts. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*. pp. 1330–1350. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). <https://doi.org/10.18653/v1/2024.findings-acl.78>, <https://aclanthology.org/2024.findings-acl.78/>
39. Souibgui, M.A., Choi, C., Barsky, A., Jung, K., Valveny, E., Karatzas, D.: Docvxqa: Context-aware visual explanations for document question answering. *arXiv preprint arXiv:2505.07496* (2025)
40. Tanaka, R., Nishida, K., Yoshida, S.: Visualmrc: Machine reading comprehension on document images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13878–13888 (2021)
41. Tannert, S., Feighelstein, M.G., Bogojeska, J., Shtok, J., Arbellet, A., Staar, P.W.J., Schumann, A., Kuhn, J., Karlinsky, L.: FlowchartQA: The first large-scale benchmark for reasoning over flowcharts. In: Aggarwal, P., Alaçam, Ö., Silberer, C., Zarrieß, S., Zesch, T. (eds.) *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*. pp. 34–46. Association for Computational Linguistics, Ingolstadt, Germany (Sep 2023), <https://aclanthology.org/2023.limo-1.5/>
42. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025)
43. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. *arXiv preprint arXiv:2504.07491* (2025)
44. Team, Q.: Qwen3 technical report (2025), <https://arxiv.org/abs/2505.09388>
45. Van Landeghem, J., Tito, R., Borchmann, L., Pietruszka, M., Joziak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Anckaert, B., Valveny, E., et al.: Document understanding dataset and evaluation (dude). In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19528–19540 (2023)
46. Venna, V.K., Gunda, S.M., Jinka, J.S., Rachakonda, H.S., Srinivasan, A., Sarvadevabhatla, R.K.: M3grounder: Mask-based multi-span and multi-granular grounding for document qa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23685–23695 (June 2026)

47. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533 (2022)
48. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
49. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., et al.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems* **37**, 113569–113697 (2024)
50. Wenhui Chen, Hongmin Wang, J.C.Y.Z.H.W.S.L.X.Z., Wang, W.Y.: Tabfact : A large-scale dataset for table-based fact verification. In: *International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia (April 2020)
51. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2087–2098 (2025)
52. Xu, Z., Du, S., Qi, Y., Lu, S., Xu, C., Yuan, C., Guo, J.: Chartpoint: Guiding mllms with grounding reflection for chart reasoning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 426–436 (2025)
53. Xu, Z., Du, S., Qi, Y., Xu, C., Yuan, C., Guo, J.: Chartbench: A benchmark for complex visual reasoning in charts. arXiv preprint arXiv:2312.15915 (2023)
54. Xu, Z., Qu, B., Qi, Y., Du, S., Xu, C., Yuan, C., Guo, J.: Chartmoe: Mixture of diversely aligned expert connector for chart understanding. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) *International Conference on Learning Representations*. vol. 2025, pp. 78550–78572 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/c33cd281f8cd784626a57de340e43fe4-Paper-Conference.pdf
55. Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: Lisa++: An improved baseline for reasoning segmentation with large language model. arXiv preprint arXiv:2312.17240 (2023)
56. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Leizhang, Li, C., et al.: Llava-grounding: Grounded visual chat with large multimodal models. In: *European Conference on Computer Vision*. pp. 19–35. Springer (2024)
57. Zhang, L., Hu, A., Xu, H., Yan, M., Xu, Y., Jin, Q., Zhang, J., Huang, F.: TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. pp. 1882–1898. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). <https://doi.org/10.18653/v1/2024.emnlp-main.112>, <https://aclanthology.org/2024.emnlp-main.112/>
58. Zhang, P., Dong, X., Zang, Y., Cao, Y., Qian, R., Chen, L., Guo, Q., Duan, H., Wang, B., Ouyang, L., et al.: Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. arXiv preprint arXiv:2407.03320 (2024)
59. Zhao, W., Feng, H., Liu, Q., Tang, J., Wu, B., Liao, L., Wei, S., Ye, Y., Liu, H., Zhou, W., et al.: Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems* **37**, 7185–7212 (2024)
60. Zhao, Y., Wang, C., Li, C., Cohan, A.: Can multimodal foundation models understand schematic diagrams? an empirical study on information-seeking qa over

- scientific papers. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 18598–18631 (2025)
61. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **36**, 46595–46623 (2023)
 62. Zheng, M., Feng, X., Si, Q., She, Q., Lin, Z., Jiang, W., Wang, W.: Multimodal table understanding. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9102–9124 (2024)
 63. Zhou, B., Gao, Z., Wang, Z., Zhang, B., Wang, Y., Chen, Z., Xie, H.: Syntab-llava: Enhancing multimodal table understanding with decoupled synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24796–24806 (June 2025)
 64. Zhou, Y., Chen, Y., Lin, H., Wu, Y., Yang, S., Qi, Z., Ma, C., Zhu, L.: Dogr: Towards versatile visual document grounding and referring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3596–3606 (2025)
 65. Zhou, Y., Chen, Y., Lin, H., Wu, Y., Yang, S., Qi, Z., Ma, C., Zhu, L., Shan, Y.: Dogr: Towards versatile visual document grounding and referring. GitHub repository, <https://github.com/Tencent/DOGR> (2025), accessed: 2026-02-25
 66. Zhu, F., Liu, Z., Yao, N.X., Wu, H., Wang, W., Feng, F., Wang, C., Luan, H., Chua, T.S.: Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding and grounding. In: Lokoč, J., Peška, L., Zahálka, J., Rudinac, S., Kastner, M., Chen, J., Hu, M.C., Wu, J., Sharma, U. (eds.) *MultiMedia Modeling*. pp. 74–88. Springer Nature Singapore, Singapore (2026)