Unveiling Inherent Visual Grounding in Multimodal LLMs for Text-Rich Images

Shijie Zhou[1*], **Jihyung Kil**[2], **Ming Li**[3], **Jiuxiang Gu**[2], **Curtis Wigington**[2]
Rajiv Jain[2], **Changyou Chen**[1], **Ruiyi Zhang**[2†]
[1]University at Buffalo, [2]Adobe Research, [3]University of Maryland
shijiezh@buffalo.edu, ryzhang.cs@gmail.com

Abstract

Visual text grounding provides interpretable evidence for document question answering. Due to the complex layouts and mixed visual-text contents in text-rich images, effective visual text grounding requires strong visual and spatial reasoning to localize multiple referenced regions. Existing multimodal large language model (MLLM) approaches often struggle to align query tokens with visual–text patches, heavily relying on lengthy OCR inputs. To tackle this problem, we propose Doc-AGround, an OCR-free approach that leverages the MLLM's inherent multi-head attention for multi-patch grounding. Doc-AGround extracts a patch-wise attention map as the grounding prediction. Concurrently, it introduces an effective multi-head weighting mechanism to amplify the attention heads' intrinsic role in connecting vision and text. Empirical results of Doc-AGround show state-of-the-art performance on challenging document grounding benchmarks, demonstrating the effectiveness of the proposed attention-based grounding design.

1 Introduction

Visual grounding aims to map natural-language expressions to their corresponding visual regions, providing contextual references for downstream applications such as referential dialogue (You et al., 2023). It has been studied across diverse settings, ranging from object-level phrase grounding in natural images (Li et al., 2022; Liu et al., 2024b) to implicit actionable-region grounding for graphical user interfaces (GUI) (Wang et al., 2025), where grounded evidence supports explainable verification for user-agent interaction. Recently, this task has been extended to text-rich document images with Multimodal Large Language Models (MLLMs) (Liu et al., 2023b; Alayrac et al., 2022).

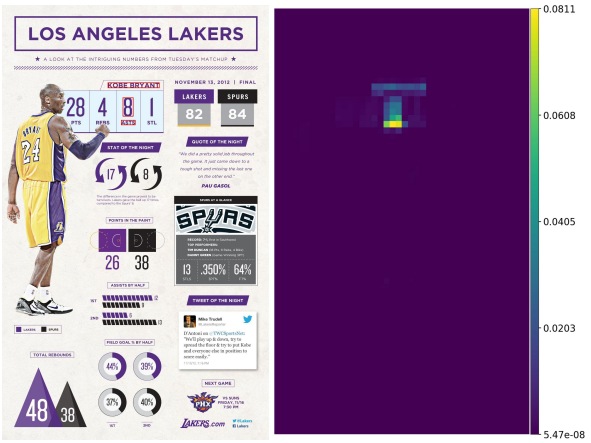


Figure 1: An example of aggregated <g>-token attention map for visual text grounding on InfographicVQA. Ground truth bounding boxes are shown in red. **Question:** How many assists did Kobe make? **Answer:** 8.

Document pages often combine heterogeneous visual elements (Zhong et al., 2020; Mathew et al., 2022), such as charts, infographics, pictograms, and forms, together with textual content, requiring joint semantic, layout, and spatial reasoning (Feng et al., 2023; Appalaraju et al., 2021). Unlike natural images, grounded evidence for document answering is often structured and dispersed across multiple disjoint regions (Li et al., 2025), demanding comprehensive multimodal understanding.

In typical document question-answering settings, explicitly referenced regions can provide a coarse region of interest, and MLLM reasoning further discovers additional supporting context from non-contiguous areas. Existing approaches broadly follow two paradigms. Embedding-based grounding typically matches query representations with image regions (Liu et al., 2024b; Li et al., 2025), which is fragile when the query does not explicitly mention the final evidence to be localized (e.g., question-only grounding). Another line of work prompts an MLLM to generate grounding results before answering (Wang et al., 2023b; Li et al., 2025). However, coordinate-based supervision (Li et al., 2025) (i.e., predicting bounding boxes [x_min, y_min,

[*]Work done while SZ interned at Adobe Research.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 352–370
July 2-7, 2026

x_max, y_max]) primarily emphasizes boundary fitting and underrepresents overlap and region coverage (Tang et al., 2025). Meanwhile, language-only supervision can be data-intensive due to indirect alignment between linguistic semantics and spatial reasoning (Wu et al., 2025).

To improve visual text grounding on document images, many prior methods additionally rely on Optical Character Recognition (OCR) (Cui et al., 2025) inputs with spatial information of textual elements, effectively turning grounding into selecting from OCR-provided candidates (Liu et al., 2024c; Li et al., 2025). However, OCR inputs can be lengthy and introduce inefficiency, and they are less effective on pages with sparse text, such as charts (Masry et al., 2022).

In this work, we propose Doc-AGround, an OCR-free and attention-based method for visual text grounding. Doc-AGround treats the MLLM's multi-head self-attention (MHSA) (Vaswani et al., 2017) between document image patches and question tokens as a patch-wise grounding signal, enabling direct supervision with patch-wise grounding labels. Specifically, following prior practice (Zhou et al., 2025; Lin et al., 2024; Wu et al., 2025) that uses tail tokens to summarize preceding context, we insert a <g> token after the visual-text inputs and before generation. Owing to causal attention, this token aggregates the preceding multimodal semantics. We then extract the attention weights between <g> and all visual patch tokens as head-wise patch-level grounding predictions. An example of the aggregated attention map of <g> is shown in Fig. 1.

To aggregate predictions across heads, we upweight attention heads whose cross-modal correlation patterns are consistent with the model's global multimodal behavior (Kang et al., 2025). Concretely, we estimate a global visual-text correlation pattern from hidden states and compare it with each head's cross-modal attention distribution to compute head importance weights. This design enables fine-grained head contributions without disrupting pretrained grounding generalization. Our contributions can be summarized as follows:

- We introduce Doc-AGround, an **OCR-free**, **attention-grounding** framework for complex document question-answering. Doc-AGround treats grounding as direct supervision on an MLLM's multi-head self-attention, enabling end-to-end evidence localization *without* bounding-box generation.
- Doc-AGround inserts a <g> token to aggregate multimodal context and uses its patch-wise MHSA map for grounding, facilitating full-region coverage beyond bounding-box fitting. We further propose a novel head-weighted merge strategy that favors heads whose visual-text correlation patterns align with a global pattern inferred from the hidden state, enabling fine-grained head contributions.
- Experiments on document grounding benchmarks show that Doc-AGround achieves state-of-the-art in-domain and zero-shot performance, outperforming prior OCR-free and OCR-based methods, and demonstrating the robustness of attention-based grounding for document pages.

2 Related Works

Document Understanding via Vision-language Models Document understanding for text-rich images, such as tables (Zhong et al., 2020), charts (Masry et al., 2022), and infographics (Mathew et al., 2022), has transitioned from traditional supervised pipelines (Wang et al., 2021; Xu et al., 2020a,b; Yu et al., 2023) toward generative Vision-language frameworks (Liu et al., 2024c), enabled by recent MLLMs (Liu et al., 2023b; Bai et al., 2025; Team, 2024) for visual–language tasks ranging from visual-text parsers (Dhouib et al., 2023) to document visual question answering (Zhou et al., 2024a). Approaches are commonly grouped as OCR-dependent (Tang et al., 2023) and OCR-free (Ye et al., 2023a): TRIG (Li et al., 2025) shows that the OCR inputs substantially boost document-grounding performance for MLLMs lacking text-recognition pretraining. LLaVAR (Zhang et al., 2023) utilizes off-the-shelf OCR tools (Cui et al., 2025) with GPT-4 (OpenAI, 2023) to generate synthetic instruction-following data to adapt MLLMs for text-rich images. OCR-free methods (Ye et al., 2023b; Kim et al., 2022), such as UniDoc (Feng et al., 2023), Pix2Struct (Lee et al., 2023), enhance the MLLM's text detection and recognition capacity on text-dense images during both pretraining and fine-tuning. Recent MLLM-based retrieval method SV-RAG (Chen et al., 2024a) improves multi-page document understanding without OCR guidance. In this work, we focus on enhancing patch-level document understanding of MLLMs to generate reliable VQA grounding results.

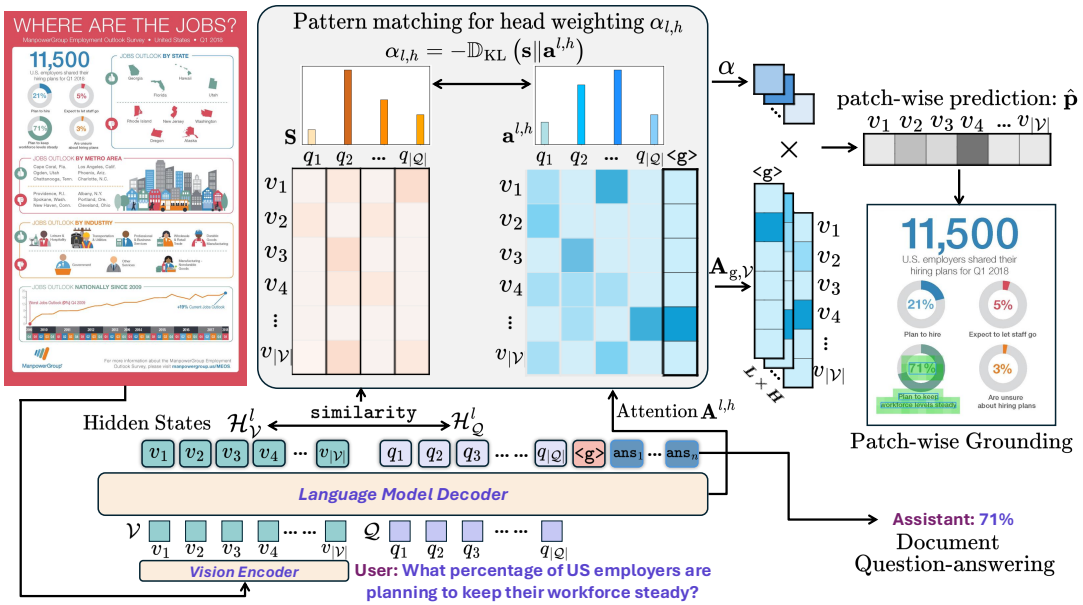


Figure 2: **Architecture of Doc-AGround. (i)** Given a MLLM fed with a text-rich image and a user question, Doc-AGround inserts a <g> token before the answer as the context aggregator for multimodal inputs, and extract the attention vector $\mathbf{A}^{l,h}_{\mathrm{g},\mathcal{V}}$ between <g> and visual tokens as the patch-wise grounding prediction; **(ii)** For each attention head (l, h), we quantify how well its inter-modality pattern (from Eq. (3)) aligns with the general pattern inferred from hidden states (from Eq. (2)) via a negative KL-divergence score $\alpha_{l,h}$; these $\alpha_{l,h}$ weights enable a principled merge of multi-head grounding predictions; **(iii)** Unlike prior embedding-based methods, Doc-AGround naturally supports VQA after grounding.

MLLMs for Visual Grounding Visual grounding aims to connect language particles to their referents in vision (Ma et al., 2024). Early attempts at visual language models focus on the token-level visual grounding on natural images (Li et al., 2022; Liu et al., 2024b). With the advancement of MLLMs (Liu et al., 2023b; Alayrac et al., 2022), recent MLLM-based visual grounding methods such as Ferret (You et al., 2023), Kosmos-2 (Peng et al., 2023), Shikra (Chen et al., 2023), Gpt4roi (Zhang et al., 2024b), combine the object-level grounding into the visual question answering as the referential dialogue, while others, e.g., GRIT (Fan et al., 2025), use the intermediate grounding as the reasoning chain for answering. Another group of work, e.g., SeeClick (Cheng et al., 2024), UI-Tars (Qin et al., 2025), focuses on the visual grounding of Graphical User Interfaces (GUIs). Among them, GUI-Actor (Wu et al., 2025) and TAG (Xu et al., 2025) propose the coordinate-free grounding manner of using MLLMs. Besides the visual grounding on natural images or for GUI agents, recent works (Wang et al., 2023b) on text-rich images primarily treat the visual grounding as the intermediate result for more accurate responses. TextCoT (Luan et al., 2024) utilizes a separate MLLM to extract grounded local text. P2G (Chen et al., 2024b) uses an OCR model to provide extra textual clues for answering. Besides, TextMonkey (Liu et al., 2024c) also includes text and VQA groundings into pretraining tasks. In this work, we concentrate on the well-formatted visual text grounding task introduced by TRIG (Li et al., 2025) to improve the grounding capacity of the MLLM on document question-answering.

3 Methods

3.1 Task Formulation: Visual Text Grounding on Text-rich Images

Given a text-rich image $\mathcal{I}$ and a user query $\mathcal{Q} = [q_1, q_2, \ldots, q_{|\mathcal{Q}|}]$, we follow standard MLLM tokenization by patchifying the image into visual tokens $\mathcal{V} = [v_1, v_2, \ldots, v_{|\mathcal{V}|}]$. Visual text grounding aims to localize the evidence regions in $\mathcal{I}$ that support answering the query, represented as a patch-level target over $\mathcal{V}$.

We focus on the query-only setting and study grounding as evidence localization that precedes answer generation. This setting is challenging for prior embedding-based grounding methods, since their decoupled vision and language modules (see Section 1) provide limited cross-modal conditioning for question-only inputs. We further adopt an OCR-free input format by excluding OCR text and bounding boxes, so that grounding performance reflects the model's intrinsic ability to localize visual

evidence rather than leverage external OCR tools.

In Doc-AGround, we propose a bounding-box-free method with direct alignments on a single MLLM's multi-head self-attention instead of on the similarity map computed from token-wise embeddings. After grounding-specific fine-tuning, the patch-wise localization in the text-rich image can be inferred from the inherent aligned self-attention to support the input questions, while allowing the followed answering in the same model.

3.2 Visual Text Grounding Indications in Multi-head Attention of MLLMs

With the patch tokens $\mathcal{V}$ of fixed patch size and the query tokens $\mathcal{Q}$ as the input, the L-layer transformer-based language decoder generates hidden states $\{\mathcal{H}_l\}_{l=1}^{L}$ after each layer. Given the preceding layer's hidden state $\mathcal{H}_{l-1}$, in layer l, H-head self-attention $\{\mathbf{A}^{l,h}\}_{h=1}^{H}$ is computed from query $Q(\mathcal{H}_{l-1})$ and key $K(\mathcal{H}_{l-1})$ embeddings to capture the head-wise correlation between input tokens. Specifically, for the h-th head, the correlation between question tokens $\mathcal{Q}$ and visual tokens $\mathcal{V}$ of the text-rich image can be indicated by the partial attention $\mathbf{A}^{l,h}_{\mathcal{Q},\mathcal{V}}$ crossing modalities from the complete $\mathbf{A}^{l,h}$. Compared with using the final layer's embedding $\mathcal{H}_{\mathcal{Q}}$ and $\mathcal{H}_{\mathcal{V}}$ to compute the similarity for grounding in previous embedding-based methods, each token q_i's attention vector $\mathbf{A}^{l,h}_{q_i,\mathcal{V}}$ with all image patches in $\mathbf{A}^{l,h}_{\mathcal{Q},\mathcal{V}}$ is a nature patch-wise grounding indication. If q_i is the representative token that serves as the semantic conclusion for query $\mathcal{Q}$, the magnitude of entry $\mathbf{A}^{l,h}_{q_i,v_i}$ in $\mathbf{A}^{l,h}_{q_i,\mathcal{V}}$ illustrates the importance of image patch v_i towards the input query, indicating whether patch v_i is the grounding area.

However, for different queries, the measurement of representativeness of each text token is impractical. Instead of finding the precise weighting $\{\alpha_{q_i}\}_{i=1}^{|\mathcal{Q}|}$ to get the grounding prediction $\sum_{i\in[|\mathcal{Q}|]} \alpha_{q_i} \cdot \mathbf{A}^{l,h}_{q_i,\mathcal{V}}$ from $\mathbf{A}^{l,h}$, we propose to construct a special token after the query $\mathcal{Q}$ to avoid token-wise weighting α_{q_i}.

Grounding token as the query compressor. Previous works that use a MLLM as a judge or a classifier for multimodal inputs usually utilize the single embedding of inherently final-positioned text token after the query, such as the end-of-sentence (EOS) token, as the whole input (image and text tokens)'s representative embedding for downstream tasks. Rather than reusing an existing vocabulary token—which could interfere with the tokens' original functionality, as shown in the blue attention in Fig. 2, we insert a new special token <g> after $\mathcal{Q}$ and before the potential answer for grounding. We extract the attention vector $\mathbf{A}^{l,h}_{\text{g},\mathcal{V}}$ between <g> and all visual tokens as the grounding indication of l-th layer's h-th attention head $\mathbf{A}^{l,h}$, instead of the complex $\sum_{i\in[|\mathcal{Q}|]} \alpha_{q_i} \cdot \mathbf{A}^{l,h}_{q_i,\mathcal{V}}$.

Now, we can solve the impractical grounding by adding <g> after $\mathcal{Q}$ and extracting $\mathbf{A}^{l,h}_{\text{g},\mathcal{V}}$ as the inherent grounding indication instead of finding α_{q_i} for $\mathbf{A}^{l,h}_{q_i,\mathcal{V}}$ of every query token. Besides the above token-level design, the strategy for merging grounding predictions of all attention heads for comprehensive grounding is underexplored. Considering the functional differences across different attention heads in the MLLM, treating each attention head equally and directly supervising on $\mathbf{A}^{l,h}_{\text{g},\mathcal{V}}$ given ground-truth patches $\mathcal{V}^{gt}$ is very biased. Instead, we formulate the weighting importance for h-th attention head of l-layer as $\alpha_{l,h}$ for merging:

$$\hat{\mathbf{p}} = \sum_{l,h} \alpha_{l,h}\, \mathbf{A}^{l,h}_{\text{g},\mathcal{V}} \quad \in \mathbb{R}^{|\mathcal{V}|}. \tag{1}$$

$\hat{\mathbf{p}}$ is the attention grounding prediction considering all heads. In Section 3.3, we propose a strategy for $\alpha_{l,h}$ for better multi-head control.

3.3 Multi-head Weighting Guided by General Visual-text Pattern of the MLLM

Previous extensive studies (Gould et al., 2023; Wu et al., 2024; Zhou et al., 2024b) on the multi-head attention mechanism have revealed that the importance of each attention head varies for the MLLM's different functionalities, such as safety, factuality. For visual text grounding on text-rich images, we need a measurement to identify and emphasize the group of attention heads that originally focus on the interactions between visual and text modalities in Eq. (1).

For the transformer-based MLLM, the inherent inter-modality pattern can be discovered from each layer's hidden state. For text token q_i, we can take the sum of its embedding similarity between all visual tokens as q_i's visual focus magnitude. Thus, for all text tokens, we can have the general inter-modality distribution $\mathbf{s} = [s_{q_1}, \ldots, s_{q_{|\mathcal{Q}|}}]$ as:

$$\mathbf{s} = \text{normalize}([\sum_{v_j\in\mathcal{V},\, l\in[L]} \text{Sim}(\mathcal{H}^l_{q_i}, \mathcal{H}^l_{v_j})]_{i=1}^{|\mathcal{Q}|}). \tag{2}$$

$\mathbf{s}$ can be treated as the general text-visual correlation pattern considering all layers. Similarly, for attention head h of layer l, we can leverage the similarity implied by the inherent attention value to measure the head-level text-visual correlation distribution $\mathbf{a}^{l,h}$:

$$\mathbf{a}^{l,h} = \text{normalize}([\sum_{v_j \in \mathcal{V}} \mathbf{A}^{l,h}_{q_i,v_j}]_{i=1}^{|\mathcal{Q}|}). \quad (3)$$

Although in each transformer block, the output hidden states $\mathcal{H}^l$ are forwarded from all attention heads of this layer, previous studies (Voita et al., 2019; Elhelo and Geva, 2024) have shown that the semantic pattern of $\mathcal{H}^l$ is contributed from a minority of attention heads. For the general text-visual correlation pattern we found in $\mathcal{H}$ via Eq. (2), we aim to find the minority attention heads that share a similar functionality pattern and highlight their contributions in Eq. (1) via $\alpha_{l,h}$. Here, we measure the pattern matching of each head between $\mathbf{s}$ and $\mathbf{a}^{l,h}$ as the negative Kullback-Leibler divergence $\mathbb{D}_{\text{KL}}$ for $\alpha_{l,h}$, denoted as **dense** fully-soft alignment shown as the pattern matching in Fig. 2, which is the default setting of Doc-AGround:

$$\textbf{Dense: } \alpha_{l,h} = -\mathbb{D}_{\text{KL}}(\mathbf{s} \| \mathbf{a}^{l,h}). \quad (4)$$

For the distribution matching measurement between $\mathbf{s}$ and $\mathbf{a}^{l,h}$, the dense fully-soft alignment in Eq. (4) might bring unexpected bias coming from the long tail query tokens in the general pattern distribution $\mathbf{s}$. Query text tokens lying in long tail of $\mathbf{s}$ are less semantic or visually related and their normalized magnitude in $\mathbf{s}$ is small but nontrivial. The participation of this weak tail in the KL-divergence computation of Eq. (4) could bring in unnecessary KL penalties. Besides the dense fully-soft matching above, we take the top-K globally salient query tokens from distribution $\mathbf{s}$ and formulate the **sparse** matching weight as another option for $\alpha_{l,h}$:

$$\textbf{Sparse: } \alpha_{l,h} = \sum_{q_i \in \mathcal{Q}^*} \mathbf{s}(q_i) \cdot \mathbf{a}^{l,h}(q_i), \quad (5)$$

where $\mathcal{Q}^* = \arg\text{topK}_{q_i \in \mathcal{Q}}(\mathbf{s}(q_i))$. In this way, we can highlight the attention heads matching the general text-visual correlation pattern only from informative sparse query tokens without the distraction from visual-irrelevant query tokens.

With softmax normalized head-wise weighting $\alpha_{l,h} = \exp(\alpha_{l,h}) / \sum_{l' \in [L], h' \in [H]} \exp(\alpha_{l',h'})$, we can achieve the attention grounding prediction $\hat{\mathbf{p}}$ from Eq. (1).

Training. Given ground-truth image patches $\mathcal{V}^{gt} \in \mathcal{V}$, the ground-truth patch distribution is defined as:

$$\mathbf{p} = \frac{1}{|\mathcal{V}^{gt}|} \left[\mathbb{1}_{\mathcal{V}^{gt}}(v_1), \ldots, \mathbb{1}_{\mathcal{V}^{gt}}(v_{|\mathcal{V}|})\right] \in \Delta^{|\mathcal{V}|-1}. \quad (6)$$

We compute the KL-divergence between ground-truth $\mathbf{p}$ and attention prediction $\hat{\mathbf{p}}$ as the visual text grounding loss:

$$\mathcal{L}_{\text{grounding}} = \mathbb{D}_{\text{KL}}(\mathbf{p} \,\|\, \text{normalize}(\hat{\mathbf{p}})). \quad (7)$$

When the input query only includes the question on the text-rich image, after grounding, the MLLM needs to generate the corresponding answer. The extra next-token prediction loss on tokens of the answer is added for MLLM's supervised fine-tuning:

$$\mathcal{L} = \mathcal{L}_{\text{grounding}} + \mathcal{L}_{\text{next-token}}. \quad (8)$$

For $\alpha_{l,h}$, we treat the global pattern $\mathbf{s}$ derived from hidden states as a fixed target and do not back-propagate gradients to it. Thus, gradients flow only through the head-wise pattern $\mathbf{a}^{l,h}$.

Inference. For inference, after achieving the patch-wise prediction vector $\hat{\mathbf{p}}$ from Eq. (1) via dense or sparse pattern matching in Eq. (4) or Eq. (5), we activate the patch v_i for $i \in \{1, \ldots, |\mathcal{V}|\}$ if its value exceeds an adaptive activation threshold: $\hat{\mathbf{p}}(v_i) \geq m \cdot \max(\hat{\mathbf{p}})$, with $m \in (0, 1]$ controlling the activation sparsity level.

4 Experiments

4.1 Experimental Setups

Implementation Details. For Doc-AGround, we utilize Qwen2.5-VL-3B (Bai et al., 2025) as the backbone of Doc-AGround for document visual text grounding and set the vanilla attention implementation for the extraction of full attention weight. Doc-AGround is trained for one epoch with all parameters unfrozen on NVIDIA A100 (80GB) GPUs (around 59 GPU-hours total) with a batch size of 32 and a learning rate of 1e-5. We implement the dense matching in Eq. (4) and sparse matching in Eq. (5) as two Doc-AGround variants, both in the OCR-free setting. Furthermore, Doc-AGround is fine-tuned with question-only inputs or the entire QA pair input, to show the effect of the explicit answer in grounding. The hyperparameter m for the adaptive threshold $m \cdot \max(\hat{\mathbf{p}})$ is set to 0.15 for all experiments. The details of m selection are provided in Appendix A.1.

Datasets and Metrics. Doc-AGround is fine-tuned on the training sets of three open-source document datasets with grounding annotations: DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), and InfographicVQA (Mathew et al., 2022), with a total of 71k training samples. We conduct the in-domain evaluation on the same three sets and zero-shot evaluation on the TRINS (Zhang et al., 2024a) set in TRIG (Li et al., 2025). For evaluation, we include Intersection over Union (IoU), Precision, Recall, and F1-score as the metrics for OCR-based grounding and Pixel-level IoU for OCR-free grounding. The definitions of metrics are listed in Appendix B.

Baselines. We compare Doc-AGround with three groups of baselines: (i) general open-source MLLMs for OCR-based zero-shot evaluation, including LLaVA-v1.6-Vicuna (Liu et al., 2023b,a, 2024a), Idefics2-8B (Laurençon et al., 2024), DeepSeek-VL-7B-chat (Lu et al., 2024), InternLM-XComposer2-VL-7B (Dong et al., 2024a), InternLM-XComposer2-4KHD-7B (Dong et al., 2024b), Monkey-Chat (Li et al., 2024), Phi3-V(Team, 2024), CogVLM2-Llama3-19B (Wang et al., 2023a), MiniCPM-Llama3-V 2.5 (Yao et al., 2024), InternVL2 (Chen et al., 2024c), InternVL2.5 (Chen et al., 2024c), and Qwen2.5-VL (Bai et al., 2025); (ii) closed-source MLLMs including Gemini-2.5-flash (DeepMind, 2025), and GPT-4o (OpenAI, 2024) with OCR inputs; (iii) visual text grounding-specific methods including embedding-based and TRIG (instruction) variants (Li et al., 2025) trained on all four document grounding datasets with both OCR-based and OCR-free settings, also including Qwen2.5-VL fine-tuned with ChartQA, DocVQA, and InfographicVQA in the same instruction-based manner as TRIG for backbone ablations. Extra baseline details are provided in Appendix A.2.

OCR-based Patch Filter For OCR-based baselines, all the bounding boxes and corresponding text obtained from the OCR tool are included in the prompts for MLLMs' training and inference. In this manner, the grounding task is turned into the easier multi-area selection problem.

Although Doc-AGround is OCR-free during both training and inference, we can still leverage OCR outputs to post-filter predicted image patches, which involves no model forwards. Specifically, we convert the image's OCR bounding boxes into a one-hot, patch-wise filter vector $\mathbf{m}$ and suppress patches that do not fall within any OCR box, forming the grounded prediction $\mathbf{m} \odot \hat{\mathbf{p}}$. As shown in Section 4.3, the improvement from the OCR filter is modest.

Method	ChartQA	DocVQA	InfographicVQA	TRINS
	Pixel IoU	Pixel IoU	Pixel IoU	Pixel IoU
Qwen2.5-VL-3B†	0.43	0.06	0.18	0.23
CogVLM2-Llama3-19B†	0.19	0.01	0.16	0.66
Monkey-Chat†	0.77	0.19	0.15	0.45
InternLM-XComposer2-4KHD-7B†	1.04	0.10	0.90	0.14
MiniCPM-Llama3-V 2.5†	0.44	1.40	0.65	4.96
Gemini2.5-Flash†	0.00	0.00	0.00	0.00
GPT-4o†	3.90	1.79	1.60	13.73
TRIG (Embedding) + Answer†	10.51	15.02	7.85	13.88
TRIG (Instruction)†	27.91	23.96	8.61	59.44
Qwen2.5-VL-3B (Instruction)	33.79	28.50	11.23	53.77
Doc-AGround (Dense)	**81.08**	**80.17**	62.74	**83.17**
Doc-AGround (Sparse)	78.88	79.67	**64.86**	81.11
Doc-AGround (Sparse) + Answer	85.55	83.14	71.86	85.02

Table 1: **OCR-free Document grounding results on TRIG benchmark. Pixel IoU** represents Pixel-level Intersection over Union score. Doc-AGround is fine-tuned on the first three datasets and zero-shot evaluated on TRINS. TRIG variants are trained on all four datasets. † denotes results reported from previous works. "+ Answer" means adding extra answer strings to the inputs.

4.2 Main Results

In this section, we present the evaluation results of document grounding on TRIG and baselines shown in Table 1 and Table 2.

Baselines. First, **Coordinate grounding is hard for instruction-following models in the OCR-free setting.** As shown in Table 1, both open-source and proprietary MLLMs struggle to generate coordinate-based bounding boxes without OCR, suggesting limited spatial grounding and formatting ability to infer implicit supporting regions. **Training-time adaptation is still inefficient:** TRIG (instruction) and Qwen2.5-VL-3B remain weak even in-domain, with a large gap to OCR-based proprietary models, e.g., GPT-4o.

Under the **OCR-based** setting, the extra OCR inputs simplify visual text grounding. Providing bounding boxes and corresponding content in the prompt reduces the need for generating autoregressive numeric coordinates, turning grounding into a selection problem among the given boxes, which is more explicit and straightforward. Consequently, models with grounding pretraining, such as Qwen2.5-VL-7B and InternVL2.5-8B, perform relatively well as shown in Table 2. TRIG (instruction) improves substantially with OCR-based fine-tuning, highlighting that **bridging grounding signals to explicit coordinates benefits from OCR inputs under limited supervision.**

Embedding matching methods, such as TRIG

Method	ChartQA					DocVQA					InfographicVQA					TRINS				
	IoU	P	R	F1	Avg.	IoU	P	R	F1	Avg.	IoU	P	R	F1	Avg.	IoU	P	R	F1	Avg.
LLaVA-v1.6-Vicuna-7B†	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-v1.6-Vicuna-13B†	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Idefics2-8b†	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.17	1.17	3.50	1.69	1.88
DeepSeek-VL-7B-chat†	0.84	1.35	2.25	1.23	1.42	0.19	0.19	1.50	0.34	0.56	0.00	0.00	0.00	0.00	0.00	1.60	1.60	2.00	1.67	1.72
InternLM-XComposer2-4KHD-7B†	1.00	2.00	1.00	1.25	1.31	0.25	0.50	0.25	0.33	0.33	0.00	0.00	0.00	0.00	0.00	3.67	6.80	3.79	4.65	4.73
Monkey-Chat†	3.58	5.38	9.91	5.05	5.98	0.94	1.19	1.75	1.15	1.26	0.34	0.30	2.00	0.51	0.79	0.00	0.00	0.00	0.00	0.00
Phi3-V†	2.54	3.14	5.46	3.25	3.60	0.97	1.22	2.00	1.21	1.35	0.40	0.35	2.08	0.59	0.85	3.81	4.31	5.00	4.28	4.35
CogVLM2-Llama3-19B†	1.65	2.30	3.68	2.02	2.41	1.56	1.87	3.58	1.89	2.23	0.42	0.84	1.75	0.71	0.93	6.76	7.93	8.33	7.53	7.64
Qwen2.5-VL-3B†	3.77	5.74	5.37	5.02	4.98	2.46	3.00	3.58	3.02	3.02	0.10	0.17	0.17	0.17	0.15	10.19	12.75	14.59	12.34	12.47
InternLM-XComposer2-VL-7B†	8.10	15.31	9.03	10.49	10.73	7.36	11.28	8.08	8.73	8.86	2.32	6.39	2.85	3.39	3.74	15.82	18.12	17.32	16.98	17.06
InternVL2-1B†	3.43	3.28	19.10	5.35	7.79	0.10	0.10	0.75	0.17	0.28	0.35	0.35	2.71	0.62	1.01	23.91	32.30	48.38	31.10	33.92
MiniCPM-Llama3-V 2.5†	7.40	12.50	8.40	9.24	9.39	15.78	19.22	18.42	17.48	17.73	5.71	9.87	7.34	7.14	7.52	54.06	62.06	57.89	57.95	57.99
InternVL2.5-8B†	9.99	12.54	14.24	11.51	12.07	5.71	6.50	7.58	6.25	6.51	5.11	8.06	8.46	6.70	7.08	59.51	63.98	70.73	64.05	64.57
Qwen2.5-VL-7B†	16.84	26.04	18.61	20.22	20.43	24.12	27.33	25.75	25.80	25.75	5.81	11.42	6.40	7.69	7.83	55.12	66.21	66.45	62.80	62.65
Gemini2.5-Flash†	74.65	81.31	81.06	78.92	78.99	73.67	81.86	78.00	77.71	77.81	66.28	73.11	79.58	73.12	73.02	72.78	78.43	76.85	75.80	75.97
GPT-4o†	83.80	88.80	89.24	87.47	**87.33**	82.14	87.14	89.50	86.16	**86.23**	68.19	79.57	78.81	75.82	**75.56**	89.08	96.06	91.53	92.16	**92.21**
TRIG (Embedding) + Answer†	39.97	57.26	52.89	49.98	50.03	37.82	40.09	72.96	48.42	49.82	25.58	28.68	49.14	32.93	34.08	70.01	86.38	75.94	77.32	77.42
TRIG (Instruction)†	70.38	81.58	73.78	75.78	75.38	73.52	81.67	75.13	77.02	76.83	39.23	47.29	42.86	43.90	43.32	85.48	92.17	79.94	83.62	85.30
Doc-AGround (Dense)	81.08	86.68	92.09	86.85	**86.67**	80.17	90.64	86.77	87.02	**86.15**	62.74	72.24	81.7	72.53	72.30	83.17	93.17	87.15	88.76	**88.06**
Doc-AGround (Sparse)	78.88	85.23	90.12	85.25	84.87	79.67	89.61	87.68	86.68	85.91	64.86	73.69	82.14	73.83	**73.63**	81.11	93.81	84.55	87.33	86.70
Doc-AGround (Sparse) + Answer	85.55	91.59	92.50	90.25	89.97	83.14	92.79	89.09	89.39	88.60	71.86	82.26	84.87	80.36	79.84	85.02	96.28	88.15	89.84	89.82

Table 2: **OCR-based Document grounding results on TRIG benchmark. IoU**, **P**, **R**, **F1** represent Intersection over Union score, precision, recall and F1-score. For each dataset, we denote **Avg.** as the average score on evaluation metrics. TRIG variants are trained on all four datasets. † denotes results reported from previous works. Doc-AGround is fine-tuned **without OCR inputs** on the first three datasets and zero-shot evaluated on TRINS. "+ Answer" means adding extra answer strings to the inputs.

$\alpha_{l,h}$	Method	ChartQA Avg.	DocVQA Avg.	Infographic Avg.	TRINS Avg.	Overall Avg.
Uniform	**Doc-AGround**	85.45	84.75	71.10	81.88	80.80
	Doc-AGround (w/o <g>)	83.20	83.43	67.95	78.12	78.18
Dense	**Doc-AGround**	86.67	86.15	72.30	88.06	**83.29**
	Doc-AGround (Qwen2-VL-2B)	82.17	84.15	68.82	84.57	79.93
	Doc-AGround (w/o OCR filter)	84.62	84.94	70.72	85.65	81.49
	Doc-AGround (top 30% heads)	86.55	86.02	72.27	87.71	83.13
	Doc-AGround (lowest 30% heads)	81.46	81.80	71.28	76.19	77.68
Sparse	**Doc-AGround** (top-1)	84.87	85.91	73.63	86.70	82.78
	Doc-AGround (top-3)	85.00	85.88	73.10	87.64	82.90
	Doc-AGround (top-5)	85.15	86.62	73.29	86.96	83.00
	Doc-AGround (lowest-1)	84.08	82.88	68.68	79.56	78.80

Table 3: **Ablation results on different attention head merge manners and the OCR-based filter. Avg.** represents the average score on four evaluation metrics. Doc-AGround is zero-shot evaluated on TRINS.

(embedding), already performs multi-area selection via patch-text embedding matching. Thus, extra OCR inputs provide limited benefits over the OCR-free setting (13.01 less improvement than TRIG (instruction) on IoU). In its contrastive training, only including text hard negatives is insufficient. Also, the decoupled encoders prevent text features from conditioning on images, limiting effectiveness when evidence is not explicitly provided. These factors contribute to the ∼17% average gap to TRIG (instruction), even with ground truth answers.

Doc-AGround: strong OCR-free grounding with efficient supervision. In the OCR-free setting, Doc-AGround achieves comparable performance with GPT-4o on ChartQA, DocVQA, and InfographicVQA. It also outperforms OCR-based open-source and grounding-adapted baselines (e.g., ∼16.53% over the strongest TRIG (instruction) on average), with the largest gains on InfographicVQA (+30.31%), the most visually complex benchmark. Unlike coordinate prediction, Doc-AGround directly computes correlations between each patch token and the query compressor token <g>, producing precise patch-level indications. **This direct attention supervision** improves efficiency in training (without TRINS, two fewer epochs). And the attention-grounding design enables faster inference on TRIG (0.73s/iteration compared with 1.02s for Qwen2.5-VL-3B), as it uses a single <g> token instead of multiple coordinate tokens for grounding.

Doc-AGround achieves better zero-shot performance on TRINS than the in-domain performance of both TRIG variants and Qwen2.5-VL-3B, suggesting that attention-based fine-tuning improves visual text grounding without limiting the grounding generalization of the base MLLM. This result further supports our design of using a single MLLM to model both modalities under causal conditioning: the <g> embedding accumulates context from prior visual and query tokens, making $\mathbf{A}^{l,h}_{\mathrm{g},\mathcal{V}}$ a stronger indicator than cross-modal embedding matching in TRIG (embedding). Finally, **dense** full-distribution head weighting (Eq. (4)) slightly outperforms the **sparse** variant (Eq. (5)), consistent with the claim that long-tail tokens provide marginal additional benefit. Moreover, Doc-AGround performs well with a weaker backbone (Qwen2-VL-2B in Table 3), demonstrating robustness across backbones.

4.3 Ablations

Functionality of <g> Token. As shown in Table 3, under the same uniform head weighting, removing the <g> token reduces the average score from 80.80% to 78.18%. Without <g>, the weighted merge of grounding predictions from all

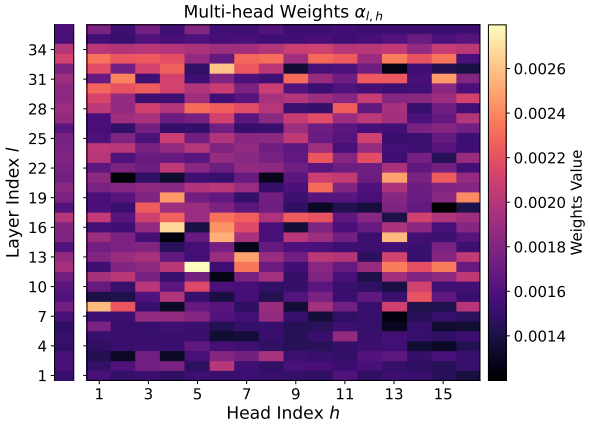


Figure 3: Visualization of multi-head weights $\alpha_{l,h}$ averaged over the TRIG test set. **Left:** layer-wise mean $\frac{1}{H}\sum_{h=1}^{H}\alpha_{l,h}$. **Right:** per-layer, per-head weights $\alpha_{l,h}$.

query tokens is necessary, for which appropriate weighting is challenging and hurts performance. In contrast, <g> simplifies this need by serving as a single context aggregator for all query tokens under causal attention. In practice, using a newly added <g> token or an existing vocabulary token makes no difference in grounding behavior. We further discuss our choice of <g> in Appendix D.

Effect of Different Head-wise Weighting $\alpha_{l,h}$ for $\mathbf{A}^{l,h}_{\mathbf{g},\mathcal{V}}$. In Table 3, we ablate head-wise weighting $\alpha_{l,h}$ strategies for the merge of head-wise attention maps in Eq. (1). Using uniform weights across heads leads to a 2.49% degradation, indicating that treating all heads uniformly is suboptimal, as grounding-irrelevant attention heads introduce biased patch predictions, and uniformly supervising them can also disrupt their pretrained roles.

We further compare dense matching (Eq. (4)) with sparse top-K matching (Eq. (5)), where $\alpha_{l,h}$ is computed from either all query tokens or only a few query tokens with the largest cumulative connections to visual tokens. We can see a flat overall declining trend when the distribution for matching is sparser from top-5 to top-1, suggesting that aligning the long-tail portions of of $\mathbf{s}$ and $\mathbf{a}_{l,h}$ in the dense variant is not critical but still provides a small benefit to overall grounding. Interestingly, sparse matching performs better on the challenging InfographicVQA in Table 2, suggesting that including visually irrelevant query tokens in the text-visual pattern matching can inject noise and even hurt head weighting in challenging cases. In contrast, using the least-salient token (lowest-1 variant), i.e., $\mathcal{Q}^* = \arg\min_{q_i \in \mathcal{Q}}(\mathbf{s}(q_i))$, leads to a 4.49% degradation as shown in Table 3, validating the importance of selecting visual-informative query tokens for computing the weight $\alpha_{l,h}$.

Effect of OCR-based Patch Filter. In Table 3, we compare the performance of Doc-AGround with or without the OCR-based patch filter described in Section 4.1. Although post-filtering the patch-wise prediction $\hat{\mathbf{p}}$ results in 1.8% improvement, Doc-AGround without post-filtering can still outperform most of the baselines. This indicates the robustness of Doc-AGround, as outlier patches rarely arise.

4.4 Analysis of Attention Head Contributions

In Fig. 3, we visualize the multi-head weights $\alpha_{l,h}$ during inference on the TRIG benchmark. The results show that heads from mid (e.g., 12, 13, 17) and late layers (e.g., 31, 33, 34) contribute more to grounding. While mid layers contain the most salient heads (e.g., head 12-5), late layers, such as 33, include more heads with higher weights.

We further conduct ablations by preserving only the top 30% or the bottom 30% of heads by $\alpha_{l,h}$ at inference time. As shown in Table 3, keeping the top 30% heads maintains nearly unchanged in-domain performance, with a 0.35% drop in zero-shot performance. In contrast, only preserving the bottom 30% heads severely reduces performance, demonstrating that $\alpha_{l,h}$ **reliably reflects each head's importance for grounding.** We include the visualization of $\alpha_{l,h}$ for each TRIG benchmark subset in Appendix E, showing that **more heads participate actively in grounding on challenging samples (e.g., InfographicVQA)**.

4.5 Qualitative Results

Qualitative examples of Doc-AGround's grounding results are shown in Appendix C. We also provide visualizations of the aggregated <g>-token attention map in Fig. 1 and Fig. 9 of Appendix F, along with analyses of both correct and failure cases.

5 Conclusion

We propose Doc-AGround, an attention-based visual text grounding method for text-rich document images. We introduce a simple yet effective framework that derives patch-wise grounding signals by inserting a context-compressor token after the multimodal inputs, and we merge per-head predictions by measuring the alignment between a global pattern and local head-wise patterns. Doc-AGround achieves significant in-domain and zero-shot improvements over prior methods on challenging document-grounding benchmarks.

6 Limitations

Doc-AGround still utilizes a single patch-wise vector in each attention head for the multi-area grounding on document images. For more complex cases, where the ground-truth multiple grounding areas are overlapped or visually distinct, a single context compressor token is not sufficiently expressive. We will explore improving the multi-area document grounding task by learning disentangled patch-wise vectors for each region and resolving assignments via Hungarian matching. We will further improve the robustness of Doc-AGround by adding a box-format loss that encourages regular box-shaped grounding predictions without isolated outliers.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A Rossi, Changyou Chen, and Tong Sun. 2024a. Svrag: Lora-contextualizing adaptation of mllms for long document understanding. *arXiv preprint arXiv:2411.01106*.

Jiaxing Chen, Yuxuan Liu, Dehu Li, Xiang An, Weimo Deng, Ziyong Feng, Yongle Zhao, and Yin Xie. 2024b. Plug-and-play grounding of reasoning in multimodal large language models. *arXiv preprint arXiv:2403.19322*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. 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.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, and 1 others. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Google DeepMind. 2025. Gemini 2.5 pro preview model card. Technical report, Google DeepMind.

Mohamed Dhouib, Ghassen Bettaieb, and Aymen Shabou. 2023. Docparser: End-to-end ocr-free information extraction from visually rich documents. In *International Conference on Document Analysis and Recognition*, pages 155–172. Springer.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024a. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, and 5 others. 2024b. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.

Amit Elhelo and Mor Geva. 2024. Inferring functionality of attention heads from their parameters. *arXiv preprint arXiv:2412.11965*.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*.

Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. 2023. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2023. Successor heads: Recurring, interpretable attention heads in the wild. *arXiv preprint arXiv:2312.09230*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9339–9350.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, and 1 others. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975.

Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. 2025. Towards visual text grounding of multimodal large language model. *arXiv preprint arXiv:2504.04974*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024c. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. 2024. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*.

Ziqiao Ma, Jiayi Pan, and Joyce Chai. 2024. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. *Preprint*, arXiv:2306.08685.

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

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 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.

OpenAI. 2024. Introducing gpt-4o. Available at: https://openai.com/index/hello-gpt-4o.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

Fei Tang, Zhangxuan Gu, Zhengxi Lu, Xuyang Liu, Shuheng Shen, Changhua Meng, Wen Wang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and 1 others. 2025. GUI-G^2: Gaussian reward modeling for gui grounding. *arXiv preprint arXiv:2507.15846*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264.

MS Phi-3 Team. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, and 1 others. 2025. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023a. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. 2023b. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. Layoutreader: Pre-training of text and layout for reading order detection. *arXiv preprint arXiv:2108.11591*.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, and 1 others. 2025. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Hai-Ming Xu, Qi Chen, Lei Wang, and Lingqiao Liu. 2025. Attention-driven gui grounding: Leveraging pretrained multimodal large language models without fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8851–8859.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, and 1 others. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. 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.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint 2408.01800*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, and 1 others. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, and 1 others. 2023b. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. Structextv2: Masked visual-textual prediction for document image pre-training. *arXiv preprint arXiv:2303.00289*.

Ruiyi Zhang, Yanzhe Zhang, Jian Chen, Yufan Zhou, Jiuxiang Gu, Changyou Chen, and Tong Sun. 2024a. Trins: Towards multimodal language models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22584–22594.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2024b. Gpt4roi: Instruction tuning

large language model on region-of-interest. In *European conference on computer vision*, pages 52–70. Springer.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer.

Shijie Zhou, Ruiyi Zhang, Yufan Zhou, and Changyou Chen. 2024a. A high-quality text-rich image instruction tuning dataset via hybrid instruction generation. *arXiv preprint arXiv:2412.16364*.

Shijie Zhou, Ruiyi Zhang, Huaisheng Zhu, Branislav Kveton, Yufan Zhou, Jiuxiang Gu, Jian Chen, and Changyou Chen. 2025. Multimodal llms as customized reward models for text-to-image generation. *arXiv preprint arXiv:2507.21391*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2024b. On the role of attention heads in large language model safety. *arXiv preprint arXiv:2410.13708*.

A Extra Experimental Details

A.1 Selection of Hyperparameter

For the hyperparameter m in the adaptive activation threshold $m \cdot \max(\hat{\mathbf{p}})$, we split a validation set of 200 samples from the training set and compute the average number of positive patches per sample from the patch labels. Then, for m ranging from 0.10 to 0.50 with a step size of 0.05, we measure the average number of activated patches per sample. We find that $m = 0.15$ leads to the closest average number of activated patches to the label statistics on the validation set. Through this process, we expect the threshold ratio to activate a reasonable number of grounding patches, avoiding overly expanded or overly sparse coverage. We fix this value for all experiments, and it also works well for our zero-shot evaluation on TRINS. This threshold selection uses only a small subset of the training data and does not involve any evaluation metrics or test data, to avoid leakage.

A.2 Extra Baseline Details

For the fine-tuned Qwen-2.5-VL-3B, we follow the standard coordinate-training format of the Qwen-2.5-VL series: images are resized to dynamic resolutions (multiples of 28 for patch alignment), and bounding-box coordinates are scaled and rounded to integer pixel coordinates, instead of high-precision floating-point coordinates. For the other training settings, we set the learning rate to 1e-5 and the batch size to 64.

B Evaluation Metrics

We report grounding performance under two evaluation settings following prior work: **OCR-free grounding** (pixel-level overlap) and **OCR-based grounding** (instance retrieval). For both settings, we denote the test set size as N. We follow the metric definitions in TRIG (Li et al., 2025).

B.1 Setting 1: OCR-free Grounding (Pixel-level IoU)

In the OCR-free setting, the model predicts one or multiple grounded regions on the image. For sample i, we convert predictions into a set of foreground pixels $\mathcal{P}_i^{\text{pix}} \subseteq \{1, \ldots, H\} \times \{1, \ldots, W\}$, and similarly denote the ground-truth foreground pixel set as $\mathcal{G}_i^{\text{pix}}$.

We evaluate by **Pixel-level IoU**:

$$\mathrm{IoU}_{\mathrm{pixel}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\mathcal{P}_i^{\mathrm{pix}} \cap \mathcal{G}_i^{\mathrm{pix}}\right|}{\left|\mathcal{P}_i^{\mathrm{pix}} \cup \mathcal{G}_i^{\mathrm{pix}}\right|}.$$

B.2 Setting 2: OCR-based Grounding (Instance-level IoU, Precision, Recall, F1)

In the OCR-based setting, the OCR system provides a set of candidate bounding boxes with text. The model selects a subset of candidate boxes as grounded evidence. For sample i, let $\mathcal{P}_i$ be the set of predicted (selected) OCR boxes and $\mathcal{G}_i$ be the set of ground-truth boxes. We treat boxes as discrete instances.

Instance-level IoU. We compute the IoU over *instances* by set overlap:

$$\mathrm{IoU}_{\mathrm{Instance}} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathcal{P}_i \cap \mathcal{G}_i|}{|\mathcal{P}_i \cup \mathcal{G}_i|}.$$

Precision and Recall. We report **micro-averaged** precision and recall by aggregating counts across the dataset:

$$\mathrm{Precision} = \frac{\sum_{i=1}^{N}|\mathcal{P}_i \cap \mathcal{G}_i|}{\sum_{i=1}^{N}|\mathcal{P}_i|},$$

$$\mathrm{Recall} = \frac{\sum_{i=1}^{N}|\mathcal{P}_i \cap \mathcal{G}_i|}{\sum_{i=1}^{N}|\mathcal{G}_i|}.$$

F1 score. The F1 score is computed from precision and recall:

$$\mathrm{F1} = \frac{2\,\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}.$$

C Visualization Examples of Doc-AGround for Document Grounding

In Figs. 5 to 8, we show the visualization examples of Doc-AGround on ChartQA, DocVQA, InfographicVQA and TRINS. The blue boxes in each figure denote the ground-truth of the document grounding. And the area in green is the positive patch-wise attention grounding result of Doc-AGround.

D Details of the <g> Token Implementation

The <g> token functions similarly to any existing special token in our design, as long as it is placed after all visual and text tokens in the user instruction to serve as the grounding token. During grounding training, the token after inputs is easily learned as a context aggregator for all instruction tokens and as an indicator of positive image patches via the attention values between the token and visual patches, enabled by causal attention in the transformer decoder.

We choose to introduce a new <g> token, rather than reuse an existing special token, such as the EOS token, in consideration of future extension of Doc-AGround to multi-target document grounding. In this task, each grounding result requires a corresponding text description. New <g> token allows inserting multiple grounding tokens (e.g., <g>, text1, <g>, text2, ..., <g>, textN) without affecting the functionality of existing special tokens or altering decoding behavior.

In Section 3.2, we abbreviate the start and end tokens for the grounding token: abbreviate [<g_START>, <g>, <g_END>] as <g>. In addition, we try to insert multiple <g> as [<g_0>, <g_1>, . . . , <g_N>] to improve the expressiveness by expanding the number of patch-wise prediction vectors for each attention head. But the naive expansion of <g> without any disentanglement objective for multi-area grounding barely improves the document grounding performance compared with the original single <g> implementation.

E Extra Visualizations of $\alpha_{l,h}$

Here, we show the extra visualizations of multi-head weight $\alpha_{l,h}$ on the four sets of TRIG benchmarks: ChartQA, DocVQA, InfographicVQA and TRINS in Fig. 4.

F Extra Visualizations of Aggregated <g>-token Attention Map and Error Analysis

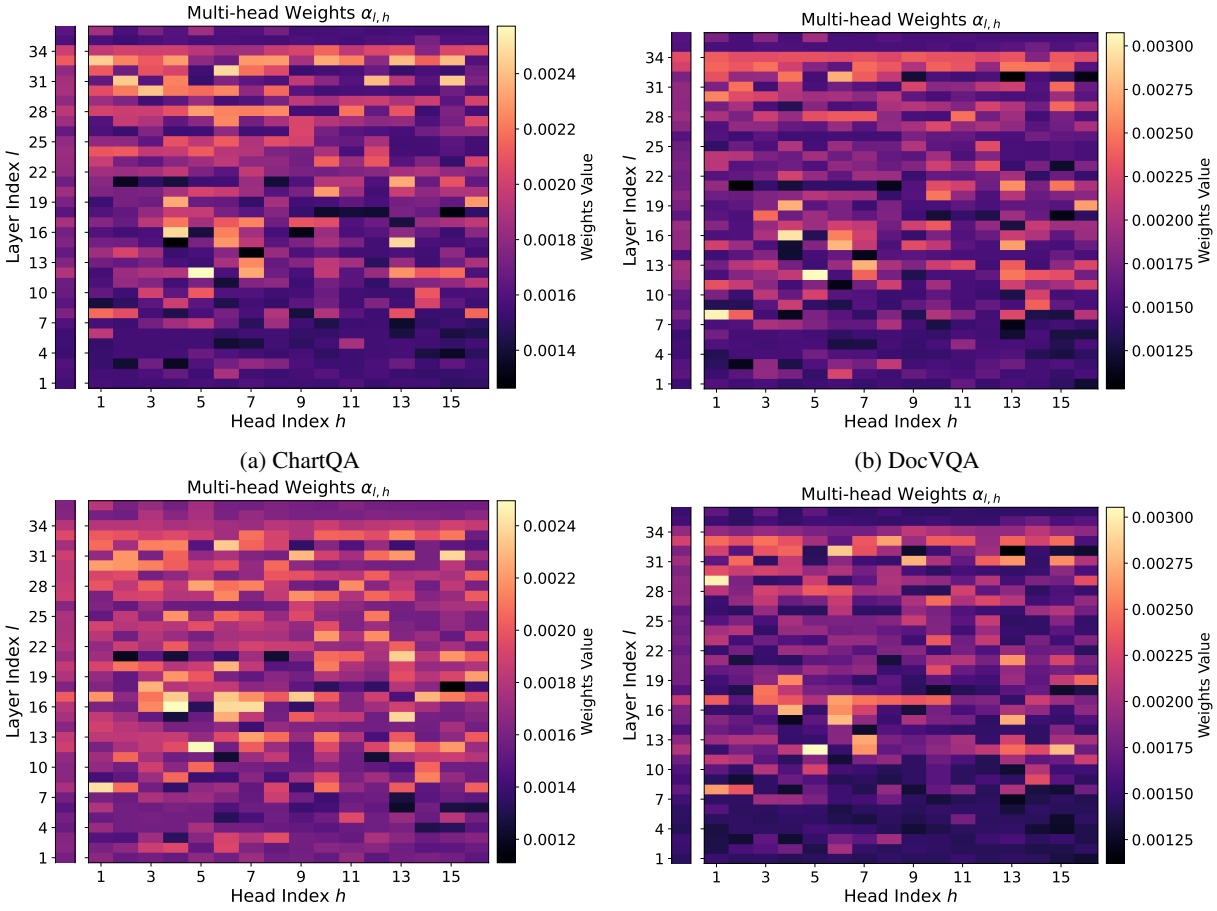
Here, we show the extra visualizations of aggregated <g>-token attention map on the TRIG benchmark in Fig. 9. The merged attention maps of <g> token in Fig. 9a,Fig. 9c and Fig. 9d successfully identify the correct grounding areas in the patch-wise manner.

"Error" Case Analysis. The visualization example of Fig. 9b seems to include extra error areas on the right side of the figure. Besides the three country names mentioned in the question, Doc-AGround also highlights the corresponding numbers. These additional regions are actually additional explainable evidence beyond incomplete

human annotations in TRIG, which are crucial for identifying the query target "second highest number", which are missing in the ground truth.

G Use Of AI Assistants

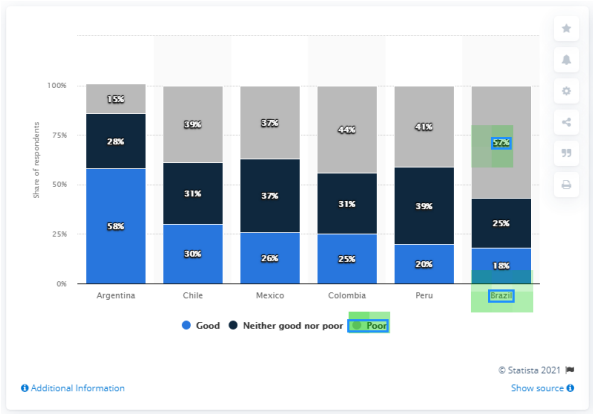
We use AI Assistants, e.g., ChatGPT, to assist with fixing the typos in writing and debugging for the code.



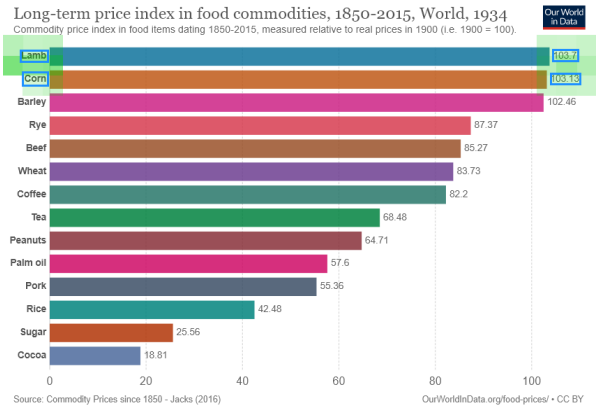
(a) ChartQA (b) DocVQA

(c) InfographicVQA (d) TRINS

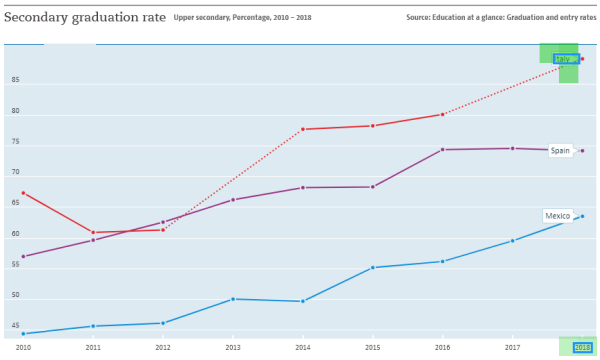
Figure 4: Visualization of $\alpha_{l,h}$ in Doc-AGround on each set of TRIG benchmark.



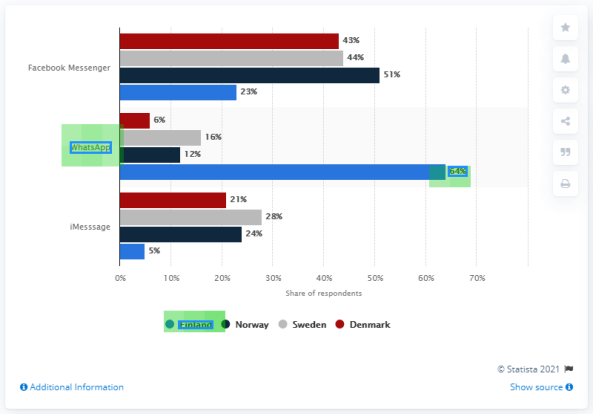
(a) **Question:** In brazil which share is high?
Answer: Poor



(b) **Question:** What is the difference in value between Lamb and Corn?
Answer: 0.57

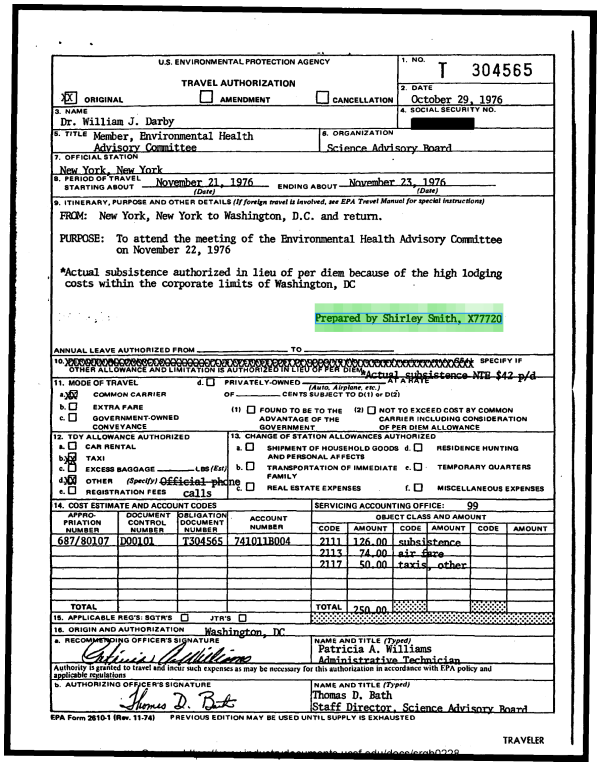


(c) **Question:** Which country has highest secondary graduation rate in 2018?
Answer: Italy



(d) **Question:** What instant messaging app has the most use in Finland?
Answer: WhatsApp

Figure 5: Visualization examples of Doc-AGround on ChartQA.



U.S. ENVIRONMENTAL PROTECTION AGENCY

TRAVEL AUTHORIZATION

1. NO. T 304565

2. DATE October 29, 1976

[X] ORIGINAL [ ] AMENDMENT [ ] CANCELLATION

3. NAME Dr. William J. Darby

4. SOCIAL SECURITY NO. [redacted]

5. TITLE Member, Environmental Health Advisory Committee

6. ORGANIZATION Science Advisory Board

7. OFFICIAL STATION New York, New York

8. PERIOD OF TRAVEL STARTING ABOUT November 21, 1976 (Date) ENDING ABOUT November 23, 1976 (Date)

9. ITINERARY, PURPOSE AND OTHER DETAILS (If foreign travel is involved, see EPA Travel Manual for special instructions)

FROM: New York, New York to Washington, D.C. and return.

PURPOSE: To attend the meeting of the Environmental Health Advisory Committee on November 22, 1976

*Actual subsistence authorized in lieu of per diem because of the high lodging costs within the corporate limits of Washington, DC

Prepared by Shirley Smith, X77720

ANNUAL LEAVE AUTHORIZED FROM _____ TO _____

10. [redacted] SPECIFY IF OTHER ALLOWANCE AND LIMITATION IS AUTHORIZED IN LIEU OF PER DIEM *Actual subsistence NTE $42 p/d

11. MODE OF TRAVEL

a. [X] COMMON CARRIER d. [ ] PRIVATELY-OWNED (Auto, Airplane, etc.) AT A RATE OF _____ CENTS SUBJECT TO D(1) or D(2)

b. [ ] EXTRA FARE

c. [ ] GOVERNMENT-OWNED CONVEYANCE (1) [ ] FOUND TO BE TO THE ADVANTAGE OF THE GOVERNMENT (2) [ ] NOT TO EXCEED COST BY COMMON CARRIER INCLUDING CONSIDERATION OF PER DIEM ALLOWANCE

12. TDY ALLOWANCE AUTHORIZED

a. [ ] CAR RENTAL

b. [X] TAXI

c. [ ] EXCESS BAGGAGE _____ LBS (Est)

d. [X] OTHER (Specify) Official phone calls

e. [ ] REGISTRATION FEES

13. CHANGE OF STATION ALLOWANCES AUTHORIZED

a. [ ] SHIPMENT OF HOUSEHOLD GOODS AND PERSONAL AFFECTS d. [ ] RESIDENCE HUNTING

b. [ ] TRANSPORTATION OF IMMEDIATE FAMILY c. [ ] TEMPORARY QUARTERS

c. [ ] REAL ESTATE EXPENSES f. [ ] MISCELLANEOUS EXPENSES

14. COST ESTIMATE AND ACCOUNT CODES

SERVICING ACCOUNTING OFFICE: 99

APPROPRIATION NUMBER	DOCUMENT CONTROL NUMBER	OBLIGATION DOCUMENT NUMBER	ACCOUNT NUMBER	OBJECT CLASS AND AMOUNT CODE	AMOUNT	CODE	AMOUNT	CODE	AMOUNT
687/80107	D00101	T304565	741011B004	2111	126.00	subsistence			
				2113	74.00	air fare			
				2117	50.00	taxis, other			
TOTAL				TOTAL	250.00				

15. APPLICABLE REG'S: SGTR'S [ ] JTR'S [ ]

16. ORIGIN AND AUTHORIZATION Washington, DC

a. RECOMMENDING OFFICER'S SIGNATURE [signature] NAME AND TITLE (Typed) Patricia A. Williams Administrative Technician

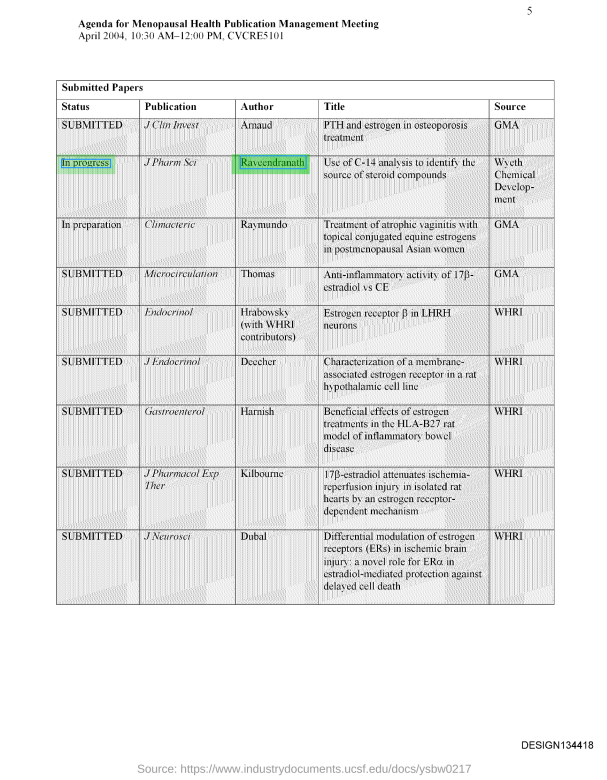
Authority is granted to travel and incur such expenses as may be necessary for this authorization in accordance with EPA policy and applicable regulations

b. AUTHORIZING OFFICER'S SIGNATURE Thomas D. Bath NAME AND TITLE (Typed) Thomas D. Bath Staff Director, Science Advisory Board

EPA Form 2610-1 (Rev. 11-74) PREVIOUS EDITION MAY BE USED UNTIL SUPPLY IS EXHAUSTED

TRAVELER

(a) **Question:** Who prepared the travel authorisation form?
Answer: hirley smith, X77720



5

Agenda for Menopausal Health Publication Management Meeting
April 2004, 10:30 AM–12:00 PM, CVCRE5101

Submitted Papers

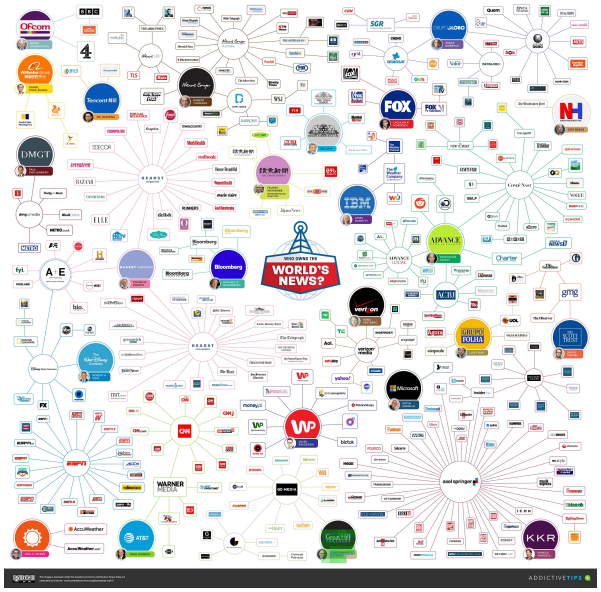
Status	Publication	Author	Title	Source
SUBMITTED	*J Clin Invest*	Arnaud	PTH and estrogen in osteoporosis treatment	GMA
In progress	*J Pharm Sci*	Raveendranath	Use of C-14 analysis to identify the source of steroid compounds	Wyeth Chemical Develop-ment
In preparation	*Climacteric*	Raymundo	Treatment of atrophic vaginitis with topical conjugated equine estrogens in postmenopausal Asian women	GMA
SUBMITTED	*Microcirculation*	Thomas	Anti-inflammatory activity of 17β-estradiol vs CE	GMA
SUBMITTED	*Endocrinol*	Hrabowsky (with WHRI contributors)	Estrogen receptor β in LHRH neurons	WHRI
SUBMITTED	*J Endocrinol*	Decther	Characterization of a membrane-associated estrogen receptor in a rat hypothalamic cell line	WHRI
SUBMITTED	*Gastroenterol*	Harnish	Beneficial effects of estrogen treatments in the HLA-B27 rat model of inflammatory bowel disease	WHRI
SUBMITTED	*J Pharmacol Exp Ther*	Kilbourne	17β-estradiol attenuates ischemia-reperfusion injury in isolated rat hearts by an estrogen receptor-dependent mechanism	WHRI
SUBMITTED	*J Neurosci*	Dubal	Differential modulation of estrogen receptors (ERs) in ischemic brain injury: a novel role for ERα in estradiol-mediated protection against delayed cell death	WHRI

DESIGN134418

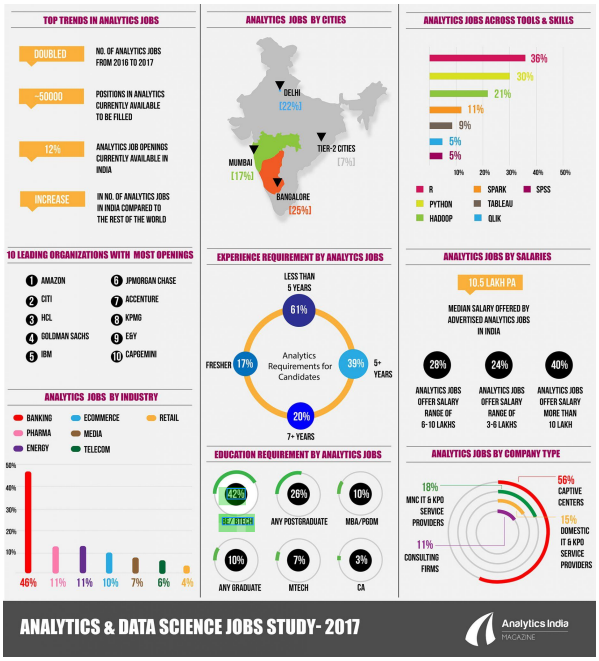
Source: https://www.industrydocuments.ucsf.edu/docs/ysbw0217

(b) **Question:** What is the name of the author whose submitted paper status is in progress?
Answer: Raveendranath

Figure 6: Visualization examples of Doc-AGround on DocVQA.

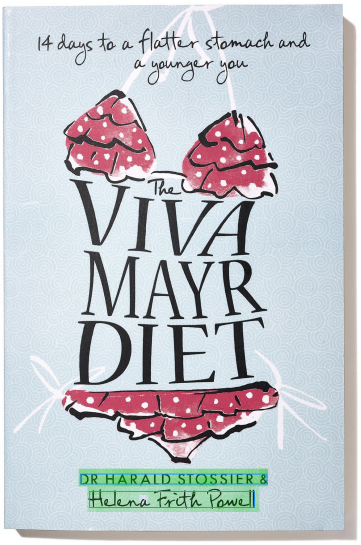


(a) **Question:** who is the founder of Great Hill Partners?
Answer: christopher s. gaffney

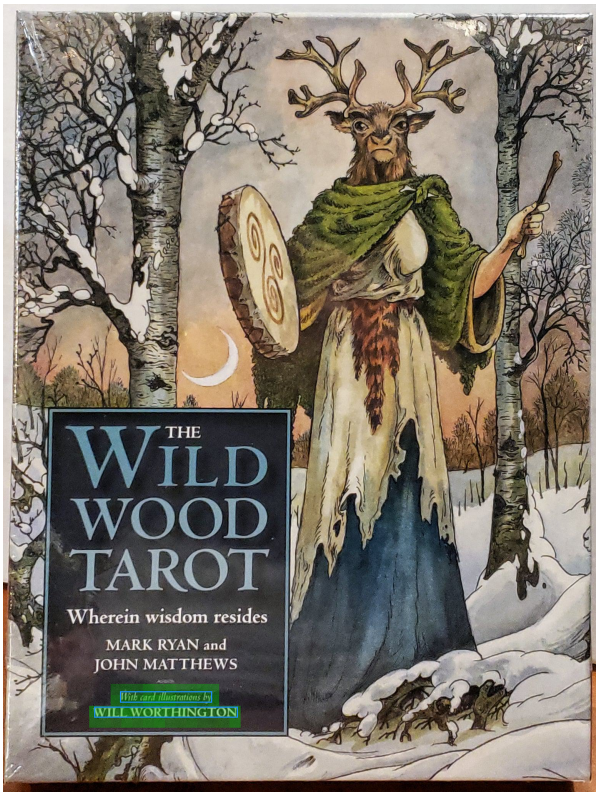


(b) **Question:** What percent of analytics jobs in India require BE/B-tech degree according to the 2017 study?
Answer: 42%

Figure 7: Visualization examples of Doc-AGround on InfographicVQA.

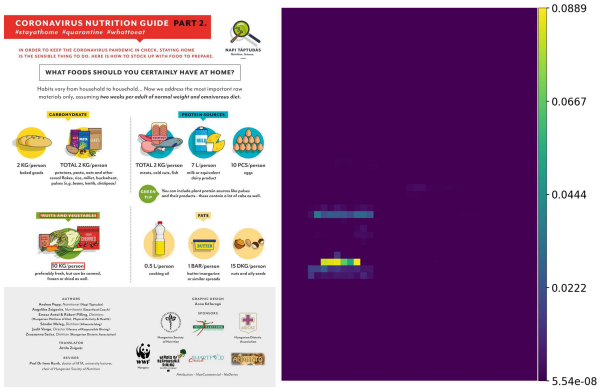


(a) **Question:** Who are the authors of the book T̈he VIVA MAYR DIET?̈
Answer: The book T̈he VIVA MAYR DIETïs written by Dr. Harald Stossier and Helena Frith Powell.

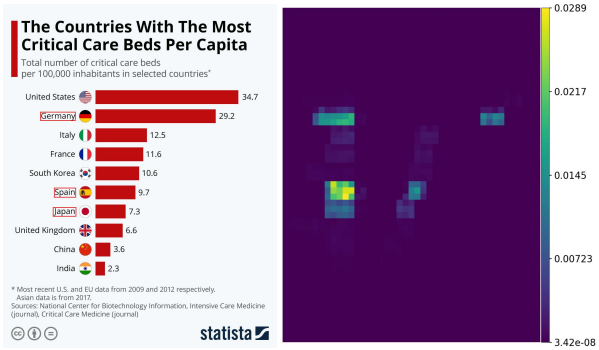


(b) **Question:** Who is the illustrator of the card illustrations in T̈he Wildwood Tarot: Wherein Wisdom Residesäs mentioned in the image?
Answer: The illustrator of the card illustrations is Will Worthington.

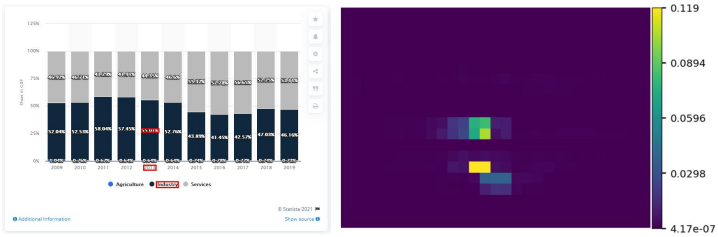
Figure 8: Visualization examples of Doc-AGround on TRINS.



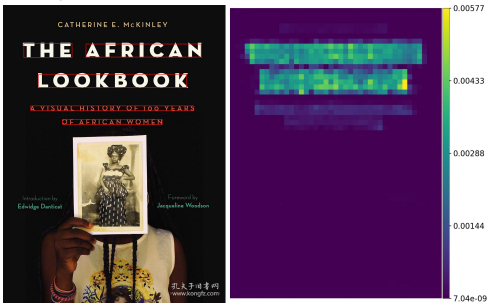
(a) **Question:** How many kilograms of fruits and vegetables need to be consumed per person, 2, 10, or 15?
Answer: 10



(b) **Question:** Which country has the second highest number of total critical care beds, Spain, Germany, or Japan?
Answer: germany



(c) **Question:** What was the industry share in GDP in 2013?
Answer: 55.01



(d) **Question:** What is the title of the book?
Answer: The title of the book is "The African Lookbook: A Visual History of 100 Years of African Women".

Figure 9: Visualization example of aggregated <g>-token attention map for visual text grounding on TRIG benchmark. Ground truth bounding boxes are shown in red.