INDOTABVQA: A Benchmark for Cross-Lingual Table Understanding in Bahasa Indonesia Documents

Somraj Gautam[1], Anathapindika Dravichi[2], Gaurav Harit[1]
[1]IIT Jodhpur, [2]Punjabi University
gautam.8@iitj.ac.in, dravichijan@gmail.com, gharit@iitj.ac.in
https://huggingface.co/datasets/NusaBharat/INDOTABVQA

Abstract

We introduce INDOTABVQA, a benchmark for evaluating cross-lingual Table Visual Question Answering (VQA) on real-world document images in Bahasa Indonesia. The dataset comprises 1,593 document images across three visual styles (bordered, borderless, and colorful) with one or more than one tables, and 1,593 question-answer sets in four languages: Bahasa Indonesia, English, Hindi, and Arabic. This enables evaluation of Vision-Language Models (VLMs) in both monolingual (Bahasa documents with Bahasa questions) and cross-lingual settings (Bahasa documents with questions in other languages). We benchmark leading open-source VLMs (Qwen2.5-VL, Gemma-3, LLaMA-3.2) and GPT-4o and reveal substantial performance gaps, particularly on structurally complex tables and in low-resource languages. Fine-tuning a compact 3B model and a LoRA-finetuned 7B model on our dataset yields 11.6% and 17.8% improvements in accuracy. Providing explicit table region coordinates as additional input further improves performance by 4-7%, demonstrating the value of Spatial priors for table-based reasoning. Our findings underscore the importance of language-diverse, domain-specific datasets and demonstrate that targeted fine-tuning can significantly enhance VLM performance on specialized document understanding tasks. INDOTABVQA provides a valuable resource for advancing research in cross-lingual, structure-aware document understanding, especially in underrepresented regions of the world. The dataset is publicly available via Hugging Face at: https://huggingface.co/datasets/NusaBharat/INDOTABVQA.

1 Introduction

Vision-Language Models (VLMs) have demonstrated strong performance on text-centric visual understanding tasks, as shown on benchmarks such as TextVQA (Singh et al., 2019), ST-VQA (Xia et al., 2023), DocVQA (Mathew et al., 2021), and OCRBench (Liu et al., 2024). Recent table-focused datasets such as TableVQA-Bench (Kim et al., 2024), TabComp (Gautam et al., 2025a), and ComTQA (Zhao et al., 2024) further assess numerical reasoning and structure-aware comprehension. However, these benchmarks share a critical limitation: they are predominantly monolingual and English-centric, providing limited insight into VLM performance on low-resource languages or cross-lingual generalization. Documents in languages like Bahasa Indonesia, Hindi, and Ara-

Figure 1: INDOTABVQA presents document images in Bahasa Indonesia, and semantically aligned QA pairs in four languages, enabling cross-lingual evaluation of VLMs.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21969–21981

bic represent billions of users globally, yet VLMs trained primarily on English data may fail to process these documents reliably. For table-based VQA specifically, models must handle both linguistic variation and structural complexity, a challenging combination that remains underexplored.

The Core Problem: Existing VQA benchmarks do not adequately test whether VLMs can (1) understand tables in low-resource languages, or (2) answer questions about these tables when queries are posed in different languages. This gap limits our understanding of true multilingual capability and hinders the development of globally applicable document AI systems.

This paper introduces INDOTABVQA, a benchmark designed to evaluate the cross-lingual and structure-aware capabilities of VLMs in the context of real-world document tables. Our benchmark comprises document images containing tables in Bahasa Indonesia, a language spoken by over 200 million people but underrepresented in vision-language research, paired with question-answer (QA) annotations in Bahasa Indonesia, English, Hindi, and Arabic, as shown in Fig. 1. Detailed statistics of our benchmark are presented in section 2.5 and table 1.

Our work provides three main contributions:

- **A novel cross-lingual benchmark** featuring real-world documents in an underrepresented language (Bahasa Indonesia) with parallel annotations in four languages, enabling systematic evaluation of cross-lingual visual reasoning.
- **Comprehensive baseline evaluation** of current VLMs, revealing specific failure modes in structure-aware reasoning and language transfer that inform future model development.
- **Analysis of Spatial priors and fine-tuning** showing that explicit table localization and domain adaptation are effective strategies for improving VLM performance on specialized document tasks.

INDOTABVQA addresses a critical gap in multilingual document AI and provides a testbed for developing more inclusive and robust vision-language systems. The dataset and evaluation code will be made publicly available upon acceptance.

2 INDOTABVQA Dataset

This section describes the construction of INDOTABVQA in detail, covering the dataset scope and design, data collection, the diversity of table types, the annotation protocol, statistics, and benchmark configuration.

2.1 Dataset Scope and Design

INDOTABVQA enables evaluation in two settings:

- **Monolingual setting:** Both documents and QA pairs are in Bahasa Indonesia, testing the model's ability to understand low-resource language content.
- **Cross-lingual setting:** Documents remain in Bahasa Indonesia while questions are posed in English, Hindi, or Arabic. This probes whether models can align visual content in one language with semantically equivalent questions in another, assessing true cross-lingual transfer rather than memorized language patterns.

This design isolates two distinct challenges: (1) visual-linguistic understanding of low-resource document content, and (2) cross-lingual alignment between visual and textual modalities.

2.2 Data Collection and Sources

We sourced table images from real-world Indonesian documents across government reports (statistical summaries, budget allocations), educational records (enrollment data, performance metrics), business documents (invoices, financial statements), public health data (demographic statistics, service records). A significant portion of our data derives from the Institutional Repository of the Ministry of Primary and Secondary Education of Indonesia[1]. We retrieved documents from the official portal and manually selected those containing well-formed tables suitable for VQA.

2.3 Visual Diversity: Table Types

To reflect real-world document variation, we categorize tables into three types based on visual presentation:

- **Bordered Tables** (500 images): Traditional tables with explicit cell borders, commonly found in official forms and reports.

[1] https://repositori.kemendikdasmen.go.id

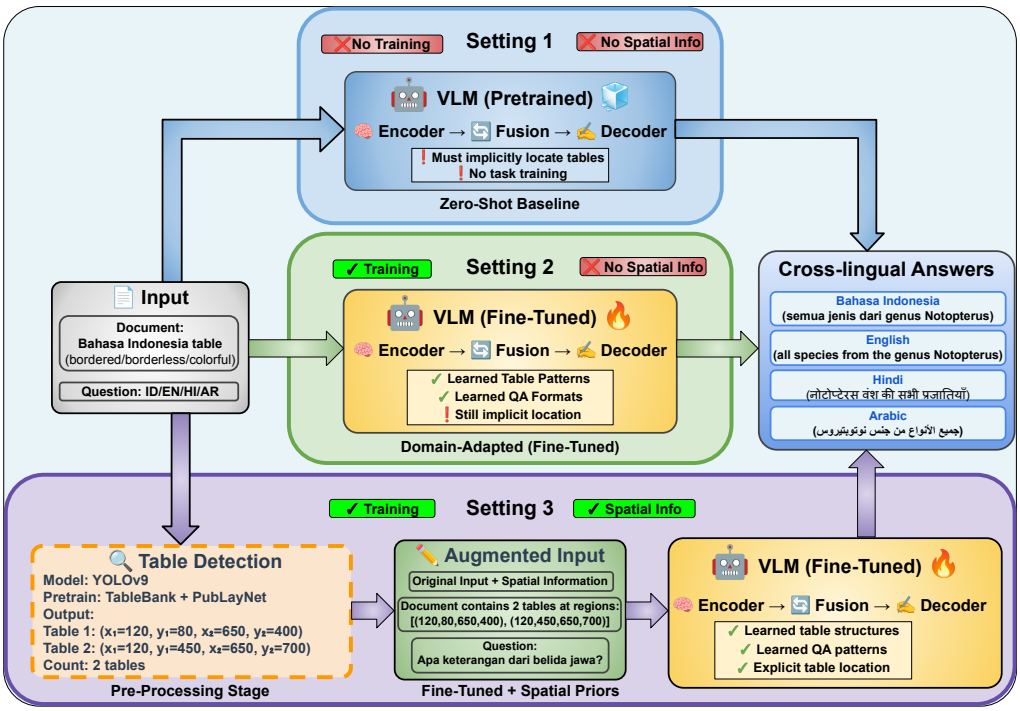

Figure 2: Architecture comparison with left-to-right pipeline flow across three evaluation settings. Each row represents a complete evaluation pipeline from input to output. **Setting 1:** Uses a pretrained VLM without fine-tuning (zero-shot). **Setting 2:** Fine-tunes the model on table QA data but without spatial information. **Setting 3:** Introduces spatial priors through table detection, enabling the model to use table locations during reasoning.

- **Borderless Tables** (602 images): Tables without explicit cell lines, requiring inference of structure from whitespace, alignment, and text positioning.
- **Colorful Tables** (491 images): Tables using background colors, cell shading, or highlighted headers for emphasis or grouping.

This taxonomy is not mutually exclusive (some tables have both borderless and colors as table present in Fig. 1), but we assign each image to its primary category for analysis purposes and give the color category a higher priority.

2.4 Annotation Protocol

Each table instance is paired with one question–answer (QA) item, authored in Bahasa Indonesia, following a controlled template designed to cover lookup, aggregation, comparison, and structural reasoning. Annotators were instructed to write unambiguous, table-grounded, and answer-contained questions. We then translated each Bahasa QA into English, Hindi, and Arabic using automatic translation, followed by human validation by native speakers. Validators corrected lexical errors, normalized number formats, ensured that entity references remained faithful to the table, and flagged ambiguous or culturally mismatched translations. Each QA underwent a two-stage quality check: (1) internal consistency (answer must exist exactly in the table region) and (2) cross-lingual equivalence (the four versions must express the same intent). Items failing either check were revised or removed. Table 7 summarises the key statistics. Extended guidelines and annotation examples appear in Appendix A. Figure 3 illustrates language coverage by country, highlighting our focus on evaluating VLMs in linguistically diverse and underrepresented regions, such as **Southeast Asia, the Middle East, South Asia, and other English-dominated countries**.

2.5 Dataset Statistics and Properties

Table 1 summarizes the key characteristics of the dataset. The visual content in all images is exclusively in Bahasa Indonesia, ensuring linguistic consistency across table elements. However, the question–answer (QA) annotations are multilingual, available in Bahasa Indonesia, English, Hindi, and Arabic, enabling cross-lingual evaluation and analysis. Each table instance is accompanied by detailed annotation metadata, including table-level

Property	IndoTabVQA
# Document Images	1,593
# Total Tables	1,910
Avg. Tables per Image	1.20
# QA Pairs	6,372 (Bahasa+English+Hindi+Arabic)
QA per Language	1,593 per language
Languages	Bahasa Indonesia, English, Hindi, Arabic
QA Annotation Style	Human-written + Translated
Table Layouts	Bordered, Borderless, Colorful table
Domains	Government, Finance, Education, Health
Image Format	JPEG (OCR-compatible resolution)
Bounding Box Annotations	Table-level bounding boxes
Cross-lingual Setting	Doc in Bahasa, QA in other languages

Table 1: IndoTabVQA dataset properties covering multilingual QA, layout styles, and domain diversity.

bounding boxes to precisely locate tables within document images and table type tags covering three distinct categories that capture structural or functional variations among tables.

2.6 Benchmark Configuration

We split the dataset into Test/Training/Validation set: 1043/500/50 samples.

We intentionally maintain a large test set to enable robust evaluation across diverse table styles and domains. Also, a small training set proves that fine-tuning with a small dataset size can improve the model's capability effectively.

3 Evaluation Methodology

3.1 Task Formulation

We formulate the task as image-grounded visual question answering: Given a document image I containing one or more tables and a natural language question Q, in language $L \in \{\text{Bahasa Indonesia}, \text{English}, \text{Hindi}, \text{Arabic}\}$, the model must generate or select the correct answer A in the same language. Formally, the task can be described as:

$$A = \text{VLM}(I, Q)$$

3.2 Input Format

Each input instance consists of a table image I (in PNG or JPEG format) and a question Q in either Bahasa Indonesia, English, Hindi, or Arabic. The answer A is a short free-form text or numeric value. Question types span factual lookup (retrieving specific cell values), numerical comparison (identifying maximum, minimum, or ranking items), aggregation (sum, count, or computing over multiple cells), and table-structure-related queries about table organization or headers.

3.3 Evaluation Settings

We evaluate models under three settings shown in Fig. 2:

- **Zero-Shot Evaluation:** Models are tested directly on IndoTabVQA without any task-specific training. This measures out-of-the-box capability for cross-lingual table understanding.
- **Fine-Tuned Evaluation:** Model is trained on the IndoTabVQA training set (500 images) and evaluated on the test set (1,043 images).
- **Fine-Tuned + Spatial Priors:** In this, we add an explicit table detection pre-processing stage (orange block) using YOLOv9 to locate table regions. These coordinates are then incorporated into an augmented prompt before VLM processing.

3.4 Table Localization as Additional Input

Motivation: A key challenge in document VQA is that tables may occupy only a small region of the full image, and documents may contain multiple tables with varying layouts and positions. Real-world document processing systems typically address this through multi-stage pipelines that first detect document regions (tables, figures, text blocks) before applying specialized models to each region. By providing explicit table bounding box coordinates to VLMs, we mirror this practical workflow and potentially help models focus their attention on relevant content rather than searching across the entire image. This approach also allows us to isolate the impact of spatial localization from other factors affecting model performance, providing insights into whether structural ambiguity, particularly in borderless tables, is a primary bottleneck for accurate table understanding.

3.5 Implementation of setting 3

Our approach consists of two stages:

- **Stage 1: Table Detection:** A separate, off-the-shelf object detection model (YOLOv9 (Wang et al., 2024a) pretrained on TableBank (Li et al., 2020) and PubLayNet (Zhong et al., 2019)) to identify table regions in document images. The detector outputs: 1) Bounding box coordinates: $[(x_1, y_1, x_2, y_2), ...]$ for each detected table. 2) Number of tables detected: N

- **Stage 2: Augmented Input:** VLM receives:1) Original input + Table bounding boxes. 2) Number of tables.

Example prompt augmentation:

Augmented Input

Document Image I.
Document contains 2 tables at regions: [(120, 80, 650, 400), (120, 450, 650, 700)].
Question: [question text]
Answer:

3.6 Evaluation Metrics

To evaluate model performance across diverse settings and languages, we employ both exact and semantic answer matching strategies.

3.6.1 In-Match Accuracy (Relaxed Matching)

We use a relaxed matching criterion where a prediction is correct if the normalized ground truth answer appears as a substring within the predicted answer.

Normalization involves converting text to lowercase, removing punctuation, collapsing whitespace, and handling number formatting variations. This relaxed matching accounts for VLMs that often generate answers with additional context (e.g., 'if the ground truth is '5 tables', a prediction of 'There are 5 tables in the document') would be considered correct. In-Match captures correct answers embedded in longer responses.

Formula:

$$\text{In-Match}(A_p, A_g) = \begin{cases} 1, & \text{Norm}(A_g) \subseteq \text{Norm}(A_p), \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

3.6.2 Semantic Textual Similarity (STS)

To better assess how well our model captures the true meaning of an answer, we go beyond simple word-for-word comparisons. We use **Semantic Textual Similarity (STS)** to measure the degree of meaning alignment between predicted answers A_p and ground truth answers A_g. STS is computed as the cosine similarity between their dense vector representations:

$$\text{STS}(A_p, A_g) = \frac{\phi(A_p) \cdot \phi(A_g)}{\|\phi(A_p)\| \cdot \|\phi(A_g)\|} \quad (2)$$

where $\phi(\cdot)$ denotes a sentence-level semantic encoder. To compute Semantic Textual Similarity (STS), we use the `paraphrase-multilingual-MiniLM-L12-v2` model from the **Sentence Transformers** library (Reimers and Gurevych, 2019), which produces language-agnostic sentence embeddings across 50+ languages. The similarity score lies in $[0, 1]$, with higher values indicating greater semantic alignment.

Breakdown by QA Type. Beyond overall metrics, to understand where models succeed and fail, we report fine-grained accuracy across all four languages, table types mentioned in section 2.3, and evaluation Settings mentioned in section 3.3

3.7 Baseline Models

We evaluate a diverse set of VLMs spanning different scales and architectures, including Open-Source Models such as: **Qwen2.5-VL [3B]** (Wang et al., 2024b): Compact VLM with strong multilingual capability. **Qwen2.5-VL [7B]** (Bai et al., 2025): Larger variant with enhanced reasoning. **Gemma-3 [12B]** (Team et al., 2025): Google's model with broad language coverage. **LLaMA-3.2 [11B]** (Grattafiori et al., 2024): Meta's vision-enabled language model, and a Closed-Source Model such as GPT-4o (OpenAI, 2024), which is a state-of-the-art proprietary VLM with strong multilingual performance.

We also evaluate Donut (Document Understanding Transformer (Kim et al., 2022)), an OCR-free document understanding model that directly maps document images to structured outputs using an encoder-decoder architecture. As it lacks multilingual pretraining and cross-lingual transfer capabilities, we expect it to serve as a lower-bound baseline, particularly in the Hindi and Arabic settings.

3.8 Fine-Tuning Configuration

Our fine-tuning strategy follows a full Instruction Fine-Tuning approach for Qwen2.5-VL [3B] and parameter-efficient finetuning (LoRA) for Qwen2.5-VL [7B]. The model was trained separately on each language variant of the dataset to isolate language-specific learning patterns. A detailed training setup is present in Appendix A.2

4 Results and Analysis

We present evaluation results across three dimensions: (1) overall performance by language, (2)

Model [#params]	In-Match accuracy(%)					STS Accuracy(%)				
	ID	EN	HI	AR	Δ	ID	EN	HI	AR	Δ
Open-source										
Donut	10.5	5.48	4.74	4.39	6.20	15.52	9.10	5.17	6.03	8.96
Qwen2.5VL [3B]	37.8	28.7	4.1	16.4	21.9	29.0	44.9	4.4	27.5	26.5
Gemma3 [12B]	40.9	27.4	19.5	17.4	26.1	41.4	31.0	27.3	26.5	31.6
Qwen2.5VL [7B]	54.8	36.2	17.3	23.0	32.9	36.5	58.1	16.1	34.3	36.3
Llama-3.2 [11B]	57.4	30.8	15.5	19.4	30.7	54.2	36.1	15.7	19.5	31.4
Closed-source										
GPT-4o	72.2	44.6	26.0	21.4	41.1	71.1	60.6	38.8	38.4	52.2
Finetuned + Spatial Priors (SP)										
GPT-4o+SP	72.6	52.7	27.2	25.5	44.6	73.4	62.2	39.1	40.0	53.6
IndoTabVQA [3B]	66.4	46.1	22.1	25.8	39.7	71.4	49.3	27.3	38.0	46.7
IndoTabVQA [7B]	71.9	51.6	26.2	28.1	44.5	77.6	64.5	31.4	46.4	54.9
IndoTabVQA [3B]+SP	73.1	54.8	27.2	31.1	46.6	75.2	61.2	36.0	40.1	53.1
IndoTabVQA [7B]+SP	**78.3**	**58.4**	**29.4**	**32.8**	**48.5**	**82.1**	**66.1**	**36.7**	**48.6**	**58.3**

Table 2: Evaluation of various VLMs on In-Match and STS Accuracy across four languages here ID is IndoTabVQA-id, EN is IndoTabVQA-en, HI is IndoTabVQA-hi, AR is IndoTabVQA-ar, and SP is Spatial Prior, Δ is average accuracy.

breakdown by table visual style, and (3) fine-grained analysis by question type. Our analysis focuses on understanding where and why models struggle, rather than simply ranking performance. Results are reported using two complementary metrics: In-Match accuracy, which measures relaxed answer inclusion, and STS accuracy, which captures semantic similarity using sentence-level embeddings. Our analysis spans both language-wise performance (Table 2) and table-type-specific behavior across languages (Table 3).

4.1 Overall Performance Across Languages

Table 2 presents In-Match and STS accuracy for all models across four languages. Several patterns emerge:

4.1.1 Performance Ranking by Model Scale (Zero-Shot)

Zero-shot performance among open-source VLMs generally increases with model scale. **Qwen2.5-VL-3B** attains 21.9% average accuracy, while **Qwen2.5-VL-7B** improves to 32.9%. Larger models such as **LLaMA-3.2-11B** and **Gemma-3-12B** achieve intermediate performance (26–31%).

However, scale alone is insufficient: **Qwen2.5-VL-7B** outperforms the larger **Gemma-3-12B**, underscoring the importance of architecture and pretraining. **GPT-4o** delivers the best zero-shot results (41.1% In-Match, 52.2% STS), reflecting the benefits of large-scale, diverse training.

4.2 The Cross-Lingual Performance Gap

Performance drops substantially in cross-lingual settings compared to monolingual (Bahasa):

4.2.1 Zero-shot degradation from ID to other languages:

GPT-4o: 72.2% → 44.6% (EN), 26.0% (HI), 21.4% (AR), Qwen2.5-VL [7B]: 54.8% → 36.2% (EN), 17.3% (HI), 23.0% (AR)

This 30-50 percentage point drop reveals a critical limitation: models struggle to align visual content in one language with questions in another. This led to two research questions (RQ) mentioned below:

RQ1: Why is Hindi particularly difficult?

Hindi shows the lowest accuracy across nearly all models (4-27.2%). Possible explanations can **Script unfamiliarity:** Devanagari script is less common in VLM pretraining, and most mainstream models use subword tokenization algorithms like SentencePiece or BPE. When applied to Devanagari, these tokenizers often fail to identify meaningful morphological units, instead splitting words into a long sequence of less meaningful, sometimes single-character, tokens. This sub-optimal segmentation has two detrimental effects: first, it creates much longer input sequences for the model, increasing computational load and making it harder to capture long-range dependencies; second, and more importantly, it fails to provide the model with consistent, semantically meaningful representations for Hindi words and concepts, thereby hindering learning and generalization (Kanjirangat et al., 2025).

Similarly, the challenges with Arabic extend beyond simple script differences. The Arabic language is a right-to-left (RTL) script, which can confound models that implicitly assume a left-to-right flow of information, especially for questions

Model [#params]	In-Match accuracy(%)			STS Accuracy(%)		
	Bordered	Borderless	Colorful	Bordered	Borderless	Colorful
Bahasa Indonesia						
Donut	11.71	10.23	9.40	21.02	17.87	8.95
Qwen2.5VL-3B	32.73	44.25	36.36	24.62	34.27	27.9
Qwen2.5VL-7B	52.55	57.29	54.55	40.54	32.74	36.36
Gemma3-12B	48.05	34.78	39.81	48.35	32.74	43.26
Llama-3.2-11B	57.36	52.43	62.38	52.20	50.50	60.10
GPT-4o	74.03	65.94	76.60	71.47	70.08	71.16
Finetuned + Spatial Priors (SP)						
GPT-4o+SP	75.23	65.94	76.6	73.38	73.45	73.34
IndoTabVQA-id [3B]	72.07	62.92	64.26	73.38	68.43	72.27
IndoTabVQA-id [7B]	72.07	69.82	74.61	79.58	73.15	81.50
IndoTabVQA-id [3B]+SP	80.78	66.75	71.79	81.38	71.59	72.73
IndoTabVQA-id [7B]+SP	80.25	73.15	81.50	87.65	76.47	83.39
English						
Donut	3.90	5.63	6.90	7.81	6.14	14.11
Qwen2.5VL-3B	20.72	33.76	31.66	44.74	45.27	44.51
Qwen2.5VL-7B	29.43	41.18	37.93	55.26	61.13	58
Gemma3-12B	27.03	23.53	31.66	32.43	26.60	33.86
Llama-3.2-11B	25.53	28.90	37.93	41.40	28.90	37.93
GPT-4o	42.34	41.18	50.16	63.96	53.96	63.95
Finetuned + Spatial Priors (SP)						
GPT-4o+SP	42.81	56.42	58.87	65.23	55.12	66.17
IndoTabVQA-en [3B]	37.84	54.73	45.77	50.15	47.31	50.47
IndoTabVQA-en [7B]	45.35	52.45	56.87	63.06	65.73	64.58
IndoTabVQA-en [3B]+SP	48.95	55.75	59.87	60.70	57.30	65.50
IndoTabVQA-en [7B]+SP	53.85	58.75	62.57	64.86	66.75	66.77
Hindi						
Donut	3.60	4.35	6.27	4.20	4.60	6.90
Qwen2.5VL-3B	3.90	4.60	3.45	2.70	6.90	3.50
Qwen2.5VL-7B	14.41	18.41	18.81	13.81	16.88	17.55
Gemma3-12B	16.50	17.40	24.50	26.43	23.53	32.29
Llama-3.2-11B	12.91	13.04	21.32	12.91	13.04	21.32
GPT-4o	20.92	26.80	30.35	35.44	39.62	40.22
Finetuned + Spatial Priors (SP)						
GPT-4o+SP	22.32	28.52	30.76	36.9	37.39	43.01
IndoTabVQA-hi [3B]	13.21	25.58	20.38	18.92	35.04	27.59
IndoTabVQA-hi [7B]	20.42	29.92	28.21	25.53	33.76	34.80
IndoTabVQA-hi [3B]+SP	14.11	28.90	21.94	31.02	36.88	40.1
IndoTabVQA-hi [7B]+SP	22.82	33.76	31.66	27.93	41.18	40.44
Arabic						
Donut	2.10	5.96	5.12	6.61	3.84	8.15
Qwen2.5VL-3B	12.01	18.93	17.87	24.30	34.30	23.80
Qwen2.5VL-7B	19.50	24.81	24.14	28.23	41.43	33.23
Gemma3-12B	18.30	14.60	19.40	26.40	24.80	28.20
Llama-3.2-11B	15.92	18.16	24.45	15.92	18.16	24.45
GPT-4o	18.92	21.48	23.82	35.44	39.62	40.22
Finetuned + Spatial Priors (SP)						
GPT-4o+SP	21.24	28.80	26.40	37.60	38.39	44.01
IndoTabVQA-ar [3B]	17.72	34.02	24.14	32.10	46.30	35.40
IndoTabVQA-ar [7B]	40.66	34.17	23.42	43.84	48.85	46.08
IndoTabVQA-ar [3B]+SP	21.32	39.64	32.29	35.44	45.78	39.18
IndoTabVQA-ar [7B]+SP	40.66	34.17	23.42	47.15	51.66	46.39

Table 3: Results of various VLMs on In-Match and STS Accuracy based on table types across four languages.

involving spatial relationships.

RQ2: Why does Bahasa perform best?

The monolingual setting removes the cross-lingual alignment challenge, as both the visual content and the question share the same language. Additionally, the fine-tuned model is directly exposed to Bahasa examples during training, giving it a distributional advantage. As shown in Table 2, language-specific In-Match gains after fine-tuning are: Bahasa Indonesia +28.6 points (highest), English +17.4, Hindi +18.0, and Arabic +9.4, demonstrating that even modest task-specific supervision over 500 training images yields meaningful improvements across all languages.

4.3 Effect of Spatial Priors (Bounding Boxes)

Adding table bounding box coordinates as additional input provides further gains:

Average improvement over fine-tuned model: Compact 3B Model: In-Match: +6.9% points (39.7% $\rightarrow$ 46.6%), STS: +6.4% points (46.7% $\rightarrow$ 53.1%). LoRA Finetuned 7B Model: In-Match: +4.0% points (44.5% $\rightarrow$ 48.5%), STS: +3.4% points (54.9% $\rightarrow$ 58.3%).

Average improvement on GPT-4o model: In-Match: +3.5% points (41.1% $\rightarrow$ 44.6%), STS:

+1.4% points (52.2% → 53.6%)

Notably, spatial priors benefit GPT-4o as well, boosting its In-Match accuracy from 41.1% to 44.6%, confirming that explicit table localization is useful regardless of model scale. Our fine-tuned 7B model with spatial priors achieves the best overall performance across both metrics (48.5% In-Match and 58.3% STS), outperforming GPT-4o+SP (44.6% and 53.6% respectively), suggesting that the combination of domain adaptation and spatial grounding is more effective than either alone.

4.4 Performance by Table Visual Style

Table 3 analyzes model performance across bordered, borderless, and colorful tables, revealing the strong influence of visual style on reasoning accuracy. Borderless tables pose the greatest challenge, as models must infer row–column relationships from whitespace and alignment, often leading to ambiguity. Accuracy improves notably with spatial priors (e.g., +3.8 points in Bahasa), showing the benefit of explicit localization. Colorful tables yield mixed results; GPT-4o performs better on them (76.6% vs. 74.0%), likely because color aids visual grouping and attention, though smaller models struggle due to limited robustness to color variation. Bordered tables provide the clearest structure and serve as a baseline (GPT-4o: 74.0%, LLaMA-3.2: 57.4%, Gemma3-12B: 48.1%, Qwen2.5-VL [7B]: 52.6%). Yet even here, performance below 75% indicates that accurate table reasoning remains a challenging task despite clear visual cues. As shown in Fig. 5, our model+SP produces correct results across all three types of tables in a cross-lingual setting.

To better understand model failures in our benchmark, we perform a detailed manual analysis of erroneous predictions in section A.1 of Appendix A. We categorize the errors into five types for English, Hindi, and Arabic, and into four types for Bahasa Indonesia (due to the absence of translation-related errors) as shown in Figure 4 and Table 8 in Appendix A.

4.5 Performance by Question Type

We further analyze model behavior across four question types (lookup, aggregation, comparison, and structural reasoning) as shown in Table 4. Consistent with earlier observations, aggregation and comparison questions achieve higher accuracy across models and languages, indicating that VLMs are relatively effective at coarse-grained reasoning

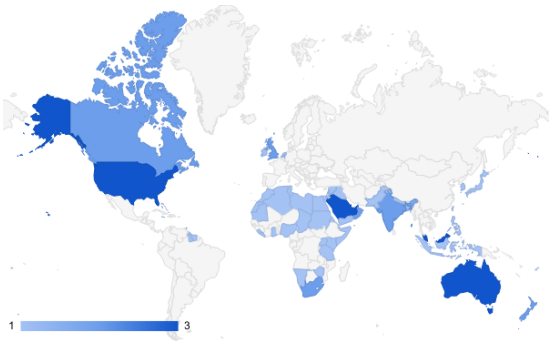

Figure 3: Global language coverage map for the IndoTabVQA benchmark. The shading intensity indicates the number of supported languages (1–3) spoken in each country. For example, Canada supports both English and Hindi. This visualization highlights the geographical and cultural reach of our cross-lingual benchmark.

where relevant values are localized or require limited structural interpretation. In contrast, lookup and structural questions remain more challenging. Lookup requires precise cell-level retrieval, while structural reasoning depends on understanding table organization, such as header alignment and row–column relationships. These challenges are further amplified in cross-lingual settings, where accurate alignment between the query language and table content is necessary. Hindi and Arabic exhibit the largest performance degradation, consistent with the cross-lingual trends discussed in Section 4.2. Fine-tuning improves performance across all question types, with the most notable gains in lookup and structural reasoning. Incorporating spatial priors provides additional improvements, particularly for lookup, by guiding the model toward relevant table regions. Overall, this analysis highlights that while current VLMs handle coarse reasoning well, fine-grained, structure-aware table understanding remains a key limitation.

5 Related Work

5.1 Table-Based Visual Question Answering

Table-Based Visual Question Answering (VQA) addresses the challenge of reasoning over tabular structures embedded in images. Benchmarks such as InfographicVQA (Mathew et al., 2022), DocVQA (Mathew et al., 2021), and ChartQA (Masry et al., 2022), TabFact (Chen et al., 2019), TAT-QA (Zhu et al., 2021), and PubTables-1M (Smock et al., 2022) emphasize reasoning over semi-structured and document tables. However,

Model [#params]	Agg.	Comp.	Look.	Str.
Indonesia				
Donut	13.13	0	4.89	12.33
Qwen2.5VL-3B	50.59	41.18	19.14	30.14
Qwen2.5VL-7B	64.58	76.47	27.56	57.53
Gemma3-12B	40.7	47.06	38.77	47.95
Llama-3.2-11B	66.23	100	43.25	53.42
GPT-4o	74.41	60.24	64.59	71.60
Finetuned + Spatial Priors (SP)				
GPT-4o + SP	75.48	64.71	78.89	72.60
IndoTabVQA-id [3B]	71.92	47.06	58.97	65.75
IndoTabVQA-id [7B]	76.62	82.35	65.55	71.23
IndoTabVQA-id [3B]+SP	74.88	70.59	71.58	71.23
IndoTabVQA-id [7B]+SP	80.39	70.59	78.54	82.19
English				
Donut	7.98	0	1.37	1.39
Qwen2.5VL-3B	40.96	41.18	11.85	12.50
Qwen2.5VL-7B	45.69	47.06	22.73	23.61
Gemma3-12B	32.39	23.53	20.43	20.83
Llama-3.2-11B	40.97	58.82	16.77	18.06
GPT-4o	54.72	41.18	35.97	37.50
Finetuned + Spatial Priors (SP)				
GPT-4o + SP	62.23	65.84	38.84	38.70
IndoTabVQA-en [3B]	58.26	41.18	26.83	31.94
IndoTabVQA-en [7B]	62.85	52.94	32.97	29.17
IndoTabVQA-en [3B]+SP	60.62	64.71	39.41	31.94
IndoTabVQA-en [7B]+SP	62.94	64.71	32.88	30.56

Model [#params]	Agg.	Comp.	Look.	Str.
Hindi				
Donut	6.50	0	0.78	0
Qwen2.5VL-3B	6.40	0	2.56	3.36
Qwen2.5VL-7B	21.60	0	7.31	4.10
Gemma3-12B	21.26	29.41	11.78	6.85
Llama-3.2-11B	16.22	64.71	5.63	6.85
GPT-4o	32.22	29.41	14.87	9.59
Finetuned + Spatial Priors (SP)				
GPT-4o + SP	35.73	34.00	17.84	13.20
IndoTabVQA-hi [3B]	29.01	29.41	5.40	5.48
IndoTabVQA-hi [7B]	36.50	29.41	13.03	4.11
IndoTabVQA-hi [3B]+SP	30.39	29.41	8.29	8.22
IndoTabVQA-hi [7B]+SP	36.46	47.06	10.41	6.85
Arabic				
Donut	6.15	0	1.56	0
Qwen2.5VL-3B	21.50	5.88	5.28	4.11
Qwen2.5VL-7B	29.75	11.76	8.01	12.33
Gemma3-12B	20.72	35.29	7.42	12.33
Llama-3.2-11B	24.73	76.47	5.86	10.96
GPT-4o	23.75	47.06	10.74	17.81
Finetuned + Spatial Priors (SP)				
GPT-4o + SP	28.33	50.28	12.84	17.81
IndoTabVQA-ar [3B]	36.24	52.94	10.41	10.96
IndoTabVQA-ar [7B]	41.17	17.65	13.14	13.70
IndoTabVQA-ar [3B]+SP	38.62	35.29	11.34	9.59
IndoTabVQA-ar [7B]+SP	43.42	41.18	12.59	16.44

Table 4: Performance (In-match) across question types in four languages, Agg is Aggregation, Comp. is Comparison, Look. is Lookup, and Str. is Table structure related questions.

most existing benchmarks remain English-centric and fail to capture the visual noise, layout diversity, and multilingual characteristics of real-world documents.

5.2 Multilingual and Cross-Lingual VQA

Several benchmarks, including MULE (Kim et al., 2020), MTVQA (Tang et al., 2025) and MaXM (Changpinyo et al., 2022), target multilingual captioning and VQA. M^{4}C (Kesen et al., 2025) considers multilingual documents but focuses primarily on scene text or scanned forms. Recent work on XT-VQA (Yu et al., 2025) demonstrates the cross-lingual gap but is linguistically limited to Chinese, English, and French, whereas MMCricBench (Gautam et al., 2025b) is limited to English and Hindi only. Both benchmarks share the goal of evaluating cross-lingual transfer, but IndoTabVQA provides complementary coverage of different languages, writing systems, and document types. Table 6 contrasts IndoTabVQA with related benchmarks.

6 Conclusions

We introduce IndoTabVQA, a novel benchmark for table-based VQA grounded in real-world document images from an underrepresented region, with cross-lingual QA pairs in Bahasa Indonesia, English, Hindi, and Arabic. Our evaluation reveals that even state-of-the-art closed-source VLMs like GPT-4o struggle with layout-aware and cross-lingual reasoning, particularly in low-resource languages. Fine-tuning a compact 3B model and LoRA-finetuning a 7B model on our dataset substantially improves performance. Our analysis shows that lookup and structural reasoning remain the hardest categories. The additional performance boost from spatial priors underscores that table localization remains a key bottleneck. IndoTabVQA enables inclusive, structure-aware document AI and supports scalable research on document intelligence.

7 Limitations

While IndoTabVQA addresses an important gap in multilingual and cross-lingual table-based VQA, our work has several limitations that point to directions for future research.

Our Benchmark is table-centric; it can be further expanded to other layouts, such as charts and histograms, which have not yet been explored. Furthermore, our spatial priors rely on table-level bounding boxes, which improve performance. However, we believe that more fine-grained supervision, such as row, column, or cell-level structure, can further enhance performance, which remains to be explored. Incorporating richer structural annotations could further disentangle visual perception errors from reasoning errors.

References

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*.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Somraj Gautam, Abhishek Bhandari, and Gaurav Harit. 2025a. Tabcomp: A dataset for visual table reading comprehension. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5773–5780.

Somraj Gautam, Abhirama Subramanyam Penamakuri, Abhishek Bhandari, and Gaurav Harit. 2025b. Mind the (language) gap: Towards probing numerical and cross-lingual limits of lvlms. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 568–584.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and representation biases in multilingual models on dialectal nlp tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24003–24021.

Ilker Kesen, Jonas F Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. 2025. Multilingual pretraining for pixel language models. *arXiv preprint arXiv:2505.21265*.

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan Plummer. 2020. Mule: Multimodal universal language embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11254–11261.

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 *Computer Vision – ECCV 2022*, pages 498–517, Cham. Springer Nature Switzerland.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

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 WACV*, pages 2200–2209.

OpenAI. 2024. Gpt-4 api documentation. OpenAI API Documentation. Accessed: 2024-02-16.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the CVPR*, pages 8317–8326.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, and 1 others. 2025. Mtvqa: Benchmarking multilingual text-centric visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7748–7763.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. 2024a. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, pages 1–21. Springer.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Haiying Xia, Richeng Lan, Haisheng Li, and Shuxiang Song. 2023. St-vqa: shrinkage transformer with accurate alignment for visual question answering. *Applied Intelligence*, 53(18):20967–20978.

Xinmiao Yu, Xiaocheng Feng, Yun Li, Minghui Liao, Ya-Qi Yu, Xiachong Feng, Weihong Zhong, Ruihan Chen, Mengkang Hu, Jihao Wu, and 1 others. 2025. Cross-lingual text-rich visual comprehension: An information theory perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9680–9688.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, and 1 others. 2024. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37:7185–7212.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Appendix

Our dataset contains:

Question Type	Percentage (%)
Aggregation	58.1
Lookup	33.2
Comparison	1.6
Table Structure	6.9

Table 5: Distribution of Question Types

A.1 Error Analysis

Error Taxonomy: Incorrect (36-45%): Answer is completely wrong, no semantic relation to ground truth. Hallucination (10-20%): Model generates plausible but unsupported information. Partial Correct (31-38%): Answer includes correct information but adds or misses components. Typo (0.7-7%): Minor lexical variation or spelling error.

Benchmark	Cross lingual	QA Language	Visual Language	Table Focus
Tabular VQA:				
DocVQA	✗	English	English	Partial
TableVQA-B	✗	English	English	✓
TabComp	✗	English	English	✓
ComTQA	✗	English	English	✓
XT-VQA	✓	(EN/FR/CH)	English	Partial
MMCricBench	✓	English	EN/HI	✓
Ours:				
IndoTabVQA (3B + 7B)	✓	4 language (ID/EN/HI/AR)	Bahasa Indonesia	✓

Table 6: **Comparison of VQA Benchmarks:** Most existing benchmarks support the English language for visual content and QA. In contrast, our benchmark focuses on underrepresented low-resource languages, such as Bahasa Indonesia in vision, and includes a variety of languages in QA (including the reading order in case of Arabic). TableVQA-B is TableVQA-Bench (Kim et al., 2024)

Translation Error (9-22%, cross-lingual only): Misunderstanding due to language-specific phrasing.

Key Insights from Error Analysis: Borderless tables produce more hallucinations: Without a clear structure, models are more likely to invent relationships between cells. Cross-lingual errors are often translation-related: Models sometimes respond in the wrong language or misinterpret culture-specific terms. Numerical errors are rare but catastrophic: When models misread numbers, errors are factually wrong (not just semantic variations). Complex tables increase all error types: Tables with merged cells, nested headers, or irregular layouts have 2× higher error rates.

Answer Format: All answers are kept concise (1-5 words typically) and consistent across languages. Numerical answers use standard formatting (e.g., "1,380,201" or "1.380.201" depending on locale conventions).

A.2 Training Setup

All experiments were conducted on a single NVIDIA RTX A6000 GPU (48GB), where we fine-tuned Qwen2.5VL-3B using mixed-precision (bfloat16) with gradient checkpointing. Training used the Hugging Face `Trainer` API with an effective batch size of 4 (per-device batch size of 1 with 4-step gradient accumulation), learning rate of 2e-5, AdamW optimizer, and a linear schedule over 3 epochs, with separate models trained for each of the 4 languages and we used the LoRA method to finetune the Qwen2.5-VL-7B model.

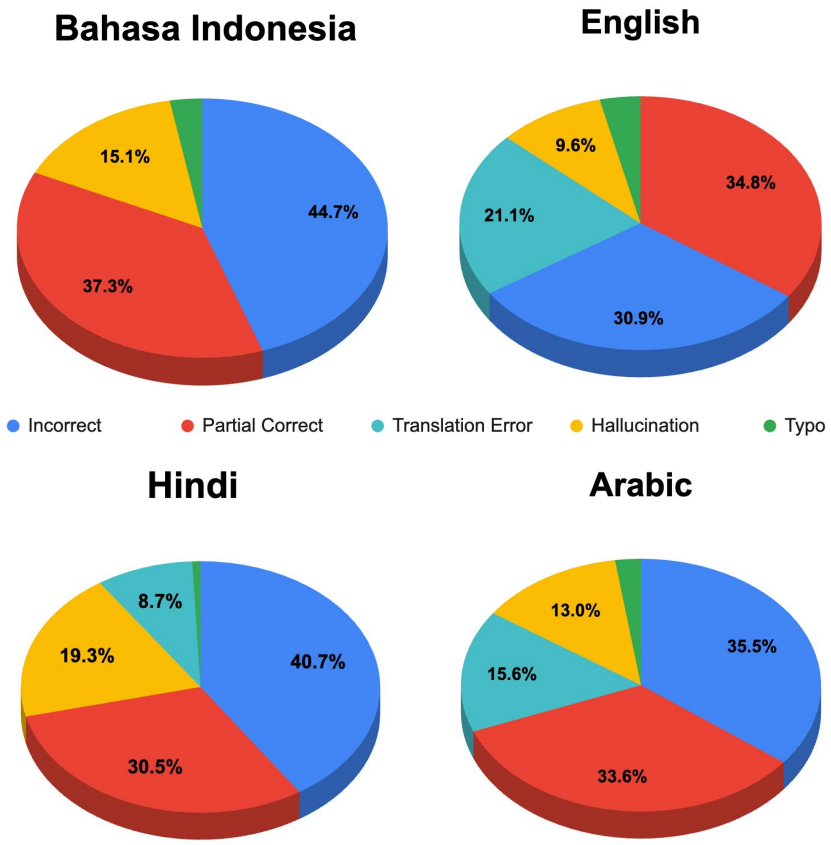

Figure 4: Comparative distribution of prediction error types across the four languages in the INDOTABVQA test set. The analysis categorizes failures into five types, revealing that 'Incorrect' and 'Partial Correct' are the dominant error modes. 'Translation Error' is a significant factor unique to the cross-lingual settings (English, Hindi, and Arabic), while 'Hallucination' and 'Typo' represent smaller but consistent sources of failure.

A.3 How INDOTABVQA is different:

- Geographic and linguistic diversity: We cover Southeast Asia (Bahasa Indonesia), South Asia (Hindi), and the Middle East (Arabic), alongside English, to ensure both regional representation and global accessibility.
- Script diversity: We include Devanagari (Hindi) and right-to-left script (Arabic), which pose different challenges than Latin/CJK scripts.
- Diverse and Real-world images: Our benchmarks consist of real-world images featuring various styles of table images.

Aspect	Protocol Summary
Question types	Lookup, aggregation, comparison, structure
Avg. tokens per question	7–10 (across languages)
Translation workflow	MT → human validation → consistency check
Discarded items	3.1% (ambiguity, mistranslation)
Annotators	3 (Bahasa), 4 validators (EN/HI/AR)

Table 7: Annotation protocol summary

Question	Answer	Predicted answer	E
Which column has 4 points as its contents?	form of learning	day (hour) (j), weight of value	H
What is the most common type of disability in Indonesia?	mentally disabled	physical handicap, multiple..	I
What is the unit of school that receives the l...	school	sekolah	T
What description is given from the falcon and eagle?	(all types of the family)	all species from the family	Ty
Which province has the highest number of persons with disabilities?	yogyakarta	west java, yogyakarta	P

Table 8: Examples of errors, H=Hallucination, I=Incorrect, T=Translation, Ty=Typo, P=Partial Correct

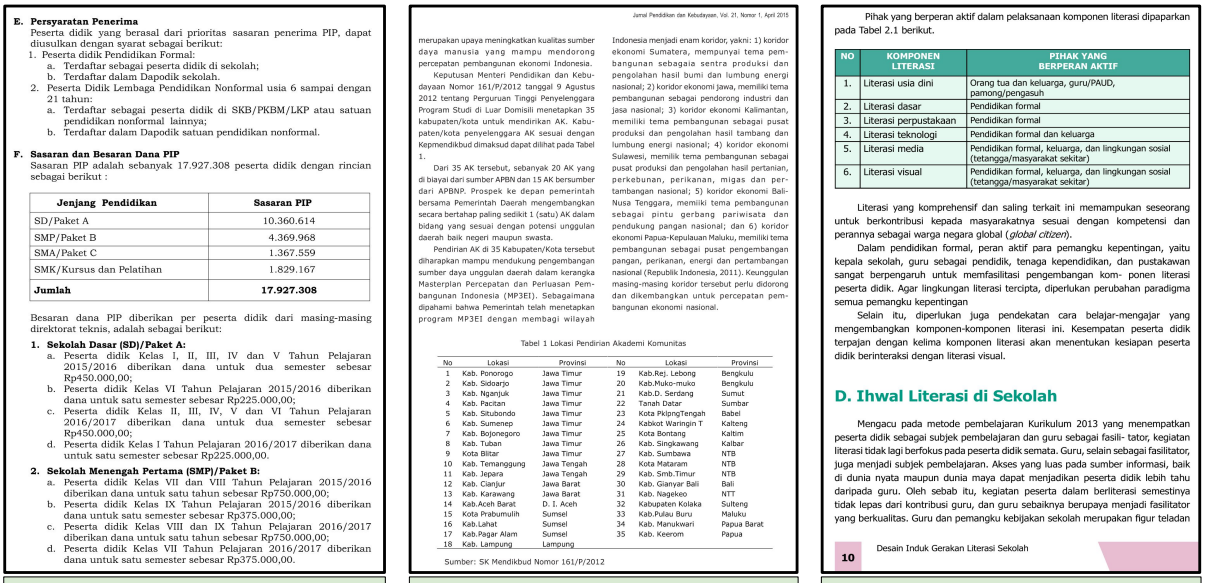

Figure 5: Example of the INDOTABVQA correct predictions on mono-lingual and cross-lingual question answering across three table formats. Bordered (left), Borderless (middle), and Colorful (right). The examples include questions in Bahasa Indonesia, English, Hindi, and Arabic.