

Do MLLMs Really Understand Low-Resource Khmer Documents? A Pilot Study on Khmer Document VQA

Nimol Thuon^{1,2*} and Panhapin Theang²

¹ University of Science and Technology of China, 230052 Hefei, China

² Université Paris Cité, 75013 Paris, France

`tnimol@mail.ustc.edu.cn, panhapin.theang@etu.u-paris.fr`

Abstract. Recent multimodal large language models (MLLMs) have advanced document understanding, visual question answering, and text extraction. However, their reliability in low-resource, non-Latin settings remains uncertain. Khmer form documents present particular challenges because they contain complex script forms, mixed Khmer–English fields, and monetary values in both Cambodian Riel and US Dollars. Available resources for Khmer Document VQA are also limited. This paper presents a pilot diagnostic evaluation of open MLLMs on Khmer document images. We construct an evaluation subset from the previously introduced KH-FUNSD collection, covering invoices, receipts, quotations, and other business forms. The subset includes questions in English and Khmer, with answers retained in their original English, Khmer, mixed-script, or numeric forms. Rather than introducing a full public benchmark, this study examines the capabilities and failure modes of existing models. We evaluate representative open Qwen-VL models using direct image-based prompting and compare parser-assisted and external OCR-assisted configurations with Qwen3-VL-8B. Direct Qwen3-VL-8B outperforms smaller models, achieving 51.9% overall accuracy, although performance remains limited for Khmer-script and mixed-script answers. External OCR produces the strongest results, reaching 61.9% with Tesseract and 61.6% with PaddleOCR. Nevertheless, Khmer-script answers remain substantially more difficult than English and numeric fields. The results indicate that current MLLMs can process visually clear English and structured numeric content, but reliable native Khmer document understanding remains an open challenge.

Keywords: Document VQA · Multimodal Large Language Models · Khmer Documents · Low-Resource Languages · Non-Latin Scripts · OCR

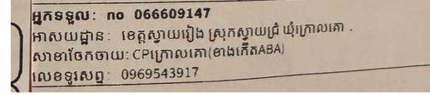
1 Introduction

Multimodal large language models (MLLMs) are increasingly used for document understanding tasks such as visual question answering, OCR-like transcription,

* Corresponding author

(a) Khmer OCR/Text Extraction

Prompt: Extract Khmer text, preserve format



Ground truth:

អ្នកទទួល: no 066609147
អាសយដ្ឋាន: ខេត្តស្វាយរៀង ស្រុកស្វាយរៀង ឃុំក្រាលគោ
សាខាចែកចាយ: CP ក្រាលគោ (ខាងកើតABA)
លេខទូរស័ព្ទ: 0969543917

Outputs:

Qwen 3.6 Plus:

អ្នកទទួល: លេខ 066609147
អាសយដ្ឋាន: 189 ផ្លូវ មុខវត្ត
ស្ថានភាព: CP (សាខាថ្មី) (តាវី&A&A)
លេខទូរស័ព្ទ: 0969543917

Gemini 3.1:

អ្នកទទួល: No 066609147
អាសយដ្ឋាន: ខេត្តស្វាយរៀង ស្រុកស្វាយរៀង ឃុំក្រាលគោ
សាខាចែកចាយ: CP (ប្រសូត្រ) ទីសក្តិ ABA
លេខទូរស័ព្ទ: 0969543917

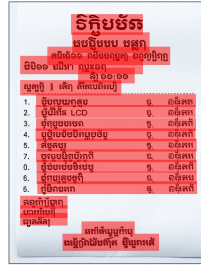
ChatGPT 5.5:

លេខកូដ: 006690147
អាសយដ្ឋាន: [blurred]
លេខអ្នកទទួល: [blurred]
លេខទូរស័ព្ទ: 0969543917

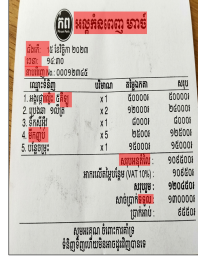
(b) Khmer Document Image Generation

Prompt: Create a realistic Khmer receipt with full Khmer text.

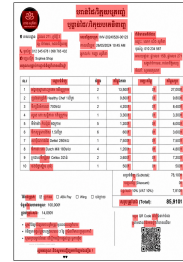
Output:



Qwen 3.6 Plus



Nano Banana 2/ Gemini 3.1



ChatGPT Images 2.0/ChatGPT 5.5

Fig. 1. Limitations of current MLLMs on Khmer document content, shown in two tasks: (a) Khmer OCR/text extraction, where models omit or corrupt native Khmer characters; and (b) Khmer document image generation, where models produce visually plausible layouts with pseudo-Khmer or linguistically invalid text.

information extraction, form interpretation, and layout-aware reasoning. Document Visual Question Answering (Document VQA or DocVQA) has become an important benchmark direction for document intelligence [9, 1, 17, 10].

Recent MLLMs, including Qwen-VL [3], Qwen2.5-VL [5], and Qwen3-VL [4], suggest that vision-language models are becoming increasingly capable at text-rich image understanding, document-style question answering, and structured information extraction. However, strong progress on mainstream document benchmarks does not necessarily imply robust understanding of low-resource non-Latin documents. Khmer is a low-resource non-Latin language with limited document AI resources compared with high-resource languages such as English, Chinese, Japanese, and major European languages. Khmer script has visually complex characteristics, including stacked consonants, dependent vowels, diacritics, dense character compositions, and limited explicit word separation. These properties make text recognition difficult, particularly in real-world documents affected by blur, low resolution, perspective distortion, compression artifacts, uneven illumination, and handwriting or stamps [20, 7, 11]. For readers unfamiliar with Khmer

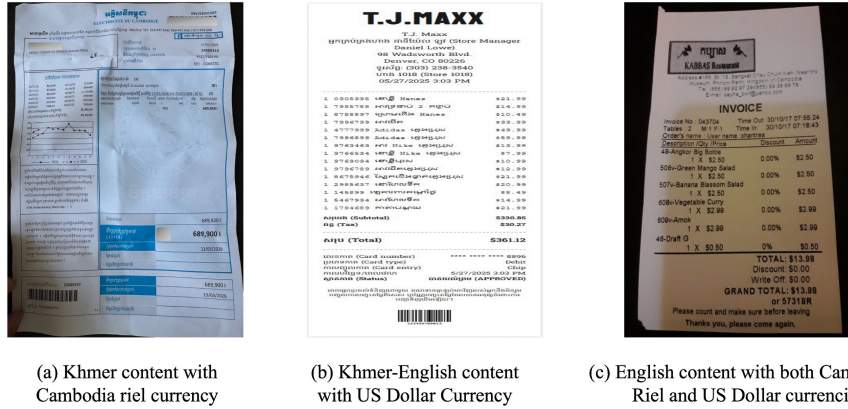


Fig. 2. Real-world Khmer business documents present mixed-language and currency challenges: (a) Khmer content with Cambodia Riel currency; (b) mixed Khmer–English content with US Dollar currency; and (c) English-dominant content listing both Cambodian Riel and US Dollar values. These examples illustrate why models must handle native Khmer text, mixed-script fields, and currency grounding simultaneously.

script, the cropped examples in Fig. 1 and the document examples in Fig. 2 illustrate the dense native-script text, stacked forms, and mixed Khmer–English fields that make this setting different from Latin-script document VQA.

Interestingly, current MLLMs may produce plausible outputs for Khmer document tasks while still making serious errors in Khmer text extraction or document generation, as shown in Figure 1. This observation motivates the present study. In OCR-like extraction, models may omit words, replace characters, misread numbers, or generate malformed Khmer strings. In document generation, models may create visually plausible receipt or form layouts while producing pseudo-Khmer or mixed Southeast Asian-looking text that is not linguistically valid. These examples indicate that visual plausibility does not necessarily imply genuine Khmer document understanding. Real-world Khmer documents introduce challenges beyond native script recognition. As shown in Fig. 2, Cambodian invoices, receipts, quotations, and administrative forms often contain mixed Khmer–English content. Khmer and English are not always separated into clean monolingual blocks; instead, they may appear in the same line, field, table, product description, company header, or address. Currency usage adds another difficulty. Cambodian business documents may use Cambodian Riel, US Dollars, or both. A model must therefore read the numeric value, identify the associated field, and ground the amount to the correct currency and document context.

Despite these challenges, Khmer document understanding remains underexplored in current Document VQA research. Existing benchmarks have advanced the field, but many are dominated by high-resource languages, relatively clean document images, or document collections that do not fully represent low-

resource non-Latin settings. Evaluating MLLMs only on such benchmarks may overestimate their ability to generalize to real-world multilingual documents from underrepresented languages.

In this paper, we present a Khmer-only pilot diagnostic study of MLLM-based Document VQA. Our goal is not to introduce a full benchmark or a broad model leaderboard. Instead, we examine whether recent open MLLMs can reliably answer questions over selected Khmer business document images and analyze where they fail. We design both English-question and Khmer-question settings and preserve answers in their original form. This allows us to analyze model behavior across question language and answer type, including English, Khmer, mixed-script, and number/date/currency answers.

Failure analysis is central to this goal. In business and financial document workflows, a system can appear useful from overall accuracy while still failing on high-risk cases such as names, native Khmer textual fields, or currency amounts. Studying failure cases therefore clarifies what task success would require: not only plausible answers, but reliable native-script reading, field grounding, and calibrated use of OCR evidence.

The main contributions of this paper are as follows:

- We present a pilot diagnostic study of MLLM performance on low-resource Khmer Document VQA.
- We analyze both English-question and Khmer-question settings while preserving answers in their original script or format.
- We compare direct prompting, MLLM-parser-assisted prompting, and external OCR-assisted prompting.
- We identify failure modes involving Khmer-script recognition, mixed-script field confusion, currency grounding, and plausible but unsupported answers.

2 Related Work

2.1 Document VQA and MLLMs for Document Understanding

Visual Question Answering (VQA) was introduced as a general vision-language task in which a model answers natural-language questions about images [1]. Document Visual Question Answering later extended this setting to document images, where models must read text, interpret layout, identify tables or forms, and ground answers in visually structured content. Unlike general image VQA, Document VQA requires close interaction among OCR-like reading, layout understanding, and visual reasoning. The DocVQA dataset introduced this task at scale, with questions defined over document images and answers extracted from document content [9]. Subsequent work extended the setting to document collections, where answering a question may require retrieving relevant documents before extracting the answer [15]. More recent work has also explored explainability in Document VQA, for example by learning visual evidence maps that justify model answers [17].

Recent MLLMs have shown strong progress in image-text understanding and document-style reasoning. Qwen-VL introduced a vision-language model family with text-reading and grounding abilities [3]. Qwen2.5-VL further emphasizes visual recognition, document parsing, structured extraction from forms and tables, and dynamic-resolution processing [5]. Qwen3-VL extends this family with stronger multimodal reasoning, expanded OCR support, long-context capability, and dense model variants such as 4B and 8B, which are suitable for practical open-model evaluation [4].

Despite these advances, document understanding remains vulnerable to fluent but incorrect outputs. Existing Document VQA benchmarks have played an important role in evaluating document understanding systems, but many are still dominated by English or other high-resource language settings. Strong benchmark performance may therefore hide weaknesses on low-resource non-Latin documents.

2.2 OCR and Parser-Assisted Document VQA

Many document understanding pipelines rely on OCR or document parsing before question answering. Early text-centric VQA work showed that text detected by OCR can provide important cues for answering questions over text-rich images [10]. In document intelligence, OCR-based representations have also been widely used to combine recognized text, layout positions, and visual features. For example, LayoutLM and LayoutLMv2 model text, layout, and image information jointly for visually rich document understanding tasks [22, 21]. Similarly, DocVQA formulates document visual question answering as answering natural-language questions from document images, often requiring the model to read text and ground answers in document regions [9].

Recent MLLMs can process document images directly, reducing dependence on explicit OCR pipelines. However, it remains unclear whether direct MLLM prompting is sufficient for low-resource non-Latin scripts. Parser-assisted prompting provides one way to test this issue. A model first produces textual evidence or structured fields from the document, and a second pass uses this evidence to answer questions. External OCR-assisted prompting provides another setting, where text extracted by a dedicated OCR engine is provided to the MLLM together with the image and question.

2.3 Khmer Document Understanding

Low-resource document understanding remains challenging because many languages lack large annotated datasets, robust OCR systems, document layout benchmarks, and pretrained language resources. For non-Latin scripts, the problem is further complicated by script-specific visual structures such as diacritics, stacked forms, character shaping, and irregular spacing. These issues become more severe in real-world documents where text may be small, blurred, distorted, stamped, handwritten, or mixed with other languages [11, 8, 20, 18].

Table 1. Statistics of the pilot Khmer Document VQA evaluation subset.

Dimension	Categories and QA pairs
Document type	Invoice (168), Quotation (103), Receipt (99)
Question language	English (249), Khmer (121)
Answer group	English (160), Khmer (50), Mixed-script (50), Number/date/currency (110)
Question type	Counting (73), Field value (76), Layout field (78), Name (70), Total amount (73)
Difficulty	Easy (192), Medium (96), Hard (82)
Total	62 document images and 370 QA pairs

Khmer document understanding is particularly underrepresented. Prior Khmer OCR studies have highlighted challenges such as stacked characters, diacritics, non-uniform character widths, limited word separation, and scarce training data [19, 11]. However, most existing work focuses on OCR or scene text recognition rather than Document VQA, where a model must interpret a question, locate the relevant field, read the answer, and preserve the correct script or numeric format.

Real-world Khmer business documents also contain mixed-language and currency rich content. In Cambodia, Khmer is the official language, while English is commonly used in business settings, and other languages such as Chinese may appear in commercial contexts [6, 19]. More generally, multilingual societies often involve code-switching or mixed-language usage [16, 2]. As a result, Cambodian invoices, receipts, and quotations may contain Khmer, English, or mixed Khmer–English fields. They may also include Cambodian Riel, US Dollars, or both, reflecting Cambodia’s highly dollarized economy [13, 12, 14].

3 Pilot Khmer Document VQA Dataset

3.1 Source Collection and Scope

The pilot evaluation subset is drawn from KH-FUNSD [19], a previously introduced collection of real-world Khmer business and administrative documents, including invoices, receipts, quotations, and other forms. These documents contain mixed Khmer–English text, structured fields, tabular layouts, company information, itemized rows, numeric identifiers, dates, and currency values.

We use KH-FUNSD only as the source of a small diagnostic Document VQA subset and do not claim to introduce a full public Khmer Document VQA benchmark. Our goal is to evaluate whether current MLLMs can reliably answer questions about low-resource Khmer documents and to identify common failure patterns.

The subset focuses on invoices, receipts, and quotations because they contain recurring business concepts, structured fields, mixed-language content, and monetary values. They also represent practical Cambodian workflows in which document understanding can support information extraction, search, and administrative processing.

3.2 Question and Answer Design

For each document image, we construct question-answer pairs covering common document understanding needs, including field extraction, name lookup, layout-dependent field lookup, counting, and amount-related queries. The evaluation includes English questions because English is commonly used in existing Document VQA benchmarks and allows controlled question writing across documents. To evaluate native-language question understanding, we additionally include Khmer questions translated or adapted from selected English questions while preserving the same answer target.

Answers are preserved in their original form as they appear in the document. Therefore, an answer may be written in English, Khmer, mixed-script form, or number/date/currency format. We keep both raw and normalized answer fields. The raw answer preserves the original document text, while the normalized answer is used only when superficial formatting differences should not affect evaluation, such as commas in numbers, date variants, or currency aliases. We avoid aggressive normalization for Khmer text to prevent distorting native-script answers.

We first describe the composition of the pilot evaluation subset used in this study. Table 1 summarizes the pilot subset. It contains 62 document images and 370 question-answer pairs across invoices, receipts, and quotations. The subset is modest in scale, but it is intentionally designed for controlled diagnostic analysis across question language, answer group, question type, and difficulty.

3.3 Diagnostic Labels

Each question is assigned a question type and a difficulty label. The question types include counting, field-value extraction, layout-field lookup, name-related questions, and total-amount questions. These categories reflect common operations in business document understanding rather than a complete taxonomy of all possible Document VQA tasks. Difficulty labels are used as diagnostic categories rather than strict cognitive difficulty levels. In this subset, hard questions often require Khmer-script name recognition or native textual-field extraction, so they mainly test native-script reading and grounding rather than complex reasoning alone.

Each question-answer instance contains a document identifier, question identifier, document type, question language, question type, difficulty label, input question, raw answer, normalized answer, answer language, and answer group. These metadata fields enable the analysis reported in Sec. 5.

4 Experimental Setup

4.1 Question Answering Protocol

We formulate Khmer Document VQA as a short-answer visual question answering task. Given a document image I and a natural-language question q , the

model generates a concise answer a based on the visible document content. The question may be written in English or Khmer, and the answer may appear in English, Khmer, mixed-script form, or number/date/currency format.

For all experiments, the model is instructed to return a short JSON object containing the predicted answer and a confidence label:

```
{"answer": "...", "confidence": "high/medium/low"}
```

The model is instructed to preserve Khmer, English, numbers, dates, and currency values exactly as they appear whenever possible. If the answer is not visible or uncertain, the model is instructed to return **Unknown**.

4.2 Evaluated Models and Evidence Settings

We evaluate representative open Qwen-VL models under direct image-based prompting. The direct-prompting models are Qwen2.5-VL-3B, Qwen3-VL-4B, and Qwen3-VL-8B. The purpose is not to build a broad leaderboard, but to diagnose whether recent practical MLLMs can reliably answer questions over low-resource Khmer document images.

We further evaluate three Qwen3-VL-8B evidence settings. In the **direct** setting, the model receives only the document image and the question. In the **MLLM-parser** setting, Qwen3-VL-8B first generates visible text, key fields, amount/date/currency fields, and layout notes; a second pass then answers using the image, question, and generated parser text. In the **external OCR** setting, text extracted by Tesseract or PaddleOCR is provided together with the image and question. In all assisted settings, the image remains available to the model and extracted text is treated as supporting evidence rather than a replacement for visual grounding.

4.3 Implementation and Prompt Details

All settings use the same JSON output format shown above and ask the model to preserve Khmer, English, numbers, dates, and currency values exactly when possible. In the direct setting, the model receives only the document image and question. In the MLLM-parser setting, Qwen3-VL-8B first generates visible text, key fields, amount/date/currency fields, and layout notes; this generated parser text is then provided to a second answering pass together with the image and question. In the external OCR setting, uncorrected text from Tesseract or PaddleOCR is provided together with the image and question. In all assisted settings, the image remains available, extracted text is treated only as supporting evidence, and no manually corrected or oracle transcript is provided.

OCR engines are used without task-specific fine-tuning or manual transcript correction. The experiments retain the original document images and do not use hand-cropped evidence regions during quantitative evaluation. Generation is run deterministically with sampling disabled. These details are intended to make clear that the assisted settings test whether ordinary parser or OCR evidence improves MLLM answering, rather than whether a fully optimized Khmer OCR pipeline can solve the task.

Table 2. Direct-prompting results on the Khmer Document VQA pilot set. Accuracy is reported in percentage.

Model	Parse	All	Question Lang.		Answer Group				
			En-Q	Kh-Q	En	Kh	Mix	Num.	
Qwen2.5-VL-3B	90.5	40.8	43.4	35.5	53.8	0.0	10.0	54.5	
Qwen3-VL-4B	90.3	41.9	48.2	28.9	56.3	0.0	12.0	53.6	
Qwen3-VL-8B	98.6	51.9	55.4	44.6	71.3	0.0	14.0	64.5	
Count	370	370	249	121	160	50	50	110	

4.4 Evaluation Metric

We report normalized answer accuracy, where a prediction is correct if it matches either the raw or normalized ground-truth answer after basic normalization: English lowercasing, whitespace cleanup, comma removal in numbers, and simple currency aliases. Khmer answers are preserved as written. Accuracy is computed at the QA-pair level over all annotated examples, with unparseable outputs counted as incorrect. We separately report **Parse**, the percentage of outputs that follow the expected JSON format. We analyze accuracy by question language, answer group, question type, difficulty, and document type.

5 Results and Analysis

5.1 Direct MLLM Performance

We first evaluate the MLLMs under a direct-prompting setting, where each model receives the document image and question and is asked to generate the answer in the required JSON format. Table 2 summarizes the direct-prompting results. Qwen3-VL-8B achieves the best direct-prompting performance, with 51.9% overall accuracy and 98.6% parsing success. This indicates that the larger model follows the requested JSON format more reliably and improves general document answering. However, the gains are uneven across answer groups. All direct models perform much better on English answers and number/date/currency fields than on Khmer-script and mixed-script answers. Most notably, all direct models obtain 0.0% accuracy on Khmer-script answers, suggesting that model scaling alone does not solve the native Khmer text recognition problem.

We further break down direct-prompting performance by question type and difficulty level. Table 3 provides a complementary view by question type and difficulty. Qwen3-VL-8B substantially improves counting accuracy, but name-question accuracy remains low. Since many name questions require reading Khmer-script person or organization names, this result is consistent with the answer-group analysis. Accuracy also drops sharply on hard questions. In this pilot subset, hard questions often require native Khmer text recognition rather than complex reasoning alone.

Table 3. Direct-prompting accuracy by question type and difficulty. Accuracy is reported in percentage over all annotated QA pairs; the Count row gives the total number of annotated QA pairs.

Model	Question Type				Difficulty			
	Count.	Field value	Layout	Name	Total	Easy	Med.	Hard
Qwen2.5-VL-3B	46.6	53.9	52.6	12.9	35.6	59.9	36.5	1.2
Qwen3-VL-4B	53.4	52.6	59.0	7.1	34.2	62.0	35.4	2.4
Qwen3-VL-8B	74.0	56.6	67.9	11.4	46.6	73.4	47.9	6.1
Count	73	76	78	70	73	192	96	82

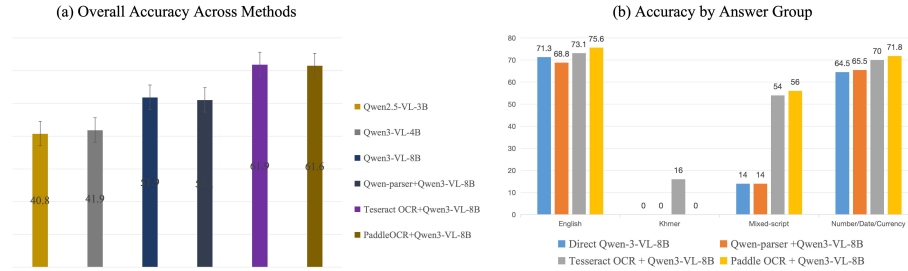


Fig. 3. Quantitative results on the pilot Khmer Document VQA subset. (a) Overall accuracy across direct MLLM prompting, MLLM-parser-assisted prompting, and external OCR-assisted prompting, with Wilson 95% confidence intervals. (b) Accuracy by answer group for Qwen3-VL-8B variants.

5.2 Parser and OCR-Assisted Evidence

To assess whether explicit textual evidence helps overcome recognition limitations, we compare direct prompting with parser-assisted and OCR-assisted settings. Figure 3 and Table 4 summarize these results. Direct Qwen3-VL-8B reaches 51.9% overall accuracy, while MLLM-parser-assisted prompting remains similar at 51.1%. This suggests that self-generated parser evidence does not substantially reduce the model’s recognition errors. In contrast, Tesseract OCR and PaddleOCR raise overall accuracy to 61.9% and 61.6%, respectively.

The improvement is especially large for mixed-script and number/date/currency answers. PaddleOCR improves mixed-script accuracy from 14.0% to 56.0%, while Tesseract reaches 54.0%. However, native Khmer-script answers remain difficult. Tesseract improves Khmer-answer accuracy to 16.0%, but PaddleOCR remains at 0.0%. This contrast suggests that external OCR is useful for structured and mixed-script evidence, but low-resource Khmer recognition remains unresolved.

The confidence intervals emphasize that small differences should not be over-interpreted. For example, the overall direct Qwen3-VL-8B result of 51.9% has a Wilson 95% interval of approximately 46.8–56.9%, while Tesseract OCR and PaddleOCR have intervals of approximately 56.8–66.7% and 56.6–66.4%, respectively. For 50-example answer groups, uncertainty is much larger: the direct mixed-script result of 14.0% corresponds to roughly 6.9–26.2%, while the Pad-

Table 4. Direct, parser-assisted, and external OCR-assisted Qwen3-VL-8B results. Accuracy is reported in percentage.

Method	Parse	All	Question Lang.		Answer Group			
			En-Q	Kh-Q	En	Kh	Mix	Num.
Direct	98.6	51.9	55.4	44.6	71.3	0.0	14.0	64.5
MLLM-parser	99.2	51.1	53.4	46.3	68.8	0.0	14.0	65.5
Tess. OCR	99.7	61.9	64.3	57.0	73.1	16.0	54.0	70.0
PaddleOCR	98.1	61.6	63.5	57.9	75.6	0.0	56.0	71.8

Table 5. Question-type comparison for Qwen3-VL-8B under direct and OCR-assisted settings. Accuracy is reported in percentage over all annotated QA pairs; the Count row gives the total number of annotated QA pairs.

Method	Question Type				
	Count.	Field	value	Layout	Name Total
Direct	74.0	56.6	67.9	11.4	46.6
Tess. OCR	74.0	68.4	73.1	37.1	54.8
PaddleOCR	80.8	69.7	64.1	30.0	61.6
Count	73	76	78	70	73

dleOCR mixed-script result of 56.0% corresponds to roughly 42.3–68.8%. Thus, the main conclusion is not a fine ranking between OCR systems, but the broader pattern that OCR evidence helps mixed-script and numeric fields while native Khmer answers remain difficult.

5.3 Question-Type Gains from OCR

We further analyze the effect of OCR assistance across different question categories. Table 5 shows that OCR assistance changes which question types can be answered. Tesseract provides the largest improvement for name questions, increasing accuracy from 11.4% to 37.1%. PaddleOCR gives the strongest result for total-amount questions, increasing accuracy from 46.6% to 61.6%. These gains suggest that OCR evidence helps the model locate and copy difficult textual or numeric fields, although name questions remain substantially harder than counting and field-value questions.

5.4 Qualitative Failure Analysis

To better interpret the quantitative results, we examine representative failure cases produced by the MLLMs. As shown in Figure 5, the observed failures can be grouped into several major categories, including Khmer-script recognition errors, mixed-script confusion, layout-field confusion, currency or amount grounding errors, OCR noise propagation, and plausible but unsupported answers. Additionally, Figure 4 provides qualitative examples that further explain these error patterns. The first example shows a Khmer-script recognition failure, where the model cannot correctly read a native Khmer textual field. The second

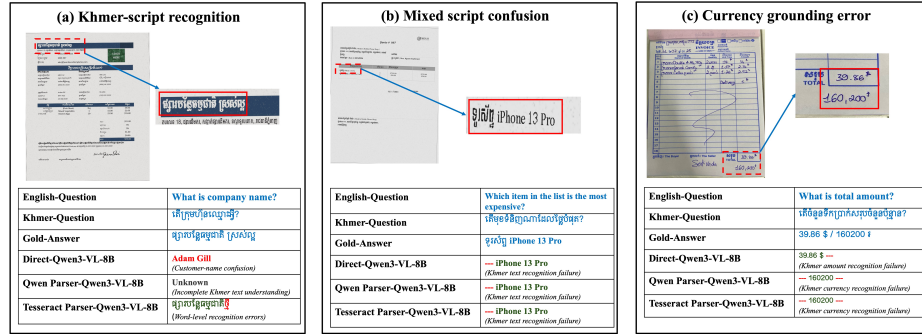


Fig. 4. Representative failure modes in Khmer Document VQA. (a) Khmer-script recognition errors occur when the model fails to read native Khmer names or textual fields. (b) Mixed-script confusion occurs when Khmer and English appear in the same field, causing the model to copy only the Latin-script portion or produce a partial answer. (c) Currency grounding errors occur when the model reads nearby numeric values but associates them with the wrong field or currency. Green text indicates correct answers, red text indicates incorrect predictions, and cropped regions highlight the relevant evidence.

example shows mixed-script confusion: the English part of the field is easier to copy, while the Khmer component is ignored or corrupted. The third example shows a currency grounding error, where the model reads a nearby amount but associates it with the wrong currency or total field. These examples support the main finding that current MLLMs are limited not only by answer formatting, but also by native-script recognition and document-field grounding.

6 Discussion and Limitations

6.1 Discussion

The results support four main findings. Larger Qwen-VL models improve direct accuracy and output-format stability, but direct prompting remains unreliable for Khmer-script and mixed-script answers. The bottleneck is not only output formatting: Qwen3-VL-8B parses reliably but still misses many native-script fields. MLLM-generated parser evidence gives little gain, suggesting that self-parsing often repeats the same recognition errors. External OCR provides the strongest improvement, especially for mixed-script and number/date/currency fields, yet Khmer-script answers remain much harder than English and numeric fields.

Overall, current MLLMs can handle some English-visible and structured numeric content in Khmer documents, especially with OCR support. However, robust native Khmer document understanding remains unresolved, and aggregate accuracy can hide severe native-script failures. Script-aware and language-aware evaluation is therefore necessary for low-resource Document VQA.

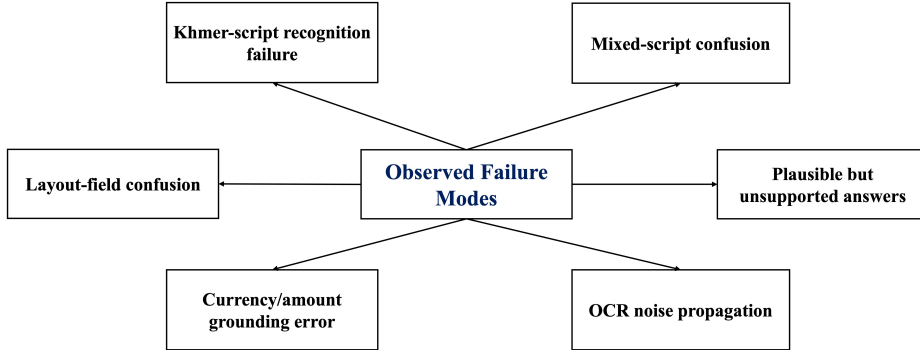


Fig. 5. Observed failure modes in Khmer Document VQA, including Khmer-script recognition failure, mixed-script confusion, layout-field confusion, currency/amount grounding errors, OCR noise propagation, and plausible but unsupported answers.

6.2 Limitations and Future Work

This study is a pilot diagnostic study rather than a full benchmark. It uses 62 document images and 370 question-answer pairs from one Khmer form-understanding source collection, so coverage across document domains is limited. The experiments also focus on open Qwen-VL models and two OCR engines. Future work should evaluate broader model families, larger MLLMs, closed-source systems such as GPT-4o or Gemini, and additional OCR systems.

The evaluation protocol also has limits. Normalized exact matching is practical for short-answer Document VQA but can mis-handle equivalent Khmer text, names, or currency formats. We do not include an oracle OCR condition with human transcripts, so OCR recognition errors cannot be fully separated from downstream MLLM comprehension or grounding errors. Confidence labels are collected but not yet calibrated. Future work should add human adjudication, oracle-transcript experiments, confidence calibration, Khmer-aware normalization, and a larger public Khmer or multilingual non-Latin Document VQA benchmark, subject to privacy and copyright constraints.

7 Conclusion

This paper presented a pilot diagnostic study of current open MLLMs for Khmer Document VQA, an underexplored low-resource non-Latin document setting. Using 62 Khmer document images and 370 question-answer pairs, we evaluated direct Qwen-VL prompting, MLLM-parser-assisted prompting, and external OCR-assisted prompting across question language, answer group, question type, difficulty, and evidence setting. The results show that Qwen3-VL-8B improves direct performance and output-format stability over smaller models, but remains weak on Khmer-script and mixed-script answers. MLLM-generated parser evidence provides little improvement, suggesting that self-parsing often repeats the

same recognition errors. In contrast, external OCR assistance improves performance, especially for mixed-script and number/date/currency fields, although native Khmer-script answers remain difficult. Overall, current MLLMs are not yet robust Khmer document readers. They can handle some structured, English-visible, and numeric content, but native-script recognition, mixed-script grounding, and currency-field interpretation remain major challenges. Future work should develop larger Khmer and multilingual non-Latin Document VQA benchmarks, stronger Khmer-aware OCR and normalization methods, and evaluation protocols that explicitly measure native-script and mixed-language document understanding.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015). <https://doi.org/10.1109/ICCV.2015.279>
2. Auer, P. (ed.): Code-Switching in Conversation: Language, Interaction and Identity. Routledge, London (1998)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint (2023). <https://doi.org/10.48550/arXiv.2308.12966>
4. Bai, S., Cai, Y., Chen, R., Chen, K., et al.: Qwen3-vl technical report. arXiv preprint (2025). <https://doi.org/10.48550/arXiv.2511.21631>
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint (2025). <https://doi.org/10.48550/arXiv.2502.13923>
6. International Trade Administration: Cambodia - business travel. <https://www.trade.gov/country-commercial-guides/cambodia-business-travel> (2025), accessed 2026-05-12
7. Keo, S., Nom, V., Bakkali, S., Luqman, M.M., Coustaty, M., Ogier, J.M.: Visual text generation in khmer language: Challenges and trends with diffusion models. In: International Conference on Document Analysis and Recognition. pp. 134–152. Springer (2025)
8. Kong, M., Buoy, R., Chenda, S., Taing, N., Iwamura, M., Kise, K.: Towards khmer scene document layout detection (2026)
9. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2199–2208 (2021). <https://doi.org/10.1109/WACV48630.2021.00225>
10. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 947–952. IEEE (2019). <https://doi.org/10.1109/ICDAR.2019.00156>

11. Nom, V., Keo, S., Bakkali, S., Luqman, M.M., Coustaty, M., Rossinyol, M., Ogier, J.M.: Wildkhmerst: a comprehensive dataset and benchmark for khmer scene text detection and recognition in the wild. In: International Conference on Document Analysis and Recognition. pp. 351–368. Springer (2025)
12. Odajima, K.: Dollarization in cambodia: Behavior of households and enterprises in a highly dollarized environment. In: Dollarization and De-dollarization in Transitional Economies of Southeast Asia, pp. 33–72. Springer (2017). https://doi.org/10.1007/978-3-319-57768-5_2
13. Odajima, K., Aiba, D.: Dollarization in cambodia: A review of recent empirical findings from a nation-wide survey. *Asian Studies* **65**(1), 24–44 (2019). https://doi.org/10.11479/asianstudies.65.1_24
14. Odajima, K., Aiba, D., Khou, V.: Currency choice in domestic transactions by cambodian households: The importance of transaction size and network externalities. Working Paper 185, JICA Ogata Sadako Research Institute for Peace and Development (2019), https://www.jica.go.jp/jica-ri/publication/workingpaper/wp_185.html
15. Pérez Tito, R., Karatzas, D., Valveny, E.: Document collection visual question answering. In: Document Analysis and Recognition – ICDAR 2021. pp. 778–792. Springer (2021). https://doi.org/10.1007/978-3-030-86331-9_50
16. Poplack, S.: Sometimes i’ll start a sentence in spanish y termino en español: Toward a typology of code-switching. *Linguistics* **18**(7-8), 581–618 (1980). <https://doi.org/10.1515/ling.1980.18.7-8.581>
17. Souibgui, M.A., Choi, C., Barsky, A., Jung, K., Valveny, E., Karatzas, D.: Docvxqa: Context-aware visual explanations for document question answering. arXiv preprint (2025)
18. Thuon, N., Du, J.: Khmerformer: Multi-scale cnns-transformer with external attention for ancient khmer palm leaf isolated glyph classification. In: 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1–6. IEEE (2024). <https://doi.org/10.1109/APSIPAASC63619.2025.10849201>
19. Thuon, N., Du, J.: Kh-funsd: A hierarchical and fine-grained layout analysis dataset for low-resource khmer business document. In: 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1886–1891. IEEE (2025)
20. Thuon, N., Du, J., Zhang, J.: Syllable analysis data augmentation for khmer ancient palm leaf recognition. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1855–1862. IEEE (2022). <https://doi.org/10.23919/APSIPAASC55919.2022.9980217>
21. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. pp. 2579–2591 (2021). <https://doi.org/10.18653/v1/2021.acl-long.201>
22. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: 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. pp. 1192–1200 (2020). <https://doi.org/10.1145/3394486.3403172>