Attention as Selector: Unlocking VLM Attention for Long Document Page Retrieval

Anonymous ACL submission

Abstract

Visual Language Models (VLMs) have become a robust foundation for document question answering. Processing long documents remains challenging due to limited context windows and computational budgets. Existing page-level retrieval methods offer a practical solution, typically encoding pages and queries into vectors and ranking them via cosine similarity. However, such embedding-based methods (i) lack query–page interaction before similarity scoring and (ii) usually require large-scale dataset to align visual and textual embeddings. In this paper, we observe that the cross-modal attention maps of well-trained VLMs are able to highlight semantically relevant regions. Building on this insight, we present CAPS (Cross-modal Attention as Page Selector), a retrieval framework that utilizes attention mechanisms inside VLMs for page selection. Specifically, CAPS first enhances attention-based retrieval capability with a small amount of contrastive data, then identifies the most effective attention head through expert head selection, and finally employs an adaptive filtering mechanism to obtain an appropriate number of relevant page candidates. Extensive experiments on four long-document benchmarks demonstrate that CAPS outperforms state-of-the-art embedding-based methods in both retrieval precision and downstream DocQA accuracy. Notably, CAPS achieves these gains using less than 10% of the training data required by competing baselines, highlighting the data efficiency of attention-based page retrieval.

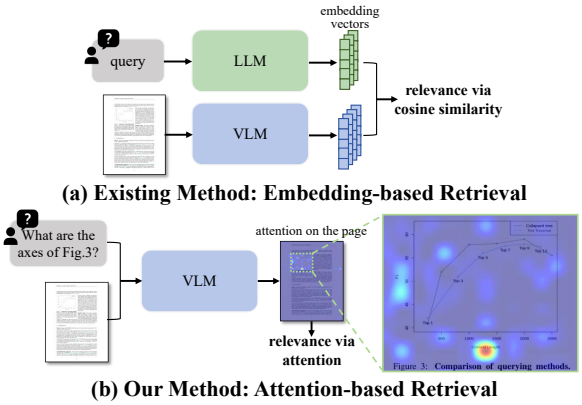

Figure 1: Comparison of page retrieval methods. (a) Embedding-based retrieval matches pages via cosine similarity in a shared vector space. However, similarity is determined by global embedding alignment, without explicit interaction between query and visual cues. (b) Attention-based retrieval utilizes internal attention signals within the VLM for retrieval, enabling fine-grained interactions between query and relevant regions (e.g., locating the axes in figure).

1 Introduction

Document Question Answering (DocQA) serves as a critical benchmark for evaluating document understanding capability (Ding et al., 2022; Zhang et al., 2025a; Suri et al., 2025; Ma et al., 2024a). Traditional Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) pipelines rely on OCR (Wang et al., 2024; Wei et al., 2024, 2025) to linearize documents into text. While effective for plain-text documents, these methods may discard vital visual cues that many DocQA questions hinge on, including complex layouts, charts, and formatting. Recent Visual Language Models offer a promising alternative by directly processing document images (Duan et al., 2025; Hu et al., 2025). However, current VLMs remain constrained by context length and computational cost, making the processing of long documents impractical.

Recent works (Chen et al., 2025a; Huang et al., 2025) adopt page retrieval as a preliminary step for DocQA: given a user query, the retrieval step selects a handful of relevant pages from a document before answering. Prevailing approaches typically rely on embedding-based retrieval. Methods such as ColPali (Faysse et al., 2025) employ VLMs as encoders to embed pages and queries into a shared vector space and compute relevance via cosine similarity (Cho et al., 2024). Building on

this, works such as MDocAgent (Han et al., 2025) and MoLoRAG (Wu et al., 2025) have introduced more precise matching mechanisms. Despite their widespread application, these embedding-based approaches employ dual-encoders to compress user query and multimodal pages into vectors for comparison, limiting their ability to model fine-grained interactions. Furthermore, they rely on large-scale contrastive training datasets to align visual and textual embeddings.

In this work, we revisit the page retrieval process for DocQA: *does a well-trained VLM already know how to retrieve?* Cross-modal attention is explicitly designed to align textual tokens with visual evidence, and recent analyses suggest that attention patterns can expose fine-grained semantic correspondences (Chen et al., 2024; Kaduri et al., 2025). This aligns with human cognitive patterns: when answering a question based on a document, one naturally locates relevant regions before formulating an answer. With a systematic analysis of the zero-shot retrieval capability encoded in VLM attention patterns, we observe that the retrieval signal is already encoded in the cross-modal attention maps.

Inspired by this insight, we propose **CAPS** (**C**ross-modal **A**ttention as **P**age **S**elector), an attention-based framework that unlocks the latent retrieval capability inside VLMs for document page selection. CAPS first enhances the attention-based retrieval capability through contrastive learning, where the model demonstrates substantial improvement in attention-based retrieval capability using only a small amount of contrastive data. Since head-wise analysis reveals that retrieval competence is not uniform across attention heads, we identify the top-performing attention head as the designated retrieval expert head. Finally, we apply an adaptive filtering strategy to select an appropriate number of pages for each query, enabling dynamic page selection.

Specifically, CAPS employs a lightweight VLM (e.g., Qwen2.5-VL-3B (Bai et al., 2025)) as the page retriever for long documents to ensure efficiency. Extensive experiments on four challenging benchmarks, including MMLongBench-Doc (Ma et al., 2024b), LongDocURL (Deng et al., 2025), PaperTab, and FetaTab (Hui et al., 2024), demonstrate that CAPS consistently outperforms existing state-of-the-art methods in both retrieval precision and DocQA accuracy, while requiring less than 10% of training data. In summary, our contributions are as follows:

- We observe that VLM intrinsic cross-modal attention signals can be directly exploited for document page retrieval, offering an alternative to existing approaches.
- We introduce CAPS, a framework that utilizes cross-modal attention in VLMs for document page selection. CAPS enhances the model's attention-based retrieval capability through contrastive learning, and subsequently utilizes the identified expert head, combined with an adaptive filtering strategy, to achieve dynamic page selection.
- Extensive experiments demonstrate that CAPS achieves superior retrieval precision and DocQA accuracy on four challenging long-document benchmarks, outperforming a wide range of existing methods.

2 Related Works

Document Question Answering. The landscape of DocQA has evolved from processing single-page, plain-text documents (Mathew et al., 2021, 2022) to analyzing multi-page, visually-rich contexts (Tito et al., 2023; Ma et al., 2024b; Deng et al., 2025). While Large Language Models (LLMs) traditionally dominated this field, their reliance on OCR introduces significant information loss and error propagation, especially in complex layouts (Kim et al., 2022; Lee et al., 2023). To mitigate this, Vision-Language Models (VLMs) have emerged as a robust, OCR-free paradigm (Ye et al., 2023; Hu et al., 2025). However, despite their superior visual understanding, standard VLMs struggle with the computational burden of long-context documents. Our work addresses this limitation by equipping VLMs with efficient page selection capabilities.

Visual Document Retrieval. RAG enhances generation by integrating external knowledge (Lewis et al., 2020). In the visual domain, traditional CLIP-based methods (Radford et al., 2021) often fail to capture fine-grained document semantics due to vector compression. To address this, pioneering works like ColPali (Faysse et al., 2025) leverage VLMs to implement late-interaction mechanisms, preserving multi-vector representations for finer matching. Building upon this paradigm, subsequent studies have tailored it to specific challenges: M3DocRAG (Cho et al., 2024) extended the approach to comprehensive page-level retrieval

benchmarks; MDocAgent (Han et al., 2025) integrated it into an agentic framework for information filtering combined with text matching; and MoLoRAG (Wu et al., 2025) utilized graph structures to further optimize page selection strategies. However, these methods fundamentally rely on decoupled encoding schemes that typically necessitate massive contrastive training data. In contrast, our approach leverages the VLM's intrinsic cross-modal attention for retrieval, avoiding parameter-heavy external modules.

Attention-based Semantic Relevance. The attention mechanism in transformer-based VLMs computes pairwise compatibility scores between all token positions, enabling flexible information routing across the sequence. The correlation between attention weights and semantic relevance in transformers is well-established. Studies, such as FastV (Chen et al., 2024) and PyramidDrop (Xing et al., 2025), leverage intrinsic attention for tasks like token pruning and visual grounding, demonstrating that attention mechanisms naturally gravitate towards informative regions (Zhang et al., 2025b; Wen et al., 2025). Further analysis of VLMs reveals that intermediate layers typically exhibit stronger semantic alignment between modalities (Kaduri et al., 2025; Jiang et al., 2025; Zhang et al., 2025c). Recent approaches have also successfully utilized attention patterns for retrieval tasks within the textual domain (Ye et al., 2025; Fang et al., 2025). In this work, we systematically apply cross-modal attention for visual document retrieval and further enhance its effectiveness through training.

3 Method

We present CAPS that unlocks the cross-modal attention in vision-language models for document page selection. Figure 2 provides an overview of our framework. This section first formalizes the interpretation of attention weights as relevance scores to probe the intrinsic retrieval capability of models, then proposes a contrastive training strategy to enhance the page retrieval capability of attention at specific architectural locations, and finally introduces an adaptive filtering mechanism to tailor the context size for answer generation.

3.1 Problem Formulation and Attention-based Relevance Scoring

Task Definition. Given a multi-page document $\mathcal{D} = \{p_1, p_2, \ldots, p_m\}$ and a natural language query q, our objective is to identify the subset of pages $\mathcal{P}_{\text{ret}} \subseteq \mathcal{D}$ that contain evidence necessary for answering the query. This retrieved subset subsequently serves as the visual context for a vision-language model to generate the response. Therefore, the size of $\mathcal{P}_{\text{ret}}$ must be constrained within a reasonable range to accommodate the model input capacity. To achieve this, we introduce a quantitative metric, the relevance score, to evaluate the alignment between the query and each document page, which subsequently serves as the core criterion for our retrieval process.

Relevance Score Formulation. To quantify the relevance, we propose a scoring mechanism based on the attention weights within transformer-based VLMs (Vaswani et al., 2017). Formally, consider a VLM processing a multimodal input token sequence $T = \{t_1, t_2, \ldots, t_N\}$ derived from the concatenation of a document page image p and a textual query q. Here, N denotes the total sequence length, and each index i ($1 \leq i \leq N$) represents a distinct position in the sequence, corresponding to either a visual patch or a textual token. In autoregressive architectures, the token at the final position t_N aggregates information from the entire preceding context to parameterize the next-token distribution, making it a natural proxy for holistic query understanding. Consequently, the distribution of its attention weights over the visual tokens reflects the model's implicit assessment of page relevance (Chen et al., 2024; Endo et al., 2025).

At layer l and attention head h, let $\mathbf{q}_{\text{final}}^{(l,h)} \in \mathbb{R}^d$ denote the query vector at the final position and $\mathbf{k}_i^{(l,h)} \in \mathbb{R}^d$ the key vector at position i. The attention weight from the final token to the token at position i is computed as:

$$A_{\text{final}\to i}^{(l,h)} = \frac{\exp\left(\frac{\mathbf{q}_{\text{final}}^{(l,h)} \cdot (\mathbf{k}_i^{(l,h)})^\top}{\sqrt{d}}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathbf{q}_{\text{final}}^{(l,h)} \cdot (\mathbf{k}_j^{(l,h)})^\top}{\sqrt{d}}\right)}. \quad (1)$$

Based on these attention weights, we define the **head-level relevance score** by aggregating the attention weights assigned to visual tokens:

$$s_{\text{head}}^{(l,h)}(p, q) = \sum_{i \in \mathcal{V}_p} A_{\text{final}\to i}^{(l,h)}, \quad (2)$$

where $\mathcal{V}_p$ is the index set of visual tokens corresponding to page p. This score quantifies the proportion of attention mass allocated to the page, serving as our retrieval signal. Furthermore, we define

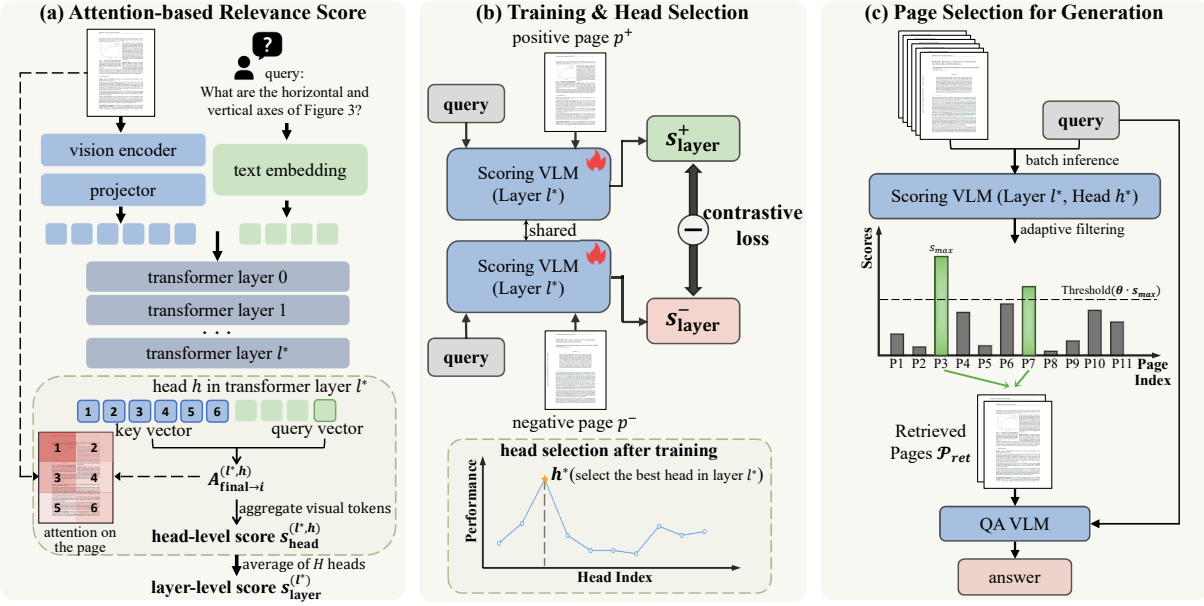

Figure 2: Overview of the CAPS method. (a) Attention-based Relevance Score: we extract the cross-modal attention weights from the final token to visual tokens within the selected layer to quantify page-query relevance. (b) Training & Head Selection: the model is trained via contrastive learning using layer-level relevance scores, and the best head h^* is subsequently selected. (c) Page Selection for Generation: we utilize head-level relevance scores to evaluate all pages, followed by adaptive filtering to obtain the retrieved pages, providing context for generation.

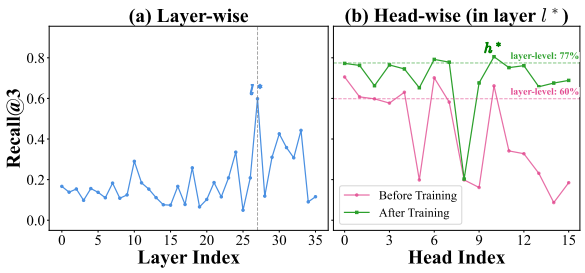

Figure 3: Quantitative analysis of attention-based retrieval capability in Qwen2.5-VL-3B, including layer-wise performance and head-wise performance within the best-performing layer (before and after training).

the **layer-level relevance score**:

$$s^{(l)}_{\text{layer}}(p,q) = \frac{1}{H}\sum_{h=1}^{H} s^{(l,h)}_{\text{head}}(p,q), \tag{3}$$

where H is the number of attention heads per layer. This score reflects the retrieval capability of a specific layer by averaging the head-level relevance scores across all heads.

3.2 Probing Intrinsic Retrieval Capability

Before introducing our training procedure, we conduct a systematic analysis to characterize the zero-shot retrieval capability encoded in VLM attention patterns. This investigation serves dual purposes: validating our core hypothesis that attention encodes retrieval-relevant information, and identifying architectural locations where this capability is most pronounced.

Evaluation Protocol. We evaluate retrieval performance using Recall@K, measuring the fraction of ground-truth evidence pages recovered within the top-K ranked candidates:

$$\text{Recall@}K = \frac{|\mathcal{P}_{\text{top-}K} \cap \mathcal{P}_{\text{gt}}|}{|\mathcal{P}_{\text{gt}}|}, \tag{4}$$

where $\mathcal{P}_{\text{top-}K}$ denotes pages ranked in the top-K positions by relevance score and $\mathcal{P}_{\text{gt}}$ represents ground-truth evidence pages.

Layer-wise Heterogeneity. We first examine how retrieval capability varies across transformer depth using layer-level relevance scores on a validation set. Figure 3 (a) reports layer-wise Recall@K for Qwen2.5-VL-3B (Bai et al., 2025), revealing layer-to-layer heterogeneity. While some layers contribute little to recall, others demonstrate substantially stronger retrieval capability, indicating selective encoding across the architecture. Examining the overall trend, almost all early layers exhibit negligible retrieval capability, consistent with their role in encoding low-level visual and textual features. Performance increases substantially in middle layers and peaks at layer 27, which we hy-

pothesize corresponds to the depth at which cross-modal semantic alignment is maximally developed. We designate this peak layer as l^* for subsequent analysis. Deeper layers show modest degradation, possibly due to increased task-specificity that diminishes general retrieval utility.

Head-wise Heterogeneity. We further analyze individual attention heads within the optimal layer l^*. As shown in Figure 3 (b), head-level performance exhibits substantial variance even within a single layer. Certain heads demonstrate markedly superior retrieval capability compared to the layer average, suggesting functional specialization. This finding carries important implications: naive aggregation across heads may dilute signals from high-performing heads with noise from less discriminative ones.

3.3 Contrastive Training for Retrieval Enhancement

Building on the preceding analysis, we propose a contrastive training procedure designed to enhance the discriminative capability of attention-based retrieval. The objective is to sharpen the model's ability to distinguish evidence pages from distractors by maximizing the score differential between positive and negative samples.

Training Data Construction. We construct training instances as triplets (q, p^+, p^-), where q is a query, p^+ is an evidence page containing information necessary to answer the query, and p^- is a hard negative page sampled from the same document. Sampling high-scoring non-evidence pages from the same document as hard negatives proves more effective than traditional in-batch negative sampling strategies. Further details can be found in Appendix A.2.

Optimization Objective. We employ a margin-based contrastive loss operating on layer-level relevance scores. Let $s^+_{\text{layer}} = s^{(l^*)}_{\text{layer}}(p^+, q)$ and $s^-_{\text{layer}} = s^{(l^*)}_{\text{layer}}(p^-, q)$ denote the relevance scores for positive and negative pages. The loss is defined as:

$$\mathcal{L} = \log\left(1 + \exp\left(\frac{s^-_{\text{layer}} - s^+_{\text{layer}}}{\tau}\right)\right), \quad (5)$$

where τ is a temperature hyperparameter controlling the sharpness of the optimization landscape. Training is conducted on the optimal layer l^*, enabling efficient adaptation by leveraging the best intrinsic zero-shot retrieval capability. Through this supervision signal, we apply full fine-tuning to all upstream parameters.

Expert Head Selection. Contrastive training yields significant improvements in performance. As illustrated in Figure 3 (b), the layer-level relevance score recall improves substantially, but individual heads exhibit widely varying improvements, indicating that different heads possess distinct training potentials for retrieval and suggesting that the training process activates latent retrieval specialization. We designate the head achieving maximum post-training recall as the retrieval expert head, denoted h^*. During inference, we deploy exclusively this head for relevance scoring:

$$s_{\text{inf}}(p, q) = s^{(l^*,h^*)}_{\text{head}}(p, q). \quad (6)$$

This deployment eliminates signal dilution from less discriminative heads while fully leveraging the retrieval capability of the trained model.

3.4 Adaptive Page Selection for Answer Generation

The final component of our framework addresses the integration of attention-based retrieval into a document question-answering pipeline. A critical challenge lies in determining the appropriate number of pages to include in the context: too few may omit essential evidence, while too many may introduce distractors and computational burden. While conventional top-K retrieval is widely adopted, this approach is suboptimal for document QA, where the number of relevant pages varies substantially across queries. For example, single-page evidence queries are contaminated by $K - 1$ distractors.

Adaptive Filtering. We propose an adaptive filtering mechanism that adjusts context size to the score distribution. Let $s_{\max} = \max_{p\in\mathcal{D}} s_{\text{inf}}(p, q)$ denote the maximum relevance score. We introduce a relative threshold $\theta \in (0, 1)$ and define the retrieved page set as:

$$\mathcal{P}_{\text{ret}} = \{p \in \text{top-}K(\mathcal{D}) \mid s_{\text{inf}}(p, q) \geq \theta \cdot s_{\max}\} \quad (7)$$

This formulation combines the computational predictability of top-K selection with adaptive pruning based on relative confidence. When the score distribution is sharply peaked, indicating high confidence in a single page, the threshold aggressively filters low-scoring candidates. When scores are more uniformly distributed, suggesting multi-page relevance, the threshold preserves broader context.

Table 1: Retrieval performance comparison (in %) under the top-K setting. The "Data Scale" indicates the volume of training data used for the base model of each method. Some baseline results are adopted from Wu et al. (2025).

Top-K	Method	Base Model	Data Scale	MMLongBench		LongDocURL	
				Recall	Precision	Recall	Precision
3	M3DocRAG	ColPali (3B)	118k samples	64.17	31.62	67.00	33.78
	MDocAgent (Text)	ColBertv2 (110M)	800k samples	43.21	20.77	58.53	29.33
	MDocAgent (Image)	ColPali (3B)	118k samples	64.74	31.97	66.67	33.62
	MoLoRAG+	ColPali + Qwen2.5-VL	118k + 5k samples	68.87	48.67	68.92	47.53
	CAPS	InternVL2.5-2B	6k samples	71.08	45.47	69.34	41.76
	CAPS	Qwen2.5-VL-3B	6k samples	**72.86**	**50.51**	**70.92**	**49.82**
5	M3DocRAG	ColPali (3B)	118k samples	72.00	22.58	74.32	23.34
	MDocAgent (Text)	ColBertv2 (110M)	800k samples	50.60	15.48	65.41	20.41
	MDocAgent (Image)	ColPali (3B)	118k samples	71.45	22.37	74.60	23.50
	MoLoRAG+	ColPali+Qwen2.5-VL	118k + 5k samples	72.37	45.34	73.69	42.47
	CAPS	InternVL2.5-2B	6k samples	77.14	39.59	75.37	34.77
	CAPS	Qwen2.5-VL-3B	6k samples	**78.89**	**45.61**	**75.76**	**44.31**

Answer Generation. The retrieved pages are concatenated with the query and provided to the VLM for response generation:

$$\mathbf{y} = \text{VLM}(\mathcal{P}_{\text{ret}}, q) \tag{8}$$

This design effectively achieves query-adaptive context selection without requiring explicit supervision for context size, thereby enabling the system to automatically balance precision and coverage based on retrieval confidence.

4 Experiments

4.1 Experimental Setup

Models. We employ Qwen2.5-VL-3B (Bai et al., 2025) and InternVL2.5-2B (Chen et al., 2025b) for retrieval, assessing downstream QA generalization on LLaVA-NeXT-7B (Li et al., 2024), DeepSeek-VL-16B (Lu et al., 2024), and the Qwen2.5-VL.

Benchmarks. Evaluations cover four benchmarks: (1) MMLongBench-Doc (Ma et al., 2024b), which requires multi-page reasoning with high information density; (2) LongDocURL (Deng et al., 2025), featuring extensive lengths but lower density; (3) PaperTab and (4) FetaTab (Hui et al., 2024), involving complex reasoning but fewer total pages. We report generalized accuracy for the former two and binary accuracy for the latter.

Baselines. We compare our approach against three categories of baselines: (1) Text RAG: Traditional pipelines using OCR for text extraction followed by text-based retrieval and generation. We also include the recent DeepSeek-OCR (Wei et al., 2025) as a stronger text-based baseline, which is capable of preserving visual information such as layout. Further details can be found in Appendix A.1. (2) VLMs Direct: Direct input of full documents into the model without prior retrieval. (3) Retrieval-based VLMs: Representative recent vision-based retrieval methods. These approaches are all based on vector similarity retrieval.

Implementation Details. The scoring VLM is trained using the AdamW optimizer with a learning rate of 3×10^{-6} and a batch size of 32. The temperature τ in the contrastive loss is set to 0.1. During inference, we set the relative threshold θ for adaptive page filtering to 0.3.

4.2 Main Results

Retrieval Performance. Table 1 compares the page retrieval performance of retrieval-based VLM methods. We report results on MMLongBench and LongDocURL, as these datasets provide ground-truth evidence page labels. The results demonstrate that CAPS outperforms state-of-the-art methods in Recall and Precision. Specifically, CAPS with Qwen2.5-VL-3B as base model achieves 78.89% on MMLongBench-Doc in terms of Recall@5, outperforming the best baseline by 6 percentage points, despite the baseline also utilizing the same Qwen2.5-VL-3B to enhance its retrieval capability. This indicates that our method more effectively unlocks and leverages the latent retrieval capability of the VLM backbone compared to previous implementation strategies. Moreover, our training method is data-efficient, allowing for low-cost method transfer as VLMs evolve.

Document Question-Answering Performance. Table 2 reports downstream DocQA results using top-3 retrieved pages, a setting balancing performance and efficiency. CAPS utilizes Qwen2.5-VL-3B as the base model. We highlight the following

Table 2: Document Question-Answering Performance Comparison (in %) under the retrieved top-3 setting. Some baseline results are adopted from Wu et al. (2025).

DocQA Model	Method	MMLongBench	LongDocURL	PaperTab	FetaTab	Avg.
Text-based						
Qwen2.5-7B	Text RAG	25.52	27.93	12.72	40.06	26.56
	DSOCR Text RAG	**30.67**	**30.87**	**17.05**	**45.57**	**31.04**
DeepSeek-V3	Text RAG	29.82	34.73	17.05	52.36	33.49
	DSOCR Text RAG	**32.81**	**35.66**	**22.65**	**52.76**	**35.97**
Vision-based						
LLaVA-NeXT-7B	Direct	7.15	10.78	3.05	11.61	8.15
	M3DocRAG	10.10	13.85	5.34	13.98	10.82
	MoLoRAG+	9.47	13.58	5.60	13.48	10.53
	CAPS (Ours)	**16.08**	**20.22**	**6.62**	**15.35**	**14.57**
DeepSeek-VL-16B	Direct	8.40	14.72	6.11	16.14	11.34
	M3DocRAG	18.12	29.60	7.89	27.07	20.67
	MoLoRAG+	25.47	37.21	10.94	41.54	28.79
	CAPS (Ours)	**25.59**	**40.27**	**12.21**	**49.31**	**31.85**
Qwen2.5-VL-3B	Direct	26.65	24.89	25.19	51.57	32.08
	M3DocRAG	29.11	44.40	24.68	53.25	37.86
	MoLoRAG+	32.47	45.27	**27.23**	58.76	40.93
	CAPS (Ours)	**34.86**	**45.60**	25.45	**62.60**	**42.13**
Qwen2.5-VL-7B	Direct	32.77	26.38	29.77	64.07	38.25
	M3DocRAG	36.18	49.03	28.50	63.78	44.37
	MoLoRAG+	41.01	51.85	31.04	69.19	48.27
	CAPS (Ours)	**44.69**	**53.78**	**36.13**	**71.36**	**51.49**
Mixed						
MDocAgent(LLaMA3.1-8B+Qwen2.5-VL-7B)		38.53	46.91	30.03	66.34	45.45

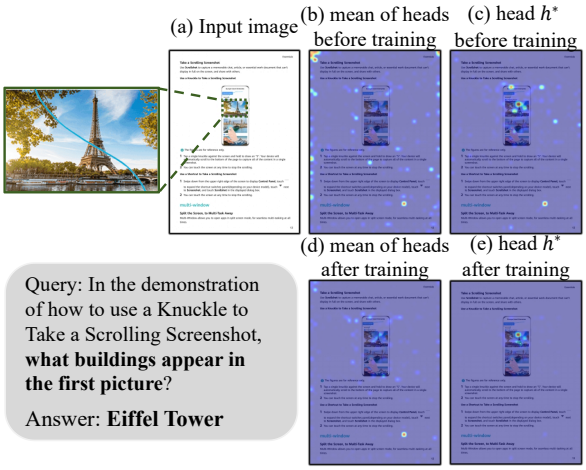

Figure 4: Visualization of attention heatmaps from Qwen2.5-VL-3B. The comparison between the expert head h^* and the layer mean highlights the focusing effect. Comparing before and after training states demonstrates that noise is significantly suppressed and the focus becomes more distinct after training.

three observations:

(1) Vision-based retrieval surpasses text-based RAG. While DeepSeek-OCR improves upon traditional text RAG by converting layouts to text, it cannot losslessly encode all visual elements. Consequently, even large-scale models like DeepSeek-V3 (DeepSeek-AI et al., 2025) lag behind VLMs due to this inherent modality gap.

(2) Retrieval is critical for long contexts. Retrieval-based methods consistently outperform direct inference. This is crucial for benchmarks like MMLongBench and LongDocURL where documents often exceed context windows, but it also benefits shorter tasks (PaperTab/FetaTab) by effectively filtering out noise to focus on valid evidence.

(3) CAPS achieves superior generalization capabilities. High-quality retrieval by CAPS consistently boosts DocQA accuracy across LLaVA, DeepSeek, and Qwen series. Notably, combining CAPS with Qwen2.5-VL-7B yields the most robust performance, ultimately surpassing even the strongest baseline by an average of 3%.

Attention Visual Analysis. We visualize cross-modal attention in Figure 4 to further analyze the attention-based retrieval capability. We observe two phenomena: (1) Expert head effect: In evidence pages, compared to the layer mean, the expert head focuses significantly more sharply on the relevant region (the Eiffel Tower in the figure). (2) Training effect: In evidence pages, contrastive training effectively suppresses background noise, concentrating attention primarily on evidence regions.

Table 3: Ablation study on the page filtering threshold θ. Experiments are conducted on the MMLongBench-Doc using Qwen2.5-VL-3B for retrieval and Qwen2.5-VL-7B for DocQA, with the top-K set to 3.

θ	Recall	Precision	Avg. Pages	Acc
without threshold				
-	73.64	36.57	3.00	43.85
relative threshold				
0.10	73.64	37.82	2.95	43.64
0.30	**72.86**	**50.51**	**2.49**	**44.69**
0.50	69.94	58.27	2.13	43.26
absolute threshold				
0.05	73.41	37.14	2.97	43.82
0.10	69.16	42.90	2.45	42.85
0.15	62.74	48.63	1.87	42.05

Table 4: Ablation study on training objectives using Qwen2.5-VL-3B. Results are reported on MMLongBench-Doc without adaptive filtering.

Score in Training	Score in Inference	Recall@3	Recall@5
training-free			
-	$s_{\text{head}}^{(l^*,6)}$	60.44	70.92
	$s_{\text{head}}^{(l^*,10)}$	52.67	63.01
training with layer-level relevance score			
$s_{\text{layer}}^{(l^*)}$	$s_{\text{head}}^{(l^*,6)}$	71.25	76.74
	$s_{\text{head}}^{(l^*,10)}$	73.64	80.50
training with head-level relevance score			
$s_{\text{head}}^{(l^*,6)}$	$s_{\text{head}}^{(l^*,6)}$	71.14	76.79
$s_{\text{head}}^{(l^*,10)}$	$s_{\text{head}}^{(l^*,10)}$	73.88	80.97

This validates the effectiveness of our method from another perspective. Further analysis can be found in Appendix B.3.

4.3 Ablation Study

Impact of threshold θ. We compare relative ($s \geq \theta \cdot s_{\max}$) and absolute ($s \geq \theta$) thresholding on MMLongBench-Doc. Table 3 shows that thresholding significantly boosts Precision and reduces the average number of pages input to the model for QA with marginal Recall degradation. The relative threshold outperforms the absolute one by better adapting to case difficulty, providing a superior Recall-Precision trade-off.

Impact of supervision signals. In Table 4, we analyze the effectiveness of different supervision signals, specifically modifying the relevance score within the loss function. The results demonstrate

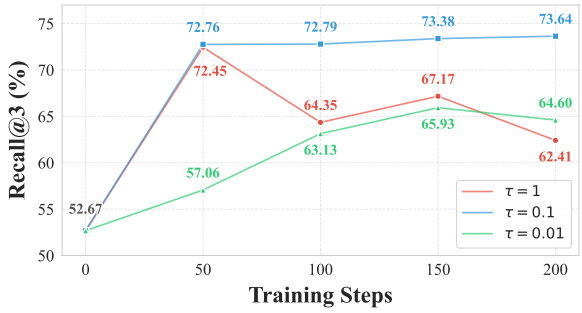

Figure 5: Ablation study on the hyperparameter τ using Qwen2.5-VL-3B for retrieval. The training curves of Recall@3 on MMLongBench-Doc are shown for different τ values, evaluated without adaptive filtering.

that supervising the layer-level relevance score effectively activates the retrieval potential of attention heads. The performance achieved via layer-level training is nearly identical to the upper bound established by directly supervising specific head-level scores (e.g., 73.64% vs. 73.88% for head 10). Crucially, identifying the optimal retrieval head prior to training is challenging, as the relative capability of heads may shift during optimization. Layer-level supervision addresses this uncertainty by allowing the most capable head to emerge naturally.

Impact of hyperparameter τ in loss function. We investigate the effect of the hyperparameter τ on training using the MMLongBench-Doc benchmark. As shown in Figure 5, the choice of τ significantly impacts training stability. A large τ leads to unstable training. Conversely, a small τ results in sluggish convergence and suboptimal performance. This hyperparameter facilitates smoother and more stable training.

5 Conclusion

In this paper, we propose CAPS, a framework that unlocks the cross-modal attention in vision-language models for document page selection. By performing contrastive training, identifying an expert head to compute relevance scores, and employing an adaptive filtering strategy, our method applies attention-based page retrieval capability to document understanding. Extensive experiments demonstrate that CAPS achieves SOTA performance among page retrieval methods across four long-document benchmarks while demonstrating high training data efficiency. This highlights the superiority of intrinsic attention as a retrieval signal for long-document understanding.

Limitations

The effectiveness of our attention-based retrieval is inherently contingent upon the visual understanding capabilities of the VLM backbone. Earlier VLM architectures with fixed-resolution encoders may struggle to resolve fine-grained text or dense charts in high-resolution documents, potentially limiting the performance of our method. However, recent VLMs are increasingly equipped with mechanisms to handle high-resolution inputs, which significantly mitigates this issue. Furthermore, our current approach operates at the page level, aligning with current document retrieval and question-answering benchmarks. While this granularity effectively filters out irrelevant contexts, future applications might benefit from finer-grained retrieval, such as selecting specific document regions or patches. We acknowledge that region-level retrieval offers higher precision but introduces significant challenges regarding context fragmentation and methodological complexity. Effectively addressing these complexities within current pipeline designs remains challenging and represents an open problem for the broader community.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan Rossi, Changyou Chen, and Tong Sun. 2025a. Sv-rag: Lora-contextualizing adaptation of mllms for long document understanding. In *International Conference on Representation Learning*, volume 2025, pages 4447–4462.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *Preprint*, arXiv:2411.04952.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. 2025. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1135–1159, Vienna, Austria. Association for Computational Linguistics.

Yihao Ding, Zhe Huang, Runlin Wang, YanHang Zhang, Xianru Chen, Yuzhong Ma, Hyunsuk Chung, and Soyeon Caren Han. 2022. V-doc: Visual questions answers with documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21492–21498.

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, Jifeng Dai, and Wenhai Wang. 2025. Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4026–4037.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. 2025. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22826–22835.

Yixiong Fang, Tianran Sun, Yuling Shi, and Xiaodong Gu. 2025. Attentionrag: Attention-guided context pruning in retrieval-augmented generation. *Preprint*, arXiv:2503.10720.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Representation Learning*, volume 2025, pages 61424–61449.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *Preprint*, arXiv:2503.13964.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mPLUG-DocOwl2: High-resolution

compressing for OCR-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834, Vienna, Austria. Association for Computational Linguistics.

De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. 2025. Frag: Frame selection augmented generation for long video and long document understanding. *Preprint*, arXiv:2504.17447.

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. Uda: a benchmark suite for retrieval augmented generation in real-world document analysis. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25004–25014.

Omri Kaduri, Shai Bagon, and Tali Dekel. 2025. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14558.

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.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18893–18912. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

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

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2024a. Visa: Retrieval augmented generation with visual source attribution. *Preprint*, arXiv:2412.14457.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems*, volume 37, pages 95963–96010. Curran Associates, Inc.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6088–6109, Albuquerque, New Mexico. Association for Computational Linguistics.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and

Conghui He. 2024. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *Preprint*, arXiv:2409.01704.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15537–15549, Vienna, Austria. Association for Computational Linguistics.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. 2025. MoLoRAG: Bootstrapping document understanding via multi-modal logic-aware retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14056, Suzhou, China. Association for Computational Linguistics.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. 2025. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14593–14603.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, Singapore. Association for Computational Linguistics.

Xiaoju Ye, Zhichun Wang, and Jingyuan Wang. 2025. Infinite retrieval: Attention enhanced llms in long-context processing. *Preprint*, arXiv:2502.12962.

Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. 2025a. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *Preprint*, arXiv:2410.21169.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2025b. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*.

Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. 2025c. Flexselect: Flexible token selection for efficient long video understanding. *Preprint*, arXiv:2506.00993.

A Implementation Details

A.1 Baseline Implementations

Text RAG Based on DeepSeek-OCR. We acknowledge the potential impact of the recently released DeepSeek-OCR on the field of document understanding. Consequently, we adopt it as a superior alternative to traditional OCR engines, establishing a stronger baseline for Text RAG, denoted as **DSOCR Text RAG**. DeepSeek-OCR offers capabilities beyond standard character recognition, specifically in layout preservation and visual element parsing. We primarily utilize the prompt "`<image>\n<|grounding|>Convert the document to markdown.`". This mode effectively retains the document's layout structure and extracts embedded images (crops), thereby generating a textual representation that maximally preserves visual information. Furthermore, to ensure a pure-text input for the RAG pipeline, we employ the "`<image>\nParse the figure.`" prompt for visual elements. This mode converts structured graphics (e.g., tables, line charts) into formal textual representations and provides natural language descriptions for natural images. We substitute the image placeholders in the generated markdown with these detailed textual descriptions to form the final corpus. Additionally, to handle rare instances where layout-aware processing fails, we implement a fallback mechanism using the "`<image>\nFree OCR.`" prompt. This corresponds to a traditional OCR mode that extracts text without preserving layout or parsing figures. The resulting text is then processed using the standard retrieval pipeline, consistent with other text-based baselines. This comprehensive approach significantly enhances the baseline's ability to address queries related to document layout and charts.

Other Baseline Implementations. Regarding other baselines, we largely follow the implementation methods provided in Wu et al. (2025). For single-image VLMs such as LLaVA-NeXT-7B, we horizontally concatenate multiple page images into a single input to simulate multi-page processing. For VLMs that support arbitrary input resolutions, such as the Qwen series, we restrict the maximum number of visual tokens to 2048 per image, considering the high-resolution nature of document images. For the "Direct" inference baseline (processing full documents without retrieval), we set a maximum input capacity of 30 pages due to context window constraints.

A.2 Training Data Construction

Data Source. Our training dataset is constructed by aggregating high-quality supervision signals from two distinct sources. First, we utilize the dataset meticulously annotated in the MoLoRAG task, which provides fine-grained query-page relevance scores ranging from 1 to 5. From this set, we select pages with relevance scores of 4 and 5 as positive evidence samples. Given the limited scale of this manually annotated dataset, we supplement it with data from the MP-DocVQA benchmark, which also provides ground-truth evidence page labels. In total, we construct a dataset of 6,400 samples. It is worth noting that we initially explored scaling the training data to approximately 10k samples and extending the training duration to multiple epochs. However, empirical monitoring of intermediate checkpoints revealed that the retrieval performance saturated rapidly. Specifically, we observed negligible gains beyond approximately 200 training steps (with a batch size of 32). Consequently, we finalized the dataset at 6.4k samples, confirming that a compact, high-quality dataset is sufficient to effectively activate the model's latent retrieval capabilities without the need for extensive data scaling or prolonged training.

Hard Negative Mining. Regarding negative sampling, we diverge from standard contrastive learning approaches that typically rely on in-batch negatives. We argue that distinguishing between pages within the same document context is significantly more challenging and beneficial than distinguishing between random documents. Therefore, we implement a model-guided hard negative mining strategy. Specifically, for each query, we compute the relevance scores of all pages within the target document using the model zero-shot capability (prior to training). We then identify the page with the highest relevance score that is not labeled as a ground-truth evidence page to serve as the hard negative. This approach forces the model to discriminate between semantically or visually similar pages within the same document, thereby facilitating the learning of more fine-grained retrieval capabilities.

A.3 Other Details

All experiments were conducted on NVIDIA A800 GPUs. The training process is computationally

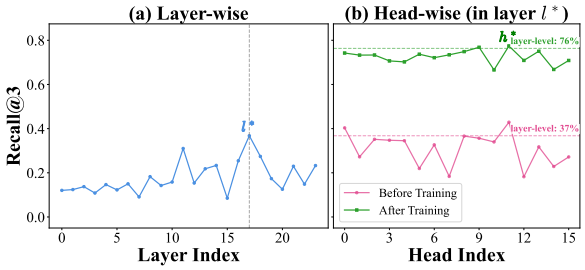

Figure 6: Quantitative analysis of attention-based retrieval capability in InternVL2.5-2B, including layer-wise performance and head-wise performance within the best-performing layer (before and after training).

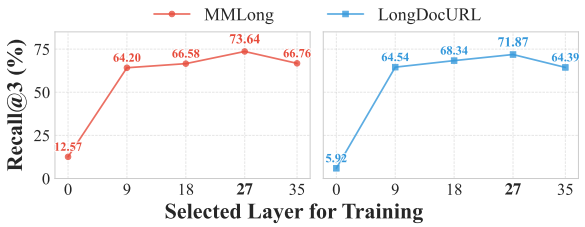

Figure 7: Ablation study on the selection layer. We compare the Recall@3 performance on MMLongBench-Doc and LongDocURL when training on different layers of Qwen2.5-VL-3B, evaluated without the adaptive filtering.

efficient; a single experimental run requires approximately 5 hours on one GPU. For optimization, we employ the AdamW optimizer with a global batch size of 32 and a learning rate of 3×10^{-6}. Regarding the specific hyperparameters introduced in our method, the contrastive loss temperature τ is set to 0.1, and the adaptive filtering threshold θ is set to 0.3. During the inference and evaluation phase, standard evaluation protocols across the utilized benchmarks involve an answer extraction step to normalize model outputs. We utilize the `qwen3-max` API as the extractor for this purpose to ensure accurate answer parsing.

B Additional Experimental Results

B.1 Analysis on InternVL

To evaluate the generalization capability of our proposed framework, we conduct identical experiments using the InternVL2.5-2B architecture, employing the same training and evaluation hyperparameters. Figure 6 illustrates the quantitative analysis of layer-wise and head-wise retrieval performance on the validation set. Our observations confirm that the findings reported for Qwen2.5-VL are generalizable across different VLM architectures. For InternVL2.5-2B, which consists of a 25-layer Transformer, we observe that the intrinsic retrieval capability peaks at the 17th layer ($l^* = 17$). Subsequent contrastive training yields a substantial improvement in attention-based retrieval performance. We identify the 11th attention head ($h^* = 11$) within this layer as the expert head for final evaluation.

The retrieval results on the MMLongBench-Doc and LongDocURL benchmarks are presented in Table 1. Although the absolute performance is slightly lower than that of Qwen2.5-VL-3B, the method still achieves highly competitive results. This demonstrates the robust generalization of our approach. Furthermore, given the low computational cost of training, our framework can be efficiently transferred to more advanced VLMs as they evolve. We anticipate that applying our method to stronger backbones will yield even superior retrieval performance, benefitting from their enhanced intrinsic visual understanding capabilities.

Table 5: Ablation study on the percentage of visual tokens used for relevance scoring. Experiments are conducted on MMLongBench-Doc using Qwen2.5-VL-3B in the training-free setting without adaptive filtering. The ratio r denotes the ratio of top-scoring tokens summed to calculate the relevance score, where "100%" represents our standard method of aggregating all visual tokens.

Score in Inference	Ratio r	Recall@3	Recall@5
$s_{\text{layer}}^{(l^*)}$	1%	27.33	36.17
	10%	38.02	48.86
	100%	42.25	53.37
$s_{\text{head}}^{(l^*,6)}$	1%	58.97	69.28
	10%	60.72	70.45
	100%	60.44	70.92
$s_{\text{head}}^{(l^*,10)}$	1%	50.03	59.32
	10%	54.39	62.28
	100%	52.67	63.01

B.2 Additional Ablation Study

Effectiveness of Layer Selection. To validate the rationale behind selecting the optimal layer l^* based on zero-shot performance, we investigate the impact of applying contrastive training to layers at various depths, rather than exclusively targeting the layer with the highest initial score. As shown in Figure 7, we track the post-training performance of the best head within each selected layer of Qwen2.5-VL-3B. The results demonstrate a strong consistency: the layer that exhibits the highest capabil-

ity in the training-free setting remains the optimal choice after training. Specifically, early layers yield poor results due to insufficient semantic interaction, while performance peaks at the deeper layers before declining in the final layers. This observation confirms that the intrinsic retrieval capability is a reliable indicator of training potential, thereby justifying the effectiveness of our layer selection strategy.

Effectiveness of Token Aggregation Ratio r. Recall that in the main text, the relevance score is defined by aggregating attention weights over the complete set of visual tokens $\mathcal{V}_p$ (Eq. 2). In this ablation study, we introduce a ratio hyperparameter r to investigate whether the retrieval signal is dominated by sparse salient regions. We define a subset $\mathcal{V}_p^{(r)} \subset \mathcal{V}_p$ consisting of the top-r percent of tokens ranked by attention weight, and restrict the summation range in Eq. 2 from the full set $\mathcal{V}_p$ to this subset $\mathcal{V}_p^{(r)}$.

As shown in Table 5, narrowing the summation scope to strictly high-confidence tokens (e.g., $r = 1\%$ or 10%) does not yield significant performance gains. This indicates that valid retrieval signals typically involve contributions from every token, rather than being concentrated on a few extreme pixels; the varying magnitudes of improvement merely result in highlighted regions during visualization. Furthermore, from an optimization perspective, dynamically selecting $\mathcal{V}_p^{(r)}$ necessitates a hard sorting operation, which is inherently non-differentiable and hinders gradient propagation. Although potential solutions exist, we can entirely circumvent this issue. Therefore, we retain the summation over the full set $\mathcal{V}_p$ as our standard configuration, ensuring both superior performance and training stability.

B.3 Visualization and Analysis

We visualize additional examples of cross-modal attention in Figure 8. We observe three key phenomena: (1) Expert head effect: In evidence pages, compared to the layer mean, the expert head focuses significantly more sharply on the relevant region in most cases, whereas the layer mean attention is dispersed as it aggregates highlights from all heads. (2) Training effect: Compared to the pre-training stage, post-training results demonstrate effective suppression of background noise in evidence pages, concentrating attention on evidence regions. This indicates that training enhances the model's discriminative capability, preventing excessive attention allocation to non-evidence areas. (3) Attention sink: All irrelevant pages exhibit a static attention sink pattern, focusing primarily on initial tokens or boundary tokens due to the lack of semantic alignment. This resembles the inherent attention sink phenomenon in LLMs, which we visualize here within the context of document pages.

Table 6: Quantitative comparison of recall on MMLongBench-Doc before and after training, based on Equation 9 using the attention weights of head h^*.

Setting	Recall@3	Recall@5
Before Training	52.67	63.01
After Training	67.44	75.11

To quantitatively validate these visual observations, we leverage the distinct "attention sink" patterns found in irrelevant pages. Let $\mathbf{P}_i$ denote the flattened attention weight vector from the final token to all visual tokens for the i-th page (refer to Eq.1), and let $\bar{\mathbf{P}}$ be the mean attention vector computed across all pages. Here, $\bar{\mathbf{P}}$ approximates the static, query-agnostic background bias (i.e., the attention sink). We define a variance-based score, s_{var}, to quantify the deviation from this static pattern:

$$s_{\text{var}} = \text{Var}(\mathbf{P}_i - \bar{\mathbf{P}}) \tag{9}$$

Intuitively, a higher s_{var} indicates that the page exhibits unique, query-specific focal points that significantly diverge from the background noise, whereas a low score suggests the page is dominated by the static attention sink. We utilize this metric to rank pages for retrieval. As presented in Table 6, the substantial improvement in Recall after training confirms our visual findings: the contrastive training effectively enhances the signal-to-noise ratio, allowing the expert head to suppress static biases and concentrate sharply on relevant evidence. However, it is important to note that this is primarily a verification experiment for our visualization analysis. For the actual retrieval task, simply aggregating all visual attention weights (as proposed in our main method) yields superior performance compared to this variance-based metric.

B.4 Comprehensive Retrieval Performance

In Table 7, we present a comprehensive evaluation of the retrieval performance using an extended set of metrics: Recall, Precision, NDCG, and MRR.

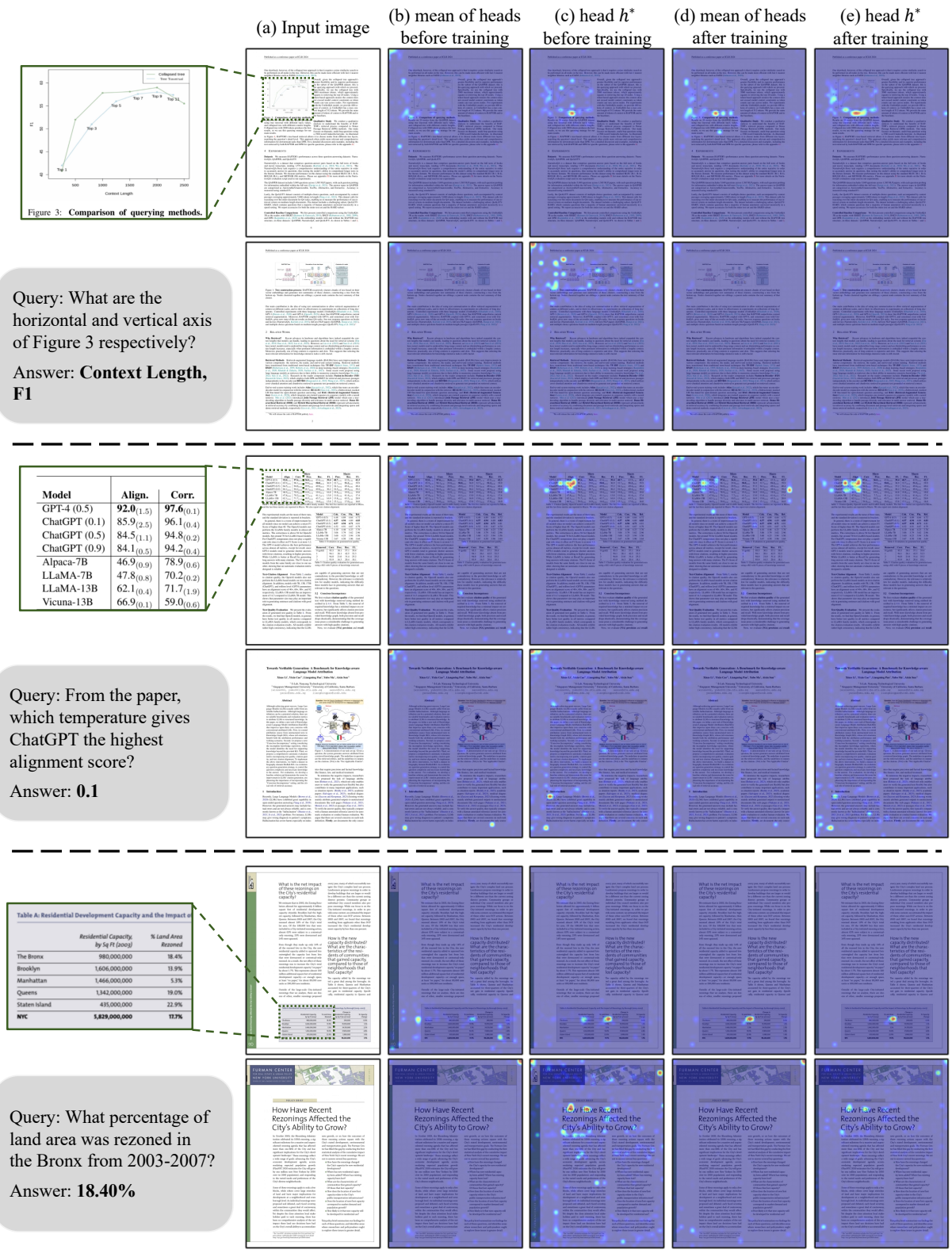

Figure 8: Visualization of additional attention heatmap cases from Qwen2.5-VL-3B. For each case, the top row displays an evidence page, while the bottom row shows an irrelevant page. We provide zoomed-in views of the evidence regions within the evidence pages. We visualize the layer mean and expert head attention heatmaps both before and after training. For the evidence page, compared to the layer mean (cols. b, d), the expert head h^* (cols. c, e) focuses more sharply on the answer region; after training (col. e), noise is suppressed compared to before training (col. c). Conversely, the irrelevant pages below exhibit a static attention sink pattern, with high-attention regions tending to be distributed along the page edges.

The results reported for our method are based on the Qwen2.5-VL-3B backbone. As demonstrated in the table, our approach consistently outperforms comparative baselines across all metrics, validating the robustness and ranking quality of the trained attention signal. To facilitate interpretation, we provide detailed definitions of the evaluation metrics used.

(1) Recall quantifies the coverage capability of the retrieval system. It measures the proportion of actual ground-truth evidence pages that the model successfully captures within its retrieved set.

$$\text{Recall} = \frac{|\mathcal{P}_{\text{ret}} \cap \mathcal{P}_{\text{gt}}|}{|\mathcal{P}_{\text{gt}}|} \tag{10}$$

A high recall indicates that the system effectively minimizes information loss, ensuring that critical evidence required for downstream reasoning is not omitted.

(2) Precision evaluates the purity of the retrieved candidates. It represents the signal-to-noise ratio within the model's output by calculating the fraction of retrieved pages that are truly relevant.

$$\text{Precision} = \frac{|\mathcal{P}_{\text{ret}} \cap \mathcal{P}_{\text{gt}}|}{|\mathcal{P}_{\text{ret}}|} \tag{11}$$

This metric reflects the model's efficiency in filtering out distractors. High precision implies that the downstream generation model receives a concise context with minimal irrelevant noise.

(3) NDCG (Normalized Discounted Cumulative Gain) assesses the quality of the ranking order. Unlike binary metrics that treat all positions equally, NDCG incorporates position bias, acknowledging that evidence appearing earlier in the list is more valuable. It rewards algorithms that prioritize relevant pages by placing them at the top of the ranking list, aligning with the preference for immediate information access.

(4) MRR (Mean Reciprocal Rank) measures the efficiency of identifying the first piece of valid evidence. It is determined by the reciprocal rank of the first correct match in the sorted list:

$$\text{MRR} = \frac{1}{\text{rank}_{\text{first}}} \tag{12}$$

This metric serves as an indicator of system responsiveness, reflecting how deep into the candidate list a user or model must search to encounter the first correct evidence page.

B.5 Fine-grained Performance Analysis

To thoroughly investigate the model's capabilities across different scenarios, we conducted a detailed analysis on the MMLongBench-Doc and LongDocURL benchmarks using the Top-3 retrieval setting. The results are presented in Table 8 and Table 9.

We categorize queries into two dimensions to evaluate distinct aspects of the model. First, the Evidence Source dimension classifies queries based on the visual modality required for the answer, including text (TXT), layout (LAY), charts (CHA), tables (TAB), and figures (FIG). This helps us assess the model's versatility in handling diverse multimodal elements. Second, the Evidence Page dimension defines the scope of information needed, distinguishing between single-page (SIN) and multi-page (MUL) reasoning. We also include an "unanswerable" (UNA) category for MMLongBench-Doc, which involves negative samples where the document contains no valid answer. This category specifically tests the model's ability to avoid hallucinations by correctly identifying the absence of information. We report both Accuracy (Acc) for semantic similarity and Exact Match (EM) for strict correctness. As shown in the tables, our method maintains robust performance across these diverse reasoning scopes and visual modalities.

B.6 Extended Analysis on Retrieval Settings

We further extend our evaluation to include Top-1 and Top-5 retrieval settings to analyze how context size impacts downstream QA performance. The detailed comparisons are provided in Table 10 (Top-1) and Table 11 (Top-5).

Our findings highlight an interesting trade-off between retrieval precision and context length, which varies by model architecture. For advanced VLMs capable of handling multi-image inputs (e.g., Qwen2.5-VL), the Top-3 setting generally offers the optimal balance, providing sufficient context for multi-page reasoning without introducing excessive noise. Conversely, for models that are less optimized for multi-page inputs or require image concatenation (e.g., LLaVA-NeXT-7B and DeepSeek-VL-16B), the Top-1 setting often yields superior results by minimizing distraction from irrelevant pages. Regardless of the specific K value (Top-1, Top-3, or Top-5), our method consistently outperforms baselines, demonstrating the robustness of our attention-based retrieval mechanism.

Table 7: Comprehensive retrieval performance comparison (in %) on MMLongBench-Doc and LongDocURL. Our method uses Qwen2.5-VL-3B as the backbone. Some baseline results are adopted from Wu et al. (2025).

Top-K	Method	MMLongBench				LongDocURL			
		Recall	Precision	NDCG	MRR	Recall	Precision	NDCG	MRR
1	M3DocRAG	43.31	56.67	56.67	56.67	46.84	64.66	64.66	64.66
	MDocAgent (Text)	29.30	38.99	38.99	38.99	42.03	58.37	58.37	58.37
	MDocAgent (Image)	43.79	57.49	57.49	57.49	46.80	64.57	64.57	64.57
	MoLoRAG+	51.32	66.86	66.86	66.86	50.82	70.08	70.08	70.08
	CAPS (Ours)	**53.41**	**69.20**	**69.20**	**69.20**	**51.35**	**70.12**	**70.12**	**70.12**
3	M3DocRAG	64.17	31.62	54.13	65.36	67.00	33.78	58.23	72.51
	MDocAgent(Text)	43.21	20.77	37.13	45.26	58.53	29.33	54.12	65.28
	MDocAgent(Image)	64.74	31.97	54.75	66.12	66.67	33.62	58.26	72.47
	MoLoRAG+	68.87	48.67	64.49	73.50	68.92	47.53	64.90	77.14
	CAPS (Ours)	**72.86**	**50.51**	**67.07**	**76.27**	**70.92**	**49.82**	**65.92**	**78.22**
5	M3DocRAG	72.00	22.58	54.06	66.92	74.32	23.34	58.05	73.83
	MDocAgent(Text)	50.60	15.48	37.19	46.98	65.41	20.41	53.97	66.55
	MDocAgent(Image)	71.45	22.37	54.58	67.53	74.60	23.50	58.06	73.90
	MoLoRAG+	72.37	45.34	64.36	73.97	73.69	42.47	64.74	77.89
	CAPS (Ours)	**78.89**	**45.61**	**66.98**	**77.25**	**75.76**	**44.31**	**65.78**	**79.03**

Table 8: Fine-grained performance analysis on MMLongBench-Doc under the retrieved top-3 setting. Some baseline results are adopted from Wu et al. (2025).

DocQA Model	Method	Evidence Source					Evidence Page			Acc	EM
		TXT	LAY	CHA	TAB	FIG	SIN	MUL	UNA		
Text-based											
Qwen2.5-7B	Text RAG	17.96	10.69	10.35	15.42	9.58	17.72	9.34	70.40	25.52	23.29
	DSOCR Text RAG	**24.36**	**17.99**	**20.07**	**25.34**	**11.28**	**28.26**	**9.64**	**72.65**	**30.67**	**27.26**
DeepSeek-V3	Text RAG	25.37	4.38	19.60	22.23	13.34	24.85	**13.03**	**69.06**	29.82	26.62
	DSOCR Text RAG	**26.77**	**17.85**	**22.47**	**29.44**	**14.39**	**32.56**	11.62	68.61	**32.81**	**29.21**
Vision-based											
LLaVA-NeXT-7B	Direct	6.54	4.38	2.12	1.53	7.38	4.17	5.07	16.59	7.15	5.73
	M3DocRAG	8.74	8.54	1.87	6.59	11.72	7.27	8.55	**17.49**	10.10	8.23
	MoLoRAG+	7.49	8.08	2.89	2.49	11.24	7.86	6.25	16.59	9.41	7.30
	CAPS (Ours)	**14.39**	**15.46**	**8.98**	**7.59**	**21.17**	**18.98**	**11.57**	16.14	**16.08**	**12.20**
DeepSeek-VL-16B	Direct	8.86	13.63	5.23	6.57	13.39	9.86	9.36	3.14	8.40	6.01
	M3DocRAG	19.75	25.02	18.38	14.31	27.55	25.91	15.84	3.59	18.12	13.49
	MoLoRAG+	27.58	32.67	21.56	23.33	**34.45**	39.40	18.74	**4.04**	25.47	19.59
	CAPS (Ours)	**28.86**	**32.80**	**22.32**	**26.01**	33.85	**39.68**	**19.68**	3.14	**25.59**	**19.96**
Qwen2.5-VL-3B	Direct	34.11	29.56	24.72	23.37	33.75	36.30	23.42	9.87	26.65	20.98
	M3DocRAG	35.11	28.06	24.04	25.58	32.23	39.62	20.62	18.83	29.11	23.66
	MoLoRAG+	38.29	**33.53**	26.48	29.39	35.34	45.12	23.17	**19.28**	32.47	26.52
	CAPS (Ours)	**41.56**	30.91	**29.95**	**37.79**	**37.73**	**48.77**	**26.53**	18.39	**34.86**	**28.37**
Qwen2.5-VL-7B	Direct	37.14	27.26	28.00	25.52	31.95	40.21	23.88	30.94	32.77	27.36
	M3DocRAG	38.83	36.56	30.46	36.24	35.83	46.85	25.29	28.70	36.18	30.96
	MoLoRAG+	42.69	**38.79**	33.26	38.53	40.73	52.90	27.59	35.87	41.01	34.94
	CAPS (Ours)	**44.51**	35.55	**37.97**	**42.76**	**44.09**	**57.20**	**28.67**	**42.15**	**44.69**	**38.35**
Mixed											
MDocAgent(LLaMA3.1-8B+Qwen2.5-VL-7B)		43.14	31.17	32.55	38.72	37.90	53.45	23.82	28.25	38.53	33.27

Table 9: Fine-grained performance analysis on LongDocURL under the retrieved top-3 setting. Some baseline results are adopted from Wu et al. (2025).

DocQA Model	Method	Evidence Source				Evidence Page		Acc	EM
		TXT	LAY	TAB	FIG	SIN	MUL		
Text-based									
Qwen2.5-7B	Text RAG	36.41	23.77	20.55	25.94	26.75	28.73	27.93	21.94
	DSOCR Text RAG	**41.14**	**25.66**	**24.93**	**30.08**	**31.58**	**30.00**	**30.87**	**24.43**
DeepSeek-V3	Text RAG	41.89	28.15	**30.84**	**35.49**	35.77	33.67	34.73	26.84
	DSOCR Text RAG	**46.03**	**29.32**	29.49	35.46	**36.60**	**33.78**	**35.18**	**27.74**
Vision-based									
LLaVA-NeXT-7B	Direct	16.79	7.39	5.28	12.12	7.87	13.43	10.78	9.29
	M3DocRAG	20.64	10.75	7.17	16.16	11.12	16.20	13.85	10.62
	MoLoRAG+	19.94	10.64	7.32	17.11	11.17	15.73	13.58	10.49
	CAPS (Ours)	**28.75**	**19.26**	**11.79**	**22.58**	**18.68**	**21.72**	**20.22**	**14.19**
DeepSeek-VL-16B	Direct	19.98	13.65	8.26	13.81	11.18	17.87	14.72	11.35
	M3DocRAG	40.61	30.78	16.19	27.56	25.54	33.31	29.60	21.29
	MoLoRAG+	44.28	32.84	29.89	**37.81**	39.19	35.58	37.21	27.74
	CAPS (Ours)	**51.05**	**36.35**	**30.44**	35.86	**42.13**	**38.64**	**40.27**	**30.32**
Qwen2.5-VL-3B	Direct	31.98	22.86	17.43	23.50	21.60	27.77	24.89	18.67
	M3DocRAG	54.07	36.97	37.97	**42.07**	46.39	42.64	44.40	34.97
	MoLoRAG+	54.24	36.62	**39.03**	41.13	**47.49**	43.31	45.27	35.53
	CAPS (Ours)	**56.12**	**38.89**	38.08	41.63	47.04	**44.26**	**45.60**	**35.87**
Qwen2.5-VL-7B	Direct	32.37	23.25	19.88	27.09	24.20	28.15	26.38	19.74
	M3DocRAG	58.16	41.24	43.75	46.04	53.35	45.13	49.03	38.88
	MoLoRAG+	61.43	42.56	45.98	49.01	55.01	49.01	51.85	40.13
	CAPS (Ours)	**64.33**	**43.30**	**47.96**	**52.35**	**57.80**	**50.09**	**53.78**	**42.71**
Mixed									
MDocAgent(LLaMA3.1-8B+Qwen2.5-VL-7B)		56.81	35.48	42.25	44.07	49.46	44.51	46.91	37.63

Table 10: Document Question-Answering Performance Comparison (in %) under the retrieved top-1 setting. Some baseline results are adopted from Wu et al. (2025).

DocQA Model	Method	MMLongBench	LongDocURL	PaperTab	FetaTab	Avg.
Text-based						
Qwen2.5-7B	Text RAG	22.11	20.75	5.34	22.64	17.71
	DSOCR Text RAG	**25.34**	**22.92**	**7.63**	**29.13**	**21.26**
DeepSeek-V3	Text RAG	25.94	24.32	**10.18**	34.55	23.75
	DSOCR Text RAG	**26.02**	**25.87**	**10.18**	**34.74**	**24.20**
Vision-based						
LLaVA-NeXT-7B	Direct	7.15	10.78	3.05	11.61	8.15
	M3DocRAG	16.32	25.25	**6.62**	15.26	15.86
	MoLoRAG+	17.15	**27.00**	6.36	17.52	17.01
	CAPS (Ours)	**18.56**	26.85	5.85	**17.91**	**17.29**
DeepSeek-VL-16B	Direct	8.40	14.72	6.11	16.14	11.34
	M3DocRAG	26.23	42.21	16.54	48.43	33.35
	MoLoRAG+	28.98	45.17	**21.88**	58.27	**38.58**
	CAPS (Ours)	**29.52**	**45.55**	18.32	**59.45**	38.21
Qwen2.5-VL-3B	Direct	26.65	24.89	**25.19**	51.57	32.08
	M3DocRAG	26.77	39.82	19.85	45.77	33.05
	MoLoRAG+	30.03	43.17	23.16	55.41	37.94
	CAPS (Ours)	**31.70**	**43.29**	20.36	**57.68**	**38.26**
Qwen2.5-VL-7B	Direct	32.77	26.38	**29.77**	64.07	38.25
	M3DocRAG	32.29	43.32	19.34	50.98	36.48
	MoLoRAG+	36.37	47.86	27.48	62.50	43.55
	CAPS (Ours)	**41.28**	**48.77**	23.16	**64.27**	**44.37**
Mixed						
MDocAgent(LLaMA3.1-8B+Qwen2.5-VL-7B)		31.73	44.42	21.63	57.78	38.89

Table 11: Document Question-Answering Performance Comparison (in %) under the retrieved top-5 setting. Some baseline results are adopted from Wu et al. (2025).

DocQA Model	Method	MMLongBench	LongDocURL	PaperTab	FetaTab	Avg.
Text-based						
Qwen2.5-7B	Text RAG	26.09	31.36	16.79	49.21	30.86
	DSOCR Text RAG	**32.21**	**33.13**	**18.32**	**51.08**	**33.69**
DeepSeek-V3	Text RAG	31.23	**39.04**	23.92	**62.01**	39.05
	DSOCR Text RAG	**34.23**	38.84	**27.48**	59.84	**40.10**
Vision-based						
LLaVA-NeXT-7B	Direct	7.15	10.78	3.05	11.61	8.15
	M3DocRAG	10.43	12.65	**4.58**	12.80	10.12
	MoLoRAG+	9.19	13.59	4.33	13.09	10.05
	CAPS (Ours)	**16.07**	**19.07**	4.33	**14.17**	**13.41**
DeepSeek-VL-16B	Direct	8.40	14.72	6.11	16.14	11.34
	M3DocRAG	18.87	29.27	8.14	27.26	20.89
	MoLoRAG+	24.86	38.02	9.67	41.44	28.50
	CAPS (Ours)	**27.00**	**39.60**	**10.94**	**48.03**	**31.39**
Qwen2.5-VL-3B	Direct	26.65	24.89	25.19	51.57	32.08
	M3DocRAG	28.38	44.67	**27.48**	55.22	38.94
	MoLoRAG+	32.41	45.13	**27.48**	58.07	**40.77**
	CAPS (Ours)	**33.25**	**46.41**	23.16	**59.94**	40.69
Qwen2.5-VL-7B	Direct	32.77	26.38	29.77	64.07	38.25
	M3DocRAG	37.19	50.33	30.53	64.37	45.61
	MoLoRAG+	40.47	52.33	31.55	69.39	48.44
	CAPS (Ours)	**42.72**	**55.20**	**35.11**	**72.05**	**51.27**
Mixed						
MDocAgent(LLaMA3.1-8B+Qwen2.5-VL-7B)		38.34	48.07	29.77	63.78	44.99