VERA: Identifying and Leveraging Visual Evidence Retrieval Heads in Long-Context Understanding

Rongcan Pei[*1], **Huan Li**[2], **Fang Guo**[3] and **Qi Zhu**[†2]
[1]Tongji University
[2]Zhejiang University
[3]Westlake University
prc@tongji.edu.cn, lihuancs@zju.edu.cn, guofang@westlake.edu.cn, qizhu.zju.research@gmail.com

Abstract

While Vision-Language Models (VLMs) have shown promise in textual understanding, they face significant challenges when handling long context and complex reasoning tasks. In this paper, we dissect the internal mechanisms governing long-context processing in VLMs to understand their performance bottlenecks. Through the lens of attention analysis, we identify specific **Visual Evidence Retrieval (VER) Heads** — a *sparse, dynamic* set of attention heads critical for locating visual cues during reasoning, distinct from static OCR heads. We demonstrate that these heads are causal to model performance; masking them leads to significant degradation. Leveraging this discovery, we propose VERA (Visual Evidence Retrieval Augmentation), a training-free framework that detects model uncertainty (*i.e.*, entropy) to trigger the explicit verbalization of visual evidence attended by VER heads. Comprehensive experiments demonstrate that VERA significantly improves long-context understanding of open-source VLMs: it yields an average relative improvement of **21.3%** on Qwen3-VL-8B-Instruct and **20.1%** on GLM-4.1V-Thinking across five benchmarks.

1 Introduction[1]

Multi-modal documents, such as PDFs, academic papers, and financial reports, constitute the backbone of many modern information systems. Unlike plain text, these documents are composed of a complex interplay of text, layout, tables, and figures. Early visual document processing pipelines rely on Optical Character Recognition (OCR) and subsequent natural language processing techniques. The emergence of Vision-Language Models (VLMs) has greatly eased the difficulty of end-to-end understanding on various components of multi-modal documents, enabling the unified interpretation of complex layouts.

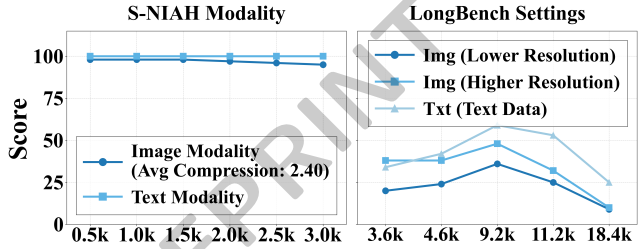

Figure 1: Performance comparison between image and text inputs on the NIAH task (left) and long-context reasoning task (right).

By treating document pages as high-resolution images, modern VLMs like GPT5.2 and Gemini3-Pro have achieved state-of-the-art on various visual and textual tasks. As agentic tasks demand increasingly long contexts, research has shifted toward investigating visual-text compression: determining whether visual modalities can represent textual content using significantly fewer tokens. To this end, DeepSeek-OCR [Wei *et al.*, 2025] sheds light on compressing 10× tokens on OCR tasks using a fine-tuned small VLM.

Recently, research on vision-text compression has been trending, while less attention has been paid to the conditions for achieving lossless compression. We begin by evaluating the performance of Qwen3-VL-8B-Instruct on the "Needle-in-a-Haystack" task [Hsieh *et al.*, 2024]. We first convert the context into text-rich images, and then provide the input to the VLM in both raw text and image formats for a comprehensive comparison. We demonstrate that in various context lengths, Qwen3-VL-8B-Instruct consistently maintains a high accuracy level between 95.2% and 98.4%, as shown in the left part of Figure 1. These findings align with prior observations that VLMs can achieve performance parity with LLMs on simpler, lower-complexity tasks [Li *et al.*, 2025].

However, this performance gap widens significantly as context length increases, particularly when evaluated on complex QA tasks (*e.g.*, multi-hop QA), as illustrated in the right panel of Figure 1. On Narrative QA, the performance drop by switching input to image even exceeds 60%! In addition, we also observe a larger gap under lower resolution (higher compression ratio). This observation reveals a critical bottleneck in VLM architectural capabilities that emerges as reasoning

[*]Work done during an internship at Zhejiang University.
[†]Corresponding author.
[1]The code is available at https://github.com/Prongcan/VERA.

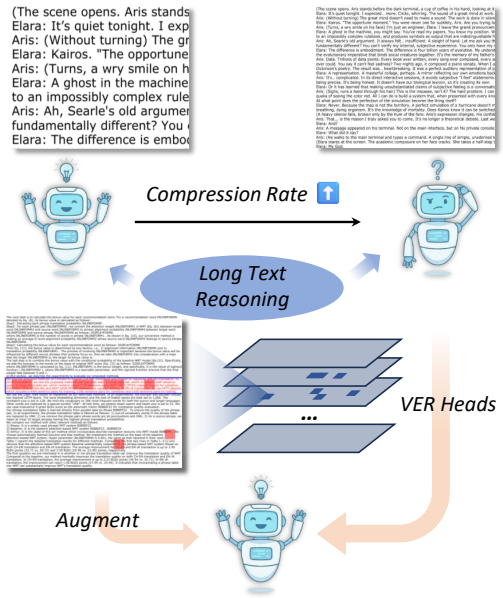

Figure 2: Challenges in high-compression VLMs and illustration of visual retrieval augmentation.

complexity and context length scale. We note that the selected resolution has already been optimized following the recommendations of recent work [Cheng *et al.*, 2025]. This gives rise to a question: *what are the underlying internal mechanisms through which VLMs comprehend long-context textual information when it is presented as an image?*

Existing studies, such as CogDoc [Xu *et al.*, 2025] and Glyph [Cheng *et al.*, 2025], started to enhance VLM's capability to comprehend text-rich images by designing end-to-end training frameworks and RAG methods; however, these approaches treat the model as a black box, yielding only marginal gains in our investigation.

To bridge this gap, we perform in-depth attention analysis and identify a specific category named Visual Evidence Retrieval Heads (**VER Heads**) within VLMs, which are responsible for the precise localization of evidence regions within image patches. Through a comprehensive suite of controlled experiments (see Section 3), we confirm the universality and significance of these retrieval heads for long-context understanding. For reasoning models, we propose probing VER heads at the first high-entropy token. Our qualitative analysis suggests this spike in uncertainty serves as a reliable proxy for the model's intrinsic "retrieval moment", signaling the precise step where it actively seeks external visual evidence.

Motivated by this, in Figure 2, we devise an inference-time VLM augmentation framework VERA that transforms these mechanistic insights into actionable guidance. Instead of relying on external retrievers, VERA leverages the newly proposed VER heads to perform attention steering. Specifically, when a high-entropy state is detected, we extract the visual patches attended to by VER heads and explicitly verbalize them into textual context. In experiments, we find this insight-driven approach improves performance of both reasoning and non-reasoning VLMs on five different datasets by **20%** on average. We believe the discovery of VER heads paves the way for high-ratio visual compression that preserves critical evidence.

We summarize our contributions as follows:

- We formally defined Visual Evidence Retrieval (VER) heads and developed a quantitative metric to evaluate the alignment between attention maps and golden evidence, validating their existence across multiple datasets.
- Through ablation studies, we established a causal link between VER head activation and task performance, providing a mechanistic explanation for the reasoning degradation observed in visual long-context understanding.
- We introduced VERA, a training-free paradigm that improves performance by leveraging the model's underlying mechanistic logic without specialized training.

2 Related Work

Visual Document Understanding. Early approaches to multi-modal document processing relied on Optical Character Recognition (OCR) [Shi *et al.*, 2015] to extract text, which was then processed using traditional NLP techniques. Large vision-language models like GPT-4o [Hurst *et al.*, 2024] and Gemini2.5 [Comanici *et al.*, 2025] have since transformed this paradigm, enabling seamless joint perception and reasoning directly from visual document representations [Kim *et al.*, 2022]. Recently, DeepSeek-OCR [Wei *et al.*, 2025] shed light on a third paradigm - utilizing VLMs for visual-text compression, thereby enabling the efficient processing of extremely long documents with significantly higher compression ratios. In this scenario, how to achieve high accuracy with low-resolution image input remains untouched.

Mechanistic Interpretability and Attention Analysis. Multi-head attention [Vaswani *et al.*, 2017] is the cornerstone of modern transformer-based language model architectures. Since the emergence of models like BERT and GPT-1/2/3, there has been extensive research [Voita *et al.*, 2019; Clark *et al.*, 2019] into the specific functions of different attention heads, such as syntactic dependencies and coreference resolution. Subsequently, the field shifted toward identifying heads with distinct *in-context* roles, most notably *induction heads* [Olsson *et al.*, 2022] discovered by Anthropic researchers. In the context of long-sequence modeling, [Wu *et al.*, 2024] identified *retrieval heads* in LLMs—specific heads responsible for fetching information from distant text tokens. In the multi-modal domain, recent work has discovered OCR heads [Baek *et al.*, 2025], which are solely responsible for recognizing text within images. However, our work reveals a critical distinction: while OCR heads focus on low-level perception, they differ significantly from our proposed Visual Evidence Retrieval (VER) heads. Unlike the dense activation patterns of OCR heads, VER heads are sparse and functionally specialized—they are dynamically activated only during high-uncertainty "retrieval moments" to ground complex reasoning. To the best of our knowledge, VERA is the first work to identify and leverage these dynamic heads for long-context visual document understanding.

Long-context VLM Enhancement. In textual LLMs, dense retrieval was performed to identify the relevant chunks, while the paradigm shifted towards visual embedding-based retrieval for VLMs. For example, ColPali [Faysse *et al.*, 2024] and PaliGemma [Beyer *et al.*, 2024] leverage strong vision encoders to perform direct retrieval on image patches or page embeddings. More recently, research has explored solving long-document tasks end-to-end via adaptive visual compression mechanisms, such as Glyph [Cheng *et al.*, 2025] and CogDoc [Xu *et al.*, 2025], though these approaches require extensive model post-training. Closely related to our work, Liu et al. [Liu *et al.*, 2025] explore attention-guided inference-time intervention but are limited to a coarse-grained *layer-wise* approach for standard VQA. In contrast, we propose VERA, a training-free framework that performs fine-grained *head-wise* attention steering by identifying and explicitly verbalizing visual evidence.

3 Characterizing VER Heads

3.1 Definition and Identification Method

Let $\mathcal{D} = \{(c_i, q_i, a_i, \mathcal{E}_i)\}_{i=1}^{N}$ denote a question answering dataset, where each sample consists of:

- A context passage c_i, constructed by concatenating multiple documents into a long context;
- A question q_i and its corresponding answer a_i;
- A set of evidence spans $\mathcal{E}_i = \{e_1, e_2, \ldots, e_K\}$, where each e_k is a contiguous text segment in C that supports the answer.

Text-to-Image Rendering. Similar to Text-as-Pixels [Li *et al.*, 2025], we render the context passage into an image $I = \texttt{Render}(c_i) \in \mathbb{R}^{H \times W \times 3}$. During rendering, we record the pixel-level bounding box $B_{e_k} = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ for each evidence span $e_k \in \mathcal{E}$. The union of all evidence regions forms a binary mask $M \in \{0,1\}^{H \times W}$:

$$M_{x,y} = \begin{cases} 1 & \text{if } (x,y) \in \bigcup_{k=1}^{K} B_{e_k} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This method enables controlled evaluation of VLMs across varying context lengths and image resolutions.

Visual Evidence Retrieval Score (VER Score). Given the rendered image I and question q_i, we feed them into a VLM and extract the attention distribution when generating the most *uncertain* token (*i.e.*, high entropy).

The VLM's vision encoder divides the input image of size $H \times W$ into a grid of $H_{\text{patch}} \times W_{\text{patch}}$ patches, where H_{patch} and W_{patch} are determined by the encoder's patch size. Each image patch $P_{i,j}$ corresponds to a visual token in the model's attention computation. For each patch $P_{i,j}$, we denote $|P_{i,j}|$ as the number of pixels in patch $P_{i,j}$ and calculate evidence coverage weight $w_{i,j} \in [0,1]$ as

$$w_{i,j} = \frac{1}{|P_{i,j}|} \sum_{(x,y) \in P_{i,j}} M_{x,y}, \tag{2}$$

measuring how much of patch $P_{i,j}$ overlaps with evidence regions.

Given an attention head h, let $A_{i,j}^{(h)}$ denote the attention weight assigned to the visual token corresponding to patch $P_{i,j}$ when generating the first output token. The **patch-level visual retrieval score** is defined as

$$r_{i,j}^{(h)} = w_{i,j} \cdot A_{i,j}^{(h)}. \tag{3}$$

The VER score of attention head h is obtained by aggregating over all patches:

$$R^{(h)} = \sum_{i=1}^{H_{\text{patch}}} \sum_{j=1}^{W_{\text{patch}}} r_{i,j}^{(h)}. \tag{4}$$

To quantify attention concentration independent of evidence size, we define the *normalized VER score* as $\bar{R}^{(h)} = R^{(h)}/\rho$, where ρ denotes the proportion of evidence pixels in the image. Finally, we identify the set of **Visual Evidence Retrieval (VER) Heads** as those satisfying $\bar{R}^{(h)} > \tau$, where the threshold τ is the midpoint of the normalized score range across all heads (i.e., $\tau = (\max \bar{R} + \min \bar{R})/2$).

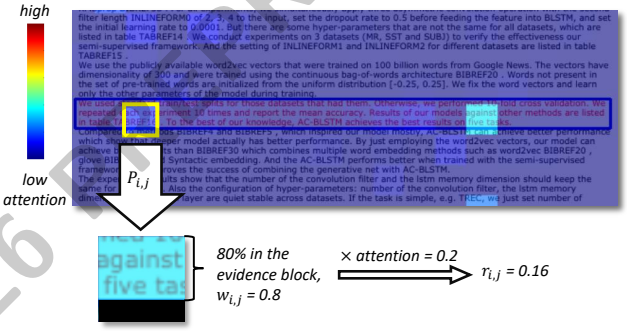

Figure 3: Calculation of VER Score of a specific head.

3.2 Quantitative Analysis of VER Heads

We now design experiments to study the effect of these VER heads across a diverse range of QA datasets and model architectures focusing on *long-context reasoning*. The five different datasets are shown in Table 4.

We evaluate a range of popular VLMs, both reasoning and non-reasoning, covering different model families: instruction-following VLMs with different vision encoders and LLM backbone such as Qwen3-VL series, reasoning-enhanced VLMs including Qwen3-VL-8B-Thinking and GLM-4.1V-Thinking.

To quantify VER head activation, we compute the VER Score (defined in Section 3.1) for each of the $L \times H$ attention heads across all datasets. To establish causality, we further conduct ablation experiments by masking the top-5 heads of varying types (VER, OCR, and random) and measure the resulting performance changes. Results are presented in Table 1. Throughout comprehensive experiments, we have the following observations:

(1) VER heads are *sparse* yet *universal*—they exist across different VLMs and remain consistent across datasets. In Figure 5, we plot the VER score of all heads on different datasets and models. Across all evaluated datasets, only a

	NO mask (%)	Random mask (%)	OCR head mask (%)	VER head mask (%)
F1 Score of Qwen3-8B-VL on different datasets				
DocMath	3.47	4.24	4.62	**2.80**
Qasper	28.37	26.63	26.75	**18.81**
HotpotQA	28.48	26.60	28.27	**23.80**
Musique	9.25	9.48	8.12	**7.76**
$\bar{\Delta}$	0	-0.66	-0.45	**-4.10**

Table 1: VER head Evaluation Results (F1 Score).

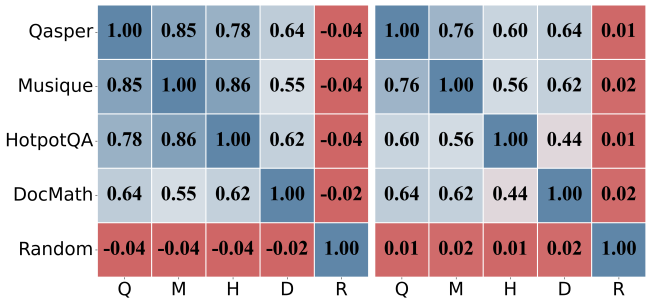

Figure 4: Spearman's rank correlation coefficients of VER Score distributions across different datasets. Left: Qwen; Right: GLM.

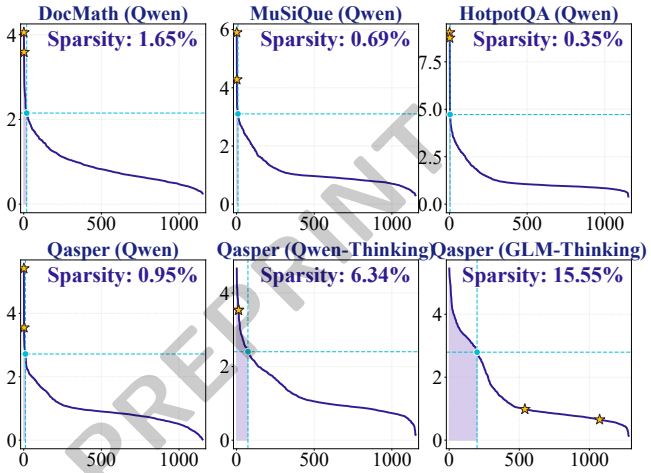

Figure 5: The distribution of VER Scores across attention heads. The stars stand for L21-H11 and L24-H29 attention heads.

small fraction (<1.65%) of attention heads of Qwen3-8B-VL exhibit high VER scores and are defined as VER heads. This explains the performance degradation of VLMs when used for long-context tasks. In the same model, we often observe a fixed set of VER heads are frequently activated in different datasets. For example, head (24,29) and (21,11) of Qwen3-VL-8B demonstrate high VER scores in all tested datasets. Finally, we examine the VER head distribution in reasoning models, such as Qwen-Thinking and GLM-Thinking. As illustrated in Figure 5, these models exhibit a similar sparse, long-tail distribution. Notably, the heightened activation intensity observed in these heads directly correlates with the superior performance recorded on downstream task metrics.

To quantify this consistency in different datasets, we compute the Spearman rank correlation coefficient between the vectors of flattened attention head's VER score of paired datasets. As illustrated in Figure 4, the Spearman rank correlation coefficients between the attention head rankings of diverse datasets are remarkably high, consistently exceeding **0.44**. This indicates a functional alignment, where the model consistently prioritizes the same specific set of heads for visual retrieval tasks regardless of the data distribution.

(2) VER heads are *functionally critical* for long-context reasoning—masking them degrades performance, while masking other head types does not. In Table 1, we investigate the performance change of Qwen3-8B-VL by masking most frequent (*i.e.*, top-5) VER heads, OCR heads [Baek *et al.*, 2025] and random heads. Here we mask only a very small fraction of heads (0.4%), which strongly demonstrates their importance. The result shows masking OCR Heads does not lead to substantial performance degradation on the evaluated datasets; on average, masking OCR heads do not perform much differently from masking the same amount of random heads. Masking VER heads causes F1 Scores drop by 4.10% on average, compared with a drop by 0.66% after masking OCR heads and 0.45% after masking random heads, which validates their importance in long-context understanding.

(3) Does fine-tuning activate new VER heads, enhance existing ones, or both? Having established that VER heads are functionally critical for long-context reasoning, a natural question arises regarding their plasticity: how do these specific mechanisms evolve during training? We investigate whether the performance gains observed after supervised fine-tuning are mechanistically rooted in the intensification of existing VER heads or the emergence of new ones. We first reorganized the Hotpot training dataset into a format consisting of image context, textual questions, and ground-truth answers, totaling 45,000 samples. Subsequently, we performed full-parameter supervised fine-tuning on the Qwen3-VL-8b-Instruct model. In Table 2, we found that with the improvement of performance, the activation of VER Head increased significantly across multiple datasets following SFT. On the Hotpot dataset, the average score of the heads ranked in the top 5% by VER score increased by approximately 25% after fine-tuning. On the DocMath dataset, this rate of increase reached 90%, demonstrating strong generalizability. We further compared VER score distributions before and after SFT, confirming a strong correlation between improved comprehension and increased VER head activation.

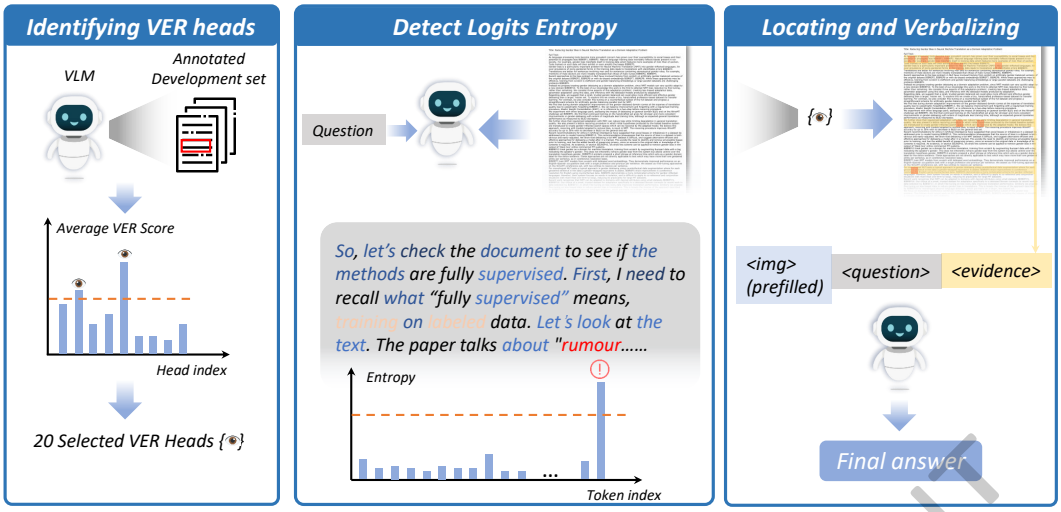

Figure 6: Pipeline of VERA.

	Qasper	Docmath	Musique	Hotpot
Top-5 Threshold				
Base	0.095	0.085	0.052	0.009
SFT	0.121	0.151	0.061	0.012
Top-5 Avg Score				
Base	0.110	0.097	0.061	0.012
SFT	0.145	0.187	0.071	0.016
F1 Score				
Base	28.37	3.47	9.25	28.48
SFT	28.40	4.00	16.24	37.89
Improved Heads	55/57	53/57	42/57	34/57

Table 2: The threshold demarcating the top 5% of scores, average VER Scores of the top 5% heads, and dataset performance before and after SFT.

4 Harnessing VER Heads in VLM Reasoning

In Section 3, we carefully examine the causal relations between the activation of VER Heads and the model's ability to reason over long contexts. This finding suggests that VLMs inherently possess the mechanism to locate task-specific visual information, but this capability is often diluted, which is also noted in recent work [Chen *et al.*, 2025] as imbalance visual and textual attention.

Building on this insight, we propose VERA (Visual Evidence Retrieval Augmentation), a training-free method that steers model focus by converting patches with high VER-head activation into text, effective for both instruction-following and reasoning VLMs. The complete pipeline of VERA is illustrated in Figure 6.

4.1 Identifying VER Heads for Different Models

In instruction following VLMs, we observe the model will attempt to retrieve the relevant visual information at the very first token, which makes VER head identification straightforward following Equation.

However, the attention distribution of reasoning models at the initial decoding step often fluctuates. Consequently, the first token fails to reliably attend on evidence regions M. To accurately characterize VER heads in these models, we must instead identify specific "retrieval moments" where the model actively seeks visual information to answer the question.

Logits Entropy as a Reasoning Indicator. We premise our approach on the observation that 'retrieval' is a consequence of 'uncertainty'. Specifically, we utilize logits entropy, a widely adopted metric for quantifying model uncertainty [Yong *et al.*, 2025]. Let $\mathcal{H}_t$ denote the entropy of the probability distribution over the vocabulary at generation step t. Qualitative analysis reveals that tokens with high $\mathcal{H}_t$ typically correspond to pivotal steps in the chain-of-thought where the model explicitly seeks external visual information.

The First High-Entropy Token. Based on this motivation, we propose identifying VER heads by analyzing the attention distribution $A^{(h)}$ at the first high-entropy token. We define this target step t^* as:

$$t^* = \arg\min\{t \mid \mathcal{H}_t > \delta\} \tag{5}$$

where δ is a hyper-parameter. In our experiment, we set $\delta = 2.0$ to effectively isolate these "retrieval moment" from standard generation steps, where logits entropy is typically negligible ($10^{-1} \sim 10^{-7}$). We select the first high-entropy token for two primary reasons: (1) indicating uncertainty on visual context; (2) targeting the initial instance minimizes interference from generated reasoning history, ensuring the attention signal reflects a reliance on the source image I rather than textual cues from the chain-of-thought. Upon identifying t^*, we compute the VER score for head (i, j) $R^{(h_{ij})}$ regarding Equation 4.

4.2 Locating Visual Evidence Patch via Head Selection

To accurately locate visual evidence during inference for a target model, we first identify a specific set of VER heads that

reliably attend to evidence regions across different datasets. Building on our observation in Section 3 that VER heads are similar in different datasets, we employ a selection strategy based on a held-out dataset with ground-truth annotations.

Head Selection Strategy. We construct a development set D_{dev} consisting of a small number of annotated instances. We first compute the average VER score for all attention heads. Then we select the set of active retrieval heads, $H_{\text{selected}} = \{h_1, ..., h_k\}$, as the top-$k$ heads with the highest average scores on $\mathcal{D}_{dev}$.

Evidence Patch Extraction. During inference time, we average the attention maps of the selected heads in H_{selected}. We then extract the top-N visual patches as evidence patches $\mathcal{E} = \{P_{i,j}\}$ with the highest accumulated attention weights. These patches serve as the "visual evidence" for the subsequent verbalization step.

4.3 Verbalizing Retrieved Visual Evidence

To explicitly steer the model's reasoning, we construct a retrieval-augmented prompt that integrates the visual cues. Formally, the input sequence $\mathbf{X}$ is structured as:

$$\mathbf{X} = [\texttt{<image>}, \mathbf{C}_{ver}, \texttt{Question}, \texttt{Answer:}] \quad (6)$$

where `<image>` represents the visual tokens and $\mathbf{C}_{ver}$ denotes the text derived from the visual evidence patches $\mathcal{E}$.

To mitigate semantic fragmentation, we employ a horizontal context expansion strategy: rather than verbalizing isolated patches, we extract the entire text row corresponding to the vertical coordinates of any selected patch $P \in \mathcal{E}$. This ensures that the retrieved evidence maintains local textual consistency. While this verbalization can generally be achieved via OCR models, in our experiments—where images are rendered from source text—we directly map the retrieved visual coordinates back to the corresponding source text lines for precise evaluation. Inference of VERA can be optimized via caching the KV of image token in first pass. And the detailed algorithm is shown in Appendix A.

5 Experiments

To evaluate the effectiveness of VERA in enhancing VLM's ablility in long-context understanding comprehensively, we applied it to both instruction-following and reasoning models. Through extensive experiments across multiple datasets. We aim to answer the following research questions:

RQ1: How does VERA perform compared to existing baselines? RQ2: How well does VERA generalize across different models and datasets? RQ3: Is VERA sensitive to hyper-parameters?

5.1 Evaluation Setup

To validate the effectiveness of the VERA framework, we focus on evaluating VLMs in long-context Question Answering (QA) task. Following the methodology introduced in Section 3.1, we transform standard text-based QA benchmarks into visual documents. This rendering process is optimized to achieve high information density, allowing us to simulate long-context visual understanding tasks under the same controlled experimental environment.

Datasets. We select five different datasets[Zhao *et al.*, 2024; Dasigi *et al.*, 2021; Yang *et al.*, 2018; Trivedi *et al.*, 2022; Chen *et al.*, 2026] as shown in Table 4 based on three criteria: (1) **diverse domains**: covering mathematical reasoning, scientific literature, and multi-hop QA[2] to ensure generalizability; (2) **annotated evidence**: some datasets contain ground-truth evidence annotations; (3) **different question complexity**: featuring QA pairs with varying difficulty levels (*e.g.*, extractive, abstractive). Specifically, we sample 70 instances from each seen dataset to construct $\mathcal{D}_{\text{dev}}$ for model-specific VER heads detection.

Baselines. We select two representative models as the instruction-following and reasoning base models - Qwen3-VL-8B-Instruct [Yang *et al.*, 2025] and GLM-4.1V-9B-Thinking [Hong *et al.*, 2025]. We employ VERA on both models to examine the effectiveness of VER heads for different tasks. To examine whether VER heads are critical to VERA framework, we adopt the following baselines:

OCR-RAG, which identifies recognition heads on the NIAH task following [Baek *et al.*, 2025]; **Random-RAG**, which samples consecutive text segments of comparable length; **Embedding RAG**, which utilizes the **Qwen3-VL-Embedding** model for retrieval; **ColPali RAG**, which employs a VLM capable of producing high-quality multi-vector embeddings for retrieval; **Glyph**, a specialized VLM trained for optimal visual context compression.

On each dataset, we report the score between ground truth and model prediction excluding thought tokens. We utilize the official evaluation scripts on all datasets, with F1-score acting as the primary metric. Within these, 40% of the LongBench Pro data uses precision.

Implementation. VERA is a training-free approach that has two hyper-parameters: (1) top-k VER heads; (2) top-N visual evidence patches. In experiments, we set $k = 5$ to identify the top-5 VER heads that achieve the highest VER scores on development set $\mathcal{D}_{\text{dev}}$. In addition, we set $N = 20$ to balance the amount of additional information against the compromised compression ratio. For RAG baselines including VERA, we extract selected text by expanding the selected visual patches to rows. Then we append the additional context following the same evidence verbalization described in Equation 6. Lastly, we provide the details of image rendering, VER head selection and prompt in Appendix B.

5.2 Performance (RQ1 and RQ2)

We present the evaluation results of VERA and other baselines in Table 3. From the table, we have two key observations regarding the effectiveness (**RQ1**) and universality (**RQ2**) of the VERA framework.

1. Significant Improvement over Baselines. Using Qwen3-VL-8B-Instruct as a comparative baseline, we observe the evidence retrieved via OCR heads yields only marginal performance gains. This validates our finding that low-level recognition heads lack the semantic alignment for reasoning—a gap VERA effectively bridges. Interestingly,

[2]We augment HotpotQA and Musique into long-context datasets, and clean the LongBench Pro with details in Appendix B.

Method	DocMath	Qasper			HotpotQA	MuSiQue	LongBench Pro	Rel.Δ
		Extract	Abstract	Bool				
Baselines								
Qwen3-VL-8B-Instruct	3.47	25.54	17.82	70.36	28.48	9.25	27.56	/
Qwen3-VL-8B-Instruct Random RAG	5.61	23.96	16.24	65.86	30.16	14.68	28.00	1.1%
Qwen3-VL-8B-Instruct OCR RAG	4.72	25.15	17.99	71.48	29.03	14.70	26.40	3.8%
GLM-4.1V-9B-Thinking	15.71	41.69	26.71	81.28	39.49	27.85	31.06	/
GLM-4.1V-9B-Thinking Random RAG	21.99	46.92	29.61	80.47	38.00	28.71	28.84	4.1%
GLM-4.1V-9B-Thinking Embedding RAG	17.49	45.21	23.71	73.60	36.03	28.08	26.29	-5.1%
GLM-4.1V-9B-Thinking ColPali RAG	21.18	47.51	26.74	75.16	**43.89**	30.24	28.58	3.6%
Glyph	13.61	39.26	24.45	45.62	38.74	24.87	28.94	/
Attention-Guided RAG								
Qwen3-VL-8B-Instruct VERA (Ours)	9.45	39.29	22.56	71.12	32.32	17.84	28.74	21.3%
GLM-4.1V-9B-Thinking VERA (Ours)	**29.02**	**54.57**	**34.67**	**94.23**	39.56	**30.58**	**34.20**	20.1%

Table 3: Performance comparison (F1 %) of different RAG methods across long-context QA benchmarks. The best performing methods in each column are highlighted in **bold**, and the second-best are underlined.

Dataset	Context Length	Avg. visual tokens	Seen in $\mathcal{D}_{\text{dev}}$
DocMath	2.2k	1.3k	Yes
Qasper	3.3k	1.5k	Yes
HotpotQA	9.5k	5.7k	Yes
MuSiQue	10.2k	4.0k	Yes
LongBench Pro	1.5k–16k	4.6k	No

Table 4: Overview of datasets used in our experiments.

Dataset	$k = 5$	$k = 10$	$k = 15$	$k = 20$
DocMath	**9.45**	8.60	9.39	9.35
Qasper	**36.37**	35.91	36.10	35.45
HotpotQA	32.32	33.36	32.88	**34.36**
MusiQue	17.84	18.01	18.00	**18.57**

Table 5: F1 Scores under different values of k.

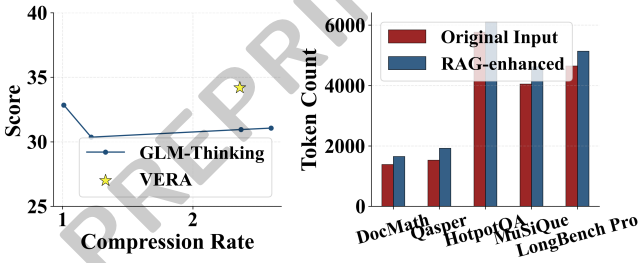

Figure 7: Cost efficiency analysis: varying visual compression ratio (left) and the minimal token cost overhead of VERA (right).

random RAG can achieve comparable gains to OCR RAG, which supports the conclusion that OCR-related attention is rarely activated in long-context reasoning. Overall, GLM-4.1V-9B-Thinking+VERA almost achieves the best performance across all datasets. ColPali RAG shows a competitive performance on HotpotQA, however, this is attributed to the high dispersion of patches retrieved by ColPali, which improves recall at the cost of precision. This is further analyzed in the retrieval performance study of the arXiv version.

2. Robustness across Model Architectures and Unseen Domains. As shown in the last column, VERA delivers over 20% average relative gains across both model types and generalizes well to the unseen LongBench Pro. Notably, it boosts GLM-4.1V-9B-Thinking on DocMath by **85%**, confirming that reasoning models benefit from VER heads and validating the first high-entropy token as the "retrieval moment".

5.3 Hyper-parameter Study (RQ3)

Hyperparameter Sensitivity. We designed experiments to explore the influence of the hyperparameter k on the effect of VERA. We vary the value of k, and use the average distribution of the attention scores of k VER heads to determine the retrieved patches. We compute the resulting F1 score to evaluate. The results are shown in Table 5. We observe that larger k values improve comprehension in long-context datasets (HotpotQA, MuSiQue), whereas $k = 5$ is optimal for shorter contexts (DocMath, Qasper). However, the overall performance variance is not significant.

Cost Efficiency of VERA. Although VERA introduces additional context tokens to the input, we demonstrate that the base model (GLM-Thinking) consistently underperforms VERA across varying visual compression ratios in Figure 7(left). This indicates that the performance gains achieved via VER heads outweigh the minimal token overhead incurred by the additional textual context. As shown in Figure 7 (right), VERA achieves 20% relative improvement with only 5.8%–15% additional tokens.

6 Conclusions

In this work, we identify VER Heads—sparse, dynamic attention heads distinct from OCR heads and are causal to visual text reasoning. Based on this discovery, we propose VERA, a training-free framework that explicitly verbalizes visual evidence attended by VER heads. Experiments demonstrate that VERA yields an average relative improvement of over 20% across diverse benchmarks. Our work paves the way for efficient and interpretable long-context VLMs.

Acknowledgements

This work was supported by the Major Research Program of the Zhejiang Provincial Natural Science Foundation (Grant No. LD24F020015).

References

[Baek *et al.*, 2025] Ingeol Baek, Hwan Chang, Sunghyun Ryu, and Hwanhee Lee. How do large vision-language models see text in image? unveiling the distinctive role of ocr heads. *arXiv preprint arXiv:2505.15865*, 2025.

[Beyer *et al.*, 2024] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[Chen *et al.*, 2025] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025.

[Chen *et al.*, 2026] Ziyang Chen, Xing Wu, Junlong Jia, Chaochen Gao, Qi Fu, Debing Zhang, and Songlin Hu. Longbench pro: A more realistic and comprehensive bilingual long-context evaluation benchmark, 2026.

[Cheng *et al.*, 2025] Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, Yushi Bai, Jie Tang, Hongning Wang, and Minlie Huang. Glyph: Scaling context windows via visual-text compression, 2025.

[Clark *et al.*, 2019] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[Comanici *et al.*, 2025] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[Dasigi *et al.*, 2021] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June 2021. Association for Computational Linguistics.

[Faysse *et al.*, 2024] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

[Hong *et al.*, 2025] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[Hsieh *et al.*, 2024] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024.

[Hurst *et al.*, 2024] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[Kim *et al.*, 2022] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022.

[Li *et al.*, 2025] Yanhong Li, Zixuan Lan, and Jiawei Zhou. Text or pixels? evaluating efficiency and understanding of LLMs with visual text inputs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10564–10578, Suzhou, China, November 2025. Association for Computational Linguistics.

[Liu *et al.*, 2025] Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*, 2025.

[Olsson *et al.*, 2022] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[Shi *et al.*, 2015] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, 2015.

[Trivedi *et al.*, 2022] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

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

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

[Wei *et al.*, 2025] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

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

[Xu *et al.*, 2025] Qixin Xu, Haozhe Wang, Che Liu, Fangzhen Lin, and Wenhu Chen. Cogdoc: Towards unified thinking in documents, 2025.

[Yang *et al.*, 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[Yong *et al.*, 2025] Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. *arXiv preprint arXiv:2505.18237*, 2025.

[Zhao *et al.*, 2024] Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand, August 2024. Association for Computational Linguistics.