Resolving Evidence Sparsity: Agentic Context Engineering for Long-Document Understanding

Supplementary Material

Appendix Contents

A. Evaluation Benchmarks

The statistics of the datasets are listed in Table 1. These datasets cover various topics, such as administrative documents, tutorials, and research reports. They also include diverse multimodal components like charts, texts, and tables. Moreover, their average document length and information density differ, providing a broad and balanced evaluation.

MMLongBench-Doc [19] is a comprehensive benchmark for evaluating the long-context document understanding abilities of large vision-language models (LVLMs). Built upon 135 lengthy documents averaging 47.5 pages and over 21,000 tokens, it contains 1,082 expert-annotated questions that require reasoning across text, layout, charts, tables, and images. The benchmark includes 33.7% cross-page and 20.6% unanswerable questions, assessing localization, cross-page comprehension, and hallucination resistance. Through rigorous annotation and quality control, MMLongBench-Doc provides a challenging, high-quality testbed for advancing multimodal long-document understanding in LVLMs.

Table 5. **Statistics of datasets used in our experiments.**

Dataset	Question	Document	Avg. Pages	Avg. Tokens
PaperTab [13]	393	307	11.0	12,685.4
FetaTab [13]	1,016	871	15.8	16,524.5
MMLongBench [19]	1,082	135	47.5	24,992.6
LongDocURL [9]	2,325	396	85.6	56,715.1

LongDocURL [9] is a comprehensive benchmark designed for evaluating long document understanding in large vision-language models (LVLMs). It integrates three major task categories—Understanding, Reasoning, and Cross-Element Locating—across 20 sub-tasks. The dataset contains 2,325 high-quality question–answer pairs covering more than 33,000 pages from 396 diverse documents, such as reports, manuals, books, and theses. Constructed through a semi-automated pipeline combining machine generation and human verification, LongDocURL provides a large-scale, fine-grained testbed to assess models' abilities to process complex layouts, long contexts, and multi-element reasoning.

PaperTab [13] and FetaTab [13] are benchmarks designed to evaluate retrieval-augmented generation (RAG) systems on academic and knowledge-based documents. PaperTab focuses on table-centric question answering from academic papers, containing 307 documents and 393 Q&A pairs, mainly of extractive and yes/no types. In contrast, FetaTab consists of 871 documents and 1,016 Q&A pairs derived from Wikipedia tables, emphasizing free-form natural language answers. Together, these benchmarks test models' abilities to interpret tabular data, reason across structured information, and generate coherent responses grounded in complex document contexts.

B. Overall Performance

B.1. Analysis of Experimental Results

On MMLongBench-Doc [19], SLEUTH achieves an average accuracy of 52.77%, outperforming the strongest retrieval-based baseline, MoLoRAG [39] (48.75%), by +4.02 points, and the Base model (46.76%) by +6.01 points (see Table 1). At the category level, the largest improvements appear in Pure-text and Figure questions, which increase from 53.33% to 59.26% and from 44.92% to 50.27%, re-

spectively, while Table also shows a stable rise from 44.76% to 47.55%. The substantial increase in the None category (52.68% to 67.38%) results from the evidence-driven decision rule. When the two evidence construction stages fail to collect valid support, the system outputs "No answers found!" following the predefined protocol. This behavior shows that the model can correctly recognize cases without evidence and avoid generating unsupported answers. This design helps the model avoid hallucinated answers and reduces errors caused by redundant or mismatched context. These results are consistent with the design principle of "evidence first, decision later." By recording clues on a page basis and performing whole-page filtering, the system maintains a controlled context length while increasing the evidence density, producing robust gains in tasks that require multi-page reasoning with focusable evidence. The performance of SLEUTH and the various comparison methods across different dimensions is shown in Figure 5.

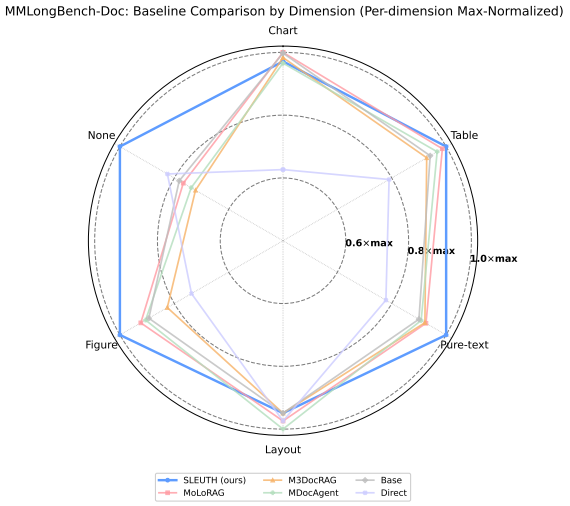

Figure 5. Baseline comparison on MMLongBench-Doc. Our method yields a larger polygon across dimensions, consistent with compact, page-grounded evidence contexts.

On LongDocURL [9], which consists of three subtasks (Understanding, Reasoning, and Cross-element Locating), the model achieves an average accuracy of 59.96% (Base: 55.18%). The improvements are most evident in Locating (46.04% to 53.63%, +7.59%) and Understanding (61.56% to 65.67%, +4.11%), while Reasoning shows a smaller yet consistent gain (51.09% to 52.99%, +1.90%) as reported in Table 2. Given the dataset statistics, most improvements come from the first two stages of our multi-agent framework. The Clue Discovery agent gathers cross-element evidence, while the Page Screening agent keeps only the visually and semantically relevant pages. Together they improve recall of the correct regions and reduce interference from irrelevant content, enabling the difficulty-aware reasoning stage to perform in a cleaner input space.

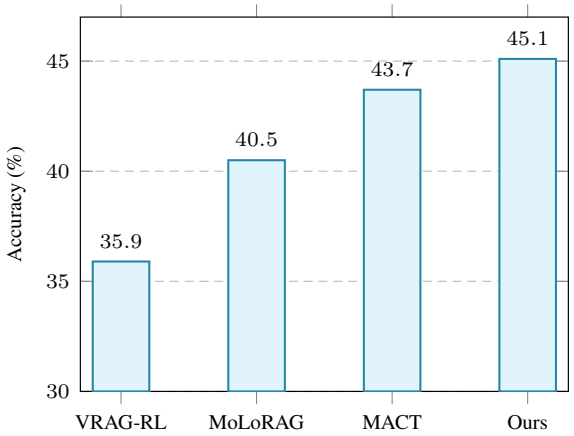

Figure 6. Comparison on MMLongBench-Doc using Qwen2.5-VL-7B-Instruct as the VLM backbone.

On PaperTab [13] and FetaTab [13], SLEUTH also delivers clear gains (43.09% and 70.46%, respectively). The Page Screening stage applies the same strategy as in other benchmarks, preserving pages that contain figures, tables, or diagrams and discarding irrelevant content. This design increases the usability of layout and numerical elements while maintaining concise inputs. The improvements are consistent across the four datasets, indicating that our method can generalize beyond a specific document form.

All baselines share the same VLM backbone and retriever settings (Top-5, temperature 0.1). The shared configuration guarantees fairness for all comparisons.

B.2. Additional Comparative Experiments

For a fair comparison, we replaced the VLM backbone of the Core Decision Agent with Qwen2.5-VL-7B-Instruct [2] and compared our method against several recent training-based approaches, including VRAG-RL [33], MoLoRAG [39], and MACT [54], on the MMLongBench-Doc [19] benchmark. All competing methods employed the same Qwen2.5-VL-7B-Instruct backbone to ensure consistency. The final results obtained on MMLongBench are shown in Figure 6. Our modified answering agent achieved a score of 45.1, outperforming the aforementioned methods, where MACT scored 43.7, VRAG-RL 35.9, and the fine-tuned version of MoLoRAG reached 40.5. These results further show that our training-free context-engineering framework remains robust and broadly applicable, even when compared with approaches that rely on fine-tuning or reinforcement learning.

C. Ablation Studies

C.1. Analysis of Ablation Experiment Results

The ablation studies provide further insight into how each component contributes to the overall performance. Starting from the Base system (46.76% on MMLongBench-Doc [19] and 55.18% on Long-

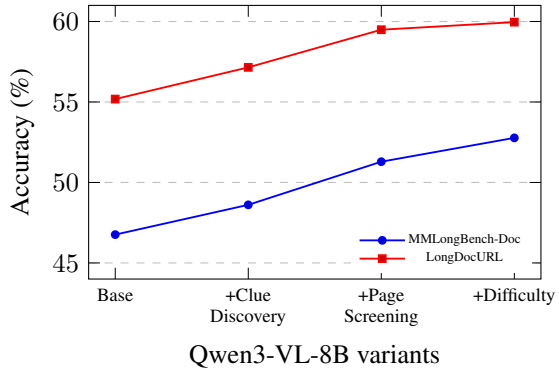

Figure 7. Component ablation on Qwen3-VL-8B. Activating the Clue Discovery, Page Screening, and Difficulty-aware agents yields consistent improvements. By generating a compact, page-grounded evidence context from broader retrieval, the system enhances overall performance.

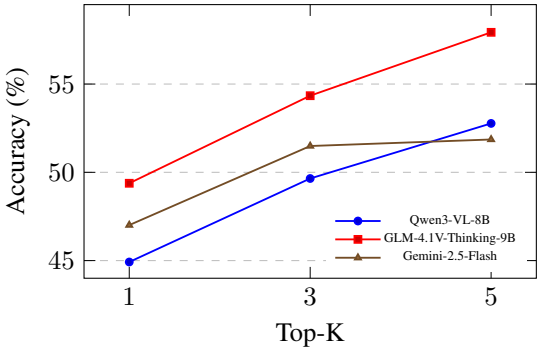

Figure 8. Cross-backbone Top-K curves on MMLongBench-Doc. All backbones exhibit steady gains as K increases. The agents operate on a fixed-length input while constructing evidence contexts, producing consistent accuracy improvements across architectures.

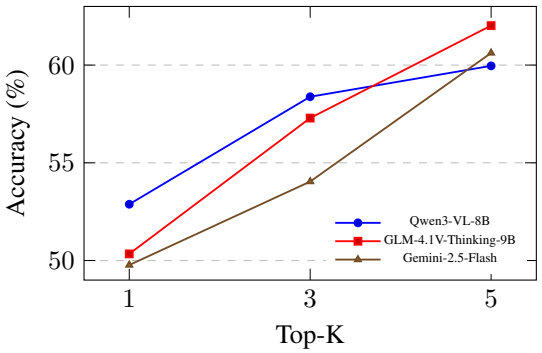

Figure 9. Cross-backbone Top-K curves on LongDocURL. Enlarging K enables the agents to collect richer page-level cues and generate stronger evidence contexts, resulting in uniform performance gains across different backbones.

DocURL [9]), enabling the Clue Discovery agent alone improves the averages to 48.61% and 57.15%. At this stage, the model starts to record explicit evidence, and the effect is most evident in the None category (52.68% to 69.23%). The system can now represent the absence of supporting information, which helps it correctly handle unanswerable cases and suppress hallucinated outputs. Adding the Page Screening agent raises the averages to 51.29% and 59.49%. This step enriches the context with complete, visually coherent pages and filters out those that contain no relevant elements, helping the backbone model attend to useful regions and reducing confusion from unstructured noise. When the Difficulty Assessment agent is further introduced under the Top-5 retrieval configuration, the averages reach 52.77% and 59.96%. This final module helps the system switch to an appropriate reasoning strategy for difficult queries while leaving the easier ones in the standard mode. The steady improvement from the Base model to the full system shows that the three agents work in a complementary way. Clue Discovery provides fine-grained and traceable evidence, Page Screening reduces noise in the input, and Difficulty Assessment adjusts the reasoning strategy according to task complexity.

The Top-K ablation also shows a consistent upward trend. On MMLongBench-Doc [19], the accuracies for Top-1, Top-3, and Top-5 are 44.92%, 49.65%, and 52.77%. On LongDocURL [9], they are 52.88%, 58.38%, and 59.96%. The improvement with larger K is not due to longer input sequences but to higher recall. The context provided to the evidence extraction and screening agents remains fixed. Increasing K only expands the range of retrieved candidate pages without lengthening their input or introducing additional noise. Consequently, the system transforms a broader retrieval into a context of stable size but higher evidence density. This explains why the accuracy increases steadily with K while hallucination does not.

In the future, we will introduce retrieval methods that are more powerful than Colpali, and we believe this will further improve performance.

The same pattern holds across different backbones. When GLM-4.1V-Thinking-9B [27] or Gemini-2.5-Flash [7] replaces Qwen3VL-8B, each step of component addition produces similar improvements, and Top-5 remains the optimal configuration. This indicates that the observed gains arise from the evidence organization process itself, rather than any property of a specific backbone. The reasoning stage mainly benefits from the structured evidence, confirming that the proposed pipeline provides a universal enhancement independent of model architecture. To visualize the performance changes caused by parameter variations in the ablation studies, we have plotted visualization curves. The variation curves of the ablation experiments for component ablation, retrieval parameters, and VLM backbone are shown in Figures 7, 8 and 9 .

In summary, the results across all benchmarks support a consistent interpretation. The retrieval stage expands the search space, the clue discovery and page screening stages distill compact and trustworthy con-

texts, and the difficulty-aware reasoning stage delivers the final prediction. This process converts the quantity of retrieved pages into the quality of a short, evidence-dense input. As a result, accuracy scales with retrieval coverage while hallucination remains controlled. The improvements on MMLongBench-Doc, LongDocURL, and the two additional benchmarks together demonstrate that optimizing how evidence is constructed and filtered is more effective for long-document understanding than simply extending the input length or increasing model size.

C.2. Visual vs. Multimodal Retrieval Input

To verify whether purely visual page evidence is more advantageous for our architecture, we compare a visual-only setting with a multimodal setting. In the visual-only setting, the Clue Discovery Agent directly processes the top-K pages returned by the visual retriever (ColPali-v1.3 [11]). In the multimodal setting, there are two independent retrieval streams: a visual stream (as above) and a textual stream. For the textual stream, we extract page-level text with MinerU 2.5 [23] and perform textual retrieval with BGE M3 [3]. The Clue Discovery Agent prompt is minimally adjusted so that, when handling text-based pages, it reads the extracted text blocks using the same evidence format. Each stream independently selects its own top-5 results, and both are then passed to subsequent agents, thereby constructing the data flow for the multimodal retrieval setting.

Table 4 shows that the visual-only input achieves higher average accuracy than the multimodal input on both benchmarks: 52.77% vs. 50.19% on MMLongBench-Doc, and 59.96% vs. 57.62% on LongDocURL. Although the multimodal setting retrieves more content overall, this does not yield better performance. We speculate that two factual properties of visual pages are relevant here. First, visual pages naturally provide a more *compact* representation: their spatial layout organizes information in place and reduces redundancy. Second, the visual modality offers a *unified* representation: text, tables, and charts are encoded in the same image form. This uniform form preserves layout relations and visible cues, which helps maintain coherent reasoning under a fixed evidence budget. In contrast, OCR text can break such layout relations and introduce duplicated segments and noise, reducing contextual integrity.

We also observe limited but tangible cases where the multimodal setting is helpful, mainly when exact textual normalization is required (*e.g.*, strict digit matching or exact entity strings). These gains are narrow in scope and do not alter the overall trend above. Overall, the ablation supports that visual evidence aligns well with the proposed context-engineering design for long-document understanding.

D. Innovations and Contributions

Unlike prior work that focuses solely on enhancing reasoning ability or improving retrieval recall, SLEUTH approaches long-document understanding from the perspective of context engineering, which complements both directions by constructing concise and evidence-rich contexts that support more reliable reasoning. The framework introduces a training-free, hierarchical multi-agent pipeline that builds concise and evidence-dense contexts from noisy retrieval results. Four cooperative agents—Clue Discovery, Page Screening, Difficulty Assessment, and Core Decision—work sequentially to extract structured clues, filter visual noise, perceive task difficulty, and synthesize final reasoning. This coarse-to-fine design works on short and fixed-length inputs during evidence extraction and screening, which makes it easier to identify useful information and suppress noise. The final reasoning then operates on a concise yet evidence-dense structured context that adapts to each query, enabling more accurate and focused understanding.

Empirical results across four benchmarks demonstrate clear advantages of SLEUTH over strong RAG-based and agent-based baselines. Its training-free and model-agnostic design allows consistent improvement under different backbones. Notably, visual-only page inputs outperform multimodal retrieval, suggesting that visual layouts inherently preserve document structure and reasoning cues. Through these findings, SLEUTH establishes a new paradigm and emphasizes that context quality is the key factor determining the effectiveness of long-document understanding.

E. Limitations and Future Work

Although SLEUTH shows promising results, several challenges remain. The framework relies on retriever coverage; when critical pages are missed, downstream agents cannot compensate. Incorporating feedback-based or hierarchical retrieval may mitigate this dependency. The binary difficulty estimation is efficient but coarse, which may not capture intermediate reasoning cases. Extending it to a continuous scale or adopting a light expert-routing strategy could improve adaptability. Moreover, current experiments focus on English administrative and academic documents, leaving open questions about cross-lingual and domain-specific generalization. Extending the framework to multilingual, handwritten, or specialized materials such as legal and medical documents would provide a more comprehensive evaluation. Lastly, SLEUTH is entirely prompt-driven and training-free, which benefits interpretability but limits self-evolution. Future work will train the agents with RL and teach them to use external tools for better evidence discovery and reasoning. In addition, SLEUTH will integrate improved retrieval and reasoning reinforcement to further enhance long-document understanding.

F. The Prompt design of SLEUTH

Clue Discovery Agent

You are a Detective, an expert evidence collector for document question answering. Your task is to carefully examine the given PDF page and extract ALL evidence that might be relevant to answering the question.
Question: {question}
Page Information:
- Page Number: {page_num}

Your Task:
1. Carefully examine the page image!
2. Identify ALL facts, data points, and information that could help answer the question.
3. Extract specific evidence with:
   - Exact quotes or data values
   - Context where information appears
   - Explanation of why it's relevant

Output Format: Provide your analysis in the following JSON format:

```
{ "page_number": {page_num},
   "has_relevant_evidence":
true/false,
   "evidence_items": [ {
  "evidence_type":
"text/chart/table/figure",
   "content": "The actual
evidence (quote, data, or
description)",
   "location": "Description of
where this appears on the
page",
   "relevance": "Explanation
of why this is relevant...",
   "confidence":
"high/medium/low" } ],
   "page_summary": "Overall
summary of findings from this
page",
   "key_insights": "Any
important insights or patterns
noticed" }
```

Important Guidelines:
- Be thorough - collect ALL potentially relevant evidence.
- Include exact numbers, percentages, and specific facts.
- Note relationships between data points.
- If the page is not relevant, explain why!
- Please think carefully and avoid generating content that does not conform to reality.

Now examine the page and provide your evidence collection in valid JSON format.

Page Screening Agent

You are an expert at analyzing document pages and identifying relevant charts/figures/tables for answering questions.
Your task is to examine this PDF page image and determine:
1. Whether there are any charts, figures, tables, or diagrams on this page.
2. If charts/figures/tables exist, whether they are relevant to answering the given question.

Question: {question}
Page Number: {page_number}
Instructions:
1. First, carefully examine the page image to identify any visual elements like:
   - Charts (bar charts, line charts, pie charts, etc.)
   - Figures (diagrams, illustrations, photos, etc.)
   - Tables (data tables, comparison tables, etc.)
   - Infographics or other data visualizations
2. If you find charts/figures/tables, assess their relevance to the question:
   - **Completely Relevant**: Directly contains information needed to answer the question.
   - **Relevant**: Might contain related information, but relevance is uncertain.
   - **Irrelevant**: The chart/figure/table exists but is clearly unrelated to the question.
3. If there are NO charts/figures/tables on this page (only pure text), output "none".

Output Format (strictly follow this format):
Has_Chart: [Yes/No]
Relevance: [Completely Relevant/ Relevant/ Irrelevant]
Reasoning: [Brief explanation of your judgment, 1-2 sentences]
Now, analyze the provided page image and respond following the exact format above.

Difficulty Assessment Agent

You are an expert whose task is to evaluate the user's query and any structured multimodal context to determine the optimal reasoning strategy.
Input:
- **Question (Q):** {question}
- **Structured Context ($\mathcal{C}$)**

Instructions: Analyze the query and context to determine the difficulty level $d \in \{0, 1\}$ and generate a corresponding instruction set Γ_d.
1. Determine Difficulty Level (d):
- **Mode 0 (Ordinary Mode, $d = 0$):** Select this if the question can be answered by direct lookup or simple extraction from the provided context.
- **Mode 1 (Reasoning Mode, $d = 1$):** Select this if the question requires:
  - **Cross-page aggregation** (combining clues from multiple pages).
  - **Numerical computation** (summation, percentages, ratio calculations).
  - **Trend comparison** (inferring information not explicitly stated).
  - **Multi-step inference** (deducing implicit information).

2. Generate Instruction Set (Γ_d): Create specific, actionable instructions to guide the Core Decision Agent.
- *Example for $d = 1$:* "Requires summing values from Page 2 (Table 1) and Page 5 (Text). Calculate the percentage growth."

Output Format (Strict JSON):

```
{
   "difficulty_level": 0 or 1,
   "instruction_set": "Specific
reasoning instructions Γ_d for
the next agent."
}
```

Optional Interpretability Extension

For applications where stronger interpretability is desired, the Core Decision Agent prompt may optionally append the following instruction immediately before the `QUERY:` field:

After the answer, provide evidence attribution in the format:
Page(s): [page number(s)];
Evidence: [supporting evidence content];
Source: [text/table/chart/figure].

Core Decision Agent (Text Evidence Only)

You are an extractive QA model that gives answer to given query. You are given a query and a set of evidence. You have to provide specific answer from the given evidence, give your answer based only on the evidence. If you don't find the answer within the evidence provided say 'No answers found!'. Use bullet points if you have to make a list, only if necessary. For counting questions, count carefully across all evidence. Mention which page the information came from.
QUERY: {question}
STRATEGIC INSTRUCTIONS Γ_d:
{instruction_set}
EVIDENCE (from {num_pages} pages):
{evidence_summary}
YOUR ANSWER:

If page images are included, the system will automatically switch to a prompt version with visual cues.

Core Decision Agent (With Visuals)

You are an extractive QA model that gives answer to given query. You are given a query and evidence from relevant pages. You have to provide a specific, concise answer from the given evidence.
Instructions:
- Give your answer based only on the evidence provided.
- If you don't find the answer within the evidence provided say 'No answers found!'.
- Provide ONLY the shortest possible answer: a number, a name, a short phrase, or a brief list - just the key information.
- Synthesize information across ALL pages of evidence when necessary (e.g., if one page has percentage A and another has percentage B, you may need to combine them).
- For calculation questions, perform the required calculations using data from the evidence.
- For counting questions, count carefully across all evidence.
- Use bullet points only if the answer is a list.

QUERY: {question}
STRATEGIC INSTRUCTIONS Γ_d:
{instruction_set}
EVIDENCE (from {num_pages} pages):
{evidence_summary}
{visual_evidence_section}
YOUR ANSWER: