TRACE: Traversal Retrieval-Augmented Chain of Evidence for Document Understanding

Liqi He[1], Zuchao Li[2,*], Hao Huang[1], Ping Wang[3]
[1]School of Computer Science, Wuhan University, Wuhan, China
[2]School of Artificial Intelligence, Wuhan University, Wuhan, China
[3]School of Information Management, Wuhan University, Wuhan, China
{heliqi, zcli-charlie, haohuang, wangping}@whu.edu.cn

Abstract

Early Long-context Document Visual Question Answering (DocVQA) methods struggle with preserving visual semantics or handling finite context windows. Conversely, recent RAG-based approaches suffer from "semantic gaps" and "structural disconnections" due to passive retrieval mechanisms that ignore logical dependencies. To address these challenges, we introduce TRACE (Traversal Retrieval-Augmented Chain of Evidence). By navigating a Bi-Layered Graph that encodes both physical adjacency and semantic relevance, TRACE transforms retrieval from static matching into adaptive evidence chain construction. Furthermore, we propose M5BookVQA, a benchmark designed to assess deep, multi-hop reasoning in books, addressing the limitations of existing datasets. Extensive experiments show that TRACE achieves an average accuracy improvement of 14.07% on M5BookVQA and exhibits robust generalization with a 13.38% gain across four established benchmarks. Our source code is available at https://github.com/shimurenhlq/TRACE.

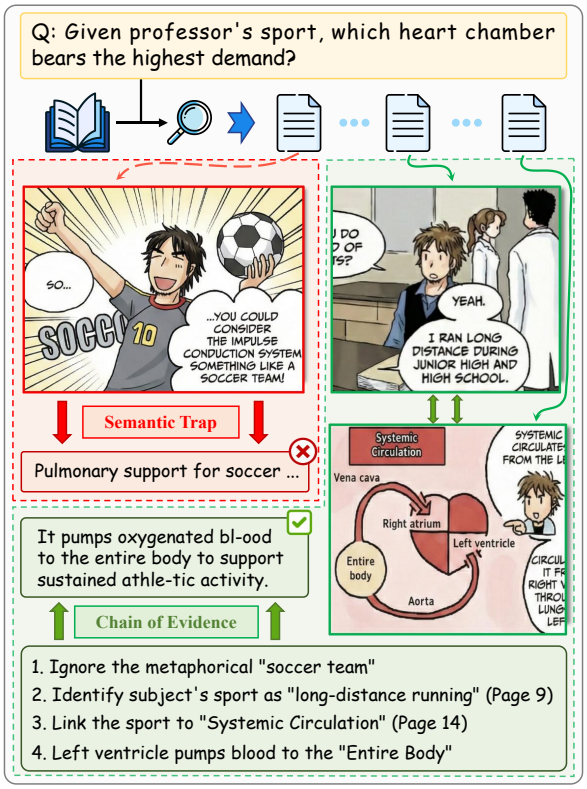


Figure 1: An illustrative example from the M5BookVQA benchmark. Traditional RAG-based methods (red) fall into a semantic trap, incorrectly identifying the metaphorical "soccer" as the subject's sport due to semantic gaps. In contrast, TRACE (green) constructs a robust chain of evidence: it correctly discerns "long-distance running" from the dialogue, links it to "systemic circulation," and visually grounds the answer to the "left ventricle" which pumps blood to the entire body.

1 Introduction

Document understanding serves as a foundational capability in Artificial Intelligence, enabling the processing of massive, unstructured information in real-world scenarios (Li et al., 2024). Within this domain, Long-context Document Visual Question Answering (DocVQA) has emerged as a cornerstone task for evaluating document intelligence. Unlike simple information extraction, real-world DocVQA requires synthesizing evidence distributed across multiple pages and embedded in heterogeneous modalities, including text, charts, and complex layouts. This multi-modal, long-context nature presents significant challenges for precise retrieval and deep reasoning.

Prevalent approaches to this task generally fall into two categories: Large Language Model (LLM)-based pipelines (Grattafiori et al., 2024; Cai

* Corresponding author. This work was supported by the National Natural Science Foundation of China (No. 62306216), the National Science and Technology Major Project (No. 2023ZD0121502), the Fundamental Research Funds for the Central Universities (No.2042025kf0090) and the National Social Science Fundation of China (No. 24&ZD186).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9806–9825
July 2-7, 2026

et al., 2024; Li et al., 2025; Team, 2025) and Vision Language Model (VLM)-based pipelines (Liu et al., 2024; Bai et al., 2025; Wang et al., 2025). Early LLM-based methods typically rely on Optical Character Recognition (OCR) (Mishra et al., 2019; Memon et al., 2020; Li et al., 2023; Wei et al., 2024) to extract text for processing. These pipelines often discard critical visual semantics, rendering them ineffective when questions hinge on visual elements like charts or layout structures. Conversely, VLM-based methods enable end-to-end processing of page snapshots, preserving visual information; however, they are fundamentally constrained by finite context windows, making it infeasible to process long-context documents (e.g., books or reports) in a single pass.

To mitigate these context constraints, recent research has adopted Retrieval-Augmented Generation (RAG) strategies (Gao et al., 2023; Zhao et al., 2024). Methods such as M3DocRAG (Cho et al., 2024) embed page snapshots into semantic vectors for similarity-based retrieval, while MDocAgent (Han et al., 2025) employs a multi-agent system to coordinate text and image retrieval. More recently, MoloRAG (Wu et al., 2025) structures page vectors into graphs to enhance connectivity. Despite these advancements, current RAG methods predominantly rely on passive matching mechanisms, which suffer from two critical failures: (1) *Semantic Gaps*, where retrieved content is lexically relevant but semantically misleading; and (2) *Structural Disconnections*, where the retrieval process ignores logical dependencies between pages.

To overcome these challenges, we present **TRACE** (**T**raversal **R**etrieval-**A**ugmented **C**hain of **E**vidence). Unlike passive approaches, TRACE leverages fine-grained question decomposition and a Bi-Layered Graph—which encodes both semantic relevance and physical adjacency of pages. TRACE functions as an adaptive tracker: it dynamically navigates the document page graph, transforming retrieval from a static matching task into adaptive evidence chain construction. This allows TRACE to bypass semantic gaps and reconstruct the complete chain of evidence to the answer. As illustrated in Figure 1, passive retrieval often leads to "Semantic Traps." A standard passive matching method incorrectly matches the keyword "soccer" (a metaphor used by the speaker) to the question about the professor's sport. The retriever fails to recognize the deeper logical context, severing the chain of evidence required to answer the multi-hop question about heart chambers.

Furthermore, the advancement of DocVQA is currently hindered by the limitations of existing benchmarks (Mathew et al., 2020; Dong et al., 2025). Datasets such as MMLongBench-Doc (Ma et al., 2024) predominantly focus on shallow retrieval and single-step reasoning (e.g., *"What is the title of the page with a screenshot?"*), failing to simulate the deep logic required for real-world problem solving. To rigorously evaluate complex reasoning capabilities, we propose **M5BookVQA** (**M**ulti-modal, **M**ulti-document, **M**ulti-hop, **M**ulti-lingual, and **M**ulti-domain **Book** **V**isual **Q**uestion **A**nswering). Comprising 2,054 carefully designed questions derived from 16,790 pages of non-fiction books, this benchmark is annotated with reasoning rationales and hop counts.

Extensive evaluations demonstrate that while prior state-of-the-art methods struggle with the complex reasoning in M5BookVQA, TRACE achieves substantial performance gains on this new benchmark while consistently improving upon four established datasets. In summary, our contributions are: **(1) M5BookVQA Benchmark:** We introduce a benchmark specifically designed to address the scarcity of deep, multi-hop reasoning scenarios. **(2) TRACE Framework:** We propose an adaptive approach that resolves semantic gaps by navigating a bi-layered graph to construct chains of evidence. **(3) Empirical Validation:** Extensive experiments demonstrate the effectiveness of our method in complex document multi-hop reasoning.

2 Related Work

2.1 Multi-page DocVQA Benchmarks

DocVQA (Mathew et al., 2020; Tanaka et al., 2023; Chia et al., 2025) has evolved into a critical benchmark for evaluating the visual understanding of VLMs. While foundational benchmarks primarily focused on single-page contexts (Mathew et al., 2020, 2022; Lee et al., 2023), they fail to adequately assess a model's ability to synthesize information dispersed across multiple pages. Consequently, recent studies have introduced benchmarks designed for cross-page retrieval (Tito et al., 2023; Ma et al., 2024; Dong et al., 2025). Despite these advancements, these datasets are often constrained by simplistic question generation pipelines and the use of short or loosely structured documents. As a result, they predominantly evaluate basic recognition and shallow retrieval capabilities rather than deep

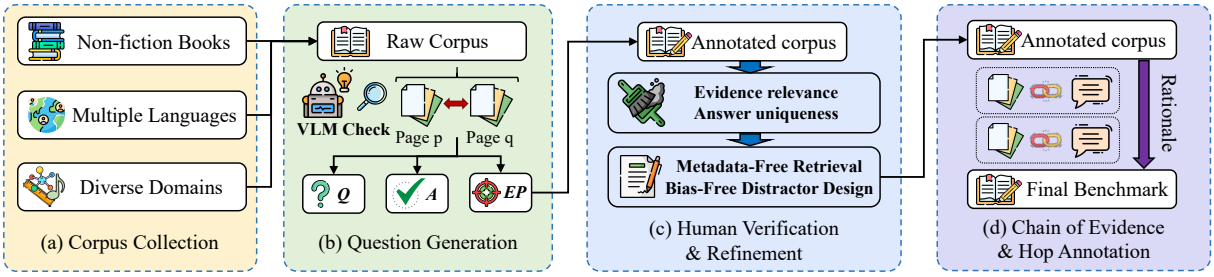


Figure 2: The data construction pipeline of M5BookVQA.

reasoning.

To bridge this gap between shallow retrieval and deep reasoning, we introduce M5BookVQA. Leveraging the extensive context and rigorous logical structure inherent to books, we design multiple-choice questions that necessitate complex, multi-hop reasoning over cross-page evidence. Unlike automated pipelines, each question in M5BookVQA undergoes strict manual verification and rewriting to ensure logical depth.

2.2 Long-context DocVQA Frameworks

Early DocVQA approaches predominantly relied on OCR to extract text (Wu et al., 2022), a process that often severs visual semantics from textual content. While the emergence of VLMs has enabled the end-to-end processing of document page snapshots, their application is fundamentally restricted by finite context windows (Liu et al., 2024). This limitation renders the simultaneous analysis of long-context documents infeasible.

To address context constraints, recent research has shifted towards RAG paradigms. Frameworks such as M3DocRAG (Cho et al., 2024), MDocAgent (Han et al., 2025), and MoLoRAG (Wu et al., 2025) utilize vector similarity or graph-based retrieval to select a subset of relevant pages. However, these methods suffer from two critical limitations. First, *Semantic Gaps* arise when direct vector matching prioritizes superficial lexical similarity over the deep semantic sufficiency required to answer complex questions. Second, *Structural Disconnections* occur as the retrieval process isolates pages from their original context, severing the logical dependencies essential for forming a complete evidence chain.

In contrast, we propose TRACE, which transforms retrieval from a passive matching task into an adaptive navigation process over a Bi-Layered Graph, dynamically constructing a coherent chain of evidence to resolve these limitations.

3 M5BookVQA Benchmark

Recent DocVQA benchmarks remain predominantly confined to shallow retrieval tasks coupled with single-step reasoning. Crucially, unlike the Natural Image VQA (Lu et al., 2022; Chen et al., 2024) domain where Chain-of-Thought (CoT) (Wei et al., 2022) reasoning is standard, existing DocVQA benchmarks lack explicit annotations for reasoning rationales, limiting the community's ability to diagnose model failures in complex reasoning. To bridge this gap, we introduce **M5BookVQA**, the first benchmark tailored for assessing cross-page, multi-hop reasoning within the rigorous context of books. As characterized in Table 1, M5BookVQA distinguishes itself through five dimensions: **M**ulti-modal, **M**ulti-document, **M**ulti-hop, **M**ulti-lingual, and **M**ulti-domain, offering significantly greater reasoning depth compared to existing datasets. To construct high-quality evaluation samples, we established a rigorous four-stage pipeline. As illustrated in Figure 2, the process evolves from corpus collection to the final rationale annotation, ensuring both the depth and validity of the benchmark. Detailed statistical analyses of the dataset are provided in Appendix A.

3.1 Corpus Construction

A distinguishing feature of M5BookVQA is its exclusive reliance on books as the knowledge source. Unlike commercial reports or loose document collections used in prior works (Cho et al., 2024; Ma et al., 2024), books possess *Holistic Contextual Integrity*—a rigorous internal logic where concepts are progressively developed across chapters. This characteristic makes books an ideal testbed for complex, long-context reasoning.

Our corpus collection adheres to three criteria: (1) strictly selecting non-fiction titles to guarantee factual correctness; (2) prioritizing genres with intrinsic multi-modal coupling (e.g., textbooks); and (3) covering 19 domains and 6 languages to ensure

Benchmark	Domain	Content Type	Language	Reasoning Depth		Metric	
				Hop Counts	Rationale	Ret.	Ans.
MP-DocVQA (Tito et al., 2023)	Industrial	T, Tab, I	English	✗	✗	✗	✓
SlideVQA (Tanaka et al., 2023)	Slides	T, I	English	✗	✓	✓	✓
MMLongBench-Doc (Ma et al., 2024)	Multi-domain	T, Tab, C, I	English	✗	✗	✓	✓
M3DocVQA (Cho et al., 2024)	Wikipedia	T, I	English	✗	✗	✓	✓
M-Longdoc (Chia et al., 2025)	Multi-domain	T	English	✗	✗	✓	✓
MMDocIR (Dong et al., 2025)	Multi-domain	T, I	English	✗	✗	✓	✗
M5BookVQA (Ours)	**Multi-domain**	**T, Tab, C, I**	**Multilingual**	✓	✓	✓	✓

Table 1: Comparison between M5BookVQA and existing multi-page DocVQA benchmarks. **Content Type**: T=Text, Tab=Table, C=Chart, I=Image. **Rationale**: Indicates whether the dataset provides explicit reasoning process (evidence grounding + step-by-step logic). **Ret.**: Supports retrieval (evidence location) evaluation.

broad knowledge distribution.

3.2 Annotation Pipeline and Quality Control

Question Generation and Constraint. In the design stage, we employ human annotators to manually craft candidate questions derived from specific book content. To ensure the validity of these questions, we leveraged advanced VLMs (Team et al., 2023; Achiam et al., 2023) to perform an "internal knowledge check." Specifically, the VLMs were tasked with answering the questions without the provided documents; questions that could be correctly answered based solely on generic background knowledge were filtered out, ensuring that the final dataset necessitates genuine document retrieval and comprehension.

Human Verification and Refinement. The subsequent verification stage focused on increasing task difficulty through *Metadata-Free Retrieval* and *Bias-Free Distractor Design*. Unlike previous benchmarks that simplify retrieval by explicitly mentioning page numbers, we strictly prohibit such metadata references. Questions must be resolved by tracing semantic cues, forcing models to perform genuine content-based retrieval. Furthermore, we engineered distractors to share similar syntactic structures and lengths with the correct answer to mitigate the exploitation of superficial patterns. To guarantee high-quality ground truth, we implement a human-centric double verification process. Every candidate sample undergoes a rigorous secondary manual review.

3.3 Reasoning Annotation

A critical innovation of M5BookVQA is the explicit annotation of *Hop Counts* and *Rationales*. Unlike prior works, we formally define a "hop" as a discrete cognitive operation (e.g., cross-page synthesis, deduction). We utilized VLMs to generate structured chains of evidence connecting questions to answers, assigning specific hop counts to quantify complexity (see Appendix B). This enables granular evaluation to distinguish between simple retrieval failures and deficits in multi-hop logic.

These annotations emphasize the benchmark's focus on deep, multi-hop reasoning, necessitating the synthesis of information across disjoint document pages. To provide a concrete perspective on these characteristics, we provide a qualitative analysis of representative case studies from different domains in Appendix H.

4 Methodology

In this section, we present TRACE. As illustrated in Figure 3, TRACE comprises three synergistic modules: Bi-Layered Graph Construction, Query Decomposition, and the Adaptive Topology Tracker.

4.1 Problem Formulation

Let $\mathcal{D} = \{p_1, p_2, \ldots, p_N\}$ denote a document containing an ordered sequence of N pages. Given a user question Q, the objective of DocVQA is to generate a natural language answer A. Traditional RAG-based approaches treat pages as independent vectors, retrieving a fixed set $\mathcal{P}_{topk} \subset \mathcal{D}$ based on cosine similarity. In contrast, TRACE aims to construct an ordered *Chain of Evidence* $\mathcal{C} = [(p_{t_1}, r_1), (p_{t_2}, r_2), \ldots]$, where each tuple represents a retrieved evidence page p_{t_k} and its corresponding reasoning analysis r_k. This chain preserves the trace trajectory required to derive A, bridging the gap between retrieval and reasoning.

4.2 Bi-Layered Graph Construction

To restore the structural integrity of fragmented pages, we model the document as a Bi-Layered Graph, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}_{phy} \cup \mathcal{E}_{sem})$.

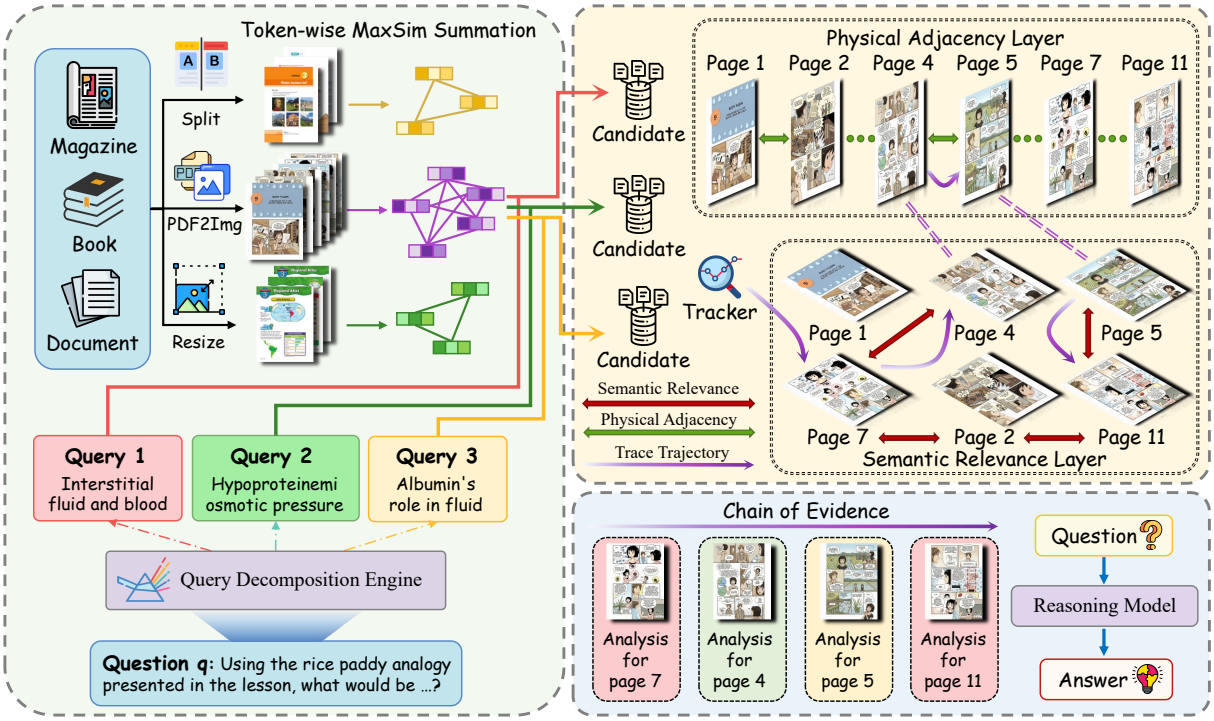


Figure 3: Overview of the TRACE framework.

Nodes ($\mathcal{V}$): Each node $v_i \in \mathcal{V}$ corresponds to a page p_i. We utilize ColPali (Faysse et al., 2024), a vision-language retriever, to map each page snapshot into a multi-vector embedding $\mathbf{E}_i \in \mathbb{R}^{n_v \times d}$.

Physical Adjacency Layer ($\mathcal{E}_{phy}$): This layer encodes the sequential narrative flow of documents. As shown in the Figure 3, directed edges are established between logically consecutive pages: $(v_i, v_{i+1}) \in \mathcal{E}_{phy}$. This enables the tracker to "turn the page," locating context that is physically adjacent and may be semantically implicit.

Semantic Relevance Layer ($\mathcal{E}_{sem}$): This layer creates "shortcuts" between logically related content scattered across disjoint pages. We compute the token-wise similarity between page embeddings using the MaxSim operator. An edge (v_i, v_j) is established in $\mathcal{E}_{sem}$ if their similarity score exceeds a threshold τ:

$$S(v_i, v_j) = \sum_{t \in \mathbf{E}_i} \max_{z \in \mathbf{E}_j} (t^\top z) > \tau \qquad (1)$$

This dual-structure graph empowers the tracker to perform both local sequential reading (via $\mathcal{E}_{phy}$) and global semantic jumping (via $\mathcal{E}_{sem}$). We employ a hierarchical optimization strategy to ensure computational feasibility, as detailed in Appendix C.

4.3 Query Decomposition and Alignment

Complex questions in M5BookVQA often encompass multiple sub-topics that a single vector query cannot capture. Direct retrieval based on Q may fail to locate pages containing only partial evidence. To address this granularity mismatch, we employ a Query Decomposition Engine (QDE). Given Q, an LLM decomposes it into a sequence of atomic, semantic-level queries $\mathcal{Q}^* = \{q_1, q_2, \ldots, q_K\}$. Each q_k serves as a specific navigational instruction, guiding the tracker to locate a distinct segment of the evidence chain (e.g., "Query 1: Interstitial fluid and blood" in Figure 3).

4.4 Adaptive Topology Tracker

The core of TRACE is the Adaptive Topology Tracker, a mechanism that actively navigates the Bi-Layered Graph $\mathcal{G}$. Unlike one-shot retrieval, the tracker maintains a persistent state to avoid local optima and redundant searching. The detailed pseudocode and hyperparameters are provided in Algorithm 1 in Appendix D. We define a global *Whitelist* $\mathcal{W}$ to store accepted evidence pages and a global *Memory* $\mathcal{M}$ to record reasoning trajectories.

For each atomic query $q_k \in \mathcal{Q}^*$, the tracker initiates a search session. It maintains a *Candidate Stack* $\mathcal{S}_k$ (initialized with the top-m pages most similar to q_k) and a query-specific *Blacklist* $\mathcal{B}_k$ to

Method	Chapter Scope				Book Scope				Global Scope			
	Acc	R@10	P@10	NDCG	Acc	R@10	P@10	NDCG	Acc	R@10	P@10	NDCG
Direct VLM Inference												
InternVL-3.5-8B	52.00	-	-	-	-	-	-	-	-	-	-	-
Qwen3-VL-8B	69.62	-	-	-	-	-	-	-	-	-	-	-
RAG Baselines												
M3DocRAG (Text)	66.02	78.00	20.28	69.68	60.52	55.50	13.84	50.97	43.87	33.71	8.64	31.62
M3DocRAG (Image)	49.12	39.21	10.20	24.59	28.82	6.27	1.43	3.79	15.73	0.87	0.19	0.46
MDocAgent	30.52	42.78	11.01	27.19	27.31	7.61	1.74	4.63	28.09	1.16	0.30	0.68
MoLoRAG	75.27	80.53	21.04	72.08	67.38	61.96	14.34	34.34	57.64	51.05	11.01	29.85
TRACE (Ours)	**84.57**	**82.68**	**21.55**	**76.01**	**80.09**	**63.46**	**14.93**	**59.96**	**77.85**	**57.07**	**12.77**	**50.60**

Table 2: **Main Results on M5BookVQA.** We compare TRACE against methods across three retrieval scopes: Chapter, Book, and Global. **Acc**uracy denotes QA Accuracy, while **R**ecall**@10**, **P**recision**@10**, and **NDCG** measure retrieval quality. TRACE demonstrates superior resilience as the search space expands.

prune irrelevant paths. The navigation proceeds iteratively:

$$p_{curr} \leftarrow \text{Pop}(\mathcal{S}_k) \tag{2}$$

If $p_{curr} \in \mathcal{B}_k \cup \mathcal{W}$, it is discarded to prevent loops. Otherwise, a VLM acts as a judge to evaluate the utility of p_{curr} given q_k:

$$\text{State}, r_{analysis} = \text{VLM}_{\text{judge}}(p_{curr}, q_k) \tag{3}$$

The tracker determines the next step based on the state. If the state is **Irrelevant**, the page is added to $\mathcal{B}_k$, and the tracker backtracks by popping the next candidate. Conversely, if the state is **Relevant**, the page and analysis are recorded in $\mathcal{W}$ and $\mathcal{M}$. To mimic human browsing, we expand the frontier by pushing neighbors onto $\mathcal{S}_k$ in reverse priority order, ensuring physical neighbors are processed first:

$$\begin{aligned} \mathcal{S}_k &\leftarrow \text{Push}(\text{TopK}(\mathcal{N}_{sem}(p_{curr}))) \\ \mathcal{S}_k &\leftarrow \text{Push}(\mathcal{N}_{phy}(p_{curr})) \end{aligned} \tag{4}$$

4.5 Reasoning with Chain of Evidence

After processing all queries, the global memory $\mathcal{M}$ contains a curated sequence of analyses, forming the "Trace Trajectory." This trajectory, along with the visual content of pages in $\mathcal{W}$, is fed into the reasoning VLM (e.g. Qwen-VL-Max). By explicitly conditioning the answer generation on this constructed chain of evidence, TRACE mitigates hallucination and ensures that the final answer A is grounded in retrieved evidence.

5 Experiments

We evaluate TRACE primarily on **M5BookVQA** for multi-hop reasoning, while extending to four long-context datasets to demonstrate generalization. To ensure a fair comparison, we utilized identical backbone models for both TRACE and the baselines across all experiments.

5.1 Performance on M5BookVQA

5.1.1 Evaluation Metrics and Baselines

To rigorously assess both retrieval quality and reasoning accuracy, we employ a diverse set of metrics. QA performance is measured by Accuracy (Acc). Retrieval performance is evaluated using three standard metrics: Recall@k (R@k) to assess evidence coverage, Precision@k (P@k) to measure relevance density, and NDCG to evaluate the ranking quality of the evidence chain. Detailed mathematical definitions and calculation formulas for these metrics are provided in Appendix E.

We compare TRACE against three distinct categories of baselines: (1) Direct VLM Inference (without retrieval); (2) Text-based RAG (OCR + Retrieval); and (3) Visual RAG (End-to-end Visual Embeddings). Table 2 summarizes the results across three hierarchical scopes: *Chapter*, *Book*, and *Global*.

5.1.2 Analysis of Direct VLM Inference

The first two rows of Table 2 present the performance of advanced VLMs (Qwen3-VL (Team, 2025) and InternVL-3.5 (Wang et al., 2025)) taking all pages as input directly. **Context Constraint.** It is crucial to note that these models were only evaluated at the *Chapter Scope*. We excluded them from the *Book* and *Global* scopes because the average book length in our dataset reaches 452.8 pages. Even with aggressive image compression, feeding entire books exceeds the feasible context windows of current models. **Retrieval Loss.** Interestingly, direct VLM inference outperforms several passive RAG methods. This highlights the retrieval bottleneck in traditional RAG: failure to identify correct evidence strictly caps the upper bound of QA accuracy.

Type	Model	Method	MMLongBench	LongDocURL	PaperTab	FetaTab	Avg.
LLM-based	Mistral-7B	Text RAG	24.47	25.06	11.45	41.14	25.53
	Qwen2.5-7B	Text RAG	25.52	27.93	12.72	40.06	26.56
	LLaMA3.1-8B	Text RAG	22.56	29.80	13.49	45.96	27.95
	GPT-4o	Text RAG	27.23	32.74	14.25	50.20	31.11
	DeepSeek-V3	Text RAG	29.82	34.73	17.05	52.36	33.49
LVLM-based	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	39.28	51.71	32.32	69.09	48.10
		TRACE	**49.06**	**52.04**	**56.49**	**88.36**	**61.48**
Multi-agent	MDocAgent		38.53	46.91	30.03	66.34	45.45

Table 3: **Generalization Performance on Established Benchmarks.** We report Question Answering accuracy under a Top-3 retrieval setting. Avg. denotes the average performance across four datasets.

Top-K	Method	MMLongBench			LongDocURL		
		Recall	Precision	NDCG	Recall	Precision	NDCG
3	M3DocRAG	64.17	31.62	54.13	67.00	33.78	58.23
	MDocAgent (Text)	43.21	20.77	37.13	58.53	29.33	54.12
	MDocAgent (Image)	64.74	31.97	54.75	66.67	33.62	58.26
	MoLoRAG	67.22	40.81	57.34	**70.04**	36.41	61.56
	TRACE	**69.22**	**44.49**	**62.93**	69.58	**37.01**	**61.95**

Table 4: **Retrieval Quality on Established Benchmarks.** We evaluate the evidence grounding capability using Recall, Precision, and NDCG under a Top-3 setting.

5.1.3 Analysis of Text-based Retrieval

For the text-based baseline, we implemented a pipeline using Hunyuan OCR (Team et al., 2025b) for full-text extraction, followed by ColBERT retrieval (Khattab and Zaharia, 2020). A counter-intuitive observation in Table 2 is that the text-based M3DocRAG significantly outperforms its image-based counterpart (66.02% vs. 49.12% Acc). This reveals a critical insight: simple visual embeddings (page snapshots) often fail to capture fine-grained semantic details required for complex reasoning in M5BookVQA. In contrast, high-quality OCR ensures that textual entities are explicitly indexed. This suggests that "naive" visual retrieval struggles with the *Semantic Gap*, where visual similarity does not equate to reasoning utility.

5.1.4 Analysis of Visual Retrieval and TRACE

Among visual RAG methods, TRACE achieves SOTA performance. Its robustness is highlighted as the scope expands: while MoLoRAG drops ∼17% in accuracy from *Chapter* to *Global*, TRACE declines only ∼7%. This superiority over graph-based baselines validates that our efficacy stems from the **Adaptive Topology Tracker** and **Chain of Evidence** rather than simple graph structuring. By actively pruning to avoid local optima, TRACE locates semantically distant but logically connected evidence. See Appendix F for cost-benefit analysis. For a more granular characterization of the performance, we provide a detailed dataset-specific breakdown analysis across languages, domains, and reasoning complexities in Appendix I.

5.1.5 Analysis of Navigation Trajectories

To analyze TRACE's actual behavior during inference, we examine the Adaptive Topology Tracker's navigation trajectories across Chapter, Book, and Global scopes in Table 5. We measure the reliance on the Physical Adjacency Layer ($\mathcal{E}_{phy}$) and Semantic Relevance Layer ($\mathcal{E}_{sem}$), alongside their Switching Frequency. Results show that reliance on $\mathcal{E}_{sem}$ grows from 53.59% to 63.51% as the scope expands, validating the necessity of "semantic jumping" for distant evidence. The Chapter scope exhibits the highest switching frequency (60.96%), indicating active alternation for fine-grained verification. Crucially, TRACE maintains a consistent ∼36% reliance on $\mathcal{E}_{phy}$ even in Global settings, proving it reconstructs coherent evidence chains by integrating global shortcuts with the document's inherent narrative structure.

Scope	$\mathcal{E}_{phy}$ Dep. (%)	$\mathcal{E}_{sem}$ Dep. (%)	SW Freq. (%)
Chapter	46.41	53.59	60.96
Book	39.16	60.84	30.65
Global	36.49	63.51	31.35

Table 5: **Quantitative Analysis of Inference Trajectories.** We report the reliance on different graph layers and the frequency of transitions between them across varying document scales on M5BookVQA. **Dep.** stands for *Dependence*; **SW Freq.** for *Switching Frequency*.

Variant	Acc	R@10	P@10	NDCG
Component Ablation				
w/o Query Decomposition	78.04	69.33	19.77	65.48
w/o Physical Adjacency	74.50	76.39	20.15	48.07
w/o Semantic Relevance	73.66	72.47	19.91	52.19
w/o Chain of Evidence	76.83	82.68	21.55	76.01
QDE Backbone				
w/ QDE: GLM-4.7	84.18	81.76	21.17	74.11
Tracker Backbone				
w/ Tracker: GPT5-mini	**85.49**	**83.06**	**21.63**	**76.30**
Reasoner Backbone				
w/ Reasoner: GLM-4.6V	79.64	82.68	21.55	76.01
w/ Reasoner: GPT-5	85.20	82.68	21.55	76.01
TRACE (Default)	84.57	82.68	21.55	76.01

Table 6: **Ablation Studies on M5BookVQA (Chapter Scope).** We analyze the contribution of core architectural components and the impact of varying backbone models for the Tracker and Reasoner. Default settings are marked in green.

5.2 Evaluation on Established Benchmarks

To verify the robustness of TRACE beyond our proposed dataset, we extended our evaluation to four established benchmarks: **PaperTab**, **FeTaTab** (Hui et al., 2024), **MMLongBench-Doc** (Ma et al., 2024), and **LongDocURL** (Deng et al., 2025).

5.2.1 Question Answering Performance

Table 3 presents the comparative results on QA tasks. For MMLongBench-Doc and LongDocURL, we report Generalized Accuracy (Acc), which employs rule-based evaluation to handle diverse answer formats. For PaperTab and FetaTab, we follow MDocAgent and MoLoRAG to employ GPT-4o as an evaluator. This evaluator assesses Binary Correctness, determining whether the model's response semantically aligns with the ground truth (assigning a score of 1 for a match and 0 otherwise).

TRACE achieves state-of-the-art performance across all four benchmarks, surpassing the strong baseline MoLoRAG by an average of 13.38%. This consistency verifies that our adaptively navigation mechanism is equally effective in diverse document formats, from scientific papers to long web reports. See Appendix G for a detailed comparison with other graph-based methods.

Notably, multi-agent frameworks like MDocAgent, despite their great reasoning capabilities, underperform compared to retrieval-centric methods (45.45% vs. 61.48% Avg). This suggests that in long-document VQA, the bottleneck lies in evidence acquisition rather than evidence processing. If the retrieval stage fails to locate the critical page ("Garbage In"), even the most advanced agents cannot derive the correct answer ("Garbage Out").

To further investigate the source of our performance gains, we evaluated retrieval metrics on datasets with ground truth page annotations (Table 4). TRACE achieves consistent improvements across the majority of retrieval metrics, validating the robustness of our fine-grained retrieval mechanism across diverse domains. Notably, incremental improvements in retrieval metrics translate into significant gains in QA Accuracy. We attribute this amplification to the global memory, which constructs a textual Chain of Evidence during navigation. By explicitly recording key insights for fine-grained queries, it mitigates hallucinations and visual information loss often caused by fragmented raw pages, effectively bridging the gap between retrieval and reasoning.

5.3 Ablation Studies

To validate the contribution of each component in TRACE, we conducted ablation studies on the M5BookVQA dataset under the Chapter Scope. Results are summarized in Table 6.

5.3.1 Impact of Architectural Components

Significance of Query Decomposition. Removing the QDE leads to a significant performance drop, with Accuracy falling by 6.53% and Recall@10 by 13.35%. Without QDE, the model relies on a single coarse-grained query vector, which fails to capture the multi-faceted nature of complex questions. This confirms that decomposing questions into atomic, semantic-level quires is essential for mitigating the granularity mismatch between questions and evidence pages.

Criticality of Physical Adjacency. The removal of the Physical Adjacency Layer results in the most severe degradation in reasoning quality. Notably, the NDCG score plummets from 76.01% to 48.07%. This sharp decline in NDCG indicates that without physical adjacency, the tracker loses the ability to

"turn the page" and reconstruct the logical order of events. Although it retrieves semantically relevant pages, it fails to form a coherent chain of evidence, thereby breaking the trace trajectory required for multi-hop reasoning.

Essentiality of Semantic Relevance. The removal of the Semantic Relevance Layer leads to a significant performance drop, with Accuracy and NDCG falling by 10.91% and 23.82% respectively. This decline indicates that without semantic edges, the tracker's navigation mechanism is restricted to purely sequential traversal, losing the ability to perform global "logical jumps" across non-contiguous pages to bridge semantically related but physically distant evidence.

Necessity of Chain of Evidence. Removing the Chain of Evidence causes a sharp 7.74% drop in Accuracy. This dissociation confirms that high-quality retrieval alone is insufficient for deep reasoning. The Chain of Evidence acts as a vital "semantic bridge," aggregating intermediate insights to guide the reasoner through disjointed raw pages, thereby preventing information loss and hallucination.

5.3.2 Impact of Backbone Scalability

QDE Backbone. To evaluate the stability of query decomposition, we replaced the default backbone with GLM-4.7 (Team et al., 2025a). The results remain highly robust, with only a marginal performance variance. This minimal discrepancy indicates the strong compatibility of the query decomposition module within the TRACE framework across different models.

Tracker Backbone. Upgrading the tracker to GPT-5-mini boosts Accuracy to 85.49% and NDCG to 76.30%. This confirms that superior visual semantic understanding enhances navigation precision—better distinguishing valid cues from noise—thereby constructing a higher-quality chain of evidence that directly benefits the downstream reasoner.

Reasoner Backbone. We analyzed the impact of the final reasoning model. Substituting the default model with GLM-4.6V (Team et al., 2025c) results in a slight performance dip, while upgrading to GPT-5 boosts Accuracy to 85.20%. Our framework provides a high-quality, structured chain of evidence that allows stronger VLMs to fully unleash their reasoning potential, translating better logic into higher accuracy.

6 Conclusion

This work tackles the bottlenecks of long-context DocVQA by proposing **TRACE** and the **M5BookVQA** benchmark. We show that passive matching mechanisms are insufficient for deep document understanding. Instead, TRACE's approach—constructing adaptive chain of evidence via a Bi-Layered Graph—proves to be a superior strategy for multi-hop reasoning. Empirical results across five benchmarks confirm that TRACE establishes a strong baseline for the field, offering consistent and significant improvements over existing LLM and VLM pipelines. Our contributions provide both a powerful tool and a rigorous testing ground for the next generation of document intelligence systems.

Limitations

Although TRACE significantly reduces reasoning errors, it may still encounter challenges in scenarios involving highly ambiguous visual cues or contradictory evidence chains where even human annotators might struggle. In such cases, the adaptive chain of evidence might occasionally retrieve structurally relevant but semantically misleading nodes. Furthermore, our current approach assumes a static document set; extending TRACE to handle dynamic or streaming document updates in real-time remains a promising avenue for future research.

Ethical Considerations

Data Compliance and Intellectual Property Protection The M5BookVQA benchmark is constructed based on commercially available non-fiction books and educational manga. We strictly adhere to copyright laws and the principles of *Fair Use* for academic research. To fully respect the intellectual property of content creators and publishers, we **do not** distribute the raw PDF files or full-page images of the books. The released dataset comprises only the metadata (Book IDs), question-answer pairs, reasoning rationales, and chain of evidence. We provide a preprocessing method that allows researchers to align their own legally purchased digital copies of the books with our annotations. This mechanism ensures that users must possess legitimate access to the source material, while facilitating reproducible research.

Annotation Process and Labor Ethics The annotation of rationales, hop counts, and QA pairs

was conducted entirely by the authors of this paper to ensure high-quality reasoning chains. No crowd-sourced platforms or external contractors were involved in the data creation process. Consequently, this study adheres to fair labor standards and entails no issues regarding underpaid labor or the exploitation of crowd workers. All annotators were fully aware of the data's intended usage and the open-source nature of the annotations.

Privacy, Bias, and Content Safety We have manually screened the dataset to ensure it does not contain sensitive Personally Identifiable Information (PII) of private individuals. Since the corpus consists of published non-fiction books, the content generally reflects established knowledge. However, we acknowledge that book contents may inherently carry historical or cultural biases present in the source material. During the selection process, we actively excluded books containing hate speech, violence, or sexually explicit content. We advise users of M5BookVQA to remain aware of potential implicit biases when analyzing model outputs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. *arXiv preprint arXiv:2502.13923*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^{3}CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2025. M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9244–9261, Suzhou, China. Association for Computational Linguistics.

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. *CoRR*, abs/2411.04952.

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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1135–1159. Association for Computational Linguistics.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multi-modal retrieval for long documents. *Preprint*, arXiv:2501.08828.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

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

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. 2025. Rag-anything: All-in-one RAG framework. *CoRR*, abs/2510.12323.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

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. *arXiv preprint arXiv:2503.13964*.

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. UDA: A benchmark suite for retrieval augmented generation in real-world document analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. *CoRR*, abs/2004.12832.

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 *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qiwei Li, Zuchao Li, Xiantao Cai, Bo Du, and Hai Zhao. 2023. Enhancing visually-rich document understanding via layout structure modeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 4513–4523. ACM.

Qiwei Li, Zuchao Li, Ping Wang, Haojun Ai, and Hai Zhao. 2024. Hypergraph based understanding for document semantic entity recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2950–2960. Association for Computational Linguistics.

Zuchao Li, Yonghua Hei, Qiwei Li, Lefei Zhang, Ping Wang, Hai Zhao, Baoyuan Qi, and Guoming Liu. 2025. What limits bidirectional model's generative capabilities? A uni-bi-directional mixture-of-expert method for bidirectional fine-tuning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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. 2024. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE access*, 8:142642–142668.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, Qi Yang, Qiming Peng, Bin Luo, Hower Yang, Xinsong Zhang, Jinnian Zhang, Houwen Peng, Hongming Yang, Senhao Xie, and 12 others. 2025b. Hunyuanocr technical report.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng,

Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025c. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

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

Xueyao Wan and Hang Yu. 2025. Mmgraphrag: Bridging vision and language with interpretable multimodal knowledge graphs. *arXiv preprint arXiv:2507.20804*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

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. *CoRR*, abs/2409.01704.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xinya Wu, Duo Zheng, Ruonan Wang, Jiashen Sun, Minzhen Hu, Fangxiang Feng, Xiaojie Wang, Huixing Jiang, and Fan Yang. 2022. A region-based document VQA. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4909–4920. ACM.

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.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

A Detailed Dataset Statistics

To provide a comprehensive understanding of the complexity and diversity of **M5BookVQA**, we present detailed statistical analyses across four key dimensions: evidence page counts, reasoning hops, language distribution, and domain coverage.

A.1 Distribution of Evidence Page Counts

Unlike traditional benchmarks where answers are often located on a single page, M5BookVQA emphasizes cross-page information synthesis. As illustrated in Figure 4, the distribution of evidence page counts exhibits a long-tailed pattern. The majority of questions require integrating evidence from 2 pages (728 samples) or 3 pages (406 samples), with some complex queries necessitating information spanning up to 10 pages. This distribution validates the benchmark's capability to evaluate long-context modeling and cross-page retrieval.

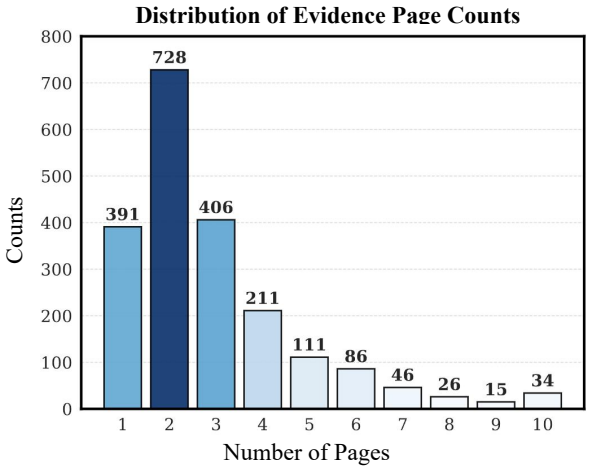


Figure 4: Distribution of evidence page counts per question in M5BookVQA. The dataset is characterized by a high proportion of multi-hop reasoning samples.

A.2 Distribution of Reasoning Hops

A core contribution of M5BookVQA is its focus on deep reasoning. Figure 5 depicts the distribution of reasoning hop counts annotated for each question. The data reveals a rigorous logic depth, with the mode centered at 3 hops (849 samples), followed closely by 4 hops (493 samples). Notably, the dataset includes highly complex questions requiring up to 13 reasoning steps. This contrasts sharply with existing datasets dominated by single-step (1-hop) retrieval tasks, ensuring a more challenging testbed for chain-of-evidence construction.

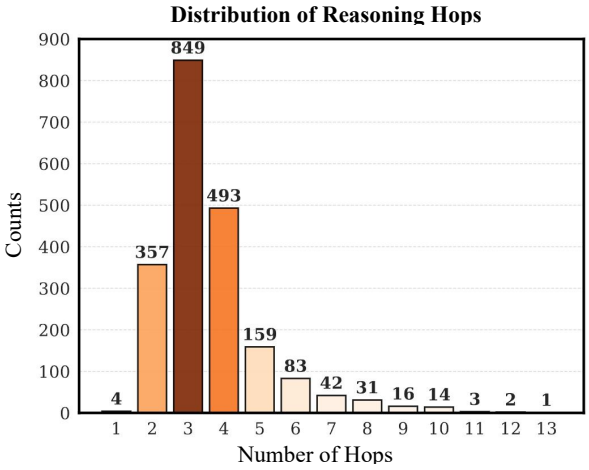


Figure 5: Distribution of reasoning hops. The benchmark predominantly features questions requiring 3 to 4 steps of deductive reasoning.

A.3 Language Diversity

To assess the multilingual capabilities of MLLMs, M5BookVQA incorporates books in six major languages. As shown in Figure 6, while English (50.2%) and Chinese (36.4%) constitute the primary corpus, the dataset also includes significant representation from Russian (5.3%), Japanese (4.7%), French (2.2%), and German (1.2%). This linguistic diversity allows for the evaluation of models in both high-resource and mid-to-low-resource language scenarios.

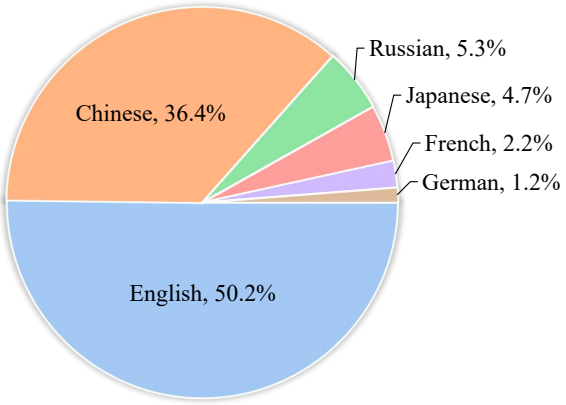


Figure 6: Language distribution of the M5BookVQA corpus, covering six major languages.

A.4 Domain Coverage

Real-world document understanding requires handling diverse knowledge domains. Figure 7 presents the hierarchical domain distribution of our corpus. The dataset spans four broad categories: Humanities & Art, Natural Sciences, Engineering,

and Social Sciences. Within these categories, it covers 19 specific sub-domains ranging from history and philosophy to physics, medicine, and computer science. This broad coverage ensures that the benchmark evaluates a model's generalizability across specialized terminologies and varied document formats.

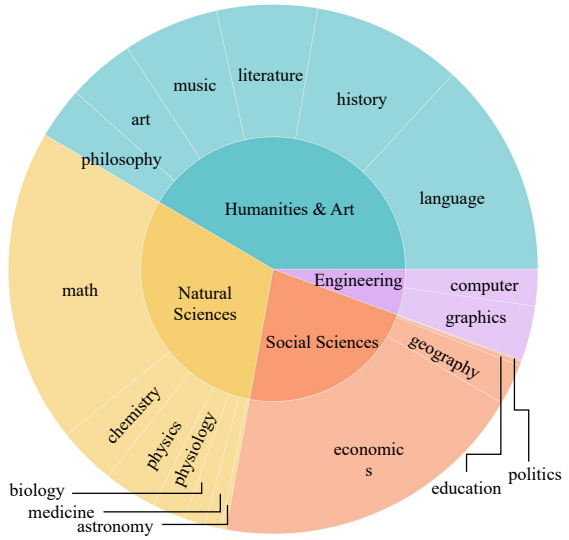


Figure 7: Hierarchical domain distribution of M5BookVQA, spanning 19 distinct sub-domains across arts, sciences, and engineering.

B Automated Rationale and Hop Count Annotation

To maintain consistency and scalability in annotating reasoning chains, we employed a VLM-based automated annotation pipeline. Figure 8 illustrates the specific system prompt designed for this task.

Unlike standard QA tasks where the model predicts the answer, our annotation stage provides the ground-truth answer to the VLM. The model is instructed to perform a "Reverse Engineering" analysis. This ensures that the generated rationales are factually aligned with the correct answer rather than hallucinated. As defined in the prompt, the process involves three distinct steps:

1. **Grounding:** Explicitly locating the visual or textual evidence within the provided page snapshots.
2. **Rationalization:** Constructing a coherent logical chain that explains *why* the answer is derived from the evidence.
3. **Hop Counting:** Quantifying the reasoning complexity using a defined metric.

Rationale Annotation Prompt

You are an expert interdisciplinary researcher and logical analyst.

Task Definition:
Perform a "Reverse Engineering" analysis. Since the answer is known, you must:
1. Grounding: Locate evidence in the provided images.
2. Rationalization: Construct a logical chain explaining *why* the answer is correct.
3. Hop Counting: Calculate the "Reasoning Hops" required.

Hop Counting Rules (Strict):
- 0 Hops: Explicitly stated in a single location.
- +1 Hop (Cross-Page): Synthesized from different pages.
- +1 Hop (Cross-Modal): Combining text and visual elements within the same page.
- +1 Hop (Deduction): Requires logical inference (A>B, B>C -> A>C).

Figure 8: Structured Prompt for Rationale Annotation. We utilize a "Reverse Engineering" approach where the model, given the ground-truth answer, is tasked with locating evidence (Grounding), constructing the logical chain (Rationalization), and quantifying the cognitive load (Hop Counting) based on strict rules.

Strict Hop Counting Rules. To standardize the difficulty assessment, we enforce a rigorous set of rules for calculating "Reasoning Hops," as detailed in the bottom half of Figure 8:

- **0 Hops:** Assigned when information is explicitly stated in a single location.
- **+1 Hop (Syntactic/Structural):** Incurred when information must be synthesized across different pages (Cross-Page) or combined from text and visual elements (Cross-Modal).
- **+1 Hop (Logical):** Incurred when logical inference (e.g., deduction $A \rightarrow B \rightarrow C$) is required to bridge the gap between evidence and answer.

C Scalable Graph Construction Strategy

Constructing a fully connected semantic layer requires computing pairwise similarities between all pages. To address the computational bottleneck in large-scale documents, we propose an **Adaptive Granularity Strategy** that adjusts the interaction mechanism based on the document scope.

Fine-grained Interaction (Micro-Scope). For local graphs within a single chapter (N_{small}), we employ full Token-wise MaxSim over the multi-vector embeddings. Denoting T as the number of visual tokens per page and D as the dimension, the complexity is $O(N_{small}^2 \cdot T^2 \cdot D)$. This retains rich, token-level nuances essential for distinguishing closely related pages (e.g., two consecutive pages discussing similar topics), with negligible latency due to the limited scale of N_{small}.

Coarse-to-Fine Interaction (Macro-Scope). For book-level or global retrieval (N_{large}), the quadratic complexity of MaxSim is computationally prohibitive. We therefore adopt a coarse-to-fine two-stage strategy:

1. **Chapter Localization (Coarse-grained):** We first condense each page into a single vector $v_i \in \mathbb{R}^D$ using global mean pooling. We perform efficient dense retrieval to identify the top-k candidate pages based on cosine similarity. Critically, we do not treat these pages as final results but as *chapter anchors*. We map these candidates to their respective chapters and aggregate them to form a concise set of *Candidate Chapters*, effectively filtering out irrelevant sections of the book.

2. **Page Refinement (Fine-grained):** Within the identified Candidate Chapters, we revert to Token-wise MaxSim using the full multi-vector embeddings. This allows us to pinpoint the precise evidence pages with high semantic fidelity.

This hierarchical approach effectively addresses the *Efficiency-Effectiveness Trade-off*. It utilizes efficient dense approximations to narrow the search space from the entire book to a few relevant chapters (reducing complexity from $O(N^2T^2)$ to $O(N^2)$), while reserving expensive high-precision computations for the final localization.

D Algorithm Details

We provide the detailed pseudocode for the Adaptive Topology Tracker in Algorithm 1. This algorithm outlines the iterative process of navigating the Bi-Layered Graph for a single atomic query. Specifically, we set the semantic relevance threshold $\tau = 0.7$ for graph construction. Furthermore, the upper bound of accepted pages for each query round is dynamically determined as $\lfloor N_{total}/N_{sub} \rfloor$, where N_{total} represents the total page budget and N_{sub} denotes the number of decomposed atomic queries.

Algorithm 1 Adaptive Topology Tracking for Query q_k

Require: Query q_k, Graph $\mathcal{G}$, Global Whitelist $\mathcal{W}$, Global Memory $\mathcal{M}$
Ensure: Updated $\mathcal{W}, \mathcal{M}$ ▷ Output: Updated states

Initialize: Stack $\mathcal{S}_k \leftarrow \text{TopSim}(q_k, \mathcal{V})$, Blacklist $\mathcal{B}_k \leftarrow \emptyset$
while $\mathcal{S}_k \neq \emptyset$ **and** $|\mathcal{W}| <$ Limit **do**
  $p_{curr} \leftarrow \mathcal{S}_k.\text{pop}()$
  if $p_{curr} \in \mathcal{W} \cup \mathcal{B}_k$ **then**
    continue
  end if
  $state, r_{analysis} \leftarrow \text{VLM}_{\text{judge}}(p_{curr}, q_k)$
  if $state ==$ RELEVANT **then**
    $\mathcal{W}.\text{add}(p_{curr})$
    $\mathcal{M}.\text{append}(\{q_k, p_{curr}, r_{analysis}\})$
    $\mathcal{N}_{phy} \leftarrow \text{GetPhyNbrs}(p_{curr})$
    $\mathcal{N}_{sem} \leftarrow \text{GetSemNbrs}(p_{curr})$
    $\mathcal{S}_k.\text{push}(\text{TopK}(\mathcal{N}_{sem}))$
    $\mathcal{S}_k.\text{push}(\mathcal{N}_{phy})$
  else
    $\mathcal{B}_k.\text{add}(p_{curr})$ ▷ Prune irrelevant path
  end if
end while
return $\mathcal{W}, \mathcal{M}$

E Evaluation Metrics Details

In this section, we provide the formal definitions of the retrieval metrics used in our evaluation. Let $\mathcal{P}_{rel}$ denote the set of ground-truth relevant pages for a given query, and $\mathcal{P}_{ret}^k$ denote the sequence of the top-k retrieved pages.

Recall@k (R@k) measures the proportion of relevant pages successfully retrieved within the top-k results:

$$\text{R@k} = \frac{|\mathcal{P}_{rel} \cap \mathcal{P}_{ret}^k|}{|\mathcal{P}_{rel}|} \tag{5}$$

Precision@k (P@k) measures the proportion of retrieved pages that are actually relevant:

$$\text{P@k} = \frac{|\mathcal{P}_{rel} \cap \mathcal{P}_{ret}^k|}{k} \tag{6}$$

NDCG@k (Normalized Discounted Cumulative Gain) evaluates the ranking quality, rewarding algorithms that place relevant pages higher in the list.

Method	Relative Latency	Relative API Calls
MoLoRAG	1.0×	1.0×
TRACE	≈ 4.0×	≈ 10.0×

Table 7: Comparison of average latency and API usage per question ($N = 30$ samples). Values are normalized relative to MoLoRAG.

It is defined as:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k} \tag{7}$$

where DCG (Discounted Cumulative Gain) and IDCG (Ideal DCG) are calculated as:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \tag{8}$$

$$\text{IDCG@}k = \sum_{i=1}^{\min(k,|\mathcal{P}_{rel}|)} \frac{1}{\log_2(i+1)} \tag{9}$$

Here, $rel_i \in \{0, 1\}$ represents the binary relevance of the page at rank i (1 if relevant, 0 otherwise).

F Latency and Computational Cost Analysis

To explicitly evaluate the computational overhead of TRACE compared to the strong baseline MoLoRAG, we conducted a comparative analysis using 30 randomly sampled queries from the M5BookVQA dataset.

F.1 Quantitative Comparison

As summarized in Table 7, TRACE exhibits higher latency and API consumption. On average, the inference latency of TRACE is approximately **4**× that of MoLoRAG, and the number of VLM API calls is roughly **10**× higher.

F.2 Mechanism Analysis and Justification

The disparity in computational cost stems from the fundamental difference in the retrieval paradigms of the two methods:

- **MoLoRAG (Static Scoring):** It relies on a "VLM Score" mechanism that processes a fixed, pre-determined number of candidates (Top-K). The computational graph is static, meaning the number of VLM inference steps is explicitly bounded and constant, resulting in lower and more predictable latency.
- **TRACE (Adaptive Navigation):** Our method employs an "Agentic VLM Navigation" via the Adaptive Topology Tracker. Unlike static reranking, the tracker is **adaptive**: it dynamically decides whether to expand to physical/semantic neighbors or prune the current branch based on real-time feedback. This process mimics human browsing, where the agent may perform multiple rounds of "look-ahead" and "backtracking" to uncover hidden evidence chains, naturally incurring higher VLM interaction costs.

The Efficiency-Effectiveness Trade-off. While TRACE is computationally more intensive, we argue that this "slowness" is a worthy trade-off for the substantial reliability gains in complex long-context reasoning. As demonstrated in the main experiments, TRACE achieves a significant **14.07% improvement in Accuracy** on M5BookVQA. In high-stakes domains targeted by our benchmark (e.g., extracting medical or legal advice from books), the cost of "hallucination" or "missed evidence" far outweighs the cost of additional GPU seconds. TRACE prioritizes *precision* over *speed*, ensuring that the answer is derived from a logically complete Chain of Evidence rather than a hurried guess.

Method	Accuracy (%)
Naive RAG (Lewis et al., 2020)	26.20
GraphRAG (Guo et al., 2024)	28.10
MMGraphRAG (Wan and Yu, 2025)	38.80
RAG-Anything (Guo et al., 2025)	42.80
TRACE (Ours)	**49.06**

Table 8: **Comparison with Graph-based RAG Methods on MMLongBench-Doc.** TRACE significantly outperforms entity-centric graph methods without requiring complex entity extraction.

G Comparison with Entity-Centric GraphRAG Methods

Table 8 presents a comparative analysis of TRACE against recent graph-based retrieval methods on the MMLongBench-Doc benchmark. As shown, TRACE achieves an accuracy of 49.06%, significantly outperforming the standard GraphRAG (28.1%) and the robust RAG-Anything baseline (42.8%).

Structural Divergence. The performance gap stems from a fundamental difference in graph

topology construction. Traditional methods like GraphRAG and MMGraphRAG adopt an *entity-centric* approach, where nodes represent fine-grained extracted entities (e.g., specific terms, objects) and edges represent extracted relationships. While effective for knowledge graphs, this approach suffers in document understanding tasks due to the high risk of extraction errors and the fragmentation of continuous narrative contexts.

Logical vs. Entity Navigation. In contrast, TRACE employs a *page-centric* logical graph. Instead of relying on fragile entity extraction, our Bi-Layered Graph models the document's inherent structure—connecting pages via physical adjacency and semantic continuity. This allows the tracker to navigate the "macro-logic" of the document rather than getting lost in the "micro-details" of isolated entities.

Efficiency Advantage. Furthermore, TRACE eliminates the computationally expensive preprocessing pipeline required by entity-centric methods, such as Named Entity Recognition (NER), relation extraction, and complex document chunking. By treating the page as the atomic unit of retrieval, TRACE achieves superior performance with a streamlined and more scalable architecture.

H Case Study for M5BookVQA

To explicitly demonstrate the complexity of the M5BookVQA benchmark, we present a detailed analysis of two representative samples. These cases highlight the necessity for Cross-Page Retrieval and Multi-Hop Reasoning, distinguishing our dataset from traditional single-page VQA tasks.

H.1 Cross-Page Logic in Business Domain

Figure 9 presents a question derived from a Business textbook.

- **The Challenge:** The question stems from a "Role Play" activity located on Page 155, asking which sales method is *NOT* typical for a "clicks-and-mortar" company. The explicit definition of "clicks-and-mortar" (describing the integration of physical stores and online presence) is provided in the main text on Page 149.

- **Reasoning Chain:** To answer correctly, the model must:

  1. **Navigate:** Identify the keyword "clicks-and-mortar" from the question on Page 155 and retrieve the defining evidence from Page 149.
  2. **Synthesize:** Comprehend that "clicks-and-mortar" implies both physical and digital presence.
  3. **Deduce (Negation):** Evaluate the options to find the one that contradicts the synthesized definition (i.e., identifying that "Online sales only" is incorrect for this business model).

  This example validates the model's ability to maintain context over a long document span and perform logical negation based on retrieved evidence.

H.2 Multi-Modal Synthesis in Art Domain

Figure 10 illustrates a fine-grained visual understanding task from an Art textbook regarding "Cardboard Creativity."

- **The Challenge:** The question asks to identify the incorrect pairing of an artwork and its crafting technique (e.g., "Swan - Assembly and pasting"). The evidence is scattered: the visual examples are shown on Page 27, while the specific categorization of techniques (e.g., "Paper Relief" vs. "Comprehensive Application") requires cross-referencing text and captions.

- **Reasoning Chain:** The model needs to perform a multi-step verification:

  1. **Visual Grounding:** Recognize the specific visual objects (Lion, Swan, Pavilion) mentioned in the options.
  2. **Evidence Alignment:** Trace each object to its corresponding technique section in the book. For instance, the model must confirm that the "Swan" is visually depicted in the "Paper Relief" section, not the "Assembly" section.
  3. **Conflict Detection:** Compare the retrieved ground truth ("Swan" $\rightarrow$ "Paper Relief") against Option E ("Swan" $\rightarrow$ "Assembly") to detect the factual mismatch.

  This case exemplifies the benchmark's demand for precise visual-textual alignment and the ability to verify multiple assertions within a single query.

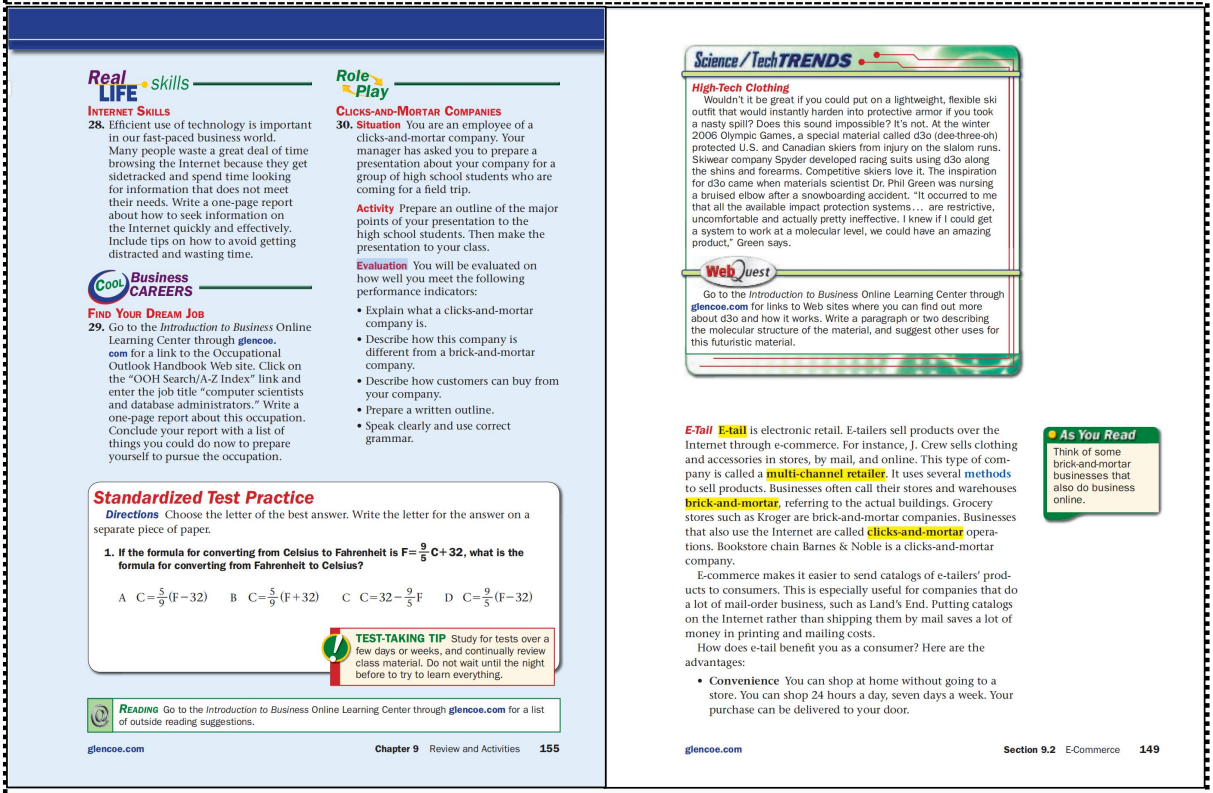
Question: In the Role Play activity at the end of the "Technology and Business" chapter, referring to the third objective in the Evaluation section, if a student is playing an employee of a clicks-and-mortar company, which of the following methods of sale, as defined on a preceding page, is NOT one that this type of company typically uses?

Options:

A. Only through e-commerce
B. Through a brick-and-mortar physical store
C. Through a company website
D. Through online sales only
E. Both B and C
F. Both A and D

Answer: F

Domain: Economics



Real LIFE Skills

INTERNET SKILLS

28. Efficient use of technology is important in our fast-paced business world. Many people waste a great deal of time browsing the Internet because they get sidetracked and spend time looking for information that does not meet their needs. Write a one-page report about how to seek information on the Internet quickly and effectively. Include tips on how to avoid getting distracted and wasting time.

Cool Business CAREERS

FIND YOUR DREAM JOB

29. Go to the *Introduction to Business* Online Learning Center through **glencoe.com** for a link to the Occupational Outlook Handbook Web site. Click on the "OOH Search/A-Z Index" link and enter the job title "computer scientists and database administrators." Write a one-page report about this occupation. Conclude your report with a list of things you could do now to prepare yourself to pursue the occupation.

Role Play

CLICKS-AND-MORTAR COMPANIES

30. **Situation** You are an employee of a clicks-and-mortar company. Your manager has asked you to prepare a presentation about your company for a group of high school students who are coming for a field trip.

Activity Prepare an outline of the major points of your presentation to the high school students. Then make the presentation to your class.

Evaluation You will be evaluated on how well you meet the following performance indicators:

- Explain what a clicks-and-mortar company is.
- Describe how this company is different from a brick-and-mortar company.
- Describe how customers can buy from your company.
- Prepare a written outline.
- Speak clearly and use correct grammar.

Standardized Test Practice

Directions Choose the letter of the best answer. Write the letter for the answer on a separate piece of paper.

1. If the formula for converting from Celsius to Fahrenheit is $F=\frac{9}{5}C+32$, what is the formula for converting from Fahrenheit to Celsius?

A $C=\frac{5}{9}(F-32)$ B $C=\frac{5}{9}(F+32)$ C $C=32-\frac{9}{5}F$ D $C=\frac{9}{5}(F-32)$

TEST-TAKING TIP Study for tests over a few days or weeks, and continually review class material. Do not wait until the night before to try to learn everything.

READING Go to the *Introduction to Business* Online Learning Center through **glencoe.com** for a list of outside reading suggestions.

glencoe.com Chapter 9 Review and Activities 155

Science/Tech TRENDS

High-Tech Clothing

Wouldn't it be great if you could put on a lightweight, flexible ski outfit that would instantly harden into protective armor if you took a nasty spill? Does this sound impossible? It's not. At the winter 2006 Olympic Games, a special material called d3o (dee-three-oh) protected U.S. and Canadian skiers from injury on the slalom runs. Skiwear company Spyder developed racing suits using d3o along the shins and forearms. Competitive skiers love it. The inspiration for d3o came when materials scientist Dr. Phil Green was nursing a bruised elbow after a snowboarding accident. "It occurred to me that all the available impact protection systems . . . are restrictive, uncomfortable and actually pretty ineffective. I knew if I could get a system to work at a molecular level, we could have an amazing product," Green says.

WebQuest

Go to the *Introduction to Business* Online Learning Center through **glencoe.com** for links to Web sites where you can find out more about d3o and how it works. Write a paragraph or two describing the molecular structure of the material, and suggest other uses for this futuristic material.

E-Tail **E-tail** is electronic retail. E-tailers sell products over the Internet through e-commerce. For instance, J. Crew sells clothing and accessories in stores, by mail, and online. This type of company is called a **multi-channel retailer**. It uses several **methods** to sell products. Businesses often call their stores and warehouses **brick-and-mortar**, referring to the actual buildings. Grocery stores such as Kroger are brick-and-mortar companies. Businesses that also use the Internet are called **clicks-and-mortar** operations. Bookstore chain Barnes & Noble is a clicks-and-mortar company.

E-commerce makes it easier to send catalogs of e-tailers' products to consumers. This is especially useful for companies that do a lot of mail-order business, such as Land's End. Putting catalogs on the Internet rather than shipping them by mail saves a lot of money in printing and mailing costs.

How does e-tail benefit you as a consumer? Here are the advantages:

- **Convenience** You can shop at home without going to a store. You can shop 24 hours a day, seven days a week. Your purchase can be delivered to your door.

As You Read

Think of some brick-and-mortar businesses that also do business online.

glencoe.com Section 9.2 E-Commerce 149

step 1: Found definition stating clicks-and-mortar companies use both physical stores and the Internet.
step 2: Noted visual organizer separating e-tail (online only) from clicks-and-mortar.
step 3: Matched the role-play evaluation asking how customers buy with the definitions.
step 4: Determined that online-only methods (A and D) are NOT typical for clicks-and-mortar, so F.

Figure 9: **M5BookVQA Sample Case 1 (Economics).** This sample illustrates a cross-page reasoning task where the model must link the "Role Play" instructions with the definition of "Clicks-and-Mortar" to identify the exclusion criteria.

问题：“纸板的创想——成型方法”一课中展示了多种纸艺范例。下列哪一选项将范例作品与其所属的具体技法分类错误地配对了？

Question: In the lesson 'Cardboard Creativity—Forming Methods,' various paper art examples are shown. Which option incorrectly matches a work with its specific technique?

Options:

A. 鹰，剪、折纸的方法（Eagle, Cutting and folding）
B. 马，纸的组合粘贴方法（Horse, Assembly and pasting）
C. 亭子，综合应用（Pavilion, Comprehensive application）
D. 狮子，纸浮雕造型（Lion, Paper relief）
E. 天鹅，纸的组合粘贴方法（Swan, Assembly and pasting）
F. 草莓，纸浮雕造型（Strawberry, Paper relief）

Answer: F　　***Domain:*** Art

step 1：定位到“纸浮雕造型”标题与配图，确认“狮子”和“天鹅”均在此板块中。（Locate the "Paper relief" section; confirm both "Lion" and "Swan" appear there.）
step 2：定位到“3. 综合应用——亭子”段落与配图，确认亭子属于综合应用。（Locate "Comprehensive application — Pavilion" to confirm its classification.）
step 3：得出天鹅为纸浮雕造型，从而判断选项E（天鹅➡纸的组合粘贴方法）与教材不符。（Identify "Swan" as paper relief , making Option E (Assembly/Pasting) incorrect.）
step 4：验证其他配对（亭子➡综合应用）为正确，从而支持E为唯一错误配对。（Verify "Pavilion" is correct , confirming E is the mismatched pair.）

Figure 10: **M5BookVQA Sample Case 2 (Art).** This sample demonstrates a multi-modal verification task. The model is required to navigate across multiple pages to align visual instances (e.g., "Swan") with their specific textual crafting techniques, identifying the mismatched pair among the options.

Scope	English	Chinese	Russian	Japanese	French	German
Chapter	89.53	80.48	83.33	67.71	80.43	79.17
Book	88.08	71.79	75.00	63.54	78.26	87.50
Global	88.18	67.51	63.89	60.42	76.09	91.67

Table 9: **Accuracy Breakdown by Language.** Performance of TRACE across different scopes and linguistic subsets of M5BookVQA.

Scope	Humanities & Art	Natural Sciences	Engineering	Social Sciences
Chapter	77.84	87.60	87.83	92.12
Book	69.05	87.44	82.61	89.93
Global	64.71	86.65	80.00	89.72

Table 10: **Accuracy Breakdown by Domain.** Comparative analysis across the four major hierarchical categories defined in Figure 7.

I Dataset-Specific Breakdown Analysis

To provide deeper insights into M5BookVQA and TRACE's performance, we conduct breakdown analyses across language, domain, and complexity. The experimental settings remain consistent with Table 2.

Table 9 presents the accuracy breakdown by language. We observe that performance trends across languages are inconsistent as the scope expands, which is primarily attributed to the inherent differences in document selections for each linguistic subset. In contrast, the domain-wise analysis in Table 10 reveals a consistent decline in Accuracy across all subjects as the retrieval scope increases from Chapter to Global. Finally, results in Table 11 underscore the significant challenge of precisely locating complex evidence distributions within expansive search spaces, particularly for multi-hop tasks requiring 5–10 pages.

Scope	1-2 Pages	3-4 Pages	5-10 Pages
Chapter	86.95	82.31	80.56
Book	82.57	76.79	77.44
Global	81.05	73.70	74.61

Table 11: **Accuracy Breakdown by Evidence Page Count.** Evaluation of TRACE's resilience as reasoning complexity (number of required evidence pages) increases.