SV-RAG: LoRA-Contextualizing Adaptation of MLLMs for Long Document Understanding

Jian Chen[1]*, **Ruiyi Zhang**[2]*, **Yufan Zhou**[2], **Tong Yu**[2], **Franck Dernoncourt**[2]
Jiuxiang Gu[2], **Ryan Rossi**[2], **Changyou Chen**[1], **Tong Sun**[2]
University at Buffalo[1], Adobe Research[2]
ruizhang@adobe.com

Abstract

Multimodal large language models (MLLMs) have recently shown great progress in text-rich image understanding, yet they still struggle with complex, multi-page visually-rich documents. Traditional methods using document parsers for retrieval-augmented generation suffer from performance and efficiency limitations, while directly presenting all pages to MLLMs leads to inefficiencies, especially with lengthy ones. In this work, we present a novel framework named **S**elf-**V**isual **R**etrieval-**A**ugmented **G**eneration (SV-RAG), which can broaden horizons of *any* MLLM to support long-document understanding. We demonstrate that **MLLMs themselves can be an effective multimodal retriever** to fetch relevant pages and then answer user questions based on these pages. SV-RAG is implemented with two specific MLLM adapters, one for evidence page retrieval and the other for question answering. Empirical results show state-of-the-art performance on public benchmarks, demonstrating the effectiveness of SV-RAG.

1 Introduction

Documents serve as a critical medium for the preservation and dissemination of information, with millions produced annually. These documents are not limited to simple text; they encompass complex layouts and a variety of modalities such as text, tables, charts, and images. Visually-rich document understanding (VDU) is thus an essential and challenging area of research. Recently, Multimodal Large Language Models (MLLMs) has emerged, showcasing remarkable abilities to process and understand documents. These models span both proprietary and open-source domains, like GPT-4o (OpenAI, 2023), Gemini-1.5 (Team et al., 2023), and Claude-3 among closed-source models, and InternLM-XC2-4KHD (Dong et al., 2024), InternVL-Chat (Chen et al., 2023b), LLaVA-NeXT (Liu et al., 2024a), Mini-CPM (Hu et al., 2024), mPLUG-DocOwl (Ye et al., 2023b), and TextMonkey (Liu et al., 2024d) in open-source space. Their performance has been particularly notable in single-page DU tasks demonstrated on datasets like DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022) and InfoVQA (Mathew et al., 2022).

In real-world applications, they often present documents that are much longer, containing dozens or hundreds of pages(Ma et al., 2024d; Tanaka et al., 2023; Islam et al., 2023; Zhu et al., 2021). Addressing the understanding of such lengthy documents presents MLLMs with new challenges (Ma et al., 2024d). One way is to utilize a classical document parser (Rausch et al., 2021) to extract information and formulate a prompt for LLM (Wang et al., 2023; Lamott et al., 2024), which is difficult to recover the layout in prompts and suffers performance degeneration from the document parser. The other way is to exploit the long context windows of large models, allowing them to take multiple pages at once. However, most of the input pages are not relevant to user requests, and efficiency will be compromised when the document contains hundreds of pages Ma et al. (2024d); Islam et al. (2023) or there is a document collection (Tito et al., 2021).

In this work, we first retrieve evidence pages to obtain relevant information within a vast and varied landscape of content. Unlike using a classical document parser, we propose using MLLMs as the information encoder, which have shown great generalization ability as they have been trained on a huge text corpus. After obtaining the embedding of each page, we further utilize contextualized

[*] Equal contribution, work done when JC is at Adobe Research.

late interaction for relevance scoring (Khattab & Zaharia, 2020). This design shows significantly better efficiency and accuracy than using the classical document parser to extract information. Top-k pages are then selected from hundreds of pages and provided to MLLMs to answer user questions on documents.

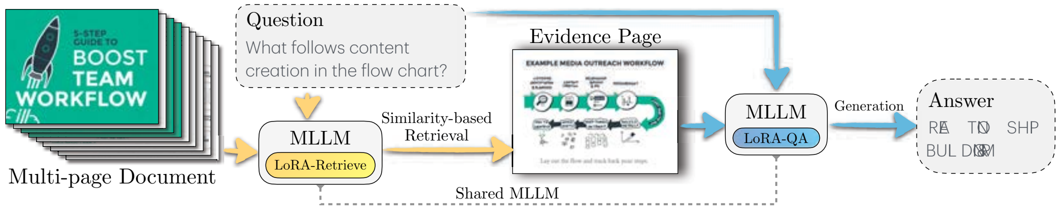


Figure 1: Overview of the SV-RAG pipeline. The multi-page document and query are encoded by a customized MLLM (yellow). The most relevant page is retrieved through similarity-based matching, and a fine-tuned MLLM (blue) generates the final answer from the evidence.

Based on this design demonstrated in Figure 1, we introduce the SV-RAG framework for multi-page document understanding, which includes modules for evidence page retrieval and answer generation. Our contributions can be summarized as follows.

- We propose a novel framework named SV-RAG to broaden the horizons of MLLMs, where we use intermediate MLLMs hidden embedding for **efficient** question-based evidence page retrieval.
- We finetune MLLMs through dual LoRA adapters for evidence page retrieval and question answering, respectively, enabling SV-RAG to be edge-friendly with great memory efficiency.
- We collect a visually-rich document QA dataset, VisR-Bench, comprising nine domains including magazine, flyer, newsletter, product manual, and presentations, etc. This dataset is built upon web-crawl documents, containing 226 documents and 471 question answer pairs.
- We empirically show that SV-RAG, with only 4B parameters, achieves state-of-the-art performance on VisR-Bench and four public benchmarks, rivaling Gemini-1.5-pro on MMLongBench-Doc and demonstrating its effectiveness.

2 Related Work

Visually-rich Document Understanding Visual Document Understanding (VDU) is the field focused on interpreting text-rich images such as tables (Zhong et al., 2019), charts (Masry et al., 2022), and webpage screenshots (Liu et al., 2024c; Tanaka et al., 2021). These images are complex, featuring a mix of text and visual elements that convey abundant information (Gu et al., 2024). To evaluate multimodal document understanding, tasks range from low-level recognition tasks, such as information extraction, to high-level cognitive tasks, such as visual question answering (Mathew et al., 2020). Models in VDU are typically divided into two categories: OCR-dependent (Xu et al., 2020) and OCR-free, based on their reliance on Optical Character Recognition (OCR). OCR-dependent models are often trained to synchronize textual and visual data. For instance, UDoP (Tang et al., 2023) is pre-trained to restore obscured textual and layout details using both image and partial text inputs. OCR-free approaches must include text recognition training. Dount (Kim et al., 2022) is an example of an OCR-free model that focuses on producing unbroken text sequences, disregarding structural details. In contrast, Pix2Struct (Lee et al., 2023a), another OCR-free model, focuses on interpreting the structure by creating HTML DOM trees from webpage screenshots. However, this technique does not easily transfer to other image types. Our method focuses on the visual question-answering task, specifically targeting questions over long documents consisting of multiple pages of multimodal information.

Multimodal Retrieval-Augmented Generation Augmenting language models with information from various knowledge sources has been found to boost their performance in different NLP tasks. The Dense Passage Retriever (DPR) (Karpukhin et al., 2020) trains its retrieval mechanism with documents from within the same batch, using contrastive learning with negatively sampled examples, which enhances its capabilities in open-domain question answering. Document Screenshot Embedding (DSE) (Ma et al., 2024c) uses MLLMs as encoders for both document screenshots and queries, training through contrastive learning to achieve enhanced multimodal retrieval. Both REALM

(Guu et al., 2020) and Retrieval-Augmented Generation (RAG) (Gao et al., 2023b) consider the passages they retrieve as hidden variables and train the retrieval and generation components together, improving the efficiency of the retrieval-augmented generation approach. Taking cues from textual RAG, the Plug-and-play (Chen et al., 2024d) approach uses GradCAM (Selvaraju et al., 2020) to fetch pertinent image segments corresponding to a given query. The MuRAG (Chen et al., 2022) model introduces a multimodal retrieval-augmented Transformer that utilizes an external multimodal memory for language generation enhancement. Unlike other approaches that retrieve information from various knowledge sources, SV-RAG focuses on retrieving relevant evidence pages from a given document. This helps MLLMs generate accurate and explainable answers based on the retrieved content. MM-GEM Ma et al. (2024a) trains an MLLM with a hybrid loss for similarity computation and caption generation on natural images. In contrast, our approach targets visually rich documents, using two LoRA modules to specialize in each task.

Multimodal Large Language Models While Large Language Models (LLMs) excel at text-only question answering (QA) (Dasigi et al., 2021; Lee et al., 2023b), they cannot process other modalities. To enable multimodal tasks like Visual Question Answering (VQA), MLLMs transform images and videos into visual tokens that LLMs can understand. To train these MLLMs, MiniGPT-4 (Zhu et al., 2023) leverages ChatGPT to produce data compliant with high-quality instructions, while LLaVA (Liu et al., 2023b) prompts GPT-4 with image captions and bounding boxes. Chen et al. (2023a; 2024a) have prompted OpenAI GPT-4V to generate more than 1M pieces of quality data to train MLLMs. LLaMA-Adapter (Zhang et al., 2023; Gao et al., 2023a) and mPLUG-Owl (Ye et al., 2023b) align text and image features with large-scale image-text pairs. InstructBLIP (Dai et al., 2023) has restructured 13 vision-language tasks to fit an instruction-based approach. mPLUG-Owl (Ye et al., 2023a;b) implements multi-task instruction fine-tuning with public document datasets. Recent research (Liu et al., 2023a; 2024a; Bai et al., 2023; Dong et al., 2024; Xu et al., 2024; Luo et al., 2024) improves visual encoders by increasing resolution, leading to significant advances in downstream applications but also raising memory costs, especially in multi-page tasks. TextMonkey Liu et al. (2024d) compresses visual tokens using a token resampler. Our method extends MLLMs to handle multi-page documents by retrieving relevant pages, reducing computation and distractions from long token sequences.

3 SV-RAG Method

Multi-page document understanding aims to answer questions related to long and complex documents containing both text and images from users. We denote a document of n-pages as a sequence of images, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Text token sequence of a question q is denoted as $\{q_1, q_2, \ldots, q_n\}$. Traditional approaches that begin with a parsing step to extract content elements such as images, tables, and forms from documents, then generate answers based on these contents using LLMs (Saad-Falcon et al., 2023; Wang et al., 2023). Here, we first consider using MLLMs to handle this task and avoiding the heuristic document parsing process, where we directly convert each page into a single image. It is not desired, as most pages in a document are irrelevant to user questions and performing an evidence page retrieval can further enhance the efficiency.

We introduce SV-RAG, a method that efficiently leverages the capabilities of pre-trained MLLMs for long document question-answering (QA). SV-RAG can broaden the horizon of MLLMs to answer questions over long documents or document collections with hundreds of pages. This finding is based on the fact that hidden states of MLLMs can be effective page representations for question-based retrieval, as shown in Section 5.6. This representation ability can be further enhanced with contrastive training using a LoRA adapter, demonstrating surprising retrieval performance of MLLMs. Furthermore, we can finetune a LoRA-adpter of QA to further enhance the performance of SV-RAG on specific domains. In summary, we first retrieve evidence pages to rank these images based on their relevance score to a given question q, then select the most relevant images, which are then fed into the MLLM to generate the answer. In this section, we introduce the SV-RAG architecture in Section 3.1, retrieval training in Section 3.2 and dual-adapter designs in Section 3.3.

3.1 Architecture

Figure 2 presents an overview of our model architecture, which comprises two MLLM-based modules for the retrieval of evidence pages and question answers.

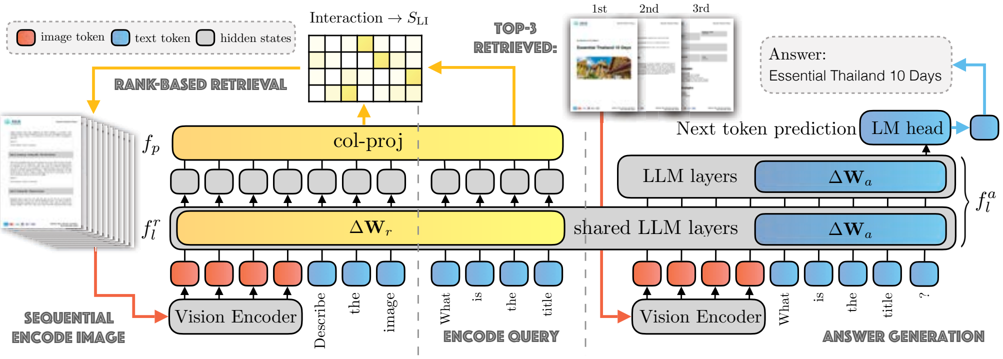


Figure 2: Model overview of SV-RAG. It contains two modules, which are finetuned using LoRA (Hu et al., 2021), sharing the **same** pretrained multimodal LLM backbone. The retrieval module selects evidence pages for the other QA module, which provides responses to user questions.

Col-Retrieval Module Building on the approach introduced in ColPali (Faysse et al., 2024), we employ a modified MLLM for retrieval, comprising a vision encoder f_v, a large language model (LLM) f_l^r, and a Col-projection layer f_p. For an input image $\mathbf{X}$, the vision encoder computes a sequence of visual embeddings $f_v(\mathbf{X})$, which are then concatenated with token embeddings $\mathbf{y}_v$ derived from a fixed text prompt: "\nDescribe the image." This combined input is fed into the LLM. The projection layer f_p then transforms the LLM's last hidden states into a low-dimensional feature space, resulting in feature sequences that can be represented as $\mathbf{E}_v = f_p(f_l^r(f_v(\mathbf{X}), \mathbf{y}_v))$. Similarly, for an input question q, the question is first augmented into y_q using a prompt template. Then, its token embedding $\mathbf{y}_q$ is processed without visual input as $\mathbf{E}_q = f_p(f_l^r(\mathbf{y}_q))$. Finally, a late-interaction score $s_{\text{LI}}(\mathbf{E}_q, \mathbf{E}_v)$ is computed between the feature sequences, measuring the relevance of a page image to the question text. More details about scoring method is provided in section 3.2.

Question-Answering Module The QA module uses a classic LLaVA-like architecture (Liu et al., 2024b), utilizing a vision encoder f_v to compute visual embeddings, which are combined with token embeddings and processed by an LLM f_l^a. The LLM then generates text answers autoregressively through next-word prediction.

3.2 Contextualized Late Interaction

We utilize the contextualized late interaction (Col) technique (Khattab & Zaharia, 2020) to compute relevance scores for evidence retrieval. Unlike traditional single-vector encoders, such as CLIP (Radford et al., 2021), the Col technique introduces an inter-sequence similarity metric called the late-interaction score, which captures more fine-grained question-image relevance. Formally, the late-interaction score between a text feature sequence $\mathbf{E}_q = \{\mathbf{e}_{q_1}, \ldots, \mathbf{e}_{q_n}\}$ of length n and a visual feature sequence $\mathbf{E}_v = \{\mathbf{e}_{v_1}, \ldots, \mathbf{e}_{v_m}\}$ of length m is defined as:

$$s_{\text{LI}}(\mathbf{E}_q, \mathbf{E}_v) = \sum_{i=1}^{n} \max_{j \in \{1,..,m\}} \mathbf{e}_{q_i} \cdot \mathbf{e}_{v_j}^T. \tag{1}$$

We use it as a similarity score in contrastive learning to facilitate ranked retrieval. Specifically, we train our retrieval module to maximize the late-interaction score between a question and its corresponding evidence image, considering these as positive pairs. We then identify the most similar, but unassociated, image within the batch to form the hardest negative pair and minimize the score for this pair. Figure A.1 shows a training pair example. The loss function is defined as:

$$\mathcal{L} = \log(1 + \exp(s_{\text{LI}}(\mathbf{E}_q, \mathbf{E}_v^-) - s_{\text{LI}}(\mathbf{E}_q, \mathbf{E}_v^+))). \tag{2}$$

The training process of the Col-retrieval module is summarized in Algorithm 1.

3.3 Parameter Sharing via Dual-Adapter

To reduce memory usage, we optimize the model by sharing a single MLLM that includes both the vision encoder f_v and the language model f_l across both the retrieval and QA modules. To

Algorithm 1 Col-retrieval training

```
Require: Pre-trained MLLM $\{f_v,\ f_l^r\}$, training batch of evidence image and question pairs $\{(\mathbf{X}_1, \mathbf{y}_1), \cdots, (\mathbf{X}_b, \mathbf{y}_b)\}$.
Initialize the Col-projection layer $f_p$.
while not converged do
    Get $\mathbf{E}_v^i = f_p(f_l^r(f_v(\mathbf{X}_i), \mathbf{y}_i))$, $i \in \{1, ..., b\}$.
    Get $\mathbf{E}_q^i = f_p(f_l^r(\mathbf{y}_i))$, $i \in \{1, ..., b\}$.
    Compute $\mathbf{S}_{i,j} = s_{\text{LI}}(\mathbf{E}_q^i, \mathbf{E}_v^j)$.
    Get negative image index $\hat{i}$ for each $\mathbf{y}_i$: $\hat{i} = \arg\max_{j \in \{1,\dots,b\}, j \neq i}(\mathbf{S}_{i,j})$
    Gradient update using loss function Eq.(2),
     where $\mathbf{E}_v^{j-} = \mathbf{E}_{\hat{i}}$.
end while
```

accommodate the different tasks required by each module, we insert two sets of adapters into the f_l using the LoRA method (Hu et al., 2021). In the retrieval module, we use a set of adapters $\Delta\mathbf{W}_r$ to create the retrieval-LLM, f_l^r. For the QA module, a different set of adapters $\Delta\mathbf{W}_a$ is added to the f_l, creating the QA-LLM, f_l^r. In this way, we support both tasks using a single LLM and vision encoder, adding only $\sim$2% additional parameters.

4 VisR-Bench

Visually-rich Document Selection About 4,000 PDF documents are crawled from the Web and contents of these documents are extracted via a document parser[1]. We keep the document with figures and throw away text-only or scan documents. To select documents with specific types of figures, we build a figure scheme that includes 19 figure types after reviewing different documents. We find some types of figures are not informative, such as logo and banner. We use the pretrained CLIP model `ViT-L/14-336` (Radford et al., 2021) to perform a figure classification on the extracted figures and keep 6 out of 19 types of figures, including tables, maps, diagrams, infographics, data charts, workflows, and screenshots. After that, we also annotate the document types for all selected documents.

Question-Answer Collection Document parser returns all document elements in JSON format and the figures are saved separately as image files. We retrieve the JSON file for the document to obtain the contexts of each figure. Then we combine the figures with their contexts and use GPT-4o (API version 2024-02-15-preview) to generate QA pairs. For the GPT-4o prompts, we provide two demonstrations and ask GPT-4o to generate a QA pair. In addition, we perform automatic verification using GPT-4o to ensure the quality of the generation. Specifically, we only provide the figure to GPT-4o and ask it with the generated question; if GPT-4o can answer it correctly, we will keep that QA pair in the SV-RAG Bench. This heuristic filter ensures that the answers are from document figures and double-checks the correctness of generated answers.

Dataset Statistics VisR-Bench contains 226 documents and 471 human-verified question-answer pairs. Figure 3 shows the distributions of the document types and the length distribution by document type. VisR-Bench has a great diversity of documents compared to previous work (Tanaka et al., 2023; Islam et al., 2023; Ma et al., 2024d).

5 Experiments

We assess the performance of SV-RAG in evidence page retrieval and visual question answering capabilities. We first evaluate the retrieval accuracy of the Col-retrieval module within SV-RAG and compare it with several baselines on SlideVQA (Tanaka et al., 2023), MM-LongBench (Ma et al., 2024d), DUDE (Van Landeghem et al., 2023), DocVQA (Mathew et al., 2020; 2021) and VisR-Bench. We then conduct experiments on question answering using SV-RAG and compare the results with other LMM baselines, inlcuding single-page and cross-page VQA. All experiments are implemented

[1] Adobe Extract API: https://developer.adobe.com/document-services/apis/pdf-extract/

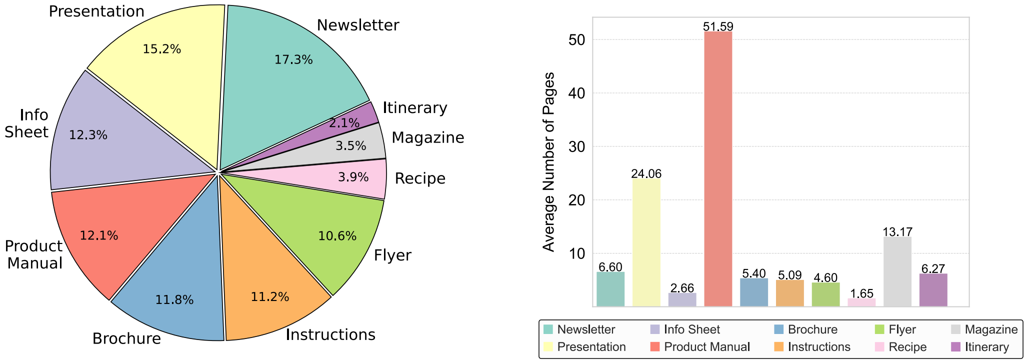


Figure 3: Distribution of document types (left) and average document lengths in each types (right).

with PyTorch and conducted on Nvidia A100 GPUs. The Col-retrieval modules are fine-tuned for 4 epochs with a batch size of 32 and a learning rate of 5e-5, using the AdamW optimizer and LoRA adapters on all linear layers in the LLM. The LoRA rank is set to 32.

5.1 Datasets

Finetuning Dataset We train our Col-retrieval modules using the original training data of ColPali (Faysse et al., 2024), which includes 39,463, 10,074, 13,251, 10,000, and 45,940 question-page pairs filtered from DocVQA, InfoVQA (Mathew et al., 2022), TATDQA (Zhu et al., 2024), arXivQA (Li et al., 2024), and synthetic data across various topics, including government reports, healthcare, artificial intelligence, and energy. We incorporated DocMatix-IR (Ma et al., 2024b) and PFL-DocVQA (Tito et al., 2023b), using GPT-4o to filter out duplicate images and unsuitable questions. The expanded dataset improved top-1 retrieval accuracy on MMLongBench-Doc by $\sim$1% without affecting other benchmarks. We fine-tuned our QA modules using the training split of the SlideVQA dataset (Tanaka et al., 2023). The SlideVQA dataset contains 1,919 slides in the training set, 400 in the test set, and 300 in the development set, with each slide consisting of 20 pages. The training split includes 10,290 samples, each annotated with questions, answers, and corresponding evidence.

Evaluation Dataset We evaluated our method's performance on four public datasets—SlideVQA, MMLongBench-Doc (Ma et al., 2024d), DocVQA (Mathew et al., 2021), and DUDE (Van Landeghem et al., 2023)—along with our proposed VisR-Bench dataset. The evaluation was conducted in both single-evidence (SP) and cross-evidence (MP) settings, where questions require information from either a single page or multiple pages within a long document. For DocVQA, we used 5,349 SP and 5,187 MP QA pairs from the validation split. Similarly, we combined the test and dev splits of SlideVQA to form 2,995 SP and 763 MP QA pairs for evaluation. For DUDE, we evaluated 6,307 QA pairs from the validation split.

MMLongBench-Doc, which consists of 135 PDF documents averaging 50.4 pages (ranging from 9 to 468 pages), contains 1,081 QAs in total. From these, we extracted 488 single-evidence QAs to assess the performance of MLLMs designed for single-image tasks. Additionally, we report the results of our best-performing model across all categories in MMLongBench-Doc, providing a comprehensive comparison against state-of-the-art LMMs.

5.2 Evaluation Metrics

We evaluate our model's performance on evidence retrieval and question-answering using several key metrics: Top-k Accuracy, Exact Match (EM) (Tanaka et al., 2023), Generalized Accuracy (G-Acc) (Ma et al., 2024d), Average Normalized Levenshtein Similarity (ANLS) (Biten et al., 2019), and Partial Normalized Levenshtein Similarity (PNLS) (Chen et al., 2024b). A detailed explanation of each metric can be found in Appendix B.

5.3 Comparative Retrieval Accuracy Analysis

We evaluated the accuracy of the Col-retrieval module in SlideVQA, MMLongBench-Doc, SP-DocVQA, and VisR-Bench, comparing it with the baseline methods including CLIP (ViT-L/14) (Radford et al., 2021), SigLip (so400m-patch14-384) Zhai et al. (2023), BM25 (Robertson et al., 2009), SBERT (Reimers & Gurevych, 2019), BGE-M3 Chen et al. (2024c), BGE-large Xiao et al. (2023), and NV-Embed-v2 Lee et al. (2024). For encoder models, we used their text and image encoders to compute the cosine similarity between the feature of the question and the page. For text-based methods, the text content in the MMLongBench-Doc and SV-RAG Bench datasets is extracted using a document parser to ensure higher accuracy. For SlideVQA and SP-DocVQA, where only scanned images are available, the text is extracted using Paddle-OCR[2].

	SlideVQA		**MMLong**		**VisR-B**		**SP-DocVQA**	
accuracy	top1	top5	top1	top5	top1	top5	top1	top5
Text-based Methods								
BM25	69.3	91.1	25.3	47.6	32.2	57.5	30.9	61.7
SBERT	73.0	91.0	44.7	70.2	38.8	72.1	47.4	74.0
BGE-M3	74.3	92.0	42.7	66.6	47.7	78.1	47.8	77.5
Bge-large	81.3	93.3	47.4	71.5	53.7	80.3	56.7	81.5
NV-Embed-v2	82.2	94.3	47.4	69.0	55.2	82.7	51.7	80.2
Encoder Models								
CLIP	58.4	86.9	32.4	63.4	33.4	62.1	37.1	69.4
SigLip	66.2	90.1	44.9	69.4	53.2	81.3	39.3	71.9
Col-Retrieval Modules								
Col-Paligemma	89.0	98.7	60.7	82.0	67.9	90.8	62.3	85.9
Col-InternVL2	88.5	98.3	61.3	83.0	69.3	90.7	63.2	85.9
Col-Phi-3-vision	90.6	98.8	64.8	84.8	71.9	91.8	65.1	87.0

Table 1: Retrieval accuracy results on four datasets, where MMLong refers to MMLongBench-Doc, SV-RAG-B refers to SV-RAG-Bench. Bold font indicates the best model.

The results indicate that Col-retrieval outperforms all baselines, achieving more than 98% in top-5 retrieval accuracy on the SlideVQA dataset, where each slide consists of 20 pages. However, performance decreases on other datasets as the data become more complex and document lengths increase significantly.

5.4 Main Results

We compared the performance of our method with popular lightweight LMMs on document question answering tasks, using PaliGemma (Beyer et al., 2024), Phi-3-v (Abdin et al., 2024), and InternVL2-4B (Chen et al., 2023b) as the backbone LMMs for both retrieval and QA modules, following the dual adapter design from Section 3.3. We fine-tuned the retrieval module using the 118,695 training question-page pairs used in ColPali (Faysse et al., 2024). The QA module is fine-tuned using SlideVQA's training split. We reported the original evaluation metrics used in prior works, including EM, G-Acc, and ANLS, and additionally reported PNLS, which better evaluates LLM-generated responses.

Table 2 presents the comparison results. We first evaluate SV-RAG on single-evidence questions from SP-SlideVQA, MMLongBench-Doc, and SP-DocVQA, where the required information is on a single page. To demonstrate the question-answering capabilities of LMMs, we include four "cheating" baselines where models are given the ground truth evidence page. Next, we test SV-RAG on cross-evidence questions from MP-SlideVQA, MP-DocVQA, and DUDE, where information spans multiple pages. We only test SV-RAG with InternVL2-4B backbone, since the other two LMM are pre-trained for single-page understanding. SV-RAG's performance is compared with classical encoder-only and encoder-decoder models, including BERT (Kenton & Toutanova, 2019), Longformer (Beltagy et al., 2020), Big Bird (Zaheer et al., 2020), T5 (Raffel et al., 2020), Hi-VT5 (Tito et al., 2023a), and LayoutLMv3 (Huang et al., 2022), with results taken from the best settings in the original

[2] PaddleOCR: https://github.com/PaddlePaddle/PaddleOCR

papers. InternVL2-8B and GPT-4o, processing all pages, serve as the state-of-the-art baselines for open-source and proprietary multipage LMMs, respectively. We demonstrate how the limitations of the retrieval and QA modules can impact overall performance through challenging examples from the SlideVQA dataset, as shown in Appendix C. Additional comparisons with text-only baselines utilizing a document parser are provided in Appendix G.1.

Table 2: **Quantitative Results in Multi-Page QA**: "#Param" refers to number of parameters. "Evidence" reports evidence setting: T (true evidence page), A (all pages), and Rk (top-k retrieved). Reported metrics include PNLS, Exact Match, Generalized Accuracy, and ANLS. † indicates models with LoRA adapter on QA module. Results for all encoder/decoder models are taken from their respective papers, with "-" indicating missing or not applicable results. Bold font indicates the best open-source model, excluding cheating baselines.

Method	#Param	Evidence	**SP-SlideVQA**		**MMLongBench**		**SP-DocVQA**	
			EM	PNLS	G-Acc	PNLS	ANLS	PNLS
Single-Page Evidence								
Cheating Baselines								
PaliGemma	3B	T	37.30	0.63	23.9	0.38	0.65	0.79
Phi-3-v	4B	T	13.72	0.80	33.7	0.52	0.65	0.85
InternVL2	4B	T	15.03	0.58	40.4	0.55	0.84	0.88
GPT-4o	-	T	30.59	0.84	56.8	0.62	0.87	0.94
Multi-image MLLMs								
InternVL2	8B	A	12.62	0.65	14.1	0.22	0.50	0.55
GPT-4o	-	A	27.28	0.81	54.5	0.57	0.69	0.80
SV-RAG Models (Proposed)								
SV-RAG-PaliGemma	3B	R1	35.03	0.60	23.9	0.35	0.56	0.69
SV-RAG-PaliGemma†	3B	R1	49.75	0.65	23.1	0.38	0.56	0.68
SV-RAG-Phi-3-vision	4B	R1	12.85	**0.78**	30.7	**0.50**	0.55	0.75
SV-RAG-Phi-3-vision†	4B	R1	**58.13**	0.77	28.4	0.44	0.68	0.73
SV-RAG-InternVL2	4B	R5	16.40	0.58	33.2	0.48	0.70	**0.76**
SV-RAG-InternVL2†	4B	R5	45.07	0.77	**34.0**	0.49	**0.71**	0.75
Cross-Page Evidence								
Method	#Param	Evidence	**MP-SlideVQA**		**MP-DocVQA**		**DUDE**	
			EM	PNLS	ANLS	PNLS	ANLS	PNLS
Encoder/Decoder models								
BERT-Large	334M	-	-	-	0.53	-	0.25	-
Longformer	148M	-	-	-	0.55	-	0.27	-
Big Bird	131M	-	-	-	0.58	-	0.26	-
T5-Base	223M	-	-	-	0.51	-	0.42	-
LayoutLMv3	125M	-	-	-	0.55	-	0.20	-
Hi-VT5	316M	-	-	-	0.62	-	0.23	-
Multi-image MLLMs								
InternVL2	8B	A	17.04	0.53	0.68	0.75	0.37	0.56
GPT-4o	-	A	16.09	0.73	0.67	0.79	0.54	0.70
SV-RAG Models (Proposed)								
SV-RAG-InternVL2	4B	R5	24.25	**0.61**	0.70	**0.76**	0.36	**0.57**
SV-RAG-InternVL2†	4B	R5	**31.98**	0.59	**0.71**	**0.76**	**0.45**	0.54

Retrieval vs Multipage We observe SV-RAG consistently outperforms InternVL2-8B, across various settings. The primary issue with LMMs is that long documents are transformed into excessively long visual token sequences, leading to significant memory burdens, as reported later in section 5.5. In datasets like MMLongBench-Doc and DocVQA, some documents exceed hundreds of pages, causing out-of-memory errors, even on servers with 8 × A100 (80GB) GPUs. In such cases, we assigned a zero score in our experiments. In contrast, GPT-4o exhibits strong multi-page reasoning capabilities. However, the accuracy of the cheating baseline slightly surpasses that of using all pages, as providing only the evidence pages helps GPT-4o avoid distractions from irrelevant information in the longer context. Moreover, SV-RAG with InternVL2-4B backbone perform slightly better than the one with Phi-3-vision backbone on MMLongBench-Doc and SP-DocVQA, possibly due to the improvement in retrieval accuracy by using top-5 pages, which is more crucial for longer documents.

Impact of Fine-tuning We observe that SV-RAG QA modules with PaliGemma and InternVL2-4B backbones show a significant increase in EM on the SlideVQA dataset, surpassing their cheating baselines after fine-tuning on SlideVQA. The model with the Phi-3-vision backbone shows notable improvements in Exact Match (EM) scores without gains in PNLS, suggesting that fine-tuning primarily enhanced the model's attention and answer formatting. This could be because the pre-trained model was already optimized for these question types. Nevertheless, as shown in Figure D.1, we empirically find that fine-tuning still improves answering performance. However, we notice a performance drop for fine-tuned SV-RAG-Phi-3-vision on MMLongBench-Doc, indicating that fine-tuning can harm LLM generalization. A similar trend is seen with the InternVL2-4B backbone on the DUDE dataset.

Comparison with SOTA LMMs Finally, we present the complete results of SV-RAG-InternVL2-4B on the MMLongBench-Doc dataset to highlight the advantages of our method. As shown in Table 3, our model, with only 4 billion parameters, outperforms all open-source LMMs and achieves performance comparable to proprietary models such as Claude-3 Opus and Gemini-1.5-Pro.

Method	#Param	Evidence Source					Evidence Page			ACC	F1
		TXT	LAY	CHA	TAB	FIG	SIN	MUL	UNA		
Open-source Models											
DeepSeek-VL-Chat	7.3B	7.2	6.5	1.6	5.2	7.6	5.2	7.0	12.8	7.4	5.4
Idefics2	8B	9.0	10.6	4.8	4.1	8.7	7.7	7.2	5.0	7.0	6.8
MiniCPM-Llama3-V2.5	8B	11.9	10.8	5.1	5.9	12.2	9.5	9.5	4.5	8.5	8.6
InternLM-XC2-4KHD	8B	9.9	14.3	7.7	6.3	13.0	12.6	7.6	9.6	10.3	9.8
mPLUG-DocOwl 1.5	8.1B	8.2	8.4	2.0	3.4	9.9	7.4	6.4	6.2	6.9	6.3
Qwen-VL-Chat	9.6B	5.5	9.0	5.4	2.2	6.9	5.2	7.1	6.2	6.1	5.4
Monkey-Chat	9.8B	6.8	7.2	3.6	6.7	9.4	6.6	6.2	6.2	6.2	5.6
CogVLM2-LLaMA3-Chat	19B	3.7	2.7	6.0	3.2	6.9	3.9	5.3	3.7	4.4	4.0
InternVL-Chat-v1.5	26B	14.0	16.2	7.1	10.1	16.6	14.9	12.2	**17.5**	14.6	13.0
EMU2-Chat	37B	6.1	9.7	2.6	3.8	7.7	5.7	6.1	16.5	8.3	5.5
SV-RAG Models (Proposed)											
SV-RAG-InternVL2 (R5)	4B	**26.5**	18.8	22.3	19.6	23.6	33.2	**13.1**	12.4	22.2	22.8
SV-RAG-InternVL2† (R5)	4B	26.3	**22.1**	**25.0**	**20.7**	**25.2**	**34.0**	10.6	15.7	**23.0**	**24.2**
Proprietary Models											
Claude-3 Opus	-	24.9	24.7	14.8	13.0	17.1	25.6	13.8	7.6	17.4	18.1
Gemini-1.5-Pro	-	21.0	17.6	6.9	14.5	15.2	21.1	11.1	69.2	28.2	20.6
GPT-4V	-	34.4	28.3	28.2	32.4	26.8	36.4	27.0	31.2	32.4	31.2
GPT-4o	-	46.3	46.0	45.3	50.0	44.1	54.5	41.5	20.2	42.8	44.9

Table 3: **Performance of various models on MMLongBench-Doc.** Questions are categorized in two ways: (1) by evidence source type—text (TXT), layout (LAY), chart (CHA), table (TAB), and image (IMG); and (2) by evidence pages—single-page (SIN), cross-page (MUL), and unanswerable (UNA). Models using LoRA adapters fine-tuned on SlideVQA for the QA module are marked with †. Bold font indicates the best open-source model. The results of baseline models are adopted from the original MMLongBench-Doc paper Ma et al. (2024d).

5.5 Efficiency of Different Models

To evaluate the efficiency of SV-RAG, we conducted experiments on the SlideVQA dataset, which has 20 pages per question with a resolution of 1024x768. We recorded peak GPU memory usage and time costs for retrieval and QA modules separately. The GPU memory is manually recorded using the `nvidia-smi` command, which tends to report higher numbers than the actual memory required by the application due to overhead and memory management processes. We tested backbones including PaliGemma, Phi-3-v, and InternVL2-4B, all equipped with LoRA adapters. Since PaliGemma and Phi-3-v are single-page models, we used top-1 retrieved image as input. InternVL2-4B, however, supports multi-image input, allowing us to test with the top-1, 5, and 12 retrieved images.

As shown in Table 4, the QA module's memory consumption increases with the number of evidence pages, with 13 images (1024x768) exceeding the 80GB limit on an A100 GPU resulting in out-of-memory error. In contrast, the retrieval module maintains low memory usage, as SV-RAG processes pages independently, with costs equivalent to single-page reasoning. Although multi-evidence QA

SV-RAG-Backbone	Page	Retrieval		QA	
		Time	Mem	Time	Mem
Paligemma	R1	2.3	9.2	1.0	12.4
Phi-3-vision	R1	4.1	11.6	0.9	12.9
InternVL2-4B	R1	9.2	14.2	1.4	14.6
InternVL2-4B	R5	9.2	14.2	2.8	40.8
InternVL2-4B	R12	9.2	14.2	4.1	76.4

Table 4: Time (s) cost and Peak GPU memory (GB) cost of SV-RAG models with different backbones.

requires more memory, SV-RAG remains efficient and compact, making it well-suited for answering questions from fewer evidence pages in resource-constrained environments. This demonstrates SV-RAG's ability to balance performance and resource usage, ensuring scalability across diverse deployment scenarios. Additional results on the retrieval efficiency are presented in Table F.1.

5.6 Ablation

In our experiment, we use the hidden states from the last transformer layer (index 31) as the feature sequence. However, LLMs consist of multiple transformer layers, each encoding different types of information. To assess the impact of layer selection, we conduct an ablation study on the hidden states used to compute the late interaction score in Eq.(1). Given the high computational cost of training the col-retrieval module across all layers, we instead evaluate top-1 accuracy on the MMLongBench-Doc dataset using hidden states from different layers of the Phi-3-vision model with pre-trained weights.

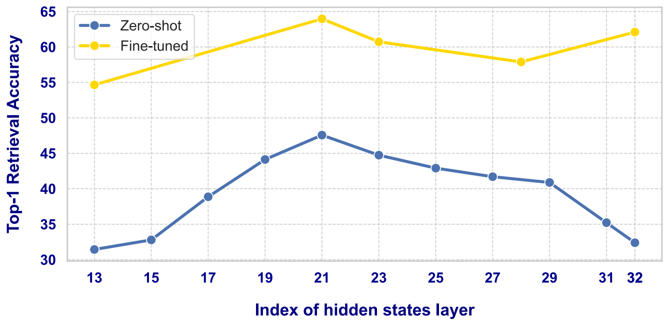


Figure 4: Top-1 retrieval accuracy on MMLongBench-Doc using different hidden states across all layers of Phi-3-vision.

Figure 4 shows that the hidden states of the 21th layer yield the highest accuracy. After fine-tuning a model with hidden states from this layer, we observed improved accuracy compared to using hidden states of the final layer. In particular, using hidden states from earlier layers can significantly reduce computational costs, enabling faster retrieval during inference.

6 Conclusions

In this paper, we propose SV-RAG, a lightweight MLLMs for visually-rich document understanding. SV-RAG has a unique design to facilitate multi-page document understanding using dual LoRA adapters. The research highlights that small open-source models are great at processing multipage documents and underscored the importance of efficient retrieval mechanisms in filtering irrelevant pages. Furthermore, we collect the VisR-bench dataset for document understanding, and empirical results on benchmarks demonstrated the effectiveness of SV-RAG. We hope these findings provide valuable insights for optimizing lightweight MLLMs, aiming to improve accuracy and efficiency in visually-rich document understanding.

7 Limitations

SV-RAG is the first MLLM that can perform visual-retrieval generation for document question answering using a single model. However, it still requires computational resources for training and inference, which may limit its practical applicability in resource-constrained environments. SV-RAG should be mobile friendly, as it only requires a single base model. This base model can be Phi-3-Silica within MS operating systems or an Apple on-device model within Apple IOS 18. A routing mechanism in Apple Intelligence can better balance computational cost and performance. However, our experiments are not performed on these real-world devices, which are necessary for pushing forward document intelligence.

8 Ethics Statement

The VisR-Bench dataset was curated with careful consideration of ethical and legal concerns. All documents are sourced from publicly available data with licenses explicitly permitting research use. To ensure data integrity and compliance, we provide links to the original sources instead of distributing the documents. Additionally, all QA pairs have been manually reviewed to exclude harmful content and personally identifiable information. The dataset does not expose sensitive user data, and experimental results are reported as aggregate statistics to prevent information leakage while ensuring reproducibility. These measures uphold ethical and legal standards while supporting responsible AI research.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

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

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024a.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Ryan Rossi, Jiuxiang Gu, and Changyou Chen. Mmr: Evaluating reading ability of large multimodal models. *arXiv preprint arXiv:2408.14594*, 2024b.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024c.

Jiaxing Chen, Yuxuan Liu, Dehu Li, Xiang An, Ziyong Feng, Yongle Zhao, and Yin Xie. Plug-and-play grounding of reasoning in multimodal large language models. *arXiv preprint arXiv:2403.19322*, 2024d.

Lin Chen et al. Sharegpt4v: Improving large multi-modal models with better captions, 2023a.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*, 2021.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

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

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023a.

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

Jiuxiang Gu, Xiangxi Shi, Jason Kuen, Lu Qi, Ruiyi Zhang, Anqi Liu, Ani Nenkova, and Tong Sun. ADOPD: A large-scale document page decomposition dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=x1ptaXpOYa.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4083–4091, 2022.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2. Minneapolis, Minnesota, 2019.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. *Computer Vision – ECCV 2022*, pp. 498–517, 2022. ISSN 1611-3349. doi: 10.1007/978-3-031-19815-1_29. URL http://dx.doi.org/10.1007/978-3-031-19815-1_29.

Marcel Lamott, Yves-Noel Weweler, Adrian Ulges, Faisal Shafait, Dirk Krechel, and Darko Obradovic. Lapdoc: Layout-aware prompting for documents. *arXiv preprint arXiv:2402.09841*, 2024.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023a.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*, pp. 19036–19052. PMLR, 2023b.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*, 2024c.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024d.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024.

Feipeng Ma, Hongwei Xue, Guangting Wang, Yizhou Zhou, Fengyun Rao, Shilin Yan, Yueyi Zhang, Siying Wu, Mike Zheng Shou, and Xiaoyan Sun. Multi-modal generative embedding model. *arXiv preprint arXiv:2405.19333*, 2024a.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. *arXiv:2406.11251*, 2024b.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*, 2024c.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*, 2024d.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2020.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

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

OpenAI. Gpt-4 technical report, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Johannes Rausch, Octavio Martinez, Fabian Bissig, Ce Zhang, and Stefan Feuerriegel. Docparser: Hierarchical document structure parsing from renderings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4328–4338, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A Rossi, and Franck Dernoncourt. Pdftriage: question answering over long, structured documents. *arXiv preprint arXiv:2309.08872*, 2023.

Peter H Sellers. The theory and computation of evolutionary distances: pattern recognition. *Journal of algorithms*, 1(4):359–373, 1980.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13878–13888, 2021.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19254–19264, 2023.

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

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pp. 778–792. Springer, 2021.

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

Rubèn Tito, Khanh Nguyen, Marlon Tobaben, Raouf Kerkouche, Mohamed Ali Souibgui, Kangsoo Jung, Lei Kang, Ernest Valveny, Antti Honkela, Mario Fritz, and Dimosthenis Karatzas. Privacy-aware document visual question answering. *arXiv preprint arXiv:2312.10108*, 2023b.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023.

Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. Layout and task aware instruction prompt for zero-shot document image question answering, 2023.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. pp. 1192–1200, 2020.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023a.

Qinghao Ye et al. mplug-owl: Modularization empowers large language models with multimodality, 2023b.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. *arXiv preprint arXiv:1911.10683*, 2019.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*, 2021.

Fengbin Zhu, Chao Wang, Fuli Feng, Zifeng Ren, Moxin Li, and Tat-Seng Chua. Doc2SoarGraph: Discrete reasoning over visually-rich table-text documents via semantic-oriented hierarchical graphs. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 5119–5131, Torino, Italia, 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.456`.

A Example of training pairs for retrieval module

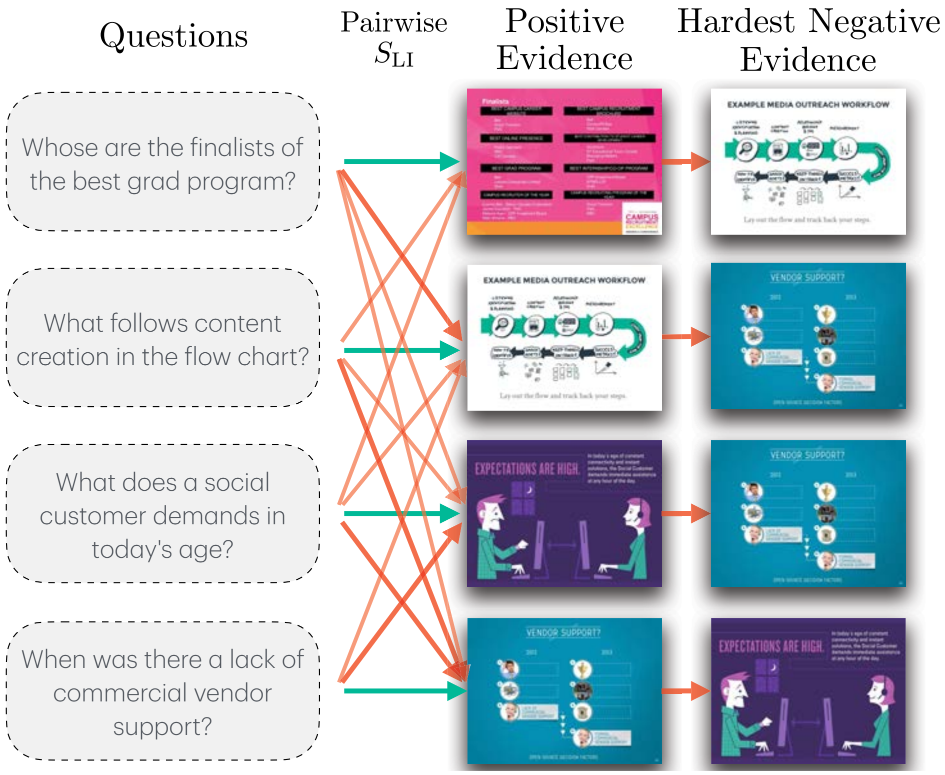

Questions
Pairwise S_{LI}
Positive Evidence
Hardest Negative Evidence
Whose are the finalists of the best grad program?
What follows content creation in the flow chart?
What does a social customer demands in today's age?
When was there a lack of commercial vendor support?
EXAMPLE MEDIA OUTREACH WORKFLOW
VENDOR SUPPORT?
EXPECTATIONS ARE HIGH

Figure A.1: Example of training pairs within a batch (batch size: 4) for contrastive training, using samples from the SlideVQA dataset.

B Evaluation Metrics

We evaluate the model's performance on evidence retrieval and question-answering using five metrics explained as follows:

Top-k Accuracy In our experiment, we focus on questions that have evidence from a single page. We use top-k accuracy to evaluate retrieval methods, which measures the percentage of times the evidence image appears within the top k most similar images.

Exact Match Following (Tanaka et al., 2023), we report exact match (EM) frequency between generated answers and the ground truth, allowing for case insensitivity and extra spaces. While effective for fine-tuned models, this metric is less suited for LLM responses, which often include full sentences. Correct answers with extra context may thus be unfairly penalized.

Generalized Accuracy We report generalized accuracy (G-Acc) from MMLongBench-Doc (Ma et al., 2024d), a GPT-dependent, rule-based evaluation protocol . Model responses are simplified using GPT-4o and scored based on answer-type-specific rules. However, G-Acc has two limitations: it introduces randomness from GPT's stochastic outputs and relies on answer-type annotations, limiting its applicability across datasets.

ANLS Average Normalized Levenshtein Similarity (ANLS) (Biten et al., 2019) measures the similarity between predicted and ground truth text using the Levenshtein distance, normalized by the longer string's length. It outputs a similarity score between 0 and 1. ANLS allows mismatches, insertions, and deletions making it useful for OCR and document understanding tasks when exact matches are not required.

PNLS The *partial normalized Levenshtein similarity* (PNLS) (Chen et al., 2024b) generalizes ANLS by not penalizing extra prefixes or suffixes while allowing mismatches, insertions, and deletions within the matched region. This makes it more suitable for evaluating LLM responses, which are often verbose to improve user experience.

The PNLS metric is formally defined as follows: String $\mathcal{T}_{1,m} = t_1 \ldots t_m$ represents the true answer and $\mathcal{S}_{1,n} = s_1 \ldots s_n$ is a model generated string. We first use using the approximate string matching algorithm (Sellers, 1980) to identify the sub-string of $\mathcal{S}$ that has the minimum edit distance to $\mathcal{T}$. Specifically, we first construct a scoring matrix $\mathbf{F}$ of size $(m+1) \times (n+1)$, where $F_{i,j}$ stores the smallest edit distance between the i-prefix $\mathcal{T}_{1,i}$ and any sub-string $\mathcal{S}_{x,j}$, $\forall x \in \{1, \ldots, j-1\}$ that ends at position j. The scoring matrix can be computed recursively

$$F_{i,j} = \begin{cases} 0 & \text{if } i = 0 \\ m & \text{if } j = 0 \\ \min \begin{pmatrix} F_{i-1,j-1} + c(t_i, s_j) \\ F_{i-1,j} + 1 \\ F_{i,j-1} + 1 \end{pmatrix} & \text{otherwise,} \end{cases}$$

where c is the substitution cost that takes a value of 0 if $t_i = s_j$ and 1 otherwise. Once $\mathbf{F}$ is computed, the minimum value in the last row is the optimal edit distance and the end index of the matched sub-string $j' = \arg\min_j(F_{m+1,j})$. The start index i' can be found by tracing back the the computation of Eq.(B) using $\arg\min$ operation. Finally, the PNLS is computed as: $m/(m + j' - i' + 1)$. In our experiments we use binary cost function: $c(t_i, s_j) = 0$ if $t_i = s_j$ else $c(t_i, s_j) = 1$

C Example of Inference Failure Scenario

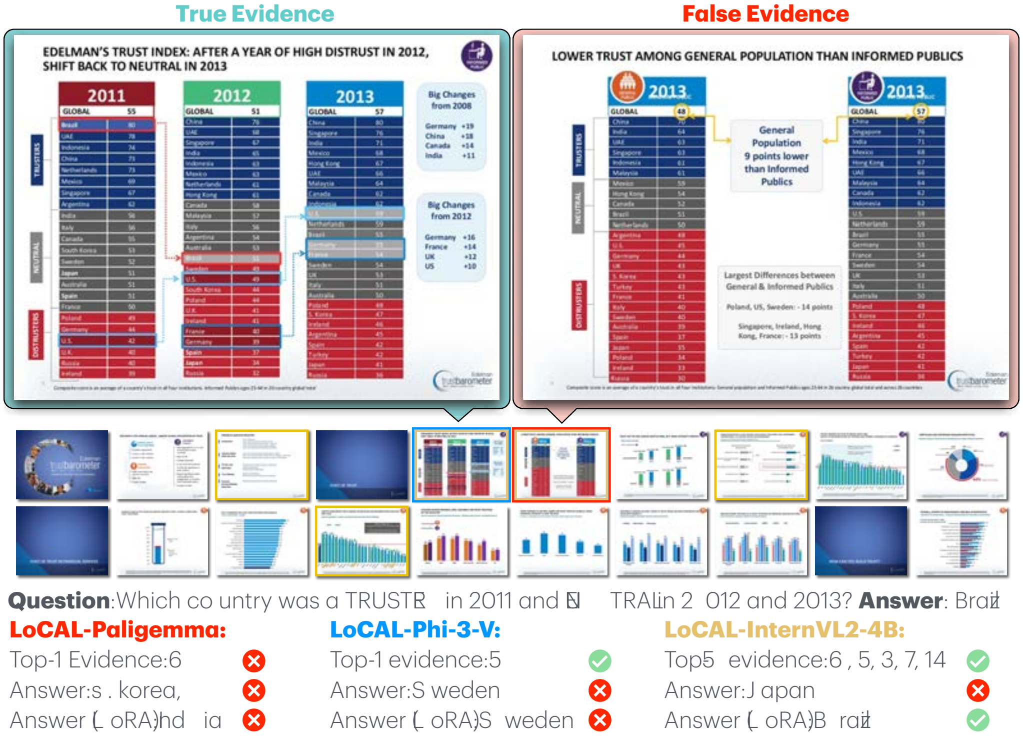
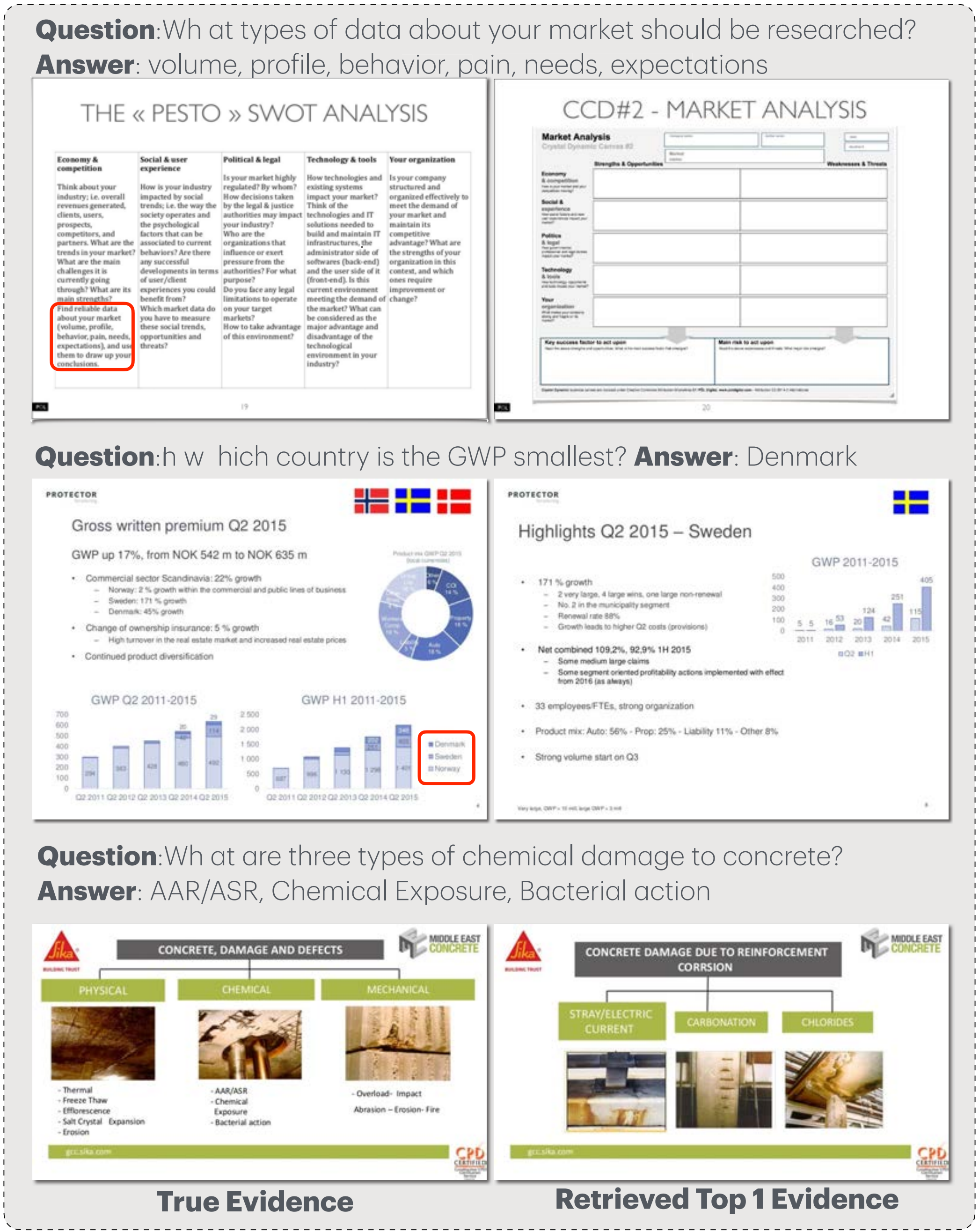

True Evidence
False Evidence
EDELMAN'S TRUST INDEX: AFTER A YEAR OF HIGH DISTRUST IN 2012, SHIFT BACK TO NEUTRAL IN 2013
2011
2012
2013
Big Changes from 2008
Big Changes from 2012
TRUSTERS
NEUTRAL
DISTRUSTERS
LOWER TRUST AMONG GENERAL POPULATION THAN INFORMED PUBLICS
General Population 9 points lower than Informed Publics
Largest Differences between General & Informed Publics
Question: Which co untry was a TRUSTEE in 2011 and NE TRAL in 2 012 and 2013? Answer: Brazil
LoCAL-Paligemma:
Top-1 Evidence: 6 ✗
Answer: s . korea, ✗
Answer (LoRA): India ✗
LoCAL-Phi-3-V:
Top-1 evidence: 5 ✓
Answer: S weden ✗
Answer (LoRA): Sweden ✗
LoCAL-InternVL2-4B:
Top5 evidence: 6 , 5, 3, 7, 14 ✓
Answer: J apan ✗
Answer (LoRA): Brazil ✓

Figure C.1: Inference example of a challenging case in the SlideVQA dataset. SV-RAG-Paligemma retrieved the wrong evidence page due to limitations in its retrieval module, leading to an incorrect answer. SV-RAG-Phi-3-V retrieved the correct page but provided a wrong answer due to limitations in its QA module. Meanwhile, SV-RAG-InternVL2-4B also assigned the highest relevance score to an incorrect page. However, since it processes multiple pages (top 5), the correct evidence page was included in the input, allowing its fine-tuned QA module to deliver the correct answer.

C.1 Additional Examples of Retrieval Failures



Question:Wh at types of data about your market should be researched?
Answer: volume, profile, behavior, pain, needs, expectations

THE « PESTO » SWOT ANALYSIS

Economy & competition	Social & user experience	Political & legal	Technology & tools	Your organization
Think about your industry; i.e. overall revenues generated, clients, users, prospects, competitors, and partners. What are the trends in your market? What are the main challenges it is currently going through? What are its main strengths? Find reliable data about your market (volume, profile, behavior, pain, needs, expectations), and use them to draw up your conclusions.	How is your industry impacted by social trends; i.e. the way the society operates and the psychological factors that can be associated to current behaviors? Are there any successful developments in terms of user/client experiences you could benefit from? Which market data do you have to measure these social trends, opportunities and threats?	Is your market highly regulated? By whom? How decisions taken by the legal & justice authorities may impact your industry? Who are the organizations that influence or exert pressure from the authorities? For what purpose? Do you face any legal limitations to operate on your target markets? How to take advantage of this environment?	How technologies and existing systems impact your market? Think of the technologies and IT solutions needed to build and maintain IT infrastructures, the administrator side of softwares (back-end) and the user side of it (front-end). Is this current environment meeting the demand of the market? What can be considered as the major advantage and disadvantage of the technological environment in your industry?	Is your company structured and organized effectively to meet the demand of your market and maintain its competitive advantage? What are the strengths of your organization in this context, and which ones require improvement or change?

19

CCD#2 - MARKET ANALYSIS

Market Analysis
Crystal Dynamic Canvas #2

	Strengths & Opportunities	Weaknesses & Threats
Economy & competition		
Social & experience		
Politics & legal		
Technology & tools		
Your organization		

Key success factor to act upon

Main risk to act upon

20

Question:h w hich country is the GWP smallest? **Answer**: Denmark

PROTECTOR

Gross written premium Q2 2015

GWP up 17%, from NOK 542 m to NOK 635 m

- Commercial sector Scandinavia: 22% growth
  - Norway: 2 % growth within the commercial and public lines of business
  - Sweden: 171 % growth
  - Denmark: 45% growth
- Change of ownership insurance: 5 % growth
  - High turnover in the real estate market and increased real estate prices
- Continued product diversification

PROTECTOR

Highlights Q2 2015 – Sweden

- 171 % growth
  - 2 very large, 4 large wins, one large non-renewal
  - No. 2 in the municipality segment
  - Renewal rate 88%
  - Growth leads to higher Q2 costs (provisions)
- Net combined 109,2%, 92,9% 1H 2015
  - Some medium large claims
  - Some segment oriented profitability actions implemented with effect from 2016 (as always)
- 33 employees/FTEs, strong organization
- Product mix: Auto: 56% - Prop: 25% - Liability 11% - Other 8%
- Strong volume start on Q3

Question:Wh at are three types of chemical damage to concrete?
Answer: AAR/ASR, Chemical Exposure, Bacterial action

CONCRETE, DAMAGE AND DEFECTS

PHYSICAL
- Thermal
- Freeze Thaw
- Efflorescence
- Salt Crystal Expansion
- Erosion

CHEMICAL
- AAR/ASR
- Chemical Exposure
- Bacterial action

MECHANICAL
- Overload- Impact
Abrasion – Erosion- Fire

gcc.sika.com

CONCRETE DAMAGE DUE TO REINFORCEMENT CORRSION

STRAY/ELECTRIC CURRENT

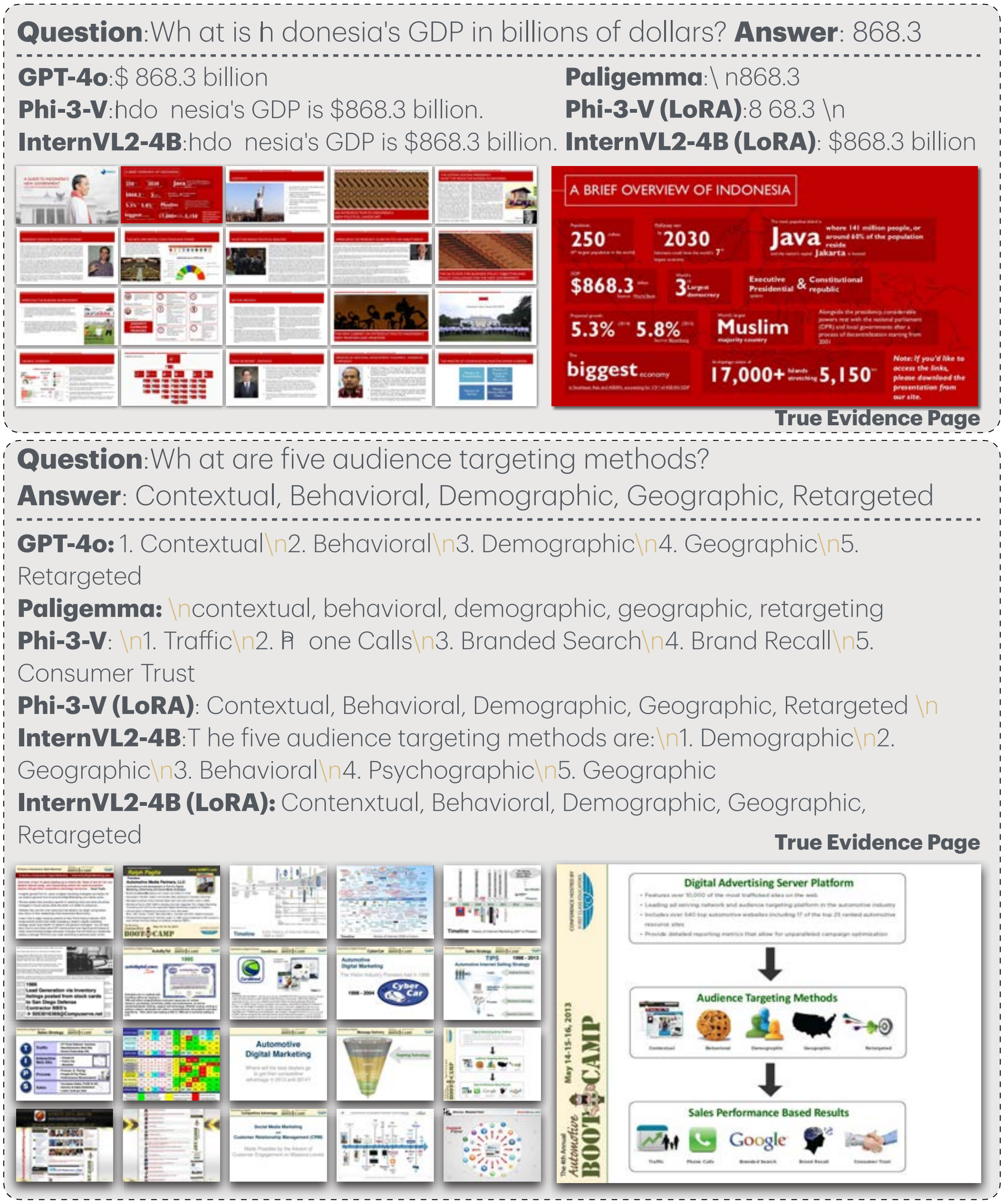
CARBONATION

CHLORIDES

gcc.sika.com

Figure C.2: Failure cases from the SlideVQA dataset, highlighting retrieval module errors. In the first two examples, some of the relevant information (highlighted in red boxes) on the true evidence pages is difficult even for human eyes to detect. In the third example, the retrieved page has a high similarity to the true evidence page, making it challenging to rank correctly. Additionally, answering the question accurately requires a deep understanding of the concept of chemical damage and related topics.

D Qualitative Results in Question-Answering



Question:Wh at is h donesia's GDP in billions of dollars? **Answer**: 868.3

GPT-4o:$ 868.3 billion

Phi-3-V:hdo nesia's GDP is $868.3 billion.

InternVL2-4B:hdo nesia's GDP is $868.3 billion.

Paligemma:\ n868.3

Phi-3-V (LoRA):8 68.3 \n

InternVL2-4B (LoRA): $868.3 billion

A BRIEF OVERVIEW OF INDONESIA
250
2030
Java
$868.3
3
Executive Presidential & Constitutional republic
5.3% 5.8%
Muslim
biggest
17,000+ 5,150
True Evidence Page

Question:Wh at are five audience targeting methods?

Answer: Contextual, Behavioral, Demographic, Geographic, Retargeted

GPT-4o: 1. Contextual\n2. Behavioral\n3. Demographic\n4. Geographic\n5. Retargeted

Paligemma: \ncontextual, behavioral, demographic, geographic, retargeting

Phi-3-V: \n1. Traffic\n2. Ph one Calls\n3. Branded Search\n4. Brand Recall\n5. Consumer Trust

Phi-3-V (LoRA): Contextual, Behavioral, Demographic, Geographic, Retargeted \n

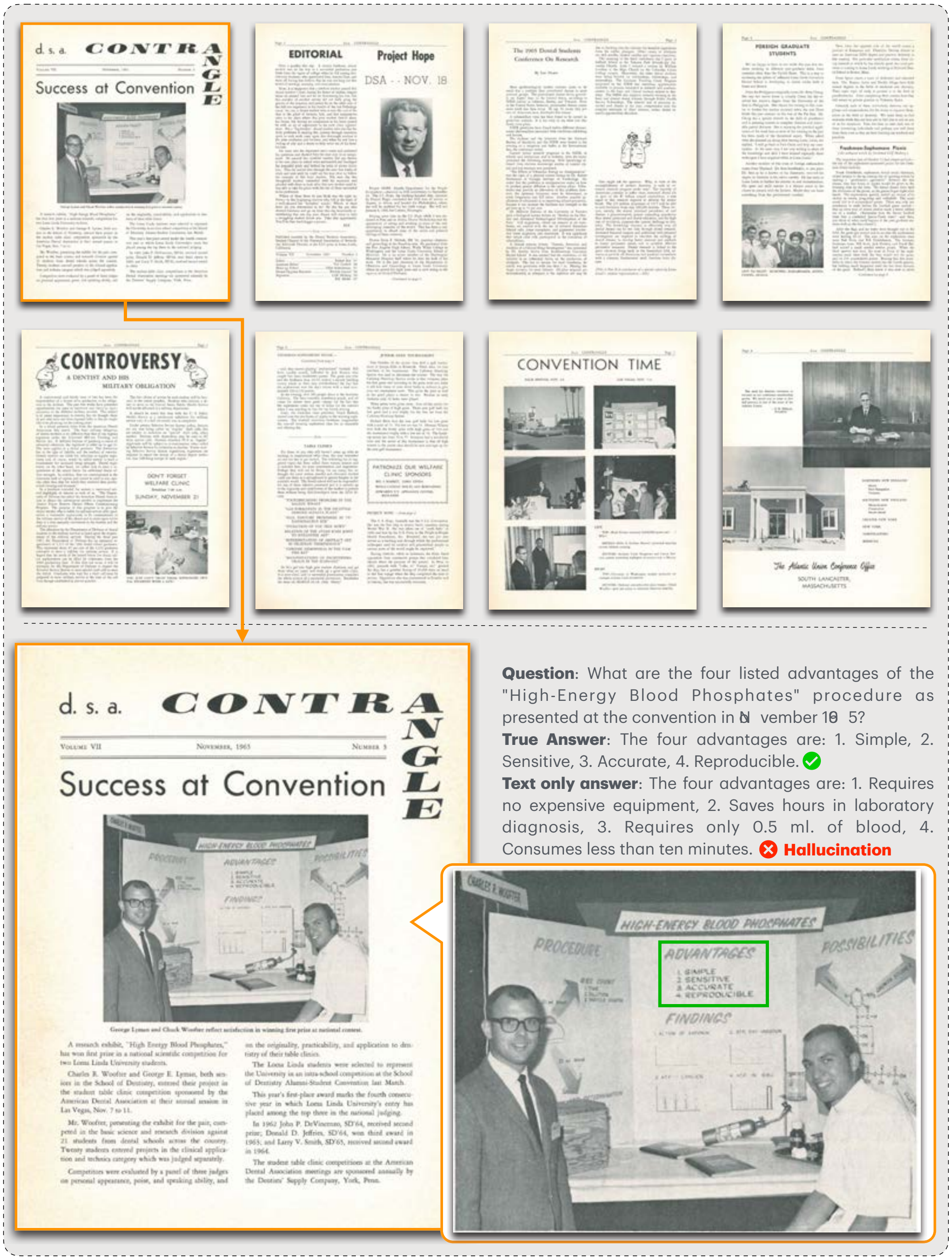
InternVL2-4B:T he five audience targeting methods are:\n1. Demographic\n2. Geographic\n3. Behavioral\n4. Psychographic\n5. Geographic

InternVL2-4B (LoRA): Contenxtual, Behavioral, Demographic, Geographic, Retargeted

True Evidence Page
Digital Advertising Server Platform
Audience Targeting Methods
Sales Performance Based Results
Google
May 14-15-16, 2013
Automotive BOOT CAMP

Figure D.1: Question answering examples on the SlideVQA dataset using different QA modules. Models without fine-tuning, such as Phi-3-V and InternVL2-4B, tend to produce verbose and error-prone responses. However, in the second example, fine-tuning with the LoRA adapter significantly improves the accuracy of Phi-3-V and InternVL2-4B.

E Examples from the VisR-Bench dataset



d. s. a. CONTRA ANGLE

Volume VII November, 1965 Number 3

Success at Convention

George Lyman and Chuck Woofter reflect satisfaction in winning first prize at national contest.

A research exhibit, "High Energy Blood Phosphates," has won first prize in a national scientific competition for two Loma Linda University students.

Charles R. Woofter and George E. Lyman, both seniors in the School of Dentistry, entered their project in the student table clinic competition sponsored by the American Dental Association at their annual session in Las Vegas, Nov. 7 to 11.

Mr. Woofter, presenting the exhibit for the pair, competed in the basic science and research division against 21 students from dental schools across the country. Twenty students entered projects in the clinical application and technics category which was judged separately.

Competitors were evaluated by a panel of three judges on personal appearance, poise, and speaking ability, and on the originality, practicability, and application to dentistry of their table clinics.

The Loma Linda students were selected to represent the University in an intra-school competition at the School of Dentistry Alumni-Student Convention last March.

This year's first-place award marks the fourth consecutive year in which Loma Linda University's entry has placed among the top three in the national judging.

In 1962 John P. DeVincenzo, SD'64, received second prize; Donald D. Jeffries, SD'64, won third award in 1963; and Larry V. Smith, SD'65, received second award in 1964.

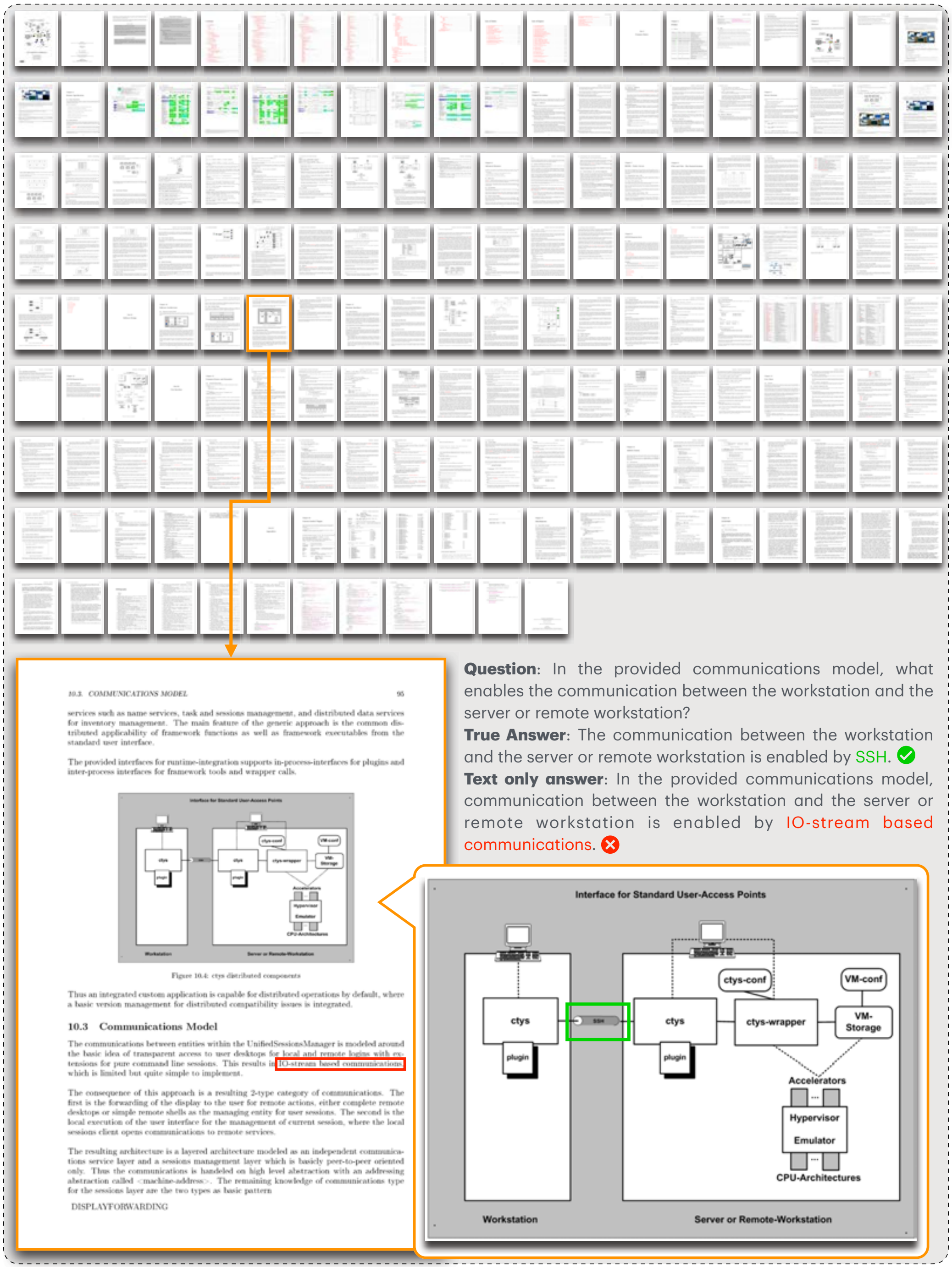
The student table clinic competitions at the American Dental Association meetings are sponsored annually by the Dentists' Supply Company, York, Penn.

Question: What are the four listed advantages of the "High-Energy Blood Phosphates" procedure as presented at the convention in No vember 196 5?

True Answer: The four advantages are: 1. Simple, 2. Sensitive, 3. Accurate, 4. Reproducible. ✓

Text only answer: The four advantages are: 1. Requires no expensive equipment, 2. Saves hours in laboratory diagnosis, 3. Requires only 0.5 ml. of blood, 4. Consumes less than ten minutes. ✗ **Hallucination**

Figure E.1: Example question-and-answer pair from the VisR-Bench dataset, highlighting the reliance on both image and surrounding text for accurate responses.



10.3. COMMUNICATIONS MODEL 95

services such as name services, task and sessions management, and distributed data services for inventory management. The main feature of the generic approach is the common distributed applicability of framework functions as well as framework executables from the standard user interface.

The provided interfaces for runtime-integration supports in-process-interfaces for plugins and inter-process interfaces for framework tools and wrapper calls.

Figure 10.4: ctys distributed components

Thus an integrated custom application is capable for distributed operations by default, where a basic version management for distributed compatibility issues is integrated.

10.3 Communications Model

The communications between entities within the UnifiedSessionsManager is modeled around the basic idea of transparent access to user desktops for local and remote logins with extensions for pure command line sessions. This results in IO-stream based communications, which is limited but quite simple to implement.

The consequence of this approach is a resulting 2-type category of communications. The first is the forwarding of the display to the user for remote actions, either complete remote desktops or simple remote shells as the managing entity for user sessions. The second is the local execution of the user interface for the management of current session, where the local sessions client opens communications to remote services.

The resulting architecture is a layered architecture modeled as an independent communications service layer and a sessions management layer which is basicly peer-to-peer oriented only. Thus the communications is handeled on high level abstraction with an addressing abstraction called <machine-address>. The remaining knowledge of communications type for the sessions layer are the two types as basic pattern

DISPLAYFORWARDING

Question: In the provided communications model, what enables the communication between the workstation and the server or remote workstation?

True Answer: The communication between the workstation and the server or remote workstation is enabled by SSH. ✓

Text only answer: In the provided communications model, communication between the workstation and the server or remote workstation is enabled by IO-stream based communications. ✗

Figure E.2: Example question-and-answer pair from the VisR-Bench dataset, highlighting the reliance on both image and surrounding text for accurate responses.

F Comparison of Retrieval Method Efficiency

Text Extraction		Text Encoding	
PaddleOCR	0.275	BM25	0.0001
		BGE-m3	0.131
PDF Parser	0.762	BGE-large	0.137
		NV-embed-v2	0.117

Multimodal Encoding	
CLIP	0.022
SigLip	0.109
Col-Paligemma	0.140
Col-Phi-3-V	0.230
Col-InternVL2	0.581

Table F.1: Per-page time cost of retrieval methods: The left table presents time cost (seconds) of text-based methods that rely on text extraction techniques, such as OCR models, followed by text encoders to compute page embeddings. The right table presents time cost (seconds) of multi-modal methods that encode the entire page as an image.

G Additional Experiment results

We compare our method with text-only baselines using a document parser[3] to highlight the advantages of MLLMs in multi-modal understanding. QA results are reported for the VisR-Bench and MMLongBench-Doc datasets, where PDF files are available.

Table G.1 presents QA results on VisR-Bench and MMLongBench-Doc datasets. To evaluate answer quality for VisR-Bench, where true answers are long and detailed, we introduce the Mean GPT Score (MGS), as string-matching methods often penalize variations in wording for lengthy answers. Instead, we prompt GPT-4o to compare a model's answer with the ground truth and assign a binary score based on detail alignment.

QA Module	Retrieval Module	Evidence	VisR-B	MMLong
			MGS	G-Acc
Text only QA methods				
Phi-3 + parser	Col-Phi-3-V	R5	14.1	29.2
GPT-4o + parser	Col-Phi-3-V	R5	24.9	43.2
GPT-4o + parser	-	A	27.6	42.4
MLLM QA models				
PaliGemma	Col-PaliGemma	R1	12.2	23.9
Phi-3-V	Col-Phi-3-V	R1	24.2	30.7
SV-RAG-InternVL2	Col-InternVL2	R5	25.2	33.2
GPT-4o	Col-Phi-3-V	R5	47.2	55.1
GPT-4o	-	A	43.2	54.5

Table G.1: parser results

Our results indicate that using image evidence consistently outperforms text-only evidence. On VisR-Bench, text-only baselines showed a significant performance drop, emphasizing the dependency of questions on both image and text. However, the MGS of text-only baselines is not zero, likely because the model leverages text from a broader context rather than relying solely on the surrounding text, enabling it to extract relevant information even in the absence of visual input.

Additionally, reducing input pages with the retrieval module improved GPT-4o's performance with image evidence, aligning with the findings in Table 2. In contrast, retrieval did not enhance GPT-4o's performance on VisR-Bench in the text-only setting, likely because the evidence pages lacked sufficient information to fully address the questions. Including additional context in such cases might yield better results.

[3] Adobe Extract API: https://developer.adobe.com/document-services/apis/pdf-extract/