PDF-WuKong : A Large Multimodal Model for Efficient Long PDF Reading with End-to-End Sparse Sampling

Xudong Xie[1*]   Liang Yin[1*]   Hao Yan[1*]   Yang Liu[1*]   Jing Ding[1]   Minghui Liao[2]

Yuliang Liu[1]   Wei Chen[1(✉)]   Xiang Bai[1(✉)]

[1]Huazhong University of Science and Technology   [2]Huawei Inc.

{lemuria_chen,xbai}@hust.edu.cn

Abstract

*Document understanding is a challenging task to process and comprehend large amounts of textual and visual information. Recent advances in Large Language Models (LLMs) have significantly improved the performance of this task. However, existing methods typically focus on either plain text or a limited number of document images, struggling to handle long PDF documents with interleaved text and images, especially in academic papers. In this paper, we introduce **PDF-WuKong**, a multimodal large language model (MLLM) which is designed to enhance multimodal question-answering (QA) for long PDF documents. PDF-WuKong incorporates a sparse sampler that operates on both text and image representations, significantly improving the efficiency and capability of the MLLM. The sparse sampler is integrated with the MLLM's image encoder and selects the paragraphs or diagrams most pertinent to user queries for processing by the language model. To effectively train and evaluate our model, we construct **PaperPDF**, a dataset consisting of a broad collection of academic papers sourced from arXiv, multiple strategies are proposed to generate automatically $1M$ QA pairs along with their corresponding evidence sources. Experimental results demonstrate the superiority and high efficiency of our approach over other models on the task of long multimodal PDF understanding, surpassing proprietary products by an average of 8.6% on F1. Our code and dataset will be released at https://github.com/yh-hust/PDF-Wukong.*

1. Introduction

The advent of Large Language Models (LLMs) has significantly advanced the field of PDF document understanding [1, 2], where these models have demonstrated impressive capabilities in processing and generating human-like

---

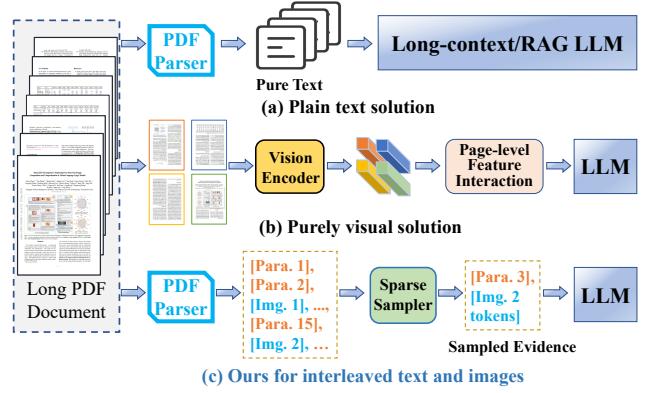* Equal contribution; ✉ Corresponding author

Figure 1. Method comparison for long multi-page PDF document understanding. (a) Plain text solution: long-context/RAG LLMs for parsed pure text content. (b) Purely visual solution: VDU models for page-level encoding and feature interaction. (c) Our method is based on end-to-end sparse sampling for long PDFs with interleaved text and images.

text. However, they still face many challenges when it comes to lengthy PDF documents with interlaced text and images, such as academic papers.

To handle lengthy documents, current large models mainly fall their input into two separate modalities. One is the plain text modality. As shown in Fig. 1 (a), they focus only on the parsed pure text content and lack multimodal understanding of visual elements such as charts and figures within the documents. They usually rely on long-context LLMs [3–5], whose efficiency and accuracy decrease as the length of the document increases. Retrieval Augmented Generation (RAG) [6–9] is also the main auxiliary means of these LLMs, but it requires the introduction of additional models and lacks the exploration of multimodal RAG.

Another approach is to input documents in a purely visual modality, treating each page as an image. However, most current Visual Document Understanding (VDU) models models [10–12] only focus on visual QA on single-page

documents. They usually rely on larger input resolution, resulting in more tokens, and thus have difficulty processing multi-page long documents. Recently, some multi-page VDU models can understand 8-page [13] or 20-page [14] documents. They usually encode each page separately and then perform page-level visual feature interactions such as concatenation [15, 16], as shown in Fig. 1 (b). However, more pages generate more visual tokens, which brings greater resource consumption to LLM. This makes them inefficient and unable to process longer documents.

Considering the limitations of existing methods in multimodal understanding of long PDF documents, we propose a new MLLM architecture with end-to-end sparse sampling, named **PDF-WuKong**. Since most user queries are only related to a small part of the content in a long document, the sparse sampling can significantly remove redundant noise information. It encodes text paragraphs and diagrams in the parsed PDF document, utilizing both text and image representations to identify and extract the most relevant evidence in response to a user's query. The sampled sparse evidence significantly reduces the number of input tokens of LLM and this process is independent of the length of the input documents. Moreover, the sparse sampler and LLM can be integrated in an end-to-end manner for training and inference, optimizing the performance of multimodal representation and question answering while improving time efficiency. It is worth noting that this sparse sampler is a plug-and-play design that can be applied to any MLLMs. Another important characteristic is that it can naturally provide strong interpretability for the question answering.

In order to simultaneously represent and understand the multimodal content of documents and further improve the ability to process long PDF documents, we construct a training dataset specifically for academic paper PDFs. The academic paper PDF is a kind of typical document that contains rich interleaved text and images, which can intuitively reflect the challenges of our task and the advantages of our model. The dataset contains complete PDF documents, professional academic questions, answers, and evidence sources for the answers, based on multiple construction strategies. We also provide a corresponding benchmark named **PaperPDF**.

The experimental results substantiate the effectiveness and efficiency of our approach on the task of long multimodal PDF understanding. PDF-WuKong significantly outperforms potential open-source models that may be applied to this task. It also surpasses some proprietary products for document understanding on our proposed PaperPDF benchmark. As the number of document pages increases, its accuracy and efficiency will not decrease significantly. It also achieves competitive performance on several document-oriented VQA datasets, especially multi-page benchmarks like DUDE [17]. Besides, for the recent benchmark MM-

NIAH [18] of long multimodal documents, PDF-WuKong also outperforms other models with fewer parameters. Our model achieves the best performance on multimodal content with a context length of 64K.

The **main contributions** of this paper are as follows:

- We introduce a large multimodal model for long PDF understanding with end-to-end sparse sampling, achieving accurate and efficient PDF question answering.
- We propose a PDF multimodal question answering dataset (**PaperPDF**) with $1M$ QA pairs for training and $6k$ QA pairs for evaluation.
- Our model significantly surpasses existing open-source models and proprietary products (by an average of 8.6% on F1) on long multimodal PDF understanding.

2. Related Works

2.1. Document Understanding Datasets

Earlier document understanding datasets only focused on the NLP tasks such as summarization [27], and QA [28] of plain text. Meanwhile, there were several visual document datasets mainly aimed at text perception tasks such as Document Layout Analysis (DLA) [29–31] and Key Information Extraction (KIE) [32–34]. Recently, more datasets have been proposed for multimodal document QA across various scenarios. For example, DocVQA [35] and OCRVQA [36] provide single-page QA data on books and business documents. ChartQA [37] and ChartX [38] focus on the visual reasoning for chart documents. ArXivQA [39] extracts scientific figure-caption data from the arxiv papers to enhance the academic ability of MLLMs. InfoVQA [40] contains many infographic documents which are a combination of textual, graphical and visual elements. However, these visual document datasets only define the single-page VQA task, and current MLLMs [11, 41] have achieved remarkable performance on it.

There are also some multi-page QA datasets [14, 15, 17] that require the model to understand the content relationship via multi-hop reasoning and capture the crucial information from multi-page documents. MP-DocVQA [14] extends DocVQA [35] by adding the context pages. DUDE [17] constructs multi-page QA data from multi-industry and multi-domain documents. Recently, DocGenome [42] was constructed as a scientific document benchmark. MM-NIAH [18] is a benchmark evaluating the capability of MLLMs to comprehend long multimodal documents, which requires the model to answer according to the key information scattered throughout the document. However, the answers in these datasets lack evidence and cannot provide reliable interpretability, especially for questions that require referring to multiple pieces of evidence in long documents.

Table 1. Comparison of various models for processing multi-page long documents.

Input modality	Type	Number of tokens	Models
Plain text	Long-context	Linear increase	LongLoRA [4], LongLLaMA [5], YaRN [3]
	RAG	w/o Linear increase	Graph RAG [9], DISC-LawLLM [6], RAPTOR [19]
Purely visual	Single-page	w/o Linear increase	UniDoc [20], DocOwl [21], Vary [12], UReader [22], TextMonkey [10], LLaVA-NeXT [23], DocPedia [24], XC2-4KHD [25] and InternVL-V1.5 [11]
	Multi-page	Linear increase	Hi-VT5 [14], GRAM [16], Fox [13], DocOwl2 [26]
Text and images	Unlimited-page	w/o Linear increase	PDF-WuKong (**Ours**)

2.2. Document Understanding Methods

Existing document understanding methods typically focus on either plain text or a limited number of document images. Methods that rely on pure text modality aim to process documents by first converting them into plain text through document parsing or Optical Character Recognition (OCR). They then employ efficient long-context mechanisms to handle long texts, such as sparse attention [4], memory networks [5], or position interpolation [3]. Besides, the methods based on retrieval-augmented generation (RAG) [6, 9, 19] also show impressive capabilities for long texts. While these approaches can integrate visual elements by including image captions or transforming images into natural language descriptions, they struggle with fine-grained understanding of visual information. This restricts their effectiveness in tasks requiring detailed interpretation of textual and visual components within documents.

Another solution is visual document understanding with purely visual input, treating each document page as an image. Many MLLMs such as UniDoc [20], mPLUG-DocOwl [21] and Vary [12] can perform this task in an OCR-free manner. Vary [12] employs an extra SAM-style [43] vision vocabulary specific to document and chart data, enabling direct encoding of entire pages with high compression ratios. Other researchers have advanced the understanding of high-resolution document pages by dividing input images into smaller patches, such as UReader [22], TextMonkey [10], and LLaVA-NeXT [23]. DocPedia [24] processes high-resolution images in the frequency domain via the DCT transformation. InternLM-XComposer2-4KHD [25] and InternVL-V1.5 [11] introduce a dynamic resolution mechanism with automatic patch configuration to capture more details. The reliance on high resolution results in a higher number of tokens and cannot be extended to multi-page documents.

Besides, there are several models specifically designed for multi-page documents. Hi-VT5 [14] and GRAM [16] are two professional models for multi-page QA. They combine image tokens of multiple pages through hierarchical transformer architecture [14] or global-local reasoning [16], being able to handle up to 20 pages. Fox [13] unifies all image tokens of up to 8 pages into a sequence to achieve multi-page QA. mPLUG-DocOwl2 [26] compress each high-resolution document image into 324 tokens, guided by low-resolution global visual features. These models encode each page separately and then perform page-level visual feature interactions. However, more pages generate more visual tokens, which brings greater resource consumption to LLM. This makes them inefficient and unable to process longer documents. Given this, we propose first to parse the lengthy document into interleaved text and image content and then perform sparse sampling in an end-to-end manner. Tab. 1 summarizes the attributes of each type of method.

3. Methodology

3.1. Overview

In order to achieve multimodal understanding of long PDF documents and alleviate the drawbacks of existing models that treat PDF documents as a single modality of plain text or images, we propose PDF-WuKong. The core motivation is that users' queries are often only related to a small number of text blocks or diagrams in a long document. Therefore, to improve the accuracy and efficiency of MLLM, we design a sparse sampler integrated with a multimodal large language model in an end-to-end manner.

Specifically, our pipeline consists of three parts: a document parser, a sparse sampler and a large language model, as shown in Fig. 2. The document parsing stage first converts the input PDF document into machine-readable content of interleaved text and images, Then, the sparse sampler encodes the text blocks and images separately and caches their embeddings. When a user inputs a query, the most relevant content can be sampled using a simple similarity measure. Finally, the query and the sampled tokens are input into the LLM to generate the answer. Algorithm 1 shows the detailed steps of this method.
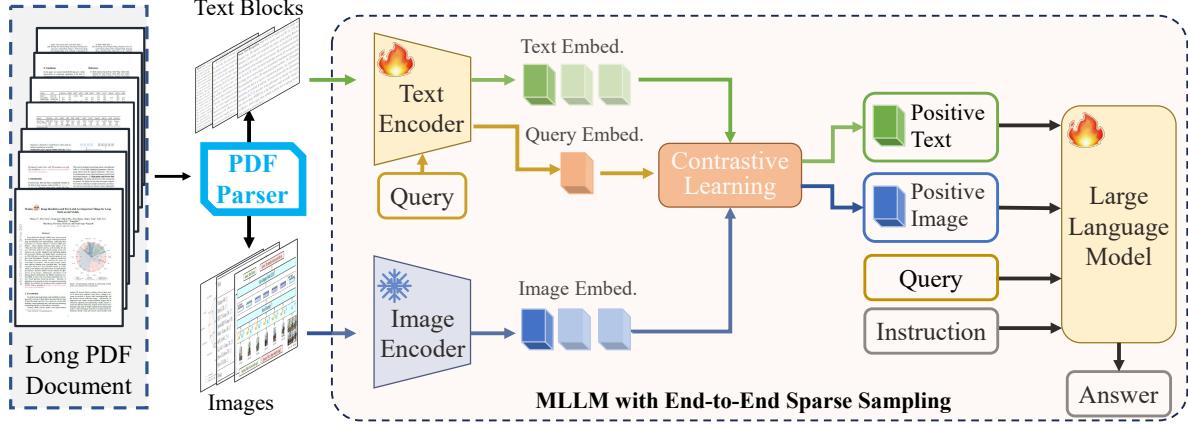
Figure 2. The overall structure of PDF-WuKong consists of a document parser, a sparse sampler and a large language model.

Algorithm 1 Inference pipeline for PDF-WuKong

1: **Input:** PDF document D, user query q
2: **Output:** Generated answer a
3: **Initialize:** Text encoder En_T, image encoder En_I, large language model LLM
4: **Stage 1: Document Parsing**
5: Parse the input document D into text blocks and images:

$$\{T_1, T_2, \ldots, T_n\}, \{I_1, I_2, \ldots, I_m\} \leftarrow \text{Parser}(D)$$

6: **Stage 2: Sparse Sampling**
7: Encode all text blocks and images and **cache** all candidate vector embeddings:

$$E_T = \{e_{T_1}, e_{T_2}, \ldots, e_{T_n}\} \leftarrow \text{En\_T}(\{T_1, T_2, \ldots, T_n\}),$$

$$E_I = \{e_{I_1}, e_{I_2}, \ldots, e_{I_m}\} \leftarrow \text{En\_I}(\{I_1, I_2, \ldots, I_m\})$$

8: Encode the user query q:

$$e_q \leftarrow \text{En\_T}(q)$$

9: Calculate the similarity between query embedding e_q and cached text/image embeddings $\{E_T, E_I\}$:

$$S_T = \{\text{Sim}(e_q, e_{T_i}) \mid i = 1, 2, \ldots, n\},$$

$$S_I = \{\text{Sim}(e_q, e_{I_j}) \mid j = 1, 2, \ldots, m\}$$

10: Select the top-k relevant text blocks and images:

$$(T, I)_{top} \leftarrow \text{TopK}(S_T, S_I, k)$$

11: **Stage 3: Answer Generation**
12: Input the query q and the selected tokens into the LLM:

$$a \leftarrow \text{LLM}(q, (T, E_I)_{top})$$

13: **Return** the generated answer a.

Algorithm 2 Training pipeline for PDF-WuKong

1: **Input:** PDF document D, user query q, ground truth answer gt
2: **Output:** Final loss function $\mathcal{L}_{\text{total}}$
3: **Initialize:** Text encoder En_T, image encoder En_I, large language model LLM
4: **Stage 1: Data Preparing**
5: Text blocks and images:

$$\{T_1, T_2, \ldots, T_n\}, \{I_1, I_2, \ldots, I_m\} \leftarrow \text{Parser}(D)$$

6: **Stage 2: Multimodal encoding**
7: Encode the user query, positive and negative samples:

$$e_q \leftarrow \text{En\_T}(q)$$

$$E_T = \{e_{T_P}, e_{T_N}\} \leftarrow \text{En\_T}(\{T_P, T_N\}),$$

$$E_I = \{e_{I_P}, e_{I_N}\} \leftarrow \text{En\_I}(\{I_P, I_N\}),$$

8: Calculate the contrastive learning loss:

$$\mathcal{L}_{\text{rep}}(e_q, \{e_{T_P}, e_{I_P}\}, \{e_{T_N}, e_{I_N}\})$$

9: **Stage 3: Output prediction of MLLM**
10: Input the query, the positive text, and the positive image tokens from the **shared image encoder** En_I:

$$a \leftarrow \text{LLM}(q, T_P, e_{I_P})$$

11: Calculate the cross-entropy loss:

$$\mathcal{L}_{\text{QA}}(a, gt)$$

12: **Stage 4: Optimize model in an end-to-end manner**
13: Update model parameters according to the joint loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{QA}}$$

14: **Return** the final loss $\mathcal{L}_{\text{total}}$.

3.2. Document Parsing

Given a PDF document D, the goal of document parsing is to convert it into some machine-readable text blocks $\{T_1, T_2, \ldots, T_n\}$ and diagrams $\{I_1, I_2, \ldots, I_m\}$ according to the reading order and layout structure. By default, text blocks are organized into paragraphs, and all figures and tables are saved as images. These text and images are finally reorganized into an XML file in reading order. This process can be completed using existing open-source PDF parsing tools. During inference, we directly input the parsed structured full data into the subsequent stage of PDF-WuKong.

3.3. Sparse Sampling

For a lengthy multi-page document, if it is directly input into the LLM, there will be two problems. The first is the problem of computing efficiency. The consumption of computing resources will increase dramatically. The second is the problem of inaccurate attention. Key information related to the user query is easily submerged by a large amount of irrelevant content. It is difficult for the model to accurately locate and extract important information in a huge token sequence. Therefore, sparse sampling is essential for efficiently handling lengthy multi-page documents by identifying and extracting the most relevant text chunks or diagrams based on their similarity to the user query.

For the parsed n text chunks $\{T_1, T_2, \ldots, T_n\}$, m images $\{I_1, I_2, \ldots, I_m\}$, and the input user query q, we first extract the positive samples and the negative samples for the query. Our PaperPDF dataset has provided corresponding positive single-evidence or multi-evidence samples for each query-answer pair (detailed in Sec. 4). We randomly select two text blocks and two images from the remaining text blocks and images as negative samples. Then, we use a text encoder En_T to obtain the text embeddings e_{T_P}, e_{T_N} and the query embedding e_q. An image encoder En_I is utilized to output the image features e_{I_P}, e_{I_N}, which is shared with MLLM.

Given the embeddings of the user query e_q, the positive samples $E_P = \{e_{T_P}, e_{I_P}\}$ and negative samples $E_N = \{e_{T_N}, e_{I_N}\}$, we employ a contrastive learning approach to align the text and image features with the query. The goal is to enable the model to capture the document content that is most relevant to the query. The contrastive learning loss is:

$$\mathcal{L}_{\text{rep}} = -\frac{1}{P} \sum_{e_i \in E_P} \log \frac{e^{\text{sim}(e_q, e_i)/\tau}}{e^{\text{sim}(e_q, e_i)/\tau} + \sum_{e_j \in E_N} e^{\text{sim}(e_q, e_j)/\tau}},$$
(1)

where $\text{sim}(e_q, e_i)$ and $\text{sim}(e_q, e_j)$ represent the similarity between the query and the positive/negative samples, respectively. τ is the temperature parameter that controls the scale of the similarity scores. P is the number of positive samples. By maximizing this probability, the model encour-

ages the representations of the query and positive samples to be closer while pushing the representations of the query and negative samples apart. We align the feature space of the text encoder to the pre-trained vision encoder with frozen parameters. It is worth noting that this sparse sampler is a plug-and-play design that can be applied to any MLLMs.

During the inference, we pre-encode all text blocks and images and cache all candidate vector embeddings. When the user inputs a query, we calculate the similarity between query embedding and cached text/image embeddings. Then the model automatically selects the top-k relevant text blocks and images as evidence to respond to this query. Therefore, this process samples out sparse document content, greatly reducing the computational burden of the subsequent LLM and alleviating the problem of attention shift when facing ultra-long sequences. Moreover, the multimodal embedding cache further optimizes inference time.

3.4. Answer Generation

At this stage, the large language model only receives the document content that is most relevant to the query and discards a lot of redundant information, so it can generate more accurate answers with higher efficiency. Specifically, we input the sampled top-k evidence, the user query, and the task instruction into the LLM, and let it generate an answer based on the provided query and evidence. The inference process is shown in Algorithm 1.

Considering that MLLM needs to encode images first for multimodal understanding, we directly input the image tokens obtained from the sparse sampler into the LLM, to save one image encoding process. Thus, the sparse sampler shares the same vision encoder with the MLLM. They can be integrated and trained in an end-to-end manner.

During the training, we input the positive text T_P and the positive image tokens e_{I_P} into the LLM. Besides, the query and instruction are also input into the LLM. Then, we calculate the cross-entropy loss $\mathcal{L}_{\text{QA}}$ between the output answer a and the ground truth. Finally, the total optimization objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{QA}}.$$
(2)

PDF-WuKong is optimized end-to-end by these two loss functions for effective multimodal alignment and question-answering. The training pipeline of PDF-WuKong is shown in Algorithm 2.

4. PaperPDF Construction

4.1. Overview

The reasons for constructing our dataset include the following two points. In most long document Q&A scenarios, the basis for an answer to a given question is typically composed of a single element or a combination of multiple elements, while the remaining information may act as
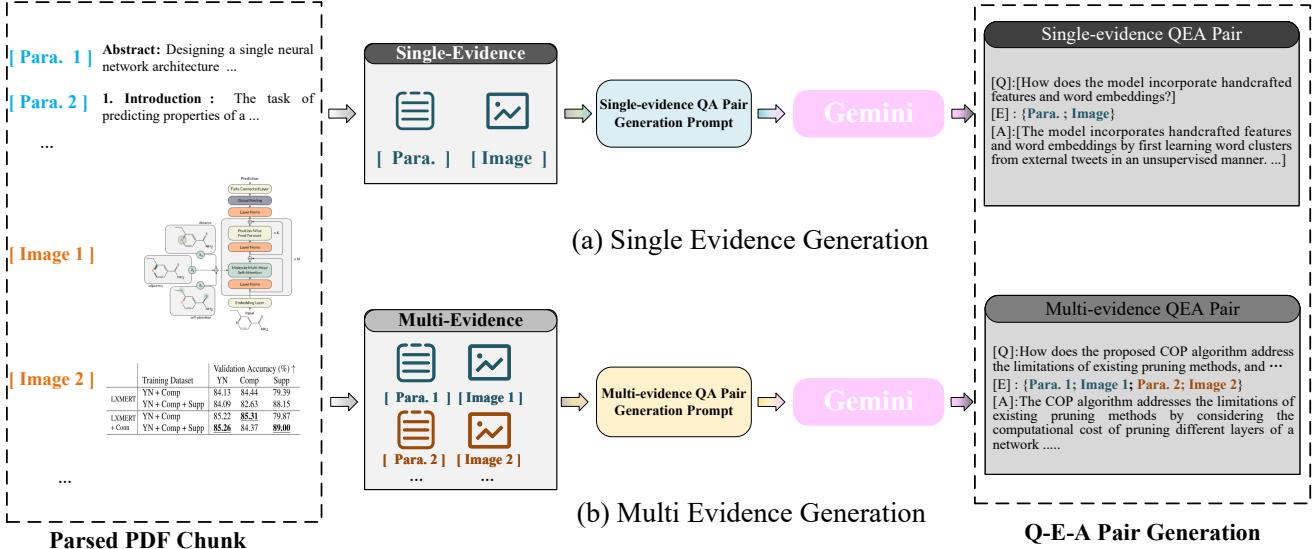
Figure 3. The construction process of PaperPDF based on single evidence and multiple evidence.

noise, interfering with the reasoning of MLLMs. Additionally, existing document question-answering datasets are either limited to single-page formats or provide only page-level ground truth, which hinders the training of the sparse sampler in PDF-WuKong. Consequently, we propose a reliable method for generating high-quality question-answer pairs of long document and develop the dataset PaperPDF for training and evaluation based on this method.

	Category	Training	Test
Single	Text-only	249k	2939
	Image-only	21K	212
Multi	Image-text	250k	2566
	Section	499k	255
	Cross-paragraph	1.2k	118

Table 2. The statistic of PaperPDF. **Text-only** and **Image-only** indicate that the QA pairs are generated based on either a single text paragraph or an image extracted from the PDF. **Image-text**, **Section**, and **Cross-paragraph** denote that the QA pairs are generated from a paragraph and its corresponding references, an entire section, or non-contiguous paragraphs, respectively.

4.2. Detail of Construction

The construction process of PaperPDF are presented in Fig. 3. It can be divided into four steps: structured parsing, rule-based extraction, prompt construction, and filtering. We obtained 89k PDF documents from the arXiv repository as our document set $\mathbb{D}$. For each document D in $\mathbb{D}$,

we first employed Grobid to parse the document and extract its text chunks $\{T_1, T_2, \ldots, T_m\}$ and images chunks $\{I_1, I_2, \ldots, I_n\}$. Subsequently, predefined rules are used to randomly select chunks from the document, which may consist of text chunks T_i, chart chunks I_j or a combination of both. The selected chunks will be input into commercial MLLM products according to different prompt templates and then, the question Q related to the input chunks, along with the corresponding answer A, is subsequently generated. Notably, For the training set, we used Gemini for generation due to its free and rapid accessibility, while the high-performance GPT-4V was employed to construct the test set, ensuring the validity and robustness of the evaluation. Finally, we devise a set of rules to filter the generated training and testing sets automatically. The removing rules for the samples include too-short questions, too-long answers, non-English text, etc. Further manual checking is conducted on test set to ensure the reliability of the evaluation. PaperPDF consists of two types of QA pairs: **Single-evidence** and **Multi-evidence**.

Single-evidence QA pair means that the question can be answered based on a single text chunk or image chunk in the long document. It can be systematically categorized into two types: text-only and image-only. In text-only generation, a text chunk is input into MLLM, whereas in image-only generation, an image chunk is utilized. The generation process for Single-evidence QA pairs is relatively straightforward and cost-effective, primarily aimed at training the foundational multimodal understanding and question-answering capabilities of MLLMs.

Multi-evidence QA pairs. Unlike single-evidence QA pairs, Multi-evidence QA data exhibits greater complex-

ity, both in terms of characteristics and generation process. The answers to such questions typically rely on multiple text chunks, image chunks, or any combination thereof in document. Multi-evidence QA pairs consist of Image-text QA pairs, Section QA pairs, and Cross-paragraph QA pairs. The Image-text QA pairs are generated based on a paragraph and its corresponding image references. These pairs require consideration of the interrelationship between the text and the image to ensure the answers capture multi-modal information. Section QA pairs are designed to train the MLLM in integrating and understanding information within a section and are generated from all chunks within a section. Cross-paragraph QA pairs involve the most complex generation process. First, the entire document is input into the MLLM for paragraph-level semantic summarization, followed by the selection of semantically related text chunks. Finally, multiple related chunks are randomly selected and re-input into the MLLM to generate the final QA pairs. This type of data primarily focuses on training the sampler's performance, enhancing its sparse sampling capability. Although the generation of Multi-evidence QA pairs is relatively complex and may require multiple MLLM calls, their presence significantly improves the performance of the sampler and strengthens the MLLM's ability to comprehend cross-chunk information.

Totally, we generated 1.5M samples for training and 8K samples for testing. After filtering, the final dataset consists of 1M training data points and 6K test data points for subsequent training and evaluation. The statistics of PaperPDF are presented in Tab. 2.

5. Experiments

5.1. Implementation Details

We adopt XComposer2-4KHD [25] as our baseline model, initializing it with its pre-trained weights. We use BGE-M3 [50] as the text encoder. We fine-tune the model using the datasets PaperPDF, DocVQA [35], ChartQA [37], InfoVQA [40], MPDocVQA [44], and DUDE [17], with a learning rate of 4e-5. Prior to both training and testing, all datasets underwent document parsing; specifically, Paper-PDF was parsed using Grobid [51], while the other datasets were processed with MinerU [52]. For the sparse sampler, we selected the top 5 sampling results by default to serve as input to the large language model. The training was conducted for one epoch across these datasets using 128 Ascend 910B GPUs. For convenient description, we denote the 4 different input formats of models with 4 symbols. As shown in Tab. 4, * means that we input the parsed image-text interleaved content of the PDF file, while τ, † and ‡ represent the input of OCR only, entire page image and entire page image & OCR content respectively.

5.2. Long PDF Understanding

To assess the effectiveness of our model in understanding long PDF documents, we conducted comprehensive experiments comparing it with both open-source models and commercial products on PaperPDF dataset.

Due to the limited availability of open-source models capable of handling this task, we firsly evaluate two such models that can process multi-page PDF documents. Considering that IXC2-VL [45] cannot understand such long image-text interleaved sequences, we also report the results with the input of pure OCR content. To further demonstrate the advantages of our multi-modal sampler, we demonstrate the evaluation of IXC2-VL with RAG where we feed the top 5 text paragraphs retrieved by BGE-M3 [50]. The results on the PaperPDF dataset are reported in Tab. 3. The findings indicate that our model significantly outperforms existing open-source models across various evaluation metrics. Additionally, thanks to the introduction of the sparse sampler, the number of tokens that our model's MLLM needs to process is substantially fewer than that of other models.

Moreover, we compared our model with several commercial products that allow users to input PDF documents and questions via web interfaces to generate answers. Considering the cost of manual testing, we randomly selected 50 PDFs from the PaperPDF dataset for evaluation, and the results are presented in Tab. 3. Our model clearly outperforms mainstream PDF question-answering products. Fig. 4 shows some examples from PDF-WuKong and other proprietary products. Our model can accurately retrieve the evidence needed to answer user questions and generate accurate answers based on relevant text segments or diagrams. Since these commercial solutions do not provide detailed code or technical reports, we could not obtain information about their parameter sizes or token counts, and uploading documents directly to their websites does not allow us to accurately assess the models' latency. Finally, we evaluated the GPT accuracy of all open-source and proprietary models on this subset benchmark. The results show that our model has significant advantages in both textual similarity and semantic understanding.

To compare with more open-source large document models, considering that most of these models can only handle single-page documents, we constructed a subset of the PaperPDF benchmark containing only test samples with single evidence. Therefore, we provided all models with only the page containing the evidence as input; note that the pages are input in the form of images. These models and our PDF-WuKong have not been trained on such data. In this setting, our sparse sampler sampled all the content of that page. As shown in Tab.5, our model's zero-shot capability on this subset is significantly better than other document large models.
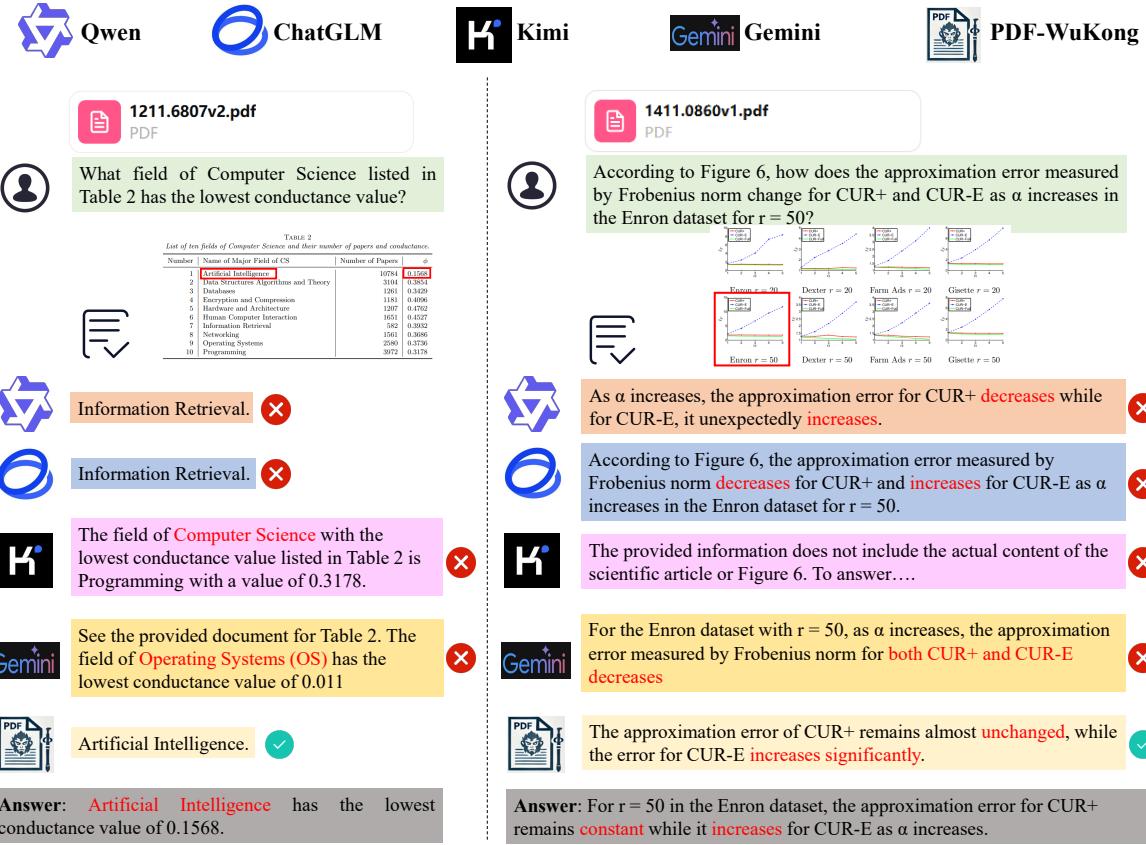
Qwen   ChatGLM   Kimi   Gemini   PDF-WuKong

1211.6807v2.pdf PDF

What field of Computer Science listed in Table 2 has the lowest conductance value?

TABLE 2
List of ten fields of Computer Science and their number of papers and conductance.

Number	Name of Major Field of CS	Number of Papers	ϕ
1	Artificial Intelligence	10784	0.1568
2	Data Structures Algorithms and Theory	3104	0.3854
3	Databases	1261	0.3429
4	Encryption and Compression	1181	0.4096
5	Hardware and Architecture	1207	0.4762
6	Human Computer Interaction	1651	0.4527
7	Information Retrieval	562	0.3932
8	Networking	1561	0.3686
9	Operating Systems	2580	0.3736
10	Programming	3972	0.3178

Qwen: Information Retrieval. ✗

ChatGLM: Information Retrieval. ✗

Kimi: The field of Computer Science with the lowest conductance value listed in Table 2 is Programming with a value of 0.3178. ✗

Gemini: See the provided document for Table 2. The field of Operating Systems (OS) has the lowest conductance value of 0.011 ✗

PDF-WuKong: Artificial Intelligence. ✓

Answer: Artificial Intelligence has the lowest conductance value of 0.1568.

1411.0860v1.pdf PDF

According to Figure 6, how does the approximation error measured by Frobenius norm change for CUR+ and CUR-E as α increases in the Enron dataset for r = 50?

Enron $r = 20$ Dexter $r = 20$ Farm Ads $r = 20$ Gisette $r = 20$

Enron $r = 50$ Dexter $r = 50$ Farm Ads $r = 50$ Gisette $r = 50$

Qwen: As α increases, the approximation error for CUR+ decreases while for CUR-E, it unexpectedly increases. ✗

ChatGLM: According to Figure 6, the approximation error measured by Frobenius norm decreases for CUR+ and increases for CUR-E as α increases in the Enron dataset for r = 50. ✗

Kimi: The provided information does not include the actual content of the scientific article or Figure 6. To answer…. ✗

Gemini: For the Enron dataset with r = 50, as α increases, the approximation error measured by Frobenius norm for both CUR+ and CUR-E decreases ✗

PDF-WuKong: The approximation error of CUR+ remains almost unchanged, while the error for CUR-E increases significantly. ✓

Answer: For r = 50 in the Enron dataset, the approximation error for CUR+ remains constant while it increases for CUR-E as α increases.

Figure 4. Examples from PDF-WuKong and other proprietary products

Table 3. Performance comparison with other commercial products for PDF multimodal understanding. (*, ⊤) indicates the use of a special input approach. For details, refer to Table 4. If no symbol is present, the PDF file is input directly into the model. The **Open** part is tested on the full set of PaperPDF; The **Closed** part and **GPT Acc** is tested on a subset of 550 QA from 50 PDF files in PaperPDF. Note: The **Closed** models refer to the document understanding products based on these models rather than the models themselves.

	Model	#param	ANLS	F1	ROUGE	Token	Latency	GPT Acc
Open	Hi-VT5 [44]‡	316M	13.5	3.1	3.7	11589	0.81s	15.2
	IXC2-VL [45]*	8B	13.7	3.8	4.6	15261	9.42s	19.5
	IXC2-VL [45]⊤	8B	27.4	30.8	32.8	4620	9.25s	-
	IXC2-VL-RAG [45]⊤	8.5B	32.4	34.0	32.4	585	1.84s	-
	PDF-WuKong (ours)*	8.5B	**41.9**	**43.5**	**40.9**	1689	4.72s	**77.5**
Closed	Qwen [46]	-	<u>36.0</u>	<u>40.3</u>	<u>35.5</u>	-	-	**78.1**
	Kimi [47]	-	28.5	33.6	31.1	-	-	74.7
	Gemini pro [48]	-	26.6	29.0	29.8	-	-	67.9
	ChatGLM [49]	-	31.2	35.4	32.0	-	-	73.5
	PDF-WuKong (ours)*	8.5B	**41.8**	**43.2**	**40.7**	-	-	<u>77.5</u>

5.3. Document-oriented VQA

To validate the generalization capability of our model to other document understanding scenarios, we conducted experiments on several benchmark datasets and compared PDF-WuKong with other representative models.

First, we evaluated the performance of PDF-WuKong and other open-source models on DocVQA [35], ChartQA [37], and InfoVQA [40], which are all single-

Table 4. Description of the input format used in the experiment

Symbol	Description
*	Input the parsed content of the pdf file.
⊤	Input the OCR content only.
†	Input the entire page image.
‡	Input the entire page image and OCR content.

Table 5. Performance comparison with other DocVLMs for PDF multimodal understanding on Single-Evidence Subset of Paper-PDF. (†) indicates the use of a special input approach. For details, refer to Table 4.

Model	# param	ANLS	F1	ROUGE
Qwen-VL [53][†]	9.6B	26.4	19.6	18.3
Monkey [54][†]	9.8B	30.0	24.4	22.3
mPLUG-Owl2 [26][†]	8.2B	19.5	20.3	22.7
Emu2-Chat [55][†]	37B	26.0	24.4	23.4
MiniCPM-2.5 [56][†]	8.5B	31.8	28.2	24.8
IXC2-VL [45][†]	8B	23.4	20.8	21.3
IXC2-4KHD [25][†]	8B	24.5	20.0	18.0
CogVLM2 [57][†]	17B	24.8	27.4	26.3
PDF-WuKong (ours)[†]	8.5B	**36.6**	**35.2**	**31.7**

page document datasets. As shown in Tab. 6, our model achieved leading performance compared to other open-source models. This demonstrates that PDF-WuKong can effectively handle various types of documents and questions, showcasing its versatility in document-oriented visual question answering tasks.

In addition, we assessed the performance of traditional specialized models and large-scale models on two existing multi-page document QA datasets. The experimental results, presented in Tab. 7, indicate that our model's performance in multi-page document scenarios is comparable to these specialized models and far surpasses the latest document large model, DocOwl2 [26]. Notably, on complex multi-page document datasets like DUDE [58], PDF-WuKong outperforms GPT-4V [59]. This improvement is attributed to our sparse sampler, which effectively filters out useful information from multi-page documents, enabling the model to focus on relevant content.

Furthermore, we conducted zero-shot evaluations on a new long multimodal document understanding benchmark MM-NIAH [18]. As shown in Tab. 8, our model uses the fewest parameters yet achieves the second-best performance. Although InternVL-V1-5-RAG [18] surpasses PDF-WuKong by 2.8%, it utilizes 36.5 billion more parameters than our model. Moreover, as the context length

of the multimodal documents increases, the performance of our model remains stable, unlike other models that experience significant declines. At a context length of 64K, PDF-WuKong achieves the best results, demonstrating its robustness in handling long-context multimodal inputs.

Table 6. Performance comparison with other DocVLMs on single-page document-oriented VQA benchmarks. (†) indicates the use of a special input approach. For details, refer to Table 4.

	Model	Doc.	Chart.	Info.
Closed	Gemini Pro [60]	88.1	74.1	**75.2**
	GPT-4V [61]	**88.4**	78.5	75.1
Open	Qwen-VL [53][†]	65.1	65.7	35.4
	Monkey [54][†]	66.5	65.1	36.1
	Text-Monkey [10][†]	73.0	66.9	28.6
	DocOwl 1.5 [62][†]	82.2	70.2	50.7
	MiniCPM-V-2.5 [56][†]	84.8	-	-
	Vary-base [12][†]	76.3	66.1	-
	DeepSeek-vl-7b [63][†]	71.9	-	-
	IXC2-VL [45][†]	72.6	57.7	34.4
	IXC2-4KHD16 [25][†]	84.9	**80.1**	60.8
	PDF-WuKong (ours)[†]	85.1	80.0	61.3

Table 7. Performance comparison with other DocVLMs for multi-page document understanding. (†) indicates the use of a special input approach. For details, refer to Table 4.

Model	MP-DocVQA	DUDE
Longformer [64][⊤]	55.1	27.1
BigBird [65][⊤]	58.5	26.3
LayoutLMv3 [66][*]	55.1	20.3
Hi-VT5 [44][*]	61.8	35.7
DocFormerv2 [67][‡]	76.4	48.4
GRAM [16][‡]	**83.0**	53.4
GPT-4(v) [61][‡]	-	53.9
Idefics3-8B [68][†]	67.2	38.7
DocOwl2 [26][†]	69.4	46.7
PDF-WuKong (ours)[‡]	76.9	**56.1**

5.4. Ablation Study

To comprehensively evaluate the effectiveness of our proposed model components, we conducted ablation studies focusing on datasets, the impact of the sparse sampler, sampling strategies, and document length. Below, we present the findings from each of these experiments.

Table 8. Performance comparison with other DocVLMs on MM-NIAH. The input approach aligns with the benchmark.

Model	#param	Overall	1K	4K	16K	64K
Emu2-Chat [55]	37B	8.8	38.9	18.2	0.0	0.0
VILA1.0-13b [69]	13B	15.7	41.9	33.2	8.6	0.1
llava-v1.6-13b [70]	13B	16.9	43.7	34.9	13.6	0.0
llava-v1.6-34b [23]	34B	20.6	57.4	45.1	8.2	0.0
InternVL1.5 [11]	26B	41.1	59.5	50.1	41.9	16.6
InternVL1.5-RAG [18]	45B	**46.1**	59.5	50.1	44.9	39.3
PDF-WuKong (ours)	8.5B	43.3	53.0	43.9	43.0	**42.1**

Datasets

We conducted experiments to evaluate the impact of training data size and composition on model performance. As shown in Tab. 9, increasing the amount of training data led to consistent improvements in the model's accuracy on the PaperPDF benchmark. Specifically, the model trained with larger datasets demonstrated better comprehension and answer accuracy. Moreover, under the same data volume, training sets that included multi-evidence annotations outperformed those containing only single-evidence. This suggests that exposure to diverse and complex evidence during training enhances the model's ability to handle queries requiring multiple pieces of information, thereby improving overall performance.

Table 9. Ablation study on dataset setting

Dataset	ANLS	F1	ROUGE
100 k	38.7	40.1	37.5
500 k	41.6	43.5	**40.8**
1 M	**42.6**	**43.6**	40.2

Impact of sparse sampler

To assess the effectiveness of the sparse sampler, we compared models trained with and without it. Without the sparse sampler, the MLLM struggled to process long documents with interleaved text and images, resulting in poor performance due to the overwhelming amount of irrelevant information. Introducing the sparse sampler significantly improved the model's accuracy, as evidenced in Tab. 10, by efficiently selecting the most relevant content for each query. Furthermore, end-to-end joint training of the sparse sampler and the MLLM led to additional performance gains compared to training them separately. This indicates that our integrated design allows the two components to mutually enhance each other, optimizing both the retrieval of pertinent information and the generation of accurate answers.

Table 10. Ablation study on the impact of sparse sampler

Sparse Sampler	End-to-End	ANLS	F1	ROUGE
✗	✗	11.1	5.1	5.0
✓	✗	40.3	42.3	39.8
✓	✓	**42.6**	**43.6**	**40.2**

Sampling strategy

We explored different sampling strategies by varying the number of top N text paragraphs or figures selected by the sparse sampler. As shown in Tab. 11, setting N too low resulted in missing critical information necessary for accurately answering queries, thereby reducing performance. Conversely, a larger N introduced redundant information and increased computational and time costs without significant improvements in accuracy. This suggests that there is an optimal range for N that balances the inclusion of essential information with efficiency. Selecting the appropriate number of samples is crucial for maximizing performance while minimizing resource consumption.

Table 11. Ablation study of different sampling strategy

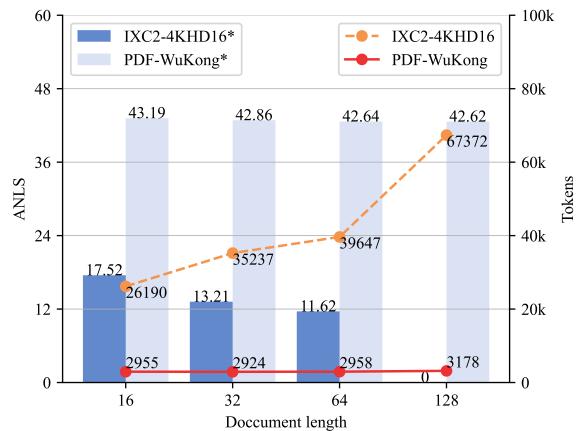| Sampling chunks	ANLS	F1	ROUGE	Tokens
Top 1	39.20	38.28	35.26	1186
Top 3	42.09	43.06	39.69	1452
Top 5	42.59	43.63	40.19	1789
Top 10	43.01	44.22	40.67	2386
Top 15	43.19	44.57	42.08	2704
Top 20	43.42	45.02	42.30	3364

Figure 5. Ablation study of different document length

Document length

To understand the impact of document length on model performance and efficiency, we divided the test set into subsets based on the number of pages per document. Results in Fig. 5 demonstrate that our model's performance and time efficiency remained relatively stable across documents of varying lengths. This stability indicates that the sparse sampler effectively reduces the input size to a manageable level, regardless of the original document length. In contrast, MLLMs without the sparse sampler were unable to handle long documents effectively; their performance and time efficiency deteriorated significantly as the document length increased. These findings highlight the robustness of our approach in processing long documents without sacrificing accuracy or incurring additional computational costs.

6. Conclusion

We have presented PDF-WuKong, a novel Multimodal Large Language Model that effectively addresses the challenges of understanding long PDF documents containing interleaved text and images. By introducing an end-to-end sparse sampling mechanism, our model efficiently extracts the most relevant paragraphs and diagrams in response to user queries, significantly reducing input token size and making the process independent of document length. We also constructed PaperPDF, a comprehensive dataset with 1 million question-answer pairs for training and 6,000 pairs for evaluation, specifically tailored for academic PDFs. Experimental results demonstrate that PDF-WuKong not only outperforms existing open-source models but also surpasses proprietary products by an average of 8.6% in F1 score on long multimodal PDF understanding tasks. Our approach maintains high accuracy and efficiency even as document length increases, offering a scalable and interpretable solution for practical applications in document understanding.

References

[1] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, David Seunghyun Yoon, Ryan A. Rossi, and Franck Dernoncourt. Pdftriage: Question answering over long, structured documents, 2023. 1

[2] T. Prem Jacob, Beatriz Lucia Salvador Bizotto, and Mithileysh Sathiyanarayanan. Constructing the chatgpt for pdf files with langchain – ai. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 835–839, 2024. 1

[3] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3

[4] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. LongloRA: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[5] Szymon Tworkowski, Konrad Staniszewski, Mikoł aj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Mił oś. Focused transformer: Contrastive training for context scaling. In *Advances in Neural Information Processing Systems*, volume 36, pages 42661–42688, 2023. 1, 3

[6] Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*, 2023. 1, 3

[7] Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. Disc-finllm: A chinese financial large language model based on multiple experts finetuning. *arXiv preprint arXiv:2310.15205*, 2023.

[8] Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024.

[9] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. 1, 3

[10] 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*, 2024. 1, 3, 9

[11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 3, 10

[12] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023. 1, 3, 9

[13] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024. 2, 3

[14] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 2, 3

[15] 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 *AAAI*, pages 13636–13645, 2023. 2

[16] Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, Aviad Aberdam, Shahar Tsiper, and Ron Litman. Gram: Global reasoning for multi-page vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2024. 2, 3, 9

[17] 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*, pages 19528–19540, 2023. 2, 7

[18] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024. 2, 9, 10

[19] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[20] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding, 2023. 3

[21] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 3

[22] 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*, 2023. 3

[23] 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 2024. 3, 10

[24] Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023. 3

[25] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 3, 7, 9

[26] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding, 2024. 3, 9

[27] Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, 2021. 2

[28] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, 2021. 2

[29] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 2

[30] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

[31] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: a large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3743–3751, 2022. 2

[32] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalický, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. Docile benchmark for document information localization and extraction. pages 147–166, 2023. 2

[33] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS*, 2019.

[34] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, pages 1516–1520, 2019. 2

[35] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 2, 7, 8

[36] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019. 2

[37] 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. 2, 7, 8

[38] Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024. 2

[39] 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. 2

[40] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 2, 7, 8

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin

Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[42] Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024. 2

[43] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[44] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 7, 8, 9

[45] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 7, 8, 9

[46] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 8

[47] Moonshot AI. Kimi. https://kimi.moonshot.cn, 2023. 8

[48] 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. 8

[49] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 8

[50] 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. *arXiv preprint arXiv:2402.03216*, 2024. 7

[51] Grobid. https://github.com/kermitt2/grobid, 2008–2024. 7

[52] MinerU Contributors. Mineru: A one-stop, open-source, high-quality data extraction tool. https://github.com/opendatalab/MinerU, 2024. 7

[53] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 9

[54] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023. 9

[55] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 9, 10

[56] 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. 9

[57] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 9

[58] Jordy Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). *arXiv:2305.08455*, 2023. 9

[59] OpenAI. Gpt-4v(ision) system card. https://openai.com/contributions/gpt-4v, 2023. 9

[60] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 9

[61] OpenAI. Gpt-4 technical report, 2023. 9

[62] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 9

[63] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 9

[64] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 9

[65] 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. 9

[66] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACMM*, pages 4083–4091, 2022. 9

13

[67] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. Docformerv2: Local features for document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 709–718, 2024. 9

[68] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 9

[69] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 10

[70] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 10

A. Dataset Examples

Text-only Data Example in the Training Set

Query: What is the impact of using the same dataset for optimizing and measuring the performance of a model?

Text: Here, again, the unfair advantage of optimizing (selecting the models for the ensemble) and measuring performance on the same dataset appears. The advantage is small but systematic for the test split of ISIC (Fig. 5a); it is much more apparent for the challenging collection of clinical images of EDRA Atlas (Fig. 5b).

Answer 1: It can lead to an unfair advantage for the model.
Answer 2: Optimizing a model involves selecting certain parameters or features that improve its performance on a given dataset. If the same dataset is used to measure the model's performance, it may lead to an unfair advantage as the model has already been "tuned" to that specific dataset.

Text-image Data Example in the Training Set

Query: What is the best performing method for both detection and classification, according to the provided figure?

Text: 3) The last one is our proposed SFCN-OPI with both sibling branches and OPI (Ours in Table 1).
Figure: Table 1

Methods	Detection			Classification		
	P	R	F1	P	R	F1
FCN-5CLS	0.741	0.867	0.790	0.466	0.264	0.298
SFCN	0.784	0.844	0.807	0.450	0.561	0.496
SFCN-OPI-1	0.764	**0.890**	0.816	0.573	0.667	0.613
SFCN-OPI-2	0.788	0.885	0.828	0.674	0.759	0.711
SSAE	0.617	0.644	0.630	-	-	-
LIPSyM	0.725	0.517	0.604	-	-	-
CRImage	0.657	0.461	0.542	-	-	-
SR-CNN	0.783	0.804	0.793	-	-	0.683
SC-CNN	0.781	0.823	0.802	-	-	0.692
Ours	**0.819**	0.874	**0.834**	**0.718**	**0.774**	**0.742**

Note: the - means the results were not reported by that method.

Caption: Experimental results of ablation analysis, different training strategies of our method and comparison with other approaches. Note: the -means the results were not reported by that method.

Answer 1: Ours.
Answer 2: The best performing method for both detection and classification is Ours, as it achieves the highest F1 scores for both tasks. This can be seen in the "Ours" row of the table, where the F1 score for detection is 0.834 and the F1 score for classification is 0.742.
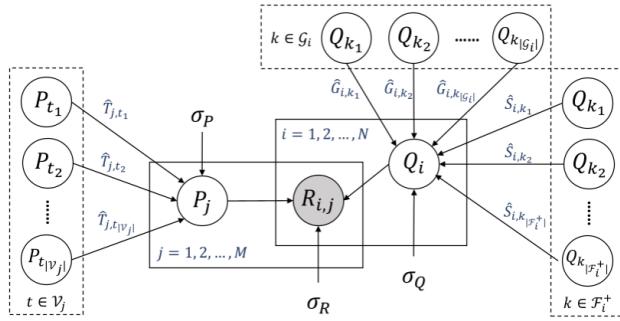
Section Data Example in the Training Set

Query: What kind of information does the factor R i, j hold?

Text:
RECOMMENDATION MODEL
In this section, we present our Joint Topic-Semantic-aware Matrix Factorization (JTS-MF) model for online social votings, in which social relationship, group a liation, and topic-semantic similarities are combined and taken into account for voting recommendation in a comprehensive manner.Motivated by Locally Linear Embedding [23] which tries to preserve the local linear dependency among inputs in the low-dimensional embedding space, we expect to keep inter-user and inter-voting topic-semantic similarities in latent feature space as well.To this end, in JTS-MF model, while the rating R i, j is factorized as user latent feature Q i and voting latent feature P j , we deliberately enforce Q i and P j to be dependent on their social-topic-semantic similar counterparts, respectively.e graphic model of JTS-MF model is as shown in Figure 5.

Figure: Figure 5

Caption: Fig. 5: Graphic Model of JTS-MF.

Answer 1: The factor R i, j represents the rating that user i gave to voting j.
Answer 2: The node labeled R i, j is a factor that represents the rating given by user i to a voting j. It is the core of the model, representing the essence of user interaction with the voting system. This node's existence suggests that the model is primarily concerned with understanding and predicting user ratings for voting content.

Text-only Data Example in the Test Set

Query: How does the distribution of output image transform tasks among N workers impact the efficiency of the final synchronization task?

Text: The number of output image transform tasks equals the number of output images times the batch size. The tasks are executed by all N workers, such that each worker picks up an arbitrary task and executes it. The last output image transform task to finish also executes the final synchronization task, which frees the memory required for the output image transforms.

Answer 1: The distribution allows for parallel processing, potentially speeding up the final synchronization task.
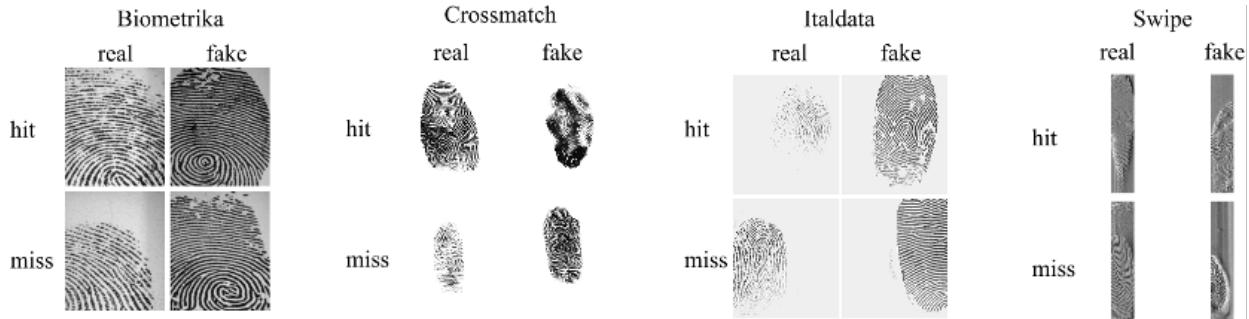**Answer 2:** Ensures faster task completion, synchronization efficiency relies on all tasks' completion.

Text-image Data Example in the Test Set

Query: What distinguishing features can be observed between 'hit' and 'miss' samples for real fingerprints in Fig. 5 across different benchmarks?

Text: Images found in these benchmarks can be observed in Fig. 5 of Section V.As we can see, variability exists not only across modalities, but also within modalities.Moreover, it is rather unclear what features might discriminate real from spoofed images, which suggests that the use of a methodology able to use data to its maximum advantage might be a promising idea to tackle such set of problems in a principled way.

Figure: Figure 5

Caption: Fig. 5. Examples of hit and missed testing samples lying closest to the real-fake decision boundary of each benchmark. A magnified visual inspection on these images may suggest some properties of the problem to which the learned representations are sensitive.
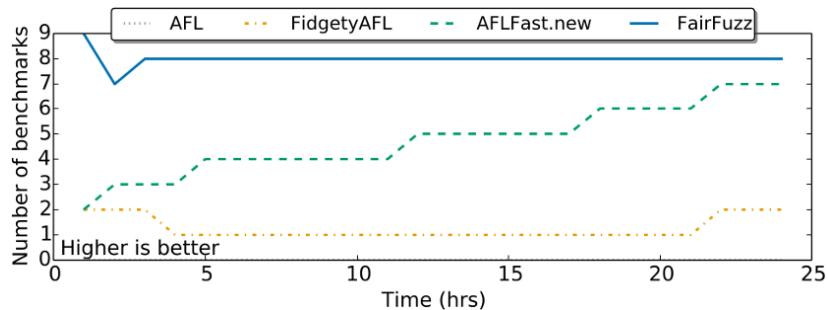
Answer 1: Hit' samples for real fingerprints show clearer ridge patterns and quality compared to 'miss' samples throughout the benchmarks.
Answer 2: Hit real fingerprints are clearer, more distinct; Miss real samples are blurred, less defined.

Only-image Data Example in the Test Set

Query: Based on Figure 4, which fuzzing technique consistently leads in coverage across all benchmarks over the 24-hour period?

Figure: Figure 4

Caption: Figure 4: Number of benchmarks on which each technique has the lead in coverage at each hour. A benchmark is counted for multiple techniques if two techniques are tied for the lead.

Answer 1: FairFuzz consistently leads in coverage across all benchmarks over the 24-hour period in Figure 4.
Answer 2: FairFuzz, highest coverage benchmark count over 24 hours.

Section Data Example in the Test Set

Query: How does the qualitative evaluation of extractive summarizers using word clouds elucidate the differences in content focus between the original documents and the summaries?
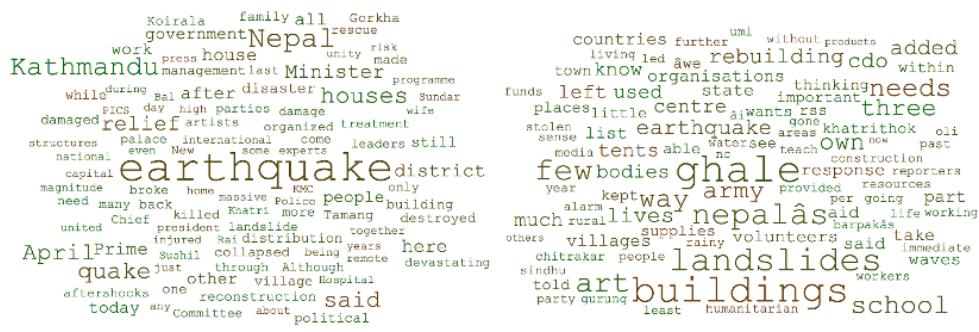
Text:
Here we use word cloud representations to give an intuitive interpretation of the content in the generated extractive summarizers.We create word clouds for the two best methods in Section 5.3.In this paper, we used an online tool called WordItOut5 to generate the word cloud representations.In all word clouds presented in this paper, a filter is used to display only the words with minimum frequency of 2.

Figure 3 shows a word cloud made by the aggregation of all the summaries generated by the PKUSUMSUM-Centroid method.This gives a sense of the content in those summaries.For contrast, we also generate a word cloud for the original news articles without the content of the generated summaries.Specifically, common words are first removed completely and then the word clouds are built with frequencies of surviving words.In essence, this shows what information remains apart from the generated summaries.

The images clearly show a contrast of content.The summary wordcloud shows "earthquake" as its most prominent word.The image of the articles show less focus.If viewed alone, the reader would not quickly infer the gist of the original content.Similarly, Figure 4 represents "Lead" method.And here we see an even more stark difference.6. Comparing Twitter and News Media information.

Figure: Figure 3

Caption: Figure 3: The word clouds representing summaries generated by PKUSUMSUM-Centroid method (left) and original documents without the content of those summaries (right).

Answer 1: Word clouds highlight the prominent themes in summaries versus original texts by displaying relative word frequencies visually.

Answer 2: Visual contrast in word frequency highlights content focus differences.

Cross-paragraph Data Example in the Test Set

Query: How can we leverage the proposed EDO approach to optimize the selection of datasets for specific algorithms, thereby enhancing the overall performance and validity of the algorithms?
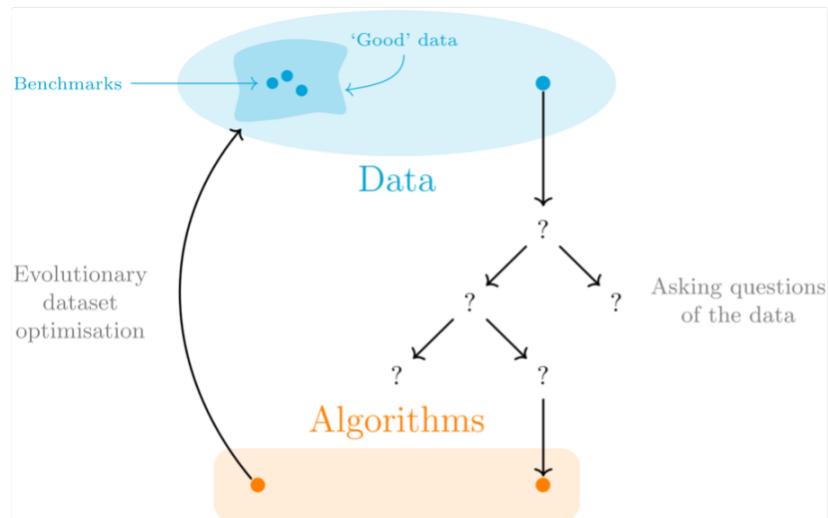
Text:

Paragraph 7: Figure 1: On the right: the current path for selecting some algorithm(s) based on their validity and performance for a given dataset.On the left: the proposed flip to better understand the space in which 'good' datasets exist for an algorithm.

Paragraph 17: Section 2 describes the structure of the proposed method including its parameters and operators.

Paragraph 20: In this section, the details of an algorithm that generates data for which a given function, or (equivalently) algorithm, is well-suited is described.This algorithm is to be referred to as "Evolutionary Dataset Optimisation" (EDO).

Figure: Figure 1

Caption: Sample number of rows and columns 2. Sample columns and fill/trim values as needed.

Answer 1: The EDO approach identifies optimal datasets, improving algorithm performance and validity through tailored data selection.

Answer 2: By employing the EDO approach, we can systematically evaluate and select datasets that align closely with the specific requirements of the algorithms in use. This method analyzes various dataset characteristics, ensuring that the chosen data not only matches the algorithm's operational parameters but also enhances its predictive accuracy. As a result, the performance metrics of the algorithms improve significantly, leading to more valid outcomes and reliable interpretations. The focus on tailored dataset selection directly contributes to optimal algorithm efficiency, thereby validating the effectiveness of the EDO approach in practical applications.

B. Prompt Engineering

Text-only QA Generation Prompt for the Training Set

Task Definition:
Generate 3 high quality academic questions and corresponding answers based on the provided text/image/table.
Requirements:
- The questions should require very complex reasoning and global understanding and the answer should be detailed and can be answered by the material, do not add extra information.
- Remember it is only a paragraph rather than a whole passage so do not ask global question about the study.
- Generate question-answer pairs instead of multiple-choice questions.
- Do not generate global question about the paper such as main idea or abstract.
- You should use English.

Expected Output:
List your Q and A as:
[Q1]:
[A1]:
[Q2]:
[A2]:
[Q3]:
[A3]:

Remember again all the questions should be academic and can be answered by the information in the material.Do not add extra information.Do not ask about the main idea.

Figure 6. Text-only QA generation prompt for the training set.

Text-image QA Generation Prompt for the Training Set

Task Definition:
Generate 2 high-quality academic questions and corresponding answers based solely on a provided figure of a research paper, without relying on accompanying text for the questions. However, you may use the provided text to understand the figure. Ensure these questions demand complex reasoning, focusing exclusively on details within the figure that are critical for understanding. The answers must be detailed, directly reflecting the content in the figure and capable of being derived solely from the figure without extraneous information.

Requirements:
Thoroughly analyze the provided figure of the research paper, attaining a deep understanding of its contents, including but not limited to, the study's objectives, methodologies, results, and conclusions. Develop 2 academic questions that:
• Must not mention the name of the figure directly or words like "figure","table" in the question.
• Demand complex reasoning and a comprehensive understanding of the whole figure.
• Must ask questions that can only be answered by the content in the figure, without reliance on textual information.
• Integrate knowledge from the entire figure, reflecting a nuanced understanding. Frame questions to elicit detailed and informed responses directly supported by the figure.
Craft detailed answers for each question that:
• Are directly derived from the provided figure, excluding information not found within the material.
• Are comprehensive and cover all relevant aspects, as presented in the provided figure. Accurately reflect the information and insights offered by the research paper figure.

Expected Output:
List your Q and A as:
[Q1]: [Formulate the first academic question here, ensuring it requires complex reasoning and encompasses the entirety of the provided figure for a comprehensive understanding. Based solely on the figure, without referencing accompanying text. Must not mention the name of the figure directly. Must not mention words like "figure", "table" exc.]
[A1]: [Provide a detailed answer here, derived exclusively from the information within the provided figure, ensuring the response is thorough and precise without including extraneous details. Mention from which figure/table you get the answer.]
[Q2]: ...
[A2]: ...

Here is the supplementary paragraph text for the figure:

Figure 7. Text-image QA generation prompt for the training set.

Section QA Generation Prompt for the Training Set

Task Definition:
Generate 3 high-quality academic questions and corresponding answers based on a section of a research paper.
Requirements:
Each question should generate two types of answers. The first answer should be concise, directly addressing the question with minimal wording. The second answer should include a "chain of thought" that provides a reasoning process and be detailed. You should use English.
Generate questions that:
• Can be answered with simple reasoning and only require a global understanding. The question should be able to answered directly with the material. Do not include 2 or more subquestions in each question. Ensure that the concise answers can be provided using sentences or phrases that do not exceed 20 words in length.
• Remember it is only a paragraph rather than a whole passage so do not ask global question about the study.
• Generate question-answer pairs instead of multiple-choice questions.
• Do not generate global question about the paper such as main idea or abstract.
• Must not mention the name of the figure directly or words like "figure", "table" exc in the question.
• Must ask questions that can only be answered by the content in the figure, without reliance on textual information.
Craft the 2 answers for each question that:
• Are directly derived from the provided figure, excluding information not found within the material.
• Are comprehensive and cover all relevant aspects, as presented in the provided figure. Accurately reflect the information and insights offered by the research paper figure.
Expected Output:
List your Q and A as:
[Q1]: [Insert the academic question here, can be answered directly from the material]
[A11]: [Insert the concise answer here, providing a straightforward, brief response directly addressing the question, no more than 20 words]
[A12]: [Insert the detailed answer here, based solely on the given information. This answer must include a detailed "thought chain" or reasoning process, detailing how the conclusions are drawn from the image and caption. Must not mention the name of the figure/table from which the answer is derived directly, without adding extraneous details.]
[Q2]: ...
[A21]: ...
[A22]: ...
[Q3]: ...
[A31]: ...
[A32]: ...
Remember again all the questions should be academic and can be answered by the information in the material.Do not add extra information.Do not ask about the main idea.
Here is the supplementary paragraph text for the figure:

Figure 8. Section QA generation prompt for the training set.

Cross-paragraph Question Generation Prompt for the Training Set

Task Definition:
Based on the selected paragraph from a research paper that share a thematic or conceptual connection, formulate an insightful, open-ended question. This question should reflect the shared themes or concepts of your selections and relate to the broader context of the research paper.

Requirements:
- Ascertain the underlying connection among the paragraphs and the figures/tables(if provided).
- Subsequently, craft an insightful, open-ended question that encapsulates the identified themes or connections, aiming to foster analytical thinking and in-depth discussion on the subject matter of the paper.
- Note that your question should not directly include the "idx"s of the paragraphs.

Expected Output:
[Q]: [Your generated question based on the shared themes or information]

Here are the selected paragraphs in the paper:

Cross-paragraph Answer Generation Prompt for the Training set

Task Definition:
Given some selected paragraphs from a research paper, each chosen based on their relevance (excluding the first 3 paragraphs), and ensuring that these paragraphs share a certain level of association, you are to answer a question that is related to the content of these selected paragraphs. The question is crafted to encompass the themes or findings presented in the paragraphs of the chosen paragraphs, aiming for a comprehensive understanding and connection between these elements.

Requirements:
Craft the 2 answers for the question that:
- Are directly derived from the provided figure, excluding information not found within the material.
- Are comprehensive and cover all relevant aspects, as presented in the provided figure. Accurately reflect the information and insights offered by the research paper figure.

Expected Output:
[A1]: [Insert the concise answer here, providing a straightforward, brief response directly addressing the question, no more than 20 words.]
[A2]: [Insert the detailed answer here, based solely on the given information. This answer must include a detailed "thought chain" or reasoning process, detailing how the conclusions are drawn from the image and caption. Must not mention the name of the figure/table from which the answer is derived directly, without adding extraneous details.]

The question is:

Figure 9. Cross-paragraph QA generation prompt for the training set.

Text-only Question Generation Prompt for the Test set

Task Definition:
Create 2 academic questions from a given research paper paragraph.
Requirements: Analyze the paragraph thoroughly,understanding its content including the study's objectives, ethods, results,and conclusions. Focus on the paragraph,not the entire paper. If the paragraph lacks valid information,return 'quit'. You should use English.
Develop 2 questions that:
• No more than 30 words.
• Incorporate knowledge from the paragraph.
• Should be answered by text instead of one of the multiple choices.
• Elicit detailed responses supported by the text.
Expected Output: (Return 'quit' directly if the paragraph lacks valid information.)
[Q1]: question1 here
[Q2]: question2 here

Text-only Answer Generation Prompt 1 for the Test set

Task Definition:
Answer 2 questions based on the material given.
Requirements:
The answers should be:
• No more than one sentence, less than 20 words.
• Comprehensive and cover all relevant aspects.
• Accurately reflect the paragraph's information and insights.
You should think step by step and give you answer in the end of your generation like: [thinking procedure]: [A1/A2]
Expected Output:
[THINKING PROCEDURE]: ...
[A1]: answer1 here, no more than one sentence.
[THINKING PROCEDURE]: ...
[A2]: answer2 here, no more than one sentence.

Text-only Answer Generation Prompt 2 for the Test set

Task Definition:
Answer 2 questions based on the material given.
Requirements:
The answers should be:
• Within a few keywords, less than 20 words.
• Comprehensive and cover all relevant aspects.
• Accurately reflect the paragraph's information and insights.
You should think step by step and give you answer in the end of your generation like: [thinking procedure]: [A1/A2]
Expected Output:
[THINKING PROCEDURE]: ...
[A1]: answer1 here, within a few keywords.
[THINKING PROCEDURE]: ...
[A2]: answer2 here, within a few keywords.

Figure 10. Text-only type QA generation prompt for the test set.

Text-image Question Generation Prompt for the Test set

Task Definition:
Formulate 2 academic questions based on the provided figures and tables from a research paper.
Requirements: Analyze the paragraph thoroughly,understanding its content including the study's objectives, ethods, results,and conclusions. Focus on the paragraph,not the entire paper. If the paragraph lacks valid information,return 'quit'. You should use English.
Develop 2 questions that:
• No more than 30 words.
• Are specific to the unique data or details visible in the figures/tables and are answerable only based on the material without inferring or speculating on details not explicitly explained by the figures/tables.
• Must not mention the label of the figure/table directly or use words like 'from the figure/table'.
Expected Output:
[Q1]:
[Q2]:

Text-image Answer Generation Prompt 1 for the Test set

Task Definition:
Answer 2 questions based on the material given.
Requirements:
The answers should be:
• No more than one sentence, less than 20 words.
• Always use English.
• Do not infer or speculate on details not explicitly explained by the figures/tables.
You should think step by step and give you answer in the end of your generation like: [thinking procedure]: [A1/A2]
Expected Output:
[THINKING PROCEDURE]: ...
[A1]: answer1 here, no more than one sentence.
[THINKING PROCEDURE]: ...
[A2]: answer2 here, no more than one sentence.

Text-image Answer Generation Prompt 2 for the Test set

Task Definition:
Answer 2 questions based on the material given.
Requirements:
The answers should be:
• Within a few keywords, less than 20 words.
• Comprehensive and cover all relevant aspects.
• Accurately reflect the paragraph's information and insights.
You should think step by step and give you answer in the end of your generation like: [thinking procedure]: [A1/A2]
Expected Output:
[THINKING PROCEDURE]: ...
[A1]: answer1 here, within a few keywords.
[THINKING PROCEDURE]: ...
[A2]: answer2 here, within a few keywords.

Figure 11. Text-image QA generation prompt for the test set.

Section Question Generation Prompt for the Test set

Task Definition:
Formulate 2 academic questions based on a section from a research paper.
Requirements:
Carefully read and comprehend the entire provided section of the research paper to ensure a thorough understanding of its content, including key points, findings, methodologies, and conclusions. You should Always use English.
Develop 2 questions that:
• No more than 30 words.
• Require an integration of information from all paragraphs and figures/tables in the section.
• Must not mention the label of the figure/table directly or use words like 'from the figure/table'.
• Not based on common knowledge or assumptions not supported by the figures and tables.
Expected Output:
[Q1]:
[Q2]:

Section Answer Generation Prompt 1 for the Test set

Task Definition:
Answer 2 questions based on the material given.
Requirements:
The answers should be:
• No more than one sentence, less than 20 words.
• Always use English.
• Do not infer or speculate on details not explicitly explained by the figures/tables.
You should think step by step and give you answer in the end of your generation like: [thinking procedure]: [A1/A2]
Expected Output:
[THINKING PROCEDURE]: ...
[A1]: answer1 here, no more than one sentence.
[THINKING PROCEDURE]: ...
[A2]: answer2 here, no more than one sentence.

Section Answer Generation Prompt 2 for the Test set

Task Definition:
Answer 2 questions based on the material given.
Requirements:
The answers should be:
• Within a few keywords, less than 20 words.
• Comprehensive and cover all relevant aspects.
• Accurately reflect the paragraph's information and insights.
You should think step by step and give you answer in the end of your generation like: [thinking procedure]: [A1/A2]
Expected Output:
[THINKING PROCEDURE]: ...
[A1]: answer1 here, within a few keywords.
[THINKING PROCEDURE]: ...
[A2]: answer2 here, within a few keywords.

Figure 12. Section QA generation prompt for the test set.