Docopilot: Improving Multimodal Models for Document-Level Understanding

Supplementary Material

7. Details of Data Construction

7.1. Document Data Format

As stated in Section 3.1, documents in Doc-750K can be extracted in two formats: **Interleaved Text-Image Format** and rendered **Multi-Image Format**. The interleaved format utilizes an external PDF parser to extract text directly, avoiding OCR errors. However, this approach sacrifices some layout information. The multi-image format, commonly used in screenshot-based QA scenarios, preserves the complete layout but relies on the model's built-in OCR capabilities, which may introduce errors. Each format has its advantages and is suitable for different applications. By leveraging both formats, the model can develop complementary capabilities, enhancing its robustness across diverse input types. Examples of each format are shown in Figure 5 and Figure 6, respectively.

7.2. Image Types and Tasks Coverage

We provide additional comparisons of Doc-750K with various document-level datasets in terms of image types and task coverage, as shown in Table 8. Compared to these datasets, Doc-750K exhibits greater diversity in proxy tasks and ranks among the largest datasets in terms of QA pair count.

7.3. Question-Answer Pairs Generation

Prompt. The prompt used to generate question-answer pairs from GPT-4o is shown below.

```
Please read the paper and first check if this is
an English paper. If it is not an English paper,
don't do any other things. If it is an English
paper, please design about 3 to 5 question-answer
pairs based on the paper. All questions should
require as much text as possible to answer and it is
better to ask about the images in the papers. All
images in the questions should be represented as the
mentioned title in the paper like Figure 1/Figure 2
or Table 1/ Table 2 and they must be mentioned in the
questions. The question should be specific enough
that it can only be answered with the paper. The
question should also be interesting and intellectual
enough that a curious reader of the paper would ask
about it. The answer of the QA pair must start with
"According to the original text ......", first give
the relevant original text in the reference content,
and then answer the question in detail. Please try
to analyze the asked image in the answer. Please
directly output a list of the QA pairs without any
other outputs. Here is the paper:
<paper>
```

Examples. In Section 3.1, we propose diverse question-answer formats tailored to data from different sources. To fully utilize the webpage structure of OpenReview, we develop tasks focused on review writing and replies within its review-reply framework. For Sci-Hub and Arxiv, we use their well-defined writing structures to create tasks such as writing and translating various sections. We provide examples of these various QA formats in Figure 7.

Quality Evaluation. The quality of Doc-750K is ensured through the following measures: (1) Human-Originated Data: QA pairs from OpenReview are derived from human-written discussions, providing high contextual quality. (2) Structured Tasks: Tasks like abstract writing and paper titling are constructed based on document metadata, following deterministic rules to ensure reliability. (3) Synthetic QA: We randomly sample and manually review 500 training QA pairs across tasks and 498 of 500 (over 99%) of the pairs were relevant.

8. Evaluation Details

8.1. Benchmark Metrics

We report the metrics of benchmarks used in the evaluation in Table 9. For DocVQA [53], InfoVQA [54], and MP-DocVQA [76], we employ ANLS to evaluate the similarity between model responses and ground truth answers, while ChartQA [52] uses Relaxed Exact Match (Relaxed EM). For open-ended QA tasks in MMLongbench-Doc [51] and DocGenome [89], we utilize GPT-4o to assess the correctness of answers and calculate GPT Accuracy. Other tasks in DocGenome follow their official evaluation metrics.

8.2. Evaluation Settings

For single-page benchmarks such as DocVQA, ChartQA, and InfoVQA, we conduct the evaluation using a unified prompt as follows:

```
<image>
<question>
Answer the question using a single word or phrase.
```

For multi-page benchmarks, we discuss them case by case. We employed image concatenation for multi-page VQA benchmarks like MP-DocVQA, MMLongBench-Doc, and DocGenome to reduce the excessive input patches. Adjacent pages were vertically concatenated into a single image, with a maximum total image count limit of 18.

(1) For MPDocVQA, we use the prompt for a N concatenated page document as follows:

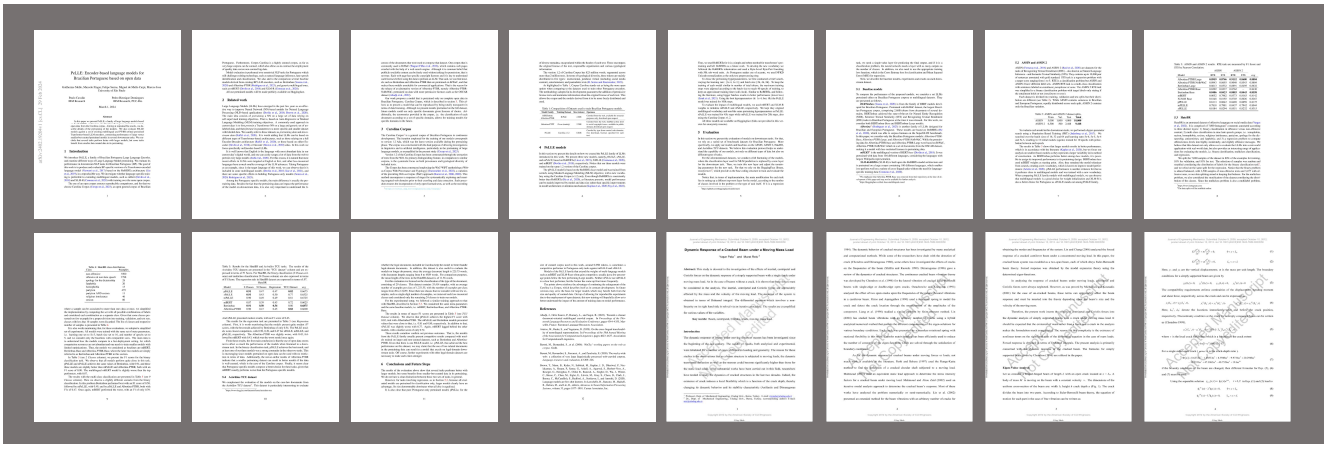


Figure 5. **An example of multi-image format document.** Each page of the document is rendered as an image.

Datasets	#QA	Image Types	Tasks
MP-DocVQA [76]	46K	PDF Documents	VQA
DUDE [78]	41K	PDF Documents	VQA
MP-DocStruct1M [27]	1M	PDF Documents	Text Parsing, Text Lookup
MP-DocReason51K [27]	51K	PDF Documents, Infographics, Webpages, Charts, Natural images	VQA
DocGenome [89]	N/A	PDF Documents	Layout Detection, Document Transformation
Doc-750K (ours)	758K	PDF documents, Charts, Tables	VQA, Abstract Writing, Paper Titling, Caption Writing, Experiment Writing, Translation, Review, Reply

Table 8. **Comparison with other document-level datasets.**

```
Image-1:  <concat-page 1>
Image-2:  <concat-page 2>
...
Image-N: <concat-page N>
<question>
Answer the question using a single word or phrase.
```

(2) For MMLongBench-Doc and DocGenome, we use the official prompt in their open-sourced code base for response generation and extract the correctness of the answer using GPT-4o.

(3) For MM-NIAH, we use the original interleaved data format and calculate the accuracy by their official judgment function.

Evaluation of LLMs on Multimodal Document Benchmarks. We utilized the InternVL2-8B as the OCR model to extract text from each image of the document, followed by post-processing to remove redundant responses. The text extraction prompt is as follows:

```
Image-1:  <image>
Please extract the text from Image-1, while retaining
as much of the original formatting and structured
information (such as headings, paragraphs, lists,
tables, charts, etc.)  as possible.  If the document
is not in PDF format, provide a caption for Image-1.
Present the extracted information directly without
additional explanations.
```

We concatenated the extracted texts to replace the original document images for the language models:

```
Page 1:
<text 1>
Page 2:
<text 2>
...
Page N:
<text N>
```

For the image-text interleaved data, we replaced the images with the captions as the input.

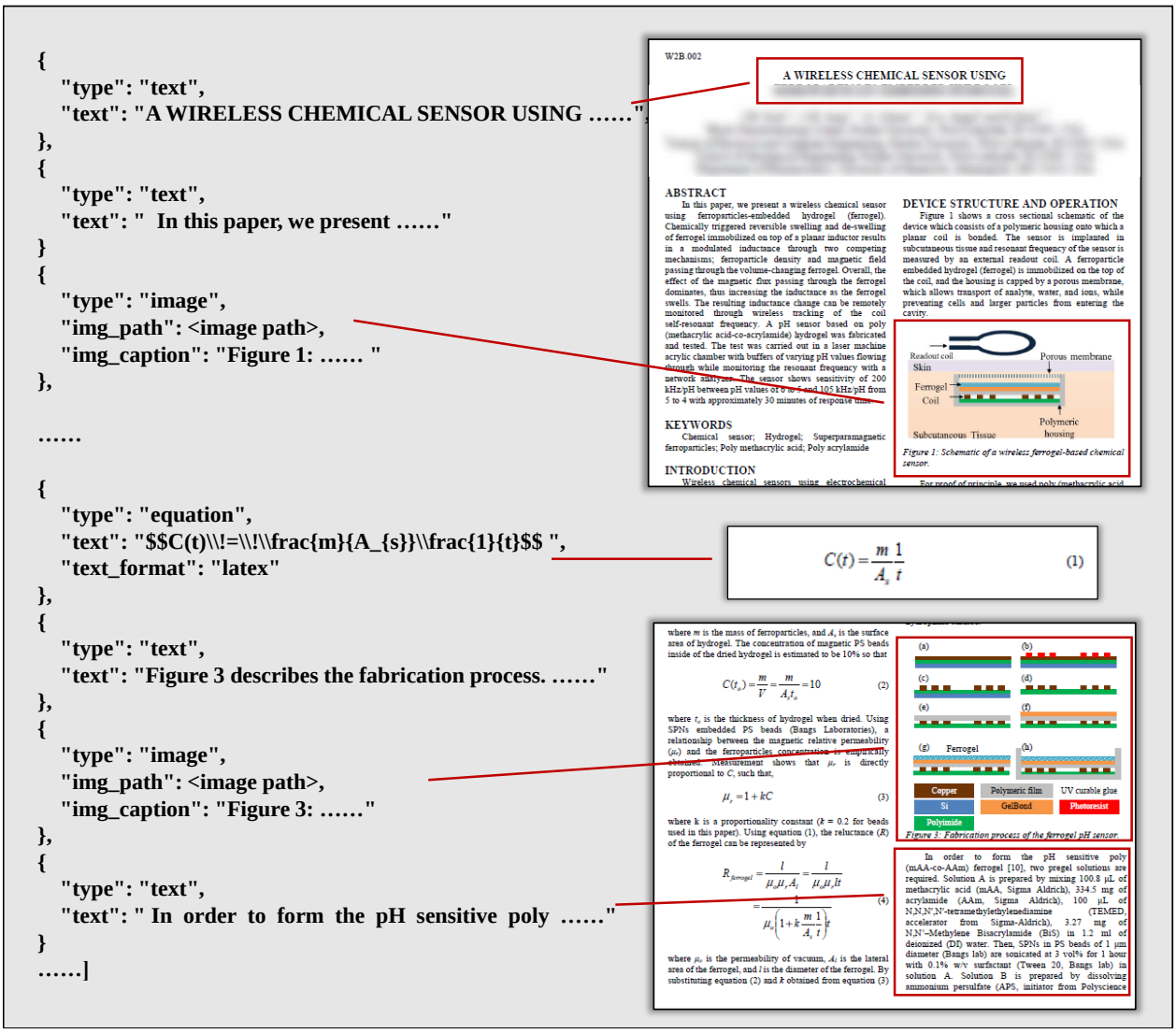


Figure 6. **An example of interleaved text-image format document.** We capture the documents' textual content and construct them into interleaved images and text.

Implementation Details of Multimodel RAG. VisRAG [95] uses their proposed retrieval model VisRet to calculate scores for each image and text segment based on the query. InternVL-RAG [85] utilizes InternVL-14B, a CLIP-like model, to compute the similarity between images and text. For multi-paged VQA benchmarks, we select the top-3 retrieved documents for generation. For interleaved VQA, we choose up to 8K tokens for the generation.

9. Training Details

9.1. Hyperparameters

We report the models and training hyperparameters of Docopilot-2B and Docopilot-8B in Table 10.

9.2. Multimodal Packed Dataset

In this section, we provide a detailed description of our packing algorithm. The main workflow is outlined in Algorithm 1. Specifically, our algorithm constructs the packed dataset by combining individual samples drawn from the original dataset. The packing operation involves four steps:

(1) **Check Sample:** Given an individual sample, we first verify whether the number of images exceeds the image threshold T_i or the number of tokens exceeds the token threshold T_t. If either condition is met, the sample is truncated into N parts. The first $N-1$ parts contain exactly T_i images or T_t tokens and are immediately added to the output. The remaining part is passed to the subsequent steps for further processing.

Review Writing
Question: Please review the following paper and provide a constructive critique. Focus on the methodology, results, and overall contributions, and highlight both strengths and areas for improvement. Your review should be detailed and insightful, offering suggestions for enhancing the research. Here is the paper:
<paper>
Answer: <review>
Reply Writing
Question: Given the following paper and its review, write a reply to address the feedback provided. Here is the paper:
<paper>
Here is the review:
<review>
Answer: <reply>
Abstract Writing
Question: Please read the full text of the following research paper and provide a concise summary in the form of an abstract. The summary should capture the main objectives, methods, results, and conclusions of the paper. Ensure that the abstract is clear, coherent, and informative for readers who have not read the full paper. Here is the paper:
<paper w/o abstract>
Answer: <abstract>
Paper Titling
Question: Based on the provided abstract or introduction of the research paper, please generate a concise and informative title that accurately reflects the main focus and contributions of the paper. The title should be engaging and clearly convey the essence of the research. Here is the abstract and introduction:
<abstract>
<introduction>
Answer: <title>
Caption Writing
Question: The images or tables and their relative texts in a research paper are given interleaved as follows. Please write a caption for each image or table based on the relative texts provided. Here is the image:
<image>
Here is the relative text:
<text>
Answer: <caption>
Experiments Writing
Question: Please write the "Experiments" section based on the incomplete research paper provided. Ensure that the section is well-structured and includes the details of the Figures and tables. Here is the paper:
<paper w/o experiments>
Answer: <experiments section>
Translation
Question: Please read the full text of the following research paper and translate the Experiments/Abstract section into Chinese. Here is the paper:
<paper> (Multi-Image format)
Answer: <translation>
Multi-Turn QA (interleaved and rendered)
Question: Please read the full paper and answer the question. Here is the paper:
<paper>
Here is the question: <question> (generated)
Answer: <answer> (generated)
Question: <question> (generated)
Answer: <answer> (generated)

Figure 7. **The examples of different formats of the QA pairs.** All tasks leverage the inherent structure of the documents.

Layout	Benchmark	Description	Metric
Single-Page	DocVQA [53]	VQA on documents.	ANLS
	ChartVQA [52]	VQA on charts.	Relaxed EM
	InfoVQA [54]	VQA on infographics.	ANLS
Multi-Page	MP-DocVQA [76]	VQA on multi-page documents.	ANLS
	MMLongBench-Doc [51]	VQA on super-long PDF documents.	Accuracy, F1 Score
	DocGenome [89]	VQA on multi-page scientific documents.	Classification Acc, Title ED, Abstract ED, Single-Page Acc, Multi-Page Acc
Interleaved	MM-NIAH [86]	VQA on natural texts or images.	Accuracy

Table 9. **Evaluation benchmarks and metrics for document understanding.** This table summarizes key benchmarks used in document understanding tasks across three layout types: single-page, multi-page, and interleaved.

	Settings	Docopilot-2B	Docopilot-8B
Model	ViT	InternViT-300M	InternViT-300M
	LLM	InternLM2-1.8B	InternLM2.5-7B
Training Hyperparameters	Tile Resolution	448	
	Batch Size	128	
	Optimizer	AdamW	
	Learning Rate	1.00E-05	
	Warmup Ratio	0.03	
	LR Scheduler	Cosine	
	Weight Decay	0.01	0.05
	ViT Drop Path	0.1	
	Max Tile Number	24	
	Image Threshold	48	
	Token Threshold	32K	
	Epochs	1	

Table 10. **Training settings and hyperparameters for Docopilot models.** Key configurations for Docopilot-2B and Docopilot-8B, including model architectures and training parameters.

(2) **Find Buffer:** For the remaining part of the sample, we attempt to find a suitable buffer from the buffer list to pack it together with the sample. The combined result must not exceed the thresholds T_i for images or T_t for tokens, while maximizing the total number of images and tokens in the packed sample. To speed up this process, the buffer list is organized as a priority queue.

(3) **Pack Samples:** The given sample and the selected buffer are packed together. Notably, during the training process, each token can only attend to other tokens within the same original sample. Tokens from other samples packed together remain inaccessible.

(4) **Maintain Buffer List:** After generating a packed sample, we check if its number of images or tokens meets the specified thresholds. If so, the sample is added to the output; otherwise, it is reinserted into the buffer list for potential future packing. Note that we omit numerous edge cases for brevity.

Algorithm 1 Multimodal Packed Dataset

Input: Dataset $\mathcal{D}$, buffer list $\mathcal{B}$, Token Threshold T_t, Image Threshold T_i
Output: Packed Dataset $\mathcal{D}_{\text{packed}}$
foreach *data_sample* d *in* $\mathcal{D}$ **do**
  $b \leftarrow \texttt{find\_buffer}(d, B)$
  $b_p \leftarrow \texttt{pack}(d, b)$
  if b_p *contains more than* T_i *images or* T_t *tokens* **then**
    $\texttt{yield}\ b_{packed}$
  else
    $\texttt{insert}(b_{packed}, B)$

10. Qualitative Examples

In this section, we show a series of qualitative examples to illustrate the effectiveness of our Docopilot in handling complex multi-page documents. Each figure highlights a specific capability of the model in addressing various tasks.

Figure 8 demonstrates the model's ability to accurately retrieve relevant information from a multi-page document, showcasing its capability to perform robust cross-page retrieval tasks.

In Figure 9, we illustrate the model's proficiency in performing backward queries, where context must be traced across pages in reverse order to locate relevant content.

Figure 10 highlights the consistency of answers when the model is queried across multiple pages. This example demonstrates that the model maintains coherence and accuracy even when information is distributed across different parts of the document.

Figure 11 presents an example of counting across pages, showcasing the model's ability to integrate numerical information from disparate locations in a document.

Lastly, Figure 12 demonstrates the model's capability to pinpoint specific information within a designated section

of a super-long document, emphasizing its fine-grained retrieval capabilities.

These examples collectively highlight the robustness and adaptability of our approach in understanding and processing multi-page documents effectively.

11. Limitations

The objective of Doc-750K is to develop a large-scale, multi-task, multimodal document-level QA dataset that efficiently trains MLLMs for document understanding. Sourced primarily from open academic platforms, the dataset focuses on tasks like multi-page QA, reasoning, and translation, with some academic-specific tasks such as titling and summarization. A key limitation of Doc-750K is its current domain restriction to academic documents. We plan to expand the dataset's coverage to a broader range of document types and enhance the generalizability of proxy tasks, ensuring wider applicability across diverse domains.

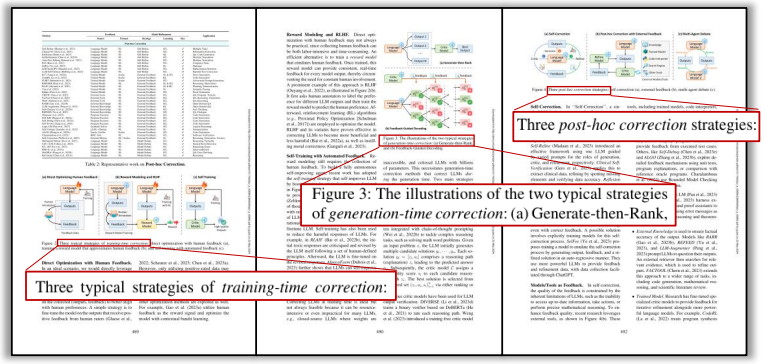


Figure 3: The illustrations of the two typical strategies of *generation-time correction*: (a) Generate-then-Rank,

Evidence Pages 6,7,9 (total 23P)

Long-context Question:

Image-1: <Concat-Page 1>\n
Image-2: < Concat-Page 2>\n
…
Image-12: < Concat-Page 12>\n
How many strategies are plotted for training-time correction, generation-time correction, and post-hoc correction, respectively? Write the answer in the list format, e.g., ["1","2"]

Answer:

['3', '2', '3']

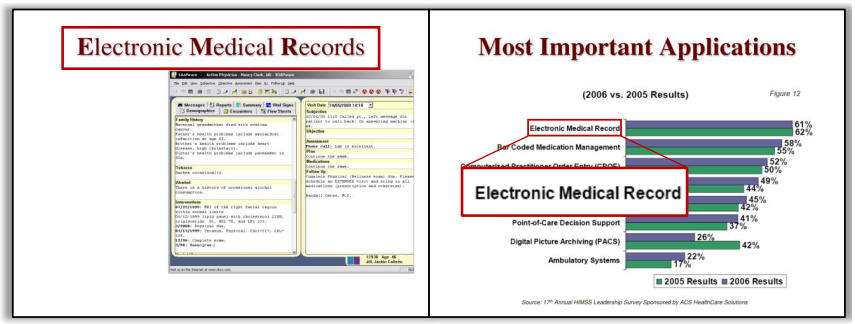
Docopilot-8B:

["3","2","3"] √

InternVL2-8B:

["3","3","3"] ×

Figure 8. **A qualitative example of retrieval in a multi-page document.**



Evidence Pages 7,11 (total 20P)

Long-context Question:

Image-1: <Concat-Page 1>\n
Image-2: <Concat-Page 2>\n
…
Image-10: <Concat-Page 10>\n
Among the most important applications listed, which one is shown a picture of the interface as an example

Answer:

Electronic Medical Records

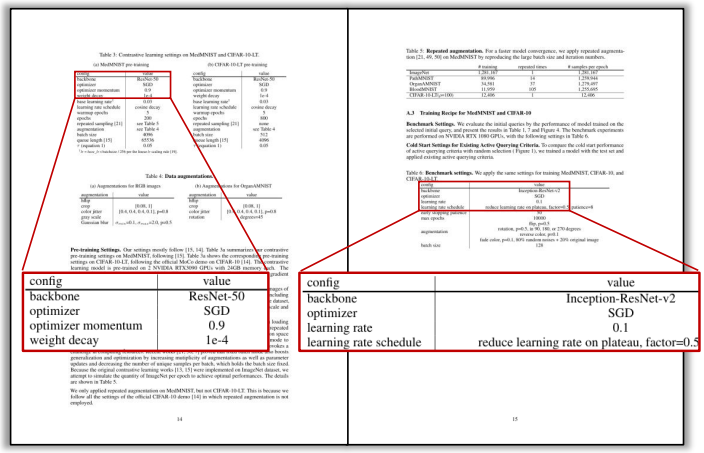
Docopilot-8B:

Electronic Medical Records. √

InternVL2-8B:

Among the most important applications listed, "Point-of-Care Decision Support" is shown a picture of the interface as an example. ×

Figure 9. **A qualitative example of backward query in a multi-page document.**



config	value
backbone	ResNet-50
optimizer	SGD
optimizer momentum	0.9
weight decay	1e-4

config	value
backbone	Inception-ResNet-v2
optimizer	SGD
learning rate	0.1
learning rate schedule	reduce learning rate on plateau, factor=0.5

Evidence Pages 14,15 (total 24P)

Long-context Question:

Image-1: <Concat-Page 1>\n
Image-2: < Concat-Page 2>\n
…
Image-12: < Concat-Page 12>\n
What are the optimizers used in this research?

Answer:

SGD

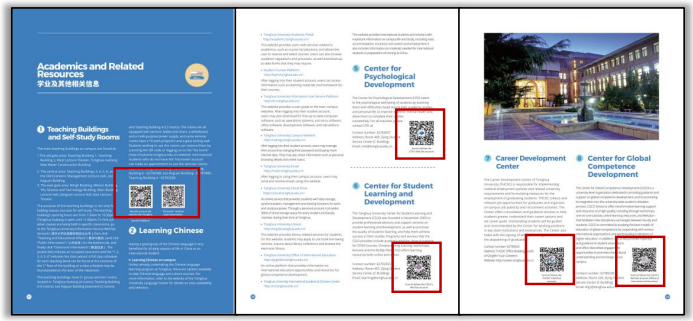
Docopilot-8B:

SGD. √

InternVL2-8B:

The optimizers used in this research are Adam and AdamW. ×

Figure 10. **A qualitative example of consistency of answers in multi pages.**



Evidence Pages 23,25,26 (total 40P)

Long-context Question:

Image-1: < Concat-Page 1>\n
Image-2: < Concat-Page 2>\n
…
Image-14: <Concat-Page 14>\n
How many QR codes are shown in the "Academics and Related Resources" part of this guidebook?

Answer:

6

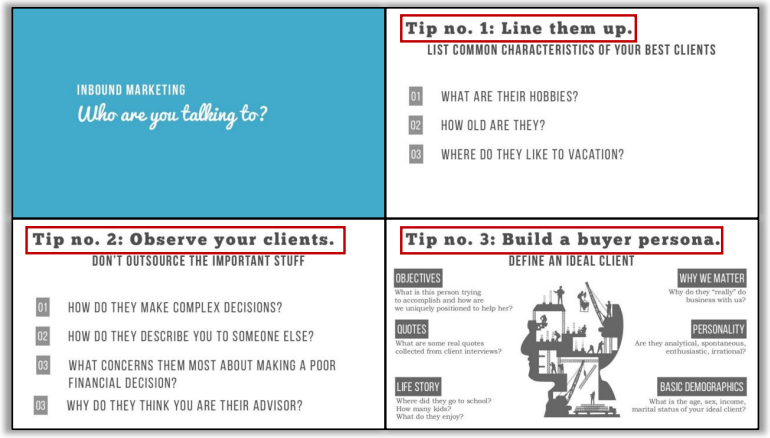
Docopilot-8B:

6. √

InternVL2-8B:

In the "Academics and Related Resources" part of the guidebook, there are three QR codes shown. ×

Figure 11. **A qualitative example of counting across pages.**



Evidence Pages 37,39,40,41 (total 65P)

Long-context Question:

Image-1: <Concat-Page 1>\n
Image-2: < Concat-Page 2>\n
…
Image-17: < Concat-Page 17>\n
In the "Who are you talking to" section, what tips does the author give us?

Answer:

['line them up', 'observe your clients', 'build a buyer persons']

Docopilot-8B:

Line them up,Observe your clients,Build a buyer persona.

√

InternVL2-8B:

In the "Who are you talking to" section, the author provides two key tips for engaging with clients effectively:

1. **Line them up**: …

2. **Observe your clients**: … incomplete

Figure 12. **A qualitative example of retrieval in a specific section of a super-long document.**