GRAM: Global Reasoning for Multi-Page VQA

Supplementary Material

Group	Parameter Name	Parameter Value
fine-tune	batch size	8
	training steps	200K
	warmup steps	1000
	fp16	True
	training number of pages	4
	evaluation number of pages	unlimited
	number of image tokens	128
DocFormer$_{concat}$ [1]	encoder learning rate	3e-5
	decoder learning rate	3e-5
	training text tokens per page	600
	inference text tokens per page	400
HiVT5* [4]	encoder learning rate	3e-5
	decoder learning rate	3e-5
	training text tokens per page	800
	inference text tokens per page	8000
	number of compression tokens per page	10
GRAM	encoder learning rate	3e-5
	decoder learning rate	3e-5
	global encoder learning rate	1e-4
	training text tokens per page	800
	inference text tokens per page	8000
	number of global tokens	32
	bias adaptation constant 'c'	20
GRAM$_{C-Former}$	encoder learning rate	3e-5
	decoder learning rate	3e-5
	global encoder learning rate	1e-4
	C-Former learning rate	1e-4
	training text tokens per page	800
	inference text tokens per page	8000
	number of global tokens	32
	bias adaptation constant 'c'	20
	compression length	256

Table 1. **Hyper-Parameters**.

A. Parameters

We present in Tab. 1 all of the relevant hyperparameters.

B. Inference Resources Consumption

We compare three key properties of MP-DocVQA baselines and our method: inference time, memory consumption, and maximal document length. The latency and memory consumption are illustrated in Fig. 1 and Fig. 2, respectively, both as functions of the number of pages in the document. We compare the following baselines: DocFormerv2$_{concat}$ [1], Hi-VT5 * [4], and our GRAM and GRAM$_{C-Former}$, utilizing the same computational resources employed in all experiments— $8 \times A100$ GPUs with $40GB$ of memory.

The memory consumption of DocFormerv2$_{concat}$ [1] reaches its maximum capacity for documents with only 20 pages, while our method efficiently processes documents, spanning hundreds of pages. Moreover, the presented figures demonstrate that GRAM$_{C-Former}$ maintains a comparable memory footprint to the GRAM model. Nevertheless, there is potential for improvement, as HiVT5* exhibits lower memory consumption. Despite this, we achieve inference times similar to HiVT5* [4], accompanied by a noteworthy enhancement in ANLS.

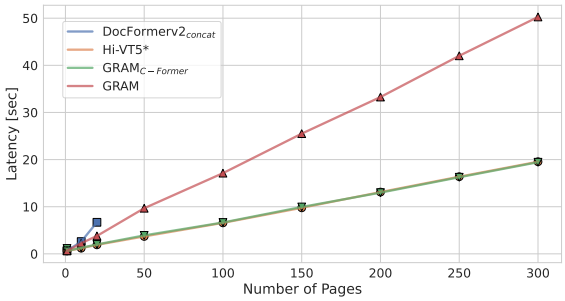


Figure 1. **Latency comparison**. We compare the dependency between overall latency and the number of pages in input document.

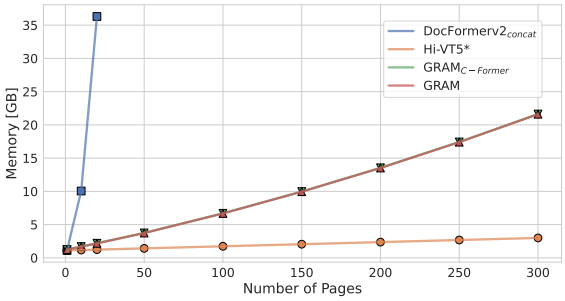


Figure 2. **Memory consumption comparison**. We compare the dependency between overall memory consumption and the number of pages in input document.

C. Qualitative Results

Finally, we present a few qualitative results on the DUDE dataset in Fig. 3, showcasing the advantages of our approach over Hi-VT5* [4]. In the first three examples, we demonstrate cases where GRAM is correct and HiVT5* is wrong. The last two examples present cases where both our method and HiVT5* are incorrect.

D. Comparison with DocFormerV2$_{\text{concat}}$

We provide additional qualitative examples with DocFormerV2$_{\text{concat}}$. Examples demonstrate the effectiveness of GRAM in tackling questions that involve multiple pages in the document.

Method	ANLS by Number of Pages DUDE validation dataset				
	All	1	2-4	5-10	11-end
GRAM	**47.88**	**49.29**	**49.90**	**45.90**	**43.94**
DocFormerv2$_{\text{concat}}$	44.32	46.08	47.05	42.81	38.17
DocFormerV2$_{\text{Longformer}}$	45.88	47.01	47.75	43.22	43.13
DocFormerV2$_{\text{AliBi}}$	34.73	36.55	37.00	30.99	31.25

Table 2. **Comparison to NLP methods**. Results on DUDE validation comparing GRAM with LongFormer [2] and AliBi [3].

E. Comparison with NLP-based Approaches

We present additional experiments, comparing GRAM with two NLP-based approaches: the sparse attention-based LongFormer [2], and the bias-based AliBi [3]. Both approaches are implemented on top of DocFormerv2 for fair comparison. Results in Tab. 2 shows an advantage in our local-global approach of utilizing existing powerful models for single-page and extending them to support the multi-page scenario.

References

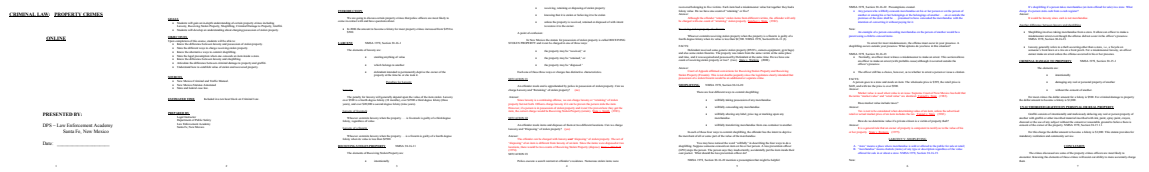
[1] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*, 2023. 1

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 2

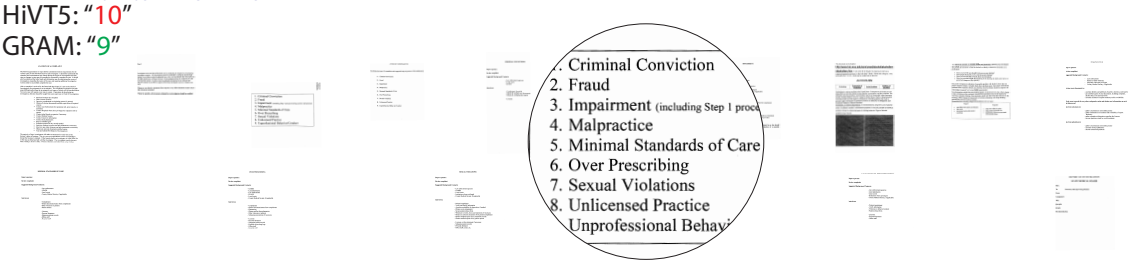
[3] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021. 2

[4] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docvqa. *arXiv preprint arXiv:2212.05935*, 2022. 1, 3

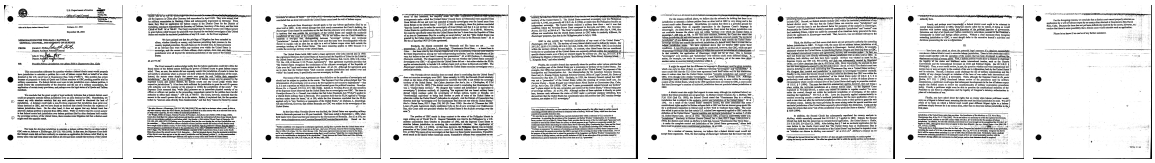
How many chapters are in the books? correct answer: ""
HiVT5: "4"
GRAM: ""



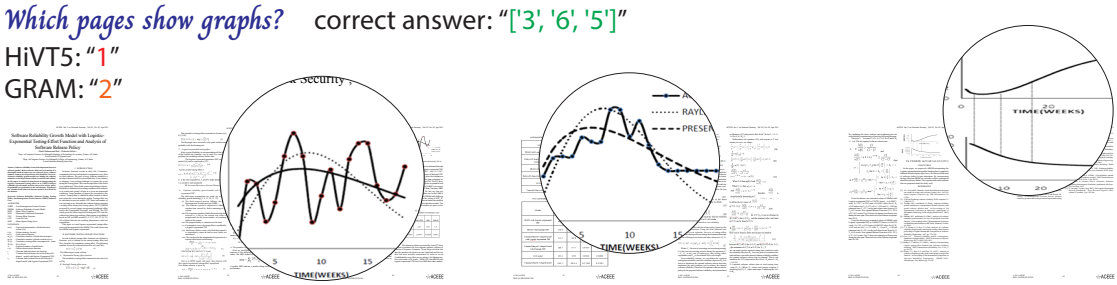
How many types of complaints were listed in the document? correct answer: "9"
HiVT5: "10"
GRAM: "9"



how many pages are there in this text? correct answer: "9"
HiVT5: "1"
GRAM: "9"



Which pages show graphs? correct answer: "['3', '6', '5']"
HiVT5: "1"
GRAM: "2"



Can real property be improved within a mortgage loan? correct answer: "The within mortage does not cover real property improved"
HiVT5: ""
GRAM: ""

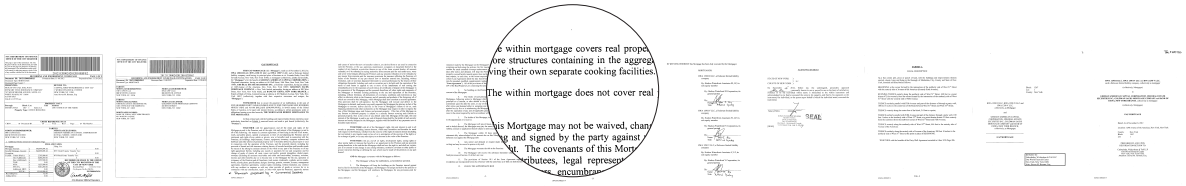


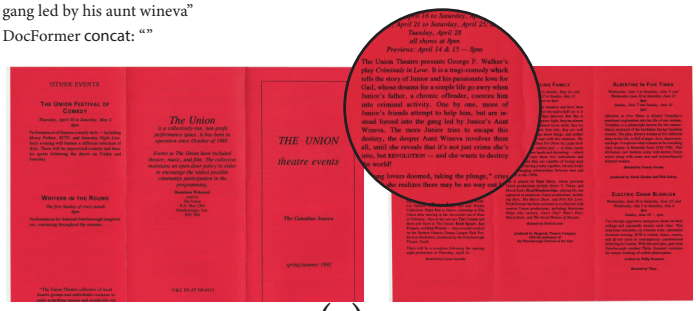
Figure 3. Qualitative comparison between our approach and Hi-VT5 [4] indicates that the integration of our global-local encoder enhances reasoning capabilities, especially when inquiries require multi-page context.

Question: Explain Criminals in Love

GT answer: “Tragi-comedy play. Tells the story of Junior who is forced into criminal activity by his father”

GRAM: “It is a tragi-comedy which tells the story of junior and his passionate love
for gail, whose dreams for a simple life go awry when junior's father, a chronic offender, coerces him into criminal activity. One by one, more of junior's friends attempt to helphim, butare forcedinto the gang led by his aunt wineva”

DocFormer concat: “”



(a)

Question: How many paragraphs are 1 page 1

GT answer: “3”

GRAM: “3”

DocFormer concat: “6”

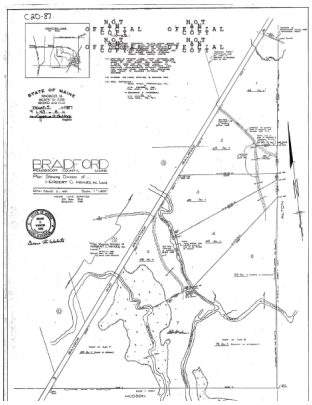
(b)

Question: Which subject use in this document?

GT answer: “”

GRAM: “Location Map”

DocFormer concat: “”



(c)

Question: How many pages with different colors are there in the document?

GT answer: “3”

GRAM: “3”

DocFormer concat: “2”

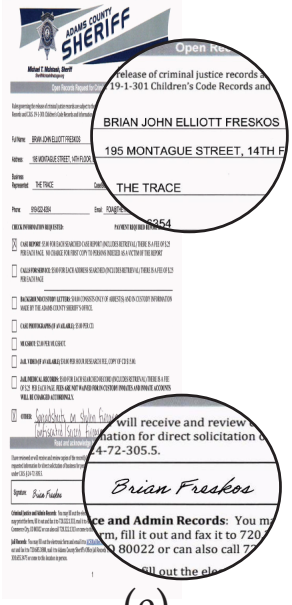
(d)

Question: Does the signature belong to the same person listed at the top of the form?

GT answer: “Yes”

GRAM: “Yes”

256 tokens GRAM: “ BRIAN FRESKO”



(e)

Figure 4. Comparisons between DocFormerV2$_{\text{concat}}$ and GRAM.