

Hierarchical multimodal transformers for Multi-Page DocVQA

Rubèn Tito

Dimosthenis Karatzas

Ernest Valveny

Computer Vision Center, UAB

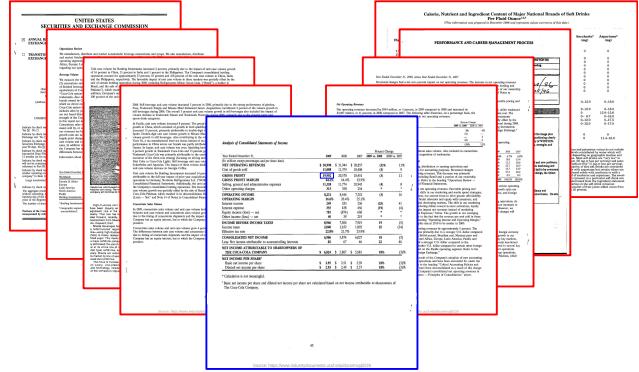
{rperez, dimos, ernest}@cvc.uab.es

Abstract

Document Visual Question Answering (DocVQA) refers to the task of answering questions from document images. Existing work on DocVQA only considers single-page documents. However, in real scenarios documents are mostly composed of multiple pages that should be processed altogether. In this work we extend DocVQA to the multi-page scenario. For that, we first create a new dataset, MP-DocVQA, where questions are posed over multi-page documents instead of single pages. Second, we propose a new hierarchical method, Hi-VT5, based on the T5 architecture, that overcomes the limitations of current methods to process long multi-page documents. The proposed method is based on a hierarchical transformer architecture where the encoder summarizes the most relevant information of every page and then, the decoder takes this summarized information to generate the final answer. Through extensive experimentation, we demonstrate that our method is able, in a single stage, to answer the questions and provide the page that contains the relevant information to find the answer; which can be used as a kind of explainability measure.

1. Introduction

Automatically managing document workflows is paramount in various sectors including Banking, Insurance, Public Administration, and the running of virtually every business. For example, only in the UK more than 1 million home insurance claims are processed every year. Document Image Analysis and Recognition (DIAR) is at the meeting point between computer vision and NLP. For the past 50 years, DIAR methods have focused on specific information extraction and conversion tasks. Recently, the concept of Visual Question Answering was introduced in DIAR [15–17]. This resulted in a paradigm shift, giving rise to end-to-end methods that condition the information extraction pipeline on the natural-language defined task. DocVQA is a complex task that requires reasoning over typed or handwritten text, layout, graphical elements such as diagrams and figures, tabular structures, signatures and the semantics that these convey.



Q: What was the gross profit in the year 2009?

A: \$19,902

Figure 1. In the **MP-DocVQA task**, questions are posed over multi-page documents where methods are required to understand the text, layout and visual elements of each page in the document to identify the correct page (blue in the figure) and answer the question.

All existing datasets and methods for DocVQA focus on single page documents, which is far from real life scenarios. Documents are typically composed of multiple pages and therefore, in a real document management workflow all pages of a document need to be processed as a single set.

In this work we aim at extending single-page DocVQA to the more realistic multi-page setup. Consequently, we define a new task and propose a novel dataset, MP-DocVQA, designed for Multi-Page Document Visual Question Answering. MP-DocVQA is an extension of the Single-DocVQA [16] dataset where the questions are posed on documents with between 1 and 20 pages.

Dealing with multiple pages largely increases the amount of input data to be processed. This is particularly challenging for current state-of-the-art DocVQA methods [9, 18, 28, 29] based on the Transformer architecture [25] that take as input textual, layout and visual features obtained from the words recognized by an OCR. As the complexity of the transformer scales up quadratically with the length of the input sequence, all these methods fix some limit on the number of input tokens which, for long multi-page documents, can lead to truncating a significant part of the input

Dataset	Questions	Documents	Pages (Images)	Avg. pages per question	Question Avg. length	Answer Avg. length	Document Avg. OCR Tokens
SingleDocVQA [16]	50K	6K	12K	1.00	9.49	2.43	151.46
VisualMRC [22]	30K	10K	10K	1.00	10.55	9.55	182.75
InfographicsVQA [15]	30K	5.4K	5.4K	1.00	11.54	1.60	217.89
DuReaderVis [19]	15K	158K	158K	1.3K	9.87	180.54	1968.21
DocCVQA [23]	20	14K	14K	14K	14.00	12.75	509.06
TAT-DQA [31]	16K	2.7K	3K	1.07	12.54	3.44	550.27
MP-DocVQA (ours)	46K	6K	48K	8.27	9.90	2.20	2026.59

Table 1. Comparison between MP-DocVQA and main DocVQA datasets.

data. We will empirically show the limitations of current methods in this context.

As an alternative, we propose the Hierarchical Visual T5 (Hi-VT5), a multimodal hierarchical encoder-decoder transformer build on top of T5 [20] which is capable to naturally process multiple pages by extending the input sequence length up to 20480 tokens without increasing the model complexity. In our architecture, the encoder processes separately each page of the document, providing a summary of the most relevant information conveyed by the page conditioned on the question. This information is encoded in a number of special [PAGE] tokens, inspired in the [CLS] token of the BERT model [7]. Subsequently, the decoder generates the final answer by taking as input the concatenation of all these summary [PAGE] tokens for all pages. Furthermore, the model includes an additional head to predict the index of the page where the answer has been found. This can be used to locate the context of the answer within long documents, but also as a measure of explainability, following recent works in the literature [23, 26]. Correct page identification can be used as a way to distinguish which answers are the result of reasoning over the input data, and not dictated from model biases.

To summarize, the key contributions of our work are:

1. We introduce the novel dataset MP-DocVQA containing questions over multi-page documents.
2. We evaluate state-of-the-art methods on this new dataset and show their limitations when facing multi-page documents.
3. We propose Hi-VT5, a multimodal hierarchical encoder-decoder method that can answer questions on multi-page documents and predict the page where the answer is found.
4. We provide extensive experimentation to show the effectiveness of each component of our framework and explore the relation between the accuracy of the answer and the page identification result.

The dataset, baselines and Hi-VT5 model code and weights are publicly available through the DocVQA Web portal¹ and GitHub project².

¹rrc.cvc.uab.es/?ch=17

²github.com/rubenpt91/MP-DocVQA-Framework

2. Related Work

Document VQA datasets: DocVQA [17, 24] has seen numerous advances and new datasets have been released following the publication of the SingleDocVQA [16] dataset. This dataset consists of 50,000 questions posed over industry document images, where the answer is always explicitly found in the text. The questions ask for information in tables, forms and paragraphs among others, becoming a high-level task that brought to classic DIAR algorithms an end purpose by conditionally interpreting the document images. Later on, InfographicsVQA [15] proposed questions on infographic images, with more visually rich elements and answers that can be either extractive from a set of multiple text spans in the image, a multiple choice given in the question, or the result of a discrete operation resulting in a numerical non-extractive answer. In parallel, VisualMRC [22] proposed open-domain questions on webpage screenshots with abstractive answers, which requires to generate longer answers not explicitly found in the text. DuReaderVis [19] is a Chinese dataset for open-domain document visual question answering, where the questions are queries from the Baidu search engine, and the images are screenshots of the webpages retrieved by the search engine results. Although the answers are extractive, 43% of them are non-factual and much longer on average than the ones in previous DocVQA datasets. In addition, each image contains on average a bigger number of text instances. However, due to the big size of the image collection, the task is posed as a 2-stage retrieval and answering tasks, where the methods must retrieve the correct page first, and answer the question in a second step. Similarly, the Document Collection Visual Question Answering (DocCVQA) [24] released a set of 20 questions posed over a whole collection of 14,362 single page document images. However, due to the limited number of questions and the low document variability, it is not possible to do training on this dataset and current approaches need to rely on training on SingleDocVQA. Finally, TAT-DQA [31] contains extractive and abstractive questions on modern financial reports. Despite that the documents might be multi-page, only 306 documents have actually more than one page, with a maximum of 3 pages.

Instead, our proposed MP-DocVQA dataset is much bigger and diverse with 46,176 questions posed over 5,928 multi-page documents with its corresponding 47,952 page images, which provides enough data for training and evaluating new methods on the new multi-page setting.

Methods: Since the release of the SingleDocVQA dataset, several methods have tackled this task from different perspectives. From NLP, Devlin *et al.* proposed BertQA [16] which consists of a BERT [7] architecture followed by a classification head that predicts the start and end indices of the answer span from the given context. While many models have extended BERT obtaining better results [8,11,13,21] by changing key hyperparameters during training or proposing new pre-training tasks, T5 [20] has become the backbone of many state-of-the-art methods [2,14,18] on different NLP and multimodal tasks. T5 relies on the original Transformer [25] by performing minimal modifications on the architecture, but pre-training on the novel de-noising task on a vast amount of data.

On the other hand, and specifically designed for document tasks, LayoutLM [28] extended BERT by decoupling the position embedding into 2 dimensions using the token bounding box from the OCR and fusing visual and textual features during the downstream task. Alternatively, LayoutLMv2 [29] and TILT [18], included visual information into a multimodal transformer and introduced a learnable bias into the self-attention scores to explicitly model relative position. In addition, TILT used a decoder to dynamically generate the answer instead of extracting it from the context. LayoutLMv3 [9] extended its previous version by using visual patch embeddings instead of leveraging a CNN backbone and pre-training with 3 different objectives to align text, layout position and image context. In contrast, while all the previous methods utilize the text recognized with an off-the-shelf OCR, Donut [10] and Dessert [6] are end-to-end encoder-decoder methods where the input is the document image along with the question, and they implicitly learn to read as well as understand the semantics and layout of the images.

However, the limited input sequence length of these methods make them unfeasible for tasks involving long documents such as the ones in MP-DocVQA. Different methods [1,5,30] have been proposed in the NLP domain to improve the modeling of long sequences without increasing the model complexity. Longformer [1] replaces the common self-attention used in transformers where each input attends to every other input by a combination of global and local attention. The global attention is used on the question tokens, which attend and are attended by all the rest of the question and context tokens, while a sliding window guides the local attention over the context tokens to attend the other locally close context tokens. While the standard self-attention has a complexity of $O(n^2)$, the new combina-

tion of global and local attention turns the complexity of the model into $O(n)$. Following this approach, Big Bird [30] also includes attention on randomly selected tokens that will attend and be attended by all the rest of the tokens in the sequence, which provides a better global representation while adding a marginal increase of the complexity in the attention pattern.

3. MP-DocVQA Dataset

The Multi-Page DocVQA (MP-DocVQA) dataset comprises 46K questions posed over 48K images of scanned pages that belong to 6K industry documents. The page images contain a rich amount of different layouts including forms, tables, lists, diagrams and pictures among others as well as text in handwritten, typewritten and printed fonts.

3.1. Dataset creation

Documents naturally follow a hierarchical structure where content is structured into blocks (sections, paragraphs, diagrams, tables) that convey different pieces of information. The information necessary to respond to a question more often than not lies in one relevant block, and is not spread over the whole document. This intuition was confirmed during our annotation process in this multi-page setting. The information required to answer the questions defined by the annotators was located in a specific place in the document. On the contrary, when we forced the annotators to use different pages as a source to answer the question, those become very unnatural and did not capture the essence of questions that we can find in the real world.

Consequently, we decided to use the SingleDocVQA [16] dataset, which already has very realistic questions defined on single pages. To create the new MP-DocVQA dataset, we took every image-question pair from SingleDocVQA [16] and added to every image the previous and posterior pages of the document downloaded from the original source UCSF-IDL³. As we show in Fig. 2a most of documents in the dataset have between 1 and 20 pages, followed by a long tail of documents with up to 793 pages. We focused on the most common scenario and limited the number of pages in the dataset to 20. For longer documents, we randomly selected a set of 20 pages that included the page where the answer is found

Next, we had to analyze and filter the questions since we observed that some of the questions in the SingleDocVQA dataset became ambiguous when posed in a multi-page setup (e.g. asking for the page number of the document). Consequently, we performed an analysis detailed in Appendix A to identify a set of key-words, such as ‘document’, that when included in the text of the question, can lead to ambiguous answers in a multi-page setting, as they origi-

³<https://www.industrydocuments.ucsf.edu/>

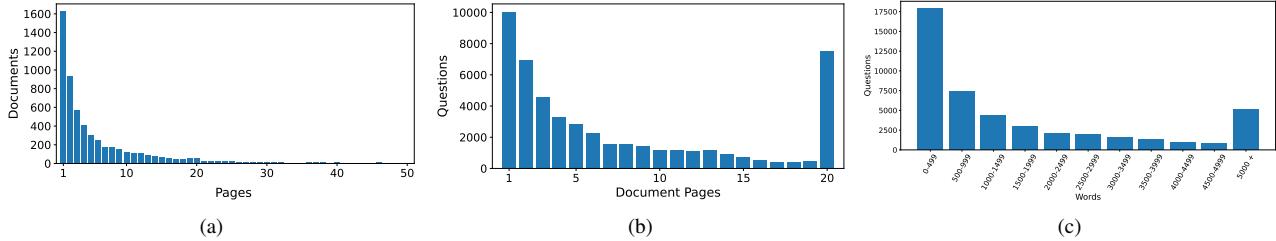


Figure 2. **MP-DocVQA statistics.** (a): Distribution of the document length in term of pages of the documents included in MP-DocVQA before applying the limit of 20 pages. (b): Distribution of the document length in term of pages along the posed questions in the dataset. (c): Number of recognized OCR words per question.

nally referred to a specific page and not to the whole multi-page document.

After removing ambiguous questions, the final dataset comprises 46,176 questions posed over 47,952 page images from 5,928 documents. Notice that the dataset also includes documents with a single page when this is the case. Nevertheless, as we show in Fig. 2b, the questions posed over multi-page documents represent the 85.95% of the questions in the dataset.

Finally, we split the dataset into train, validation and test sets keeping the same distribution as in SingleDocVQA. However, following this distribution some pages would appear in more than one split as they originate from the same document. To prevent this, we trim the number of pages used as context for such specific cases to ensure that no documents are repeated between training and validation/test splits. In Fig. 2b we show the number of questions according to the final document length.

To facilitate research and fair comparison between different methods on this dataset, along with the images and questions we also provide the OCR annotations extracted with Amazon Textract⁴ for all the 47,952 document images (including page images beyond the 20 page limit to not limit future research on longer documents).

3.2. Dataset statistics

As we show in Tab. 1, given that MP-DocVQA is an extension of SingleDocVQA, the average question and answer lengths are very similar to this dataset in contrast to the long answers that can be found in the open-domain datasets VisualMRC and DuReader_{Vis}. On the contrary, the main difference lies in the number of OCR tokens per document, which is even superior to the Chinese DuReader_{Vis}. In addition, MP-DocVQA adopts the multi-page concept, which means that not all documents have the same number of pages (Fig. 2b), but also that each page of the document may contain a different content distribution, with varied text density, different layout and visual elements that raise unique challenges. Moreover, as we show in Figs. 2b

and 2c the variability between documents is high, with documents comprising between 1 and 20 pages, and between 1 and 42,313 recognized OCR words.

4. Hi-VT5

Although documents contain dense information, not all of them is necessary to answer a given question. Following this idea, we propose the Hierarchical Visual T5 (Hi-VT5), a hierarchical encoder-decoder multimodal transformer where given a question, the encoder extracts the most relevant information from each page conditioned to the question and then, the decoder generates the answer from the summarized relevant information extracted from the encoder. Figure 3 shows an overview of the model. We can see that each page is independently processed by the encoder taking as input the sequence of OCR tokens (encoding both text semantics and layout features), a set of patch-based visual features and the encoded question tokens. In addition, a number of learnable [PAGE] tokens are introduced to embed at the output of the encoder the summary of every page. These [PAGE] tokens are concatenated and passed through the decoder to get the final answer. Moreover, in parallel to the answer generation, the answer page identification module predicts the page index where the information to answer the question is found, which can be used as a kind of explainability measure. We utilize the T5 architecture as the backbone for our method since the enormous amount of data and their novel de-noising task utilized during pretraining makes it an excellent candidate for the model initialization. In this section, we first describe each module, then how they are integrated and finally, the training process followed.

Textual representation: Following recent literature on document understanding [9, 18] which demonstrates the importance of layout information when working with Transformers, we utilize a spatial embedding to better align the layout information with the semantic representation. Formally, given an OCR token O_i , we define the associated word bounding box as $(x_0^i, y_0^i, x_1^i, y_1^i)$. Following [2], to embed bounding box information, we use a lookup table

⁴<https://aws.amazon.com/textract/>

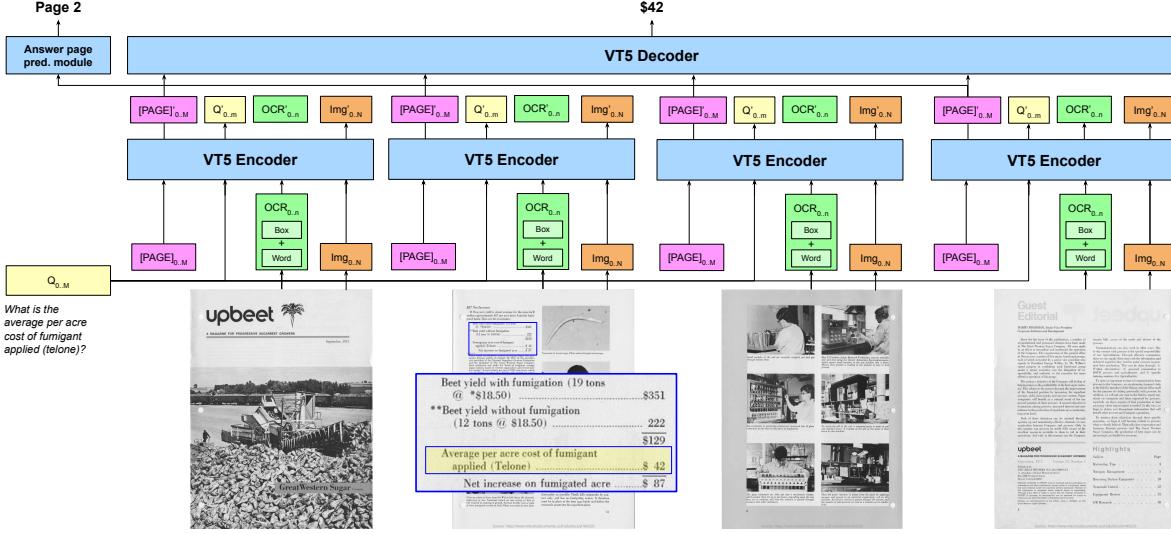


Figure 3. **Architecture of Hi-VT5 model.** The architecture is based on T5 with 2D layout features. Each page passes through the encoder to represent in the contextualized $[\text{PAGE}]'$ tokens the most relevant information necessary to answer the posed question. Then, the $[\text{PAGE}]'$ tokens of all pages are concatenated to provide the decoder with a holistic representation of the document at the time of generating the answer. In addition, a classification layer in the page answer page identification module outputs the page where the answer to the question is found, providing the model with an explainability measure of the answers which allows, among others, to understand if the answer has been inferred from the actual input data, or from a prior learned bias.

for continuous encoding of one-hot vectors, and sum up all the spatial and semantic representations together:

$$\mathcal{E}_i = E_O(O_i) + E_x(x_0^i) + E_y(y_0^i) + E_x(x_1^i) + E_y(y_1^i) \quad (1)$$

where \mathcal{E}_i is the encoded representation for the OCR token O_i , and E_O , E_x and E_y are the learnable look-up tables.

Visual representation: We leverage the Document Image Transformer (DiT) [12] pretrained on Document Intelligence tasks to represent the page image as a set of patch embeddings. Formally, given an image I with dimension $H \times W \times C$, is reshaped into N 2D patches of size $P^2 \times C$, where (H, W) is the height and width, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the final number of patches. We map the flattened patches to D dimensional space, feed them to DiT, pass the output sequence to a trainable linear projection layer and then feed it to the transformer encoder. We denote the final visual output as $V = \{v_0, \dots, v_N\}$.

Hi-VT5 hierarchical paradigm: Inspired by the BERT [7] $[\text{CLS}]$ token, which is used to represent the encoded sentence, we use a set of M learnable $[\text{PAGE}]$ tokens to represent the page information required to answer the given question. Hence, we input the information from the different modalities along with the question and the learnable tokens to the encoder to represent in the $[\text{PAGE}]$ tokens the most relevant information of the page conditioned by the question. More formally, for each page $p_j \in P = \{p_0, \dots, p_K\}$, let $V_j = \{v_0, \dots, v_N\}$ be the patch visual features, $Q = \{q_0, \dots, q_m\}$ the tokenized question, $O_j = \{o_1, \dots, o_n\}$ the page OCR tokens and $K_j = \{k_0, \dots, k_M\}$ the learnable $[\text{PAGE}]$ tokens. Then,

we embed the OCR tokens and question using Eq. (1) to obtain the OCR \mathcal{E}_j^o and question \mathcal{E}^q encoded features. And concatenate all the inputs $[K_j; V_j; \mathcal{E}^q; \mathcal{E}_j^o]$ to feed to the transformer encoder. Finally, all the contextualized K' output tokens of all pages are concatenated to create a holistic representation of the document $D = [K'_0; \dots; K'_K]$, which is sent to the decoder that will generate the answer, and to the answer page prediction module.

Answer page identification module: Following the trend to look for interpretability of the answers in VQA [26], in parallel to the the answer generation in the decoder, the contextualized $[\text{PAGE}]$ tokens D are fed to a classification layer that outputs the index of the page where the answer is found.

Pre-training strategy: Since T5 was trained without layout information, inspired by [2] we propose a hierarchical layout-aware pretraining task to align the layout and semantic textual representations, while providing the $[\text{PAGE}]$ tokens with the ability to attend to the other tokens. Similar to the standard de-noising task, the layout-aware de-noising task masks a span of tokens and forces the model to predict the masked tokens. Unlike the normal de-noising task, the encoder has access to the rough location of the masked tokens, which encourages the model to fully utilize the layout information when performing this task. In addition, the masked tokens must be generated from the contextualized K' $[\text{PAGE}]$ tokens created by the encoder, which forces the model to embed the tokens with relevant information regarding the proposed task.

Training strategy: Even though Hi-VT5 keeps the same

Model	Size	Parameters	Max Seq. Length	Setup	Accuracy	ANLS	Ans. Page Accuracy
BERT [7]	Large	334M	512	Oracle	39.77	0.5904	100.00
				Max Conf.	34.78	0.5347	71.24
				Concat	27.41	0.4183	51.61
Longformer [1]	Base	148M	4096	Oracle	52.48	0.6177	100.00
				Max Conf.	45.87	0.5506	70.37
				Concat	43.91	0.5287	71.17
Big Bird [30]	Base	131M	4096	Oracle	55.31	0.6450	100.00
				Max Conf.	49.57	0.5854	72.27
				Concat	41.06	0.4929	67.54
LayoutLMv3 [9]	Base	125M	512	Oracle	58.81	0.6729	100.00
				Max Conf.	42.70	0.5513	74.02
				Concat	38.47	0.4538	51.94
T5 [20]	Base	223M	512	Oracle	59.00	0.6814	100.00
				Max Conf.	32.68	0.4028	46.05
				Concat	41.80	0.5050	—
Hi-VT5 (Ours)	Base	316M	20480	Oracle	50.01	0.6572	100.00
				Multipage	48.28	0.6201	79.23

Table 2. **Baselines and proposed method Hi-VT5 results on MP-DocVQA dataset.** Baselines are evaluated on three different setups: oracle, concat and ‘max conf.’. The proposed method is evaluated only on the oracle setup and the realistic multi-page setting. We highlight in bold the best results for the oracle and any multi-page (oracle and ‘max conf.’) setup.

model complexity as the sum of their independent components ($T5_{BASE}$ (223M) + DiT_{BASE} (85M)) and despite being capable to accept input sequences of up to 20480 tokens, the amount of gradients computed at training time scales linearly with the number of pages since each page is passed separately through the encoder and the gradients are stored in memory. Consequently, it is similar to have a batch size P times bigger in the encoder compared to a single page setting. While this could be tackled by parallelizing the gradients corresponding to a set of pages into different GPUs, we offer an alternative strategy using limited resources. We train the model on shortened versions of the documents with only two pages: the page where the answer is found and the previous or posterior page. Even though this drops the overall performance of the model, as we show in Appendix C, training with only 2 pages is enough to learn the hierarchical representation of the model achieving results close to the ones using the whole document, and offers a good trade-off in terms of memory requirements. However, after the training phase the decoder and the answer page identification module can’t deal with the full version of the documents of up to 20 pages. For this reason, we perform a final fine-tuning phase using the full-length documents and freezing the encoder weights.

5. Experiments

To evaluate the performance of the methods, we use the standard evaluation metrics in DocVQA, accuracy and Average Normalized Levenshtein Similarity (ANLS) [4]. To assess the page identification we use accuracy.

5.1. Baselines

As Multi-Page DocVQA is a new task, we adapt several state-of-the-art methods as baselines to analyze their limitations in the multi-page setup and compare their performance against our proposed method. We choose BERT [7] because it was the first question-answering method based on transformers, and it shows the performance of such a simple baseline. Longformer [1] and Big Bird [30] because they are specially designed to deal with long sequences, which might be beneficial for the multi-page setting. In the case of Big Bird it can work following two different strategies. The former, Internal Transformer Construction (ITC) only sets the global attention over one single token, while the Extended Transformer Construction (ETC) sets the global attention over a set of tokens. Although the latter strategy is the desired setup for question-answering tasks by setting all the question tokens with global attention, the current released code only supports the ITC strategy and hence, we limit our experiments to this attention strategy. We also use LayoutLMv3 [9] because it is the current public state-of-the-art method on the SingleDocVQA task and uses explicit visual features by representing the document in image patches. Finally, T5 [20] because it is the only generative baseline and the backbone of our proposed method.

However, all these methods are not directly applicable to a multi-page scenario. Consequently, we define three different setups to allow them to be evaluated on this task. In the ‘oracle’ setup, only the page that contains the answer is given as input to the transformer model. Thus, this setup aims at mimicking the Single page DocVQA task. It

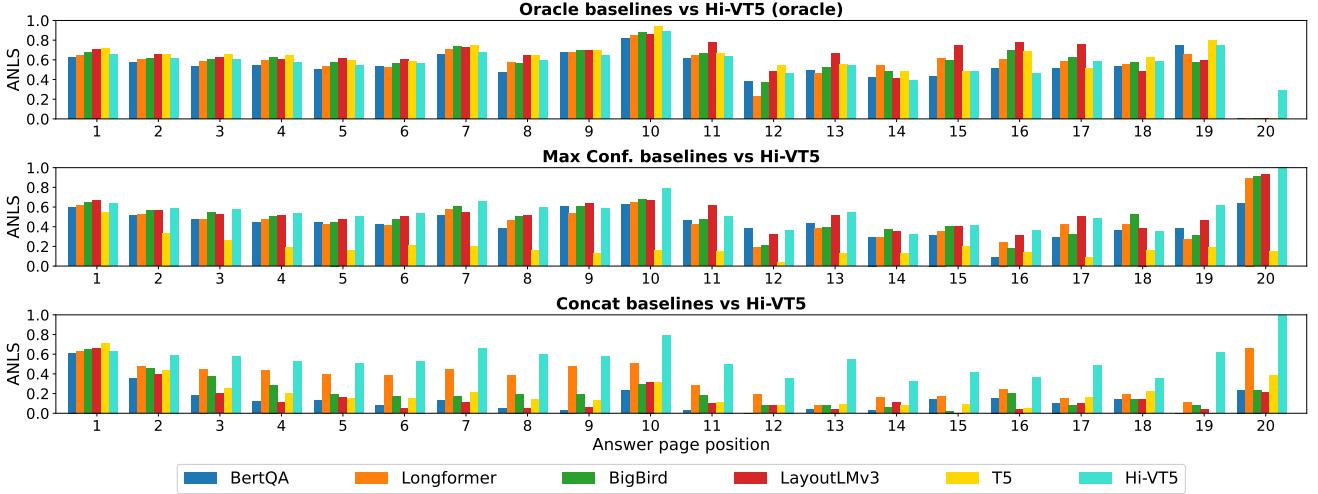


Figure 4. Methods ANLS by answer page position. The figure shows the answering performance of the different baselines and Hi-VT5 in the oracle setup (top), and the baselines in the ‘max conf.’ (middle) and concat (bottom) setup against Hi-VT5 using its answer page identification module. Notice that the breakdown of the scores is NOT performed on the number of the document pages, but in which page the answer is found.

shows the raw answering capabilities of each model regardless of the size of the input sequences they can accept. So, it should be seen as a theoretical maximum performance, assuming that the method has correctly identified the page where the information is found. In the ‘concat’ setup, the context input to the transformer model is the concatenation of the contexts of all the pages of the document. This can be considered the most realistic scenario where the whole document is given as a single input. It is expected that the large amount of input data becomes challenging for the baselines. The page corresponding to the predicted start index is used as the predicted page, except for T5, since being a generative method it does not predict the start index. Finally, max conf is the third setup, which is inspired in the strategy that the best performing methods in the DocCVQA challenge [23] use to tackle the big collection of documents. In this case, each page is processed separately by the model, providing an answer for every page along with a confidence score in the form of logits. Then, the answer with the highest confidence is selected as the final answer with the corresponding page as the predicted answer page.

For BERT, Longformer, Big Bird and T5 baselines we create the context following the standard practice of concatenating the OCR words in the image following the reading (top-left to bottom-right) order. For all the methods, we use the Huggingface [27] implementation and pre-trained weights from the most similar task available. We describe the specific initialization weights and training hyperparameters in Appendix D.

5.2. Baseline results

As we show in Tab. 2, the method with the best answering performance in the oracle setup (i.e. when the an-

swer page is provided) is T5, followed by LayoutLMv3, Big Bird, Longformer and BERT. This result is expected since this setup is equivalent to the single page document setting, where T5 has already demonstrated its superior results. In contrast, in the ‘max conf.’ setup, when the logits of the model are used as a confidence score to rank the answers generated for each page, T5 performs the worst because the softmax layer used across the vocabulary turns the logits unusable as a confidence to rank the answers. Finally, in the concat setup, when the context of all pages are concatenated Longformer outperforms the rest, showing its capability to deal with long sequences as seen in Fig. 4, which shows that the performance gap increases as long as the answer page is placed at the end of the document. The second best performing method in this setting is T5, which might seem surprising due to its reduced sequence length. However, looking at Fig. 4 it is possible to see that is good on questions whose answers can fit into the input sequence, while it is not capable to answer the rest. In contrast, Big Bird is capable to answer questions that require long sequences since its maximum input length is 4096 as Longformer. Nevertheless, it performs worse due to the ITC strategy Big Bird is using, which do not set global attention to all question tokens and consequently, as long as the question and the answer tokens become more distant, it is more difficult to model the attention between the required information to answer the question.

5.3. Hi-VT5 results

In our experiments we fixed the number of [PAGE] tokens to $M = 10$, through experimental validation explained in detail in Appendix B. We observed no significant improvements beyond this number. We pretrain Hi-VT5 on

hierarchical aware de-noising task on a subset of 200,000 pages of OCR-IDL [3] for one epoch. Then, we Train on MP-DocVQA for 10 epochs with the 2-page shortened version of the documents and finally, perform the fine-tuning of the decoder and answer page identification module with the full length version of the documents for 1 epoch. During training and fine-tuning all layers of the DiT visual encoder are frozen except a last fully connected projection layer.

Hi-VT5 outperforms all the other methods both on answering and page identification in the concat and ‘*max conf.*’ setups, which are the most realistic scenarios. In addition, when looking closer at the ANLS per answer page position (see Fig. 4), the performance gap becomes more significant when the answers are located at the end of the document, even compared with Longformer, which is specifically designed for long input sequences. In contrast, Hi-VT5 shows a performance drop in the ‘*oracle*’ setup compared to the original T5. This is because it must infer the answer from a compact summarized representation of the page, while T5 has access to the whole page representation. This shows that the page representation obtained by the encoder has still margin for improvement.

Finally, identifying the page where the answer is found at the same time as answering the question allows to better interpret the method’s results. In Tab. 2 we can see that Hi-VT5 obtains a better answer page identification performance than all the other baseline methods. In addition, in Fig. 5 we show that it is capable to predict the correct page even when it cannot provide the correct answer. Interestingly, it answers correctly some questions for which the predicted page is wrong, which means that the answer has been inferred from a prior learned bias instead of the actual input data. We provide more details by analyzing the attention of Hi-VT5 in Appendix F.

		Answer		
		ansls = 1	0.5 ≤ anls < 1	ansls < 0.5
Answer page prediction	Correct	1613	1286	1098
	Wrong	242	284	496

Figure 5. Matrix showing the Hi-VT5 correct and wrong answered questions depending on the answer page prediction module result.

6. Ablation studies

To validate the effectiveness of each feature proposed in Hi-VT5, we perform an ablation study and show re-

sults in Tab. 3. Without the answer page prediction module the model performs slightly worse on the answering task, showing that both tasks are complementary and the correct page prediction helps to answer the question. The most significant boost comes from the hierarchical de-noising pre-training task, since it allows the [PAGE] tokens to learn better how to represent the content of the document. The last fine-tuning phase where the decoder and the answer page prediction module are adapted to the 20 pages maximum length of the MP-DocVQA documents, is specially important for the answer page prediction module because the classification layer predicts only page indexes seen during training and hence, without finetuning it can only predict the first or the second page of the documents as the answer page. Finally, when removing the visual features the final scores are slightly worse, which has also been shown in other works in the literature [2, 9, 18], the most relevant information is conveyed within the text and its position, while explicit visual features are not specially useful for grayscale documents.

Method	Accuracy	ANLS	Ans. Page Acc.
Hi-VT5	48.28	0.6201	79.23
-2D-pos	46.12	0.5891	78.21
-Vis. Feat.	46.82	0.5999	78.22
-APPM	47.78	0.6130	00.00
-Pretrain	42.10	0.5864	81.47
-Fine-tune	42.86	0.6263	55.74

Table 3. **Hi-VT5 ablation studies.** We study the effect of removing different components independently from Hi-VT5 namely the 2D position embedding (2D-pos), visual features (Vis. Feat.), the answer page prediction module (APPM), the pretraining (Pretrain) and the last fine-tuning (Fine-tune) phase of the decoder and answer page prediction module.

7. Conclusions

In this work, we propose the task of Visual Question Answering on multi-page documents and make public the MP-DocVQA dataset. To show the challenges the task poses to current DocVQA methods, we convey an analysis of state-of-the-art methods showing that even the ones designed to accept long sequences are not capable to answer questions posed on the final pages of a document. In order to address these limitations, we propose the new method Hi-VT5 that, without increasing the model complexity, can accept sequences up to 20,480 tokens and answer the questions regardless of the page in which the answer is placed. Finally, we show the effectiveness of each of the components in the method, and perform an analysis of the results showing how the answer page prediction module can help to identify answers that might be inferred from prior learned bias instead of the actual input data.

Acknowledgements

This work has been supported by the UAB PIF scholarship B18P0070, the Consolidated Research Group 2017-SGR-1783 from the Research and University Department of the Catalan Government, and the project PID2020-116298GB-I00, from the Spanish Ministry of Science and Innovation.

References

- [1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. [3](#), [6](#)
- [2] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikanth Apalaraju, and R Manmatha. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16548–16558, 2022. [3](#), [4](#), [5](#), [8](#)
- [3] Ali Furkan Biten, Ruben Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. Ocr-idl: Ocr annotations for industry document library dataset. *arXiv preprint arXiv:2202.12985*, 2022. [8](#)
- [4] Ali Furkan Biten, Rubén Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4291–4301, 2019. [6](#)
- [5] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019. [3](#)
- [6] Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. End-to-end document recognition and understanding with dessurt. *arXiv e-prints*, pages arXiv–2203, 2022. [3](#)
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. [2](#), [3](#), [5](#), [6](#), [11](#)
- [8] Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. Lambert: layout-aware language modeling for information extraction. In *International Conference on Document Analysis and Recognition*, pages 532–547. Springer, 2021. [3](#)
- [9] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022. [1](#), [3](#), [4](#), [6](#), [8](#)
- [10] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. [3](#)
- [11] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. [3](#)
- [12] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. *arXiv preprint arXiv:2203.02378*, 2022. [5](#)
- [13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. [3](#)
- [14] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. [3](#)
- [15] Minesh Mathew, Viraj Bagal, Rubén Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. [1](#), [2](#)
- [16] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021. [1](#), [2](#), [3](#), [11](#)
- [17] Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Document visual question answering challenge 2020. *arXiv preprint arXiv:2008.08899*, 2020. [1](#), [2](#)
- [18] Rafal Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. In *International Conference on Document Analysis and Recognition*, pages 732–747. Springer, 2021. [1](#), [3](#), [4](#), [8](#)
- [19] Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaozhao She, Hua Wu, Haifeng Wang, and Ting Liu. Dureaderivs: A: A chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1338–1351, 2022. [2](#)
- [20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. [2](#), [3](#), [6](#)
- [21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. [3](#)
- [22] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021. [2](#)

- [23] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *International Conference on Document Analysis and Recognition*, pages 778–792. Springer, 2021. [2](#), [7](#)
- [24] Rubèn Tito, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2021 competition on document visual question answering. In *International Conference on Document Analysis and Recognition*, pages 635–649. Springer, 2021. [2](#)
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [1](#), [3](#)
- [26] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. [2](#), [5](#)
- [27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. [7](#)
- [28] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020. [1](#), [3](#)
- [29] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, 2021. [1](#), [3](#)
- [30] 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. [3](#), [6](#)
- [31] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022. [2](#)

A. MP-DocVQA construction process

As described in Sec. 3.1, the source data of the MP-DocVQA dataset is the SingleDocVQA [16] dataset. The first row of Tab. 4 shows the number of documents, pages and questions in this dataset. The first step to create the MP-DocVQA dataset was to download and append to the existing documents their previous and posterior pages, increasing the number of page images from 12,767 to 64,057, as shown in the second row of Tab. 4.

	Documents	Pages	Questions
SingleDocVQA	6,071	12,767	50,000
MP-DocVQA (full)	6,071	64,057	50,000
MP-DocVQA (filtered)	5,928	60,884	46,176
MP-DocVQA (20 page limit)	5,928	47,952	46,176
MP-DocVQA (multi-page)	3,824	39,688	39,688

Table 4. Statistics of the MP-DocVQA during its construction process.

However, not all questions are suited to be asked on multi-page documents. Therefore, we performed an analysis based on manually selected key-words that appear in the questions, searching for those questions whose answer becomes ambiguous when they are posed over a multi-page document. Some of the selected key-words are shown in table Tab. 6, along with some examples of potentially ambiguous questions containing those key-words. The most clear example is with the word 'document'. When looking at each document page separately, we can observe that many times they start with a big text on the top that can be considered as the title, which is actually the answer in the single page DocVQA scenario when the question asks about the title of the document. However, this pattern is repeated in every page of the document, making the question impossible to answer when multiple pages are taken into account. Moreover, even if there is only one page with a title, the answer can still be considered wrong, since the title of the document is always found in the first page like in the example in Fig. 1. On the other hand, when we analyzed more closely other potentially ambiguous selected key-words such as 'image', 'appears' or 'graphic' we found out that the answers were not always ambiguous and also the amount of questions with those words was negligible compared to the entire dataset. Thus, we decided to keep those questions in our dataset. Finally, we found that the key-word 'title' was mostly ambiguous only when it was written along with the word 'document'. Hence, we decided to remove only the questions with the word 'document' in it, while keeping all the rest. This filtered version, which is represented in the third row of Tab. 4 is the dataset version that was released and used in the experiments.

Nevertheless, it is important to notice that not all the

questions in MP-DocVQA are posed over multi-page documents. We keep the documents with a single page because they are also a possible case in a real life scenario. However, as showed in the fourth row of Tab. 4, the questions posed over multiple pages represent the 85.95% of all the questions in the dataset.

B. Number of [PAGE] tokens

Hi-VT5 embeds the most relevant information from each page conditioned by a question into M [PAGE] tokens. However, we hypothesize that contrary to BERT [7], which represents a sentence with a single [CLS] token, Hi-VT5 will require more than one token to represent a whole page, since it conveys more information. Consequently, we perform an experimental study to find the optimum number of [PAGE] tokens to use. We start by defining the maximum number of tokens M that can be used, which is limited by the decoder input sequence length S , and the number of pages P that must be processed. Formally,

$$M = \text{int} \left(\frac{S}{P} \right) \quad (2)$$

We can set M as an hyperparameter to select depending on the number of pages we need to process, where in the extreme cases we can represent a single page with 1024 [PAGE] tokens, or a 1024 page document with a single token for each page.

Constraining to the 20 pages documents scenario of MP-DocVQA, the maximum possible number of tokens M would be 51. We performed a set of experiments with different [PAGE] tokens to find the optimal value. As we show in Tab. 5, the model is able to answer correctly some questions even when using only one or two tokens. However, the performance increases significantly when more tokens are used. Nevertheless, the model does not benefit from using more than 10 tokens, since it performs similarly either with 10 or 25 tokens. Moreover, the performance decreases when using more. This can be explained because the information extracted from each page can be fully represented by 10 tokens, while using more, not only does not provide any benefit, but also makes the training process harder.

[PAGE] Tokens	Accuracy	ANLS	Ans. Page Accuracy
1	36.41	0.4876	79.87
2	37.94	0.5282	79.88
5	39.31	0.5622	80.77
10	42.10	0.5864	81.47
25	42.16	0.5896	81.35
50	30.63	0.5768	59.18

Table 5. Results of Hi-VT5 with different [PAGE] tokens.

Document (3824)	Image (72)	Appears (15)	Title (1836)
What is the subject of the document /letter?	What is the number of calories written in the image ?	Whose name appears on top of the schedule?	What is the title of this document?
What is the title of the document ?	What does the image say?	What is the name of registered agent as it appears of record?	What is the title of the table?
What date is the meeting scheduled to develop the overall structure of the document ?	In the image of the man with a trophy, what is the name of the awards given?	Who appears in the photograph at the top of the document standing alone with Nehru?	Which are prescribed earlier in the treatment of type 2 diabetes under the title of "critical success factors"?
What is the subject of the document ?	What type of product is on the image ?	Which company appears first among the attendees?	What is the title of the diagram?
What 'council' is mentioned in the document ?	In the image of the playing card pack, what is the number on the card of diamonds?	Which is the numerical rating that appears most number of times?	Who prepared the controversial report entitled "Dietary Goals for the United States"?
Which date is mentioned at the end of the ' document '?	What is the name of the company in the image ?	Which is the page number greater than 28, that appears only once?	What is the title of this page?

Table 6. Key-words used to find inadequate questions over multi-page documents. In the title row, following each key-word is showed the number of questions in SingleDocVQA with that word.

C. Document pages during training

As described in Sec. 4, it is not feasible to train with 20 page length documents due to training resource limitations. However, as we show in Tab. 7, even though the model performs significantly worse when trained with a single page, the returns become diminishing when training with more than 2. Thus, as explained in Sec. 4 we decided to use 2 pages in the first stage of training.

Trained pages	Acc	ANLS
1	22.96	0.3860
2	33.37	0.5577
5	34.08	0.5730
10	34.25	0.5792

Table 7. Experiments showing the results when training with different number of document pages and tested with the document original length.

D. Hyperparameters

	BERT	Longformer	BigBird	T5	Hi-VT5 [†]
Model size	large	base	base	base	base
Parameters	334M	148M	131M	223M	316M
Model initial weights	SingleDocVQA	SQuADv1	TrivaQA	C4	C4
Max Seq. Length	512	4096	4096	512	20480
Training Loss	CE	CE	CE	CE	CE
batch size	32	8	8	20	8
lr	5e-5	1e-4	3e-5	2e-4	2e-4
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
scheduler	linear	linear	linear	linear	linear
warmup iterations	1000	1000	1000	1000	1000
training epochs	1	10	10	10	1 - 10 - 1

Table 8. Hyperparameters of the baselines and the proposed method that were used to train and evaluate on MP-DocVQA. [†]: Hi-VT5 refers to all three pre-training, training and fine-tune stages. The only difference is the number of epochs: 1, 10 and 1 respectively. Training loss CE denotes CrossEntropy loss.

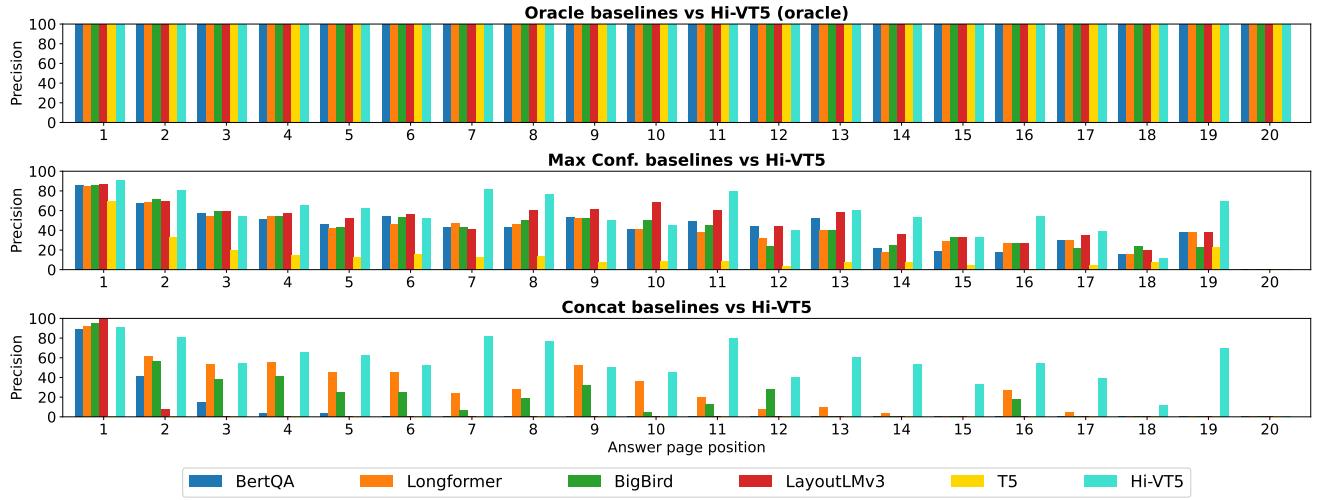


Figure 6. **Accuracy of page identification as a function of answer page position.** The figure shows the page identification accuracy of the different baselines and Hi-VT5 in the oracle setup (top), and the baselines in the ‘max conf.’ (middle) and concat (bottom) setup against Hi-VT5 using the page identification module. Notice that the breakdown of the scores is NOT performed on the number of pages the document, but in which page the answer is found.

E. Page identification accuracy by answer page position

In Fig. 6 we show the answer page identification accuracy of the different baselines and the proposed method, as a function of the page number of the answer. The overall performance follows a similar behavior as the answer scores. Longformer is the baseline that performs the best in the concat setting, and the performance gap between this and the rest of the baselines becomes more significant as the answer page is located in the final pages of the document. However, Hi-VT5 outperforms all the baselines by a big margin.

F. Hi-VT5 attention visualization

To further explore the information that Hi-VT5 embeds into the [PAGE] tokens, we show the attention scores for some examples in MP-DocVQA. The attention of Fig. 7a, corresponds to the first [PAGE] token, which usually performs a global attention over the whole document with a slight emphasis on the question tokens, which provides a holistic representation of the page. Other tokens like in Fig. 7c focuses its attention over the other [PAGE], and question tokens. More importantly, there is always a token that focuses its attention to the provided answer like in Figs. 7b and 7d.

[PAGE]
 - question : What is the date when the approval form was filled ? context :

RJRT/RESEARCH & DEVELOPMENT		DOCUMENTATION APPROVAL FORM	
Log-in No. 88-			
NAME: William M. Coleman, III	TYPE OF DOCUMENTATION: Abstract	TITLE: Sr. R&D Chemist	Research Manuscript
DIVISION: Analytical Research	Oral Presentation		Other
DATE: June 7, 1988	RJRT/R&D R&D or R&D Number & Date		R&D R&D R&D or R&D Number & Date
MAIL ADDRESS: 611-13E/002	Extension: 5177	MAIL ADDRESS: 611-13E/002	Extension: 5177
DOCUMENT TITLE: "Mainstream Particulate Phase Comparison of a Reference Cigarette and a Cigarette That Heats Rather Than Burns Tobacco".			
AUTHOR(S)/CO-AUTHOR(S): W. M. Coleman, III, H. L. Chung, D. S. Moore, E. L. White, B. M. Gordon, M. S. Uhrig, J. A. Giles, J. F. Elder, Jr., M. F. Borgerding, and R. D. Hicks.			
PURPOSE OF DOCUMENT			
PUBLICATION IN:			
PRESENTATION AT: 42nd Tobacco Chemists' Research Conference			
LOCATION: Lexington, Kentucky Date: October 2-5, 1988			
Other pertinent information: Abstract due no later than June 24, 1988.			
Your signatures on this form indicate that, to the best of your knowledge, the information contained in the document to be published/presented is not proprietary and the technical quality reflects positively on the image of RJRT and its R&D Departments.			
Author(s): <u>W.M.Coleman</u>	Date: <u>7 Jun 1988</u>	Manager: <u>JC Long L. Fugay</u>	Date: <u>6-9-88</u>
Director:	Date:		

5136 8959

(a) Global attention over all the text in the page

[PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE]
 - question : What is the total costs for proposed project period ? context :

SECTION II - PRIVILEGED COMMUNICATION HARVARD UNIVERSITY ROCHE ALEX F. 370 56 0985							
BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE DIRECT COSTS ONLY (Omit Costs)							
DESCRIPTION	1ST PERIOD TAKED BUDGET	ADDITIONAL YEARS SUPPORT REQUESTED (Indicate only)	2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	7TH YEAR
PERSONNEL COSTS	17,730	18,971	20,299	21,720			
CONSULTANT COSTS (Include fees, travel, etc.)	****						
EQUIPMENT	****						
SUPPLIES	400	400	400	400			
TRAVEL	DOMESTIC: **** FOREIGN: ****						
PATIENT COSTS	****						
ALTERATIONS AND RENOVATIONS	****						
Other Expenses	1,870	1,870	1,870	1,870			
Total Direct Costs	20,000	21,241	22,569	23,990			
Indirect Costs	6,300	6,741	7,313	7,718			
Total Costs	26,300	27,982	29,882	31,708			
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD (Enter on Page 1, Item 4) →							\$115,872.
REMARKS: Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increase in other categories. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation sheet if needed.)							
1. Valadian: Will select and plan content of the three investigations. Will direct the writing and review the literature. R. Reed: Will plan and supervise the statistical analysis of the three topics. K. Halvorsen: Will carry out statistical analysis, coding and programming (statistician) under the direction of Dr. Robert Reed. Supplies: Pens, pencils, paper, xerocopying. Computer and Computer Time: coding and programming of data collected. Storing of data for projects.							

Rev. 1/73

Page 22 January, 1980

(c) Attention focused over the rest of the [PAGE] and question tokens.

[PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE]
 - question : What is the date when the approval form was filled ? context :

RJRT/RESEARCH & DEVELOPMENT		DOCUMENTATION APPROVAL FORM	
Log-in No. 88-			
NAME: William M. Coleman, III	TYPE OF DOCUMENTATION: Abstract	TITLE: Sr. R&D Chemist	Research Manuscript
DIVISION: Analytical Research	Oral Presentation		Other
DATE: June 7, 1988	RJRT/R&D R&D or R&D Number & Date		R&D R&D R&D or R&D Number & Date
MAIL ADDRESS: 611-13E/002	Extension: 5177	MAIL ADDRESS: 611-13E/002	Extension: 5177
DOCUMENT TITLE: "Mainstream Particulate Phase Comparison of a Reference Cigarette and a Cigarette That Heats Rather Than Burns Tobacco".			
AUTHOR(S)/CO-AUTHOR(S): W. M. Coleman, III, H. L. Chung, D. S. Moore, E. L. White, B. M. Gordon, M. S. Uhrig, J. A. Giles, J. F. Elder, Jr., M. F. Borgerding, and R. D. Hicks.			
PURPOSE OF DOCUMENT			
PUBLICATION IN:			
PRESENTATION AT: 42nd Tobacco Chemists' Research Conference			
LOCATION: Lexington, Kentucky Date: October 2-5, 1988			
Other pertinent information: Abstract due no later than June 24, 1988.			
Your signatures on this form indicate that: to the best of your knowledge, the information contained in the document to be published/presented is not proprietary and the technical quality reflects positively on the image of RJRT and its R&D Departments.			
Author(s): <u>WM Coleman</u>	Date: <u>7 Jun 1988</u>	Manager: <u>JC Long L. Fugay</u>	Date: <u>6-9-88</u>
Director:	Date:		

5136 8959

(b) Attention focused over the OCR tokens corresponding to the answer (7 June, 1988)

[PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE] [PAGE]
 - question : What is the total costs for proposed project period ? context :

SECTION II - PRIVILEGED COMMUNICATION HARVARD UNIVERSITY ROCHE ALEX F. 370 56 0985							
BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE DIRECT COSTS ONLY (Omit Costs)							
DESCRIPTION	1ST PERIOD TAKED BUDGET	ADDITIONAL YEARS SUPPORT REQUESTED (Indicate only)	2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	7TH YEAR
PERSONNEL COSTS	17,730	18,971	20,299	21,720			
CONSULTANT COSTS (Include fees, travel, etc.)	****						
EQUIPMENT	****						
SUPPLIES	400	400	400	400			
TRAVEL	DOMESTIC: **** FOREIGN: ****						
PATIENT COSTS	****						
ALTERATIONS AND RENOVATIONS	****						
Other Expenses	1,870	1,870	1,870	1,870			
Total Direct Costs	20,000	21,241	22,569	23,990			
Indirect Costs	6,300	6,741	7,313	7,718			
Total Costs	26,300	27,982	29,882	31,708			
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD (Enter on Page 1, Item 4) →							\$115,872.
REMARKS: Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increase in other categories. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation sheet if needed.)							
1. Valadian: Will select and plan content of the three investigations. Will direct the writing and review the literature. R. Reed: Will plan and supervise the statistical analysis of the three topics. K. Halvorsen: Will carry out statistical analysis, coding and programming (statistician) under the direction of Dr. Robert Reed. Supplies: Pens, pencils, paper, xerocopying. Computer and Computer Time: coding and programming of data collected. Storing of data for projects.							

Rev. 1/73

Page 22 January, 1980

(d) Attention focused over the OCR tokens corresponding to the answer (\$115,872)

Figure 7. Visualization of the Hi-VT5 attention scores.