DocVLM: Make Your VLM an Efficient Reader
Supplementary Material

A. Additional Implementation Details

DocVLM efficiently leverages OCR data to enhance VLMs' reading capabilities. We extract text and layout information from document images using an OCR system, which is then processed by an OCR encoder as discussed in Sec. 3. Specifically, we utilize the encoder component of DocFormerV2 [7], omitting the visual branch of this encoder, as detailed in the main paper. The encoder is initialized with pretrained weights from DocFormerV2, which was pretrained on the Industry Document Library (IDL) dataset [13]. For details on this pretraining process, refer to [7].

A.1. Optimization and Hyperparameters Details

As discussed in Sec. 3.2, our training process comprises two stages: 1) OCR-LLM alignment and 2) Vision alignment. For both stages, we utilize AdamW optimization algorithm with a cosine learning scheduling and 1000 warmup steps. For the OCR-LLM alignment stage with our learned queries component, we trained for 140K steps. We used learning rates of 10^{-4} for the projection layer and query tokens, and $5 \cdot 10^{-5}$ for the OCR encoder. To preserve the pretrained weights of the OCR encoder while optimizing the randomly initialized components, we initially froze the encoder for the first 10K steps. In experiments without our learned queries component (*i.e.*, without OCR compression), we adjusted the process, training the OCR encoder and projection layer for 100K steps using the same learning rates. In the Vision alignment stage, we trained all components for an additional 100K steps with a learning rate of $5 \cdot 10^{-6}$. Unlike the previous stage, this phase included visual features as input to the LLM, allowing the model to align the OCR modality with the visual one.

B. Datasets

B.1. Training Datasets

Tab. 5 details all the datasets used to fine-tune DocVLM. For the OCR-LLM alignment stage, our dataset selection focuses on text-related tasks, including approximately 990K queries, including document VQA datasets (DocVQA [40], InfoVQA [41], ChartQA [39], TAT-DQA [62]), scene text VQA datasets (TextVQA [48], ST-VQA [11], OCR-VQA [42]), and a captioning dataset (TextCaps [47]). The vision alignment stage incorporates additional visual-centric datasets: COCO Caption [15] and VQA-V2 [24], bringing the total training set to approximately 2M queries.

Task	Dataset	Subsplit	Visual Only	# Queries
Document VQA	DocVQA [40]	train	×	39463
	InfoVQA [41]	train	×	46883
	ChartQA [39]	train (H)	×	7398
	TAT-DQA [62]	train	×	13246
Scene Text VQA	TextVQA [48]	train	×	34602
	ST-VQA [11]	train	×	26308
	OCR-VQA [42]	train	×	800000
Captioning	TextCaps [47]	train	×	21953
	COCO Caption [15]	train	✓	566747
General VQA	VQA-V2 [24]	train	✓	443757
Total Examples				**2000357**

Table 5. **Training Datasets for DocVLM Fine-tuning.** Datasets used for fine-tuning DocVLM, categorized by task type. The 'Visual Only' column indicates datasets that are not text-centric. The total number of queries across all datasets is shown at the bottom.

Task	Dataset	Subsplit	Metric	Zero-Shot	# Queries
Document VQA	DocVQA [40]	Test	ANLS	×	5188
	InfoVQA [41]	Test	ANLS	×	6573
Scene Text VQA	TextVQA [48]	Val	VQAScore	×	5000
	ST-VQA [11]	Test	ANLS	×	4163
Captioning	TextCaps [47]	Val	CIDEr	×	3166
Multipage VQA	MP-DocVQA [50]	Test	ANLS	×	5019
	DUDE [51]	Test	ANLS	✓	11402
Total Examples					**40511**

Table 6. **Evaluation Datasets for DocVLM.** Datasets used for evaluating DocVLM, categorized by task type. The table includes the dataset split used, evaluation metric, zero-shot status, and number of queries for each dataset.

B.2. Evaluation Datasets

Tab. 6 details all the datasets used to evaluate DocVLM's performance across a diverse range of document understanding tasks, including document VQA, scene text VQA, captioning, and multipage document understanding. While our training focused on single-page documents, we extended our evaluation to include multipage datasets: MP-DocVQA [50] and DUDE [51]. It is important to note that although both multipage datasets were not included in our training set, we only consider DUDE as a true zero-shot evaluation, as MP-DocVQA is an extension of DocVQA, which was included in our training data.

C. Qualitative Results

Figures 6 and 7 showcase DocVLM's enhanced document understanding capabilities through representative examples. Figure 6 focuses on document images from the DocVQA [40] test set, while Figure 7 presents infographic images from the InfoVQA test set [41]. We present results for LLaVA-OneVision with a 1.5K visual token limitation, InternVL2 with 256 and 1280 visual token limitations, and Qwen2VL with 256 and 512 visual token limitations. As can be seen, the baselines' errors occur in scenarios that demand superior reading comprehension capabilities. Notably, by only utilizing 64 OCR compressed tokens, DocVLM effectively corrects errors and provides the correct responses. This improvement is consistent across different VLM architectures and visual token limitations, highlighting the efficiency and versatility of our approach.

D. Ablation Study on Visual Features

In this section, we explore how visual features contribute to DocVLM's performance by first evaluating DocVLM *without visual input* and then assessing the impact of adding visual features.

DocVLM's OCR Encodings Without Visual Input. We evaluate DocVLM based on Qwen2VL after the OCR-LLM Alignment stage, using only OCR encodings as input to the LLM, without visual tokens. This approach allows us to assess how well the encodings capture OCR data and their sufficiency for document question answering tasks. Our architecture consists of inputting DocVLM's encodings or compressed encodings to the Qwen2 LLM along with the query prompt. Tab. 7 presents our results on DocVQA [40] and InfoVQA [41] test sets compared to baselines that also rely solely on OCR information [52, 54, 55]. We can see that DocVLM's OCR encodings effectively capture OCR information, yielding the best results in the comparison. Remarkably, using only 64 learned queries (compressed encodings) achieves competitive performance, significantly surpassing the OCR words baseline, despite being much shorter (64 compared to 1K tokens).

Contribution of Visual Features. In Tab. 8, we compare the results from the previous text-only evaluation to those obtained when adding 256 visual tokens to the input of the same model checkpoint. The results demonstrate that incorporating visual information improves performance across both datasets, with a particularly notable enhancement when using compressed OCR encodings. This comparison highlights the complementary nature of textual and visual information in DocVLM's architecture.

Method	**LLM OCR Input**	DocVQA	InfoVQA
Alpaca	Latin Prompt	42.0	–
ChatGPT-3.5	Latin Prompt	82.6	49.0
LayoutLM$_{\text{LARGE}}$	OCR Encodings	72.6	27.2
DocLLM	OCR Encodings	69.5	–
Qwen2	OCR Words	76.4	44.5
DocVLM$_{\text{Qwen2}}$	OCR Encodings	**89.2**	**62.9**
DocVLM$_{\text{Qwen2}}$	64 Compressed Encodings	85.5	56.8

Table 7. **Effectiveness in LLMs (no visual input).** Comparison of DocVLM's full and compressed OCR encodings as *sole* input to Qwen2 LLM against OCR-only baselines, showing DocVLM's OCR encodings effectiveness even without visual features.

Visual Features	**64 Compressed Encodings**		**OCR Encodings**	
	DocVQA	InfoVQA	DocVQA	InfoVQA
×	85.5	56.8	89.2	62.9
✓	90.2	60.2	91.9	65.3
Δ	**+4.7**	**+3.4**	**+2.7**	**+2.4**

Table 8. **Contribution of Visual Features in DocVLM.** Comparison of DocVLM's performance in text-only mode (without visual features) versus full multimodal operation, using both compressed (64 tokens) and full OCR encodings. Results highlight the complementary benefits of visual information in DocVLM's architecture.

D.1. Exploring LLM Fine-tuning for Text-Only

Impact of LLM Fine-tuning with LoRA. To assess the potential for further improvement in DocVLM's text processing capabilities, we fine-tuned the LLM for an additional 100K steps using LoRA, focusing on the text-only mode of operation. Tab. 9 presents the results of this experiment, including a comparison with the baseline of inputting OCR words directly. Our results show that LoRA significantly improves the OCR words baseline performance. However, both compressed and full OCR encodings outperform this improved baseline even without LoRA fine-tuning. Notably, we observed only minor performance improvements when applying LoRA to the LLM with our OCR encodings, both compressed and full. Based on these findings in the text-only scenario, we decided against additional fine-tuning in our full multimodal DocVLM method. This decision helps maintain the vision and LLM alignment achieved through the extensive pretraining of the original VLM, ensuring that DocVLM enhances the existing VLM abilities without disrupting its pretrained knowledge.

LoRA	**OCR Words**		**64 Compressed Encodings**		**OCR Encodings**	
	DocVQA	InfoVQA	DocVQA	InfoVQA	DocVQA	InfoVQA
×	76.4	44.5	85.5	56.8	89.2	62.9
✓	80.3	49	85.7	56.8	89.4	63
Δ	**+3.9**	**+4.5**	**+0.2**	**+0**	**+0.2**	**+0.1**

Table 9. **Effect of LoRA Fine-tuning on Text-Only Performance.** Comparison before and after LoRA fine-tuning in text-only mode for OCR words baseline, compressed and full OCR encodings. Results show minimal gains for DocVLM's encodings.

E. Robustness to OCR Systems

We investigate DocVLM's robustness to various OCR systems, with particular emphasis on open-source OCR models that typically exhibit higher error rates due to limited training data. While our primary implementation uses Amazon Textract for text and layout extraction, we evaluated DocVLM's generalizability across different OCR architectures. Specifically, we tested text localization models including DB-ResNet50 and DB-FAST-base, combined with text recognition models such as CRNN-VGG16 and ParSeq. Notably, these evaluations involved only OCR system substitution without any additional model training.

Our experimental results, presented in Table 10, demonstrate that DocVLM consistently outperforms the baselines across all OCR system combinations tested, highlighting its robustness to varying OCR quality. This is particularly significant as it shows that our method remains effective even with open-source OCR systems and without any system-specific fine-tuning.

Visual Features	LLM OCR Input	Baseline	OCR System			
			[1]	[2]	[3]	[4]
256	OCR Words	84.4	87.4	87.4	87.6	89.3
	DocVLM (64)		**89.1**	**89.4**	**89.6**	**91.2**
512	OCR Words	91.5	91.3	91.1	91.2	92.0
	DocVLM (64)		**91.9**	**92.2**	**92.1**	**92.8**

Table 10. **DocVLM$_{\text{Qwen2-VL}}$ performances on DocVQA with different OCR systems.** OCR systems: [1] DocTR: DB-ResNet50 & CRNN-VGG16, [2] DocTR: DB-FAST-base & ParSeq, [3] DB-ResNet50 & ParSeq, [4] Our system (Textract). DocVLM consistently improves performance over the baseline and OCR words in prompt, even with lower-quality OCR systems.

F. Complexity Analysis

We compare DocVLM in the setting of 512 visual tokens + 64 learned OCR queries and compare it to the full-resolution Qwen2VL baseline which uses 16k visual tokens to highlight DocVLM's computational advantage. We consider (I) theoretical FLOP estimates, (II) empirical latency, and (III) GPU memory usage.

(I) FLOP Estimation: To quantify computational costs, we evaluate the number of FLOPs of all attention-based components (Vision encoder, OCR encoder, and LLM) and the OCR system. The FLOPs per attention layer for N tokens and dimension d are estimated by:

$$\underbrace{4\,N\,d^2}_{\text{Q,K,V,O projections}} + \underbrace{2\,N^2\,d}_{\text{attention inner product}} + \underbrace{m\,(h/d)\,N\,d^2}_{m \text{ projections } (d \to h)}.$$

For the OCR system, we reference LSGSpotter [38], which requires 194 GFLOPs for 1600×960 images. As a conservative upper bound, we estimate OCR processing at < 1 TFLOPs per image. Assuming 100 prompt tokens and up to 800 OCR tokens for DocVLM, our analysis shows that **DocVLM achieves a 72× reduction** in computational complexity (7.8 vs. 565.3 TFLOPs), as detailed in Table 11.

Component	L	Dim	Qwen2VL (16k)		DocVLM (576)	
			Seq. Length	TFLOPs	Seq. Length	TFLOPs
OCR System	–	–	–	–	–	< 1
OCR Encoder	24	1024	–	–	800	0.3
Vision Encoder	32	1280	$16{,}384 \times 4$	393	512×4	1.6
LLM	28	3584	$16{,}384 + 100$	172.3	$(512 + 64) + 100$	4.9
Total	–	–	–	**565.3**	–	**7.8**

Table 11. **Complexity comparison between Qwen2VL and DocVLM.** DocVLM achieves a 72× reduction in total computational complexity (7.8 TFLOPs) compared to Qwen2VL (565.3 TFLOPs), while maintaining high performance as shown in previous experiments.

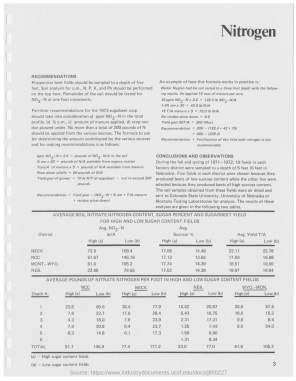
(II) Empirical Latency: We evaluate inference latency on an NVIDIA A100 GPU. The results are as follows:

- DocVLM achieves an inference time of 0.30s, compared to 1.46s for Qwen2VL.
- The OCR system, using DocTR without optimization, has a latency of 0.128s per DocVQA page, which is comparable to LSGSpotter's 0.14s.
- Overall, **DocVLM is 3.4× faster** in inference speed compared to Qwen2VL.

(III) GPU Memory Usage: In addition to computational savings, DocVLM significantly reduces peak GPU memory usage:

- DocVLM: 18.6GB
- Qwen2VL: 35.1GB

This reduction enables better scalability and lower hardware requirements, making DocVLM a more practical choice for real-world document understanding tasks.



Question: How many fields were sampled in each factory district?

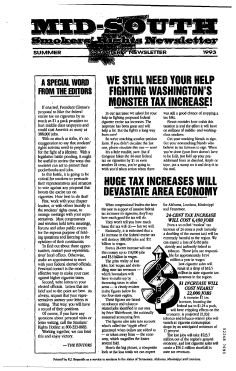
VLM: LLaVA-OneVision [1.5K image tokens]

Baseline: 3

DocVLM (Ours): 10

Zoomed Answer

CONCLUSIONS AND OBSERVATIONS
During the fall and spring of 1971–1972, 10 fields in each factory district were sampled to a depth of 5 feet (6 feet in Nebraska). Five fields in each district were chosen because they



Question: How many jobs will be lost due to a 24-cent tax increase?

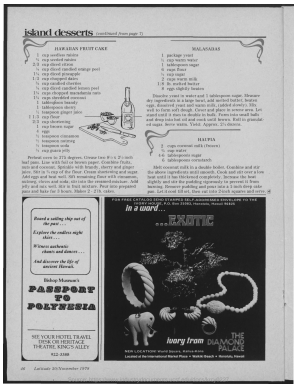
VLM: Qwen2-VL [256 image tokens]

Baseline: 300,000

DocVLM (Ours): 6,450

Zoomed Answer

24-CENT TAX INCREASE WILL COST 6,450 JOBS



Question: In the Hawaiian fruit cake, how many teaspoons of cinnamon are needed?

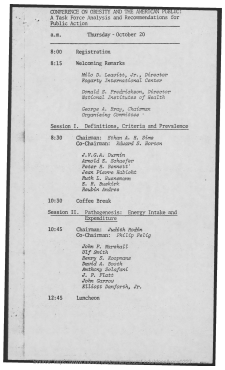
VLM: Qwen2-VL [512 image tokens]

Baseline: 1 teaspoon

DocVLM (Ours): 1/2 teaspoon

Zoomed Answer

½ teaspoon cinnamon



Question: Who is the co-chairman in the first session?

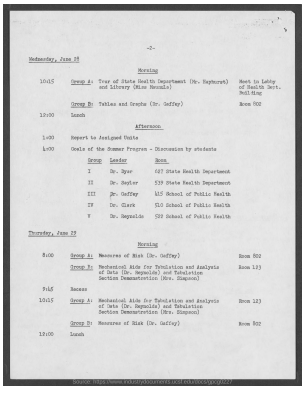
VLM: LLaVA-OneVision [1.5K image tokens]

Baseline: Edwin S. Horton

DocVLM (Ours): Edward S. Horton

Zoomed Answer

Session I. Definitions, Criteria and Prevalence
8:30 Chairman: Ethan A. H. Sims
Co-Chairman: Edward S. Horton



Question: Who is the leader in group two?

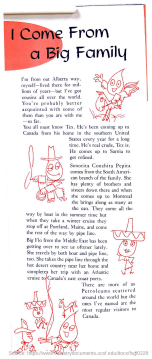
VLM: InternVL2 [256 image tokens]

Baseline: Dr. Gaffey

DocVLM (Ours): Dr. Saylor

Zoomed Answer

Group	Leader	Room
I	Dr. Dyar	627 State Health Department
II	Dr. Saylor	539 State Health Department
III	Dr. Gaffey	415 School of Public Health
IV	Dr. Clark	510 School of Public Health
V	Dr. Reynolds	522 School of Public Health



Question: Who comes to Canada from southern united states every year?

VLM: InternVL2 [256 image tokens]

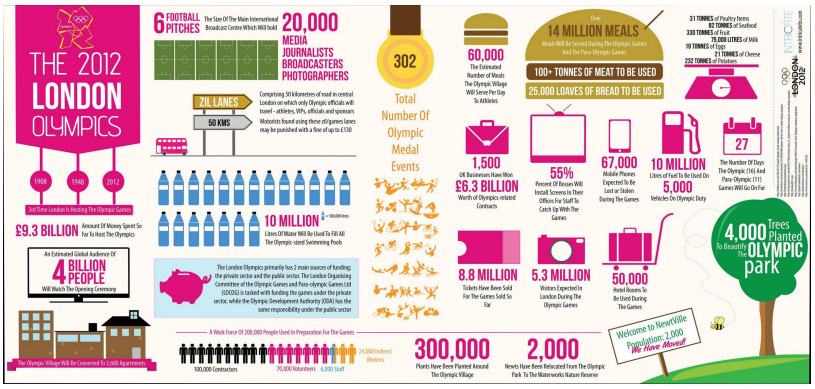
Baseline: Senator Concha Peña

DocVLM (Ours): Tex

Zoomed Answer

You all must know Tex. He's been coming up to Canada from his home in the southern United States every year for a long time. He's real crude, Tex is.

Figure 6. **Qualitative Results on Text-Heavy Documents.** Representative examples of DocVLM's performance on text-dense documents compared to baseline models (LLaVA-OneVision, InternVL2, and Qwen2VL). Each example shows an image-instruction pair with baseline and DocVLM predictions, demonstrating DocVLM's enhanced reading comprehension using only 64 OCR compressed tokens.



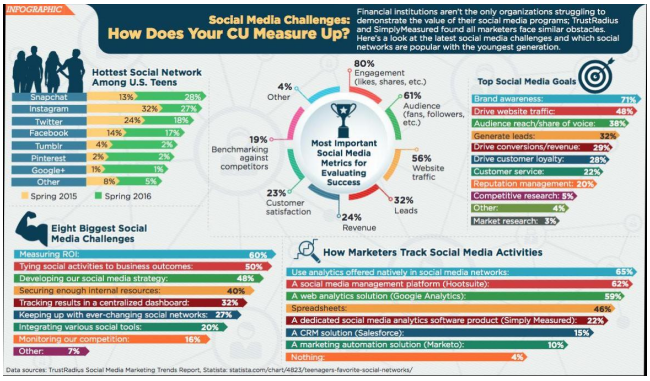
Question: How many days the Para-Olympic games will go on for?

VLM: Qwen2-VL [256 image tokens]

Baseline: 10

DocVLM (Ours): 11

Zoomed Answer



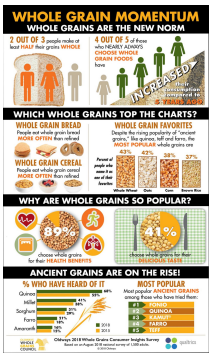
Question: What ranks as the third top social media goal?

VLM: InternVL2 [256 image tokens]

Baseline: Drive website traffic

DocVLM (Ours): Audience engagement

Zoomed Answer



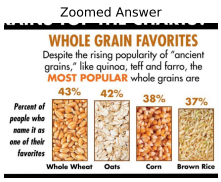
Question: What percentage of people listed oats as their favorite whole grain?

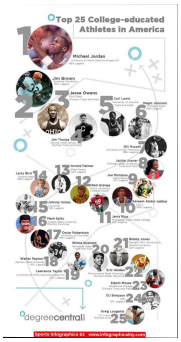
VLM: Qwen2-VL [256 image tokens]

Baseline: 47%

DocVLM (Ours): 42%

Zoomed Answer





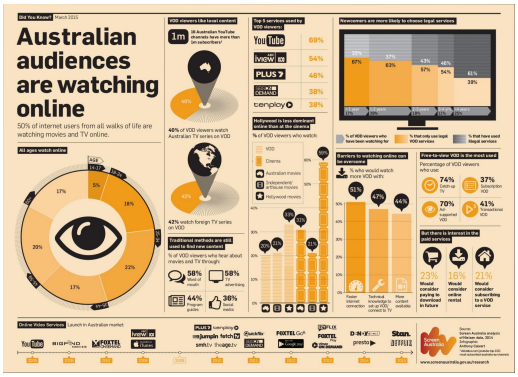
Question: Which university did Eric Heiden go to?

VLM: InternVL2 [1280 image tokens]

Baseline: The University of Oregon

DocVLM (Ours): Pennsylvania State University

Zoomed Answer



Question: How many people would consider online rentals?

VLM: Qwen2-VL [256 image tokens]

Baseline: 21%

DocVLM (Ours): 16%

Zoomed Answer

Figure 7. **Qualitative Results on Infographics.** Representative examples of DocVLM's performance on infographic-style documents compared to baselines under various visual token constraints, demonstrating improved handling of complex layouts and visual information.