

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 Doc-FormerV2 [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
	DocVQA [40]	train	×	39463
Desument VOA	InfoVQA [41]	train	×	46883
Document VQA	ChartQA [39]	train (H)	×	7398
	TAT-DQA [62]	train	×	13246
	TextVQA [48]	train	×	34602
Scene Text VQA	ST-VQA [11]	train	×	26308
	OCR-VQA [42]	train	×	800000
Captioning	TextCaps [47]	train	×	21953
Captioning	COCO Caption [15]	train	\checkmark	566747
General VQA	VQA-V2 [24]	train	\checkmark	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
Dooumont VOA	DocVQA [40]	Test	ANLS	×	5188
Document VQA	InfoVQA [41]	Test	ANLS	×	6573
Same Test VOA	TextVQA [48]	Val	VQAScore	×	5000
Scene Text VQA	ST-VQA [11]	Test	ANLS	×	4163
Captioning	TextCaps [47]	Val	CIDEr	×	3166
Maltine VOA	MP-DocVQA [50]	Test	ANLS	×	5019
Multipage VQA	DUDE [51]	Test	ANLS	\checkmark	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 _{LARGE}	OCR Encodings	72.6	27.2
DocLLM	OCR Encodings	69.5	_
Qwen2	ŌŪR Ŵords	76.4	44.5
DocVLM _{Qwen2}	OCR Encodings	89.2	62.9
DocVLM _{Qwen2}	64 Compressed Encodings	<u>85.5</u>	<u>56.8</u>

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 Comprese DocVQA	ssed Encodings InfoVQA	OCR En DocVQA	icodings InfoVQA
×	85.5	56.8	89.2	62.9
\checkmark	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 finetuning. 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.

I-DA	OCR	Words	64 Compre	ssed Encodings	OCR Encodings		
LOKA	DocVQA	InfoVQA	DocVQA	InfoVQA	DocVQA	InfoVQA	
×	76.4	44.5	85.5	56.8	89.2	62.9	
\checkmark	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 textonly 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 systemspecific fine-tuning.

Visual	LIMOCD Input	Dacalina		OCRS	System	
Features	LLW OCK Input	Dasenne	[1]	[2]	[3]	[4]
256	OCR Words	84.4	87.4	87.4	87.6	89.3
250	DocVLM (64)	04.4	89.1	89.4	89.6	91.2
512	OCR Words	01.5	91.3	91.1	91.2	92.0
512	DocVLM (64)	11.5	91.9	92.2	92.1	92.8

Table 10. **DocVLM**_{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:

$4 N d^2$	+	$2 N^2 d$	+	$m (h/d) N d^2$.
\sim		\sim		
Q,K,V,O projections		attention inner product		m projections $(d \rightarrow h)$

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	L	L	L	L	L	L	L	L	L	L	L	L	L	L	т	L	т	L	L	L	L	L	L	T	L	т	т	Dim	Qwen2VL	(16k)	DocVLM (5	i76)
component	1	Dim	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 $72 \times$ 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.

							Nitro	ogen	Quest each f	ion
									VLM: LLa	∍VA-
RECOMMEND	DATIONS									
her, Soit analy on the top feo MDg-N at and	el for o.m. el Pleneinde e Saet inceer	H, P, F, and Ph r of the soil pho- metric.	should be parfo id be tasted for	erned Prices ing m 24/pp	Nogram Aud I In No appli In NO ₂ -N + 2	his soid too had too a had 1.8 keess of mar 1.6 = 1.22 d ib 10	done foor diget we per are Og-AGA	with the Adham	Baselin	e: 3
Pertilizer reside should take int	nanendation ta comidera	s for the 1972 s. tion all post MD,	garbeet stop p-N in the tarts	15 F.		- 757.0 mA				
	ide. No mo fed from the g the amounts a recontence		200 pounds of The formula to the various sec-	N Face	neundetter	200 - (122.4 - 200 - (200.4) Fertilization of	42 + 751 Mill Terip with with	regentia mat	DocVL	м (
Descrate	nine - Kinin		5 en + 734 man	The sol		aired from the	to field: you's al	bied and		
	PRASE BO	Nor ploy down 6, NITRATE NI FOR P	TROGEN CON	Morran analyse TENT, BURAN / SJEAR CON	PERCENT TENT FILL	beratories for a the following t AND BUCARE DS	ratysis, The res we tables. RET YIELO	alloc of these		
AV Dista	· m	And provident R, NITPATE NI FOR 5 Ang. M BA 1004 OI	TROSEN CON IGH AND LOI Dg-M	Mortan analyse TENT, SUGAR / SUGAR CON	Toting La are given in PEPICENT TENT FIEL Aug. Success %	borstories far a the following t And BUCARE 26	Aug. YI	end T/A		
Overal Stock	ENACE DO	Aller providend R, NITPATE NI FOR 5 Ang, M BJ High OI 72.5	TRODEN CON IGH AND LOI Dg-N Low (k) 165.4	Morran andyse TENT SUGAR PSUGAR CON High 17.0	a Toting La are plan in PERCENT TENT FIEL Arg. Success % OI	Law Ibl 1445	Aug. Yi High ful 22.11	eld T/A Low (b) 22.76		
Dishi Dishi NCK NCC MONTWYO.	THE STATE	1600 ptor-Joant 61, NITPATE NI FOR 5 Arg, M Arg, M High GI 72,5 51,87 61,6 22,68	1790389 CON IGH AND LOI 03-M Lon 00 1654 1654 1652 29.50	Montan analyse TEMT, BURAN I SUBAR 005 I SUBAR 005 I T/R 17,0 17,1 17,5 17,5	a Toting Lil um given in PERCENT TENT FIEL Aug. Success % OI 5 2 8 2	Law Rd 14.45 13.30 14.45 13.30 14.30	Angola, The res we tables. EET VIELO Ango Yi High Sol 22,11 17,50 18,51 19,37	eld T/A Low (b) 22.76 19.80 19.90 19.90	CONCLU	JSIC
AV Desits NECK NECK NECK NER AVER	ACE POUN	Alex provident IL INITIATE NI FOR N Area M BA High OI 72.5 \$1.87 \$1.6 22.68 OS OF AITHAT	1710389 CDN 95H AND LOR 03-M 1054 1054 1054 1052 2050	Morran analyse TEMT, BURGAN I SUGAN COS I SUGAN COS I T/D I 7/D I	a Testing La are given in PERCENT TENT FIEL Ang Success % GI S 2 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	Law Ibi 14,45 12,80 14,30 14,30 14,30	Angola, The onit we tables. EET YIELD Ango YI High-Sci 22,11 17,50 16,51 19,97 2000TENT FIE	eld T/A Low Rd 22,76 19,86 19,90 19,84 10,90	CONCLU During th	USIC
AV Desits NEOK NEOK NEOK NEOK NEOK AVER Depth 6.	AGE POUR MADE POUR	Ann provident R, NITPATE RI FOR 5 Ang, M BA Migh OI 72,5 \$1,83 \$1,6 \$23,48 OS OF AITAATI HOC Low-W	THOGEN COM IGH AND LOR 023-M LON 00 1054 1054 1052 2930 E MITHOGEN I <u>NEE</u> Hagt 10	Morran analyse TENT, BURAN PSUGAR CON 17.0 17.1 17.0 17.1 17.0 17.1 17.0 17.1 17.0 17.1 17.0 17.1 17.0 17.1 17.0 17.0	a Toding La are given in PERIODAT TONT FACL Ang Shoreen % COL 9 2 4 2 4 4 2 4 4 2 4 4 2 4 4 4 4 4 4 4	Low IN 14.45 12.60 14.45 12.60 14.39 LOW EXCAR I 558 LOW EXCAR I 558 LOW EXCAR	August, The on we tables. EET VIELD Aug. YI High Od 22.11 31.50 36.51 39.37 CONTRICT FIEL WYY High-Od	eld T/A Low Bil 22.76 99.88 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.84 10.99 59.95 59.84 10.99 50.84 10.99 50.84 10.99 50.84 10.99 50.84 10.99 50.84 10.99 50.84 10.99 50.84 10.99 50.90 50.90 50.90 50.90 50.90 50.90 50.90 50 50 50 50 50 50 50 50 50 50 50 50 50	CONCLU During the	JSIC ne fa
AV Dealers NECK NOC MORT, WYO, NER AVER Depth &	ACE POUR 22.5	Ann arter deen Si, NITPATE AS FOR 5 Ang, M BA Migh OI 72.5 51.83 51.83 51.8 22.68 OS OF AITPATT NOC Low 30 60.5	1990 28% CDM ISH AND LOB 2g=M LOB 00 165.4 140.30 165.2 30.50 E MITHOLEN I <u>High 16</u> 35.4	Morran andyce TTMC REGAR (SUGAR COS) (SUGAR COS) (SUGA	e Testing Lai europeen in PERSONT TENT FOL Ang Secret 5 01 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	Line Ibl And States Ibl Line Ibl Line Ibl Ibl Ibl Ibl Ibl Ibl Ibl Ibl Ibl Ibl	Angust, The muse we station. RET VIELO Angust, The muse statistics Reg. VIELO 22,11 13,59 18,51 18,52 20,011 19,501 20,011 20,011 19,501 20,011 20,011 20,011 20,011 20,011 20,011 20,011 20,011 20,011 20,011 20,011 20,011 20,011	eld 17/A. Low Bd 227/6 19.80 19.90 19.34 LowCet LowCet Ecore/Bd 87.8	CONCLU During th factory d	JSIC ne fa
Ave Desnist NEOK NOC MONTWYO. NEE Depth b: 1 2 3	* 10 101402 00 AGE POUR 203 23 43	Ann arten denn 10, 1917PATC NI FOR 1 FOR 1 Ang M 10, 1940 172,5 51,83 22,8 05 0P ATTPATC 1000 05 0P ATTPATC 1000 00,6 22,3 15,0	1700289 C000 1564 AND L09 32-14 L09 30 1664 146.39 18654 186555 186555 186555 186555 186555 186555 186555 186555 186555 186555 186555 186555 186555 186555 186555 186555 1865555 1865555 1865555 1865555 1865555 1865555 18655555 18655555 186555555 186555555 18655555555 186555555555555555555555555555555555555	Horan Horan High High High High 17.0 17.1 17.5	a Testing Lai are given in PERCENT TEXT FIGU Arg. Success % GI S S S S S S S S S S S S S S S S S S	Low Iol 1448 Sciences An Low Iol 1448 Sciences An 1448 Sciences An 1448 Sciences An 1430 Science	Angust, The muse we sublex. RET VIELO Ang, Yi High-Sol 22,11 15.59 16.31 20.011 17.59 18.31 20.011	ell 17/A. Low Bal 22.70 19.80 19.90 19.34 Low Bal Low Bal 19.90 19.34 19.45 Low Bal 19.45 19.4	CONCLU During th factory d Nebraska	JSIC ne fa listri
AV Desna NECK NOC MONT-WYO, NEE AVER 1 3 4	* 10 1014.00 00 AGE POUR 22.8 2.6 2.6 2.6	60 stradeni 60 stradeni FOR 6 FOR	1700381 COM 654 AND LOI 554 AND LOI 555 105,4 105,2 70,500 105,2 70,500 105,4 105,2 70,500 105,4 10,6 7,8 9,4	Moran ani/yee TEMT, BUGAR, COS (SJGAR, COS	a Testing Lai are given in PERCENT TEXT FIGU Arg. Success % 01 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	Low Ed. Low Ed. Low Ed. Link Education of a Article Spectaries DS Link Ed. 14,45 14,45 14,30 14,	Aniput, The me we show we show. RET VIELD Anip.Y1 High fol 22.10 10.81 10.81 10.81 20.8 10.8 20.8 0.8 0.8	end 7/A Low Bd 22.78 10.99 10.90 10.	CONCLU During th factory d Nebraska	USIC ne fa listri I). Fi
Deshal NEOK NGC MEDI AVER J 2 3 4 5	* m EPLACE BC ACE POUR <u>Rightal</u> 22.8 2.4 3.4 3.4 8.3	600 a stree deard 81, NTRATE NU FOR 5 FOR 5 FOR 5 Area, 84, 80 87, 5 51, 87 51, 87 51	1700389 CON 654 AND LOI 554 AND LOI 554 1053 1052 1055 1055 1055 1055 1055 1055 1055 1055 1055 1055	Horan Horan Hap 17.0 28.4 29.7 17.3 17.5	e Testing Lai are given in PEPREENT TEST FIGL Arg. Success % old 5 2 HIGH AMD 18.62 5.43 2.53 18.62 5.43 2.53 19.55 19.57	Law Jbl Law Jbl Law Jbl Law Jbl Law Jbl 14,45 12,60 14,30	Aniput, The me we show. Map Y1 High GI 22,11 105,81 105,81 105,81 105,81 105,81 20,11 105,81 <tr< td=""><td>eld 17/8 Low Bit 22/86 99/96 99/96 99/96 99/96 10/96 Low/Bit 0.05 15/26 15/</td><td>CONCLU During th factory d Nebraska</td><td>JSIC ne fa listri l). F</td></tr<>	eld 17/8 Low Bit 22/86 99/96 99/96 99/96 99/96 10/96 Low/Bit 0.05 15/26 15/	CONCLU During th factory d Nebraska	JSIC ne fa listri l). F
0404 04044 NECK NECK NECK NECK NECK NECK NECK NECK	* m PERAGE SC ACE POUR 22.8 7.8 4.3 7.8 8.3 81.7	600 a strendend 81, NUTRATE NU FOR 5 FOR	1790389 CCN 1564 AND LCD 253-N Low (b) 16054 16053 16054 17054 170555 170555 170555 170555 170555 170555 170555 170555 170555 170555 170555 1705555 1705555 1705555 1705555 1705555 1705555 1705555 1705555 17055555 17055555 1705555555 170555555555555555555555555555555555555	Horan Horan Hap TEM, Buckan /SUSAR CON TEM, BUCKAN TEM TEM TEM TEM TEM TEM TEM TEM	1 Toting Lill are given in PEPGEDET PERCENT Solution Solu	Line ID Line ID <td< td=""><td>Ang-Y1 EET Y1ELD Ang-Y1 High-Sol 22.11 15.60 10.63 10.</td><td>elia 1774. Loss (ki) 1922 1924 1936 1934 1936 1934 1936 1934 1935 1934 1935 1934 1935 1934 1935 1934 1935 1935 1935 1935 1935 1935 1935 1935</td><td>CONCLU During th factory d Nebraska</td><td>JSIC ne fa listri l). F</td></td<>	Ang-Y1 EET Y1ELD Ang-Y1 High-Sol 22.11 15.60 10.63 10.	elia 1774. Loss (ki) 1922 1924 1936 1934 1936 1934 1936 1934 1935 1934 1935 1934 1935 1934 1935 1934 1935 1935 1935 1935 1935 1935 1935 1935	CONCLU During th factory d Nebraska	JSIC ne fa listri l). F
Ave District NDCX NDC NDC NDC NDC NDC NDC NDC NDC NDC NDC	* m TERLAGE SC ACE POUS 22.8 2.4 1.3 2.4 1.3 2.4 1.3 1.3 1.3 1.3 1.3 1.3 1.3 1.3	40, NITPATE NI 508 + Aug. M 508 + Aug. M 508 + Aug. M 72,5 51,80 72,5 51,80 72,5 51,80 72,5 51,80 72,5 51,80 72,5 51,80 72,5 51,80 72,5 51,80 72,5 73,5 74,5 74,5 74,5 74,5 74,5 74,5 74,5 74	THOSEN CON SCH AND LOD Jg-N Low (b) 165,4 165,30 185,2 30,50 185,2 30,50 185,2 30,50 187,70 20,4 17,8 7,8 9,4 6,1 77,4	10000000000000000000000000000000000000	a Tosting Lai are given to are given to reproduct TEAT FIGL Arg. Scores % oil 5 2 1100 1100 1100 1100 1100 1100 1100	Line IDI Line IDI Line IDI 14.45 Line IDI 14.35 Line IDI 15.31 J.3.31 3.43 EXH 5.34 T7.0 14.35	Ang-Y1 EET Y1ELD Ang-Y1 High-Sol 22.11 11.50 10.531 2007TENT P4EL 2013 0.53 0.53 0.53 0.53 0.53 0.53 0.53 0.5	eld T/A Low Bil 22:25 10:30 10:30 10:30 10:30 10:34 Low Bil 10:30 10:30 10:34 Low Bil 10:30 10:3	CONCLU During th factory d Nebraska	JSIC ne fa listri). F

: How many fields were sampled in ory district? OneVision [1.5K image tokens]

....

Ours): 10

Zoomed Answer IS AND OBSERVATIONS

and spring of 1971–1972, 10 fields in each t were sampled to a depth of 5 feet (6 feet in e fields in each district were chosen because they

stion. In the Hawaiian fruit cake



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



Question: Who is the co-shairman in the first

stand desserts (restricted (terr page 1)	many teaspoons of cinnamon are needed?	A Tack Force Analysis and Broomendations for Robic Action	session?
IDEGRAVE PERIFICACE INCASSORS I according index I package yiel in according index in acco	VLM: Qwen2-VL [512 image tokens]	8.m. Thursday October 20 8:00 Registration 8:15 Malconing Samarka	VLM: LLaVA-OneVision [1.5K image tokens]
any should also any should also any should be also any should also any sh	Baseline: 1 teaspoon	Milo D. Learist, Jr., Director Pagard, Xizarnatical Contar Donaid J. Tradicham, Director Bardonal Jourthucas of Sailth	Baseline: Edwin S. Horton
To you have a get the second of the sec	DocVLM (Ours): 1/2 teaspoon	Session F. Definitions, Criteria and Provilence 8.30 Chairman Fabor & S. Sore O-Chairman Fabor & S. Sore	DocVLM (Ours): Edward S. Horton
<text></text>	Zoomed Answer 1/2 teaspoon cinnamon	Andra David State	Zoomed Answer Session I. Definitions, Criteria and Prevalence 8:30 Chairman: <i>Bakars A. B. Stame</i> Co-Chairman: <i>Bakard S. Borton</i>
	Question: Who is the leader in group two?		Question: Who comes to Canada from souther united states every year?
-2- artiy <u>, Jun 21</u> <u>Bening</u>	VLM: InternVL2 [256 image tokens]	I Come From a Big Family	VLM: InternVL2 [256 image tokens]
Didd Owner A: Twee of distant Boalth Departments Div. Rephraveh) Here's to Ledge of Boalth Series, Sould Ange 0.0015 Owner Di Halas and Department Div. Soffwy) Boalth Series Boalth Series 0.2002 Di Halas and Department Div. Soffwy) Boant Boalth Series Boalth Series	Baseline: Dr. Gaffey	Fue from our Alberty wep, equal-back force for end- land of start-back for end- traction perhality better	Baseline: Senator Concha Peña
Literano. 100 Report to Assent Program - Opportunity profession L00 Outs at the Source Program - Opportunity profession L00 Description Description Description Description Description Description Description Description Description	DocVLM (Ours): Dr. Saylor	arguinted and the stores of the stores are an available of the stores of the stores of the cash of the stores Tex. He's force are using a store cash of the stores of the stores of the stores of the stores. He's many stores of the stores of	DocVLM (Ours): Tex
 Br. Seyler 533 Disks Kealth Department Br. Gaffery M.S. Shaoal of Public Mealth Dr. Garra 510 School of Public Mealth T. Dr. Shayalda 527 School of Public Mealth 	Zoomed Answer	Services Condition Payline one from the Should Ameri- ene brends of the family. Site has prevery all housines and the consense up to Montmal the foregoing and the same y and	
rfay, June 22 Norming	Group Leader Room	the ann. They entry all the way by hant in the surrange time has when they take a wirner crucia they may off as Pereland, Maine and come	
decomp A: Neareness of Bish (Dr. Osffray) Room 502 decomp A: Neareness of Bish (Dr. Torthalico and Scalaria Room 11)	I Dr. Dyar 627 State Health Department	We rest of the maps by pipe line. Eq. The dawn the Middle East has been gaming over an set on obstant linely.	Zoomed Answer
of Data (Dr. Meynold) on failure states Soften Demantertic (Dr. Sapon)	II Dr. Saylor 539 State Health Department	 of three by controls and page has, two She relates the pape here through the box doars county must har hour and ownphyre, her thy with an Adamic 	You all must know Tex. He's been coming up to Canada from his home in the southern United
015 droup A: Hechanical Aids for Tabelation and Enalysis Room 12) of Data (Dr. Reports) and Tabulation	III Dr. Gaffey 115 School of Public Health	errite to Canada's not construction of an There are more of an Perceducan centered	States every year for a long
owebbon Demonstration (New Engenes) <u>Group D</u> : Measures of Risk (Dr. Gaffay) Reem 802	TV Dr Clark 510 School of Public Hoolth	around the model for the ones five named are the return regular violater to Canada	∩ M time. He's real crude, Tex is.
2100 Sociali	1. Dr. derk jib School of Public Reside	AN (A)	
	V Dr. Reynolds 522 School of Public Health	W. HERE	

In

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: Owen2-VL [256 image tokens]

Baseline: 10

DocVLM (Ours): 11





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

VLM: InternVL2 [256 image tokens]

Baseline: Drive website traffic

DocVLM (Ours): Audience engagement











VLM: InternVL2 [1280 image tokens]

Baseline: The University of Oregon

DocVLM (Ours): Pennsylvania State University





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.