Supplementary Material for VisFocus: Prompt-Guided Vision Encoders for OCR-Free Dense Document Understanding

Ofir Abramovich[1*], Niv Nayman[2†], Sharon Fogel[*], Inbal Lavi[*], Ron Litman[2], Shahar Tsiper[2], Royee Tichauer[2], Srikar Appalaraju[2], Shai Mazor[2], and R. Manmatha[2]

[1] Reichman University, Israel [2] AWS AI Labs

A VisFocus Visualization

A.1 LMPM Pre-training

To further elucidate the efficacy of LMPM pre-training in focusing where prompt-related textual regions are, we conduct an additional extensive visualization in Fig. 1, showing multiple text tokens' aggregated attention maps across the document. It can be seen that most of the attention is activated where the sampled text snippet (served as the prompt) originally lies.

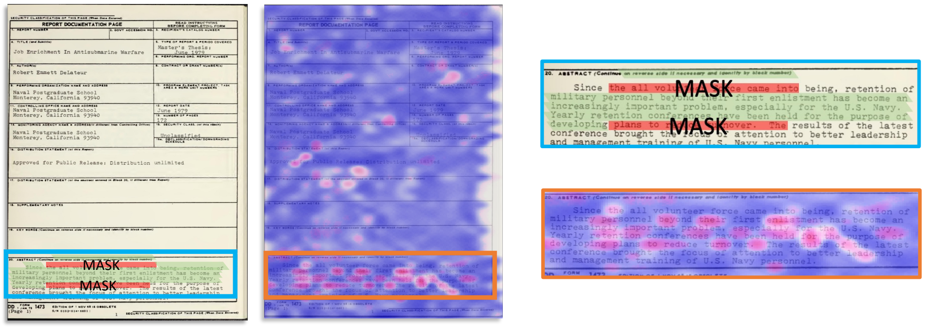


Fig. 1: Aggregated Token Visualizations for LMPM. We combine multiple attention maps between textual and visual token from the last layer of our document encoder. 🟩 denotes randomly sampled text snippets.

*Work done during an internship\employment at Amazon

† Corresponding author: nivnay@amazon.com

A.2 VQA Fine-tuning

As discussed, our novel LMPM pre-training encourages the model to focus on relevant portions of the document, concerning the input query. This is particularly demonstrated by the learned attention maps within the last ViLMA layers (Figs. 3a, 4 and 5), highlighting the correspondence between similar textual (query) and visual (context) tokens. The learned attention encompasses both literal word-level alignment ('center' ↔ 'center' in Fig. 3a, '100' ↔ '100' in Fig. 4b) and semantic relations ('birth' ↔ 'national' in Fig. 3a, 'be' ↔ 'are' in Fig. 4a, 'height' ↔ 'weight' in Fig. 5c).

B OCR-Free vs. OCR-Based

The OCR-free branch in document understanding aims to eliminate the need for external OCR systems, offering a more efficient standalone approach for processing document images. Consequently, OCR-free models' performance currently lags behind traditional OCR-based methods. This limitation stems from the absence of explicit textual information, which OCR-based models leverage as an additional input modality. Tab. 1 compares the two branches. Despite the remaining performance gaps, OCR-based methods depend on external systems, which indirectly add model parameters and are pruned to error propagation and thus heavily rely on the quality of the OCR engines.

Table 1: Comparison with OCR-based methods on VQA benchmarks. While the OCR-based approach still dominates in performance, the remaining gap with respect to OCR-free methods depends on the quality of external OCR engines which also implicitly add more parameters ($P*$) and complexity to the system.

	Method	#params	DocVQA ANLS	InfoVQA ANLS
OCR-based	LayoutLMV2-B [6]	200M + $P*$	78.1	-
	LayoutLMV2-L [6]	426M + $P*$	83.4	-
	LayoutLMV3 [6]	794M + $P*$	83.4	45.1
	UDOP [20]	794M + $P*$	84.7	47.4
	DocFormerV2-L [1]	368M + $P*$	**87.8**	**48.8**
OCR-free	Dessurt [4]	127M	63.2	-
	Donut [10]	176M	67.5	11.6
	ScreenAI-B [2]	670M	50.7	-
	Pix2Struct-B [11]	282M	72.1	**38.2**
	VisFocus-B	408M	**72.9** **(+1.2)**	31.9 **(+5.1)**

C Qualitative Comparisons

Figs. 6 to 10 provide additional examples where VisFocus-B excels in comparison to our baseline and Pix2Strcut-B and Fig. 11 extends the comparison to other OCR-free methods: Dessurt, Donut and Pix2Strcut-B. It can be seen that all but VisFocusoften predict wrong answers, extracted from somewhere in the document. This implies on the lack of focusing, as discussed in A, which leads to extraction of unrelated information and in turn to wrong predictions. Fig. 12 shows fail cases of VisFocuscompared to other OCR-free methods.

D VisFocus on Zero-shot Key-Value Extraction

VisFocus is originally designed for prompt-related document VQA tasks, but can be adapted to demonstrate its versatility on other document understanding tasks. One such task is key-value extraction, which can be reformulated as a prompt-related task. This reformulation allows leveraging VisFocus's capabilities beyond its original design scope.

To accomplish this, the key-value extraction task is reframed using a prompt template: "*What is the value of <key>?*". where *<key>* is some key in the form/reciept (Fig. 2). We refer to this task as *relaxed KV extraction*, since one should know a key in the document, and prompt it. We evaluate our proposed task on the FUNSD dataset [7], using DocVQA-finetuned checkpoints of VisFocus and previous works, and report superior performance in the zero-shot setting (Tab. 2). This evaluation strategy demonstrates the flexibility of prompt-based models like VisFocus and explores their potential for tackling diverse document understanding tasks through clever task reformulation.

Table 2: Comparison of zero-shot Relaxed KV Extraction task on FUNSD dataset. We report ANLS on the test set, applying the reformulated KV task.

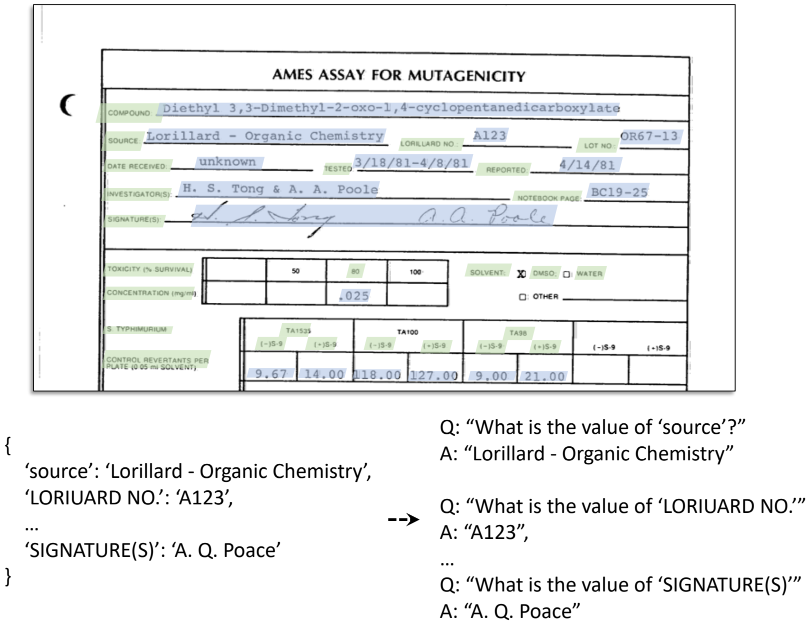
Method	ANLS
Donut	58.9
VisFocus-S	**60.2** (+1.3)
Pix2Struct-B	62.7
VisFocus-B	**63.4** (+0.7)

E Datasets and Hyperparameters

In this section we present in detail every benchmark and dataset used in our work. For more pre-training and fine-tuning details see Tabs. 3 and 4.

E.1 Pre-training Data

For pretraining data, we utilize the IDL-OCR dataset [3], comprising 26M document pages accompanied by corresponding raster-scan OCR outputs. In Tab. 5



AMES ASSAY FOR MUTAGENICITY

COMPOUND: Diethyl 3,3-Dimethyl-2-oxo-1,4-cyclopentanedicarboxylate

SOURCE: Lorillard - Organic Chemistry LORILLARD NO.: A123 LOT NO.: OR67-13

DATE RECEIVED: unknown TESTED: 3/18/81-4/8/81 REPORTED: 4/14/81

INVESTIGATOR(S): H. S. Tong & A. A. Poole NOTEBOOK PAGE: BC19-25

SIGNATURE(S): H. S. Tong A. A. Poole

TOXICITY (% SURVIVAL)		50	80	100
CONCENTRATION (mg/ml)			.025	

SOLVENT: ☒ DMSO: □ WATER

□: OTHER

S. TYPHIMURIUM	TA1535		TA100		TA98			
	(−)S-9	(+)S-9	(−)S-9	(+)S-9	(−)S-9	(+)S-9	(−)S-9	(+)S-9
CONTROL REVERTANTS PER PLATE (0.05 ml SOLVENT)	9.67	14.00	118.00	127.00	9.00	21.00		

```
{
  'source': 'Lorillard - Organic Chemistry',
  'LORIUARD NO.': 'A123',
  ...
  'SIGNATURE(S)': 'A. Q. Poace'
}
```

-->

Q: "What is the value of 'source'?"
A: "Lorillard - Organic Chemistry"

Q: "What is the value of 'LORIUARD NO.'"
A: "A123",
...
Q: "What is the value of 'SIGNATURE(S)'"
A: "A. Q. Poace"

Fig. 2: Visualization of the Relaxed KV Extraction. We re-define the key-value extraction as a prompt-based task to apply zero-shot on VQA fine-tuned models. 🟩 and 🟦 denote keys and values respectively.

we compare our pre-training data with previous methods. [4, 11] create different labels for documents, whereas we employ only the OCR text, similar to [10]. Dessurt collects textual data to create synthetic documents using open-sourced fonts. It also re-renders IIT-CDIP [5, 12] and FUNSD [7] with different fonts and layouts, while Pix2Struct scrape the web to generate structured representations of documents (HTML DOMs). Pre-training our model with text-oriented approaches, further provides an advantage for our method when dealing with dense documents of many words.

E.2 Downstream Tasks

Here we provide technical details about the downstream datasets we experimented with and some bottom line results.

DocVQA [17] is an open-ended VQA dataset consists of various types of scanned documents. It is a subset of IDL corpus, consists of $\sim 14k$ document images and $\sim 40k$ questions. We use the ANLS metric and report a boost of $+\mathbf{1.2}$ on the test split over the baseline, and $+\mathbf{0.8}$ over Pix2Struct-B, which is the current state-of-the-art on small OCR-free models.

Table 3: Model hyper-parameters for pre-training. We use AdamW [14] optimizer and Cosine Annealing [13] scheduler. We train on 8 A100 GPUs. '$*$' denotes early stopping. All reported numbers apply for all our model variants.

PT Stage	#steps	Batch Size	Base LR	Image Resolution
LtR	$200K^*$	32	$1e-4$	1536×768
LMPM	$400K$	48		

Table 4: Model hyper-parameters for fine-tuning. Same as in pre-training, in all our experiments we adopt AdamW [14] optimizer, Cosine Annealing [13] scheduler, early-stopping and train on 8 A100 GPUs.

	Dataset	#Steps	Batch Size	Base LR	Image Resolution
VisFocus-S	DocVQA	$15K$	72	$1e-4$	1536×768
	InfoVQA	$15K$	32	$5e-5$	
	ChartQA	$15K$	72	$2e-4$	
	OCR-VQA	$50K$	64	$5e-5$	
	AI2D	$30K$	512	$1e-4$	
VisFocus-B	DocVQA	$15K$	72	$1e-4$	1536×768
	InfoVQA	$15K$	144	$1e-4$	
	ChartQA	$15K$	72	$2e-4$	
	OCR-VQA	$50K$	144	$1e-4$	
	AI2D	$30K$	32	$5e-5$	

InfographicsVQA (InfoVQA) [16] contains various infographics with annotations for questions that demand reasoning across text, layout, graphics, and data visualizations. It consists of $\sim 5k$ images and $\sim 30k$ questions. Since VisFocus was trained to encode the question with respect the question, and given that InfoVQA has more visual than textual content, along with complex numerical reasoning, its performance drops and becomes less competitive. However, our approach still beats the baseline by $+\mathbf{5.1}$ points.

ChartQA [15] is a large-scale benchmark dataset designed to evaluate models' ability to answer complex questions about charts, requiring both visual and logical reasoning. It consists of $9.6K$ human-written questions and $23.1K$ generated questions based on summaries. We follow previous works and report average Relaxed Accuracy (RA) of each split. Even though VisFocus is not trained on

Table 5: Pre-traing Data. Comparison with previous OCR-Free methods. "I" is denoted as the IIT-CDIP dataset [5], 'Form','Handwriting', and 'Wiki' are synthetic datasets presented in [4]. OCR refers to raster-scan order.

	Pre-training Datasets	Annotations	#Samples
Dessurt [4]	I+Form+Handwriting+Wiki	OCR	not reported
Donut [10]	I+SynthDog [10]	OCR	13.5M
Pix2Struct [11]	C4 [19]	HTML DOMs + OCR	80M
VisFocus	IDL-OCR [3]	OCR	25.6M

structure-related tasks, it achieves an improvement of +**4.6** over the baseline, and exceeding previous works.

OCR-VQA [18] is a large-scale dataset of ∼$200k$ book cover images and $1M$ questions. The task requires high skills of reading text. We report Exact Match (EM) on the test set and outperform our baseline by +**3.1** and +**0.6** points over the baseline and Pix2Struct-B, respectively.

AI2 Diagrams (AI2D) [9] consists of ∼$5K$ grade-school science diagrams, corresponding multiple-choice questions testing comprehension and reasoning about the diagrams. VisFocus-B achieves a +**2.2** boost over the baseline and +**6.9** compared to Pix2Struct-B, the prior state-of-the-art model in our setting.

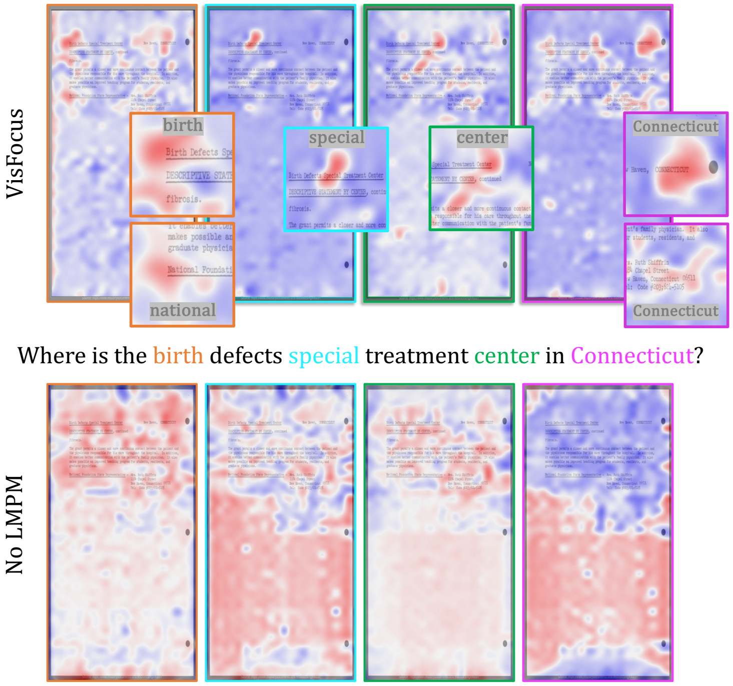
E.3 Prompt Encoding

To quantify the impact of the prompt encoding, we conduct an ablation study on several different text encoding techniques, involving both context-aware encoding (T5 encoder [19], TinyBERT [8]) and independent learned token embeddings (Tab. 6). It can be seen that designated T5 based encoders perform best, possibly due to the alignment with the T5 language model cascaded to the vision encoder. Reducing its size by about 65% decreases the DocVQA ANLS by merely 0.2. This motivates further research of smaller variants for T5 to be utilized in our framework.

Table 6: Prompt Encoding Ablation Study. 'Embedding' denotes independent token embedding and 'shared' as the VisFocus's LM encoder. 'shared' uses the T5 encoder for encoding the textual prompt in addition to the visual features. 'Cross-modal grad.' computes gradients from both paths.

Prompt Encoder	Δ #Params	DocVQA	ChartQA
	(frozen)	ANLS	RA
Embedding	-	71.4	55.4
TinyBERT [8]	8M	71.3	56.2
T5-Base Enc. (Copy of the learnt LM)	-	71.7	56.0
T5-Small Enc.	39M	72.0	56.3
T5-Base Enc.	**113M**	**72.2**	**57.1**

Fig. 3: ViLMA Attention Maps. Attention maps of the last ViLMA layer activated at words from the input prompt, the frames are colored according to the colored prompt tokens and the highly activated visual tokens are explicitly written inside the boxes. Top rows: VisFocus. Bottom rows: VisFocus **without LMPM pre-training**.



(a)

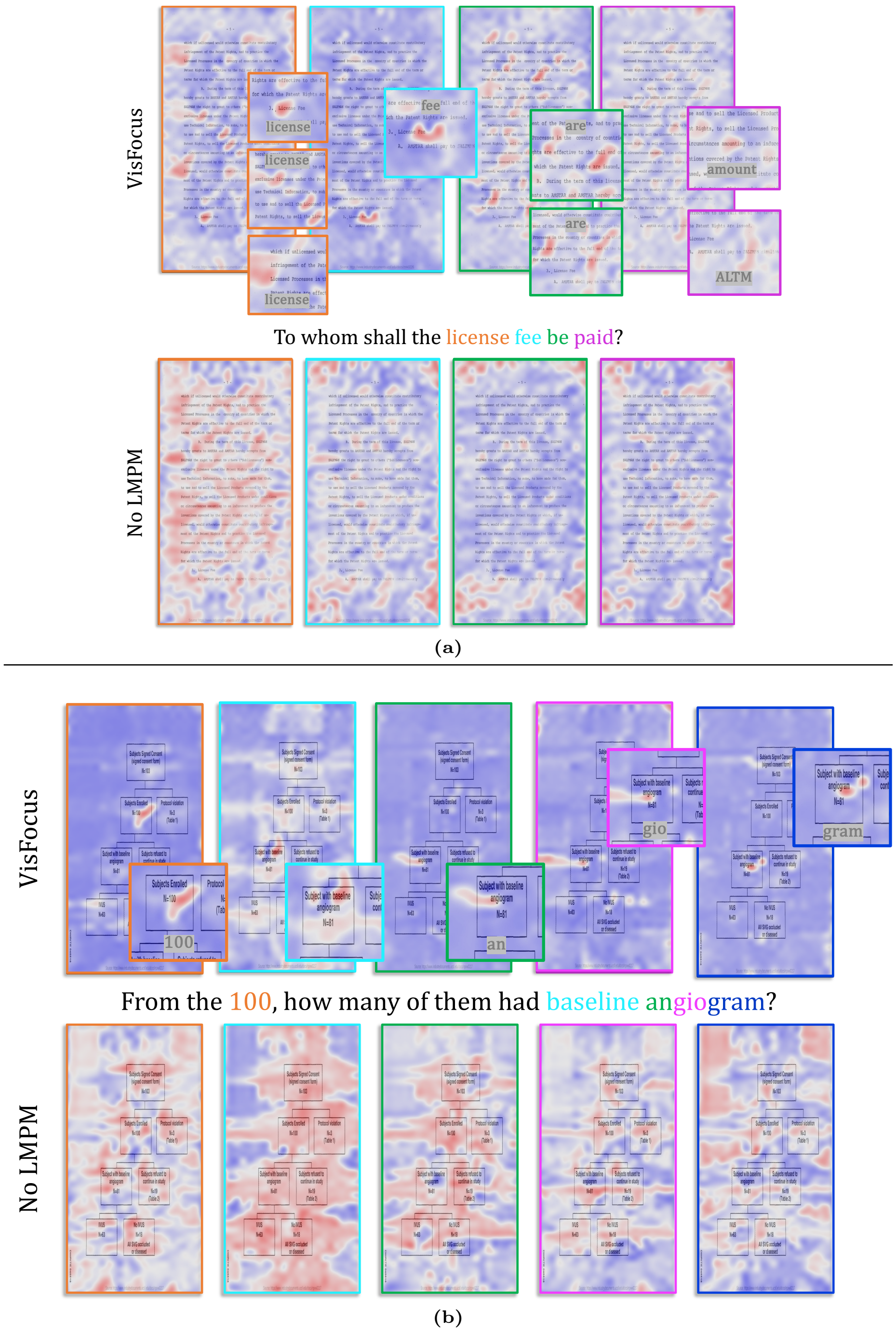
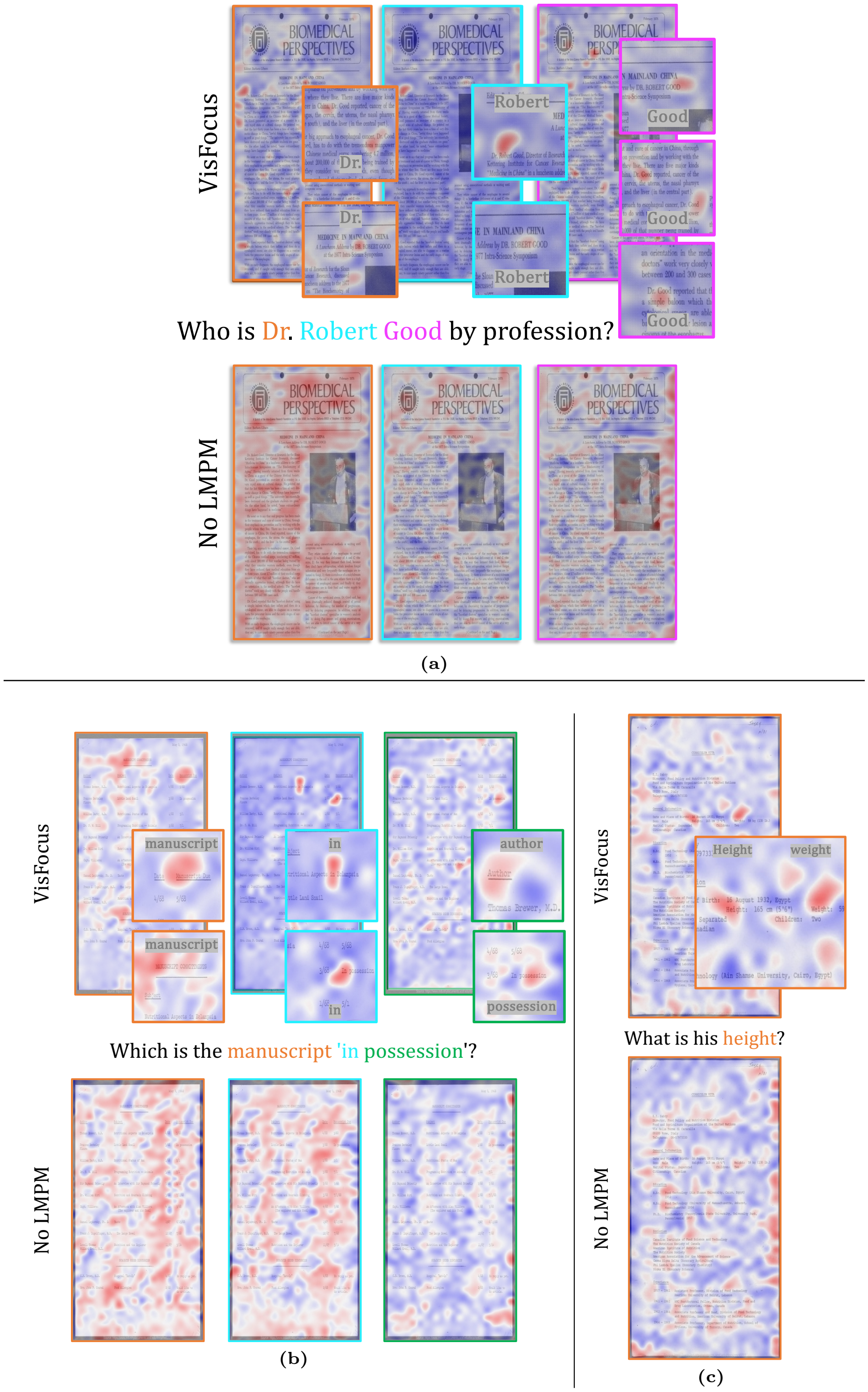

VisFocus
license
license
license
fee
are
are
amount
ALTM
To whom shall the license fee be paid?
No LMPM
(a)
VisFocus
100
gio
gram
an
From the 100, how many of them had baseline angiogram?
No LMPM
(b)

Fig. 4: Fig. 3 continued.



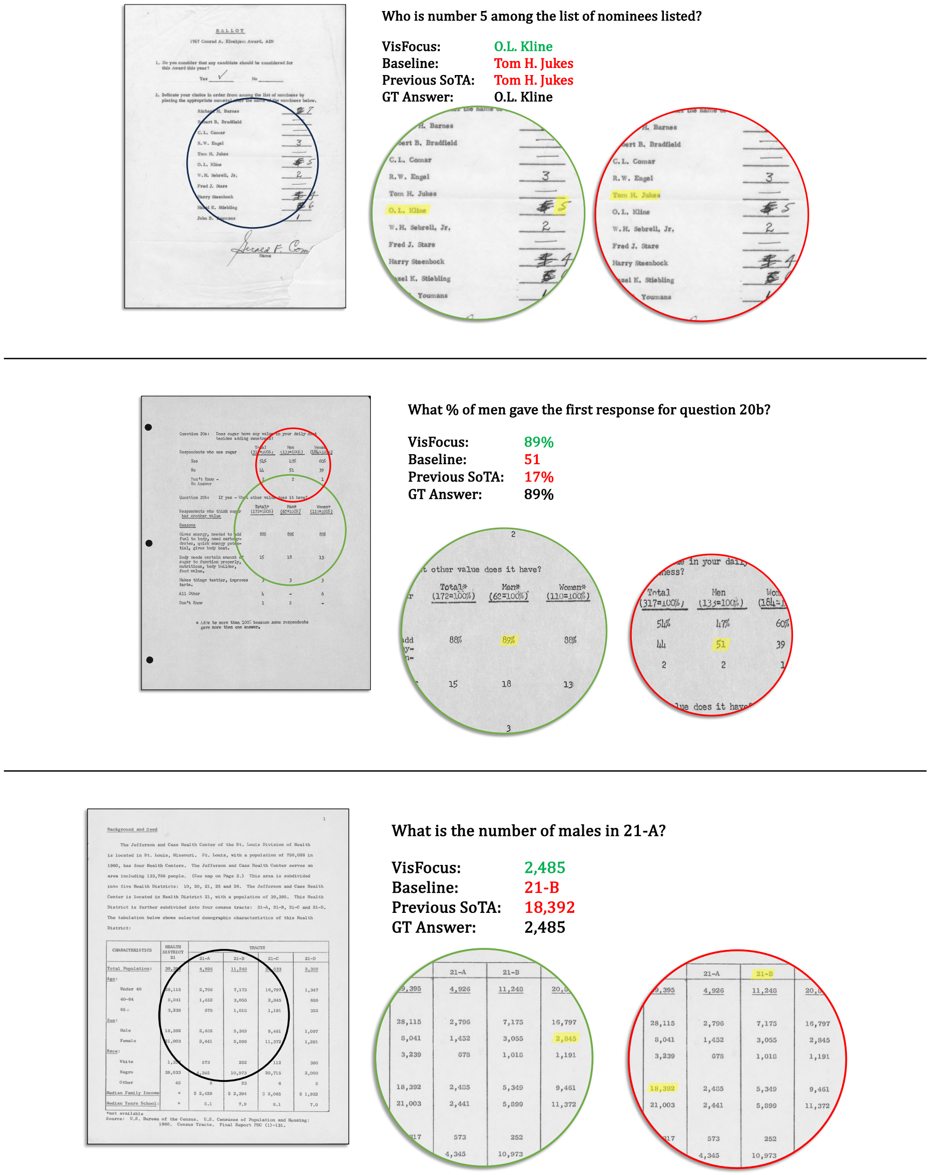
(a)

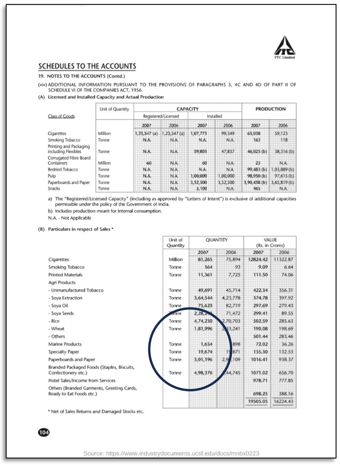
(b)

(c)

Fig. 5: Fig. 3 continued.

Fig. 6: Qualitative comparison. Further examples from the DocVQA validation set, demonstrating VisFocus's ability to accurately answer questions on denser documents, compared to previous SoTA (Pix2Struct-B) and to our baseline





ITC Limited

SCHEDULES TO THE ACCOUNTS

19. NOTES TO THE ACCOUNTS (Contd.)

(xiv) ADDITIONAL INFORMATION PURSUANT TO THE PROVISIONS OF PARAGRAPHS 3, 4C AND 4D OF PART II OF SCHEDULE VI OF THE COMPANIES ACT, 1956.

(A) Licensed and Installed Capacity and Actual Production

Class of Goods	Unit of Quantity	CAPACITY				PRODUCTION	
		Registered/Licensed		Installed			
		2007	2006	2007	2006	2007	2006
Cigarettes	Million	1,23,547 (a)	1,23,547 (a)	1,07,773	99,349	63,038	59,123
Smoking Tobacco	Tonne	N.A.	N.A.	N.A.	N.A.	163	118
Printing and Packaging including Flexibles	Tonne	N.A.	N.A.	59,803	47,837	46,025 (b)	38,316 (b)
Corrugated Fibre Board Containers	Million	60	N.A.	60	N.A.	23	N.A.
Redried Tobacco	Tonne	N.A.	N.A.	N.A.	N.A.	99,483 (b)	1,03,889 (b)
Pulp	Tonne	N.A.	N.A.	1,00,000	1,00,000	98,950 (b)	97,615 (b)
Paperboards and Paper	Tonne	N.A.	N.A.	3,52,500	3,52,500	3,90,458 (b)	3,65,819 (b)
Snacks	Tonne	N.A.	N.A.	5,100	N.A.	465	N.A.

a) The "Registered/Licensed Capacity" (including as approved by "Letters of Intent") is exclusive of additional capacities permissible under the policy of the Government of India.

b) Includes production meant for internal consumption.

N.A. - Not Applicable

(B) Particulars in respect of Sales *

	Unit of Quantity	QUANTITY		VALUE (Rs. in Crores)	
		2007	2006	2007	2006
Cigarettes	Million	81,265	75,894	12824.42	11322.87
Smoking Tobacco	Tonne	164	93	9.09	6.64
Printed Materials	Tonne	11,361	7,725	111.50	74.06
Agri Products					
- Unmanufactured Tobacco	Tonne	49,691	45,714	422.34	356.31
- Soya Extraction	Tonne	3,64,544	4,23,778	374.78	397.92
- Soya Oil	Tonne	75,623	82,719	297.69	279.43
- Soya Seeds	Tonne	2,28,243	71,472	299.41	89.55
- Rice	Tonne	4,74,230	2,70,703	502.59	285.63
- Wheat	Tonne	1,81,996	2,33,241	190.08	198.69
- Others				501.44	283.46
Marine Products	Tonne	1,654	898	72.02	36.26
Specialty Paper	Tonne	19,674	19,871	135.30	132.53
Paperboards and Paper	Tonne	3,01,196	2,90,109	1016.41	938.37
Branded Packaged Foods (Staples, Biscuits, Confectionery etc.)	Tonne	4,98,376	3,44,745	1071.02	656.70
Hotel Sales/Income from Services				978.71	777.85
Others (Branded Garments, Greeting Cards, Ready to Eat Foods etc.)				698.25	388.16
				19505.05	16224.43

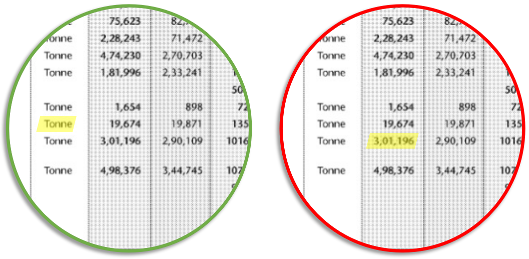
* Net of Sales Returns and Damaged Stocks etc.

104

Source: https://www.industrydocuments.ucsf.edu/docs/mnbx0223

What is the "unit of quantity" of Paperboards and paper?

VisFocus: tonne
Baseline: industrial
Previous SoTA: 3,01,196
GT Answer: Tonne



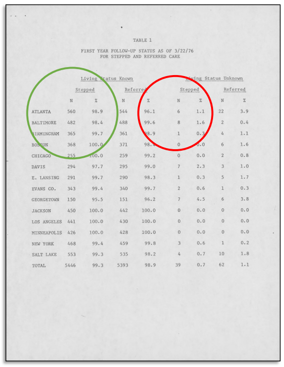


TABLE 1

FIRST YEAR FOLLOW-UP STATUS AS OF 3/22/76
FOR STEPPED AND REFERRED CARE

	Living Status Known				Living Status Unknown			
	Stepped		Referred		Stepped		Referred	
	N	%	N	%	N	%	N	%
ATLANTA	560	98.9	544	96.1	6	1.1	22	3.9
BALTIMORE	482	98.4	488	99.6	8	1.6	2	0.4
BIRMINGHAM	365	99.7	361	98.9	1	0.3	4	1.1
BOSTON	368	100.0	371	98.4	0	0.0	6	1.6
CHICAGO	255	100.0	259	99.2	0	0.0	2	0.8
DAVIS	294	97.7	295	99.0	7	2.3	3	1.0
E. LANSING	291	99.7	290	98.3	1	0.3	5	1.7
EVANS CO.	343	99.4	340	99.7	2	0.6	1	0.3
GEORGETOWN	150	95.5	151	96.2	7	4.5	6	3.8
JACKSON	450	100.0	442	100.0	0	0.0	0	0.0
LOS ANGELES	441	100.0	430	100.0	0	0.0	0	0.0
MINNEAPOLIS	426	100.0	428	100.0	0	0.0	0	0.0
NEW YORK	468	99.4	459	99.8	3	0.6	1	0.2
SALT LAKE	553	99.3	535	98.2	4	0.7	10	1.8
TOTAL	5446	99.3	5393	98.9	39	0.7	62	1.1

In Baltimore, what is the no. of stepped cases, whose living status is known?

VisFocus: 482
Baseline: by fleeing the jurisdiction
Previous SoTA: 8
GT Answer: 482

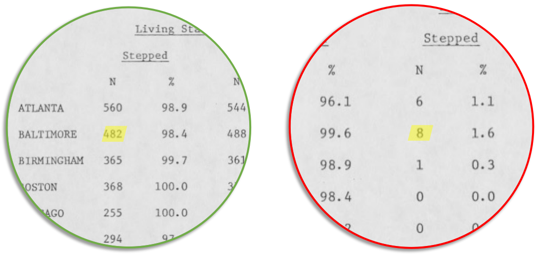
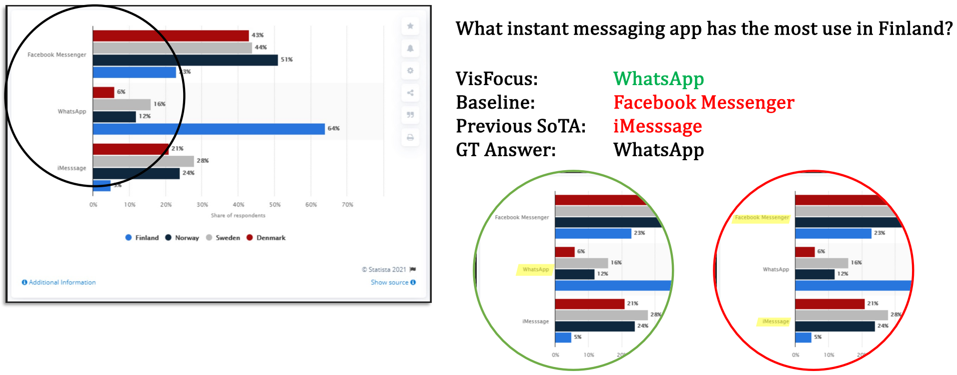


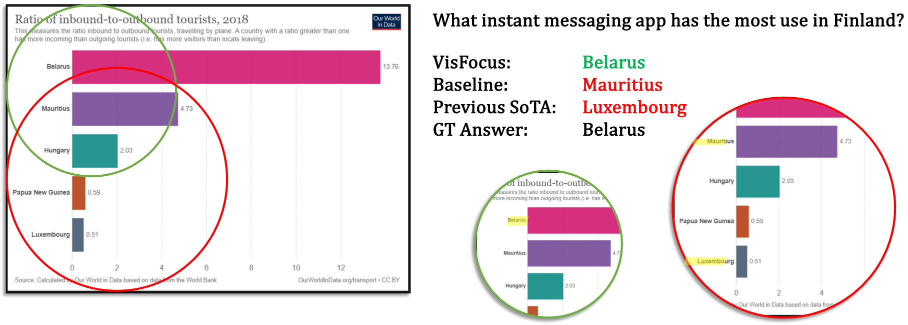
Fig. 7: Fig. 6 continued.

Fig. 8: Qualitative comparison. Further examples from ChartQA test set, demonstrating VisFocus's ability to accurately answer questions on visual data such as charts and plots, compared to previous SoTA (Pix2Struct-B) and to our baseline. The last example shows a fail case for all of the compared methods.



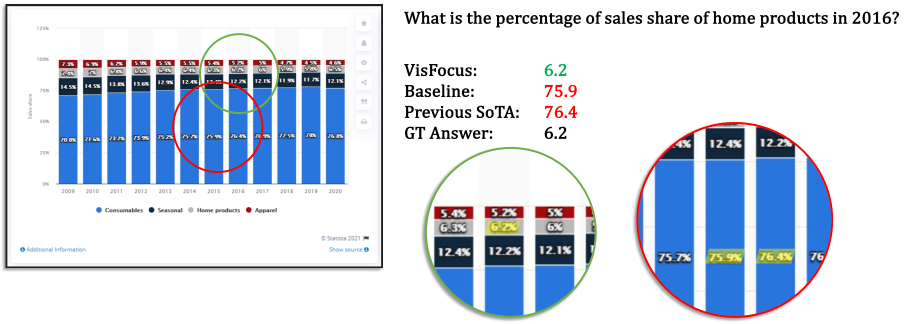
What instant messaging app has the most use in Finland?

VisFocus: WhatsApp
Baseline: Facebook Messenger
Previous SoTA: iMesssage
GT Answer: WhatsApp



What instant messaging app has the most use in Finland?

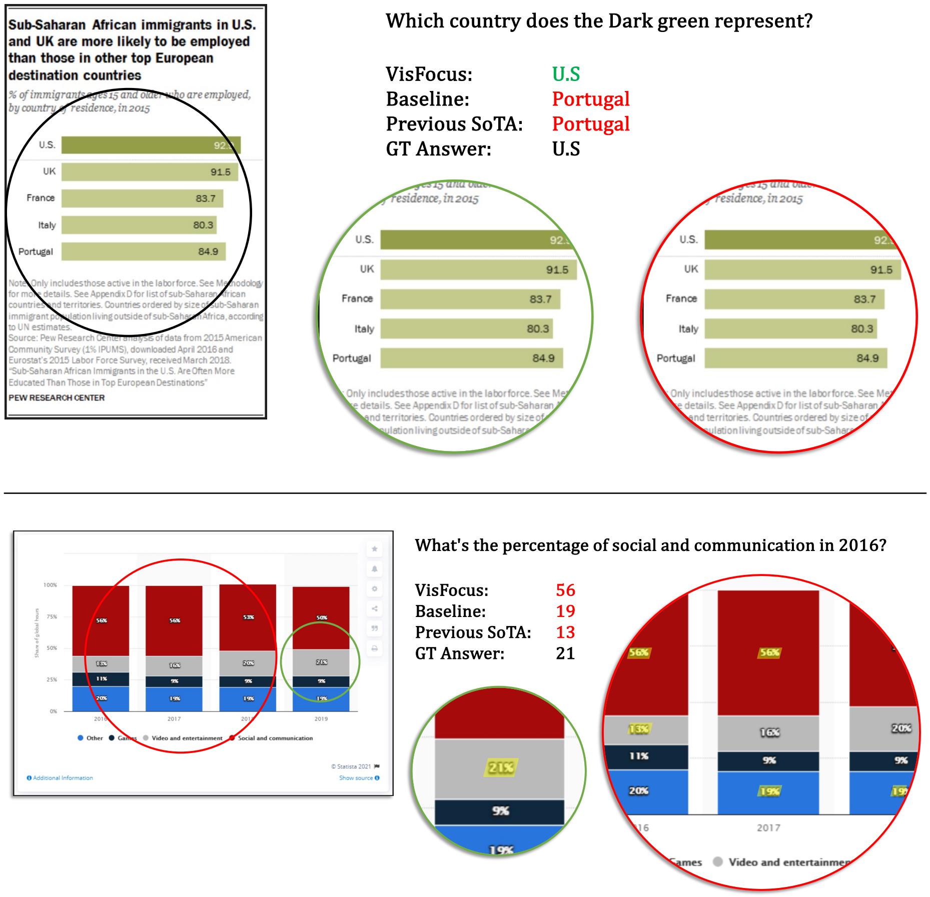
VisFocus: Belarus
Baseline: Mauritius
Previous SoTA: Luxembourg
GT Answer: Belarus



What is the percentage of sales share of home products in 2016?

VisFocus: 6.2
Baseline: 75.9
Previous SoTA: 76.4
GT Answer: 6.2

Fig. 9: Fig. 6 continued.



Which country does the Dark green represent?

VisFocus:	U.S
Baseline:	Portugal
Previous SoTA:	Portugal
GT Answer:	U.S

What's the percentage of social and communication in 2016?

VisFocus:	56
Baseline:	19
Previous SoTA:	13
GT Answer:	21

Fig. 10: Fig. 6 continued.

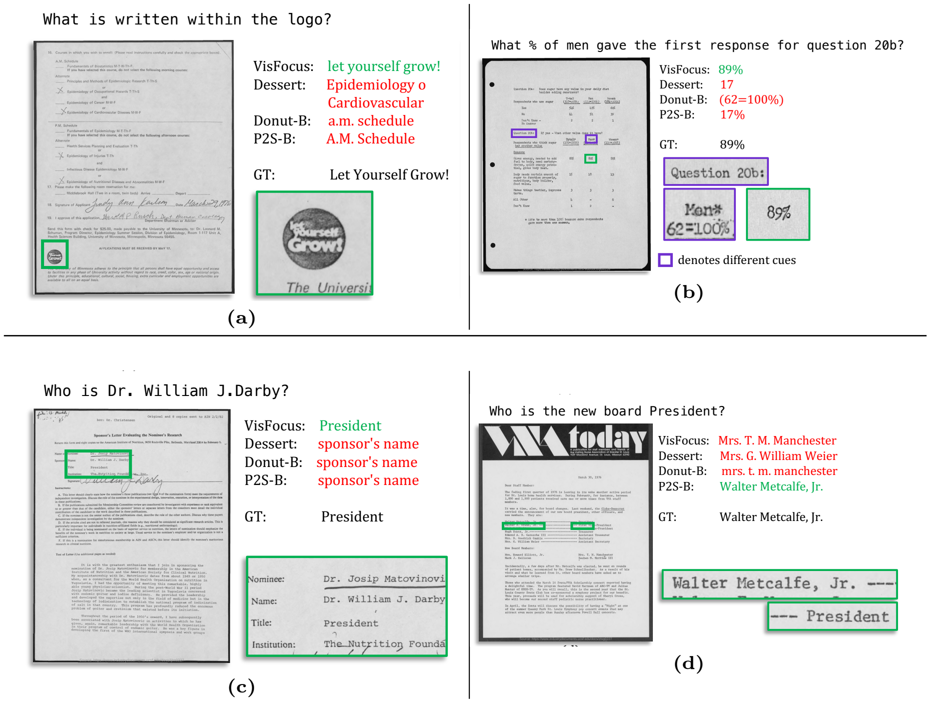

What is written within the logo?
VisFocus: let yourself grow!
Dessert: Epidemiology o Cardiovascular
Donut-B: a.m. schedule
P2S-B: A.M. Schedule
GT: Let Yourself Grow!
Let Yourself Grow!
The Universi
(a)
What % of men gave the first response for question 20b?
VisFocus: 89%
Dessert: 17
Donut-B: (62=100%)
P2S-B: 17%
GT: 89%
Question 20b:
Men*
62=100%
89%
denotes different cues
(b)
Who is Dr. William J.Darby?
VisFocus: President
Dessert: sponsor's name
Donut-B: sponsor's name
P2S-B: sponsor's name
GT: President
Nominee: Dr. Josip Matovinovi
Name: Dr. William J. Darby
Title: President
Institution: The Nutrition Founda
(c)
Who is the new board President?
VisFocus: Mrs. T. M. Manchester
Dessert: Mrs. G. William Weier
Donut-B: mrs. t. m. manchester
P2S-B: Walter Metcalfe, Jr.
GT: Walter Metcalfe, Jr.
Walter Metcalfe, Jr. ---
--- President
(d)

Fig. 11: Qualitative comparison. Further success cases of VisFocus, compared to the failures of other OCR-free methods: Dessurt, Donut and Pix2Struct-B.

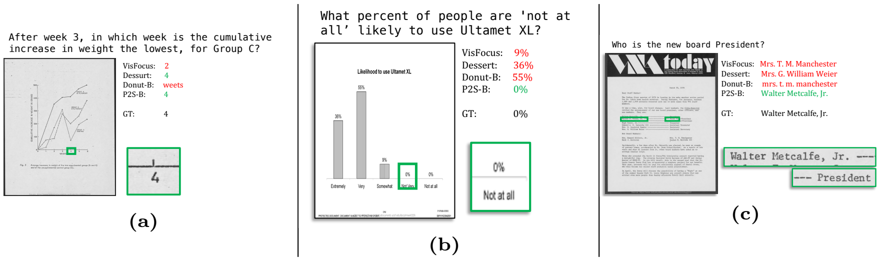

After week 3, in which week is the cumulative increase in weight the lowest, for Group C?
VisFocus: 2
Dessert: 4
Donut-B: weets
P2S-B: 4
GT: 4
4
(a)
What percent of people are 'not at all' likely to use Ultamet XL?
Likelihood to use Ultamet XL
36%
55%
9%
0%
0%
Extremely
Very
Somewhat
Not Very
Not at all
VisFocus: 9%
Dessert: 36%
Donut-B: 55%
P2S-B: 0%
GT: 0%
0%
Not at all
(b)
Who is the new board President?
VisFocus: Mrs. T. M. Manchester
Dessert: Mrs. G. William Weier
Donut-B: mrs. t. m. manchester
P2S-B: Walter Metcalfe, Jr.
GT: Walter Metcalfe, Jr.
Walter Metcalfe, Jr. ---
--- President
(c)

Fig. 12: Qualitative comparison. Failure examples of cases where VisFocus fails and other OCR-free methods: Dessurt, Donut and Pix2Struct-B succeed better.

References

1. Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., Manmatha, R.: Docformerv2: Local features for document understanding. In: AAAI Conference on Artificial Intelligence (2024) 2
2. Baechler, G., Sunkara, S., Wang, M., Zubach, F., Mansoor, H., Etter, V., Cărbune, V., Lin, J., Chen, J., Sharma, A.: Screenai: A vision-language model for ui and infographics understanding. arXiv preprint arXiv:2402.04615 (2024) 2
3. Biten, A.F., Tito, R., Gomez, L., Valveny, E., Karatzas, D.: Ocr-idl: Ocr annotations for industry document library dataset. In: European Conference on Computer Vision. pp. 241–252. Springer (2022) 3, 6
4. Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., Morariu, V.: End-to-end document recognition and understanding with dessurt. In: European Conference on Computer Vision. pp. 280–296. Springer (2022) 2, 4, 6
5. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR) 4, 6
6. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091 (2022) 2
7. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents (2019) 3, 4
8. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding (2019) 6, 7
9. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. ArXiv **abs/1603.07396** (2016), https://api.semanticscholar.org/CorpusID:2682274 6
10. Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Donut: Document understanding transformer without ocr. arXiv preprint arXiv:2111.15664 **7**, 15 (2021) 2, 4, 6
11. Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.W., Toutanova, K.: Pix2struct: Screenshot parsing as pretraining for visual language understanding. In: International Conference on Machine Learning. pp. 18893–18912. PMLR (2023) 2, 4, 6
12. Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building a test collection for complex document information processing. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 665–666. SIGIR '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1148170.1148307, https://doi.org/10.1145/1148170.1148307 4
13. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with restarts. CoRR **abs/1608.03983** (2016), http://arxiv.org/abs/1608.03983 5
14. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. CoRR **abs/1711.05101** (2017), http://arxiv.org/abs/1711.05101 5
15. Masry, A., Long, D., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 2263–2279. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.177, https://aclanthology.org/2022.findings-acl.177 5

16. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022) 5
17. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021) 4
18. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR (2019) 6
19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints (2019) 6
20. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C.Y., Bansal, M.: Unifying vision, text, and layout for universal document processing. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19254–19264 (2022), https://api.semanticscholar.org/CorpusID:254275326 2