Enhancing Visual Document Understanding with Contrastive Learning in Large Visual-Language Models

Supplementary Material

1. Multimodal Feature Extraction

Within DoCo, we utilize LayoutLMv3 [3] to extract the multimodal features of the document objects, which encompass textual embeddings, visual embeddings and layout embeddings. This section provides a comprehensive description of the multimodal feature extraction process.

Textual embeddings: We pre-process the document images with an off-the-shelf OCR toolkit PaddleOCR [2] to obtain textual content and bounding boxes of the document objects. The textual embeddings are subsequently extracted with the word embedding matrix from LayoutLMv3 [3], culminating in a sequence with a length of T and a dimension of d_m. The maximum value of T is established at 512.

Visual embeddings: Inspired by LayoutLMv3 [3], the document image is resized to $H \times W$, and subsequently divided into a sequence of $P \times P$ patches. These patches are then linearly projected to a dimension of d_m and flattened into a sequence with a length of $I = HW/P^2 = 196$.

Layout embeddings: For the layout embeddings, we follow [3] to involve 1D and 2D position embeddings to the T textual tokens and I image patches, where the 1D position refers to the index of tokens and the 2D position denotes the box coordinates of the corresponding object layouts.

As a result, these embeddings are forwarded to the multimodal encoder to aggregate the multimodal features $\widetilde{\mathbf{F}}^m = \left\{\mathbf{f}_1^m, \mathbf{f}_2^m, ..., \mathbf{f}_{N+1}^m\right\} \in \mathbb{R}^{(N+1)\times d_m}$. Note that each object is processed as a sequence of $\mathbf{f}_i^m$, and all the $N+1$ objects $\left\{\mathbf{f}_1^m, \mathbf{f}_2^m, ..., \mathbf{f}_{N+1}^m\right\}$ are processed in batch by the encoder.

2. Fine-tuning Datasets

During the fine-tuning stage, LVLMs undergo optimization utilizing approximately 0.4 million text-rich datasets, which include TextVQA [11], DocVQA [5], ChartQA [4], OCRVQA [7], InfoVQA [6], KLC [12], WTQ [8], and TextCaps [10]. These datasets can be classified into three task categories: document Image captioning, key information extraction and document visual question answering.

Document image captioning: Document image captioning involves generating descriptive text for a given document image, necessitating models to interpret and rationalize the text within these images to produce accurate captions. This process specifically requires models to integrate a novel modality of text present in the images, and to reason over both this text and the visual content within the image to generate comprehensive image descriptions. To enhance the performance of the model on the task of document image captioning, we utilize the training split of TextCaps [10].

Key information extraction: Key information extraction in document understanding, alternatively referred to as Property Extraction, denotes the methodological process of pinpointing and extracting salient or pertinent information from a specified document. This extracted information can span a wide array of data types, encompassing elements such as names, dates, geographical locations, organizational entities, financial figures, among other specific details that are integral to the comprehensive understanding of the document's content. In the present study, we leverage the training and validation splits of the KLC [12] dataset to enhance the performance of our model in executing the key information extraction task.

Document visual question answering: Document visual question answering bears a superficial resemblance to knowledge information extraction in terms of structure. However, upon closer examination, the differences become more pronounced. Visual question answering necessitates the interpretation of an open-ended set of questions and the ability to handle a variety of document types, thereby requiring superior generalization capabilities. Moreover, the specific content under analysis necessitates a more profound understanding of visual elements, as the questions often pertain to figures and graphics that accompany the formatted text. To enhance the performance of our model, we utilize a variety of highly recognized public question-answering benchmark datasets, which include TextVQA [11], DocVQA [5], ChartQA [4], OCRVQA [7], InfographicVQA [6] and WTQ [8].

3. More Interpretability on DoCo

In Fig. 1 and Fig. 2 of this appendix, we visualize more heat-maps derived from the image encoder and the generated tokens at each decoding phase between CLIP [9] and DoCo based on the Qwen-VL-Chat [1] framework.

Fig. 1 elucidates the interpretability within text-rich document scenarios. It is observed that the attention heat-maps of DoCo adeptly capture pertinent information associated with the correct response, whereas the maps of CLIP display a significant drift and weak semantic interrelation deviated from the ground truth, which is evident in the second and fourth instances of the figure. Moreover, aligning the multimodal features to the visual representations in the document object level enhances comprehension of fine-grained text details and effectively mitigates recognition errors, which is corroborated by the first, third and fifth instances of the figure.

Fig. 2 further illustrates additional results within text-rich natural scenes, following a similar pattern. For instance, in the first case, the CLIP model generates incorrect tokens "be-a-ure-gard" from steps 7 to 10, which bear no relevance to the question "type". Furthermore, the attention maps of CLIP display a significant lack of focus and drift from step 7, resulting in the absence of fine-grained features. By comparison, the maps of DoCo tend to focus on the continuous fine-grained textual features, yielding satisfactory results. Additionally, our DoCo leverages the context information to predict the subsequent results, as demonstrated by the image features of "2002" in step 11, which are fuzzy but can be inferred to the correct tokens from the context "en 2002 par" below.

In summary, Fig. 1 and Fig. 2 compellingly demonstrate that DoCo can assist the image encoder in capturing more effective visual cues in text-rich scenarios.

4. More Qualitative Results

Fig. 3 presents an expanded set of qualitative outcomes derived from a variety of benchmark datasets between CLIP and DoCo based on Qwen-VL-Chat [1]. The illustrations underscore the superior generalization capacity of DoCo, which assists the vision encoder of LVLMs in acquiring more efficacious cues and enhances comprehension in text-rich scenes.

We also delineate the failure cases of our method in Fig. 4. It is evident that our DoCo still struggles with document-related commonsense reasoning and mathematical calculations, which furnishes invaluable insights for the enhancement of document comprehension capabilities with LVLMs in this domain. Future research endeavors will investigate these problems and attempt to attack them by further improving visual understanding performance.

5. Broader Impact

The remarkable proficiencies of LVLMs hold vast potential for facilitating more robust document analysis and comprehension in text-rich environments, but the significance of fine-grained features remains largely unexplored within the LVLM community. Thanks to the document object discrimination between visual and multimodal representations, our proposed DoCo tailored for the fine-grained feature collapse issue can yield more precise results in text-rich scenarios. The acquisition of more efficient fine-grained visual representations opens up a plethora of potential applications and opportunities for visual document understanding tasks. We advocate for researchers to develop LVLMs integrated with DoCo for text-rich tasks, as we anticipate this to be particularly advantageous.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[2] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 1

[3] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022. 1

[4] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1

[5] 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

[6] Minesh Mathew, Viraj Bagal, Rubèn 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

[7] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 1

[8] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[10] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 1

[11] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1

[12] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 1

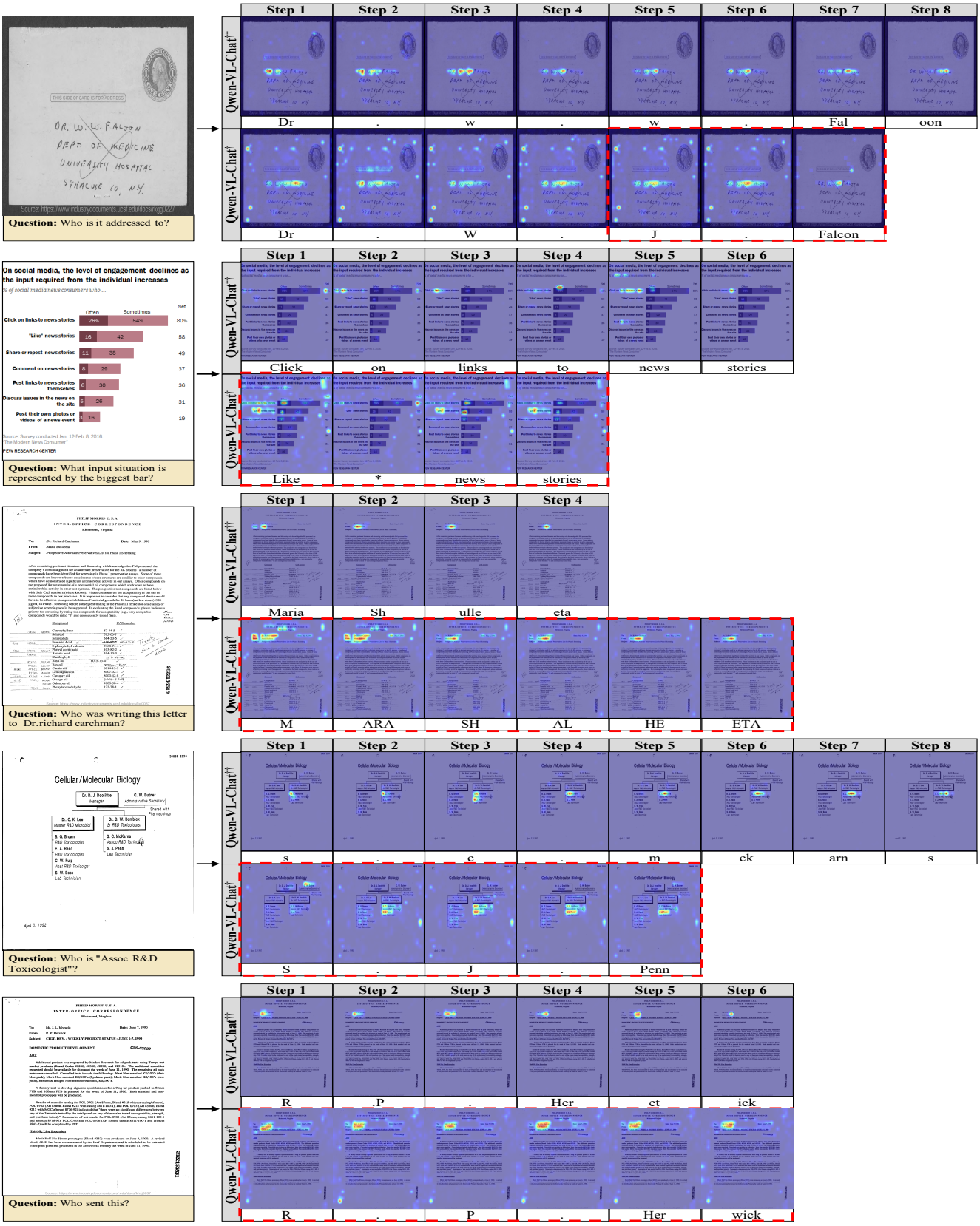
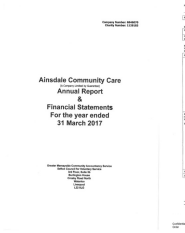


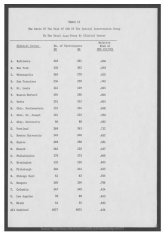
Figure 1. Visualization of the heat-maps and generated tokens from CLIP ("†") and DoCo ("††") in text-rich document scenarios.

Figure 2. Visualization of the heat-maps and generated tokens from CLIP ("†") and DoCo ("††") in text-rich natural scenes.

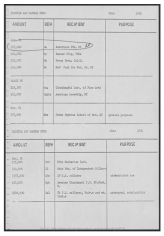


Question: What is the value for the charity_number?
Qwen-VL-Chat†: 1139183
Qwen-VL-Chat††: 1129183

Question: What is the Target date for the 'supporting activities'?
Qwen-VL-Chat†: current status / comments / next action
Qwen-VL-Chat††: Mid 1999



Question: What is the no. of participants in UC group in Miami?
Qwen-VL-Chat†: 211
Qwen-VL-Chat††: 51

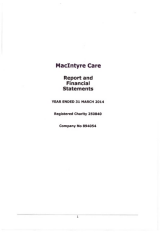


Question: What is REF# for RECIPIENT 'Ohio mechanics inst.'?
Qwen-VL-Chat†: 51
Qwen-VL-Chat††: Evt

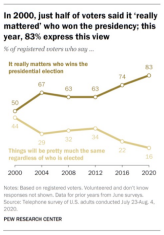
Question: What is the code written as footer at the right of the document
Qwen-VL-Chat†: Dwh05404
Qwen-VL-Chat††: DWRITE 066042

Question: What is the 'subtotal'?
Qwen-VL-Chat†: 2 year inflation @ 15% per year
Qwen-VL-Chat††: $353.5

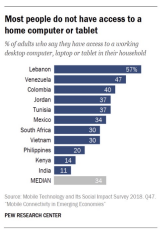
Question: What is the P. O. Drawer no.?
Qwen-VL-Chat†: 2955
Qwen-VL-Chat††: 1734



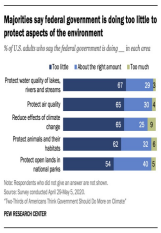
Question: What is the value for the charity_number?
Qwen-VL-Chat†: 894054
Qwen-VL-Chat††: 250840



Question: What's the peak value of dark brown graph?
Qwen-VL-Chat†: 63
Qwen-VL-Chat††: 83

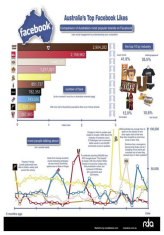


Question: What is the percentage value in the Jordan who say they have access to a working desktop computer, laptop, or tablet in their household?
Qwen-VL-Chat†: 40
Qwen-VL-Chat††: 37

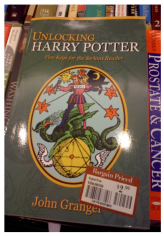


Question: Which answer has the majority of votes? (Too little, about the right amount or too much)
Qwen-VL-Chat†: about the right amount
Qwen-VL-Chat††: Too little

Question: Experiences of the which branch of the clan will you be most interested in?
Qwen-VL-Chat†: african
Qwen-VL-Chat††: alberta

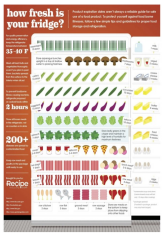


Question: What percentage of Australians are interested in Clothing/Apparrel?
Qwen-VL-Chat†: 10.6%
Qwen-VL-Chat††: 35.9%



Question: How much does this book cost?
Qwen-VL-Chat†: 7.99
Qwen-VL-Chat††: 2.99

Question: What time is it according to this radio?
Qwen-VL-Chat†: 12:41
Qwen-VL-Chat††: 12:54



Question: How long yogurt can be stored fresh in the refrigerator?
Qwen-VL-Chat†: 7 days
Qwen-VL-Chat††: 2 weeks

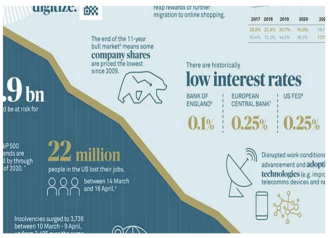
Rank	Cyclist	Team	Time	UCI ProTour Points
1	Alejandro Valverde (ESP)	Caisse d'Epargne	5h 29' 10"	40
2	Alexandr Kolobnev (RUS)	Team CSC Saxo Bank	s.t.	30
3	Davide Rebellin (ITA)	Gerolsteiner	s.t.	25
4	Paolo Bettini (ITA)	Quick Step	s.t.	20
5	Franco Pellizotti (ITA)	Liquigas	s.t.	15
6	Denis Menchov (RUS)	Rabobank	s.t.	11
7	Samuel Sánchez (ESP)	Euskaltel-Euskadi	s.t.	7
8	Stéphane Goubert (FRA)	Ag2r-La Mondiale	+ 2"	5
9	Haimar Zubeldia (ESP)	Euskaltel-Euskadi	+ 2"	3
10	David Moncoutié (FRA)	Cofidis	+ 2"	1

Question: What was the total number of points by franco pellizotti?
Qwen-VL-Chat†: 20
Qwen-VL-Chat††: 15

Question: Who was the next opponent after david wheaton?
Qwen-VL-Chat†: Amos Mansfield
Qwen-VL-Chat††: Amos Mansdorf

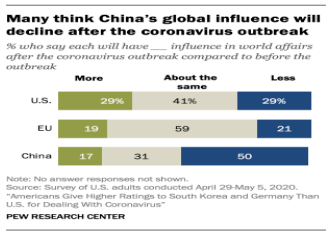
Contestant	Age	Height	Home City	Rank
Isabelle Raisa	16	170 cm (5 ft 7 in)	Vienna	Eliminated in Episode 1
Alina Chlebecek	18	170 cm (5 ft 7 in)	Deutsch-Wagram	
Sabrina Angelika Rauch [1]	21	175 cm (5 ft 9 in)	Graz	Eliminated in Episode 2
Katharina Mihalović	23	179 cm (5 ft 10.5 in)	Vienna	
Nataša Marić	16	175 cm (5 ft 9 in)	Liefering (originally from Serbia)	Eliminated in Episode 3
Michaela Schopf	21	172 cm (5 ft 7.5 in)	Salzburg (originally from Germany)	Quit in Episode 4
Christine Riener	20	181 cm (5 ft 11.25 in)	Bludenz	Eliminated in Episode 4
Teodora-Mădălina Andreica	17	177 cm (5 ft 9.5 in)	Romania	Eliminated in Episode 6
Yennsi Rieger	17	177 cm (5 ft 9.5 in)	Vienna	Eliminated in Episode 7
Izabela Pop Kostić	20	170 cm (5 ft 7 in)	Vienna (originally from Bosnia)	Eliminated in Episode 8
Bianca Ebelsberger	24	179 cm (5 ft 10.5 in)	Aurach am Hongar	Eliminated in Episode 9
Nadine Trinker	21	183 cm (6 ft 0 in)	Bodensdorf	
Dzejlana "Lana" Baltić	20	179 cm (5 ft 10.5 in)	Graz (originally from Bosnia)	1st Eliminated in Episode 10
Melisa Popanicić	16	175 cm (5 ft 9 in)	Wörgl	2nd Eliminated in Episode 10
Gina Zeneb Adamu	17	175 cm (5 ft 9 in)	Bad Vöslau	Runner-Up
Antonia Maria Hausmair	16	175 cm (5 ft 9 in)	Siegendorf	Winner

Question: How long in height is isabelle raisa?
Qwen-VL-Chat†: 160 cm
Qwen-VL-Chat††: 170 cm (5 ft 7 in)

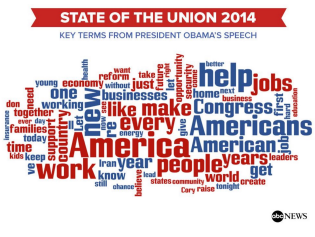


Question: What is the interest rates of European Central Bank and US FED?
Qwen-VL-Chat†: 0.1% and 0.25%
Qwen-VL-Chat††: 0.25%

Figure 3. More qualitative results between CLIP ("†") and DoCo ("††") based on the Qwen-VL-Chat model in text-rich scenes.



Question: What's the average of all the values in the green bars (round to one decimal)?
Qwen-VL-Chat††: 30.33
Ground Truth: 21.6



Question: Which is the biggest word in the infographic?
Qwen-VL-Chat††: STATE
Ground Truth: America

Question: what time is on the watch?
Qwen-VL-Chat††: 10:10
Ground Truth: 12:35

Year	Film	Role	Language	Notes
2008	*Moggina Manasu*	Chanchala	Kannada	Filmfare Award for Best Actress - Kannada Karnataka State Film Award for Best Actress
2009	*Olave Jeevana Lekkachaara*	Rukmini	Kannada	Innovative Film Award for Best Actress
2009	*Love Guru*	Kushi	Kannada	Filmfare Award for Best Actress - Kannada
2010	*Krishnan Love Story*	Geetha	Kannada	Filmfare Award for Best Actress - Kannada Udaya Award for Best Actress
2010	*Gaana Bajaana*	Radhey	Kannada	
2011	*Huduguru*	Gayithri	Kannada	Nominated, Filmfare Award for Best Actress - Kannada
2012	*Alemari*	Neeli	Kannada	
2012	*Breaking News*	Shraddha	Kannada	
2012	*Addhuri*	Poorna	Kannada	Udaya Award for Best Actress Nominated — SIIMA Award for Best Actress Nominated — Filmfare Award for Best Actress - Kannada
2012	*18th Cross*	Punya	Kannada	
2012	*Sagar*	Kajal	Kannada	
2012	*Drama*	Nandini	Kannada	
2013	*Kaddipudi*	Uma	Kannada	
2013	*Dhwaja*	Preethi	Kannada	
2013	*Bahaddoor*	Anjali	Kannada	Filming
2014	*Mr. & Mrs. Ramachari*			Announced
2014	*Endendigu*			Filming

Question: how many total films have they already appeared in?
Qwen-VL-Chat††: 14
Ground Truth: 17

Figure 4. Failure cases of DoCo on document-related commonsense reasoning and mathematical calculations.