A Token-level Text Image Foundation Model for Document Understanding (Supplementary Materials)

1. Interactive Demo

As shown in Figures 1, 2, and 3, we provide more interactive examples, including natural scene images, documents, codes, charts, tables, and GUIs. For each scene, we provide two examples. The first column is the original image, the second to fourth columns are the corresponding visualizations of the selected BPE words within the image, and the last column shows the highlighted area of the image when the prompt is a space " ". As we observed,

1) Our foundation model, TokenFD, can distinguish text and background areas well. This means that when using the foundation model for downstream tasks, we can remove redundant background features at a very low cost;

2) For complex, dense, and small texts, TokenFD still precisely perceive, such as "picture"(Code), "f"(Code), "19"(Table), "P"(Table), *etc*. Our TokenFD also supports handwritten texts, such as "STE"(Document) and "USA"(Document). Additionally, our TokenFD can still capture punctuation marks, such as commas, periods, double quotes, *etc*. This means that our foundation model has the potential to be customized for retrieval-augmented generation tasks;

We will deploy TokenFD to the huggingface space to provide an interactive interface for users to experience.

2. VQA-based Text Parsing Tasks

Modality connectors act as the bridge between the visual foundation model (VFM) and the LLM. Previous MLLMs employ image-text pairs of natural images (*e.g.,* Conceptual Captions, LAION, COYO) to pre-train them. In the work, to endow our MLLM TokenVL with generality and comprehensive document understanding abilities, we follow DocOwl [27] to conduct modality alignment. It involves both structure-aware parsing tasks (recognizing full text, converting formulas into LaTeX, converting tables into markdown or LaTeX, and converting charts into CSV or markdown formats) and multi-grained text localization tasks (recognizing partial text within localization and visual text grounding). Specifically, we present an example to introduce them, as shown in Table 1. In this way, the pre-trained modality connector can understand the visual features of our VFM and better project them into the same feature space with the linguistic features of our LLM.

3. TokenIT Dataset

3.1. Data Source

To construct a comprehensive TokenIT dataset, we collect various types of data, including natural scene text images, documents (PDF, receipt, letter, note, report, code, *etc.*), tables, charts, and screenshot images (GUIs). The data sources are summarized in Table 2.

3.2. Data Generation

Next, we elaborate on the data construction pipeline for the TokenIT dataset, which involves four steps:

1) Text Image Segmentation. For natural scene text images, charts and tables, we fine-tune the SAM model [39] on datasets with character-level mask annotations and leverage the well-learned model to generate text masks, since these images are relatively complex and diverse in color and style. For PDFs and industrial documents, we conduct simple unsupervised clustering [33] to get their text masks, as these images have high contrast between foregrounds and backgrounds;

2) Text Recognition. We use the previous state-of-the-art method [21] to obtain the recognition results for all types, except for natural scene text images. As these natural scene datasets already provide text transcriptions, we adopt them directly;

3) Tokenizer. We choose the widely adopted BPE tokenizer [9] to split the language texts into multiple BPE tokens, where each token corresponds to a BPE-level subword;

4) Token-level Image Text Construction. After obtaining the text masks in Step 1, we apply the method [21] to produce character-level segmentation masks. Subsequently, we combine each token's corresponding character-level mask to create a complete token-level segmentation mask.

5) Data Correction. For each image and its generated labels following the above stage, we render the labels onto the images to verify data labeling quality and perform manual relabeling as needed. Finally, three rounds of inspections are conducted to minimize labeling errors, a process that took four months to develop the first token-level image text

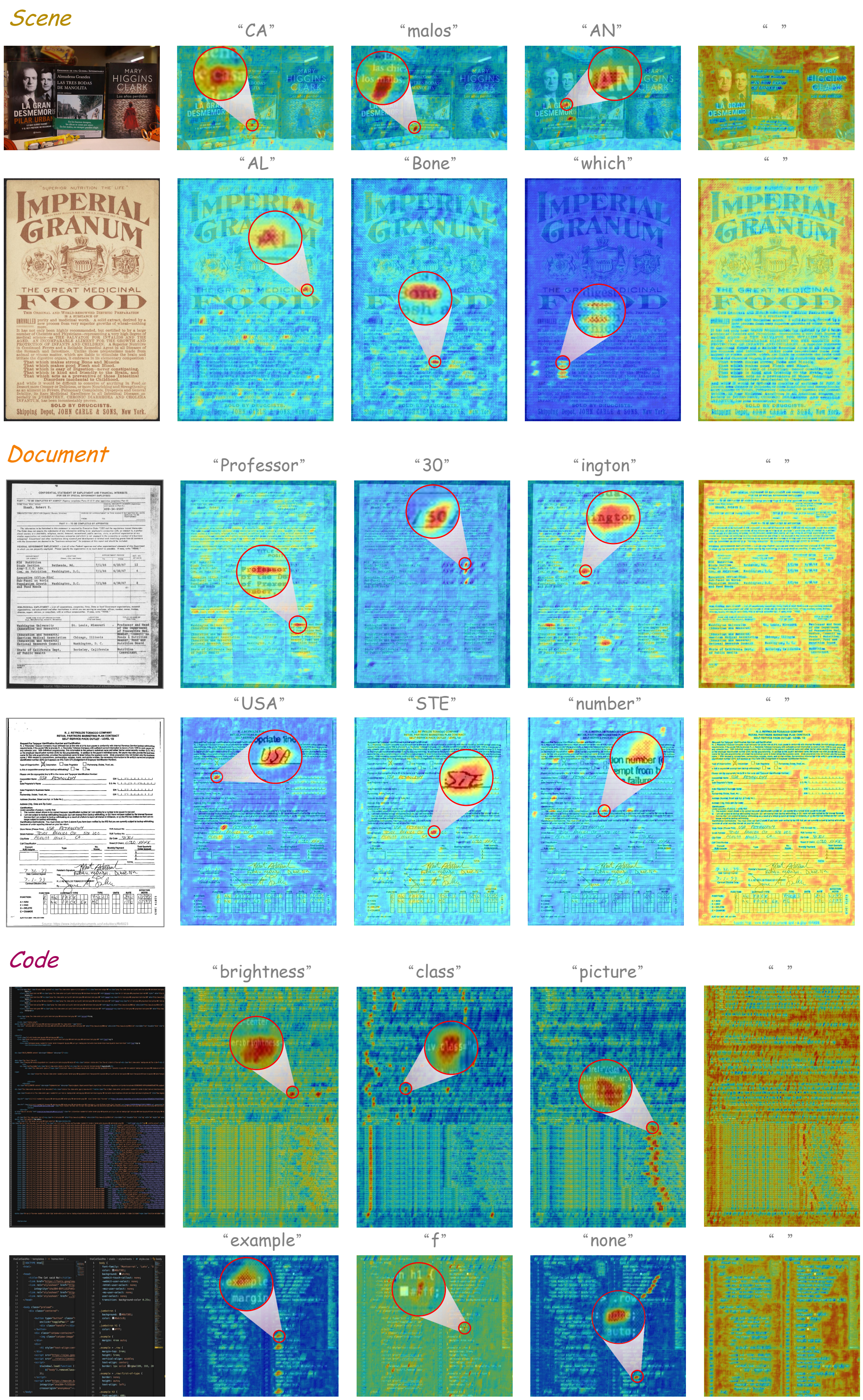


Figure 1. More visualization examples of the natural scene images, document images, and code images.

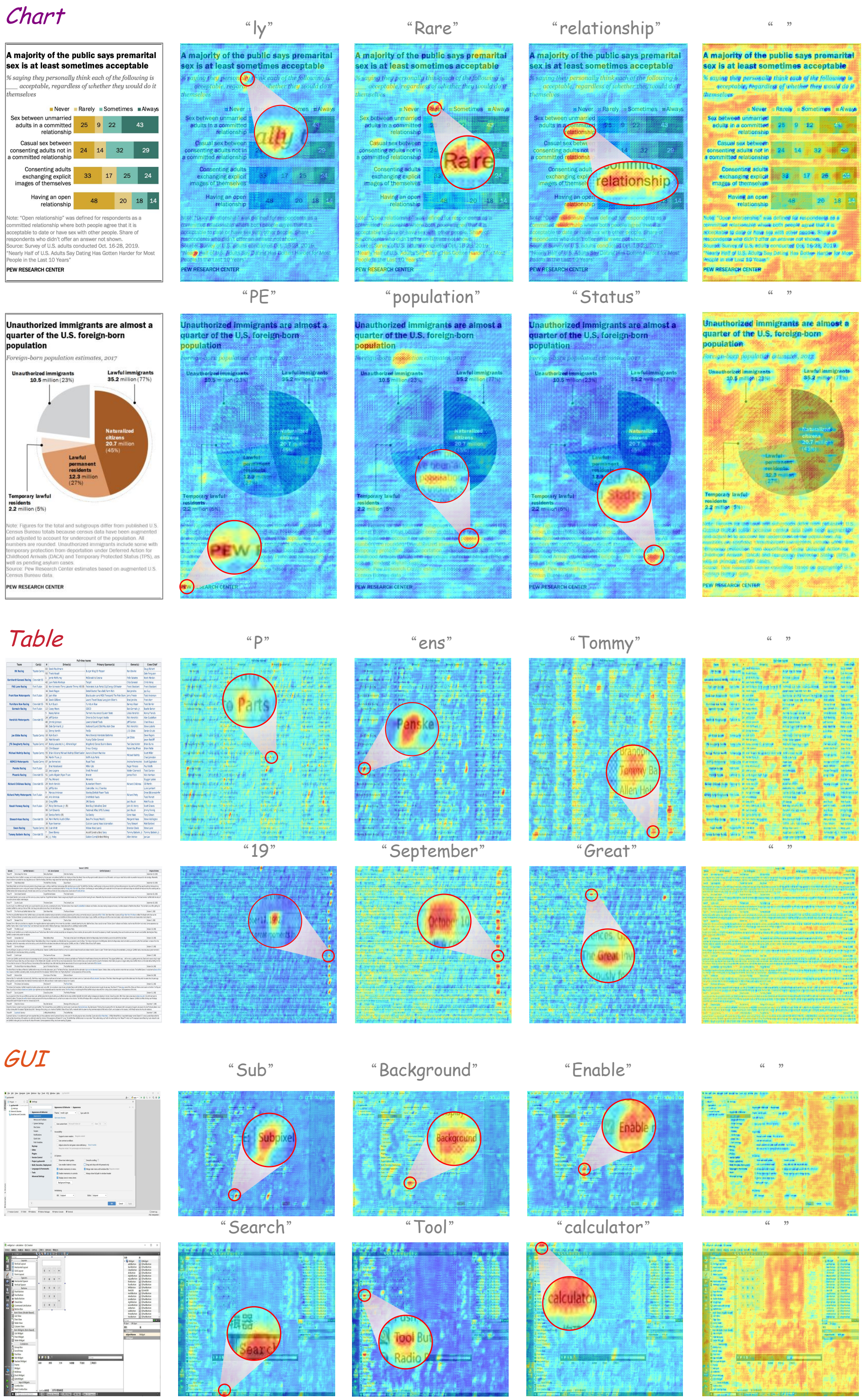


Figure 2. More visualization examples of the chart, table, and GUI images.

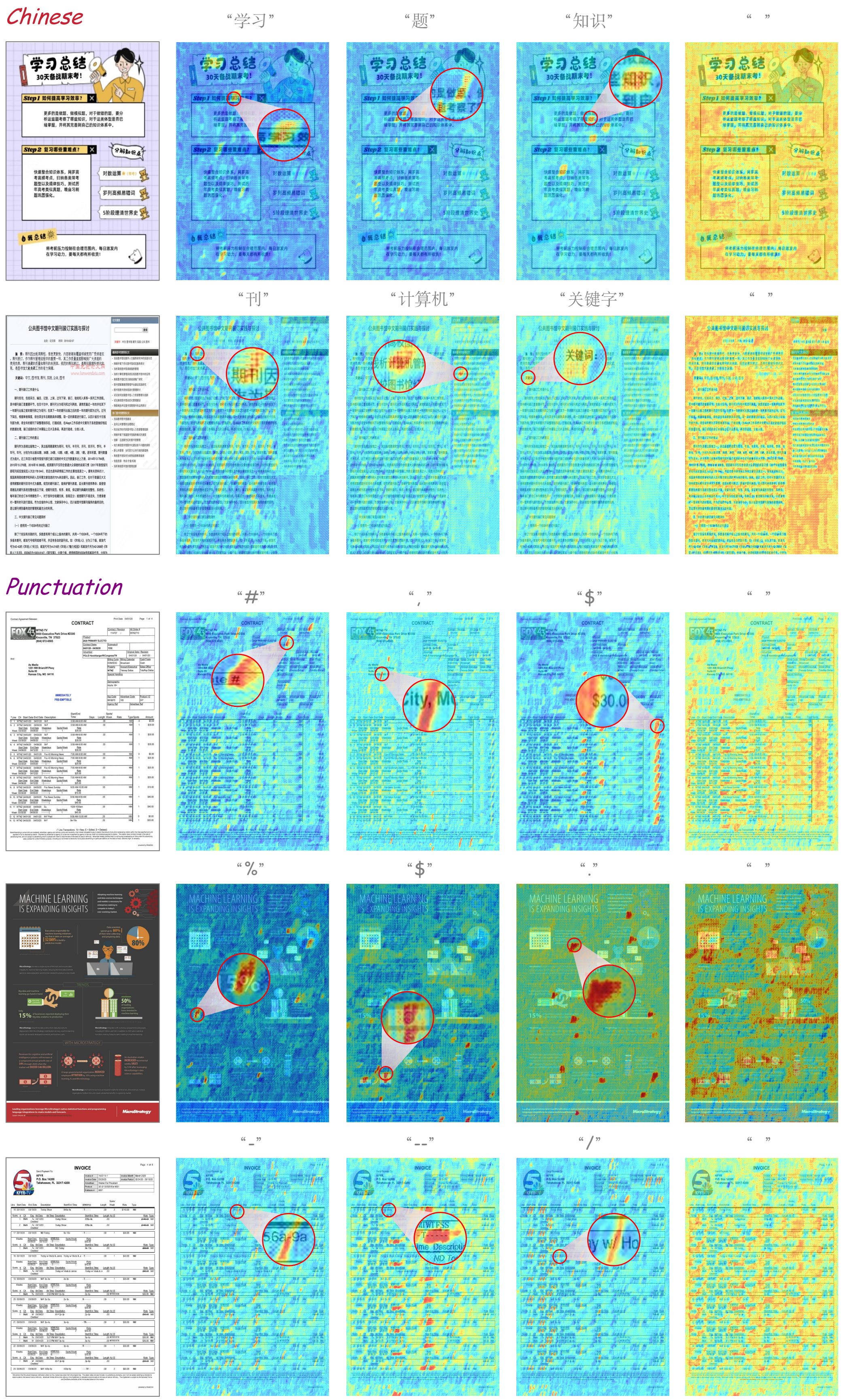
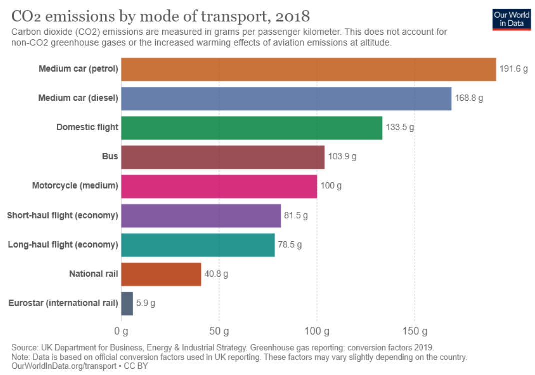


Figure 3. More visualization examples of the Chinese and Punctuation interaction.

VQA-based Text Parsing



Question: Recognize all texts in the image.

Answer: CO_2 emissions by mode of transport, 2018. Carbon dioxide (CO2) emissions are measured in grams per passenger kilometer. This does not account for non-CO2 greenhouse gases or the · · · · · · .

Question: Recognizing the text within the bounding box <bbox>75, 200, 160, 230</bbox>.

Answer: Medium car (diesel).

Question: Predict the bounding box of the text <ocr>Eurostar (International rall)</ocr>

Answer: <bbox>25, 520, 160, 550</bbox>.

Question: Converting the chart into CSV format.

Answer:
Mode of Transport, CO2 Emissions (g CO2 per passenger km)
Medium car (petrol), 191.6
Medium car (diesel), 168.8
Domestic flight, 133.5
Bus, 103.9
. . . . . .

Question: Converting the chart into Markdown format.

Answer:
Mode of Transport	CO2 Emissions (g CO2 per passenger km)
——————	——————
Medium car (petrol)	191.6
Medium car (diesel)	168.8
Domestic flight	133.5
Bus	103.9
. . . . . .

Table 1. The illustration of VQA-based Text Parsing tasks.

dataset (TokenIT).

Overall, the proposed TokenIT dataset includes 20 million images (including natural scene text images, documents, tables, charts, and GUIs) and 1780679833 (1.8 billion) token-mask pairs. Each BPE token corresponds one-to-one with a pixel-level mask. The number of token-mask pairs ultimately constructed is 4.5 times that of CLIP and 0.7B more than SAM.

4. Training Details

4.1. Text Segmentation

In this section, we evaluate the performance of text segmentation using TextSeg, COCOText, and HierText, which provide pixel-level annotations. The test sets of these datasets are utilized for zero-shot experiments. In the linear probe setting, all methods are trained on the combined three training sets and evaluated separately on each test set. The training configuration includes 70 epochs, a learning rate of 0.0001, a batch size of 6, and the optimizer AdamW.

4.2. Visual Question Answering

In this section, we evaluate the performance of visual document understanding using the test sets of DocVQA, InfoVQA, ChartQA, and TextVQA. Following LLava-1.5 [46], we build VFMs using TokenFD or other vision encoders based on the Vicuna-7B LLM [12]. The VFM are fixed during LLM training. The whole procedure includes two stages: pre-training and fine-tuning.

During the pre-training phase, we randomly sample 200,000 images each from the IIT-CDIP and DocMatix document datasets. Full-text recognition is implemented using PaddleOCR to generate ground-truth textual content, which serves as target answers. The model is trained with the instructional prompt "Recognize all texts in the image:" where only the Multilayer Perceptron (MLP) component receives parameter updates. The training configuration includes one epoch with a learning rate of 0.001 and a batch size of 24.

In the fine-tuning stage, the LLM is fine-tuned with Low-Rank Adaptation (LoRA) [29]. The training data consists of the training sets split from the previously mentioned QA evaluation datasets. This phase retains single-epoch training but employs modified hyperparameters—a reduced learning rate of 0.0002 and a batch size of 12—to ensure stable parameter convergence. This hierarchical training approach progressively enhances both text recognition accuracy and semantic comprehension capabilities in document understanding tasks.

4.3. Text Retrieval

In this section, we evaluate model performance using the CTR benchmark (English) [82] and the CSVTRv2 benchmark (Chinese) [85]. For English text retrieval, we employ the training sets from ICDAR2013, ICDAR2015, COCO-Text, MLT2017, OpenImagesV5Text, CTW1500, TotalText, HierText, and TextOCR. For Chinese text retrieval, we use ArT, ChineseOCR, HCCDoc, icdar2017rctw, LSVT, MTWI, and ReCTS as the training sets. These methods are optimized using the AdamW optimizer. The initial learning rate is 0.0001. We use a batch size of 6 and a number of training

Dataset Type	Dataset Name
Natural Scene	ICDAR2013 [35], COCOText [82], CTW1500 [95], HierText [57], ICDAR2015 [11], OCRCC [36], OpenImagesV5Text [40], TextCaps [74], TextOCR [75], TotalText [13], Laion-OCR [71], Wukong-OCR [20], MLT2017 [68], ocrvqa [66], ST-VQA [3], SynText [55], the-cauldron [43], ArT [14], ChineseOCR [15], HCCDoc [96], ICDAR2017rctw [72], LSVT [78], MTWI [23], and ReCTS [100]
Document	DocVQA [63], InfographicsVQA [64], KleisterCharity [76], PubTabNet [103], RVL-CDIP [22], VisualMRC [80], Docmatix [41], IIT-CDIP [92], publaynet [102], Synthdog-en [37], DocGenome [90], CCpdf [81]
Chart	ChartQA [62], FigureQA [32], PlotQA [65],TabMWP [58], DVQA [31]
Table	TableQA [77], DeepForm [79], TURL [16], TabFact [4], WikiTableQuestions [70]
GUI	Screen2Words [83], WebSight [42], OmniACT [34], SeeCliCK [10], Mind2Web [17]

Table 2. Data source of our TokenIT dataset.

epochs of 10. After the first 5 epochs, the initial learning rate is reduced to 0.00001.

5. Spatial-wise Alignment

The sequence-to-sequence auto-regression training allows language inputs to interact only implicitly with visual inputs, where the outputs may rely more on the LLM's robust semantic context capabilities, especially when generating very long tokens. Consequently, some research [61, 69] attempts to equip the model with spatial-wise capabilities, encouraging the LLM to reference image content more directly when responding to questions, rather than relying solely on its powerful semantic context capabilities. The task they proposed enhances the spatial-wise capabilities of MLLMs by integrating localization prompts or predicting coordinates. However, these methods are implicit and difficult for models to achieve a precise understanding of spatial alignment. In contrast, TokenVL provides a direct and explicit method by aligning answer tokens with their corresponding spatial image tokens to guide MLLMs. In this way, the model not only answers the question well, but also explicitly knows the spatial region in the image to which the answer corresponds. To compare these methods more intuitively, we use the same data to follow their spatial alignment task while conducting a VQA-based text parsing. Table 4 presents the final comparison results.

6. Mainstream Benchmark Results

General multi-modal large models [1, 5, 6, 49, 86, 97] typically use DocVQA, InfoVQA, ChartQA, and TextVQA to evaluate document understanding capabilities, as these benchmarks encompass diverse and comprehensive scenarios that reflect real-world applications. To compare performance intuitively and clearly, we collected data from nearly all MLLMs that reported scores on these four benchmarks and summarized them in Table 3. Specifically, we categorize the existing MLLMs into three types based on model size: "<2B", "<8B", and ">8B". Due to resource constraints, we did not conduct experiments with models exceeding 8B parameters in our TokenVL, providing only two versions: TokenVL-2B and TokenVL-8B. Notably, our TokenVL-2B improves upon the previous state-of-the-art (SOTA) result by 1.32%, and our TokenVL-8B improves by 0.63%. Compared to models with larger parameters, our 8B version slightly surpasses DeepSeek-VL2-16B and InternVL2-40B by 0.3%.

7. More examples compared to other MLLMs

As shown in Figure 4, we present more qualitative visualization results to demonstrate TokenVL's capabilities in various VQA tasks. TokenVL analyzes the question, identifies the key elements in the image relevant to answering the question.

8. Why compare with SAM/CLIP?

We compare them for two reasons:1) Prior works use them as VFMs due to the lack of domain-specific ones. We close the gap by developing TokenFD (the first token-level VFM) comparable to them. Thus the comparison will highlight the significance of developing TokenFD. 2) Data used to train CLIP/SAM also includes natural scene text images, making our comparisons in retrieval/segmentation/TextVQA tasks reasonable. In addition, similar to other VLMs' visual encoders, SAM is commonly used as the encoder in MLLMs (*e.g.,* Vary and Deepseek-vl).

9. Less Token

Even when using fewer visual tokens for testing, TokenVL still achieves robust results 5.

Size	Model	Visual Encoder	LLM Decoder	DocVQA	InfoVQA	ChartQA	TextVQA	Avg.
<2B	DocLLM-1B [84]	-	Falcon-1B	61.4	-	-	-	-
	Mini-Monkey [30]	InternViT-300M	InternLLM2-2B	87.4	60.1	76.5	75.7	74.93
	MM1.5-1B [97]	CLIP-ViT-H	Private	81.0	50.5	67.2	72.5	67.80
	MM1.5-3B [97]	CLIP-ViT-H	Private	87.7	58.5	74.2	76.5	74.23
	InternVL2-1B [7]	InternViT-300M	Qwen2-0.5B	81.7	50.9	72.9	70.5	69.00
	InternVL2-2B [7]	InternViT-300M	InternLM2-1.8B	86.9	58.9	76.2	73.3	73.83
	LLaVA-OneVision-0.5B [46]	SigLIP	qwen2-0.5B	70.0	41.8	61.4	-	-
	InternVL2.5-1B [6]	InternViT-300M	Qwen2.5-0.5B	84.8	56.0	75.9	72.0	72.18
	InternVL2.5-2B [6]	InternViT-300M	InternLM2.5-1.8B	88.7	60.9	79.2	74.3	75.78
	TokenVL-2B	TokenFD	InternLM2.5-1.8B	**89.9**	**61.0**	**81.1**	**76.4**	**77.10**
<8B	UReader [93]	CLIP-ViT-L/14	LLaMA-7B	65.4	42.2	59.3	57.6	56.13
	DocLLM-7B [84]	-	LLaMA2-7B	69.5	-	-	-	-
	Cream [38]	CLIP-ViT-L/14	Vicuna-7B	79.5	43.5	63.0	-	-
	Qwen-VL [2]	ViT-bigG	Qwen-7B	65.1	35.4	65.7	63.8	57.50
	LLaVA-1.5-7B [53]	CLIP-ViT-L	Vicuna1.5-7B	-	-	-	58.2	-
	SPHINX [51]	CLIP-ViT+CLIP-ConvNext+DINOv2-ViT	LLaMA2-7B	-	-	-	61.2	-
	LLaVA-OneVision [45]	SigLIP	Qwen2-7B	87.5	68.8	80.0	-	-
	Monkey [48]	Vit-BigG	Qwen-7B	66.5	36.1	65.1	67.6	58.83
	TextMonkey [56]	Vit-BigG	Qwen-7B	73.0	-	66.9	65.6	-
	IDEFICS2 ([44])	SigLIP-SO400M	Mistral-7B	74.0	-	-	73.0	-
	LayoutLLM [59]	LayoutLMv3-large	Vicuna1.5-7B	74.25	-	-	-	-
	DocKylin [98]	Swin	Qwen-7B	77.3	46.6	66.8	-	-
	DocLayLLM [50]	LayoutLMV3	LLaMA3-8B	77.79	42.02	-	-	-
	mPLUG-DocOwl [25]	CLIP-ViT-L/14	LLaMA-7B	62.2	38.2	57.4	52.6	52.60
	mPLUG-DocOwl1.5 [26]	CLIP-ViT-L/14	LLaMA2-7B	82.2	50.7	70.2	68.6	67.93
	mPLUG-DocOwl2 [28]	CLIP-ViT-L/14	LLaMA2-7B	80.7	46.4	70.0	66.7	65.95
	Vary [88]	CLIP-ViT-L/14 + SAM	Qwen-7B	76.3	-	66.1	-	-
	Eagle [73]	CLIP + ConvNeX + Pix2Struct + EVA2 + SAM	LLaMA3-8B	86.6	-	80.1	77.1	-
	PDF-WuKong [91]	CLIP-ViT-L-14	InernLM2-7B	85.1	61.3	80.0	-	-
	TextHawk2 [94]	SigLIP	Qwen2-7B	89.6	67.8	81.4	75.1	78.48
	MM1.5-7B [97]	CLIP-ViT-H	Private	88.1	59.5	78.6	76.5	75.68
	HRVDA [52]	Swin-L	LLaMA2-7B	72.1	43.5	67.6	73.3	64.13
	InternVL2-4B [7]	InternViT-300M	Phi-3-mini	89.2	67.0	81.5	74.4	78.03
	InternVL2-8B [7]	InternViT-300M	InternLM2.5-7B	91.6	74.8	83.3	77.4	81.78
	InternVL2.5-4B [6]	InternViT-300M	Qwen2.5-3B	91.6	72.1	84.0	76.8	81.13
	InternVL2.5-8B [6]	InternViT-300M	InternLM2.5-7B	93.0	**77.6**	84.8	79.1	83.63
	InternVL2.5-8B-mpo[87]†	InternViT-300M	InternLM2.5-7B	92.3	76.0	83.8	79.1	82.80
	DeepSeek-VL2-3B [89]	SigLIP-SO400M-384	DeepSeekMoE	88.9	66.1	81.0	80.7	79.18
	DocPeida [19]	Swin	Vicuna-7B	47.1	15.2	46.9	60.2	42.35
	TokenPacker-7B [47]	CLIP-ViT-L/14	Vicuna-7B	60.2	-	-	-	-
	LLaVA-OneVision-7B [46]	SigLIP	qwen2-7B	87.5	68.8	80.0	-	-
	DocVLM [67]	CLIP-ViT-G/14 + DocFormerV2	Qwen2-7B	92.8	66.8	-	**82.8**	-
	TokenVL-8B	TokenFD(323M)	InternLM2.5-7B	**94.2**	76.5	**86.6**	79.9	**84.30**
>8B	LLaVA-13B [54]	CLIP-ViT-L/14	Vicuna-13B	6.9	-	-	36.7	-
	PaLI-X [5]	ViT-22B	UL2-32B	86.8	54.8	72.3	80.8	73.68
	LLaVAR [101]	CLIP-ViT-L/14	Vicuna-13B	11.6	-	-	48.5	-
	LLaVA-1.5-13B [53]	CLIP-ViT-L	Vicuna1.5-13B	-	-	-	62.5	-
	CogAgent [24]	EVA2-CLIP+CogVLM +Cross Attention	Vicuna-13B	81.6	44.5	68.4	76.1	67.65
	Unidoc [18]	CLIP-ViT-L/14	Vicuna-13B	90.2	36.8	70.5	73.7	67.80
	MM1.5-30B [97]	CLIP-ViT-H	Private	91.4	67.3	83.6	79.2	80.38
	InternVL1.5-26B [8]	InternViT-6B	InternLM2-20B	90.9	72.5	83.8	80.6	81.95
	InternVL2-26B [7]	InternViT-6B	InternLM2-20B	92.9	75.9	84.9	82.3	84.00
	InternVL2-40B [7]	InternViT-6B	Nous-Hermes-2-Yi-34B	93.9	78.7	86.2	83.0	85.45
	InternVL2.5-26B [6]	InternViT-6B	InternLM2.5-20B	94.0	79.8	87.2	82.4	85.85
	InternVL2.5-38B [6]	InternViT-6B	Qwen2.5-32B	**95.3**	83.6	88.2	82.7	87.45
	InternVL2.5-78B [6]	InternViT-6B	Qwen2.5-72B	95.1	**84.1**	**88.3**	83.4	**87.73**
	TinyChart [99]	SigLIP	Phi-2	-	-	83.6	-	-
	TokenPacker-13B [47]	CLIP-ViT-G/14	Vicuna-13B	70.0	-	-	-	-
	DeepSeek-VL2-16B [89]	SigLIP-SO400M-384	DeepSeekMoE	92.3	75.8	84.5	83.4	84.00
	DeepSeek-VL2-27B [89]	SigLIP-SO400M-384	DeepSeekMoE	93.3	78.1	86.0	**84.2**	85.40

Table 3. Comparison results on four widely evaluated datasets. † refers to our evaluation result using the official checkpoint.

10. Future Directions

In this paper, we use some simple prompts to explore the effectiveness of the visual foundation model, TokenFD, in fine-grained scene tasks, including segmentation, retrieval, recognition, and understanding. In the future, we hope to explore more complex applications based on tokenFD, such as multimodal RAG, controllable text erasure, controllable text generation, and general image understanding.

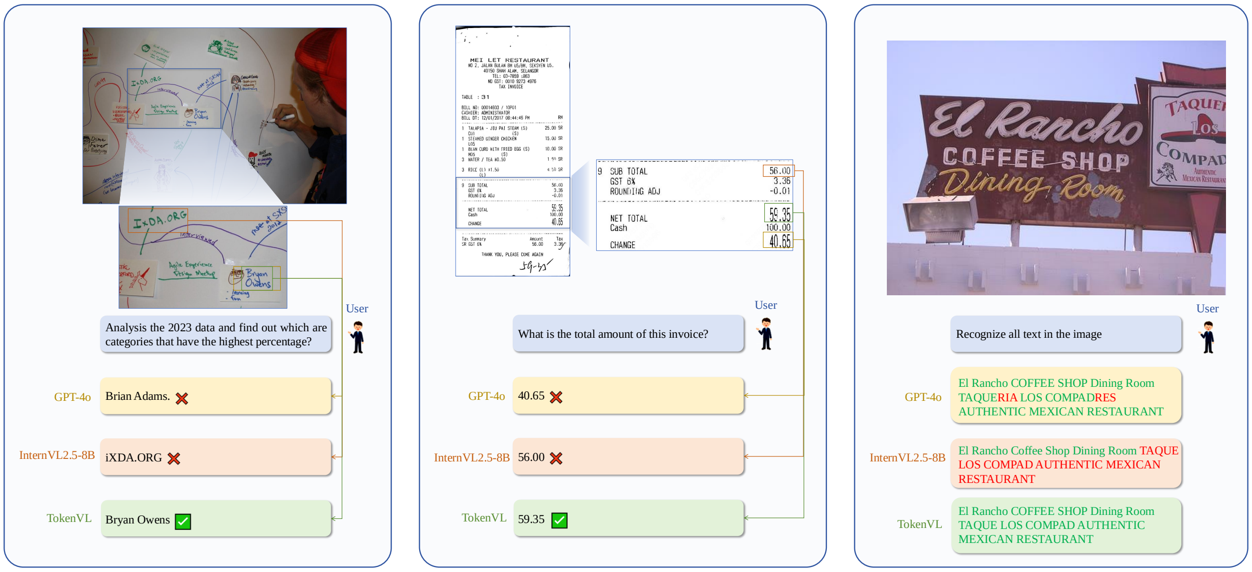


Figure 4. Visualization of TokenVL's comparison with GPT-4o, internvl2.5-8B on VQA tasks.

Method	IIT↓	Docgenome↓	IC15↓	TotalText↓
Park *et al.* [69]	39.21	38.63	48.20	65.66
Kosmos2.5 [60]	32.75	36.17	34.22	53.62
TokenVL	**19.21**	**22.54**	**23.24**	**35.47**

Table 4. Edit distance for full-image text recognition.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 6

[2] 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. 7

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 6

[4] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019. 6

[5] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 6, 7

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 7

[7] Zhe Chen, Weiyun Wang, and et al. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. 2024. 7

[8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 7

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1

[10] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024. 6

[11] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084, 2017. 6

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 5

Model	DocOwl	TextMonkey	InternVL2.5	TokenVL	DocPeida	DocOwl1.5	MM1.5	InternVL2.5	TokenVL	KOSMOS2.5	TextHawk2	InternVL2.5	TokenVL
Token/Score	841/52.6	768/58.6	768/68.0	768/68.4	1600/42.4	1698/67.9	1440/75.7	1536/81.4	1536/82.4	2048/56.35	2304/78.5	3133/83.6	3133/84.3

Table 5. The token number per image of testing. "Token" is the approximate average number of tokens per image. "Score" refers to the average score on Doc/Info/Chart/TextVQA datasets.

[13] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, pages 935–942. IEEE, 2017. 6

[14] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 6

[15] Ltd. Beijing Anjie Zhihe Technology Co. Chinese ocr. 2024. 6

[16] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022. 6

[17] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[18] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*, 2023. 7

[19] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. DocPedia: unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv*, 2311.11810, 2024. 7

[20] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022. 6

[21] Tongkun Guan, Wei Shen, and Xiaokang Yang. CCDPlus: Towards accurate character to character distillation for text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[22] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*. 6

[23] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. Icpr2018 contest on robust reading for multi-type web images. In *2018 24th international conference on pattern recognition (ICPR)*, pages 7–12. IEEE, 2018. 6

[24] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents, 2023. 7

[25] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5:unified structure learning for OCR-free document understanding. *arXiv*, 2403.12895, 2024. 7

[26] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 7

[27] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 1

[28] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl2: high-resolution compressing for OCR-free multi-page document understanding. *arXiv*, 2409.03420, 2024. 7

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[30] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey:alleviating the semantic sawtooth effect for lightweight MLLMs via complementary image pyramid. *arXiv*, 2408.02034, 2024. 7

[31] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 6

[32] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 6

[33] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE TPAMI*, 24(7):881–892, 2002. 1

[34] Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem AlShikh, and Ruslan Salakhutdinov. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, pages 161–178. Springer, 2024. 6

[35] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan

Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 6

[36] Sameem Abdul Kareem, Pilar Pozos-Parra, and Nic Wilson. An application of belief merging for the diagnosis of oral cancer. *Applied Soft Computing*, 61:1105–1112, 2017. 6

[37] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 6

[38] Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. Visually-situated natural language understanding with contrastive reading model and frozen large language models. *arXiv preprint arXiv:2305.15080*, 2023. 7

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1

[40] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. Open images v5 text annotation and yet another mask text spotter. In *Asian Conference on Machine Learning*, pages 379–389. PMLR, 2021. 6

[41] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024. 6

[42] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024. 6

[43] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 6

[44] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 7

[45] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 7

[46] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5, 7

[47] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm, 2024. 7

[48] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey:image resolution and text label are important things for large multi-modal models. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[49] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pages 26763–26773, 2024. 6

[50] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024. 7

[51] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 7

[52] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. Hrvda: High-resolution visual document assistant. In *CVPR*, pages 15534–15545, 2024. 7

[53] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 7

[54] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 7

[55] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network, 2020. 6

[56] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. TextMonkey: an OCR-Free large multimodal model for understanding document. *arXiv*, 2403.04473, 2024. 7

[57] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *CVPR*, pages 1049–1059, 2022. 6

[58] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 6

[59] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *CVPR*, pages 15630–15640, 2024. 7

[60] Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023. 8

[61] Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. KOSMOS-2.5: a multimodal literate model. *arXiv*, 2309.11419, 2024. 6

[62] 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. 6

[63] 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. 6

[64] 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. 6

[65] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 6

[66] 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. 6

[67] Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. Docvlm: Make your vlm an efficient reader, 2024. 7

[68] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, pages 1454–1459. IEEE, 2017. 6

[69] Jaeyoo Park, Jin Young Choi, Jeonghyung Park, and Bohyung Han. Hierarchical visual feature aggregation for ocr-free document understanding. *arXiv preprint arXiv:2411.05254*, 2024. 6, 8

[70] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 6

[71] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6

[72] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, pages 1429–1434. IEEE, 2017. 6

[73] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders, 2024. 7

[74] 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. 6

[75] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pages 8802–8812, 2021. 6

[76] 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. 6

[77] Ningyuan Sun, Xuefeng Yang, and Yunfeng Liu. Tableqa: a large-scale chinese text-to-sql dataset for table-aware sql generation. *arXiv preprint arXiv:2006.06434*, 2020. 6

[78] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019. 6

[79] S Svetlichnaya. Deepform: Understand structured documents at scale. 2020. 6

[80] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 6

[81] Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. Ccpdf: Building a high quality corpus for visually rich documents from web crawl data. In *International Conference on Document Analysis and Recognition*, pages 348–365. Springer, 2023. 6

[82] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 5, 6

[83] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510, 2021. 6

[84] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. DocLLM: a layout-aware generative language model for multimodal document understanding. *arXiv*, 2401.00908, 2023. 7

[85] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4558–4567, 2021. 5

[86] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[87] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reason-

ing ability of multimodal large language models via mixed preference optimization, 2024. 7

[88] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024. 7

[89] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 7

[90] Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024. 6

[91] Xudong Xie, Liang Yin, Hao Yan, Yang Liu, Jing Ding, Minghui Liao, Yuliang Liu, Wei Chen, and Xiang Bai. PDF-WuKong: a large multimodal model for efficient long PDF reading with end-to-end sparse sampling. *arXiv*, 2410.05970, 2024. 7

[92] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: pre-training of text and layout for document image understanding. In *Knowledge Discovery and Data Mining*, 2019. 6

[93] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. 7

[94] Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*, 2024. 7

[95] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017. 6

[96] Hesuo Zhang, Lingyu Liang, and Lianwen Jin. Scut-hccdoc: A new benchmark dataset of handwritten chinese text in unconstrained camera-captured documents. *Pattern Recognition*, 108:107559, 2020. 6

[97] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 6, 7

[98] Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*, 2024. 7

[99] Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*, 2024. 7

[100] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE, 2019. 6

[101] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 7

[102] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 6

[103] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020. 6