DREAM: Document Recognition with Explicit Adaptive Memory

Supplementary Material

Table 6. Experiment results on the DreamDoc dataset

Method	Edit Distance↓	F1-score↑	Precision↑	Recall↑	BLEU↑	METEOR↑
GOT [43]	0.473	0.744	0.723	0.794	0.354	0.657
MonkeyOCR [27]	0.435	0.798	0.770	0.848	0.414	0.641
DeepSeekOCR [44]	0.384	0.801	0.770	0.857	0.528	0.666
Paddle-VL [8]	0.353	0.823	0.812	0.855	0.521	0.724
DREAM(Ours)	**0.248**	**0.869**	**0.863**	**0.893**	**0.735**	**0.819**

7. Additional Comparison Results

This section provides additional results comparing the proposed document recognition model with other methods on DreamDoc dataset and OmniDocBench Dataset, complementing Section 4.1.3.

7.1. Comparison With Other Methods on DreamDoc Dataset

To validate the effectiveness of our proposed method on the self-built DreamDoc dataset, we evaluate not only GOT [43] but also several mainstream VLM-based document recognition models, including MonkeyOCR [27], DeepSeekOCR in the Gundam mode [44], and Paddle-VL [8]. The experimental results on DreamDoc test set are summarized in Table 6, where bold indicates the best performance and underline indicates the second-best.

Across all evaluated models, our DREAM model achieves the state-of-the-art performance on the DreamDoc dataset.

7.2. Comparison With Other Methods on OmniDocBench Dataset

To further validate our methods on public dataset, we conducted additional experiments on OmniDocBench dataset [31]. The OmniDocBench dataset is a comprehensive benchmark for evaluating OCR performance across diverse real-world document types. It comprises 1355 PDF pages spanning 9 document types, 4 layout styles, and 3 language categories, offering a diverse testbed that covers both structured and unstructured documents.

We compare our method with several mainstream VLM-based document recognition models, including expert VLMs such as GOT [43] and Nougat [5], as well as general-purpose VLMs such as GPT-4o [18], InternVL2-76B [7], and Qwen2-VL-72B [40]. Evaluation is conducted using the edit distance metric across 9 document categories. The results are reported in Table 7.

The results show that our DREAM model achieves overall performance ranking closely behind the best model. Notably, our method achieves the best performance in the Newspaper category, a document type characterized by highly complex layouts and diverse font styles. That demonstrates that our prototype memory module effectively captures these factors.

8. Additional Ablation Study

This subsection provides additional details complementing the ablation results presented in Sec. 4.1.4.

We assume that local regions of the visual feature map contain multiple compositional factors related to spatial layout structures and visual styles. We refer to each independent contributing factor as a prototype, and model them using our explicit prototype memory module. In the proposed DREAM model, the prototype memory module consists of a set of memory slots, each storing a prototype representation that is updated via an Exponential Moving Average (EMA) mechanism during training.

However, the number of possible layout structures and stylistic patterns in real documents is large, and thus determining an appropriate capacity for the prototype memory is important. Table 8 in this section reports our ablation study on the number of prototype memory slots (i.e., memory size) used in the multiscale memory module within the document recognition model described in Sec. 4.1.

The results show that the prototype memory model achieves the best performance when the memory size is set to 2048. This is likely because the spatial structures in documents exhibit substantial complexity. When the memory size increases to 4096, however, the performance drops. A possible reason is that an excessively large number of memory slots makes it difficult for the model to assign a semantically independent prototype to each slot, and satisfy the sparsity regularization loss $\mathcal{L}_{\text{sparse}}$. Therefore, for all experiments presented in the previous sections of this paper, we adopt this best-performing memory configuration and fix the memory size to 2048.

Table 7. Experiment results on the OmniDocBench dataset, evaluation using edit distance

Method	Book	Slides	Financial Report	Textbook	Exam Paper	Magazine	Academic Papers	Notes	Newspaper	Overall
GOT [43]	0.111	0.222	0.067	0.132	0.204	0.198	0.179	0.388	0.771	0.267
Nougat [5]	0.734	0.958	1.000	0.820	0.930	0.83	0.214	0.991	0.871	0.806
GPT4o [18]	0.157	0.163	0.348	0.187	0.281	0.173	0.146	0.607	0.751	0.316
InternVL2-76B [7]	0.216	0.098	0.162	0.184	0.247	0.150	0.419	0.226	0.903	0.3
Qwen2-VL-72B [40]	**0.096**	**0.061**	**0.047**	0.149	**0.195**	**0.071**	**0.085**	**0.168**	0.676	**0.179**
DREAM(Ours)	0.123	0.115	0.144	**0.128**	0.293	0.096	0.278	0.239	**0.574**	0.231

Table 8. Ablation study of different memory size on the Fox dataset

Memory Size	Edit Distance↓		F1-score↑		Precision↑		Recall↑		BLEU↑		METEOR↑	
	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN
256	0.119	**0.080**	0.949	0.939	0.961	**0.933**	0.940	0.947	0.756	0.906	0.868	0.916
512	0.145	0.103	0.941	0.934	0.950	0.927	0.934	0.945	0.736	0.890	0.850	0.910
1024	0.106	0.104	0.949	0.927	0.959	0.920	0.939	0.939	0.759	0.888	0.871	0.917
2048	**0.101**	0.082	**0.964**	**0.939**	**0.974**	0.933	**0.953**	**0.955**	**0.767**	**0.909**	**0.881**	**0.942**
4096	0.114	0.122	0.939	0.901	0.950	0.895	0.932	0.918	0.759	0.868	0.866	0.895

9. Additional Visualization

In Section 4.1.5, we visualized how the prototype memory module responds during the cross-attention routing process. For each patch in the image feature map, we displayed the index of the prototype with the highest attention weight, indicating which prototype the patch is most strongly associated with.

However, a patch may be influenced by multiple underlying prototypical factors. To illustrate how our model captures this phenomenon, we conducted an additional visualization experiment: for each patch, we plot the full attention-weight distribution over prototypes. Figure 5 presents the attention distributions of two representative patches.

The results show that both patches exhibit strong peaks at prototype indices 2 and 15, indicating that they respond most strongly to the same two prototypes. Both patches span a table boundary—blue background in the upper region and white background in the lower region—suggesting that these two prototypes jointly encode this type of table-edge structural pattern. At the same time, their attention distributions differ: the patch on the right activates a larger number of prototypes with non-zero weights, likely because it contains more textual content, which triggers additional text-style-related prototypes.

10. DreamDoc Dataset

10.1. Dataset Overview

Our self-built DreamDoc Dataset is a dataset for document recognition contains 7 categories, containing 4800 pages for trainning and 108 pages for testing.

10.2. Data Collection and Filtering

To construct the DreamDoc Dataset, we collect documents from a diverse set of real-world sources that reflect common scenarios encountered in document recognition tasks.

We obtain listed-company announcements and publicly disclosed reports from open regulatory platforms. University textbooks, middle and high school slides, and primary/secondary school textbooks are sourced from campus-internal digital repositories and publicly available electronic teaching materials. Newspapers and magazines are collected from publicly accessible digital archives and online reading platforms. Handwritten notes are sourced from naturally occurring student note-taking materials.

To ensure high-quality images, we remove pages with duplicated content, corrupted files, extreme blur, incomplete scans, or low-information pages (e.g., blank or near-blank). After filtering, we manually verify a subset of the data to ensure content integrity and category correctness.

10.3. Annotation Protocol

We adopt a semi-automatic annotation pipeline that combines OCR-based text extraction with manual verification to ensure both efficiency and high-quality annotations.

Automatic Text Extraction: We used Chandra [9] to process the images and obtain the OCR results as an original annotation.

Human Verification and Correction: To guarantee annotation accuracy, all OCR-generated text is manually

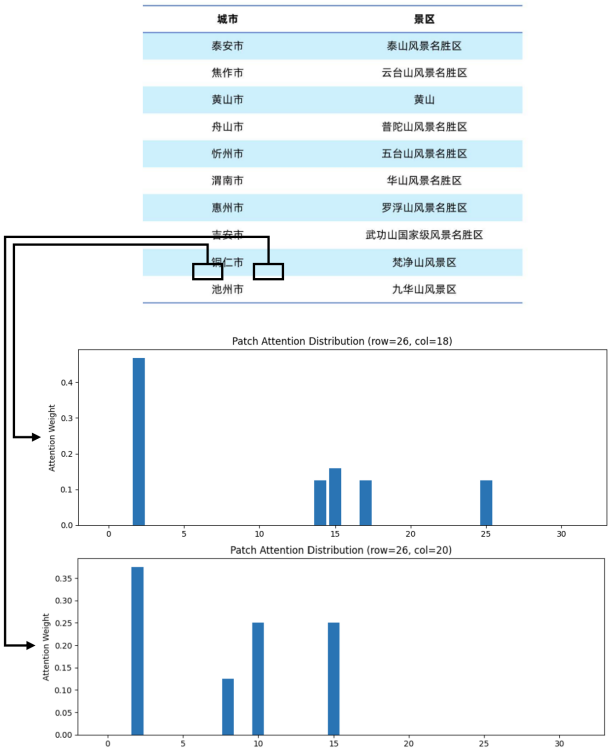


Figure 5. Additional visualization of attention weights for the prototype memory at scale $M^{(32)}$. The top shows a local region of the original image, and the bottom displays the attention weights distributions for two image patches. For clarity, the memory size is set to 32.

checked by trained annotators. The annotators correct recognition errors such as missing characters, hallucinated words, incorrect punctuation, and mis-segmented lines. Special attention is given to pages containing dense typesetting, complex layouts (e.g., tables, multi-column text), or handwriting, where OCR tends to be less reliable. Annotators also ensure that text ordering follows the natural reading flow of the document.

10.4. Dataset Statistics

As depicted in Fig. 6, the entire DreamDoc dataset (combining training and test sets) consists of 4,908 images. Specifically, there are 1,824 images of listed-company announcements, 1,016 images of university textbooks, 184 images of newspapers, 274 images of handwritten notes, 424 images of middle and high school slides, 1,084 images of primary/secondary school textbooks, and 102 images of magazines.

Among these, the listed-company announcements, newspapers, handwritten notes, middle and high school slides, and primary/secondary school textbooks are in Chinese, totaling 3,790 images, while the university textbooks and magazines

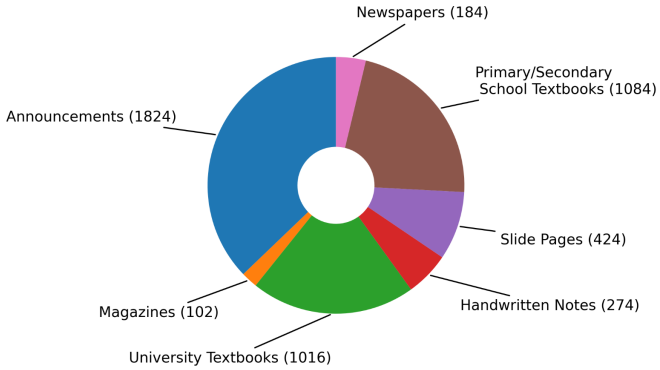


Figure 6. Category distribution of the DreamDoc dataset. The dataset contains seven document categories.

are in English, totaling 1,118 images.

10.4.1. Representative Sample Images

Representative sample images from all seven categories are shown in Fig. 7. As illustrated, our dataset covers a wide range of layout styles, including single-column, multi-column, and highly complex layouts, as well as tables. It also exhibits diverse font styles and includes handwritten text.

公司于 2025 年 3 月 18 日、2025 年 9 月 25 日与重整投资人北京华基科技有限公司、云南国际信托有限公司（代表“云南信托-云智 235 号单一资金信托”）、吉富马类投资（广东）有限公司、北京汇智云创企业管理中心（有限合伙）、北京福石重整管理咨询有限公司、深圳市招商平安资产管理有限公司、国民信托有限公司（代表“国民信托-东易日盛重整财务投资单一资金信托”）、长城（天津）股权投资基金管理有限责任公司、上海宏翼私募基金管理有限公司-宏翼红林 2 号私募证券投资基金、北京大有星瀚企业管理中心（有限合伙）、四川资本市场纾困发展证券投资基金合伙企业（有限合伙）及西银重整企业管理（北京）有限公司分别签署了重整投资协议。

公司于 2025 年 11 月 6 日与深圳市招商平安资产管理有限公司（以下简称“招商平安资产”）、深圳市招平珠光二号投资合伙企业（有限合伙）（以下简称“招平珠光二号”）签署协议书，与长城（天津）股权投资基金管理有限责任公司（以下简称“长城股权基金”）、芜湖长财投资中心（有限合伙）（以下简称“芜湖长财”）签署协议书。招商平安资产、长城股权基金的基本情况详见公司 2025 年 9 月 27 日编号 2025-065《关于通达重整投资人结果及签署重整投资协议暨预重整进展的公告》，招平珠光二号、芜湖长财的基本情况及协议书主要内容具体如下：

（一）招平珠光二号的基本情况

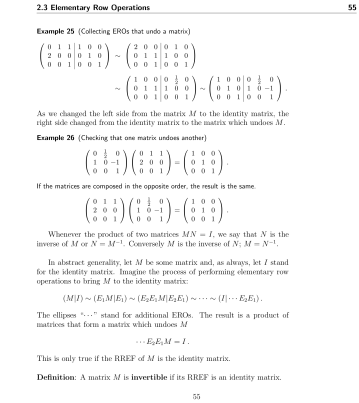
1、基本工商信息

企业名称	深圳市招平珠光二号投资合伙企业（有限合伙）
统一社会信用代码	91440300MAEGJ9FT8N
成立日期	2025-4-17
注册地址	深圳市南山区招商街道水湾社区太子路 51 号太子广场 3701
注册资本	14005 万（元）人民币
执行事务合伙人	深圳市平盈投资有限公司
经营范围	以自有资金从事投资活动。（除依法须经批准的项目外，凭营业执照依法自主开展经营活动）
出资结构	见下表

出资结构：

序号	合伙人名称	拟认缴出资额（万元）	持股比例

(a) Announcement



2.3 Elementary Row Operations 55

Example 25 (Collecting EROs that undo a matrix)

$$\begin{pmatrix}0&1&1&1&0&0\\2&0&0&0&1&0\\0&0&1&0&0&1\end{pmatrix}\sim\begin{pmatrix}2&0&0&0&1&0\\0&1&1&1&0&0\\0&0&1&0&0&1\end{pmatrix}\sim\begin{pmatrix}1&0&0&0&\frac{1}{2}&0\\0&1&1&1&0&0\\0&0&1&0&0&1\end{pmatrix}\sim\begin{pmatrix}1&0&0&0&\frac{1}{2}&0\\0&1&0&1&0&-1\\0&0&1&0&0&1\end{pmatrix}.$$

As we changed the left side from the matrix M to the identity matrix, the right side changed from the identity matrix to the matrix which undoes M.

Example 26 (Checking that one matrix undoes another)

$$\begin{pmatrix}0&\frac{1}{2}&0\\1&0&-1\\0&0&1\end{pmatrix}\begin{pmatrix}0&1&1\\2&0&0\\0&0&1\end{pmatrix}=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}.$$

If the matrices are composed in the opposite order, the result is the same.

$$\begin{pmatrix}0&1&1\\2&0&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&\frac{1}{2}&0\\1&0&-1\\0&0&1\end{pmatrix}=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}.$$

Whenever the product of two matrices $MN = I$, we say that N is the inverse of M or $N = M^{-1}$. Conversely M is the inverse of N; $M = N^{-1}$.

In abstract generality, let M be some matrix and, as always, let I stand for the identity matrix. Imagine the process of performing elementary row operations to bring M to the identity matrix:

$$(M|I) \sim (E_1M|E_1) \sim (E_2E_1M|E_2E_1) \sim \cdots \sim (I|\cdots E_2E_1).$$

The ellipses "$\cdots$" stand for additional EROs. The result is a product of matrices that form a matrix which undoes M

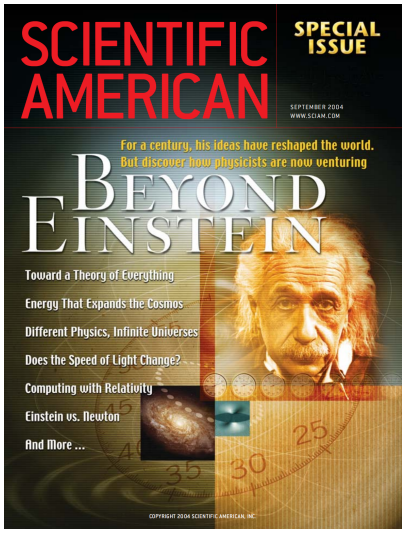
$$\cdots E_2E_1M = I.$$

This is only true if the RREF of M is the identity matrix.

Definition: A matrix M is **invertible** if its RREF is an identity matrix.

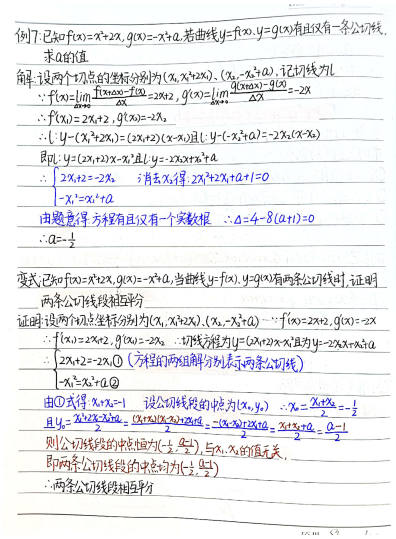
55

(b) University Textbook

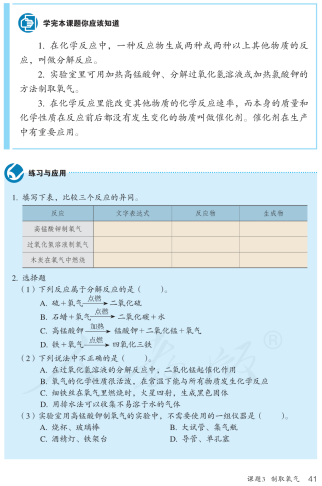


(c) Magazine

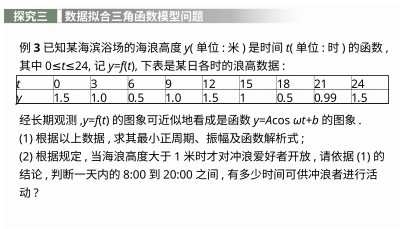
(d) Newspaper



(e) Handwritten Note



(f) Primary/Secondary School Textbook



(g) Slide

Figure 7. Representative sample images of DreamDoc dataset

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023. 6

[2] Ayan Banerjee, Sanket Biswas, Josep Lladós, and Umapada Pal. Swindocsegmenter: An end-to-end unified domain adaptive transformer for document instance segmentation. In International Conference on Document Analysis and Recognition, pages 307–325. Springer, 2023. 2

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In ACL Workshops, pages 65–72, 2005. 7

[4] Manuele Barraco, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. With a little help from your own past: Prototypical memory networks for image captioning. In ICCV, pages 3021–3031, 2023. 3

[5] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418, 2023. 6, 1, 2

[6] 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. Sci. China Inf. Sci., 67(12):220101, 2024. 6

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern-VL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In CVPR, pages 24185–24198, 2024. 1, 2

[8] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. arXiv preprint arXiv:2510.14528, 2025. 1, 2

[9] Datalab-to. Chandra: An ocr tool for image processing, 2025. 2

[10] Jonathan Daume, Jan Kamiński, Andrea GP Schjetnan, Yousef Salimpour, Umais Khan, Michael Kyzar, Chrystal M Reed, William S Anderson, Taufik A Valiante, Adam N Mamelak, et al. Control of working memory by phase–amplitude coupling of human hippocampal neurons. Nature, 629(8011):393–401, 2024. 3

[11] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines, 2014. 3

[12] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626):471–476, 2016. 3

[13] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In ECCV, 2020. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 6

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, pages 9729–9738, 2020. 5

[16] Jiarong Huang, Dezhi Peng, Hongliang Li, Hao Ni, and Lianwen Jin. SegCTC: Offline handwritten Chinese text recognition via better fusion between explicit and implicit segmentation. In ICDAR, pages 332–349, 2023. 8

[17] Marta Huelin Gorriz, Masahiro Takigawa, and Daniel Bendor. The role of experience in prioritizing hippocampal replay. Nat. Commun., 14(1):8157, 2023. 3

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024. 1, 2

[19] Yire Jeong, Hye-Yeon Cho, Mujun Kim, Jung-Pyo Oh, Min Soo Kang, Miran Yoo, Han-Sol Lee, and Jin-Hee Han. Synaptic plasticity-dependent competition rule influences memory formation. Nat. Commun., 12(1):3915, 2021. 3

[20] 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 ECCV, pages 498–517, 2022. 1, 2

[21] VI Lcvenshtcin. Binary coors capable or 'correcting deletions, insertions, and reversals. In Soviet physics-doklady, 1966. 7

[22] Minhyeok Lee, Suhwan Cho, Seunghoon Lee, Chaewon Park, and Sangyoun Lee. Unsupervised video object segmentation via prototype memory network. In WACV, pages 5924–5934, 2023. 3

[23] Tianqin Li, Zijie Li, Andrew Luo, Harold Rockwell, Amir Barati Farimani, and Tai Sing Lee. Prototype memory and attention mechanisms for few shot image generation. In ICLR, 2022. 3

[24] Tao Li, Shilian Wu, and Zengfu Wang. Mask guided selective context decoding for handwritten Chinese text recognition. In ICASSP, pages 1–5, 2023. 8

[25] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In ECCV, pages 280–296. Springer, 2022. 6

[26] 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. 1, 2

[27] Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218, 2025. 1

[28] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. arXiv preprint arXiv:2405.14295, 2024. 2, 7

[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, OCR, and world knowledge, 2024. 6

[30] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: Towards real-world vision-language understanding, 2024. 1, 2

[31] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In CVPR, pages 24838–24848, 2025. 1

[32] Randall C O'Reilly, Rajan Bhattacharyya, Michael D Howard, and Nicholas Ketz. Complementary learning systems. Cognitive science, 38(6):1229–1248, 2014. 3

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL, pages 311–318, 2002. 7

[34] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, pages 3743–3751, 2022. 2

[35] Claude E Shannon. A mathematical theory of communication. BSTJ, 27(3):379–423, 1948. 5

[36] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. In International Conference on Document Analysis and Recognition, pages 131–146. Springer, 2021. 2

[37] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. MovieChat+: Question-aware sparse memory for long video question answering. arXiv preprint arXiv:2404.17176, 2024. 3

[38] Jochen Triesch, Anh Duong Vo, and Anne-Sophie Hafner. Competition for synaptic building blocks shapes synaptic plasticity. Elife, 7:e37836, 2018. 3

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NeurIPS, 30, 2017. 2, 4, 6

[40] 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. 1, 2

[41] 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 ECCV, pages 408–424. Springer, 2024. 2, 6

[42] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Small language model meets with reinforced vision vocabulary. arXiv preprint arXiv:2401.12503, 2024. 6

[43] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704, 2024. 1, 2, 6, 7

[44] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-OCR: Contexts optical compression. arXiv preprint arXiv:2510.18234, 2025. 1

[45] 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 KDD, pages 1192–1200, 2020. 1, 2

[46] Shi Yan, Jin-Wen Wu, Fei Yin, and Cheng-Lin Liu. Recognizing handwritten Chinese texts with insertion and swapping using a structural attention network. In ICDAR, pages 557–571, 2021. 8

[47] Gang Yao, Ning Ding, Tianqi Zhao, Kemeng Zhao, Pei Tang, Yao Tao, and Liangrui Peng. Visual prompt learning for chinese handwriting recognition. In ICDAR, pages 119–133, 2024. 6, 8

[48] Matthew D Zeiler. Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012. 8

[49] Hesuo Zhang, Lingyu Liang, and Lianwen Jin. SCUT-HCCDoc: A new benchmark dataset of handwritten Chinese text in unconstrained camera-captured documents. PR, 108: 107559, 2020. 2, 7

[50] 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. 2

[51] Yuanzhi Zhu, Zecheng Xie, Lianwen Jin, Xiaoxue Chen, Yaoxiong Huang, and Ming Zhang. SCUT-EPT: New dataset and benchmark for offline Chinese text recognition in examination paper. IEEE Access, 7:370–382, 2018. 2, 7, 8