Nanbeige-VL: MMFM Version Technical Report

Rui Xiao * James Wu * Emily Zhou * Thomas An * Michael Du *
David Li * Daniel Cheng *

June 13, 2024

Abstract

Multimodal Foundation Models (MMFMs) have achieved remarkable success across various computer vision tasks, yet their application to specific domains such as document understanding remains challenging. In this report, we introduce Nanbeige-VL (MMFM Version), including architecture, data, training and final evaluation results in MMFM challenge. The results show that we achieved the highest scores in 10 out of 13 sub-assessments in the two phases of this challenge, and won the first place in this challenge.

1 Introduction

Multimodal Foundation Models (MMFMs) have emerged as pivotal tools in computer vision, demonstrating exceptional prowess across diverse tasks. However, their effective application to specialized domains such as document understanding presents significant challenges. In this report, we present Nanbeige-VL (MMFM Version), a novel architecture tailored for the MMFM challenge. This model integrates state-of-the-art techniques from Vision Transformers (ViT) and Large Language Models (LLM), specifically the pre-trained DFN ViT-H[6] and our self-developed Nanbeige2-8B[22] LLM, augmented by a custom projector module.

Our approach addresses the complexities of document analysis through a dynamic resolution strategy, enabling robust token representation across varying image sizes and aspect ratios. This adaptation not only enhances computational efficiency by reducing visual tokens but also preserves ViT's performance integrity. Key to our methodology is a meticulous three-stage training regimen encompassing pretraining, multitask learning, and supervised fine-tuning, each leveraging distinct datasets meticulously curated for alignment, multitask proficiency, and challenge-specific competence.

The efficacy of Nanbeige-VL (MMFM Version) is underscored by its exceptional performance in the MMFM challenge, where it attained the highest scores in 10 out of 13 sub-assessments across two phases. This achievement culminated in securing the first-place position, affirming its superiority in multimodal document understanding tasks. In the following sections, we delineate the architecture, datasets, training methodologies, and comprehensive evaluation results that substantiate Nanbeige-VL's competitive edge in the burgeoning field of MMFMs.

2 Methodology

2.1 Architecture

Nanbeige-VL (MMFM version) employs a well-established architecture in open-source Multimodal Language Models (MLLMs), referred to as "ViT-Projector-LLM". Our implementation integrates the pre-trained DFN ViT-H[6] with our pre-trained LLM, Nanbeige2-8B[22], utilizing a randomly initialized projector.

Our projector solution incorporates 2 layers of convolution and MLP, effectively representing a 378×378 image with 144 visual tokens. Experimental results demonstrate minimal impact on ViT's performance while significantly reducing the number of visual tokens.

Additionally, we have adopted a dynamic resolution strategy, partitioning images into tiles ranging from 378×378 pixels across sizes 1 to 20, tailored to the aspect ratio and resolution of each input image. This approach allows for a flexible representation with visual token counts ranging from 144 to 2880.

*BOSS ZHIPIN (Kanzhun Limited)

2.2 Dataset

We divided the training into three stages: pretraining, multi-task, and supervised fine-tuning, and used different data for each stage.

Pretraining Dataset. In this stage, the main training is the alignment of ViT and LLM. The data used is shown in Table 1, with a total of 100 million image-text pairs. During training, LLM is frozen and other weights are trained.

We used some common cleaning methods on these datasets.

- Remove uncommon symbols in the text
- Remove URLs in the text
- Remove pairs with text containing non-Chinese and non-English characters
- Use CLIP score filtering
- Remove HTML tags in the text
- Remove pairs with too short text
- Remove pairs with too small or big image
- Remove pairs with text containing emoji characters

Table 1: Datasets used in pretraining stage.

Dataset
Laion[38], Coyo[2], DataComp[7], SBU Captions[33], Wukong[9]

Multi-task Dataset. At this stage, the main goal of training is to enable the model to solve multiple tasks. The datasets related to MMFM challenge are listed in Table 2. During training, all weights are trained.

Table 2: Datasets used in multitask stage.

task	**Dataset**
Document	DocBank[23], Cord[34], DocLaynet[37], funsd[14], tabfact[3], websrc[4], wildreceipt[43], docile[41], DeepForm[45], Screen2words[48], rvlcdip[10]
Science	ScienceQA[26], AI2D[18], TQA[19]
OCR	OCRVQA[32], ArT[5], COCO-Text[47], CTW[49], LSVT[44], RCTW-17[39], ST-VQA[1], VisualMRC[46], Textcaps[40], TATDQA[50], ICDAR-19[12], Synthdog[20] gen
Grounding	GRIT[36], VisualGenome[21], RefCoco[17], RefCoCo+, RefCocog
Chart	DVQA[15], WTQ[35], PlotQA[31], MMC-Ins[25], LRV-Ins[24]
General QA	VQAv2[8], GQA[13], iconQA[27], OKVQA[29]

SFT Dataset. At this stage, in order to cope with MMFM challenge, we used some new data to replace the original data used for SFT. The data used are shown in Table 3. These data are used to train the document, chart, table comprehension capabilities required for this challenge. During this stage of training, ViT is frozen and other weights are trained.

Table 3: Datasets used in SFT stage.

Dataset
Dataset provided by MMFM, DocReason[11], TextVQA[42], ChartQA[30], PubTabNet[28], Chart-to-text[16]

3 Results and Training Settings

The MMFM challenge was conducted in two phases, with the results presented in Table 4. Our model achieved the highest scores in 10 out of 13 sub-assessments across both phases of the challenge. Notably, the three tests in Phase 2 utilized private test sets provided by MMFM, where our model also secured the highest scores. This demonstrates the exceptional generalization capability of our model.

We used a GPU with 40 A800 units for training. The learning rate was set to 2e-5, and the batch size was 2 with gradient accumulation steps of 4. Additionally, a warmup ratio of 0.03 was applied during training.

Table 4: Evaluation results.

Phase	Eval	Acc	Phase Overall Acc
Phase 1	iconqa fill	97%	76.7%
	funsd	87.5%	
	iconqa choose	86.5%	
	wildreceipt	91%	
	textbookqa	69%	
	tabfact	72%	
	docvqa	79.5%	
	infographicvqa	41%	
	websrc	99.5%	
	wtq	44%	
Phase 2	mydoc	73.5%	56.5%
	mychart	10.5%	
	myinfographic	62.15%	

4 Conclusion

In this report, we introduced Nanbeige-VL (MMFM version), a pioneering Multimodal Foundation Model designed specifically for document understanding tasks. By integrating the robust capabilities of DFN ViT-H and our Nanbeige2-8B LLM through an innovative projector module, we achieved outstanding performance in the MMFM challenge. Our model excelled across various sub-assessments, securing the top position in 10 out of 13 categories over two phases of evaluation.

Key to our approach was the adoption of a dynamic resolution strategy and efficient token representation, which optimized computational efficiency without compromising performance. The meticulous three-stage training regimen underscored our model's adaptability and proficiency across diverse datasets, ranging from pretraining alignment to specialized task fine-tuning.

As we continue to advance the frontier of multimodal AI, Nanbeige-VL stands as a testament to the transformative potential of integrating vision transformers and large language models. Moving forward, further refinements and applications in document understanding promise to extend the utility and impact of multimodal foundation models in real-world applications.

References

[1] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE, 2019.

[2] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. *arXiv preprint arXiv:2022.12345*, 2022.

[3] 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.

[4] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021.

[5] 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] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[7] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[9] 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.

[10] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015.

[11] 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.

[12] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019.

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[14] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.

[15] 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.

[16] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022.

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[18] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[19] Daesik Kim, Seonhoon Kim, and Nojun Kwak. Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. *arXiv preprint arXiv:1811.00232*, 2018.

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

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[22] NanBeiGe LLM Lab. Nanbeige llm. `https://github.com/Nanbeige/Nanbeige`, Nov 2023.

[23] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

[24] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[25] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023.

[26] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[27] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

[28] Maksym Lysak, Ahmed Nassar, Nikolaos Livathinos, Christoph Auer, and Peter Staar. Optimized table tokenization for table structure recognition. In *International Conference on Document Analysis and Recognition*, pages 37–50. Springer, 2023.

[29] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[30] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[31] 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.

[32] 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.

[33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[34] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwal-suk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

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

[36] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[37] 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.

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[39] 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)*, volume 1, pages 1429–1434. IEEE, 2017.

[40] 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.

[41] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, pages 147–166. Springer, 2023.

[42] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[43] Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*, 2021.

[44] 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.

[45] S Svetlichnaya. Deepform: Understand structured documents at scale, 2020.

[46] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021.

[47] 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.

[48] 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.

[49] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34:509–521, 2019.

[50] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022.