FTL Detector Technical Report

Yuhui Zheng

Abstract

Multimodel Large Language Models (MLLMs) with chart image parsing and document image understanding capabilities and have a significant advantage over traditional methods. In order to better understand the core content and logic of a document in accordance with the human reading order, it is crucial to understand its tables, each section of the document and the logic of document flowchart. Before understanding them, we first need to detect the base shapes of the flowchart, the tables in the document, and the location of the various sections of the document. In order to complete the first step correctly, this project introduces ***Flowchart, Table and Layout Detector (FTL Detector)****, a specialised MLLMs obtained by fine-tuning the InternVL2 using the FTL Dataset, a multi-source targeted dataset that we constructed. In order to better detect the base shapes of flowchart, we design two data synthesis methods: rule-based and GPT4o-based, to improve the quantity and quality of fine-tuning data. By leveraging the advanced capabilities of InternVL2 and targeted fine-tuning data, this tool supports functionalities that accurately detect the coordinates of the base shapes of the flowchart and the tables in the document and the layout of the document. Furthermore, the FTL Detector boasts enhanced interpretability of object coordinate detection. Our code are available at* `https://github.com/cszhengyh/FTLDetector`.

1. Introduction

In recent years, Multimodal Large Language Models (MLLMs) have exhibited impressive capabilities in chart parsing and document understanding. These models, which combine visual and linguistic features, can tackle tasks like chart data extraction, table parsing, and the structural analysis of complex documents. However, despite these advancements, significant challenges persist, limiting their performance in real-world applications.

Some Multimodal Large Language Models (MLLMs) like mPLUG-DocOwl[29] and Donut[6] excel in tasks such as document question answering (DocVQA), chart understanding (ChartQA), and structural analysis. mPLUG-DocOwl uses modular designs for diverse tasks, while Donut adopts an OCR-free approach to improve structural parsing. However, these models struggle with precise positional and relational data extraction, particularly in complex layouts. Recent integrations of advanced object detection frameworks like DINO[32] and GroundingDINO[15] have enhanced MLLMs' ability to identify bounding boxes and text positions, critical for chart parsing and document understanding. Datasets like ChartQA and benchmarks such as Dessurt[4] provide targeted evaluation and training opportunities for structured data extraction and relational modeling.

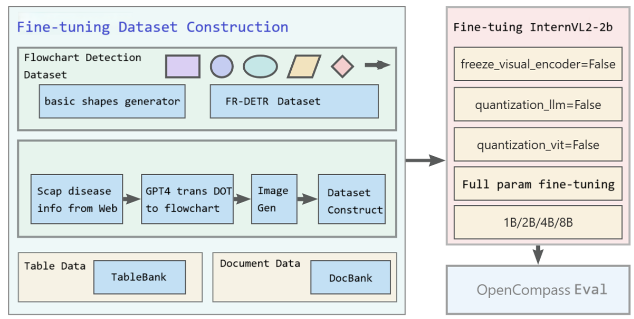

Figure 1. Structure and flow of FTL Detector.

Key challenges remain, including effective cross-modal fusion, handling diverse layouts, and improving logical and positional comprehension. These limitations hinder the models' ability to fully replicate human-like document understanding and constrain their real-world applicability.

As shown in Fig. 1, this study aims to overcome these gaps by synthesizing flowchart and document object detection data to fine-tune state-of-the-art open-source MLLMs tailored for precise chart and document parsing. We propose a new multimodal large model graph parser based on InternVL, which aims to:

- Accurately analyze the layout of document image and understand the content in human reading order.
- Accurately identify the bounding boxes of basic shapes in flowcharts and the bounding boxes of tables in document image and output the bbox coordinates.

A dataset of 7,000 output bounding boxes was constructed, and the InternVL2-2B model was full parameter fine-tuned. It is now capable of recognizing targeted object in images.

We present a summary of our key contribution:

- We propose two data synthesis methods for flowchart object detection task based on rule and GPT4o, and verify their effectiveness.
- We construct a set of high quality and efficient object detection data of document and flowchart images and get a specialised model that is capable of precise object detection in the flowchart and document images.

2. Related Work

Multimodal large language models utilize a connector to bridge the gap between large language models [3, 19, 24, 33, 36] and vision encoders [18, 20], enabling enhanced capabilities in comprehension and instruction following. Approaches like BLIP2 [8], Flamingo [1], mPLUG-Owl [30], and Qwen-VL [2] employ QFormers or Resamplers to align modalities over vast datasets of image-text pairs. LLaVA [13, 14] is a groundbreaking effort to extend the instruction-tuning paradigm to visual tasks using only text-based GPT-4 [17], achieving impressive performance with a simple MLP while maintaining visual information to refine multimodal alignment. Some studies [12, 22, 23] investigate combining various vision encoders, complementing each other to enhance visual representations and improve the fine-grained visual perception of MLLMs. Despite advances in structural design, training strategies and data quality remain critical in the further development of MLLMs.

Open-set Object Detection The field of open-set object detection has seen rapid progress with the incorporation of large language models and multimodal learning frameworks[7, 10, 25, 25–28, 31, 34, 35, 37–41]. MDETR[5] presents a model that aligns textual descriptions with visual regions using a DETR-like architecture. This approach enhances object detection by leveraging textual queries to guide detection, providing a more flexible understanding of visual content. GenerateU[11], on the other hand, focuses on identifying and naming objects within images without relying on predefined categories, thus overcoming the constraints of traditional object detection techniques. Methods like GLIP[9] and Grounding DINO[16] have showcased the potential of combining object detection with language, framing detection as a language grounding problem to learn instance-level visual representations through deep language-aware fusion. APE[21] integrates detection and grounding into a single model capable of tackling a wide range of tasks concurrently. While these advancements have greatly improved open-set object detection, training large-scale multimodal detection models still demands considerable computational resources and data. Consequently, fine-tuning existing detection frameworks for better adaptation to new downstream data emerges as a more cost-efficient approach.

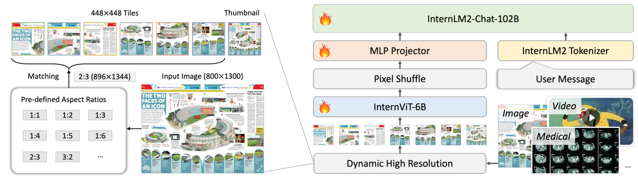

Figure 2. **Overall Architecture and Illustration of dynamic high resolution**. InternVL2 adopts the ViT MLP-LLM architecture similar to popular MLLMs,combining a pre-trained InternViT-6B with InternL,M2-20B through a MLP proiector. Here, we employ a simple pixel shufle to reduce the number of visual tokens to one-quarter. We dynamically match an optimal aspect ratio from predefned ratios, dividing the image into tiles of 448×448 pixels and creating a thumb nail for global context. This method minimizes aspect ratio distor-tion and accommodates varying resolutions during training

3. Method

3.1. Model Architecture

As shown in Fig. 2, this project adopts the advanced InternVL2 model. InternVL2 multimodal macromodel, which uses InternViT for the visual model and Qwen, Internlm and other models for the language module. The 40B and 76B versions of the model achieved about the same results as GPT-4o on Image benchmark, while the $1 \sim 8$B models all achieved fairly good results on Grounding Benchmarks, which shows that the results of ViT model are quite good. So the model was firstly used for training. Tab. 1 and Tab. 2 show the visual and language parts of InternVL2.

Table 1. Corresponding vision part for various sizes of the InternVL2 model.

Model Name	Vision Part
InternVL2-1B	InternViT-300M-448px
InternVL2-2B	InternViT-300M-448px
InternVL2-4B	InternViT-300M-448px
InternVL2-8B	InternViT-300M-448px
InternVL2-26B	InternViT-6B-448px
InternVL2-40B	InternViT-6B-448px
InternVL2-Llama3-76B	InternViT-6B-448px

3.2. Fine-tuning

We constructed the fine-tuned dataset according to the fine-tuned data format provided by the official fine-tuning documentation of InternVL2[1].

[1] https://internvl.readthedocs.io/en/latest/internvl2.0/finetune.html

Table 2. Corresponding language part for various sizes of the InternVL2 model.

Model Name	Language Part
InternVL2-1B	Qwen2-0.5B-Instruct
InternVL2-2B	internlm2-chat-1-8b
InternVL2-4B	Phi-3-mini-128k-instruct
InternVL2-8B	internlm2_5-7b-chat
InternVL2-26B	internlm2-chat-20b
InternVL2-40B	Nous-Hermes-2-Yi-34B
InternVL2-Llama3-76B	Nous-Hermes-2-Yi-34B

4. Experiment

4.1. Dataset

In this project, the supervised fine-tuning data primarily consists of self-constructed datasets due to the scarcity of existing datasets for flowchart recognition, table recognition, and PDF recognition. We plan to use these existing datasets as a test set rather than for training.

- Flowchart Recognition Dataset

  1) Basic shapes recognition dataset: Basic shapes include rectangle, diamond, parallelogram, circle, ellipse, arrow. This project designed a rule-based basic shape generator that can automatically generate shapes in batches, arranging them in $2 \times 2, 3 \times 3, 4 \times 4, 5 \times 5$, and 6×6 layouts in each image and combine them into flowcharts and store bounding boxes of basic shapes.

  2) FR-DETR Dataset[2] : The dataset contains flowchart and labels (annotations.zip) for training (train.zip), validation (val.zip) and testing (test.zip). Fig. 3 is the statistic of the FR-DETR dataset's basic shapes.

  3) GPT-4o-based synthesis: We generated flowcharts by scraping disease data from Dingxiang Doctor. Each section produced a DOT flowchart by GPT4o and converted to PNG via Graphviz. We filtered out images with an aspect ratio over 2 and sampled 10,000 for the dataset, also converting DOT to JSON formats.

- Table Dataset

  TableBank Dataset[3] is a new image-based table detection and recognition dataset built with novel weak supervision from Word and Latex documents on the internet, contains 417K high-quality labeled tables. We selected 10,000 samples for our dataset.

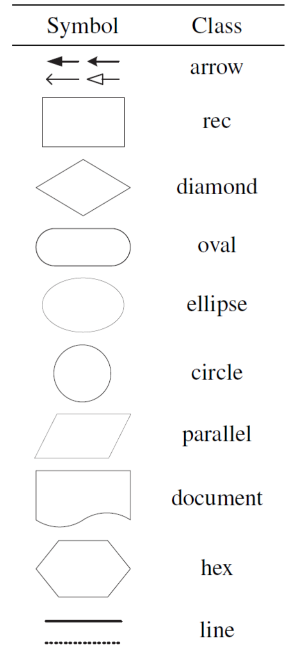

Figure 3. Statistic of the FR-DETR dataset's basic shapes.

- Document Layout Format Dataset

  The document layout dataset uses the open-source DocBank[4] dataset, selecting 10,000 samples for the dataset. DocBank is a new large-scale dataset that is constructed using a weak supervision approach. It enables models to integrate both the textual and layout information for downstream tasks. The DocBank dataset totally includes 500K document pages, where 400K for training, 50K for validation and 50K for testing.

4.2. Implementation Details

We use a A100 to fine-tune InternVL2-2B with full parameter and set freeze_visual_encoder=False, quantization_llm=False, quantization_vit=False.

The main parameters in this study are set as follows: the batch size is 4 , with gradient accumulation steps of 4, effectively resulting in a batch size of 16. The maximum input text length is set to 8192 tokens. The optimizer is AdamW with a learning rate of 1×10-5, beta values of (0.9, 0.999), and a weight decay of 0.01. Gradient clipping is applied with a max norm of 1. The training process includes 2 epochs, with a warmup ratio of 0.03. The learning rate scheduling uses a linear warmup followed by cosine annealing. Data loading is multithreaded with 4 workers per dataloader.

For outputting bounding boxes and target detection data, the size of the images needs to be normalized to the range [0,1000].

[2] https://github.com/harolddu/frdetr_dataset/tree/main

[3] https://github.com/doc-analysis/TableBank

[4] https://github.com/doc-analysis/DocBank

5. Conclusion and Future Work

The FTL Detector has demonstrated significant advancements in detecting object coordinate in flowchart and document images. In this project, we presented FTL Detector, a new specialized MLLM fine-tuned using open-source datasets and our synthetic data that can accurately detect the bounding box of the base shapes of the flowchart images and the bounding box of the tables and the content of each section in the document images. We designed two flowchart generators that can automatically batch-generate flowcharts for fine-tuning to enhance the ability of flowchart detection. Experiments show that our proposed method can accurately analyze the layout of document images, understand the document content in accordance with the human reading order.

For future work, more data synthesis routes are worth exploring. After being able to accurately detect important objects in document and flowchat images, how to futher improve the model's documents understanding ability becomes the next goat.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. In *proceedings of NeurIPS*, pages 23716–23736, 2022. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, et al. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint:2308.12966*, 2023. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *proceedings of NeurIPS*, pages 1877–1901, 2020. 2

[4] Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. End-to-end document recognition and understanding with dessurt, 2022. 1

[5] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[6] 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, 2022. 1

[7] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models, 2023. 2

[8] Junnan Li, Dongxu Li, Silvio Savarese, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *proceedings of ICML*, pages 19730–19742, 2023. 2

[9] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[10] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip, 2023. 2

[11] Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13958–13968, 2024. 2

[12] Ziyi Lin, Chris Liu, Renrui Zhang, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multimodal large language models. *arXiv preprint:2311.07575*, 2023. 2

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *proceedings of NeurIPS*, 2023. 2

[14] Haotian Liu, Chunyuan Li, Yuheng Li, et al. Improved baselines with visual instruction tuning. In *proceedings of CVPR*, 2024. 2

[15] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 1

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025. 2

[17] OpenAI. Gpt-4 technical report. *arXiv preprint:2303.08774*, 2023. 2

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. 2

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *proceedings of ICML*, pages 8748–8763. PMLR, 2021. 2

[21] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203, 2024. 2

[22] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2

[23] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2

[24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint:2302.13971*, 2023. 2

[25] Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21938–21948, 2023. 2

[26] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7031–7040, 2023.

[27] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36, 2024.

[28] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. 2

[29] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding, 2023. 1

[30] Qinghao Ye, Haiyang Xu, Guohai Xu, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint:2304.14178*, 2023. 2

[31] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection, 2022. 2

[32] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 1

[33] Susan Zhang, Stephen Roller, Naman Goyal, et al. Opt: Open pre-trained transformer language models. *arXiv preprint:2205.01068*, 2022. 2

[34] Shaokun Zhang, Xiaobo Xia, Zhaoqing Wang, Ling-Hao Chen, Jiale Liu, Qingyun Wu, and Tongliang Liu. Ideal: Influence-driven selective annotations empower in-context learners in large language models. *arXiv preprint arXiv:2310.10873*, 2023. 2

[35] Shaokun Zhang, Jieyu Zhang, Jiale Liu, Linxin Song, Chi Wang, Ranjay Krishna, and Qingyun Wu. Training language model agents without modifying language models. *arXiv e-prints*, pages arXiv–2402, 2024. 2

[36] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *proceedings of NeurIPS*, 2023. 2

[37] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2

[38] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.

[39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[40] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.

[41] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 2