TableVLM: A visual language model with multi-modal joint guided learning for end-to-end image-based table recognition

Anonymous ACL submission

Abstract

Image-based table recognition is one of the important issues in intelligent document processing. Existing solutions usually decompose it into multiple subtasks to solve them separately, but lead to shortcomings like error propagation, weak generalization, etc. Considering multi-modal large language models usually have excellent performance in image captioning and support multiple languages, we propose an innovative end-to-end solution, and construct corresponding datasets, models and evaluation metrics. Specifically, we firstly redefine the HTML representation of the table and remove some unnecessary tags for fair comparison and save limited tokens. Then, we construct a multi-modal dataset containing more than 600k question-answer pairs in total, and each image is annotated only with its HTML representation for training and evaluating the performance of the corresponding methods. In addition, to make the evaluation scheme more comprehensive, we proposed EDSC, Efficiency to evaluate the content recognition ability and cost-effectiveness of various methods. Finally, we construct a multi-modal image-based table recognition model TableVLM, including two different versions, 4B and 14B, focusing on cost-effectiveness and performance respectively. Experimental results show that the proposed TableVLM is able to recognize table images of various styles. Its recognition and generalization capabilities surpass those of existing table-related multi-modal large language models. Therefore, it is an effective and innovative end-to-end solution.

1 Introduction

Table images, as one of the main forms of tables, are extremely common in our daily life. However, since the information in them cannot be read directly, which hinders its widespread application. Therefore, image-based table recognition that can convert table images into readable files is crucial. To address this problem, existing solutions usually decompose it into four sub-tasks: table structure recognition, table content detection, table content recognition, and table reconstruction, as shown in Fig. 1. Furthermore, according to the differences in the feature representation, existing solutions can be roughly divided into two categories: visual recognition-based methods and sequence generation-based methods. Specifically, the former usually firstly recognizes the visual elements in the table, such as cells, separator lines, rows, columns, etc., and reconstructs the table structure based on their relevant information. Some representative methods include TSRFormer(Lin et al., 2022), SEM(Zhang et al., 2022), LORE(Xing et al., 2023), TGRNet(Xue et al., 2021), RobustTabNet(Ma et al., 2023), TableNet(Paliwal et al., 2019), DeepTabStR(Siddiqui et al., 2019), etc. Then, it is necessary to extract the content in the table based on text detection and text recognition methods, such as DBNet(Liao et al., 2020), DBNet++(Liao et al., 2022), SVTR(Du et al., 2022), etc., and finally embed the content into the correct cells according to pre-set rules. Unlike the former, the latter regards image-based table recognition as a sequence generation task, that is, converting table images into corresponding sequence representations, such as HTML, LaTeX, Markdown, etc. They firstly convert the table structure into the corresponding sequence representation and obtain the pixel coordinates of each cell. Some representative methods include TableMaster(Ye et al., 2021), EDD(Zhong et al., 2020), VAST(Huang et al., 2023), etc. Then, the content in the table is still extracted through text detection and text recognition methods, and embedded into the corresponding cells according to the coordinates of the cells and text blocks obtained previously. According to the above analysis, we can find that most of the existing solutions include multiple steps and require various models to cooperate with each other. However, an inherent defect

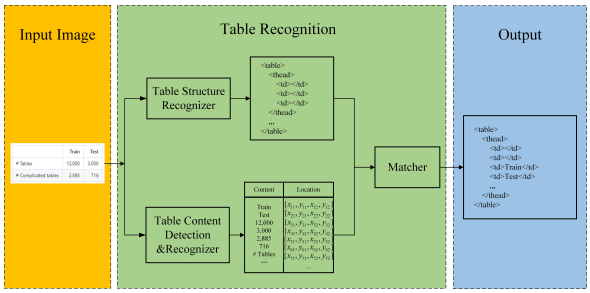


Figure 1: The traditional pipeline of image-based table recognition.

of this pipeline is that the performance degradation of any sub-task will directly lead to the decline in the quality of the final recognition result, such as the loss of table content due to low table structure recognition accuracy, the embedding of table content into the wrong cell due to low cell positioning ability, etc. In addition, the complexity and diversity of table content is also a major challenge. For example, multiple languages, mathematical formulas, subscripts, subscripts, special symbols, etc. In summary, the existing multi-stage solutions still have some limitations that need to be addressed to improve the performance of image-based table recognition.

With the emergence of ChatGPT(Achiam et al., 2023), large language model-related technologies have developed rapidly and have been widely used in various scenarios. Image-based table recognition is no exception. So far, many studies have focused on how to process the tabular data according to the natural language prompts, such as TableLLM(Zhang et al., 2024a), TableGPT(Zha et al., 2023), TableGPT2(Su et al., 2024), etc., or how to use multi-modal large language models to recognize and understand table images, such as Table-LLaVA(Zheng et al., 2024), UniTable(Peng et al., 2024), TabPedia(Zhao et al., 2024), etc., but they have not fully utilized the excellent capabilities of multi-modal large language models. Therefore, we conducts an in-depth exploration and analysis of the image-based table recognition capabilities of multi-modal large language models. Specifically, considering that a table can be represented by the HTML sequence, we regard image-based table recognition as a holistic Image-To-Seq task without having to decompose it into multiple sub-tasks. In addition, since multi-modal large language models usually have excellent image captioning capabilities, we believe that they can understand table images and the corresponding sequence representations. Finally, large language models usually have strong text processing capabilities, and support multiple languages, while OCR-related methods trained on public available datasets do not have such capabilities. Therefore, we believe that it is a good choice to use the excellent capabilities of multi-modal large language models to improve the performance of image-based table recognition. Based on the above analysis, we propose an innovative multi-modal-based end-to-end solution and construct corresponding datasets and models. Specifically, we firstly unified define the HTML representation of table images, that is, remove some unnecessary tags for fair comparison while saving limited tokens. Then, we construct a multi-modal dataset for image-based table recognition, which contains more than 600k question-answer pairs. Each image is annotated only with its HTML representation. Based on this dataset, we comprehensively evaluate the recognition capabilities of existing multi-modal large language models, select the most suitable foundation model, and further construct a multi-modal large language model TableVLM for image-based table recognition, including two different versions, 4B and 14B, focusing on cost-effectiveness and performance, respectively. Experimental results show that the proposed TableVLM is able to recognize the table images with various styles and show competitive performance. In summary, the contributions of this paper are as follows:

1. We proposed an innovative multi-modal-based end-to-end solution, which is more concise than the traditional multi-stage solution and can achieve competitive results with only sequence-level annotation.

2. We unified the HTML representation of table images, that is, deleted some tag pairs that can be added by preset rules and only retained necessary tags, which is conducive to fair comparison and also saves limited tokens.

3. We constructed a multi-modal dataset with more than 600k question-answer pairs. Each table image is annotated only with its HTML annotation.

4. We constructed a multi-modal large language model for image-based table recognition TableVLM, including two different versions, 4B and 14B, which focus on cost-effectiveness and performance respectively.

5. We introduced new evaluation metrics, including EDSC, Efficiency, to more comprehensively evaluate the image-based table recognition ability

of multi-modal large language models.

2 Related Works

In this section, we review the representative works related to image-based table recognition and multi-modal large language models. They are as follows.

2.1 Image-based Table Recognition

Image-based table recognition usually refers to extracting the structure and content in the table image and converting it into a machine-readable format, such as Excel, database, etc. It usually includes four main sub-tasks: table structure recognition, table content detection, table content recognition, and table reconstruction. They work together to achieve image-based table recognition. At present, many excellent solutions have been proposed for this problem, which can be roughly divided into two categories: visual recognition-based methods and sequence generation-based methods. The difference between the two is only in the representation of the table structure. Specifically, the former uses relevant information of visual elements such as cells, separator lines, rows and columns to describe the table structure. Some representative works include TGRNet(Xue et al., 2021), LORE(Xing et al., 2023), LORE++(Long et al., 2025), RobustTabNet(Ma et al., 2023), TSRFormer(Lin et al., 2022), SEM(Zhang et al., 2022), SEMv2(Zhang et al., 2024b), etc. The latter converts the table structure into the corresponding sequence representation and obtains the pixel coordinates of each cell. The representative methods include TableMaster(Ye et al., 2021), EDD(Zhong et al., 2020), VAST(Huang et al., 2023) etc. After completing the table structure recognition, the contents and their pixel coordinates in the table are extracted through text detection and text recognition methods. And finally, they are embedded into the corresponding cells to obtain the final recognition result. The above is the most common pipeline in image-based table recognition. Although the recognition result can be obtained accurately, it is a multi-stage solution that requires different models and data to solve the above sub-tasks respectively. The performance of any sub-task will directly affect the quality of the final recognition result. For example, due to insufficient accuracy of table content detection, the content is cut off, resulting in recognition errors, and due to low table structure recognition accuracy, the corresponding table content is lost, etc. Therefore, researchers gradually focus on how to construct a more effective pipeline.

In response to this problem, some corresponding methods have also been proposed, such as MTLTabNet(Ly and Takasu, 2023), which is a multi-task joint optimization model that uses three decoder heads for table structure recognition, cell position recognition, and cell content recognition, respectively, to directly generate HTML sequence representations of table images. In addition, GPT4V-OCR(Shi et al., 2023) also explores the potential in image-based table recognition by fine-tuning GPT-4V. This preliminarily verifies the feasibility of solutions based on multi-modal large language models.

2.2 Multi-Modal Large Language Models

Multi-Modal large language model, its main feature is combine the natural language processing capabilities of the large language model with the ability to understand and generate the data of other modalities, aiming to provide more powerful interactive capabilities by integrating multiple types of input and output such as text, images, sounds, etc. Existing multi-modal large language models mainly include Qwen-VL-Series(Bai et al., 2023), Qwen2-VL-Series(Yang et al., 2024), LLaVA-Series(Liu et al., 2024), mPLUG-Owl-Series(Ye et al., 2024), Phi-Series(Abdin et al., 2024), MiniCPM-V-Series(Hu et al., 2024b), etc. From the perspective of architecture, they can be roughly divided into three parts: pre-trained modal encoder, pre-trained large language model, and adapter. If multi-modal data generation is involved, the generator may also be included. Among them, the pre-trained large language model is very important, and its performance usually directly determines the capabilities of the corresponding multi-modal large language models. At present, many pre-trained large language models have been released one after another. They usually provide multiple different versions. Generally, the larger the number of parameters, the better the performance. The above-mentioned general multi-modal large language models perform well in tasks such as image captioning, visual question answering, OCR, etc., but their performance is not ideal when used directly for image-based table recognition, because image-based table recognition requires accurate content recognition and strict sequential output, which is a more difficult task. However, although it cannot be directly applied, it shows

Attributes		Traditional Solution	Our proposed Solution
Overall Pipeline		Multi-Stage	End-To-End
Supervision	HTML	×	✓
	HTML-Structure	✓	×
	Cell Content	✓	×
	Cell Bbox	✓	×
	Content Bbox	✓	×
Post-processing	Method	Position Matching	String modification
	Complexity	Complex	Simple
Recognition Ability	Table Structure	✓	✓
	Table Content	✓	✓
Generalization Ability	Various Styles	×	✓
	Multi-Language	×	✓

Table 1: The detailed comparison of the traditional multi-stage Image-To-HTML solution and our proposed solution.

competitive performence after fine-tuning, so some table-related multi-modal large language models have been proposed. For example, TabPedia(Zhao et al., 2024) is an innovative table-related visual language model that can seamlessly integrate multiple table-related tasks such as table detection, table structure recognition, and table question answering. A large number of experiments conducted on various public benchmarks have verified its effectiveness and superiority. UniTable(Peng et al., 2024) proposed a novel framework that unify table structure recognition, cell content recognition, and cell bounding box recognition into sequence modeling tasks, and achieved excellent performance on various public available datasets. In addition, Table-LLaVA(Zheng et al., 2024) uses LLaVA-1.5 as the foundation model and trains it on the proposed multi-modal table understanding dataset MMTab. After comprehensive evaluation, Table-LLaVA shows quite good image-based table recognition and understanding capabilities.

Although the above multi-modal large language models have excellent performance, they do not deeply analyze the influencing factors, limitations, etc. between multi-modal large language models and image-based table recognition. We believe that for image-based table recognition, considering that the size of text blocks in table images is usually small, the resolution of the input image should be large to ensure sufficient visual information. In addition, since table content usually contains multiple languages, the pre-trained large language model should also support multiple languages. Considering that the sequence representation of large tables is usually long, the model should also have a large context length. Based on the above analysis, we comprehensively evaluated existing multi-modal large language models and public available datasets, further constructed an innovative end-to-end solution with corresponding datasets and models. In the following, we introduce them in details.

3 Task Definition

Considering that most existing image-based table recognition methods and public available datasets use HTML to represent tables, we also adopts the same approach. However, there are some differences in the tag pair sets used by different methods and datasets, which in turn affects the fairness of performance comparison. For example, in TableMaster(Ye et al., 2021), the author specifically introduced <eb></eb> to represent blank cells, although the fluctuation caused by the difference may not be large. In addition, some tag pairs in HTML have fixed meanings and are not very helpful for recognition. For example, <html></html> usually indicates the beginning and end of an HTML sequence, and <thead></thead> usually specifically indicates the first row of a table. They all can be added to the HTML sequence when necessary through pre-set rules.

To solve the above problems, we uniformly define the HTML representation of tables to ensure fair comparison. Specifically, we perform the tag pruning to remove some unnecessary tag pairs, such as <thead></thead>, <html></html>, etc. The specific representation rules are set as follows:

1. We only retain the essential structural tags in HTML, including <table></table>, <tr></tr>,

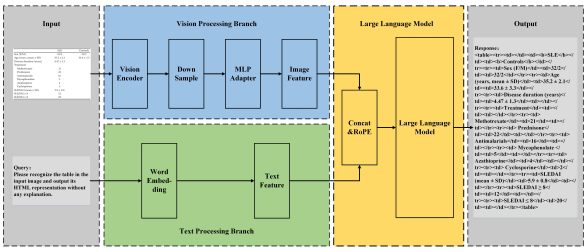


Figure 2: The overall architecture of the innovative multi-modal-based end-to-end solution.

<td></td>, which represent tables, table rows, and cells respectively. If a cell has row-span or column-span, it is achieved by setting the rowspan or colspan attributes of the cell.

2. We retain all content-related special tags, such as , , , etc., which is conducive to enhancing the description ability of HTML.

3. The HTML representation of the table should start with <table> and end with </table>. It should contain several rows, each starting with <tr> and ending with </tr>. And each row should contain several cells, each starting with <td> and ending with </td>. The cell content should be embedded between its start and end tags.

The benefit of following the above rules is that the model can focus on understanding the table structure and content, while making full use of limited tokens, simplifying the task and improving the recognition accuracy.

4 TableVLM

In this section, we introduce the details of the innovative multi-modal-based end-to-end solution and the proposed TableVLM. In the following, we introduce them in details.

4.1 Overall Architecture

The overall architecture of the innovative multi-modal-based end-to-end solution is shown in Fig. 2. It includes several main components, such as visual encoder, adapter, large language models. They are responsible for extracting the visual feature map, alignment, etc., respectively, finally converting the table image into the HTML representation. Compared with traditional solutions, it only requires HTML-level supervision to achieve image-based table recognition, which is a more concise and effective solution. Its detailed comparison with traditional multi-stage Image-To-HTML solution is shown in Table 1.

4.2 TableVLM-4B

In this section, we introduce TableVLM-4B, which includes multiple core modules such as visual encoder, projector, and small language model, with a total of about 4.2 billion parameters. The details of the main components are as follows.

Visual Encoder. Considering that CLIP has excellent image-text alignment capabilities, we use CLIP ViT-L/14(Cherti et al., 2022) as the visual encoder of TableVLM-4B to extract visual feature maps with stronger description capabilities.

Projector. For TableVLM-4B, two stacked MLP layers are used as the visual language adapters, which mainly for compressing or aligning visual feature maps and text feature maps.

Small Language Model. Due to Phi3.5-mini-Series(Abdin et al., 2024) language models perform well in various public benchmarks, such as DocVQA(Mathew et al., 2020), TextVQA(Singh et al., 2019), OCRbench(Liu et al., 2023), etc., we use Phi3.5-mini-instruct as the small language model in Table VLM-4B.

The above components together constitute TableVLM-4B. It has a small number of parameters while maintaining satisfactory performance. However, in some complex scenarios, its performance is not ideal, so it is necessary to explore the image-based table recognition capabilities of multi-modal large language models with larger parameters.

4.3 TableVLM-14B

In this section, we introduce the details of TableVLM-14B, which is a large visual language model with better performance. It also contains multiple core components such as visual encoder, projector, large language model, etc., with a total of about 13.9 billion parameters. We introduce the main components in details below.

Visual Encoder. Compared with CLIP, EVA-CLIP(Sun et al., 2023) performs better with the same number of parameters and lower training cost. Therefore, we use EVA-02-CLIP-E as the visual encoder of TableVLM-14B, which has more parameters and stronger visual feature map extraction ability. And the resolution of the input image is set to 1120×1120.

Projector. Similar to the above, we use two layers of stacked MLP as projectors to align visual feature maps and textual feature maps.

Large language model. Since GLM4-9B-

Chat(Zeng et al., 2024) performs well on various public test benchmarks and supports multiple languages, we use it as the large language model of TableVLM-14B, with a total parameter size of about 9.4B. In addition, we still use RoPE to encode the various feature maps, and all visual feature maps share the same position id. The above components together constitute TableVLM-14B.

5 Experiments and Analysis

In this section, we introduce the proposed multi-modal image-based table recognition dataset, the more comprehensive evaluation scheme, and a series of experiments to verify the effectiveness and advancement of TableVLM, including the evaluation of image-based table recognition ability, the evaluation of generalization ability, and the comparison with existing multi-modal large language models and existing image-based table recognition methods. In the following, we introduce the above in details.

5.1 Dataset

Considering the complexity of table content and morphology, our proposed dataset should contain many table images of different styles as much as possible. Therefore, we selected four distinctive datasets from the existing public available datasets, namely PubTabNet(Zhong et al., 2020), FinTabNet(Zheng et al., 2020), SciTSR(Chi et al., 2019) and TableOCR, all of which contain rich table images and accurate annotations. Based on them, we carefully designed the corresponding scripts to generate the HTML representations of table images. Then, according to the preset dialogue format, we constructed a large multi-modal image-based table recognition dataset, which contains more than 600k question-answer pairs for training and evaluating the performance of multi-modal large language models.

5.2 Evaluation Scheme

For the evaluation scheme, we require it to be able to comprehensively evaluate the image-based table recognition capabilities of the corresponding methods from various dimensions, including precision, cost-effectiveness, etc. Therefore, we proposed two innovative evaluation metrics to further supplement the existing evaluation scheme, as follows.

1.EDSC. Its full name is Edit-Distance-based Similarity for table Content, which is mainly used to evaluate the accuracy of table content recognition. It takes the average of the edit distance between the content prediction and the corresponding label of each cell in the table as the result, as shown in the following formula.

$$EDSC = \frac{1}{N}\sum_{i=0}^{N-1} EditDist(Pre_i, GT_i) \quad (1)$$

2.Efficiency. It is used to measure the cost-effectiveness of the corresponding method, the calculation formula is shown in formula 2. It should be noted that the basic unit of parameter quantity is billion.

$$Efficiency = \frac{TEDS}{Parameters} \quad (2)$$

The above evaluation metrics, together with TEDS, TEDS-Struct, and Acc, constitute the comprehensive evaluation scheme.

5.3 Results and Analysis

The performance evaluation of TableVLM. We evaluate the image-based table recognition capabilities of TableVLM on four different datasets, including PubTabNet, FinTabNet, SciTSR, and TableOCR. The corresponding evaluation metrics are shown in sub-section 5.2. The quantitative results are shown in Table 2.

According to the above, we conclude that the proposed TableVLM can accurately recognize the table images of various styles, whether Chinese tables, English tables, scanned tables, or distorted tables. Specifically, for FinTabNet and SciTSR, the TEDS and TEDS-Struct of TableVLM both exceed 95%, showing excellent performance. And for TableOCR, it also achieves great performance even though there are a large number of distorted images in this dataset. The above phenomenon shows that TableVLM can uniformly model the recognition of table images of different styles as a sequence generation task without designing solutions for each. For PubTabNet, TableVLM performs competitively but not completely satisfactory. The main reasons are as follows. Firstly, its content is more complex, including superscripts, subscripts, special symbols, etc., which increases the difficulty of image-based table recognition. Secondly, its original annotations do not contain the correspondence between the table content and the cells, which requires us

Models	Dataset	TEDS	TEDS-Struct	Acc	EDSC	Efficiency
Qwen2-VL-7B-Instruct	PubTabNet	79.12%	90.94%	54.21%	68.94%	9.53
	FinTabNet	90.93%	95.55%	75.76%	86.78%	10.97
	SciTSR	96.60%	97.36%	82.57%	95.26%	11.65
	TableOCR	93.67%	96.26%	86.03%	93.25%	11.30
LLava1.5-7B-Instruct	PubTabNet	59.31%	74.59%	11.76%	32.66%	8.35
	FinTabNet	35.07%	71.35%	7.79%	25.75%	4.94
	SciTSR	41.19%	73.61%	16.10%	27.64%	5.80
	TableOCR	39.21%	72.47%	20.13%	31.17%	5.52
LLava1.5-13B-Instruct	PubTabNet	60.21%	75.86%	12.86%	33.71%	4.49
	FinTabNet	37.08%	73.20%	10.08%	26.78%	2.77
	SciTSR	42.91%	74.79%	18.17%	28.92%	3.20
	TableOCR	42.72%	76.04%	24.59%	33.21%	3.19
MiniCPM-V.V2.5	PubTabNet	71.91%	88.03%	39.14%	55.24%	8.46
	FinTabNet	83.97%	92.31%	59.95%	77.18%	9.88
	SciTSR	88.94%	93.77%	61.53%	85.14%	10.46
	TableOCR	87.34%	92.66%	72.69%	86.09%	10.28
TableVLM-4B	**PubTabNet**	**81.86%**	**92.65%**	**59.61%**	**73.63%**	**19.49**
	FinTabNet	**93.50%**	**96.35%**	**78.07%**	**89.94%**	**22.26**
	SciTSR	**96.55%**	**97.38%**	**81.03%**	**95.15%**	**22.99**
	TableOCR	**88.22%**	**95.55%**	**82.31%**	**86.97%**	**21.00**
TableVLM-14B	**PubTabNet**	**83.90%**	**95.51%**	**71.10%**	**76.16%**	**6.03**
	FinTabNet	**95.82%**	**97.82%**	**84.94%**	**93.53%**	**6.89**
	SciTSR	**97.89%**	**98.33%**	**87.67%**	**97.02%**	**7.04**
	TableOCR	**97.70%**	**99.24%**	**95.56%**	**97.44%**	**7.03**

Table 2: The performance comparison of the proposed TableVLM with existing multi-modal large language models on different datasets.

to embed them one by one according to their arrangement order when generating the HTML representation, which may cause some errors, thereby reducing the quality of the multi-modal dataset.

In summary, TableVLM is an effective end-to-end image-based table recognition solution. Its advantage is that it does not need to split the image-based table recognition into multiple sub-tasks to be solved separately, only HTML is needed as supervision, and it can adapt to more different scenarios, etc. In addition, although its performance on PubTabNet is not ideal, we still believe that it can have excellent performance by providing enough high-quality datasets.

The performance comparison with existing multi-modal large language models. We compare the performance of the proposed TableVLM with various existing multi-modal large language models. The dataset and evaluation scheme used remain unchanged. The experimental results are shown in Table 1. It should be noted that the above models used for comparison have been fine-tuned using the proposed dataset.

According to the data shown in Table 1, it can be seen that TableVLM-14B has the best performance on all datasets, its TEDS and TEDS-Struct are both higher than 98%, far exceeding LLava-1.5-13B-Instruct with similar parameters. TableVLM-4B has the highest cost-effectiveness while taking into account performance. We believe that this should be attributed to its 128k context length, because the longer context length enables it to learn the dependencies between tokens that are far away, which is beneficial for image-based table recognition. For Qwen2-VL-7B-Instruct, LLava-1.5-7B-Instruct, and MiniCPM-V-V2.5-Chat, their performance is slightly inferior to or far inferior to TableVLM. We believe that the difference in their performance comes from different configurations, such as the pre-trained language model, the input resolution of the visual encoder, the context length, etc. Therefore, we make the following conclusion: a multi-modal large language model that is beneficial to image-based table recognition should meet

	Methods	Datasets	TEDS	TEDS-Struct	Acc
Image-based Table Recognition Methods	MTLTabNet	PubTabNet	96.67%	97.88%	-
		FinTabNet	-	98.79%	-
	UniTable	PubTabNet	96.50%	97.89%	-
		FinTabNet	-	98.89%	-
	MuTabNet	PubTabNet	96.87%	-	-
		FinTabNet	97.69%	98.87%	-
	TableMaster+PSENet+Master	PubTabNet	96.84%	-	-
Ours	**TableVLM-4B**	**PubTabNet**	**81.86%**	**92.65%**	**59.61%**
		FinTabNet	**93.50%**	**96.35%**	**78.07%**
	TableVLM-14B	**PubTabNet**	**83.90%**	**95.51%**	**71.10%**
		FinTabNet	**95.82%**	**97.82%**	**84.94%**

Table 3: The performance comparison of TableVLM and various existing image-based table recognition methods on different datasets. It is necessary to note that '-' represent there is no related data in the original paper.

the following conditions: support for multiple languages, large context length, large and flexible input image resolution, powerful OCR capabilities and image captioning capabilities. This is also the direction we are working towards.

The performance comparison with existing image-based table recognition methods. We compare the performance of the proposed TableVLM with various existing image-based table recognition methods, including MTLTabNet(Ly and Takasu, 2023), UniTable(Peng et al., 2024), etc. The experimental results are shown in Table 2.

According to the data shown in Table 2, we can find that, for table structure recognition, TableVLM shows performance close to that of existing image-based table recognition methods, although we did not fine-tune it using data related to table structure recognition. In addition, for table content recognition, TableVLM still shows good performance on FinTabNet, but its performance on PubTabNet is still not ideal. This is because the quality of the corresponding generated multi-modal data is not good enough and only HTML is used as supervision. Nevertheless, we still think this is a promising solution that can unify the recognition of table images with different styles into the same task without splitting it into multiple sub-tasks to solve separately, and requires less supervision. This is also the development trend of image-based table recognition and understanding. In the future, we will strive to improve its performance and cover more complex scenarios as much as possible. **Please note that, due to the limited space, more experimental results are shown in the appendix.**

6 Conclusion

In view of the shortcomings of existing image-based table recognition methods, such as cumbersome processes, error propagation, and weak generalization ability, considering the excellent cross-modal recognition ability and image captioning ability of multi-modal large language models, we proposed an innovative end-to-end solution and constructed corresponding datasets, models, and evaluation metrics. Specifically, we unified the HTML representation of the table and removed some unnecessary tags to make the corresponding models focus on the table structure and content. Then, we constructed a large-scale multi-modal image-based table recognition dataset and a comprehensive evaluation scheme for training and evaluating the performance of the corresponding methods. Finally, we constructed an image-based table recognition model TableVLM, including two different versions, 4B and 14B, focusing on cost-effectiveness and performance respectively. Experimental results show that the proposed TableVLM can unify the recognition of table images of different styles into a same task, and its performance is excellent, with strong recognition and generalization capabilities. The TEDS and TEDS-Struct are as high as more than 98%. In addition, we also analyzed the conditions for the adaptation of multi-modal large language models to image-based table recognition, including long context length, large input resolution, richer data, etc., which provided direction for subsequent optimization. In summary, it is an effective, innovative, and promising end-to-end solution that provides a novel and concise pipeline for image-based table recognition.

7 Limitations

Although this work has comprehensively explored the problem of image-based table recognition based on multi-modal large language models, there are still some limitations that need to be addressed in subsequent research. Firstly, the proposed dataset does not contain a large number of more difficult samples such as cross-page tables, blurred tables, rotated tables, etc., which limits the versatility of TableVLM to a certain extent. In the future, we will collect more complex table images and construct corresponding multi-modal data, such as WTW(Long et al., 2021). We believe that this can greatly enhance its application prospects. Then, the performance of the proposed TableVLM is still unsatisfactory. It still makes mistakes in low-quality table images, approximate characters, etc., and its inference process is not controllable. In the future, we will design a chain of though for image-based table recognition, so that it can gradually generate HTML sequences, improve the thinking ability of the corresponding models, and obtain better recognition results. Thirdly, the proposed TableVLM is only effective for Image-To-HTML, and cannot be generated for other common sequences such as LaTeX and Markdown. In the future, we will further expand the data volumn to make the performance of TableVLM more powerful. Finally, with the development of multi-modal large language models, more and more excellent models and fine-tune methods have emerged, such as mPLUG-DocOwl2(Hu et al., 2024a), DocPedia(Feng et al., 2023), GOT-OCR2.0(Wei et al., 2024), LLaVA-OneVision(Li et al., 2024), DeepSeek-VL2-Tiny(Wu et al., 2024), InternVL2.5-1B/2B, LoRA-GA(Wang et al., 2024), LoRA+(Hayou et al., 2024), etc. Therefore, how to improve the cost-effectiveness of TableVLM but maintain the great performance is also an important issue. In the future, we will evaluate more multi-modal large language models and select stronger foundation model to obtain higher cost-effectiveness.

8 Ethical Considerations

The multi-modal image-based table recognition dataset and model proposed in this paper are constructed based on public available datasets and models, which are usually free and open, using the MIT license. Based on the above, we carefully designed Python scripts to generate HTML sequence representations according to the original annotations in the dataset and further construct the corresponding multi-modal dataset. The multi-modal dataset will also be an open resource related to multi-modal image-based table recognition and understanding. Therefore, the research in this paper does not involve any ethical issues.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829.

Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. 2019. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*.

Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. 2022. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*.

Hao Feng, Qi Liu, Hao Liu, Wen gang Zhou, Houqiang Li, and Can Huang. 2023. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *ArXiv*, abs/2311.11810.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. *ArXiv*, abs/2402.12354.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Mingshi Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *ArXiv*, abs/2409.03420.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chaochao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024b. Minicpm: Unveiling the potential of small language models with scalable training strategies. *ArXiv*, abs/2404.06395.

Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, and Wei Peng. 2023. Improving table structure recognition with visual-alignment sequential coordinate modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11134–11143.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *ArXiv*, abs/2408.03326.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481.

Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):919–931.

Weihong Lin, Zheng Sun, Chixiang Ma, Mingze Li, Jiawei Wang, Lei Sun, and Qiang Huo. 2022. Tsrformer: Table structure recognition with transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6473–6482.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. 2023. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*.

Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. 2021. Parsing table structures in the wild. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 924–932.

Rujiao Long, Hangdi Xing, Zhibo Yang, Qi Zheng, Zhi Yu, Fei Huang, and Cong Yao. 2025. Lore++: logical location regression network for table structure recognition with pre-training. *Pattern Recognition*, 157:110816.

Nam Tuan Ly and Atsuhiro Takasu. 2023. An end-to-end multi-task learning model for image-based table recognition. *arXiv preprint arXiv:2303.08648*.

Chixiang Ma, Weihong Lin, Lei Sun, and Qiang Huo. 2023. Robust table detection and structure recognition from heterogeneous document images. *Pattern Recognition*, 133:109006.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. 2019. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 128–133. IEEE.

ShengYun Peng, Aishwarya Chakravarthy, Seongmin Lee, Xiaojing Wang, Rajarajeswari Balasubramaniyan, and Duen Horng Chau. 2024. Unitable: Towards a unified framework for table recognition via self-supervised pretraining. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*.

Shoaib Ahmed Siddiqui, Imran Ali Fateh, Syed Tahseen Raza Rizvi, Andreas Dengel, and Sheraz Ahmed. 2019. Deeptabstr: Deep learning based table structure recognition. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1403–1409. IEEE.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, et al. 2024. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*.

Quan Sun, Yuxin Fang, Ledell Yu Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *ArXiv*, abs/2303.15389.

Shaowen Wang, Linxi Yu, and Jian Li. 2024. Lora-ga: Low-rank adaptation with gradient approximation. *ArXiv*, abs/2407.05000.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jian-Yuan Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *ArXiv*, abs/2409.01704.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bing-Li Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yu mei You, Kaihong Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding.

Hangdi Xing, Feiyu Gao, Rujiao Long, Jiajun Bu, Qi Zheng, Liangcheng Li, Cong Yao, and Zhi Yu. 2023. Lore: logical location regression network for table structure recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2992–3000.

Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, and Qingyong Li. 2021. Tgrnet: A table graph reconstruction network for table structure recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1295–1304.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. 2024. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Jiaquan Ye, Xianbiao Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, and Rong Xiao. 2021. Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: table recognition to html. *arXiv preprint arXiv:2105.01848*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multimodal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Team Glm Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Ming yue Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiaoyu Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yi An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhenyi Yang, Zhengxiao Du, Zhen-Ping Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *ArXiv*, abs/2406.12793.

Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, et al. 2023. Tablegpt: Towards unifying tables, nature language and commands into one gpt. *arXiv preprint arXiv:2307.08674*.

Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, et al. 2024a. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*.

Zhenrong Zhang, Pengfei Hu, Jiefeng Ma, Jun Du, Jianshu Zhang, Baocai Yin, Bing Yin, and Cong Liu. 2024b. Semv2: Table separation line detection based on instance segmentation. *Pattern Recognition*, 149:110279.

Zhenrong Zhang, Jianshu Zhang, Jun Du, and Fengren Wang. 2022. Split, embed and merge: An accurate table structure recognizer. *Pattern Recognition*, 126:108565.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Shu Wei, Binghong Wu, Lei Liao, Yongjie Ye, Hao Liu, Wengang Zhou, et al. 2024. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *arXiv preprint arXiv:2406.01326*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*.

Xinyi Zheng, Douglas Burdick, Lucian Popa, and Nancy Xin Ru Wang. 2020. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 697–706.

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

A More details of Multi-Modal Dataset

Due to limited space, we present here the relevant details of the proposed multi-modal dataset, including the composition of the dataset, main characteristics, multi-modal dialogue format, some necessary processing, etc., as follows.

A.1 The Multi-Modal Dialogue Format

Fig. 3 shows the dialogue format used in this multi-modal image-based table recognition dataset, which includes two roles: user and assistant. Among them, the former is responsible for inputting the user's instructions, and the latter is responsible for outputting the HTML representation of the table images. Based on this, we convert table images and the corresponding HTML representations into single-round multi-modal dialogues, and use them to train and evaluate the image-based table recognition capabilities of the multi-modal large language models.

A.2 The detailed description of Multi-Modal Dataset

As mentioned above, the multi-modal image-based table recognition dataset consists of four public available datasets, namely PubTabNet, FinTabNet, SciTSR and TableOCR. They each have their own characteristics. Specifically, the table images in the first three datasets are all flat English tables, but the last dataset contains many distorted table images and its content is in Chinese. In addition, considering that except for TableOCR, the original annotations of the above datasets do not directly contain the HTML representation of the table images, we carefully design the corresponding scripts to generate the corresponding HTML representation according to the original annotations of the dataset, and further removed a small number of erroneous samples or damaged samples to ensure the high quality of the multi-modal dataset. The basic properties of the above datasets are shown in Table 4, some representative samples are shown in Fig. 4, and the methods used to generate the HTML representation of the table images are shown in Fig. 5, 6, and 7.

B Implement Details

In this sub-section, we introduce the details of experimental configuration, including the hardware environment, the hyperparameters used in the training and inference stages, etc. They are shown in Table 5.

C More Experimental Results and Analysis

Due to limited space, we put more detailed experimental results in the appendix as the supplement, including the performance comparison of more table-related methods, the analysis of samples that failed to be recognized, the evaluation of generalization ability, etc., as shown below.

C.1 The performance comparison with other existing table recognition-related methods

In this sub-section, we provide more detailed comparison of TableVLM with various existing table-related methods, including the performance comparison with existing table structure recognition methods, the performance comparison with existing table-related multi-modal large language models, etc. The quantitative results are shown in Table 6.

According to the data shown in Table 6, we can find that TableVLM has excellent performance in table structure recognition, and TEDS-Struct exceeds 95%, which is only 2% lower than the current best method SEMv2, and exceeds the table-related multi-modal large language models such as TabPedia, etc., highlighting its excellent table structure recognition ability. However, in terms of table content recognition, the performance gap between TableVLM and existing multi-stage methods is quite obvious, and further optimization is needed. However, a careful observation of the evaluation results shows that this gap is particularly obvious on PubTabNet, but not on FinTabNet. This is because the quality of the multi-modal data generated by us based on PubTabNet is not good enough, resulting in a decrease in the learning effect of the corresponding model, while FinTabNet does not have this defect. Based on this, we still believe that after improving the data quality, TableVLM can also have an excellent performance comparable to SOTA. In general, it is feasible to construct an end-to-end solution based on multi-modal large language models because they have some irreplaceable advantages, such as compatibility with more different scenarios, less supervision required, multi-language generalization capabilities, etc.

C.2 The Analysis of Failure Samples

In this sub-section, we analyze the defects of the proposed solution, mainly including the following

three points.

Firstly, considering the principle of the multi-modal large language model is to predict the next token, so the basic unit of the output is also the token. But for HTML, the tag pairs will be decomposed into multiple tokens for output. If a token is predicted incorrectly during this process, meaningless output will be generated, which often occurs when recognizing cells with row-column spans, as shown in Fig. 8a. However, after a comprehensive statistical analysis of all recognition results, we find that its probability is low, usually less than 0.1%.

Secondly, the length of the token that can be output by the multi-modal large language model is often limited. As the number of tokens that can be output increases, the performance of TableVLM can gain certain gains. However, when facing large tables, due to the extremely long length of its HTML representation, incomplete output will occur, as shown in Fig. 8b.

Finally, the hallucination of the multi-modal large language model is still an inevitable problem, such as the recognition errors, the expression preference formed in context learning, etc. But considering image-based table recognition is a very rigorous task, even if two sentences express the same meaning, they cannot be replaced arbitrarily, so hallucination will also have an adverse effect on image-based table recognition, as shown in Fig. 8c.

C.3 The comprehensive evaluation of the generalization ability of TableVLM

In this sub-section, we comprehensively evaluate the out-of-distribution generalization capabilities of TableVLM and existing multi-modal large language models. Specifically, we use the above four datasets alternately as training sets and test sets to evaluate their recognition ability for out-of-distribution samples from different dimensions such as structure, content, etc. The quantitative results are shown in Tables 7, 8, 9, and 10.

According to the above results, we can conclude the follows. Firstly, TableVLM-14B still has the best performance, far exceeding other multi-modal large language models. Specifically, for table structure recognition, after fine-tuning on large-scale datasets such as PubTabNet, FinTabNet, etc., it can still obtain excellent results on other datasets, and most of TEDS-Struct exceeds 85%, as shown in Table 7. And for table content recognition, due to the large language model has been pre-trained on large-scale textual datasets, so it has Chinese recognition capabilities even if it is fine-tuned on an English dataset, and vice versa. This is one of the important advantages that the current OCR-related methods do not have. For example, in Table 8, after fine-tuning TableVLM-14B on FinTabNet, its TEDS can still reach 74.78% and TEDS-Struct reaches 87.77% on a dataset named TableOCR that is very different from it, which highlights its excellent generalization ability. In addition, TableVLM-4B has the best cost-effectiveness ratio, but its performance is slightly weaker than TableVLM-14B, close to Qwen2-VL-7B-Instruct, and better than MiniCPM-V-V2.5 and LLaVA-1.5.

Secondly, we find that the performance of the model fine-tuned on a large-scale dataset is better than that of the model fine-tuned on a small-scale dataset. For example, in Table 8 and Table 9, we evaluate the performance of the TableVLM-14B after fine-tuned on FinTabNet on SciTSR, its TEDS reaches 92.47% and TEDS-Struct reaches 95.01%. In contrast, its TEDS and TEDS-Struct are only 72.07% and 78.36%. This is because large-scale datasets usually contain richer and more diverse samples, which enables the model to learn more feature maps with stronger descriptive ability.

Finally, we also find that even if the model is fine-tuned on a dataset containing only flat table images, it still has considerable recognition ability for distorted images, as shown in Table 10. This is because the image-based table recognition is modeled as a sequence generation task, which weakens the negative impact of image distortion. This ability is also one of the advantages that existing multi-stage solutions do not have. Because they often rely on the pixel coordinates of table cells and text blocks for table reconstruction, the changes in pixel coordinates will greatly affect the effect of table reconstruction, but the solutions based on multi-modal large language models do not rely on the pixel coordinates of table-related elements, so they are less affected.

In summary, we conclude that the proposed end-to-end image-based table recognition method based on a multi-modal large language model is an excellent solution with strong recognition and generalization capabilities. It can uniformly model the recognition of table images with different languages and styles as sequence generation, so it is an effective and promising solution that deserves further in-depth exploration by researchers.

Dataset	Image Styles	Language	Data Volumn (Train/Test)
PubTabNet	Flat, no distortion	English	500777/9115
FinTabNet	Flat, no distortion	English	91505/10627
SciTSR	Flat, no distortion	English	11970/3000
TableOCR	Some images are distorted	Chinese	12800/3200

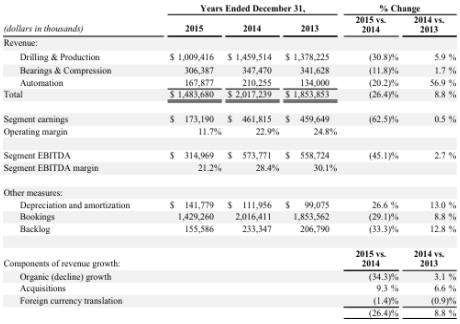
Table 4: The details of the public available datasets used to construct the multi-modal dataset, including image style, language, data volume, etc.

Configurations	Hyperparameters	Details
Hardware Environment	GPU	NVIDIA A800 (80GB)
	CPU	Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz
Training Stage	LoRA-Rank	8
	LoRA-Scaling factor	32
	Dropout Rate	0.05
	LoRA-Modules	The Language Model and Vision Module
	LoRA-Epoches	1 or 3 Epoches
	Optimizer	AdamW
	Initial Learning Rate	1e-4
	Learning Rate Schedule	Cosine Learning Rate Decay Strategy
	Beta1	0.9
	Beta2	0.999
	Weight Decay Coefficient	0.1
Inference Stage	Max New Tokens	2048

Table 5: The details of experimental configuration, including the hardware environment, the hyperparameters used in the training and inference stages, etc.

Categories	Methods	Datasets	TEDS	TEDS-Struct	Acc
Table Structure Recognition Methods	TSRFormer	PubTabNet	-	97.5%	-
		FinTabNet	-	98.4%	-
	VAST	PubTabNet	96.31%	97.23%	-
		FinTabNet	98.21%	98.63%	-
	TabStructNet	PubTabNet	-	90.1%	-
	SEMv2	PubTabNet	-	97.5%	-
	GTE	PubTabNet	-	93.0%	-
Table-Related Multi-Modal Large Language Models	TabPedia	PubTabNet	-	95.41%	-
		FinTabNet	-	95.11%	-
	GPT4V-OCR	SciTSR	-	99.19%	-
Ours	TableVLM-4B	PubTabNet	81.86%	92.65%	59.61%
		FinTabNet	93.50%	96.35%	78.07%
		SciTSR	96.55%	97.38%	81.03%
	TableVLM-14B	PubTabNet	83.90%	95.51%	71.10%
		FinTabNet	95.82%	97.82%	84.94%
		SciTSR	97.89%	98.33%	87.67%

Table 6: The performance comparison of TableVLM and various existing image-related methods. It is necessary to note that '-' represent there is no related data in the original paper.



Years Ended December 31, | % Change

(dollars in thousands)	2015	2014	2013	2015 vs. 2014	2014 vs. 2013
Revenue:					
Drilling & Production	$ 1,009,416	$ 1,459,514	$ 1,378,225	(30.8)%	5.9 %
Bearings & Compression	306,387	347,470	341,628	(11.8)%	1.7 %
Automation	167,877	210,255	134,000	(20.2)%	56.9 %
Total	$ 1,483,680	$ 2,017,239	$ 1,853,853	(26.4)%	8.8 %
Segment earnings	$ 173,190	$ 461,815	$ 459,649	(62.5)%	0.5 %
Operating margin	11.7%	22.9%	24.8%		
Segment EBITDA	$ 314,969	$ 573,771	$ 558,724	(45.1)%	2.7 %
Segment EBITDA margin	21.2%	28.4%	30.1%		
Other measures:					
Depreciation and amortization	$ 141,779	$ 111,956	$ 99,075	26.6 %	13.0 %
Bookings	1,429,260	2,016,411	1,853,562	(29.1)%	8.8 %
Backlog	155,586	233,347	206,790	(33.3)%	12.8 %

Components of revenue growth:	2015 vs. 2014	2014 vs. 2013
Organic (decline) growth	(34.3)%	3.1 %
Acquisitions	9.3 %	6.6 %
Foreign currency translation	(1.4)%	(0.9)%
	(26.4)%	8.8 %

a) Table Image

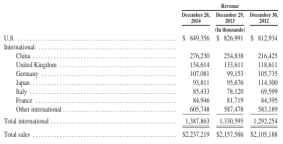
{"conversations":[
{"from": "user", "value": "<img>table_149260.png</img>Please recognize the table in the input image and output its HTML sequence representation without any extra explanation."},
{"from": "assistant", "value": "<table><tr><td></td><td colspan="3">Years Ended December 31,</td><td colspan="2">% Change</td></tr><tr><td>(dollars in thousands)</td><td>2015</td><td>2014</td><td>2013</td><td>2015 vs. 2014</td><td>2014 vs. 2013</td></tr><tr><td>Revenue:</td></tr><tr><td>Drilling & Production</td><td>$1,009,416</td><td>$1,459,514</td><td>$1,378,225</td><td>(30.8)%</td><td>5.9%</td></tr><tr><td>Bearings & Compression</td><td>306,387</td><td>347,470</td><td>341,628</td><td>(11.8)%</td><td>1.7%</td></tr><tr><td>Automation</td><td>167,877</td><td>210,255</td><td>134,000</td><td>(20.2)%</td><td>56.9%</td></tr><tr><td>Total</td><td>$1,483,680</td><td>$2,017,239</td><td>$1,853,853</td><td>(26.4)%</td><td>8.8%</td></tr><tr><td>Segment earnings</td><td>$173,190</td><td>$461,815</td><td>$459,649</td><td>(62.5)%</td><td>0.5%</td></tr><tr><td>Operating margin</td><td>11.7%</td><td>22.9%</td><td>24.8%</td><td></td><td></td></tr><tr><td>Segment EBITDA</td><td>$314,969</td><td>$573,771</td><td>$558,724</td><td>(45.1)%</td><td>2.7%</td></tr><tr><td>Segment EBITDA margin</td><td>21.2%</td><td>28.4%</td><td>30.1%</td><td></td><td></td></tr><tr><td>Other measures:</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Depreciation and amortization</td><td>$141,779</td><td>$111,956</td><td>$99,075</td><td>26.6%</td><td>13.0%</td></tr><tr><td>Bookings</td><td>1,429,260</td><td>2,016,411</td><td>1,853,562</td><td>(29.1)%</td><td>8.8%</td></tr><tr><td>Backlog</td><td>155,586</td><td>233,347</td><td>206,790</td><td>(33.3)%</td><td>12.8%</td></tr><tr><td>Components of revenue growth:</td><td></td><td></td><td></td><td>2015 vs. 2014</td><td>2014 vs. 2013</td></tr><tr><td>Organic (decline) growth</td><td></td><td></td><td></td><td>(34.3)%</td><td>3.1%</td></tr><tr><td>Acquisitions</td><td></td><td></td><td></td><td>9.3%</td><td>6.6%</td></tr><tr><td>Foreign currency translation</td><td></td><td></td><td></td><td>(1.4)%</td><td>(0.9)%</td></tr><tr><td></td><td></td><td></td><td></td><td>(26.4)%</td><td>8.8%</td></tr></table>"}
]}

b) Multi-Modal Dialogue Format

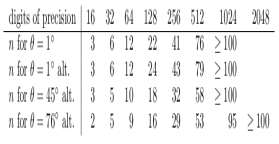
Figure 3: The dialogue template in the proposed multi-modal image-based table recognition dataset.

Table Images

Parameter name	Meaning
NB.FT	Number of fact tables
NB.DIM(f)	Number of dimensions describing fact table *#f*
TOT.NB.DIM	Total number of dimensions
NB.MEAS(f)	Number of measures in fact table *#f*
DENSITY(f)	Density rate in fact table *#f*
NB.LEVELS(d)	Number of hierarchy levels in dimension *#d*
NB.ATT(d,h)	Number of attributes in hierarchy level *#h* of dimension *#d*
HHLEVEL.SIZE(d)	Number of tuples in the highest hierarchy level of dimension *#d*
DIM.SF.ACTOR(d)	Size scale factor in the hierarchy levels of dimension *#d*



	Revenue		
	December 28,2014 (In thousands)	December 29,2013	December 30,2012
U.S.	$849,356	$826,991	$812,934
International:			
China	276,230	254,838	216,425
United Kingdom	134,614	133,611	118,611
Germany	107,081	99,153	105,735
Japan	93,811	95,676	114,300
Italy	85,433	78,120	69,599
France	84,946	81,719	84,395
Other international	605,748	587,478	583,189
Total international	1,387,863	1,330,595	1,292,254
Total sales	$2,237,219	$2,157,586	$2,105,188



digits of precision	16	32	64	128	256	512	1024	2048
n for $\theta = 1°$	3	6	12	22	41	76	≥ 100	
n for $\theta = 1°$ alt.	3	6	12	24	43	79	≥ 100	
n for $\theta = 45°$ alt.	3	5	10	18	32	58	≥ 100	
n for $\theta = 78°$ alt.	2	5	9	16	29	53	95	≥ 100

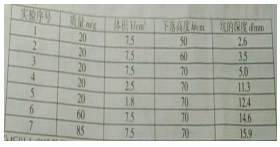
HTML Representations

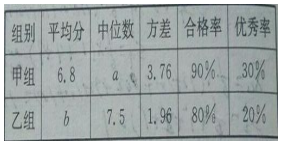
Parameter name	Meaning
N B F T	Number of fact tables
N B DIM (f)	Number of dimensions describing fact table #f
T OT N B DIM	Total number of dimensions
N B M EAS(f)	Number of measures in fact table #f
DEN SIT Y (f)	Density rate in fact table #f
N B LEV ELS(d)	Number of hierarchy levels in dimension #d
N B ATT (d, h)	Number of attributes in hierarchy level #h of dimension #d
HHLEV EL SIZE(d)	Number of tuples in the highest hierarchy level of dimension #d
DIM SF ACT OR(d)	Size scale factor in the hierarchy levels of dimension #d

	Revenue		
	December 28,2014 (In thousands)	December 29,2013	December 30,2012
U.S.	$849,356	$826,991	$812,934
International:			
China	276,230	254,838	216,425
United Kingdom	134,614	133,611	118,611
Germany	107,081	99,153	105,735
Japan	93,811	95,676	114,300
Italy	85,433	78,120	69,599
France	84,946	81,719	84,395
Other international	605,748	587,478	583,189
Total international	1,387,863	1,330,595	1,292,254
Total sales	$2,237,219	$2,157,586	$2,105,188

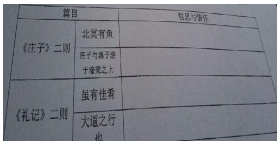
digits of precision	16	32	64	128	256	512	1024	2048
n for θ = 1◦	3	6	12	22	41	76	≥ 100	
n for θ = 1◦ alt.	3	6	12	24	43	79	≥ 100	
n for θ = 45◦ alt.	3	5	10	18	32	58	≥ 100	
n for θ = 78◦ alt.	2	5	9	16	29	53	95	≥ 100

HTML Representations





组别	平均分	中位数	方差	合格率	优秀率
甲组	6.8	a	3.76	90%	30%
乙组	b	7.5	1.96	80%	20%



篇目		暂思与情怀
《庄子》二则	北冥有鱼	
	庄子与惠子游于濠梁之上	
《礼记》二则	虽有佳肴	
	大道之行	

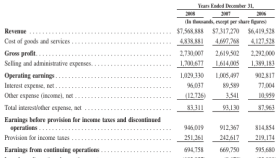
HTML Representations

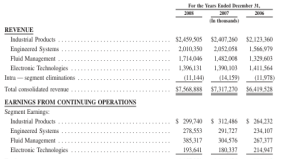
实验序号	质量m/g	体积V/cm^3	下落高度h/cm	坑的深度d/mm
1	20	7.5	50	2.6
2	20	7.5	60	3.5
3	20	7.5	70	5.0
4	20	2.5	70	11.3
5	20	1.8	70	12.4
6	60	7.5	70	14.6
7	85	7.5	70	15.9

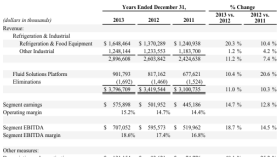
组别	平均分	中位数	方差	合格率	优秀率
甲组	6.8	α	3.76	90%	30%
乙组	b	7.5	1.96	80%	20%

篇目		暂思与情怀
《庄子》二则	北冥有鱼	
	庄子与惠子游于濠梁之上	
《礼记》二则	虽有佳肴	
	大道之行	

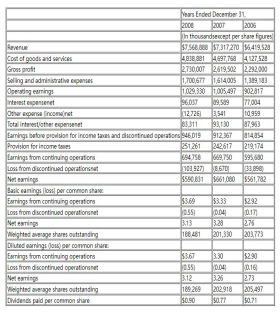
HTML Representations

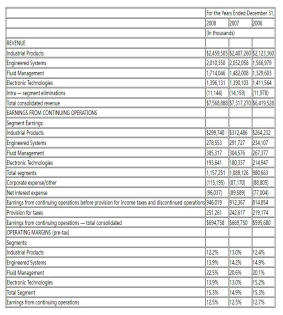






HTML Representations





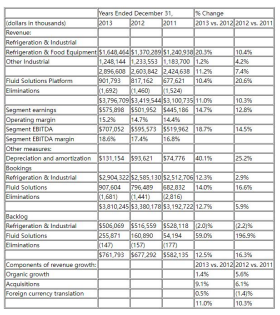


Figure 4: Some samples from the proposed multi-modal image-based table recognition dataset, including some table images with various styles and the corresponding HTML representations.

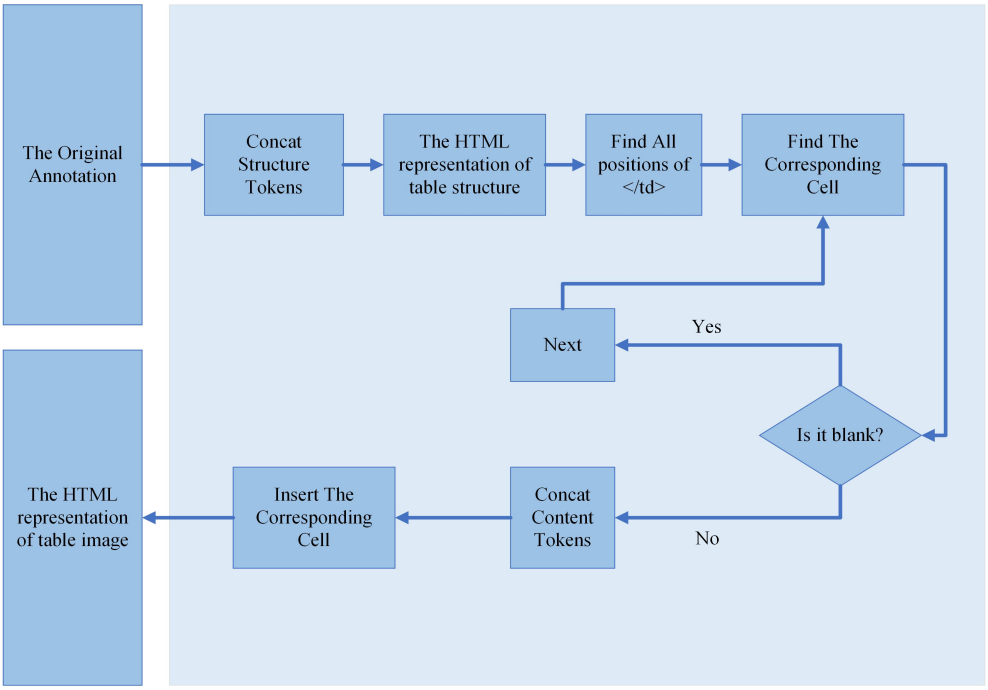


Figure 5: The method to generate the corresponding HTML representation of table images according to the original annotation of PubTabNet.

Models	Train	Test	TEDS	TEDS-Struct	Acc	EDSC	Efficiency
Qwen2-VL-7B-Instruct	PubTabNet	PubTabNet	79.12%	90.94%	54.21%	68.94%	9.53
	PubTabNet	FinTabNet	68.18%	85.40%	32.66%	61.36%	8.21
	PubTabNet	SciTSR	84.04%	94.82%	69.57%	79.37%	10.13
	PubTabNet	TableOCR	43.29%	68.29%	40.44%	64.91%	5.22
LLaVA1.5-7B-Instruct	PubTabNet	PubTabNet	59.31%	74.59%	11.76%	32.66%	8.35
	PubTabNet	FinTabNet	32.10%	62.36%	2.68%	20.64%	4.52
	PubTabNet	SciTSR	45.45%	75.05%	16.47%	25.47%	6.40
	PubTabNet	TableOCR	35.77%	54.73%	5.06%	28.48%	5.04
LLaVA1.5-13B-Instruct	PubTabNet	PubTabNet	60.21%	75.86%	12.86%	33.71%	4.49
	PubTabNet	FinTabNet	31.92%	62.55%	3.00%	20.64%	2.38
	PubTabNet	SciTSR	45.38%	75.12%	16.30%	25.56%	3.39
	PubTabNet	TableOCR	34.41%	57.24%	7.25%	29.77%	2.57
MiniCPM-V-V2.5	PubTabNet	PubTabNet	71.91%	88.03%	39.14%	55.24%	8.46
	PubTabNet	FinTabNet	52.18%	77.92%	19.96%	38.83%	6.14
	PubTabNet	SciTSR	67.95%	88.96%	48.83%	56.85%	7.99
	PubTabNet	TableOCR	45.90%	66.34%	28.50%	40.82%	5.40
TableVLM-4B	PubTabNet	PubTabNet	81.86%	92.65%	59.61%	73.63%	19.49
	PubTabNet	FinTabNet	62.29%	79.13%	26.67%	55.35%	14.83
	PubTabNet	SciTSR	80.88%	90.80%	67.33%	75.40%	19.26
	PubTabNet	TableOCR	45.64%	71.57%	45.88%	41.95%	10.87
TableVLM-14B	PubTabNet	PubTabNet	83.90%	95.51%	71.10%	76.16%	6.04
	PubTabNet	FinTabNet	64.17%	83.94%	36.81%	56.61%	4.62
	PubTabNet	SciTSR	83.24%	95.06%	72.93%	77.81%	5.99
	PubTabNet	TableOCR	53.33%	89.97%	77.41%	48.20%	3.84

Table 7: The comparison of the generalization capabilities of TableVLM and existing multi-modal large language models after fine-tuning on PubTabNet.

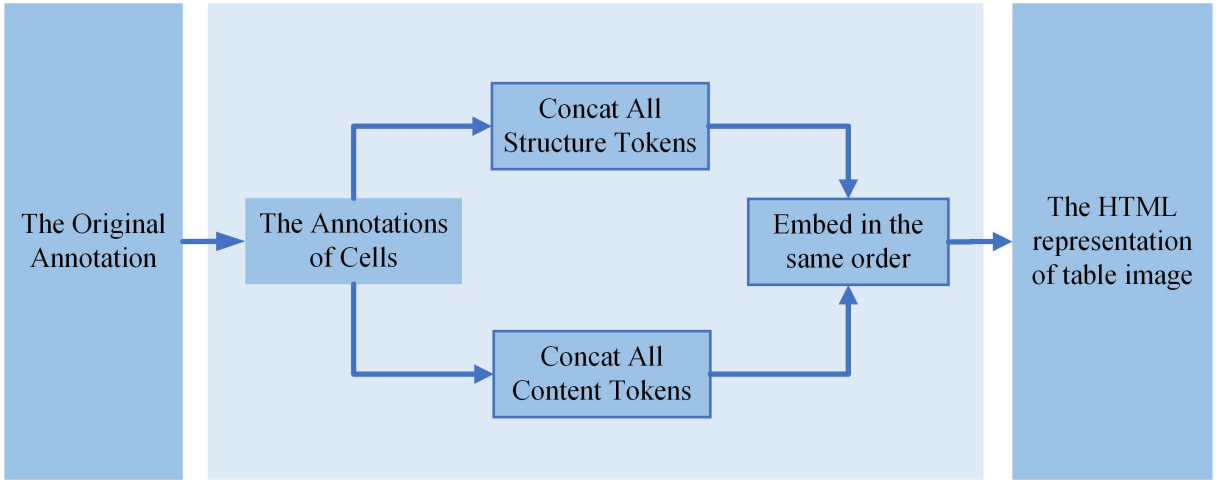


Figure 6: The method to generate the corresponding HTML representation of table images according to the original annotation of FinTabNet.

Models	Train	Test	TEDS	TEDS-Struct	Acc	EDSC	Efficiency
Qwen2-VL-7B-Instruct	FinTabNet	PubTabNet	64.60%	83.67%	36.46%	53.99%	7.78
	FinTabNet	FinTabNet	90.93%	95.55%	75.76%	86.78%	10.96
	FinTabNet	SciTSR	89.47%	92.71%	67.83%	86.28%	10.78
	FinTabNet	TableOCR	72.67%	81.24%	57.31%	75.73%	8.76
LLaVA1.5-7B-Instruct	FinTabNet	PubTabNet	35.74%	59.87%	1.90%	22.75%	5.03
	FinTabNet	FinTabNet	35.07%	71.35%	7.79%	25.75%	4.94
	FinTabNet	SciTSR	25.85%	62.82%	4.27%	19.45%	3.64
	FinTabNet	TableOCR	14.72%	48.09%	1.94%	20.52%	2.07
LLaVA1.5-13B-Instruct	FinTabNet	PubTabNet	36.67%	61.78%	2.12%	22.56%	2.74
	FinTabNet	FinTabNet	37.08%	73.20%	10.08%	26.78%	2.77
	FinTabNet	SciTSR	27.57%	64.38%	5.03%	20.02%	2.06
	FinTabNet	TableOCR	17.36%	52.48%	3.06%	20.96%	1.30
MiniCPM-V-V2.5	FinTabNet	PubTabNet	57.68%	78.73%	22.41%	43.89%	6.79
	FinTabNet	FinTabNet	83.97%	92.31%	59.95%	77.18%	9.88
	FinTabNet	SciTSR	70.14%	82.07%	37.80%	62.93%	8.25
	FinTabNet	TableOCR	51.14%	69.29%	28.28%	51.06%	6.02
TableVLM-4B	FinTabNet	PubTabNet	70.41%	88.48%	44.21%	61.05%	16.76
	FinTabNet	FinTabNet	93.50%	96.35%	78.07%	89.94%	22.26
	FinTabNet	SciTSR	89.89%	93.04%	66.37%	85.22%	21.40
	FinTabNet	TableOCR	56.06%	76.97%	47.69%	56.96%	13.35
TableVLM-14B	FinTabNet	PubTabNet	73.92%	91.44%	53.09%	66.05%	5.32
	FinTabNet	FinTabNet	95.82%	97.82%	84.94%	93.53%	6.89
	FinTabNet	SciTSR	92.47%	95.01%	71.93%	88.75%	6.65
	FinTabNet	TableOCR	74.78%	87.77%	69.41%	73.37%	5.38

Table 8: The comparison of the generalization capabilities of TableVLM and existing multi-modal large language models after fine-tuning on FinTabNet.

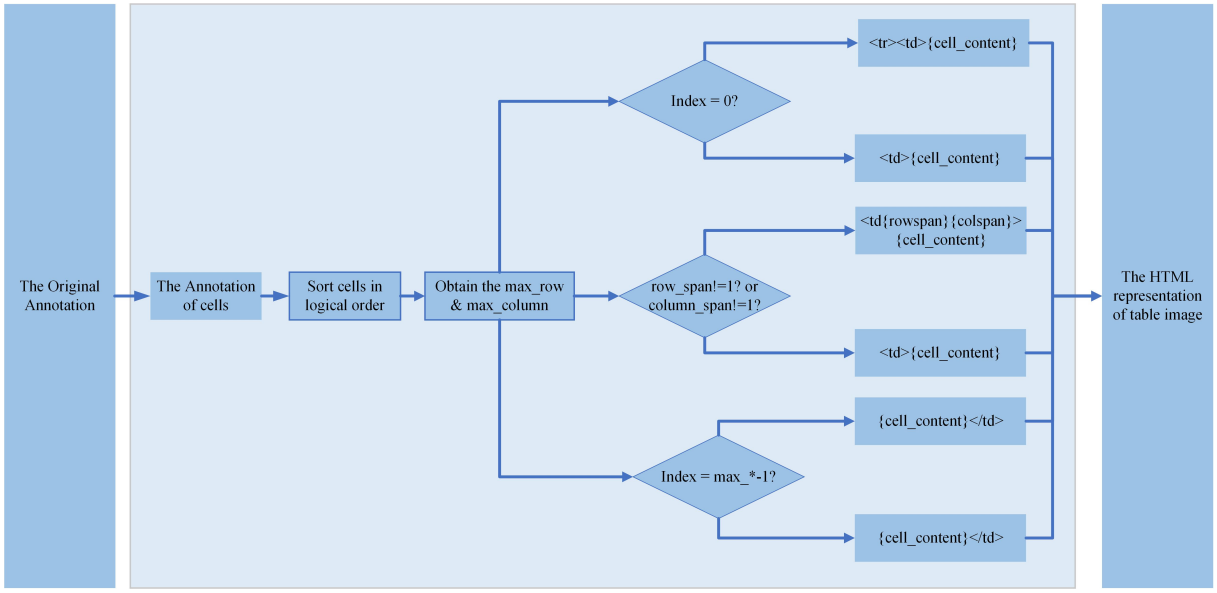
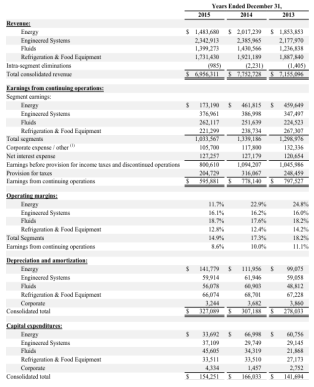


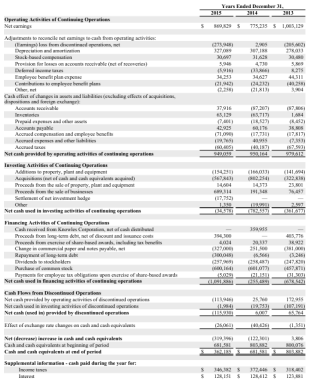
Figure 7: The method to generate the corresponding HTML representation of table images according to the original annotation of SciTSR.

Models	Train	Test	TEDS	TEDS-Struct	Acc	EDSC	Efficiency
Qwen2-VL-7B-Instruct	SciTSR	PubTabNet	63.30%	81.92%	22.16%	48.26%	7.63
	SciTSR	FinTabNet	70.47%	81.26%	21.93%	54.45%	8.49
	SciTSR	SciTSR	96.60%	97.36%	82.57%	95.26%	11.64
	SciTSR	TableOCR	73.38%	81.68%	52.72%	76.78%	8.84
LLaVA1.5-7B-Instruct	SciTSR	PubTabNet	35.93%	60.08%	3.35%	24.65%	5.06
	SciTSR	FinTabNet	17.21%	55.22%	1.42%	19.83%	2.42
	SciTSR	SciTSR	41.19%	73.61%	16.10%	27.64%	5.80
	SciTSR	TableOCR	18.43%	53.43%	5.31%	19.55%	2.60
LLaVA1.5-13B-Instruct	SciTSR	PubTabNet	36.62%	61.90%	3.41%	24.14%	2.73
	SciTSR	FinTabNet	18.01%	55.52%	1.66%	19.19%	1.34
	SciTSR	SciTSR	42.91%	74.79%	18.17%	28.92%	3.20
	SciTSR	TableOCR	18.35%	52.32%	5.59%	19.05%	1.37
MiniCPM-V-V2.5	SciTSR	PubTabNet	60.43%	79.07%	21.37%	46.46%	7.11
	SciTSR	FinTabNet	55.87%	70.63%	11.56%	45.83%	6.57
	SciTSR	SciTSR	88.94%	93.77%	61.53%	85.14%	10.46
	SciTSR	TableOCR	63.60%	72.72%	37.75%	69.79%	7.48
TableVLM-4B	SciTSR	PubTabNet	66.16%	82.91%	29.70%	56.33%	15.75
	SciTSR	FinTabNet	59.28%	68.66%	17.31%	53.47%	14.11
	SciTSR	SciTSR	96.55%	97.38%	81.03%	95.15%	22.99
	SciTSR	TableOCR	53.94%	73.23%	43.28%	59.56%	12.84
TableVLM-14B	SciTSR	PubTabNet	70.96%	88.18%	39.81%	61.59%	5.10
	SciTSR	FinTabNet	72.07%	78.36%	28.28%	60.32%	5.54
	SciTSR	SciTSR	97.89%	98.33%	87.67%	97.02%	7.04
	SciTSR	TableOCR	81.06%	88.04%	70.59%	82.68%	5.83

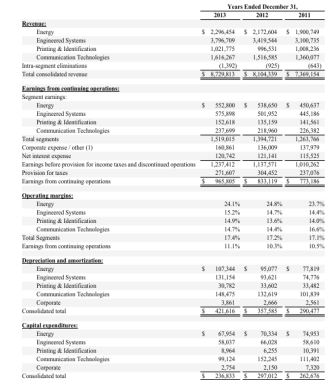
Table 9: The comparison of the generalization capabilities of TableVLM and existing multi-modal large language models after fine-tuning on SciTSR.



a) Meaningless Output



b) Incomplete Output



c) Hallucination

Figure 8: Some failed recognition samples correspond to the three main reasons mentioned above, including meaningless output, incomplete output, language preference, etc.

Models	Train	Test	TEDS	TEDS-Struct	Acc	EDSC	Efficiency
Qwen2-VL-7B-Instruct	TableOCR	PubTabNet	61.37%	74.75%	23.05%	42.39%	7.39
	TableOCR	FinTabNet	59.51%	68.64%	14.57%	43.89%	7.17
	TableOCR	SciTSR	88.97%	92.45%	64.47%	83.66%	10.72
	TableOCR	TableOCR	93.67%	96.26%	86.03%	93.25%	11.29
LLaVA1.5-7B-Instruct	TableOCR	PubTabNet	40.58%	44.75%	1.23%	23.48%	5.72
	TableOCR	FinTabNet	13.07%	37.30%	0.82%	17.07%	1.84
	TableOCR	SciTSR	24.86%	56.02%	4.53%	17.61%	3.50
	TableOCR	TableOCR	39.21%	72.47%	20.13%	31.17%	5.52
LLaVA1.5-13B-Instruct	TableOCR	PubTabNet	40.63%	47.60%	1.39%	24.32%	3.03
	TableOCR	FinTabNet	14.22%	38.67%	0.76%	17.85%	1.06
	TableOCR	SciTSR	26.57%	57.66%	6.00%	19.16%	1.98
	TableOCR	TableOCR	42.72%	76.04%	24.59%	33.21%	3.19
MiniCPM-V-V2.5	TableOCR	PubTabNet	56.79%	67.90%	16.51%	38.70%	6.68
	TableOCR	FinTabNet	42.01%	55.77%	7.84%	33.75%	4.94
	TableOCR	SciTSR	69.75%	79.35%	37.50%	62.24%	8.21
	TableOCR	TableOCR	87.34%	92.66%	72.69%	86.09%	10.28
TableVLM-4B	TableOCR	PubTabNet	61.94%	70.82%	26.08%	48.96%	14.75
	TableOCR	FinTabNet	41.42%	48.33%	7.93%	34.92%	9.86
	TableOCR	SciTSR	67.00%	73.19%	41.33%	61.53%	15.95
	TableOCR	TableOCR	88.22%	95.55%	82.31%	86.97%	21.00
TableVLM-14B	TableOCR	PubTabNet	74.28%	90.60%	51.60%	66.72%	5.34
	TableOCR	FinTabNet	63.76%	70.75%	20.02%	52.89%	4.59
	TableOCR	SciTSR	83.53%	87.28%	59.33%	78.83%	6.01
	TableOCR	TableOCR	97.70%	99.24%	95.56%	97.44%	7.03

Table 10: The comparison of the generalization capabilities of TableVLM and existing multi-modal large language models after fine-tuning on TableOCR.